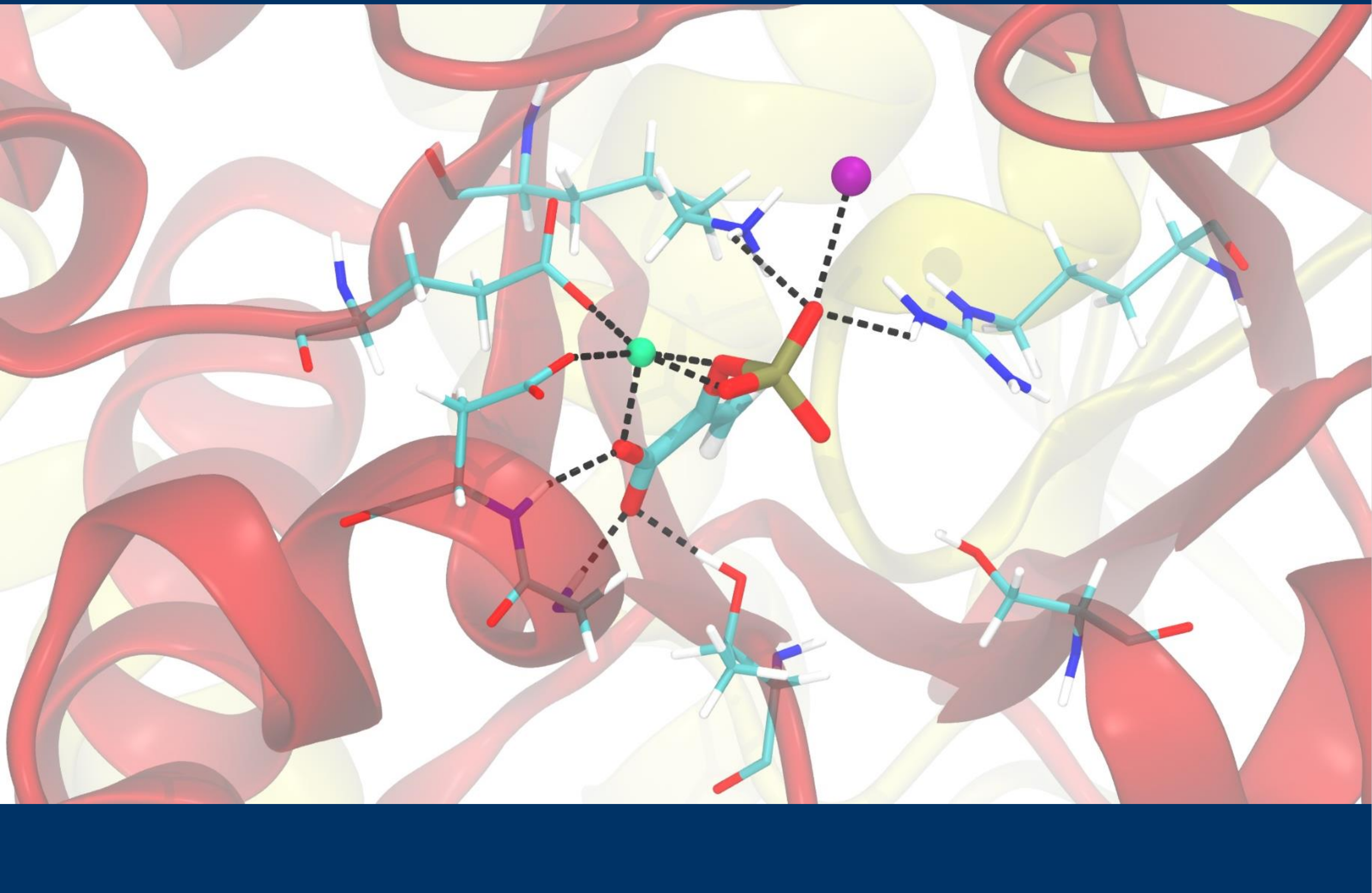
Analysis of consensus motions in proteins through molecular dynamics simulations

Luis Jordà Bordoy

Analysis of consensus motions in proteins through molecular dynamics simulations



Luis Jordà Bordoy

Doctoral Thesis

UNIVERSITAT DE BARCELONA

FACULTAT DE BIOLOGIA

DOCTORAT EN BIOMEDICINA

Analysis of consensus motions in proteins through molecular dynamics simulations

Memòria presentada per Luis Jordà Bordoy per optar al grau de doctor per la Universitat de Barcelona

Tutor i director
Josep Lluís Gelpí Buchaca

Doctorand
Luis Jordà Bordoy

Tesi adscrita al *Departament de Bioquímica i Biomedicina Molecular*
i realitzada a *Barcelona Supercomputing Center (BSC)*

Desembre 2023 – Barcelona

A la Marta i a la Núria

Acknowledgements

En primer lloc, vull expressar el meu agraïment més profund i sincer al Josep Lluís, tutor d'aquesta tesi i el meu mentor en el meu recorregut científic. Em sento molt afortunat per haver pogut treballar aquests anys amb tu. Gràcies per l'oportunitat i per apostar per mi. Gràcies per les estimulants converses posant en comú els reptes i les troballes científiques del projecte, que no han fet més que multiplicar la meva passió per la bioquímica i la bioinformàtica estructural. Entusiasta de la infraestructura informàtica i defensor ferri del seu agermanament amb la ciència, dones eines per integrar aquests dos mons en un de sol més enriquidor. Amb les teves orientacions, m'has ajudat a enrobustir el meu criteri científic i a desenvolupar una intuïció més fina per resoldre problemes. Gràcies per la confiança dipositada en mi, per deixar-me fer i desfer, deixar que m'equivoqui i fer possible que, a vegades, també l'encerti de ple. I, sobretot, gràcies per defensar el costat més humà de la ciència i el món acadèmic, per sempre posar en valor la feina feta, i per espantar els escrúpols quan calia.

Gracias a todos mis compañeros/as del INB con los que he podido coincidir todos estos años. Haciendo memoria, aparecen muchos nombres y caras de todas aquellas personas que han transitado por el equipo. Aunque no he podido compartir tantos ratos con todos como me hubiera gustado, de vosotros me llevo el ambiente tan saludable y ejemplar que convierte al INB en una pequeña comunidad de iguales, que es mucho más que una mera oficina. Gracias por el apoyo que he recibido por vuestra parte en estos últimos meses, por los consejos siempre útiles y sinceros, por todos los chistes y ocurrencias varias, y por las conversaciones de todo tipo, profundas y mundanas. Gracias, Salva, por la determinación con la que logras estimular tantas de estas cualidades en el equipo. Gràcies al team JL més clàssic: Víctor, Dmitry, Romina, Laia, Pau, Genís i Adam. Sempre sereu la primera "foto de grup" que em ve al cap en el meu pas per l'IRB/BSC/INB. Pieter i Raquel, va ser un gust poder acompanyar-vos en els vostres respectius projectes i treballar plegats. Que fàcil que ho vau fer tot, i quanta vida vau aportar en el dia a dia, a vegades solitari, de la tesi. Us desitjo el millor en les vostres respectius recorreguts. Gràcies, Adam, per ser escut i antídot contra la por. Ets i seràs un referent per a mi i molts altres.

Several institutions and individuals have provided academic, infrastructural, and technical support for my thesis. I would like to express my sincere gratitude to the Universitat de Barcelona (UB) for providing me with the opportunity and the resources to pursue my PhD research. I am especially thankful to the Programa de Doctorat en Biomedicina for their administrative support throughout my doctoral studies. Special thanks to the members of my Supervisory Committee, Baldo Oliva and Xavier Daura, for their feedback and their enthusiastic guidance on the project. I would also like to acknowledge the Barcelona Supercomputing Center (BSC) for providing me with the computational infrastructure and the working environment that have been essential for my research. I appreciate the access to the supercomputing facilities from the BSC and the Molecular Modeling and Bioinformatics (MMB) group of the Institute for Research in Biomedicine of Barcelona (IRB), as well as the grants from the Red Española de Supercomputación (RES), which have enabled me to run my simulations and analyses. I also thank the BSC staff for their technical assistance and their friendly attitude.

In addition, I would like to express my gratitude to Pablo Dans for the guidance provided with the parameterization of ions, to Dani Beltrán for his assistance with the construction of the database to upload the simulations, and to Dr. Richard van Wijk for his generosity in sharing variant annotations from personal clinical research.

Por otra parte, gracias a todas las personas que me habéis acompañado de forma directa o indirecta en esta etapa de mi vida. Aunque no puedo representaros a todos/as en este escrito, he llegado hasta aquí porque tengo vuestro apoyo.

Gràcies a la comunitat del Monestir de Sant Pere de les Puel·les, per haver estat casa meva per un temps. Gràcies, especialment, a la Georgina i a les germanes Griselda, Cesca, Catalina i Esperança, per la vostra generositat i per tenir cura de tothom. I gràcies als companys de viatge, pel petit gran miracle de coincidir i compartir l'estada. Entre d'altres, la Mercè, el Benjamín, l'Araceli, el Qays, la Glòria, la Maribel…

Gracias a los de Morat: Blanca, Edu, Sonia, Maria, Cris, Pati, y Borja. Sois mi familia. Gracias al resto de mis grandes amigos del CIM. Gracias a todos mis allegados del Casal Loiola, la Red MAG+S y la Compañía de Jesús. Gràcies, "aniki" Xavi, per ser-hi sempre. Gràcies, Marc, per fer desaparèixer màgicament la distància i el temps compartint la vida. Gracias, Diego, por iluminar el camino, por tirar de mí pacientemente, por ser mi hermano. Gracias, Antonio, por haber llenado de vida y alegría la vida. Sé que intercedes por todos nosotros, y te echamos de menos. Gracias, Tocayo, por tanto (sobran las palabras).

Gràcies als Font Garcia per acceptar-me com un més de la família des del primer moment, i per tot el suport que em doneu.

Gracias a mis padres y a mi hermano, por ser mi hogar y buscar siempre lo mejor para mí.

Gràcies, Marta, per ser el meu pilar, per la teva complicitat entregada fins a l'extrem, i per omplir de felicitat la meva vida.

Abstract

The understanding of proteins as dynamical entities rather than static structures marked a very significant advance in the interpretation of their functional role in life. The capacity of proteins to interact with their environment, sense molecular perturbations and exert responses can be explained in an effective manner by specific dynamical events. The study of proteins from this perspective has been possible in the last decades thanks to the emergence of computational approaches. Among these techniques, Molecular Dynamics (MD) simulations have emerged as a potent tool, playing a pivotal role in investigating conformational transitions at atomic resolution across diverse biomacromolecular systems.

As computational power and infrastructures keep evolving, we are increasingly able to generate longer MD simulations that are capable of capturing dynamical events at biologically relevant timescales. MD simulations typically generate an overwhelming amount of data in the form of a collection of snapshots, called a trajectory. Thus, we need to find suitable metrics to extract, quantify and present the relevant information depending on the target of the study. The scenario is even more challenging when we aim to analyze multiple trajectories and compare their similarity. Among the proposed strategies to explore the comparability between trajectories, essential dynamics analysis (EDA) approaches are a common choice, where Principal Components Analysis (PCA) or other dimensionality reduction techniques are applied to express the differential behavior between trajectories in terms of the underlying collective features of the ensemble.

The work presented in this thesis delves further into this analytical field with the aim of improving the applicability of EDA in functional studies of proteins. The developed approach, termed Consensus Essential Dynamics Analysis (CEDA), introduces a protocol to integrate the information from independent PCAs and derive a consensus set of vectors, the Consensus Principal Components (CPCs). CPCs encapsulate the most representative (consensus) collective motions of an ensemble of trajectories of the system under study, allowing for sharper descriptions and comparisons of its relevant dynamical events. The framework of CEDA also facilitates the comparative study of alternative trajectory ensembles of the same system, in terms of the reference set of CPCs. The outcomes of such comparisons may be interpreted using different data analysis techniques and graphical representations. In this thesis, a strategy was proposed to evaluate the underlying similarities and differences between trajectory ensembles by comparison of their conformational profiles and application of similarity metrics between statistical distributions.

The capacities of the CEDA protocol were demonstrated with the analysis of a collection of MD simulations of human erythrocyte pyruvate kinase (PKR) that covers multiple conditions of the enzymatic complex with its natural ligands, as well as a large array of human genomic missense variants of the protein. Pyruvate kinase is among the most studied proteins from the perspective of biochemistry, given both its role in glycolysis and its paradigmatic and complex set of allosteric properties. This study has provided new support for several of the proposed conformational changes that are associated with the transition between the inactive and active states of the enzyme. Following from the study of the wild-type protein, a second experimental part of the project revolved around the characterization of the functional effects of missense variants of the enzyme. Analysis with CEDA enabled detection of altered dynamical behavior in variants either with a previously validated pathogenic status or for which no functional details were previously known. The conducted research in this regard is presented in depth throughout this manuscript. The obtained results are discussed in the light of the potential application of this protocol in functional studies of proteins in general, and with a particular perspective on pathogenicity prediction studies.

Contents

List of Figures

List of Tables

List of Equations

Abbreviations and acronyms

ACE	Acetyl
ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
BC	Bhattacharyya coefficient
CAGI	Critical Assessment of Genome Interpretation
CEDA	Consensus Essential Dynamics Analysis
CPC	Consensus Principal Component
CV	Collective variable
dbSNP	Single Nucleotide Polymorphism database
DFT	Density Functional Theory
EAM	Ensemble allosteric model
EDA	Essential dynamics analysis
EM	Energy minimization
ESP	Exome Sequencing Project
ExAC	Exome Aggregation Consortium
FBP	Fructose 1,6-bisphosphate
FEL	Free-energy landscape
GAFF	General Amber Force Field
gnomAD	Genome Aggregation Database
GPU	Graphical Processing Unit
GUI	Graphical User Interface
GWAS	Genome Wide Association Studies
H-bond	Hydrogen bond
HF	Hartree-Fock
HGMD	Human Gene Mutation Database
HPC	High-performance computing
ICGC	International Cancer Genome Consortium
IDP	Intrinsically disordered protein
IDR	Intrinsically disordered region
IOD	Ion-oxygen distance
KDE	Kernel Density Estimation
KNF	Koshland-Némethy-Filmer (model of allostery)
LINCS	Linear Constraint Solver
LJ	Lennard-Jones
LOVD	Leiden Open Variant Database
MD	Molecular Dynamics
MEP	Molecular electrostatic potential
MK	Merz-Singh-Kollman (scheme of atomic charges)
MM	Molecular Mechanics
MP2	Second-order Møller-Plesset perturbation theory
MWC	Monod-Wyman-Changeux (model of allostery)
NGS	Next-generation sequencing
NMA	Normal Mode Analysis
NME	N-methyl amide group

NMR	Nuclear Magnetic Resonance spectroscopy
NPT	Constant particles, pressure, and temperature ensemble
NVT	Constant particles, volume, and temperature ensemble
PBC	Periodic Boundary Conditions
PC	Principal Component
PCA	Principal Component Analysis
PCM	Polarizable continuum model
PDB	Protein Data Bank
PEP	Phospho*enol*pyruvate
PGA	Phosphoglycolate
PK	Pyruvate kinase
PKD	Pyruvate kinase deficiency
PKL	Pyruvate kinase; L isoenzyme
PKM1	Pyruvate kinase; M1 isoenzyme
PKM2	Pyruvate kinase; M2 isoenzyme
PKR	Pyruvate kinase; R (erythrocytic) isoenzyme
PME	Particle Mesh Ewald
QM	Quantum Mechanics
RESP	Restrained Electrostatic Potential
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation
SAV	Single amino-acid variant
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SPE	Single-point energy
TCGA	The Cancer Genome Atlas
TIM	Triose-phosphate isomerase
VDW	Van der Waals
WT	Wild-type

Chapter 1 Introduction

1.1 Proteins

Proteins are fascinating components of the machinery of life. The vast majority of biological processes are carried out (or mediated) by proteins, a fact that explains why they have been a permanent object of study in biochemistry since their discovery and progressive characterizations in the nineteenth and twentieth centuries. Etymologically, the term “protein” was coined by Jöns Jacob Berzelius in 1838 from the Greek word πρώτειος (*proteios*) and means “primary” [1] or “standing in front” [2], which reinforces the idea of their capital importance. The range of functions exerted by proteins is spectacularly broad and diverse, including providing structural integrity to cells and tissues, catalyzing chemical reactions, sensing and transmitting stimuli, transporting molecules or even performing mechanical action, to mention a few.

Such a wide range of functions can only be explained by a high degree of structural complexity. Proteins are polymeric chains of smaller molecules called amino acids. The different (and virtually infinite) combinations of the 20 standard amino acids in sequences of varying length confer a rich variety of physicochemical properties. A given amino-acid sequence is, in fact, the principal responsible for a particular protein to achieve its features and capabilities [1]. This statement serves as a primary view of one of the fundamental challenges of biology: *what determines protein function?* However, with this level of description we are barely scratching the surface of the problem. A mechanistic understanding of protein function cannot be derived from sequence information alone.

A far better and more comprehensive view is achieved when we take notice of how proteins operate from the perspective of their three-dimensional structure and dynamics. After all, it is only intuitive that biological functions such as the aforementioned cannot be carried out in the absence of a proper architecture and movement patterns. Structure enables function, and dynamics ultimately governs it [3]–[7]. Both features are the embodiment of the information inherently encoded in sequence that has accordingly been refined through evolution [8], [9]. The capacity of understanding and studying proteins from this perspective has been possible in the last decades thanks to the emergence of structural resolution techniques and more recently the use of computational approaches to simulate dynamics. These are employed in the disciplines of structural biology and structural bioinformatics to solve problems in biology and create new knowledge by analyzing the structural and dynamical behavior of biological macromolecules [10]. The following section elaborates on the subjects of protein structure and dynamics, with special focus on the latter.

1.1.1 Protein structure and dynamics

Under biological conditions, most known proteins exhibit compactly folded states under the influence of internal non-covalent interactions and forces (Figure 1.1). The information that drives the folding process is inherently encoded in the amino-acid sequence, also called the primary structure of proteins. For instance, in soluble proteins, hydrophobic amino acids tend to form densely packed cores that make up the inner part of the three-dimensional structure. This phenomenon is due to the so-called hydrophobic effect, which can be considered the principal agent of the folding process. The rest of weak interactions cooperatively guide the local folding of protein segments into fundamental units of specific structural arrangement, the secondary structure elements, such as α-helices and β-strands. Secondary structure elements subsequently combine to form stable larger folding patterns, called motifs and folds, that constitute the tertiary structure. In addition, many proteins also exist as assemblies of multiple polypeptide chains (in this context, also called subunits or monomers) that associate through specific networks of non-covalent interactions in order to achieve their assembled functional forms (called complexes or multimers). This last level of arrangement is also known as the quaternary structure of proteins. Such an organized architecture confers the needed structural integrity and specificity to participate in recognition events with other molecules involved in biological processes (such as small ligands or nucleic acids) or simply to serve pure structural roles such as fibrous or coat proteins [6].

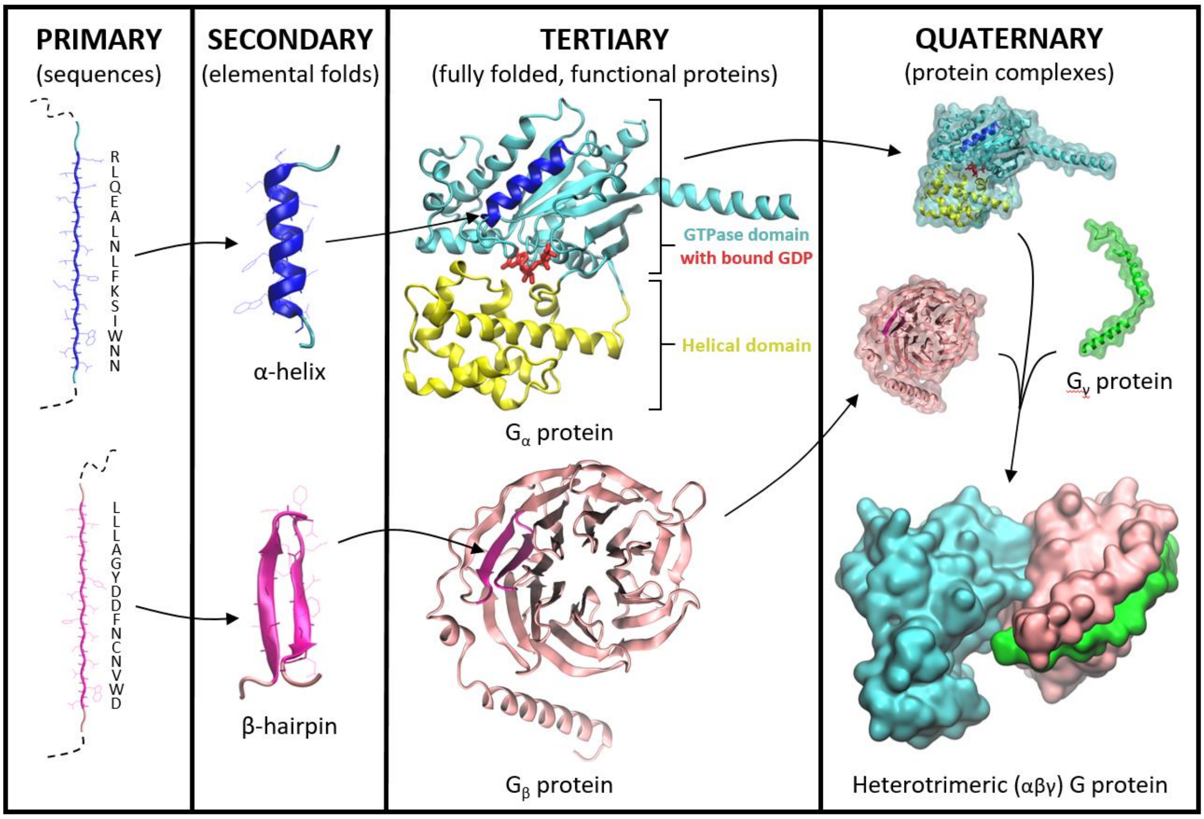


Figure 1.1. Levels of the protein structure. The concept is illustrated with the example of the structure of a heterotrimeric G protein (Protein Data Bank entry 6EG8 [11]). Primary level: a fraction of the amino-acid sequences of the G_α and G_β subunits is shown in one-letter code next to their corresponding non-folded segments. Secondary level: the segments fold into an α-helix and a β-hairpin (two antiparallel β-strands). Tertiary level: the fully folded G_α and G_β subunits. The particular location of the previously shown secondary structure elements is indicated. The G_α subunit possesses two domains that have been highlighted in different colors. The GTPase domain, colored in cyan, displays a characteristic tertiary structure motif called Rossmann fold in the form of a 3-layer "aba" sandwich. The helical domain, colored in yellow, displays an orthogonal bundle motif of α-helices. A molecule of GDP, colored in red, binds to the cleft between these two domains. The G_β protein displays a 7-bladed β-propeller motif. Quaternary level: the G_α, G_β and G_γ proteins can associate to form the functional protein complex of the heterotrimeric G protein. NOTE. Images generated with the software VMD [12].

Dynamics is intimately related to structure. Extensive evidence indicates that the more we understand the structure of a protein, the more insight we gain about its potential dynamical traits (and how they relate to its function). For instance, it is increasingly clear that families of proteins with similar folds also share similar modes of motion [9], [13]. The complex architecture of proteins is often described as a set of distinguishable modules with mechanistic and functional implications, called domains. The concept of *domain* does not have a single universally accepted definition, however it usually refers to a region of the protein that i) is independently stable and often folds in recognizable motifs, ii) has been conserved as a sequence block through evolutionary pathways, iii) may undergo movements as a single entity with respect to the rest of the protein, or iv) fulfills a particular function [6], [10], [14]. In other words, it represents a fundamental unit of evolutionary, structural, dynamical and functional significance. Proteins can consist of just a single domain or otherwise be modularized in several domains. In the latter case, domains may be visually discernible as distinct globular lobes separated by hinges or rather be more intimately attached through extensive contact interfaces. Such structural complexity results in a combination of rigid and flexible regions with the potential to undergo certain motion patterns.

The dynamical nature of proteins remained elusive for years. In fact, we still lack the means of observing the motion of macromolecules in real time with the current technology. With the advent of X-ray crystallography in the late 50s, the scientific community was able to get a first glimpse of the dynamical properties of biological macromolecules and began building new knowledge on top of this notion. X-ray crystallography revealed the existence of structural regions (even whole proteins) that are hard or unable to be resolved due to disperse electronic density (*i.e.*, they possess high mobility). This technique also provides a rough measure of structural flexibility, the B-factors [5]. Remarkably, the role of dynamics in catalytic mechanisms was validated thanks to the crystallographic data of lysozyme both in its apo and holo conditions, which revealed diverse structural arrangements for the same protein [15], [16]. Later, Nuclear Magnetic Resonance (NMR) spectroscopy was able to reaffirm this view. This technique enables the resolution of the atomic displacement of a structure over time. Subsequent comparisons between the resulting static structures help infer the potential motions that occur in timescales from picoseconds to seconds. Other more recent structure resolution techniques (see section 1.4.1) have been providing further insight in this regard since then.

Nevertheless, for the time being, we resort to computational simulation methods (see section 1.4.2.2) to obtain more accurate descriptions of dynamical behavior, with molecular dynamics (MD) simulations being the most accepted and powerful approach [17]. As a result, through the years we have gathered irrefutable evidence that dynamics is instrumental for protein function. The capacity of proteins to interact with their environment, sense molecular perturbations and exert responses can be explained in an effective manner by specific dynamical events. While this is clear in proteins that serve as molecular motors, which explicitly convert chemical energy into mechanical work [18], it is also true in many other protein functions that were traditionally believed to depend only on static structures. Allosteric regulation of enzymes is facilitated by their ability to sample alternative structural forms [19], [20]. Protein flexibility can adapt the shape of binding pockets to regulate the binding process, and even trigger the formation of a completely new binding pocket [21]. The catalytic event in enzymes can be triggered by explicit loop or domain closures that isolate the active site from solvent and provide an optimal chemical environment [22], [23]. Protein regions cataloged as loops and linkers can exert a variety of communication roles between different functional modules of multidomain proteins via sequential dynamical effects [24]. Changes in the flexibility profiles of transmembrane proteins act as switches to propagate biological information through the cell membrane [25]. These and many other examples have been decisive in abandoning the simplistic view of proteins as single structures and adopting a more realistic model that considers their dynamical nature as a powerful source of research [5], [9], [19], [26], [27].

Protein motions occur in a spectrum of time and length scales. Broadly speaking, we can distinguish between two major kinds of dynamics: local and global motions. Local motions correspond to fluctuations of small amplitudes, involving a few atoms, that generally take place in the faster timescales. For instance, bond vibrations at the femtosecond timescale, rotations of amino-acid side chains at the picosecond-nanosecond timescale or rearrangements of small regions (*e.g.*, loops or helices) at the nanosecond timescale. On the other hand, global motions refer to larger structural rearrangements that usually require longer time spans, from collective motions involving domains and subunits (nanosecond-microsecond timescale) to the folding process of whole proteins from their unstructured state (millisecond-second timescale and even beyond). Dynamical events taking place in the time range between picoseconds and milliseconds are usually the main target of protein dynamics studies, leaving behind faster atomic vibrations that are negligible. Examples of different biologically relevant motions are shown in Figure 1.2.

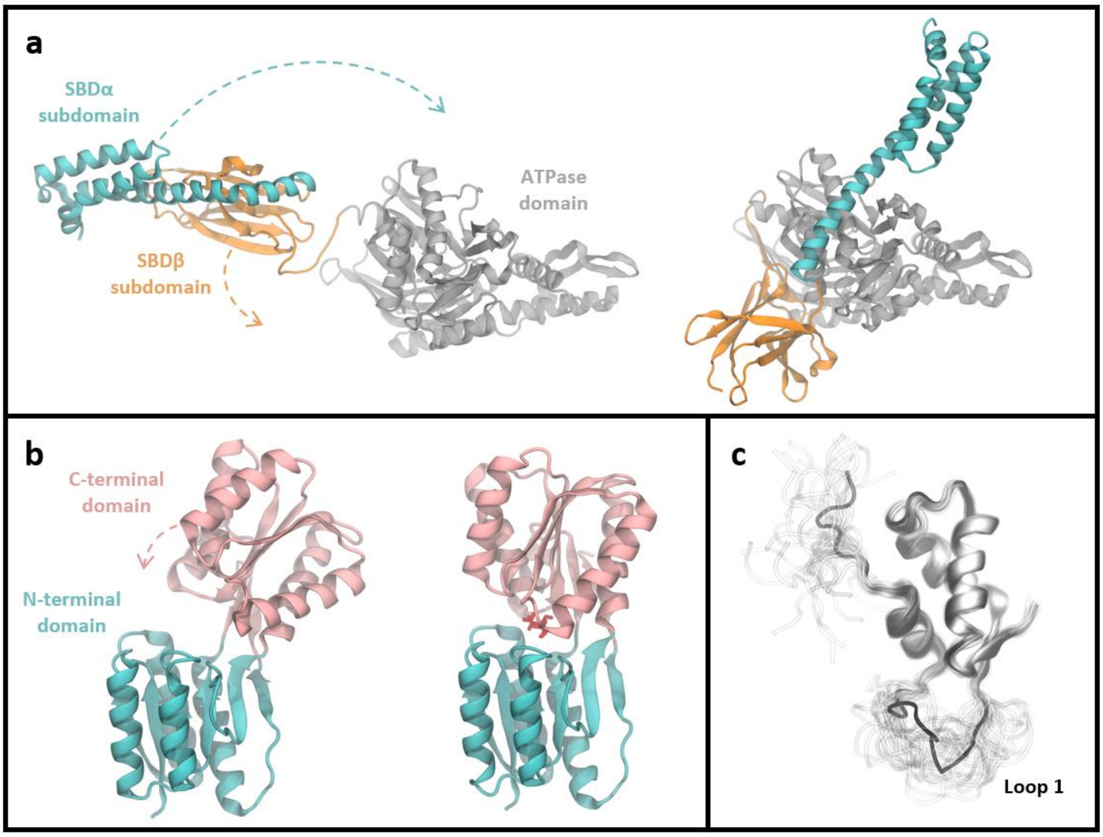


Figure 1.2. Examples of biologically relevant motions of proteins. (**a**) Large global motions of Hsp70. Hsp70 proteins are an important part of the cellular machinery for protein folding, performing chaperoning functions, and helping to protect cells from the adverse effects of physiological stresses. Left: closed state of the substrate binding domain (SBD) of Hsp70, with subdomains SBDα (cyan) and SBDβ (orange) interacting (Protein Data Bank entry 2KHO). Right: open state of the SBD, with SBDα and SBDβ detached from one another and docked to different faces of the ATPase domain (gray) as a consequence of ATP binding (Protein Data Bank entry 4B9Q). (**b**) Hinge motions of RBP. Ribose-binding protein (RBP) is a periplasmic binding protein involved in bacterial intercellular communication systems that mediate transport and chemotaxis. Left: open state of RBP in the absence of ligand (Protein Data Bank entry 1URP). Right: closed state of RBP, with the bilobal structure capturing the ligand with a hinge motion (Protein Data Bank entry 2DRI). (**c**) Local fluctuations of the C-terminal domain of the bacterial 50S ribosomal protein L11. This domain of the protein (Protein Data Bank entry 1FOX; 33 structures solved by NMR) binds tightly to a highly conserved 58 nucleotide domain of the 23S ribosomal RNA. The conformational dynamics are thought to play an important role in the binding process. Loop 1 is flexible and disordered in the RNA-unbound form; it fluctuates in a picosecond-nanosecond timescale. In the RNA-bound form, it rigidifies and adopts a specific conformation as a result of its direct contact with RNA. NOTE. Images generated with the software VMD.

The spatial arrangements that a protein can adopt are called conformations. Once in their functional folded state, proteins still possess internal degrees of freedom that allow the structure to undergo changes of conformation without altering the fold. Such an array of conformations is known as the native conformational ensemble, where the predominant conformations are usually the ones that are energetically more stable (*i.e.*, they possess the lowest values of free energy, symbolized as G) [6], [19]. The study of protein dynamics often aims to characterize both the kinetic and thermodynamic perspectives of this scenario, in other words, the quantification of the energy barrier (ΔG) associated to the transitions between particular conformations and the relative probabilities of their respective equilibrium populations [4].

In this regard, we can make use of a simple mathematical definition to fully describe and characterize protein dynamics. All possible conformations of a protein can be represented as points in a topological space where dimensionality corresponds to the number of atomic Cartesian coordinates (or to another metric with geometric information of the structure). Consequently, any conformational change, either local or global, can be depicted accordingly as a shift in spatial position. Such a space, infinite in its mathematical definition, is not explorable in its entirety in practice. There are regions in this space that would imply physically impossible twists, elongations, overlaps… These unrealistic structures are forbidden because they would lead to bond breakage or simply be absurdly penalized in energetic terms. Conversely, the allowed portion of space complies with the energetic constraints dictated by the laws of physics. In turn, this implicitly means that there are particular regions which are actually preferred in energetic terms. Following this rationale, the described scenario is equivalent to considering a continuous function that can be applied to the topological space so that every point (*i.e.*, every conformation) is associated with a particular energy value that informs about its relative stability. This is the concept of a free-energy landscape (FEL) [4], [26], [28], [29].

FELs are highly multidimensional, since proteins have many atoms and therefore there is an overwhelming number of degrees of freedom. In practice, we build low-dimensional FELs (with energy as a function of 1D or 2D conformational spaces) in order to gain representability and interpretability; often with pedagogical or qualitative purposes as in Figure 1.3. FELs are rugged surfaces, where valleys and hills (the local minima and maxima of the function) correspond respectively to metastable states and to the energy barriers (or the transition states) between them. The explorable landscape is vast: it does not only include the native conformational ensemble, but also the countless unstructured states of the unfolded protein, misfolded states and possible folding intermediates. It is organized in a hierarchy where energy valleys are, in turn, composed of smaller energy valleys, and so on [30]. This topological model successfully illustrates the dynamical nature of the range of possible local and global motions and their probability of occurrence. For instance, it helped solve the puzzle of how proteins can achieve their correct folded states without randomly sampling all the possible conformations (Levinthal's paradox [31]). Following the hypothesis of the folding funnel [32], [33], the FEL looks like a funnel (Figure 1.3a), where the outermost surface corresponds to the unfolded subspace and the native conformational ensemble is located at the bottom of a major deep well. The steep walls of the well provide multiple pathways that rapidly direct the unfolded protein downhill until reaching the states of minimum energy.

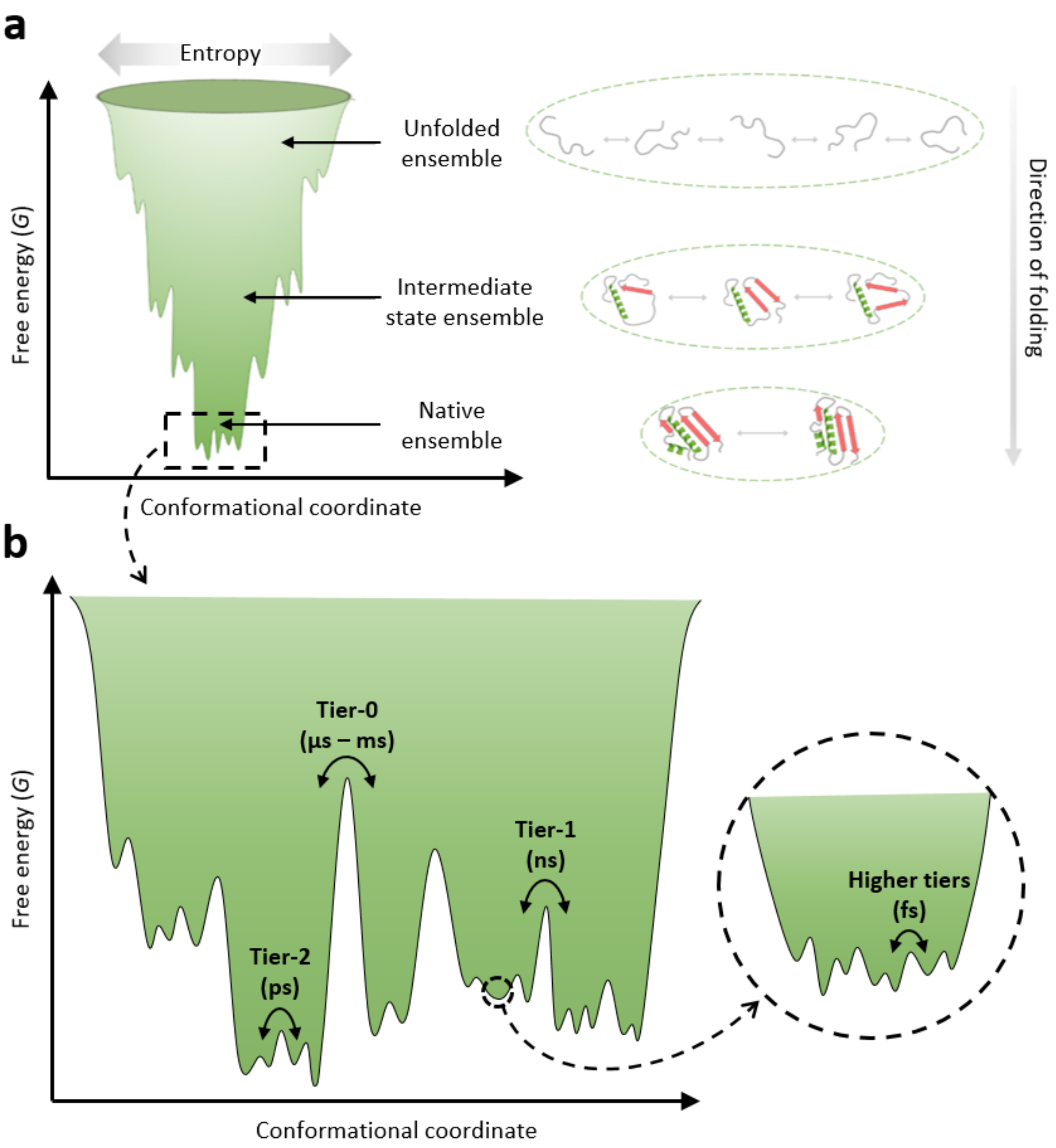


Figure 1.3. The free-energy landscape (FEL) in protein folding and dynamics. Example of the FEL of a hypothetical protein with the free energy as a function of a one-dimensional conformational coordinate. (**a**) The folding funnel. The unfolded protein can fold to its native conformation via multiple intermediates by following different routes. As the protein travels downhill in the free-energy landscape, there is a fall in conformational entropy and energy. (**b**) Dynamics of the native conformational ensemble. A progressive close-up of the native conformational ensemble reveals the valleys and hills that correspond to different conformational (sub)states and the energy barriers between them. The different tier levels of the hierarchical nature of protein dynamics [4] encompass the transitions that occur in a spectrum of timescales according to the height of the corresponding energy barrier. NOTE. (a) was adapted from [34].

The native conformational ensemble looks like a basin with multiple valleys that correspond to the conformational subpopulations that exist in equilibrium. The width and ruggedness of this level of the FEL determine the degree of conformational entropy or heterogeneity of the functional protein. Again, the topology is hierarchical: the larger valleys correspond to the major stable conformations that also inherently consist of spectrums of minor conformational substates (Figure 1.3b) [26], [34], [35]. Native proteins are constantly sampling the available space, hopping from valley to valley, undergoing transitions between conformational states and substates. Transitions mainly occur by virtue of thermal motion, which induces structural shaking and eventually provides the necessary energy to reach the up-and-down path that leads to a different valley nearby. Energy barriers can be expressed as multiples of the term $k_B T$, where k_B is the Boltzmann constant and T is the temperature of the system. The height of energy barriers between wells is correlated with the timescale at which conformational transitions take place.

In 2007, Henzler-Wildman and Kern [4] synthesized the theoretical and experimental knowledge of protein dynamics available at the time. The contents of their work have been extensively reproduced since then to illustrate the nature of dynamical events based on their timescale. More specifically, they provided a referential tier classification of dynamical processes that is illustrated in Figure 1.3b and summarized here as follows. Slow dynamical events (tier-0) correspond to transitions between kinetically distinct conformations separated by energy barriers of several $k_B T$ at physiological conditions. These occur with relatively low frequency, at the timescale of microseconds-milliseconds and slower. Typically, the transitions in this category are global motions such as collective motions of whole domains, usually involved in biologically relevant processes. Within a tier-0 state, the protein fluctuates in a faster timescale between the more closely related conformational substates that compose the major state. These fast dynamical events are separated by smaller energy barriers (often less than 1 $k_B T$ at physiological conditions) that occur with higher frequency at the picosecond-nanosecond timescale. The motions involved in this category of transitions are usually local, with small amplitudes, such as the collective fluctuation of a loop or a backbone region (tier-1) or side-chain rotations (tier-2). Bond vibrational modes would be encompassed in higher tiers, at the femtosecond timescale.

Following from the kinetic description of dynamics, it is clear that temperature has a critical impact on the shape of the FEL. However, the FEL is substantially susceptible to various other physicochemical factors such as pressure and solvent conditions (ionic strength, pH…). In fact, some studies sustain that the role of the solvent must be considered primarily in order to obtain a complete view of the hierarchical nature of dynamical events. In those models, water dynamics (the bulk solvent and the hydration shell surrounding the protein) would be the dominant agent of protein dynamics [29], [30]. Be that as it may, all evidence supports that the physicochemical environment determines the profile of the FEL, and that any perturbation would modify the energetic stability of the states and/or their rate of interconversion [4], [8], [36]. Whenever a perturbation occurs (*e.g.*, a chemical reaction or a temperature or pressure jump) the conformational equilibrium of the ensemble is altered. The system then undergoes a relaxation process from the generated non-equilibrium state towards the new equilibrium dictated by the reshaped FEL [26].

Far from being incidental, the reshaping of the FEL as a response to a perturbation is a key functional mechanism of the vast majority of proteins. Evolution has exploited this capacity and has given rise to sophisticated conformational regulation mechanisms that are able to tune protein activity [8]. Through this perspective, virtually every known process mediated by proteins can be explained. External stimuli often involve direct interactions with other biomolecules (cofactors, small ligands, substrates, and/or biomacromolecules). Other purpose-specific proteins rely on particular environmental changes, such as temperature or electric potential shifts. All these considerations are intimately linked to the concept of allostery, a topic of central interest in protein research that is covered in the following section.

1.1.2 Allosteric proteins

Allostery is a fundamental property of proteins, being one of the major means of functional regulation with which the organism achieves a careful calibration of physiological activity, adapting to the needs at any given time and in an orchestrated manner. Many biological processes require allostery, most prominently signal transduction, molecular machine function, transcriptional regulation and metabolism [37], [38]. Consequently, failure of allosteric control leads to malfunction and is associated with many diseases [39], [40]. Due to its capital role, it is no wonder that allostery is one of the most widely studied properties of proteins [41], and was even referred to as "the second secret of life" (second only to the genetic code) back in the 1960s when Jacques Monod and coworkers were presenting pioneering formulations on allosteric regulation [42], [43].

Essentially, allostery refers to the process by which a perturbation at one site of a biological macromolecule transmits an effect to a physically distinct site of it through alteration of shape and/or dynamics [40], [44]. The allosteric stimulus generally consists in the non-covalent binding of a ligand molecule (either small or large), called the allosteric modulator or effector. However, in some contexts the concept also applies (albeit with controversy [43]) to other localized events such as covalent modifications (phosphorylation, formation of disulfide bonds, glycosylation…) or light absorption [39], [40], [45], [46] because they function analogously to the mechanism described for ligand binding. The definition has been further exploited to include as allosteric perturbations particular changes in the physical environment (temperature, pH, ionic concentration…) [46], [47] or even to refer to the functional consequences of point mutations [40], [46], [48]. In such cases, this terminology may be specifically employed when there are noteworthy similarities or implications with respect to the general conceptual framework of the phenomenon. For instance, some mutations in G-protein–coupled receptors can emulate the modulation effects of their natural allosteric ligands [49].

Allosteric effectors may be either allosteric activators when they enhance protein activity, or allosteric inhibitors when they suppress or reduce it. The degree of complexity of allosteric proteins is diverse, depending on the number of agents that take part in the regulation and their mechanism of action. A relatively simple allosteric protein may have only one or two modulators. For instance, aspartate carbamoyltransferase is activated by ATP and inhibited by CTP and both ligands bind competitively to the same allosteric site. In contrast, other proteins may combine several allosteric mechanisms that turn them into systems of considerable regulatory complexity. For instance, bacterial glutamine synthetase is among the most complex regulatory enzymes known, with at least eight different modulators including reversible covalent modification and the association of other regulatory proteins [6].

Over the years, the knowledge gathered shows that all dynamical proteins have potential allosteric faculties [38], [47], [50], [51], thus leaving behind the idea that allosteric regulation is exclusive to certain enzymes or processes. In turn, this means that, potentially, many allosteric systems remain unknown. This notion has opened up the possibility to apply high-throughput screening experiments with the aim of discovering new allosteric sites and modulators by what we could call an "accelerated serendipity". Some artificial ligands can bind to latent allosteric sites (*i.e.*, binding pockets without natural ligands [52]) and selectively stabilize a protein conformation that possesses the desired biological activity. For instance, a synthetic compound named mitapivat binds to a latent allosteric site of erythrocytic pyruvate kinase and stabilizes its active conformation, enhancing its enzymatic activity.

This drug has been recently approved to treat hemolytic anemia in patients with pyruvate kinase deficiency [53]–[55]. Thus, the remarkable omnipresence of allostery provides promising opportunities from the perspective of pharmacological research. We are experiencing an explosive growth in the number of approved allosteric drugs [56] that clearly demonstrates their prominent advantages: they can cooperate with other endogenous or exogenous ligands, they offer potentially enhanced selectivity and reduced toxicity, they can target proteins that are difficult to reach pharmacologically, and they have the potential to combat drug-resistant mutations located at orthosteric sites [40], [56]–[58].

1.1.2.1 The classical views of allostery

The first formulations of allostery as such are attributed to the works of Jacques Monod, Jean-Pierre Changeux and François Jacob [59]–[61]. However, decades before their elaborations on the phenomenon and their efforts to build a plausible model to understand it, some studies had described prior observations of protein activity regulation processes. Remarkably, Christian Bohr discovered that the affinity of hemoglobin for oxygen decreased with higher concentrations of carbon dioxide and, consequently, acidic pH (the so-called "Bohr effect") [62]. Later, Adair related this behavior to the tetrameric structure of the protein [63] and Pauling proposed the first structural model to explain the positive regulation of the binding of oxygen molecules to hemoglobin [64]. Gerty and Carl Cori described how glycogen phosphorylation was regulated by the concentration of adenosine monophosphate [65]. The works of Novick and Szilard [66], Umbarger [67], and Yates and Pardee [68] provided definitive proof of the existence of feedback inhibition mechanisms in the biosynthesis of amino acids and nucleotides, where the end product inhibited the initial steps of the enzymatic pathway without sharing structural similarity with the substrates. All the gathered evidence led to an incipient need to propose a new model of activity regulation in proteins [49].

The terminology that Monod and Jacob chose to refer to the phenomenon was "allosteric inhibition" [59]. The term "allosteric" derives from the Greek words ἄλλος (*allos*), meaning "other," and στερεός (*stereos*), meaning "solid" or "shape." This etymological construction historically has had two possible interpretations [49]. The first is a reference to the fact that, in allosteric regulation mediated by ligand binding in enzymes, the shape of the allosteric effector may be different from that of the enzymatic substrate(s). Conversely, the second interpretation does not refer to the ligand but to the fact that the regulatory site of an allosteric protein lies in a separate region from the functional site and consequently also has a different shape. Even though both interpretations are practically equivalent, the latter best encompasses the formal definition of allostery as we understand it nowadays.

The first models of allosteric regulation were developed in the mid 1960s. Both the rationale and terminology employed were specially influenced by the extensive studies that had been performed on a few proteins, most notably hemoglobin. This protein has indisputably been the prototype system for the investigation of functional regulation in macromolecules [69]. Over those years, the term "cooperativity" had been used to refer to the kinetic profile displayed by proteins in which ligand binding alters the affinity for subsequent binding of the same ligand. In positive cooperativity the affinity increases and in negative cooperativity the affinity decreases. On another note, the first crystallographic structures of proteins (myoglobin in 1958 [70] and hemoglobin in 1960 [71]) allowed the models of allostery to include speculations underpinned by the observations made on the structural data. In fact, a new view of the mechanism of molecular recognition had recently arisen to address the question of conformational change of the protein upon ligand binding that had been raised by evidential findings. For instance, it was seen that the binding of the oxygen molecule to the heme group in hemoglobin required some structural rearrangements; otherwise the binding site was hardly accessible just by regular diffusion [6], [71]. In this sense, the "induced fit" model of Koshland [72] was gaining more and more acceptance as a replacement to the older "lock-and-key" model. The "lock-and-key" model (Figure 1.4a) had been introduced by Fischer in 1894 [73] and hypothesized that the substrate was recognized by the enzymatic binding site through shape complementarity, using the analogy of a key in a lock, thus neglecting any dynamical contributions of the binding process and relying solely on chemical specificity. Conversely, the "induced fit" model (Figure 1.4b) suggests that the binding of a ligand induces a local conformational change in the binding interface so that the shapes of both bodies are adjusted to provide the optimal fit and form the final complex [72]. The analogy in this case is like a hand in a glove.

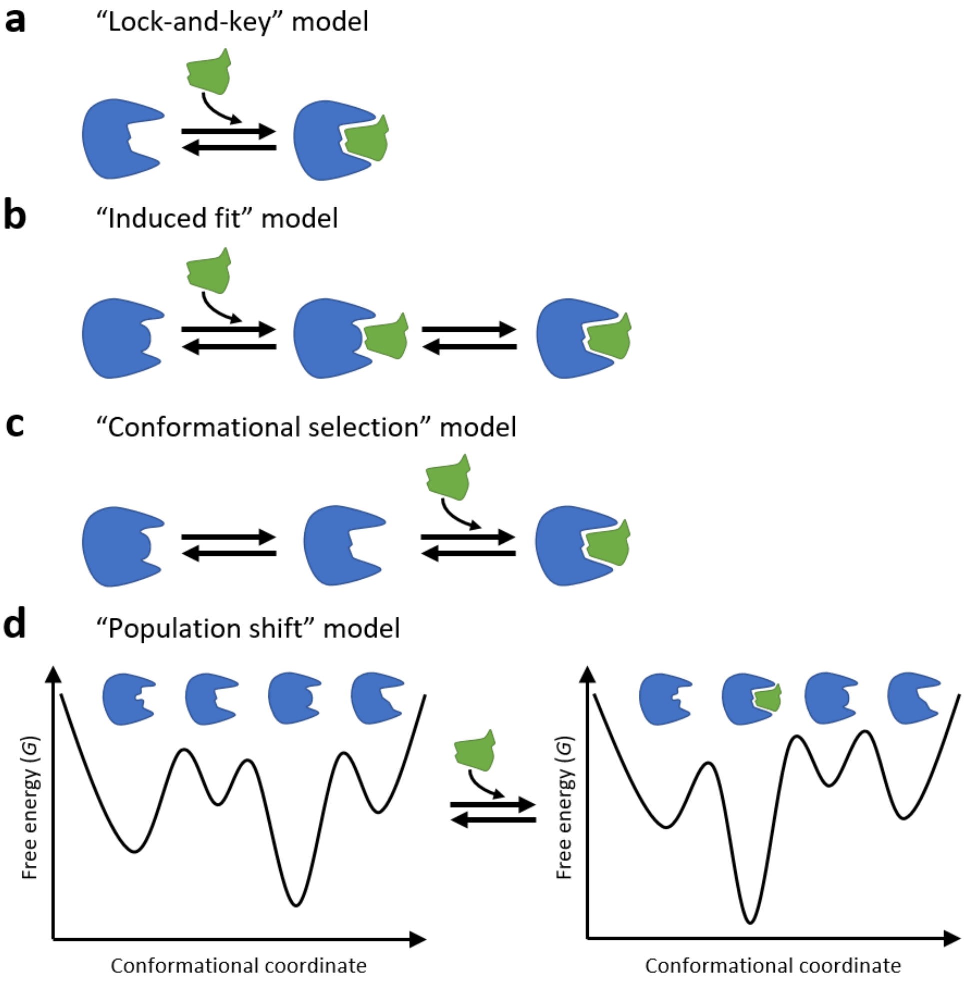


Figure 1.4. Models of ligand binding. The receptor (represented in blue) is capable of adopting different conformations with an impact on the binding affinity for the ligand (represented in green). (**a**) "Lock-and-key" model. The receptor and its ligand are complementary in shape; there is no conformational change involved. (**b**) "Induced fit" model. The ligand induces a conformational change in the binding interface to provide the optimal fit. (**c**) "Conformational selection" model. The receptor exists in an equilibrium between conformations, with different affinities for the ligand, prior to the binding event. The ligand binds to the most affine conformation. (**d**) "Population shift" model. The receptor exists in multiple states, each with an associated energetic stability (and proportional population). The binding occurs as in the "conformational selection" model, however the stabilities and populations of all receptor states (the native conformational ensemble) are altered.

In this context, two alternative models were proposed to explain the effects observed in oligomeric protein assemblies of identical subunits. Both models are phenomenological, therefore they offer explanations to observations without providing definite insight into the actual structural mechanisms that take place at an atomic level of detail [44], [69]. They were based on the assumptions that there is at least one ligand binding site per protein subunit and that they adopt two major conformations: the relaxed (R) state and the tense or tight (T) state. The R state is commonly associated with the more active form of the protein.

The MWC model, formally published in 1965 and named following the initials of its authors Monod, Wyman and Changeux [74], postulated that all protein subunits exist at the same time either in the R or the T state. In other words, the conformational interconversion between T and R occurs equivalently and simultaneously in all subunits. For this reason, this model is also known as the concerted or the symmetry model. Furthermore, the model assumes that the interconversion between states is in thermodynamic equilibrium in the absence of the ligand. The allosteric effect is achieved because when the ligand binding takes place, the conformational equilibrium becomes shifted towards the state with higher affinity. Since the conformational change is concerted, no matter which subunit hosts the ligand, all subunits undergo the transition and therefore they are more receptive to further ligand binding. Consequently, the MWC model added an alternative view of the mechanism of molecular recognition, other than the recently introduced “induced fit” model [49]. The fact that the ligand-affine conformation is available prior to the binding event contradicts the postulate of the “induced fit” model, where it is stated that the specific interaction of the ligand is required to induce the conformational change. This view is commonly known nowadays as the “conformational selection” model (Figure 1.4c). In addition, the MWC model also introduced subclasses of allosteric interaction. According to the nature of the allosteric effector, the allosteric interaction is either homotropic when it occurs between identical ligands or heterotropic when it occurs between different ligands. According to the effects produced in enzymatic systems, in the so-called K-type systems the binding of the allosteric effector alters the affinity for the substrate (as described in the general behavior of the model), whereas in V-type systems it changes the maximum rate of catalysis (in this case the T and R states differ in catalytic activity rather than affinity for the substrate) [43], [74].

The KNF model came a year later and offered an alternative view of the allosteric mechanism. It was developed by Koshland, Némethy and Filmer [75], and took up some of the considerations that had been made by Pauling [64], [69]. Contrary to the MWC model, under the perspective of the KNF model the subunits of the protein can undergo conformational transitions independently. However, prior to ligand binding, the protein exists only in the T state. When the ligand binds to a given subunit, it triggers its conversion to the previously inaccessible R state. This conformational change affects the structures of the adjacent subunits, increasing their affinity for the ligand. Successively, they will be more prone to achieving ligand binding, changing into the R state, and facilitating the same process at the remaining subunits until eventually all binding sites are occupied. Due to the sequential nature of the allosteric process, this model is also called the sequential model. The KNF model was, in fact, a natural derivation from the “induced fit” model that Koshland himself had proposed a few years earlier [72] to explain the mechanism of the binding event. An additional trait of the KNF model is that it can explain the possibility of negative homotropic cooperativity, unlike the MWC model [69], [76]. In such a case, the effect exerted on the adjacent subunits upon ligand binding is a decrease in the affinity instead of an increase.

The dichotomy brought about by the formulations of the MWC and KNF models caused that the subsequent studies of allosteric systems would focus on finding out which model best fitted the experimental data. In general, the MWC model was able to describe a broader range of cases [69], especially those for which concerted quaternary structure changes could be proved by means of techniques such as X-ray crystallography, spectroscopic techniques or even molecular dynamics simulations. Several regulatory proteins from different categories (enzymes, ligand-gated ion channels, G-protein–coupled receptors and nuclear receptors) have been found to support a "conformational selection" regime [49]. Conversely, the KNF model particularly succeeded in describing systems displaying negative cooperativity (something that the MWC model was incapable of doing by definition) [76]. In other cases, such as aspartate carbamoyltransferase, the GroEL chaperonin, the nicotinic acetylcholine receptor, ribonucleotide reductases or even the very same hemoglobin, both models have proven to be partially right by predicting well different structural and kinetic features [49], [69], [76]. Given that they were not entirely mutually exclusive, was the KNF model a particular case of a more general scenario governed by the MWC model, or the other way around? At least, the existence of a "conformational selection" did not exclude the possibility that, at a fine-structure level, local ligand-dependent movements consistent with the "induced fit" mechanism might take place [49]. Some works addressed the controversy by proposing general schemes combining features of both models [77], [78]. It is a debate that has never closed, although the scenario has evolved substantially since then. The new insight provided by the more contemporary views on the subject allows us to see beyond the constraints of the old models.

1.1.2.2 The current understanding of allostery

The subsequent discovery of allosteric properties in diverse biomacromolecular systems evinced some of the limitations of the original formulations of allostery. Remarkably, allostery began to be detected in many monomeric systems, such as signaling proteins, transport proteins, molecular motors and ion pumps [57], [69], [79], leading to the realization that proteins could be allosteric irrespective of their quaternary structure. This, in turn, implied that structural symmetry was not a required feature, and also that the conformational changes associated to the allosteric mechanism could occur at the level of tertiary structure motions. Allostery has been well-documented in a broad spectrum of possibilities of conformational heterogeneity, including large-scale conformational disorder (absence of a defined fold) [44], [80]. Furthermore, extensive evidence gained through analysis of structural data revealed that a conformational change is not even always required to transmit the allosteric signal, while such a role can be attributed to dynamical changes alone [81], [82]. Last but not least, allostery is also a regulatory mechanism of DNA and RNA-based systems, seen for instance in ribozymes and riboswitches [58], [83] or in the cooperative binding of two ligands to a DNA segment [84].

The consideration of dynamics in protein function and the first descriptions of the FEL in the 80s and the 90s [26], [28] marked a turning point in the understanding of the allosteric phenomenon. The modern views of allostery share the central notion that the allosteric behavior of a macromolecular system arises from the properties of the native conformational ensemble of the system [38]. Proteins are constantly sampling the available conformations of their free-energy landscape, depicted as a rugged surface with valleys and hills. The conformations with lower free energy (the valleys) are more stable and therefore are sampled more often than those of higher free energy (the hills). Any perturbation of the system, for instance the binding of an allosteric effector, reshapes the landscape and consequently modifies the energetic stability of the states and/or their rate of interconversion

[4], [8], [26]. The considerations of how this concept relates to allosteric proteins have been explored in depth through the years [8], [38], [44], [45], [50], [51], [85], [86].

The current understanding of allostery has been integrated in the so-called ensemble allosteric model (EAM), a framework that unifies descriptions of allostery in structured, dynamical and disordered systems [44], [80], [87]. The model describes the allosteric phenomenon entirely on the basis of the thermodynamic analysis of the FEL of allosteric systems. Accordingly, the characterization of the allosteric behavior of a system does not lie in identifying a single active conformation or a defined pathway of transmission of the signal along the structure, but in determining the conformational heterogeneity of the system. In other words, the nature of allosteric coupling is statistical and considers the ensemble-weighted contribution of all the states that exist in solution. Even if an allosteric effector binds with affinity only to a small fraction of states in the ensemble, the binding event will stabilize those particular states to the detriment of others, thus leading to a redistribution of the probabilities of the entire ensemble. Consequently, every state in the ensemble, regardless of its structural dissimilarity and/or functional activity, can be regarded as a potential allosteric state. This perspective has been very enlightening and extensively supported by experimental and computational research, especially as the number of biomolecular structural models available for study has increased [44].

In this regard, the EAM works with the following parameters: 1) the relative stabilities (or populations) of each accessible state of the ensemble, 2) the energy barriers associated with the transitions between states and their timescales, and 3) the binding affinities of the allosteric effectors or the measures of susceptibility to any other corresponding allosteric stimuli associated with each state [38], [86]. The spectrum of conformational heterogeneity of the FEL will give rise to different allosteric mechanisms, each with different properties and biological advantages, which are inherently pre-encoded in the ensemble. Thus, allosteric phenomena can take place within a dynamical continuum of possibilities (Figure 1.5) that encompasses rigid body motions (the classical T and R states), dynamical changes without conformational transition, local unfolding, and general structural disorder [44], [87], [88].

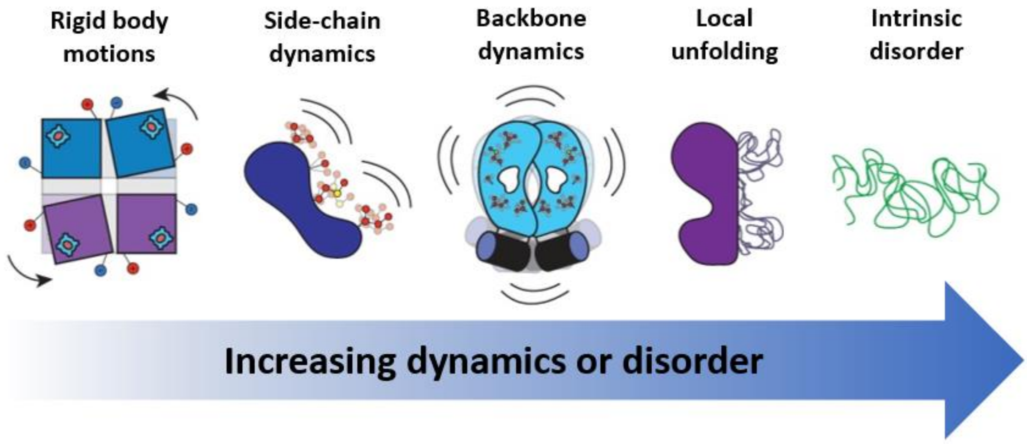


Figure 1.5. The dynamical continuum of the allosteric phenomena. The Ensemble Allosteric Model considers the whole spectrum of allosteric mechanisms in structured, dynamic and disordered systems. NOTE. Adapted from [44].

A particular feature of the EAM is that it successfully integrates in the same framework cases of allostery that lie in extreme points of the spectrum of conformational heterogeneity. Hilser and Thompson [80] were the first ones to propose and demonstrate that intrinsic disorder can be used by proteins to mediate allosteric coupling. Intrinsic disorder in proteins refers to the lack of a discernible tertiary structure. Proteins can have intrinsically disordered regions (IDRs), which can either be short, spanning a few amino acids, or constitute full domains. In some cases, the entire protein is disordered; it is an intrinsically disordered protein (IDP). IDRs and IDPs are highly dynamic systems typically involved in cellular signaling and regulation [6], [89]. They exhibit extreme conformational fluctuation (*i.e.*, smooth FELs), and typically stabilize into a more defined conformation once they bind or interact to other biomolecules [90]. Indeed, this attribute has been identified as a mechanism that allows to adjust or even reverse the cooperativity between two sites of the protein under certain conditions [44], [80].

Much of the insight contained in the EAM is the product of a collective effort of research in allostery and, in general, in structural biology. The growing embrace of the inherent dynamical nature of biomacromolecules inspired new insights and formulations. In 1999, Nussinov and coworkers proposed the “population shift” model of molecular recognition [33], which extended the previous propositions of the “conformational selection” model. The improvement consisted in arguing that the two-state scenario was in fact a reductionism, and actually, the general case comprises multiple states in equilibrium prior to ligand binding. A selective binding to the states of more affinity generates a new equilibrium of shifted populations (Figure 1.4d). “Population shift” is now the usual term employed by the scientific community to refer to the allosteric model that follows this molecular recognition regime [19], [39], [41], [50], [85]. For many years, it was claimed to be the “new view” of allostery, although it is rooted in the earlier intuitions of the MWC model. It had wide acceptance and managed to bring about an upturn of interest in research in allostery [69], and now it is integrated in the broader scheme of the EAM [44], [87].

An insightful implication of the dynamical view of proteins was brilliantly pointed out by Gunasekaran, Ma and Nussinov [50]. They stated that, if a population shift of the conformational ensemble is likely inducible by structural perturbation at any site of the protein, then virtually all dynamical proteins are potentially allosteric. Indeed, the evidence gathered in the last years of research was showing that allostery could be elicited by a rich variety of dynamical mechanisms. Furthermore, many proteins previously assumed to be non-allosteric were increasingly confirmed to display allosteric behavior, either by newly discovered putative effectors or by artificially introduced perturbations such as chemical modifications or point mutations. Following this rationale, they proposed that the binding of a ligand anywhere could potentially lead to a conformational change with an associated functional impact. The proposition has been subsequently supported [38], [41], [47], [51], [69]. Moreover, the existence of such hidden allosteric sites has been well proved [52] and has led to fruitful applications in drug discovery [56].

Perhaps the most extreme incarnation of dynamics in this subject [69] is the particular scenario of allostery without conformational change, an early hypothesis of Cooper and Dryden in 1984 [91]. Using a statistical thermodynamic formalism, they presented a rigorous demonstration of a mechanism by which the allosteric phenomenon can rely entirely on a change in the entropy of the system. In this model, termed “thermal fluctuations allostery” by the authors, the allosteric signal is transmitted solely by a modulation in the frequency and amplitude of thermal fluctuations, *i.e.*, in the absence of an appreciable conformational change of the protein backbone [44], [69], [91]. Years later, the work of Popovych and coworkers [81] was able to provide experimental validation of this theoretical proposition. They detected that the negative cooperativity in the CAP homodimer is mainly the result of an induced entropic penalty: the binding of cAMP to a subunit of the protein increases dynamics in the adjacent subunit that must be quenched to bind the second ligand [44], [81].

This model of allostery is known nowadays as "dynamic allostery" and has been further studied and discussed [38], [44], [47], [69], [82], [85], [86], [92]. Rather than a separate model, it has remained as a particular case of allostery potentially compatible with the EAM. In the definition made by Tsai *et al.* [47], allostery, as a thermodynamic phenomenon, can be classified as a process governed by enthalpy; by enthalpy and entropy; or solely by entropy. Accordingly, a change in the average structure (or "population") is not strictly required for allostery. Rather, in entropy-driven allosteric systems, it is the distribution around the average structure that changes, which in turn, affects the subsequent (binding) affinity at a distant site [69]. In other words, the principal (or the only) agent of allosteric response is the change in the vibrational profile of the protein (*e.g.*, backbone and/or side-chain dynamics).

The updated view of allostery with the consideration of the "dynamic allostery" model has been helpful in understanding allosteric mechanisms and thus in prediction of allosteric sites, allostery-related residues, allosteric drugs, and allosteric modulation [92]. Importantly, it has helped clarify the limitations of deducing the allosteric mechanism from static structure alone, clearly manifested in cases where the ligand-bound and unbound structures of a protein revealed no conformational differences [44], [49], [69]. Nevertheless, Nussinov and Tsai [93] also request extensive validation before assuming that a system follows the "dynamic allostery" model, and list several reasons for failing to observe conformational changes in some prominent allosteric systems. These reasons include crystal-packing effects, non-native crystallization conditions, inadequate accounting for disordered regions, ignoring synergistic effects between allosteric effectors, and too short molecular dynamics simulations [38], [41], [93].

Additional formulations have been proposed and discussed through the years, focusing on diverse particular aspects of allostery. For instance, is the allosteric signal from site to site propagated via a defined network of physically interconnected and/or thermodynamically linked amino acids that sequentially interact? The idea of the existence of allosteric networks that are evolutionarily conserved is intuitively very appealing and has been hinted at by many studies [41], [46], [58], [94], [95]. Moreover, if such networks exist, their identification in apparently non-allosteric proteins could provide clues to the location of possible allosteric sites [50]. Under the perspective of Tsai, del Sol and coworkers [47], [92], it is unlikely that a single defined pathway of communication between sites exists. A unified view of allostery should consider the more realistic scenario of the coexistence of pre-existing multiple pathways, where the degree of contribution of each pathway is determined by the particular physicochemical conditions [47], even in highly structured proteins [38]. Alternatively, other researchers dismiss the idea of defined pathways to state that communication occurs in a diffusely distributed manner [96], especially because of the existence of allosteric IDPs [87].

On another note, some studies have proposed other classification schemes and minor models that account for specific traits that have been observed in diverse allosteric systems. These may complement or overlap with each other or with the EAM, and can serve as a practical manner to refer to subgroups of proteins that display equivalent behaviors. Laskowski *et al.* [39] provide a classification scheme which groups cases into different categories according to the nature of the allosteric mechanism: i) open/close active site, ii) change active site conformation, iii) change active site electrostatic properties, iv) affect protein-protein complex formation, v) change protein flexibility, and vi) induce “population shift” in ensemble of conformers. The “morpheein model” explains the allosteric process of morpheeins: homo-oligomers that achieve alternative stoichiometries and require dissociation and reassociation steps [97]. The “mnemonical mechanism”, characteristic of the enzyme glucokinase, describes the concept that an enzyme “memorizes” the active conformation after the catalytic reaction and subsequently “forgets” it some time afterwards [98], [99].

All in all, over the more than five decades of research in allostery, we now possess a broader and precise view of the allosteric phenomenon. The concept of allostery has evolved substantially, thanks to a joint perspective of the structural and dynamical traits of proteins (and other biomacromolecules) that is focused on highlighting the convergent nature of the wide range of observed allosteric mechanisms [38], [44], [86]. The large bodies of experimental and computational studies have allowed us to consolidate our understanding of the strategies developed by evolution to transmit information from one part of a molecule (the allosteric site) to another (the effector site) [69]. In parallel, the growing interest in this subject has expanded considerably our therapeutic capabilities by developing effective allosteric drugs [40], [41], [56]–[58], [88], [92]. There are also initiatives devoted to integrating the vast amount of gathered knowledge, such as the AlloSteric Database [100], a central resource for the display, search and analysis of the structure, function and related annotation for allosteric molecules.

Despite the significant advances, much is still unknown about the physical properties that underpin allostery [38]. There are essential questions that remain unresolved. For instance, what mathematical parameters are needed to describe allostery both in highly structured and highly disordered systems? Or, what are the relative contributions of entropy and enthalpy to the allosteric free energy? Indeed, quantitatively understanding allosteric communication remains a great challenge, since we are still unable to derive common terms to provide a general detailed (atomic-level), quantifiable and predictive description of the allosteric mechanism [41], [43], [44], [87]. With this goal, Tsai and Nussinov developed a unified mathematical framework of the thermodynamic, “population-shift”, and structural points of view that links experimental allosteric proteins and the relative change in energy between the distinct conformational states [95]. LeVine and Weinstein explored models of biomolecular allostery through the use of Allosteric Ising Models (AIMs) in order to develop a quantitative theoretical description that bridges the features of the structural components and their interactions, to the thermodynamic allosteric parameters [101]. A recent work by Huang *et al.* developed a unified anisotropic elastic network model (uANM) to quantitatively estimate the contribution of pure “dynamic allostery” in a dataset of known allosteric proteins by excluding the conformational changes upon ligand binding [102]. Additional studies are needed to unveil the missing pieces of the puzzle. With both the available and the upcoming methods we will progressively conquer the still large territory that is open for scientific exploration.

1.1.2.3 Methods to study allosteric properties

Allostery has been largely studied by experimental and computational methods. These have allowed the discovery of a vast number of allosteric sites and effectors, as well as the characterization of the corresponding allosteric regulation mechanisms. The following summary offers just a glimpse of the actual scenario of the research in allostery, and only aims to illustrate the general traits of the spectrum of methods and techniques employed to study allosteric systems. For extensive reviews on the topic, I refer the reader to the following articles, [38], [41], [56], [58], [86], [88], [103], which list further references and describe many paradigmatic case studies.

Experimental approaches tackle the study of allostery by focusing on the analysis of local or distal conformational rearrangements and dynamical fluctuations at different timescales and possibly at the atomic level [88]. Computational approaches complement experimental methods and provide powerful tools to study allostery, with molecular dynamics simulations being a major source to provide details on dynamics. Some of the underpinnings of allostery have been elucidated by powerful computational techniques, which are also a great promise in protein engineering and drug discovery [38], [58], [86], [103]. Another advantage is that they help integrate large quantities of information resulting from extensive experimental data. However, the wide range of mechanisms encompassed by allostery cannot be modeled solely by a single technique. Rather, joint research efforts between different techniques are needed to keep advancing in our general understanding of allostery [88].

In the experimental side, X-ray crystallography, cryogenic electron microscopy (cryo-EM), and Nuclear Magnetic Resonance (NMR) spectroscopy are the most frequently used techniques [41], [44], [56], [69], [88]. X-ray crystallography and cryo-EM allow the study of biomacromolecules by providing detailed structural information on different static conformations of the same system, at atomic resolution. The new technical advances are expanding its potential by reducing the constraints of cryogenic temperatures typically employed (room-temperature X-ray crystallography [104], Multitemperature Multiconformer X-ray crystallography [105]) and enhancing the frequency of structural resolution along rapid events (time-resolved X-ray crystallography [106]). A thorough comparison of the obtained structures between different conditions can lead to the identification of pivotal regulatory mechanisms and of key residues involved. The lack of dynamical information of X-ray crystallography and cryo-EM is overcome by NMR spectroscopy methods, which can capture more transient conformations that are less populated and track motions belonging to the picosecond-second timescale. It is particularly useful to study allosterically regulated proteins for which a conformational change is rarely detected in response to allosteric modulator binding, thus providing direct experimental evidence to prove the “dynamic allostery” model in such systems [107].

Another technique that has emerged as a very useful tool to study allostery is native mass spectrometry (native MS). It enables simultaneous detection of co-existing states with different numbers of bound ligand molecules, allowing quantification of their populations and determination of binding constants [38], [88]. Further experimental approaches can be used to indirectly explore allostery in particular systems by detecting conformational changes. These include: site-directed mutagenesis methods, fluorescent and photoaffinity labeling such as Förster resonance energy transfer (FRET), and hydrogen/deuterium exchange mass spectrometry (HDX-MS) [41], [56].

In the computational side, molecular dynamics (MD) simulations have had a central role in exploring allosteric transitions at atomic resolution in all kinds of biomacromolecular systems [44], [69], [108], [109]. MD simulations generate large collections of snapshots, called trajectories, that track the dynamical behavior of biomacromolecules. They can be used to map the energy landscape of the allosteric process, compute free-energy calculations and explore any collective atomic motions possibly associated with allosteric transitions. Beside classical MD simulations, diverse other MD-derived strategies have been designed that can magnify the slower relevant motions by artificially lowering free-energy barriers. Some of these biased MD techniques that have been used to study allostery are accelerated molecular dynamics, steered molecular dynamics, and discrete molecular dynamics [38], [86]. Coarse-grained MD simulations also provide a strategy to attenuate the usually high computational cost of this technique by employing simplified models that preserve the global dynamical properties [41], [44]. It is important, however, to keep validating the reliability of the mechanistic descriptions of the conformational transitions of allosteric systems obtained with MD simulations by cross-checking the computational outcome with evidence gathered from experimental data, when possible [38].

The output data from MD simulations can be subsequently employed to conduct specialized analyses. Remarkably, graph theory derived methods can probe the mechanistic aspects of allostery on the simulated system by representing the structure of the biomacromolecule as a weighted graph or network, with individual residues as nodes, and edges expressing properties of the interconnected nodes such as the strength of residue-residue interactions or dynamical correlation metrics [58], [86], [103], [109]. Another well-established analysis derived from dynamical data consists in building Markov state models. In Markov state models, the sampled conformations of the trajectory are clustered by structural similarity, defining microstates of the system, and the transition probabilities between microstates can be estimated [86]. More recently, deep learning neural networks known as "autoencoders" have been applied to compare the time fluctuations of protein structures from MD simulations under different conditions, revealing concerted motions potentially involved in allosteric regulation [110].

Dynamical data can also be extracted from the approaches of Normal Mode Analysis (NMA) and the Elastic Network Models (ENM). This family of methods represent a reasonable first approximation to the description of correlated thermal motions. They can reveal soft collective modes of motion directly related to the functional mechanism of biomacromolecules, especially including conformational transitions associated with the allosteric regulation [86]. These modes of motion have been suggested to be like "paths" in the conformational space favored by evolution to enable the allosteric transition, thanks to the fact that they are highly collective and robustly defined by the overall architecture of the system [111]. Additionally, other more specialized methods have been developed to simulate and quantify the transmission of the allosteric perturbation by means of coarse-grain models. Some of them were reviewed recently in an interdisciplinary CECAM (*Centre Européen de Calcul Atomique et Moléculaire*) workshop, such as diverse 2D and 3D spring networks, the "allosteron" models and the "allosteric potential" metrics [38].

On the other hand, a handful of computational methods are focused on the identification of allosteric sites and effectors. These may be identified without previous experimental knowledge on conformational changes. Predictive tools can estimate the dynamical perturbation upon ligand binding to a particular site [41], [112]. The approaches for identifying the so-called "cryptic binding sites" are diverse [38], [86], [112]. Potential hotspots may be predicted by integrating the available dynamical and structural features of the analyzed systems, which can be subsequently implemented into machine-learning approaches. Alternatively, docking procedures can drive the predictions by estimating the strain exerted by different small ligand probes and selecting those with the lower energy binding poses. Among the different strategies for prediction of allosteric sites, some have been integrated in web servers, such as SPACER [113], MCPath [114] and Allosite [115].

Finally, some other bioinformatics approaches should be mentioned, which work with sequence data and seek to find patterns of evolutionary co-conservation of residues. Although it is not necessarily a property specific to allosterically coupled residues, it can inform or validate mechanistic aspects of allosteric processes [38], [56]. The increasingly available data on protein sequences and human polymorphism obtained with next-generation sequencing techniques is providing a promising scenario to exploit sequence analysis and investigate the role of evolution in shaping allosteric regulation. Nonetheless, this kind of analysis may be complemented by the usually more powerful computational approaches on the structural and dynamical level [38]. A referential approach named Statistical Coupling Analysis (SCA) was developed in 1999 by Lockless and Ranganathan [116], which measures the tendency of certain residues in a multiple sequence alignment to display correlated substitution patterns. Other methods inspired by SCA have been developed to find signals of evolutionary pressure that can lead to the identification of paths of allosteric communication [41], [86].

1.2 Pyruvate kinase

The protein known as pyruvate kinase (PK) represents a paradigmatic case of an allosteric protein that has been a focus of extensive research in the field of biochemistry, due to its pivotal role in the regulation of glycolysis. This section delves into the structural attributes of PK and the current understanding of its allosteric mechanism. Special focus will be given to the human erythrocytic form of this protein. Its distinctive features, both in the biological and the clinical contexts, make it an ideal subject for study in this thesis, aligning with the proposed hypotheses and research plan.

1.2.1 Function

Pyruvate kinase (systematic name: ATP:pyruvate 2-*O*-phosphotransferase; Enzyme Commission number 2.7.1.40) catalyzes the last step of glycolysis. In the corresponding reaction (Figure 1.6) a phosphate group from phospho*enol*pyruvate (PEP) is transferred to adenosine diphosphate (ADP), producing pyruvate and adenosine triphosphate (ATP).

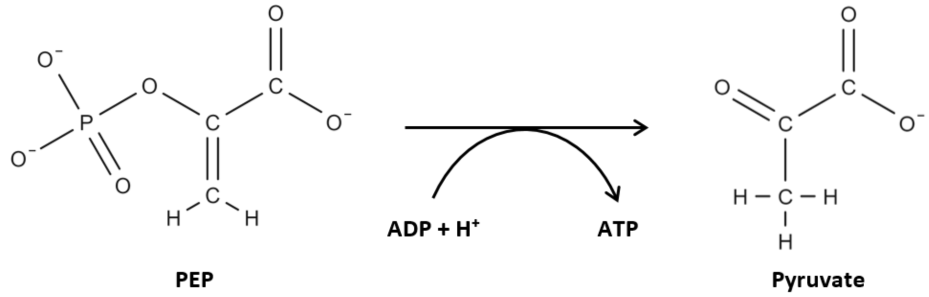


Figure 1.6. The reaction catalyzed by pyruvate kinase.

PK is a major regulation component of the flux of the glycolytic pathway in almost every cell type, together with hexokinase and phosphofructokinase. The three reactions catalyzed by these enzymes are the three rate-limiting steps of this pathway, meaning that they determine the overall rate of the pathway and thus serve as regulation points. Under physiological conditions, these reactions are energetically favorable and thus become essentially irreversible [117]–[119]. In addition, this reaction is of twofold importance as it is one of the two steps of the glycolytic pathway that synthesizes ATP, along with that of phosphoglycerate kinase, thus providing chemical energy for consumption in other metabolic or cellular processes [120].

The other product of the reaction, pyruvate, is a crucial versatile metabolite involved in multiple pathways. Under aerobic conditions, pyruvate is converted to acetyl-CoA to fuel the citric acid cycle as part of cellular respiration or, alternatively, to be utilized in fatty acid biosynthesis. Under anaerobic conditions, it is oxidized to lactic acid or ethanol. Moreover, pyruvate can be converted to alanine or, when gluconeogenesis is stimulated, to oxaloacetate to begin the synthesis of glucose and other carbohydrates [6], [117], [119]. Due to the existence of such a network of metabolic pathways, PK is subject to fine control by multiple mechanisms that favor the right processes under specific conditions and avoid possible futile cycles, for instance, between glycolysis and gluconeogenesis. Depending on the cell type and isoform, PK can be regulated by gene expression, pH, phosphorylation, and several allosteric effectors [6], [117], [120].

Given the wide occurrence of glycolysis across all life forms, PK is an ubiquitous enzyme that is found in most species and cell types [117]. It has been largely conserved throughout evolution (with up to ~40% of overall sequence identity between PKs from evolutionarily distant organisms [119]), as demonstrated by the high similarity in overall structure and active-site architecture [121]. Evolutionary divergence in this enzyme seems to be intimately connected to the acquisition of different allosteric properties that have enabled adaptation to diverse environments and metabolic strategies [119], [122]. In general, the main regulation mechanisms exhibited by most PKs comprise: i) the modulation of enzymatic activity by the presence of physiologic ions, mostly H^+, K^+ and Mg^{2+} (or Mn^{2+}), ii) the homotropic activation by its own substrate PEP, iii) the inhibition by the product ATP, and iv) the heterotropic activation by allosteric effectors whose nature depends on the organism, being fructose 1,6-bisphosphate (FBP) the most common modulator in bacterial, yeast, and mammalian PKs [117], [123].

1.2.2 Isoenzymes

Almost all organisms have at least one PK gene that corresponds to a single form of the enzyme. However, many species possess two or more PK isoenzymes (or enzyme isoforms) aimed at fulfilling specialized roles. For instance, multiple bacterial species express two isoenzymes that are modulated by different allosteric activators, while vascular plants exhibit cytoplasmic and plastic isoenzymes with disparate physical, immunological, and kinetic/regulatory characteristics [117].

In vertebrates, the use and control of PK is more sophisticated than in the rest of eukaryotes due to the specialized demands of the different tissues. Four PK isoenzymes are expressed in a tissue-specific manner by two different genes, namely, *PKM* and *PKLR*. The *PKM* gene is located on chromosome 15 (chromosomal location 15q22) and consists of 12 exons. The PKM1 and PKM2 isoenzymes are produced via alternative splicing of exons 9 and 10, respectively, differing solely in 22 amino acids within a fragment of 56 amino acids [124]. PKM1 is expressed in tissues with high catabolic demand, such as muscle, brain, heart [119], [125]. PKM2 is expressed in most adult tissues and especially in proliferative cells. It has been found in kidney, white adipose tissue, lungs, spleen, leukocytes, platelets, lymphocytes and the cells of the intestinal epithelium. Additionally, PKM2 is the dominant isoenzyme in early fetal tissues, and is progressively replaced by the more tissue-specific isoenzymes [118], [119], [124]–[126]. Moreover, elevated levels of low-active forms of PKM2 have been found in a broad range of cancer types. This anomaly aids in tumor cell proliferation and sustains their high metabolic demands by converting glucose into lactate, even under conditions of sufficient oxygen (a phenomenon known as the Warburg effect). For this reason, this isoenzyme has been the focus of considerable research in the last few decades, being considered a valuable tumor marker [55], [119], [124], [125], [127], [128].

On the other hand, the *PKLR* gene is located on chromosome 1 (chromosomal location 1q21) and consists of 12 exons. Alternate splicing of exons 1 and 2 produce the PKR and PKL isoenzymes, respectively, through the use of tissue-specific promoters, causing the former to have a larger N-terminal fragment of 31 extra amino acids [118], [129]. PKL is mainly expressed in the liver, renal cortex, and small intestine, and can also be found in pancreatic β-cells [117] and as a minor isoenzyme in the kidney [125]. PKR is exclusive to erythrocytes [118], [121]. During erythroid differentiation, the action of erythroid-specific transcription factors binding to the PKR promoter causes a switch from PKM2 to PKR [130]. A metabolic disorder arises from the functional disruption of PKR, called pyruvate kinase deficiency, which is presented in greater detail in section 1.3.4.

1.2.3 Molecular architecture

More than 120 crystallographic models of PKs have been resolved and uploaded in the public database of the three-dimensional structural data of biomacromolecules known as the Protein Data Bank [131]. Thanks to the extensive research conducted on PK, a rich characterization of its structure (functional sites, conformations, structural divergence between isoenzymes…) has been achieved. In most organisms, PK is a homotetramer of 200–240 kDa, although it may also be found in several other forms of quaternary structure, from monomeric to decameric [117], [118], [121]. The tetrameric formation has its subunits assembled in a "dimer-of-dimers" configuration (D_2 or 222 symmetry), involving three perpendicular 2-fold rotation axes that intersect at the center of the structure [118], [124] (Figure 1.7a). Each subunit is generally composed of four well-defined domains (Figure 1.7b).

The A domain is the largest, and constitutes the central region of the subunit. It folds in a $(\beta\alpha)_8$ barrel structure [121], [132], [133] that consists of an eightfold repeat of βα units, such that eight parallel β-strands on the inside are covered by eight α-helices on the outside. This is a conserved protein fold that can be found in several unrelated protein families. It is also known as the TIM barrel, named after the first enzyme where it was found, triose-phosphate isomerase (TIM) [134]. The β-strands and α-helices of the A domain can be designated with a sequential numbering from the N-terminus as Aβ1–Aβ8 and Aα1–Aα8 [120], [132], [135], following the typical naming scheme of TIM barrels. The connecting loops can then be referred to as βα loops and αβ loops [134]. In this work, a more explicit nomenclature has been adopted to designate the particular loops in the form "L-X-Y", where X and Y are the connected secondary-structure elements. For instance, the L-Aβ1-Aα1 loop follows after strand β1, or the L-Aα2-Aβ3 loop follows after α2. In PK, the barrel is characterized by three additional α-helical segments frequently named Aα6′, Aα7′, and Aα8′ that are located on βα loops and precede Aα6, Aα7, and Aα8, respectively. These helices have a central role in catalysis and allosteric regulation [125], [132].

The B domain is a protrusion of the A domain, inserted in the L-Aβ3-Aα3 loop, thus splitting the sequence of the A domain into two separate fragments. This domain folds in a combination of seven β-strands, random coils, and one or two short helical fragments; this topology has also been described as a mixed β-barrel [119], [124], [132]. The two fragments that interconnect the two domains (called the hinge fragments) are very flexible, enabling the B domain to behave as a lid that can bend and cover the region of the barrel of the A domain where the active site is located [119], [124].

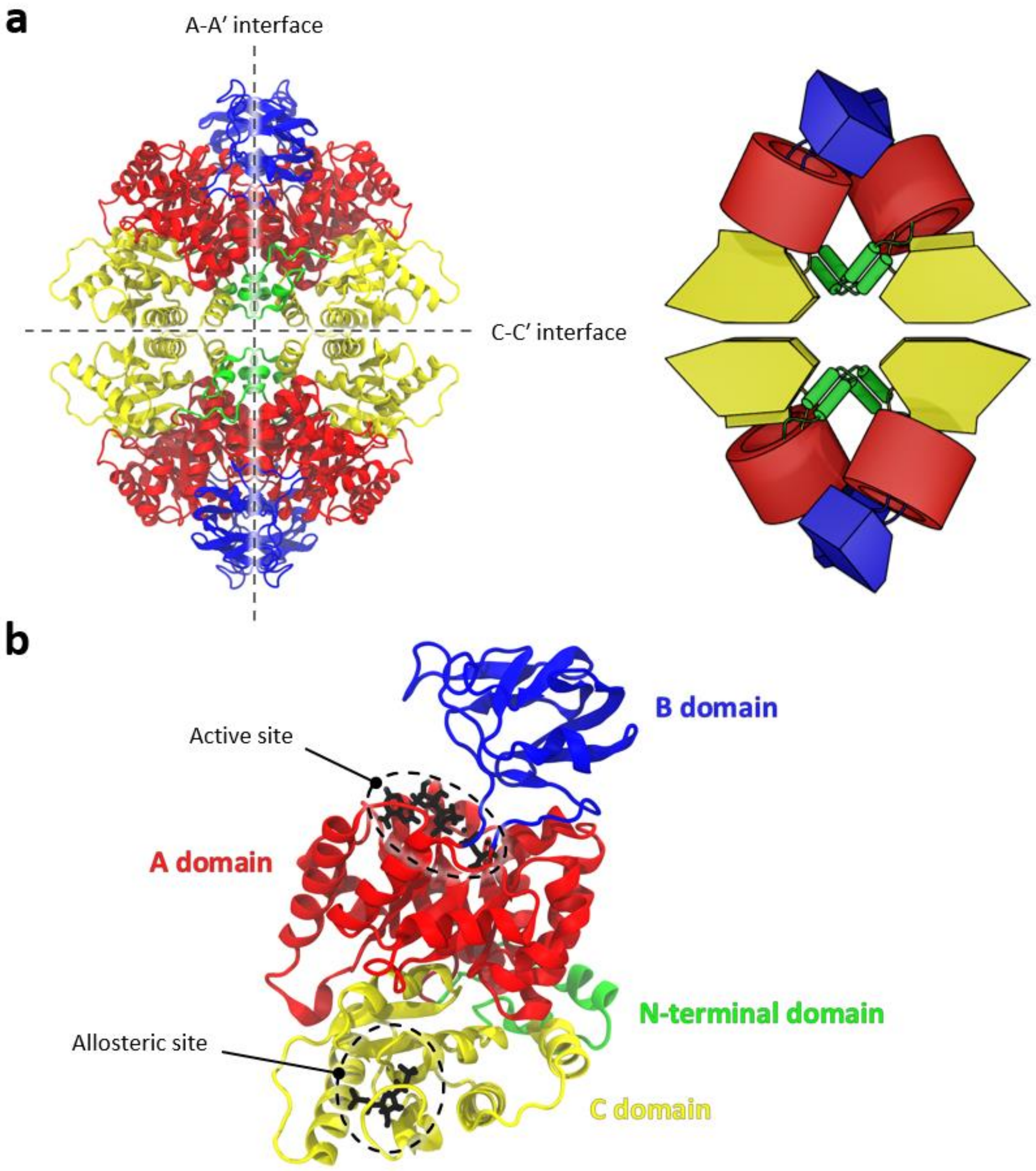


Figure 1.7. The structure of pyruvate kinase. The model corresponds to the human PKR isoenzyme (Protein Data Bank entry 2VGB [121]), shown in ribbon representation. (**a**) View of the tetrameric structure. The dashed gray lines show the two perpendicular 2-fold rotation axes that coincide with the A-A′ and C-C′ interfaces between subunits. The image of the protein structure (left) is accompanied by a 3D schematic diagram of the arrangement of the subunits and domains of the protein (right). (**b**) View of a subunit. The natural ligands that bind to the active and allosteric sites are shown in a black licorice representation. The ligands come from both the 2VGB structure and the holoenzyme complex that was modeled and employed in this study. NOTE. The images of the protein structure were generated with the software VMD. The 3D schematic model was built with the software Blender [136].

The C domain is found on the opposite side of the A domain and displays an α+β structure with ααβαβαβαββ topology. The five β-strands form a central β-sheet with the first four strands being parallel and the last strand being antiparallel [132], [133]. Finally, the N-terminal domain is a smaller helical domain characterized by a helix-turn-helix motif preceded by an initial fragment that has never been successfully crystallized and, thus, has been suggested to be intrinsically disordered [121], [124].

The tetrameric assembly is stabilized through extensive intersubunit interactions between the A and C domains of adjacent subunits, defining two large contact areas. The A-A′ interface coincides with a 2-fold rotation axis (vertical axis in Figure 1.7a) and is characterized by a network of contacts between the adjacent A domains, especially involving the Aα6, Aα7, and Aα8 helices. The C-C′ interface corresponds to the contact between adjacent C domains, coinciding with another 2-fold rotation axis (horizontal axis in Figure 1.7a). Tight interactions between the Cα2 helices and the Cβ5 strands of both subunits maintain the attachment. In fact, the C-C′ interface joins the central β-sheets of each C domain to generate a single continuous 10-stranded intermolecular β-sheet, antiparallel between subunits [119], [121], [132]. The folded fragment of the N-terminal extends towards the center of the tetramer where it reinforces the intricate network of contacts of the assembly by interconnecting the A and C domains of both its own subunit and the adjacent subunits [120].

In general, the A domain is the region with highest sequence identity among PKs, followed by the B and C domains [119]. The N-terminal domain is characteristic of eukaryotes, being absent in many bacterial species. It is a sequence of a variable number of residues, being especially lengthy in PKL and PKR isoenzymes [117], [121], [132]. Several bacterial isoenzymes, especially from the genus *Bacillus*, feature an additional C-terminal domain that is predicted to interact with adjacent subunits and be involved in a specific allosteric mechanism among PKs [122].

1.2.4 Active site and reaction mechanism

The active site of the enzyme is located in a cleft between the A and B domains, at the C-terminal ends of the β-strands of the barrel (Figure 1.7b). This is, indeed, the universal location of active sites in all known enzymes consisting of a TIM barrel [134]. The evolutionary conservation of the active site between PK isoenzymes is remarkably high, with very limited variation between species [117], [119]. Thus, the descriptions of both the binding geometry of ligands and the residues involved in catalysis are strongly equivalent between the diverse crystallographic models of this protein [120], [121], [124], [126], [132], [133], [135], [137]–[140].

The main sequence differences that exist in this regard lead to a distinction between two groups of PKs: those that show absolute requirement for K^+, and those that do not. Most PKs are from the first type, with the second type being found mainly in bacterial species [117], [119], [141]. In the binding site of K^+ (Figure 1.8a), the cation coordinates to the side-chain oxygen atoms of residues Asn118, Ser120, and Asp156, and the backbone carbonyl oxygen atom of Thr157 [126] (residue numbering according to the sequence of human PKR). In the absence of further ligands at the active site, the coordination complex is completed with two water molecules in a distorted octahedral geometry [137], [139]. Even though Na^+ may also bind to the site with barely distinguishable crystallographic structures, catalytic activity is dramatically reduced, which suggests that Na^+ somehow alters the optimal catalytic environment [133], [139], [142].

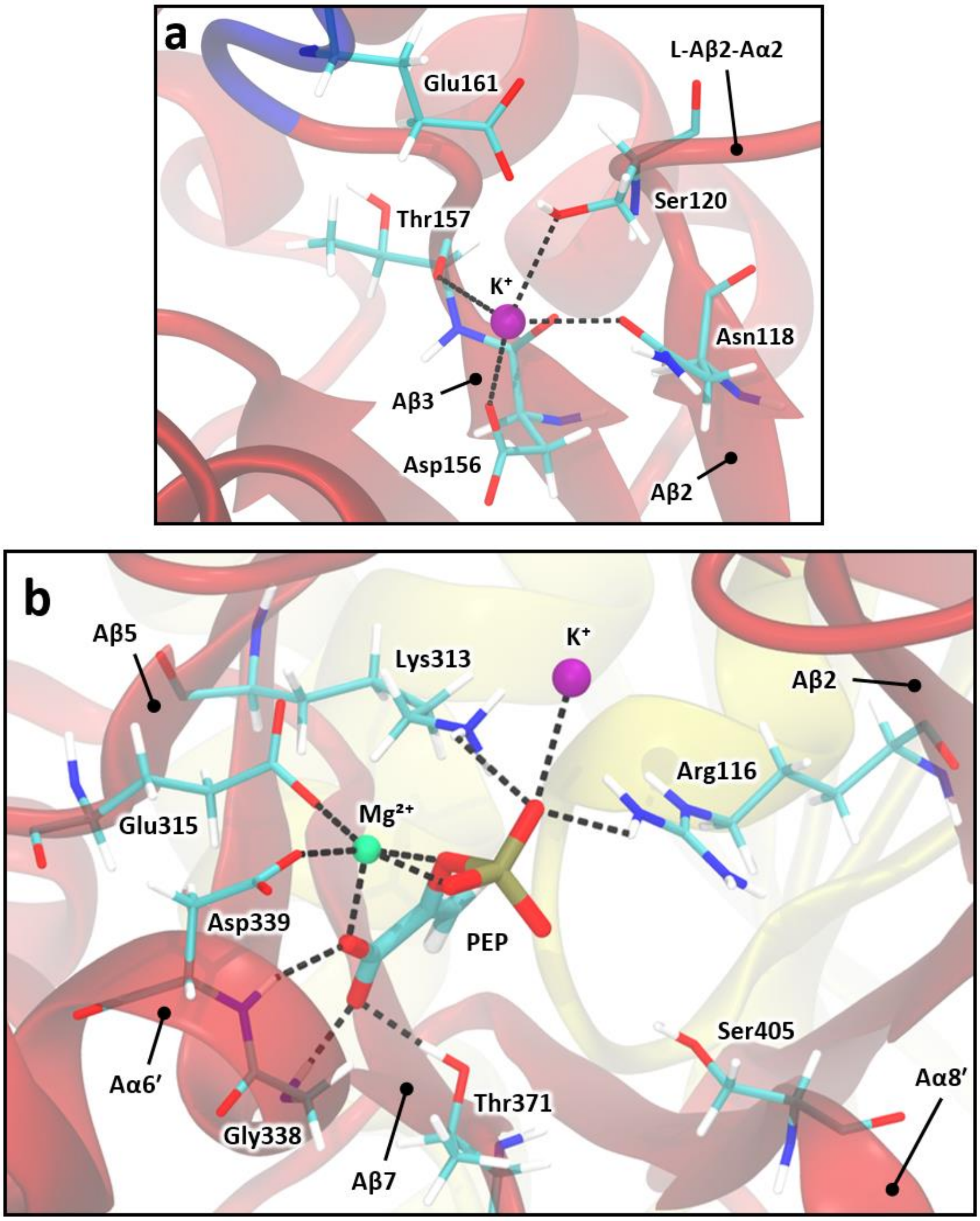
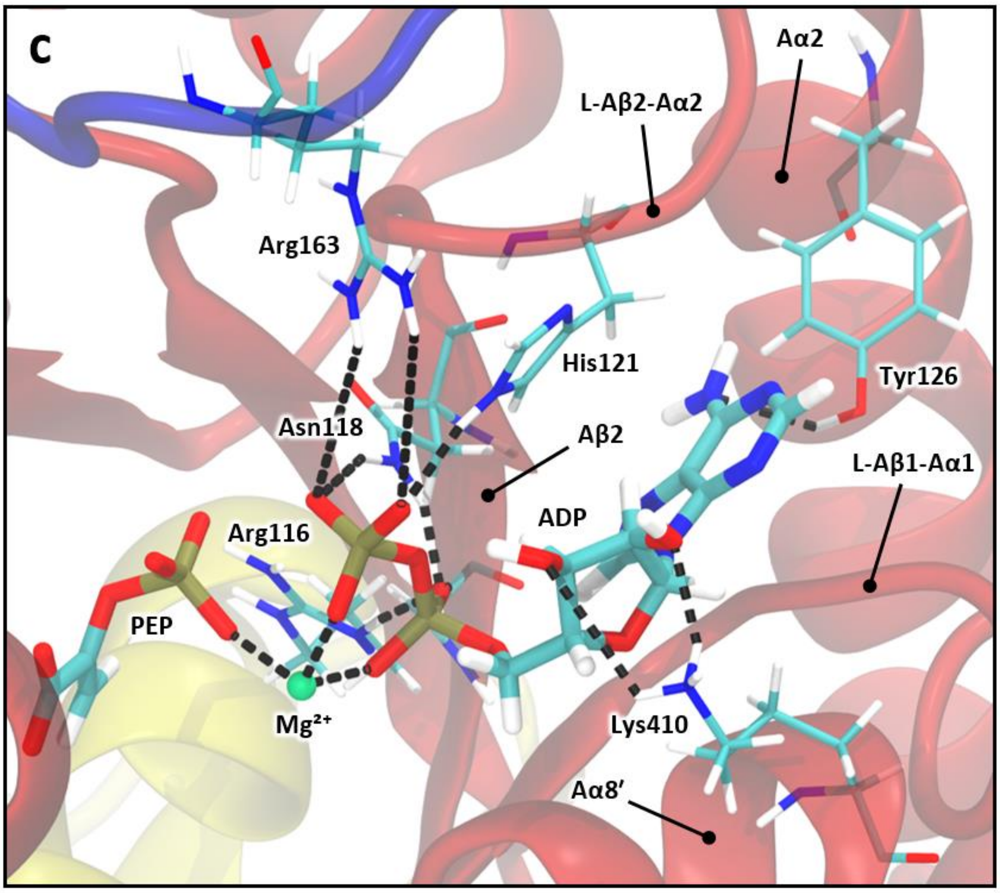


Figure 1.8. The active site of pyruvate kinase. The model corresponds to the human PKR isoenzyme (Protein Data Bank entry 2VGB). The backbone of the protein is shown in ribbon representation, colored according to the domains of the protein: A domain in red, B domain in blue, and C domain in yellow. The ligands are depicted with spherical or thick licorice representations, and come from both the 2VGB structure and the holoenzyme complex that was modeled and employed in this study. The amino acids that interact with the ligands are depicted with a thinner licorice representation. Atoms are colored by species. Coordination bonds and hydrogen bonds are depicted with black dashed lines. All the relevant amino acids, ligands, and structural regions of the protein are labeled. (**a**) The binding site of the cofactor K^+. (**b**) The binding site of the cofactor Mg^{2+} and the substrate PEP. K^+ is also included to show the coordination bond with the phosphate group of PEP. (**c**) The binding site of ADP and its complexed Mg^{2+} ion. PEP is also included to show its coordination bond with Mg^{2+}. NOTE. The images were generated with the VMD software.

Figure 1.8 (Continued)



The side chain of Glu161, located at the hinge between the A and B domains, also establishes electrostatic interactions with K^+ but is not canonically described as part of the coordination complex. In K^+-independent PKs, position 161 instead corresponds to a lysine residue that provides the positive charge internally via the protonated ε-amino group. In fact, it has been demonstrated that K^+ dependence can be removed by directed mutagenesis of the Glu161Lys replacement [119], [143]. K^+-independent PKs can still bind K^+ given that they conserve 3 of 4 of the coordination residues, with the only change being a substitution of Thr157 for a leucine (which nevertheless maintains coordination to K^+ via the backbone carbonyl oxygen atom) [141].

PK also requires a divalent cation as a cofactor, usually Mg^{2+} (or Mn^{2+}). The divalent cation coordinates to the carboxylate groups of Glu315 and Asp339 [121], [126], [139]. Mg^{2+} does not have more amino-acid coordination ligands because it forms a tridentate complex with the substrate PEP, specifically with oxygen atoms from the carboxylate, phosphoester, and phosphate groups (Figure 1.8b) [121], [126], [137], [139]. The coordination complex completes its octahedral geometry with one water molecule [124], [137], [139]. PEP also binds to the enzyme via interactions of its carboxylate moiety with the backbone amide groups of Gly338 and Asp339, which are located in the Aα6′ helix, and the hydroxyl group of Thr371 [119], [121], [137]. When PEP is bound, its phosphate group becomes coordinated to K^+, replacing one water molecule in the distorted octahedral geometry [126], [137]. The nearby side chain groups of Arg116 and Lys313 also establish interactions with the phosphate group of PEP.

ADP (bearing its own complexed Mg^{2+} ion) binds to the active site with the β-phosphate group aligned towards the phosphate group of PEP to provide an optimal orientation for the phosphoryl-transfer reaction (Figure 1.8c). The adenine ring of ADP fits into a pocket defined by the L-Aβ1-Aα1 and L-Aβ2-Aα2 loops and the Aα8′ helix. Multiple side chains establish interactions with ADP: Arg116 with the α-phosphate, Asn118 with both the α- and β-phosphates, His121 with the β-phosphate, and Lys410 with the 2′- and 3′-hydroxyl groups of the ribose moiety. Additionally, Arg163 from the B domain also interacts with the β-phosphate when the B domain adopts the corresponding closed conformation to better capture the substrate [119], [137], [139]. The exocyclic amino group of the adenine ring potentially makes a H-bond with the hydroxyl group of Tyr126. However, the absence of a definite H-bonding interaction for this group is consistent with the rather broad nucleotide specificity of the enzyme [139]. Incidentally, this ADP/ATP-binding site does not fit the patterns of binding interactions that are characteristic of the superfamilies of nucleotide-binding proteins or ATP-binding structural motifs [139]. Crystallographic structures with bound MgATP show that this second Mg^{2+} ion does not interact with the protein, but rather coordinates to the α-, β-, and γ-phosphates and has its octahedral coordination sphere completed by three water molecules [137], [139]. Since the phosphate of PEP is equivalent to the γ-phosphate before the occurrence of the reaction, it can be inferred that in the initial state, with the two substrates bound, PEP concurrently interacts with the three metal ions of the active site [139].

Due to the abundance of K^+ under physiologic conditions in the cytosol, the metal is most likely constitutively bound to the enzyme [142], [144], [145]. It has been demonstrated that with K^+ bound, PEP and MgADP can bind independently to their respective sites in a random sequential mechanism [125], [140], [146]. This is the usual kinetic mechanism in type I K^+-activated enzymes [142]. The binding of the cofactor Mg^{2+} is also required for the stable and catalytically active configuration of the active site. The process by which the bound state of both Mg^{2+} and PEP is eventually reached is not clear. However, Mg^{2+} has been hypothesized to be retained after product release to prime the active site to accept the next PEP substrate molecule [137].

The physiologic reaction takes place in two steps (Figure 1.9). The first step consists in the phosphoryl transfer from PEP to ADP, producing ATP and the enolate of pyruvate (Figure 1.9a). The cofactors K^+ and Mg^{2+} assist by providing electrostatically optimal binding of the substrates. These and the ADP-bound Mg^{2+} are coordinated to each peripheral oxygen of the phosphate of PEP, thus screening electrostatic repulsion between the anionic reactants and lowering energy barriers. Under this influence, the aligned β-phosphate of ADP accomplishes the nucleophilic attack on the phosphorus atom of PEP (S_N2 reaction). During the phosphoryl-transfer process, K^+, Arg116, and Lys313 serve to compensate the developing negative charge of the corresponding pentacoordinate transition state and stabilize the enolate intermediate for its detachment from the phosphate group [124], [125], [133], [139], [142].

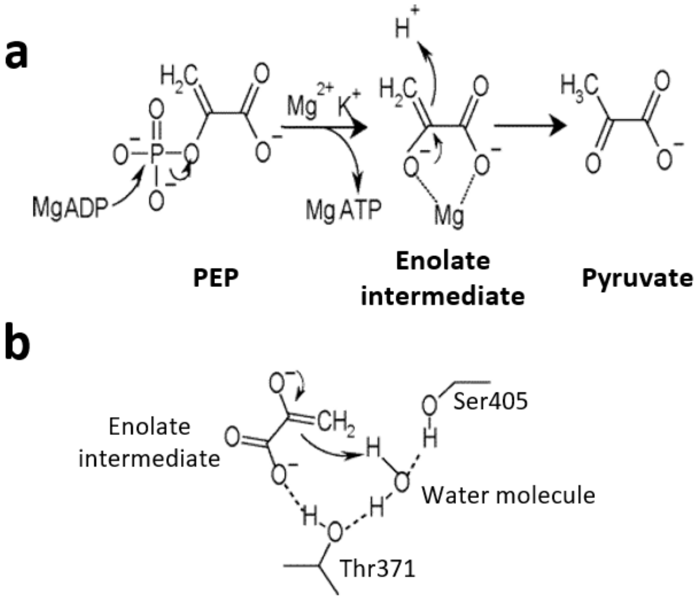


Figure 1.9. The two steps of the reaction catalyzed by pyruvate kinase. Curved arrows indicate the flow of pairs of electrons as commonly used in diagrams of reaction mechanisms. (**a**) The first step involves the formation of an enolate intermediate and the release of ATP as a product, while the second step involves the ketonization of the enolate intermediate to release pyruvate as a product. (**b**) The proposed H-bonding scheme responsible for the protonation of the enolate, with a water molecule as the donor, stabilized by Thr371 and Ser405. Additional water molecules and residues may also participate as general proton donors based on a H-bonding network. NOTE. The image was adapted from [124].

In the second step, the energetically less-stable enol form of pyruvate is subsequently protonated to produce the corresponding α-keto carboxylate form of the molecule, *i.e.*, pyruvate (Figure 1.9b). This process is also known as the tautomerization from the enol to the keto form, or simply ketonization. Ketonization of *enol*pyruvate is highly energetically favorable and, thus, renders the overall reaction exergonic and irreversible in physiologic conditions. The process occurs when *enol*pyruvate accepts a proton from a water molecule at the 2-*si* face of the double bond. Crystallographic evidence shows that a specific water molecule can be held in position by the conserved active site residues Thr371 and Ser405. A proton-relay system has been also hypothesized whereby Thr371, Ser405 and even Arg116 and Lys313 could participate as general proton donors of the reaction via the H-bonding network of water molecules nearby [124]–[126], [133], [139].

1.2.5 Allosteric sites and regulation mechanisms

The major allosteric activator in PKs in a diverse range of species is FBP, only being replaced by the slightly different effector fructose 2,6-bisphosphate (FDP) in trypanosomatid protozoans [117], [119], [125], [137], [147]. Binding of FBP (or FDP) increases the affinity of the enzyme for PEP without altering the catalytic rate (k_{cat}) or the affinity for MgADP. Therefore, regulation by FBP follows the classical description of K-type allosteric mechanisms [43], [74], [123]. The allosteric center is located within the C domain, next to the C-C′ interface and approximately 40 Å from the catalytic site. The binding site of FBP (Figure 1.10) is a pocket between the L-Cβ1-Cα3 loop and the first two turns of the Cα5 helix. The L-Cβ4-Cβ5 loop is a mobile fragment that surrounds the cavity. Upon binding of FBP, it covers the molecule via stabilizing interactions [119], [121], [124], [126], [147], [148].

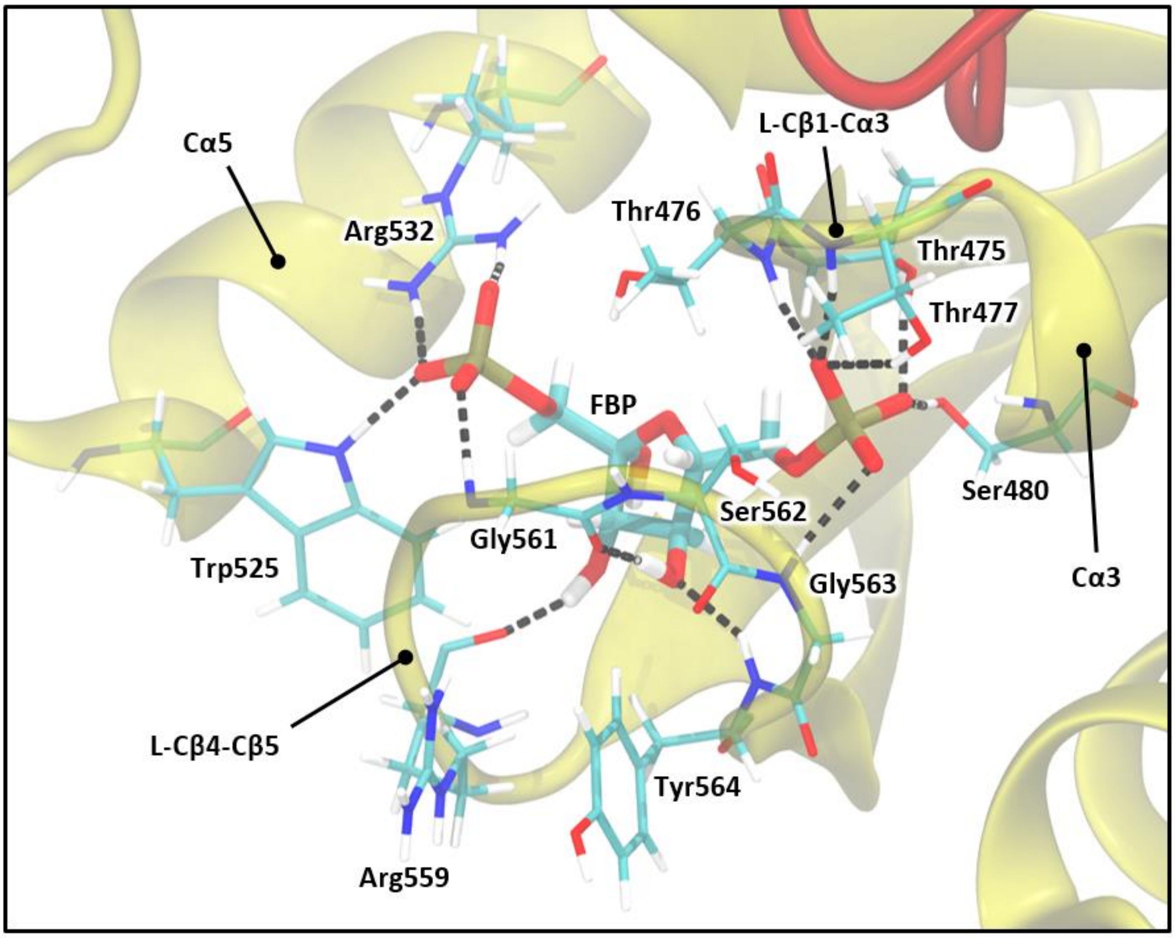


Figure 1.10. The allosteric site of pyruvate kinase. The model corresponds to the human PKR isoenzyme (Protein Data Bank entry 2VGB). The backbone of the protein is shown in ribbon representation, colored according to the domains of the protein: A domain in red and C domain in yellow. The ligand FBP is depicted with a thick licorice representation. The amino acids that interact with the ligand are depicted with a thinner licorice representation. Atoms are colored by species. Hydrogen bonds are depicted with black dashed lines. All the relevant amino acids and structural regions of the protein are labeled. NOTE. The image was generated with the VMD software.

The FBP molecule is engaged in an extensive network of contacts at the binding site. The 6′-phosphate group of FBP makes a series of H-bonds with i) the backbone amide groups of Thr476, Thr477 and Gly563; ii) the hydroxyl groups of Thr475, Thr477, Ser480, and Ser562. The 1′-phosphate group makes H-bonds with i) the side-chain groups of Thr476, Trp525, and Arg532, and ii) the backbone amide group of Gly561. The 3′-hydroxyl and 4′-hydroxyl groups make H-bonds with the backbone oxygen atoms of Arg559 and Gly561, and with the backbone amide group of Tyr564. These interactions of the furanose ring are a result of the orientation of the L-Cβ4-Cβ5 loop locking FBP bound at the allosteric site [119], [121], [124], [126], [147], [148].

The PK isoenzymes that are not regulated by neither FBP nor FDP belong to bacterial species and coincide with the PK subtype of isoenzymes that have a K^+-independent reaction mechanism. These are instead regulated by adenosine monophosphate (AMP), which binds to the same site as FBP, or sugar monophosphates. In addition, PKs of some species are not even subject to allosteric control [117], [119], [132].

The four PK isoenzymes from vertebrates display disparate kinetic and allosteric properties that reflect the different metabolic requirements of the tissues. PKM1 is the only constitutively active isoenzyme;

it is not allosterically regulated by FBP and neither displays cooperativity by PEP [118], [120], [132]. This is due to the sequence differences introduced by the alternatively spliced exon with respect to PKM2, which affect the specific regions of the allosteric site and the intersubunit contacts involving the C domain. FBP is not able to bind PKM1 due to the structural differences. A glutamic acid replaces a threonine/serine residue (Thr476 in PKR) in the allosteric site that is characteristic of FBP-regulated PKs. This glutamic acid may either hinder FBP binding due to electrostatic repulsion, or act to mimic the effect of FBP binding [117], [124]–[126], [149].

On the other hand, PKM2 exhibits a unique regulation mechanism that is based on the transition between different oligomeric states during allosteric activation. Unlike the other isoenzymes, PKM2 is not a constitutive tetramer; it exists in an equilibrium between monomers, dimers and tetramers, the latter being the least abundant in the absence of the allosteric activator FBP [119], [127], [128], [149], [150]. The binding of FBP effectively shifts the equilibrium towards an active tetrameric formation. In the monomeric or dimeric state, the enzymatic activity of PKM2 is very low in comparison to the FBP-bound tetramer. FBP may bind to monomeric or dimeric forms to induce tetrameric formation, or may bind to the available sites of a tetramer and stabilize its assembly in an active conformation. The binding occurs in a highly positively cooperative manner [125], [150]. Monomers first dimerize along the A-A′ interface and, subsequently, dimers associate along the C-C′ interface [124]. In addition, PKM2 is also allosterically regulated by the thyroid hormone 3,3′,5-triiodo-ʟ-thyronine (T_3). In this case, T_3 inhibits PKM2 by binding to the inactive monomeric form and stabilizing it. The inhibitory effects of T_3 are overcome by FBP binding [125], [149].

In addition to the regulation by FBP, a range of individual amino acids in solution have been described to be allosteric effectors of PKs. Amino-acid modulators bind to a second allosteric site, different from that of FBP, which is known as the amino-acid binding site. It is located in each subunit at the interface between the A and C domains, in a pocket delimited by Aβ1, L-Aα1-Aβ2, L-Aα2-Aβ3, Cβ3, and Cα4 (Figure 1.11). This region is highly conserved in sequence between PKM1, PKM2, PKL, and PKR. The ʟ-2-aminopropanaldehyde substructure shared by all amino acids (except for glycine) has been proposed to be the chemical moiety that is required for successful binding of the amino-acid effectors in PKM1 and PKL. However, particular amino acids achieve different extents of allosteric response between these two isoenzymes, raising the idea that allosteric pathways are not always conserved between homologues of a protein family [151]. As observed in crystallographic structures, the carboxylate group of the amino-acid ligand interacts with the side-chain groups of Asn113, Arg149 and His507, while the amino group interacts with the backbone oxygen atom of Val512 (PKR numbering) [128], [149], [151].

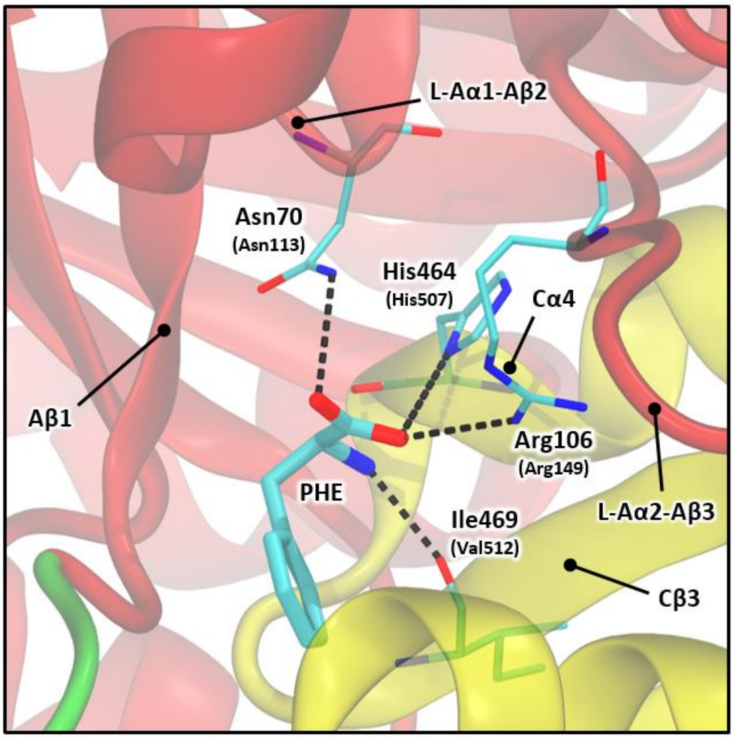


Figure 1.11. The amino-acid binding site of pyruvate kinase. The model corresponds to the human PKM2 isoenzyme (Protein Data Bank entry 6GG4 [128]) in complex with the allosteric inhibitor phenylalanine (PHE). The backbone of the protein is shown in ribbon representation, colored according to the domains of the protein: N-terminal domain in green, A domain in red and C domain in yellow. The ligand PHE is depicted with a thick licorice representation. The amino acids that interact with the ligand are depicted with a thinner licorice representation (hydrogen atoms not depicted). Atoms are colored by species. Hydrogen bonds are depicted with black dashed lines. All the relevant amino acids and structural regions of the protein are labeled. Amino-acid labels corresponding to human PKR numbering are included in parenthesis. NOTE. The image was generated with the VMD software.

In PKL and PKR, the amino-acid effector that exerts the most significant modulation is alanine, which is an allosteric inhibitor. Kinetic assays with directed mutagenesis on several PKL positions suggest that inhibition by alanine is not simply the reverse of activation by FBP, *i.e.*, different networks of residues contribute to the two allosteric functions [152]. Other allosteric inhibitors with weaker effect in PKL and PKR are cysteine, proline, valine, and phenylalanine [151]. In PKM1, alanine, proline, and phenylalanine are allosteric inhibitors, with the latter exerting a stronger effect than the others. FBP binding overcomes their inhibitory effects [122], [125], [128], [149], [151]. In PKM2, phenylalanine, alanine, and tryptophan are inhibitors. Interestingly, phenylalanine does not oppose the tetramer-promoting activity of FBP, but rather stabilizes the tetrameric formation albeit in a less-active state. There is conflicting evidence on whether alanine and tryptophan inhibit the enzyme via the same mechanism or, alternatively, by inducing dissociation of the tetramer [125], [128], [149]. Serine is an allosteric activator and induces tetrameric formation in an active state, similarly to FBP. Other amino acids are either activators or inhibitors that exert weaker effects [128], [153], [154].

Finally, several post-translational modifications that regulate the different tissue-specific isoenzymes have also been described. The activity of PKL and PKR can be regulated by phosphorylation by protein kinase A. This phosphorylation event affects the residue Ser43 (PKR numbering), at the N-terminal domain, and has been suggested to interrupt a constitutive interaction between the N-terminal domain and the main body of the protein, leading to a decrease in apparent PEP affinity [155], [156].

Phosphorylation of Ser43 has also been correlated with increased affinity for inhibitors ATP and alanine and decreased affinity for FBP [157], although the real physiologic implications of the phosphorylation event remain unanswered [158]. Regarding PKM2, several other post-translational modifications have been identified. These are caused by the action of either enzymes (phosphorylation of serine and tyrosine residues, acetylation of lysine residues, hydroxylation of proline residues) or reactive oxygen species (oxidation of cysteine residues). Some of these modifications disrupt PKM2 tetramerization and thus are inhibitory, whereas others have been correlated with more complex responses that promote proliferation in tumor cells [119], [125], [127], [128].

1.2.6 Conformational changes

Extensive evidence has been gathered over the years through the interpretation of comparative studies of crystallographic structures of PKs in complex with various ligands, revealing the diverse conformations of the enzyme. This has enabled the deduction of the conformational transitions induced by the binding of cofactors, substrates, and allosteric effectors. The tetrameric structure exists in an equilibrium between two main conformations, namely, the T and R states, which correspond to catalytically inactive and active states, respectively. Shifts in this conformational equilibrium are correlated with the binding events of either the substrate PEP or the allosteric effectors [135], [159].

The conformational transition involves a reorientation of the subunits and their domains that occurs in a symmetric fashion along the tetramer. The overall structure of each domain does not undergo major internal rearrangements in the process; instead, the different structural modules behave as rigid bodies that move relative to each other. The corresponding concerted motions have been interpreted with two different models (Figure 1.12). Mattevi *et al*. [132] compared the crystallographic structures of the type I PK from *Escherichia coli* without ligands and the constitutively active PKM1 from rabbit in complex with Mn^{2+}, K^{+}, and pyruvate [133]. The structural differences between the former (T state) and the latter (R state) were described as a combination of rotations both of the individual domains within each subunit and of each subunit within the tetramer (Figure 1.12a). This model has been referred to as the “domain-rotation” model by other studies and reviews [119], [160]. In contrast, Morgan *et al*. [147] worked with crystallographic data of PKs from trypanosomatid parasites and interpreted the conformational change as a joint rigid-body rotation (rocking motion) of the A and C domains as a single block (Figure 1.12b). The pivot point of the rotation was identified as the interface between these two domains, coinciding with the location of the Cα4 helix. The model was termed the “rock-and-lock” model by the authors, and later has been referred to as the “rigid-body–reorientation” model [119], [160]. This model was further expanded in subsequent studies with structures of human PKM2 [128], [149].

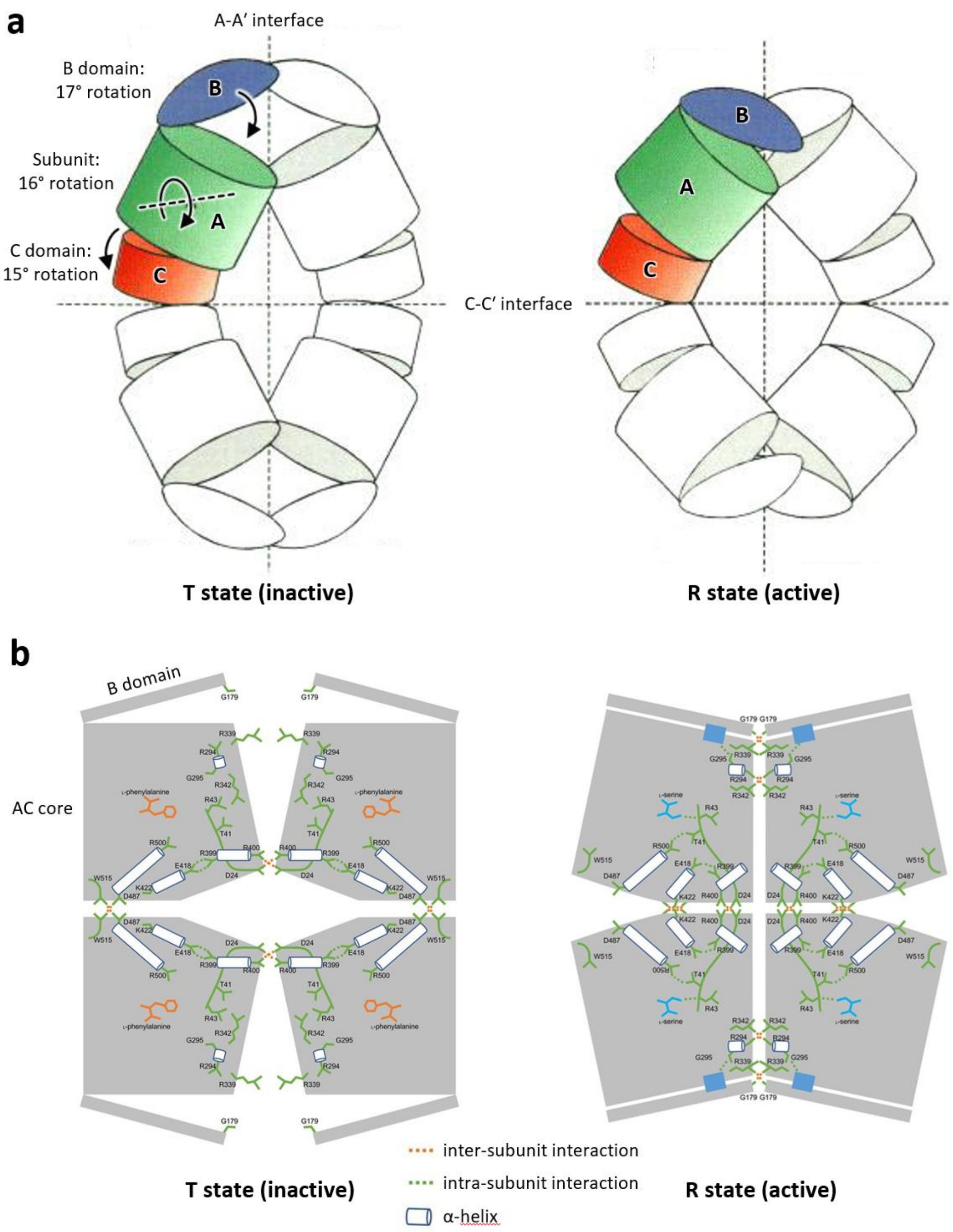


Figure 1.12. The main conformational transition of pyruvate kinase. Schematic diagrams of the arrangements of the subunits and domains in the inactive (T) and active (R) states of the enzyme. (**a**) The “domain-rotation” model, proposed in the study of Mattevi *et al*. [132]. The diagrams represent the T state of type I PK from *Escherichia coli* and the R state of rabbit PKM1. The A, B, and C domains of a subunit are colored in green, blue and red, respectively. The set of motions affecting the domain and subunit orientations involved in the T-to-R conformational transition are indicated with arrows. (**b**) The “rock-and-lock” model, proposed in the studies of Morgan *et al*. [147], [149] and further characterized in the study of Yuan *et al*. [128]. The diagrams represent human PKM2 in the tetrameric formation. The T and R states feature the allosteric effectors phenylalanine (inhibitor; shown in orange) and alanine (activator; shown in cyan), respectively, bound to the amino-acid binding site. Each AC core of PKM2 is represented as an irregular pentagonal block. B domains are represented by narrow rectangles. Active sites are represented as blue rectangles. The diagrams show a network of the important interactions (dashed lines) established between different residues (shown in green) that were identified in [128]. The figure included here is an adaptation from the one of the original work that only has the purpose of showing the overall reorientations of the structure. Readers are referred to the original work to find the corresponding full-size figure and scrutinize the network of interactions in detail. NOTE. The images were adapted from [132], [161] (a), and [128] (b).

Irrespective of the particular interpretation of the underlying concerted motions, both models offer an explanation for the molecular basis of the coupling between the active and allosteric sites. Importantly, the same conformational transition is achieved by the independent binding of PEP or FBP, which simultaneously explains both homotropic cooperativity and heterotropic activation and evinces the existence of a communication pathway that interconnects both mechanisms [132], [159], [162]. The original studies in which these models were proposed, in conjunction with subsequent extensive research, have provided consistent descriptions of concurrent changes in the network of interactions at the A-A′ and C-C′ interfaces that result from such subunit/domain reorientations. The mechanistic interpretations suggested in the literature [124], [127], [128], [132], [147], [148], [159], [162]–[165] can be summarized as follows.

Ligand binding at either the active or the allosteric site exerts local rearrangements at the nearest type of subunit interface that subsequently propagate to the rest of the structure, driving the integral conformational change. The binding event of PEP to the active site of a subunit stabilizes a symmetrical network of interactions between the Aα6′ and Aα7 helices across the A-A′ interface, priming the adjacent (unoccupied) active site to accept a molecule of PEP with higher affinity. In turn, this symmetrical local rearrangement of the A-A′ interface results in the subunit rotations exhibited by T- and R-state crystallographic structures. The rotational motions of the structures are accompanied by the corresponding local rearrangements at the C-C′ interface, manifested in the formation and breakage of different intersubunit interactions. The region close to the center of the tetramer involving the Cα1 and Cα2 helices gains tighter interactions between the adjacent C domains. In parallel, the intersubunit interactions at the opposite section of the C-C′ interface that involves the Cα5 helix, the L-Cβ4-Cβ5 loop, and the Cβ5 strand are broken.

This very same process can be triggered from the complementary perspective that begins with FBP binding to a subunit of the T-state tetramer. Initially, the L-Cβ4-Cβ5 loop folds to lock FBP at the allosteric site. This event interrupts the intersubunit interactions at that region of the C-C′ interface that stabilize the T state, thus promoting the formation of the ones near the center of the tetramer. The changes propagate with the corresponding rotation of the structure, reaching the active site and inducing a local environment with higher affinity for PEP. Interestingly, the particular interactions that stabilize the T and R states may be different between species due to sequence variation, but the underlying principle of subunit rotation to stabilize the R-state tetramer upon effector binding appears to be a shared allosteric strategy [122], [149]. Amino-acid allosteric inhibitors and activators also promote the T or R states, respectively, via a network of interactions that originates in the amino-acid binding site and converges on the aforementioned mechanism [128] (Figure 1.12b).

On the other hand, the conformational changes of the B domain were also originally included in the “domain-rotation” model [132]. Considering the general notion of allosteric enzymes, the binding of allosteric effectors should be expected to modulate the behavior of the B domain. Contrary to this expectation, subsequent studies have shown that the conformational changes of the B domain are not directly coupled to the main conformational transition of the core structure of PK. The B domain is a particularly mobile region due to the flexible pair of linker fragments that connect it to the A domain and that serve as a hinge mechanism [164]. The ensemble of orientations of this domain is dominantly governed by the presence or absence of the ligands at the active site. The B domain tends to decrease mobility upon substrate binding, adopting partially closed conformations when PEP is bound and a fully closed conformation when ADP is bound [121], [124], [135], [137], [139], [140], [147], [160].

As a final remark, it is worth noting that most of the accumulated evidence regarding the conformational transitions of PK has been primarily derived from the comparison of static structures of the enzyme. Currently, there is limited insight about the speculated transition events from a dynamical point of view. To the best of my knowledge, only a few dynamical studies of PK structures have been conducted [141], [160], [164], [166]–[170], with the study by Naithani *et al*. [164] being the only one specifically aimed at elucidating the dynamical nature of the allosteric transitions in PK via MD simulations. The implementation of further dynamical studies is crucial for attaining a holistic view of the mechanism of the enzyme by providing a more direct identification of the corresponding dynamical events.

1.3 The clinical significance of protein variants

1.3.1 Single amino-acid variants

The central role of proteins in biology may be highlighted from a different perspective when we consider their implication in human health. A significant portion of the known diseases are directly related to the deficiency and/or dysfunction of one or several proteins. We have seen how the functional capabilities of proteins can be understood from the intricacies of their structural traits and dynamical behavior. In turn, such macromolecular features primarily emerge from the specific constitution of their sequence [1], [8], [9]. It is the amino-acid chain that confers the combination of physicochemical properties that are needed to adopt particular folds and to form regions with specific functional purposes (flexible hinges, pockets with chemical complementarity with a ligand, chemical groups optimally oriented for enzymatic reactivity…). Given this degree of complexity and fine-tuning of protein structure and dynamics, it is not surprising that even the subtlest variations in protein sequence may result in functional alterations. Since the information to produce proteins is encoded in genes (*i.e.*, at the DNA level), the occurrence of disease-causing protein variants is rooted in the phenomenon of genetic variation.

Mutations are alterations in the DNA sequence that occur from time to time due to a number of events such as DNA replication errors or contact with mutagenic agents [171]. Depending on the DNA region where the change in sequence is produced, a mutation will have diverse effects at the protein level. A broad classification of mutations consists in distinguishing those that affect coding and non-coding regions of DNA. The former introduce changes at the sequence of genes, and thus may possibly perturb the native constitution of protein sequences. The latter affect regulatory elements and other regions between genes, and generally do not modify the sequence of proteins but may alter their expression levels [172].

Among the different types of mutations that affect the coding region, some may have a dramatic impact whereas others may only produce a minor perturbation or even be totally innocuous. Of course, there is a deductible general correspondence between the portion of the sequence that is affected by the mutation and its potential functional consequences. The so-called nonsense mutations introduce a change in the DNA sequence that prematurely stops the production of the protein, generating a shortened protein that is likely non-functional. The omitted fragment will be shorter or larger depending on the position of the mutation. Frameshift mutations disrupt the reading frame of the gene, causing the misreading of the DNA encodings for all subsequent amino acids of the chain and producing a meaningless peptide from the mutation point onwards. Both nonsense and frameshift mutations likely produce non-functional proteins due to the significant loss of information. Generally, they can be automatically classified as damaging [173]–[175]. Nevertheless, a certain tolerance to the alteration of such mutation types has been recognized when they occur near the carboxyl terminus of the protein, thus minimizing their magnitude [174], [176]. Also with high probability of being damaging, mutations that occur at the splice sites of the gene disrupt the splicing machinery that is responsible for processing the final mature RNA after transcription. This kind of mutation leads to the production of an abnormal protein that either skips extensive fragments or includes non-functional peptides[177].

The most commonly observed type of mutation affecting protein-coding genes are missense variants [176], [178]. These cause the replacement of an amino acid for a different one at a given position of the protein sequence. For this reason, missense mutations are also known as single amino-acid variants (SAVs), especially in the clinical context. SAVs are particularly interesting from a medical point of view since they are implicated in a wide range of human diseases, as they are often the direct cause for a protein to partially or totally lose its function [173], [179]. Amino-acid replacements can affect to a different extent the stability, catalytic efficiency and regulatory properties of proteins. The possible functional consequences of such molecular perturbations range from having no effect at all (innocuous, benign or neutral) to completely impeding protein function (damaging or pathogenic). In the majority of cases, elucidating the relationship between the nature and location of the replaced amino acid and the type of molecular perturbation is not always evident [121]. Thus, the exact functional effects of SAVs are often difficult to predict or assess [173], [175].

Pathogenic SAVs may disrupt sites that are critical in protein function. They can impair biochemical function by hindering catalytic or binding efficiency at active sites or ligand-binding pockets [173], [180]. At protein-protein interfaces, they can prevent the formation of protein complexes [181]. They can alter the flexibility profile of certain regions that are key for native dynamics such as hinges [182]. They can prevent the sampling of the proper conformations by altering local or global folding, impeding the formation of disulfide bonds or in general by decreasing structural stability [173], [175], [180]. They may also affect regulatory properties by impairing protein expression or affecting post-translational modification sites [175], [180], [183]. On the other hand, if the given amino-acid substitution maintains the needed physicochemical properties of the position where it occurs, the change may be imperceptible. This is the molecular basis of benign SAVs. For instance, a benign SAV could consist in the substitution of an amino acid that only provides structural support (*i.e.*, it forms non-specific interactions with its surroundings) with another that displays similar capacities. As a final remark of SAVs, it is also interesting to note that they could also lead to the enhancement of the original function of the protein or even drive the acquisition of a new one [175]. These possibilities lie outside of the clinical setting, but are the basis of positive selection and genetic diversification in evolution.

Finally, indels (*i.e.*, insertions and deletions) of one or a few amino acids anywhere in the protein sequence can also be a product of mutations. As with missense mutations, the functional consequences of indels are hard to assess because they can give rise to a wide range of effects [175]. The study of such variants in a comprehensive manner has been hampered by their lower frequency of occurrence. For an indel to occur at the protein level, the corresponding genetic mutation must involve three adjacent positions (or multiples of three), so that the reading frame of the gene is maintained. The majority of small natural indels of amino acids that are tolerated affect either solvent-exposed regions, intrinsically disordered regions, or the termini regions of the protein [174].

1.3.2 Towards personalized medicine

The types of protein variants reviewed above can be involved in diseases as the resulting products of mutations with a genetic origin. Thus, from a medical point of view it is of paramount importance to delve into the prognosis, diagnosis and treatment of genetic diseases. Contemporary medicine benefits from an interdisciplinary effort that tackles the study of health problems from diverse angles. In particular, the mission of the discipline known as medical genetics is to study hereditary diseases. In other words, to detect and monitor the incidence and transmission of disease-causing mutations both at the levels of population and individuals.

The elucidation of pathogenic mutations at the genetic level is not always straightforward since they coexist among a rich pool of neutral variants. Intrinsic DNA variation is a natural part of life. Every single person has a unique genome. Evolution takes place thanks to this genetic diversity, which maximizes the chance of survival of populations (and, in general, of life beings) by making us able to adapt to the environment [171], [184]. Between human individuals, genomes diverge in a few million sites [176], [185]. The first complete picture of human genetic diversity was achieved in the unprecedented endeavor of the Human Genome Project that achieved full sequencing of the complete human genome in the early 2000s [186]. Later on, several other projects such as the 1000 Genomes initiative (1000G) [185], the Exome Sequencing Project (ESP) [187], the Exome Aggregation Consortium (ExAC) [188] or the Genome Aggregation Database (gnomAD) [189] have achieved important milestones in completing, enriching and refining the study of DNA variation, with particular emphasis on coding regions.

Sequencing methods have been experiencing constant growth in the last couple of decades. These technological advances, known as next-generation sequencing (NGS), have brought cheaper and faster sequencing methods; they have facilitated an explosion in the number of species and human genomes sequenced in the last years [171], [172], [183]. As this trend has been increasing, so has the expectation towards personalized medicine [190]. Personalized or precision medicine can be understood as the customized medical care for the unique condition of each patient with respect to their genetic makeup [191]. Thus, personalized medicine benefits directly from the generalized application of NGS techniques in different areas of medical practice, such as diagnosis, prognosis and therapy. Sequencing biological or clinical samples offers insights into the relationships between the human genomic variants and the observable traits of disease symptoms and responses to treatment. This medical strategy not only enables more accurate diagnosis and the identification of diseases at the molecular level, but also has a major role in unveiling which genetic profiles respond better to certain drug types and treatments [172], [184], [190], [192]. In summary, the purpose is to take into account the specific needs and singularities of every patient to administer personalized treatment.

The exponential growth of human genomic information is gathered and integrated in databases that help us understand and classify the patterns and occurrences of genetic variants. One of the most comprehensive and well-known public repositories is the Single Nucleotide Polymorphism database (dbSNP) [193] which serves as both a publication and permanent archive for variation data. The Ensembl Variation database [194], [195] is another useful infrastructure that integrates genetic variation data from several other sources and provides a set of interconnected components for data visualization and analysis. Variants can be classified according to the type of genetic alteration and the frequency of observation within the population. Traditionally, a variant that is observed with at least 1% of frequency in the population is termed a polymorphism. Now, in the era of NGS and personalized medicine, this terminology needs reassessment because rare variants may become polymorphisms or *vice versa* according to the population analyzed, which is particularly confusing on the basis of their disease-causing capacities [192]. For instance, the most common mutation type is the single nucleotide variant (SNVs), that accounts for around 90% of the whole genetic variation between human individuals [173], [185], [188], [196]. When a given SNV is observed with at least 1% of population frequency, it has been traditionally called a single nucleotide polymorphism (SNP). SNPs have been generally described as harmless, being markers of ancestries and human subpopulations, however some have been associated with clinically relevant profiles consisting in the predisposition to certain traits, including diseases [172], [192]. Given that it is not possible to classify the functional role of variations according to frequency in the population or their capability to cause a disease [192], in the present work the most generic nomenclature is adopted, *i.e.*, SNV. The second most common type of genetic variation in human genomes after SNVs are insertions and deletions (indels). Indels that affect just one position have the highest occurrence. In general, the larger the length of the affected fragment, the less frequent the indel [174]. Accordingly, the smallest portion of genetic variation comes in the form of larger altered DNA segments, called structural variants. These encompass mutation events of diverse types and sizes, such as translocations, duplications, and inversions [196].

Mutations can arise in any cell of the organism, and they will be hereditary when they affect the germline cells. Genetic variation is subjected to the effects of evolutionary pressure, meaning that it follows the rules of both positive and negative selection [171]. This means that the prevalence of a given mutation will depend on the effects that it produces. Neutral mutations with no functional significance persist by chance in the absence of selective pressure. Conversely, mutations with deleterious effects, either on viability before reproduction or on fertility, are less frequently transmitted to subsequent generations. By this reasoning, the observed frequency of pathogenic mutations (especially those that are more severe) will be necessarily lower than predicted by the generic rate of occurrence of mutations, or directly not observable if they are lethal [196]. This phenomenon can be particularly evidenced in the fact that the observed mutation rate is higher in non-coding regions than in coding regions since the former will not alter the protein product. In coding regions, the observed mutation rate is approximately half of the general genome, with synonymous variants being more prevalent than missense or nonsense ones [188]. The case of genetic indels at the coding region is very illustrative. Indels of one or two consecutive positions at a gene will alter the reading frame and give rise to frameshift protein variants, whereas indels of three consecutive positions will not (they will produce an amino-acid indel instead). Since frameshift mutations are likely more deleterious than amino-acid indels, genetic indels of one or two positions are less common even if they have more odds of happening (as they affect a shorter DNA fragment) [174].

Of course, the strategies of variant annotation are in turn influenced by this scenario. The information obtained from large-scale sequencing projects such as 1000G or gnomAD gives an idea of the prevalence of benign variants. On the other hand, the same kind of initiative has also been conducted for the study of complex diseases; the so-called Genome Wide Association Studies (GWAS). GWAS initiatives collect the genetic sequences of individuals suffering from certain diseases, enabling the contrast of the genomic information from healthy and sick individuals. In other words, they focus on revealing the potentially pathogenic variants associated with those diseases and enriching their annotation [174]. For instance, The Cancer Genome Atlas (TCGA) [197] and the International Cancer Genome Consortium (ICGC) [198] have sequenced thousands of samples from patients of cancer to better understand this disease and find markers at the level of genetic sequence.

1.3.3 Diagnosis of genetic disorders

Analyzing the sequence of patients suffering from certain diseases that are suspected to have a genetic origin is a valuable asset to confirm diagnosis. This practice is increasingly becoming available and prevalent in the clinical setting, outside large-scale sequencing projects, thanks to NGS techniques [175], [192], [199]. In order for this practice to fully fulfill the promise of personalized medicine [190], clinical genomics and genome informatics face the colossal challenge of the accurate interpretation of genetic variation [172], [175]. The major problem in ascribing functional roles and/or clinical-significance labels to variants identified from sequencing studies lies in being able to distinguish harmful mutations from the background variants without significant effect on human health [172], [200]. Indeed, only accurate discrimination of variants will help in understanding the etiology of mutation-related diseases [179] and providing proper diagnostic and progression assessments [201], [202].

The discovery of novel variants in clinical research starts with the sequencing of the particular genes and/or regulatory elements potentially suspected to be involved in the impaired physiologic functions of the disease. Then, the interpretation of variants begins with the annotation step, which consists in providing the location of any found genetic variants using genome coordinates [172]. The obtained sequence information of the patient should be contrasted with the reference genome information and the known variants that may have been characterized previously. Whether a novel variant becomes a candidate for being causative of the disease or not depends on the prediction of their impact on protein functions or genetic regulation properties. Such an assessment is primarily based on the location of the variant in the coding or non-coding regions [172]. Variants in the non-coding regions are often connected to complex diseases. However, generally they are not prioritized for further assessment because currently there is insufficient understanding of the regulatory machinery encrypted in non-coding DNA, leading to poor accuracy of prediction [176].

Most well-annotated genetic diseases are directly associated with coding variants, which fortunately allow further structural analysis and thus are the basis of variant interpretation [172], [175]. Among the coding variants, synonymous changes (*i.e.*, those that do not alter amino-acid encoding) rarely influence the protein product, however they have been associated with non-neutral effects, such as influencing alternative splicing, mRNA stability, or translational efficiency [176]. On the other hand, the impact of non-synonymous variants can be predicted in terms of the expected alteration of the amino-acid encoding pattern, with nonsense and frameshift variants usually being assigned damaging labels by default [176]. Indeed, SAVs are the protein variants that are the most particularly interesting from a medical point of view, given their widespread incidence in the average human genome [173], [176], [178], [179]. SAVs are implicated in a large portion of the known inherited diseases, being causative for more than half of them [203]. Therefore, they constitute an important focus in clinical research [204]. As mentioned earlier, amino-acid substitutions can result in a wide range of possible molecular perturbations that may be not trivial to identify, hampering the accurate prediction of their possible functional effects [121], [173], [175], [179], [201]. Since (novel) SAVs will often be found upon sequencing and there may be discrepancies between their predicted and actual clinical consequences [178], there is an active field of research devoted to further improving our ability to discriminate between pathogenic and neutral SAVs [201].

The interdisciplinary community of professionals involved in medical genetics have provided consensus guidelines for the evaluations of NGS applications for the diagnosis of genetic disorders [202], [205]. A clear message given by such evaluations is that the variant-disease relationship cannot be conclusive from the information of sequence alone. As a matter of fact, various additional assessments must be conducted to gather sufficient evidence to support or reject pathogenicity propositions. Ultimately, connecting variants either to disease or innocuousness is a complex, multistep process where expert review should always be required [176]. In the consensus guidelines issued by the American College of Medical Genetics and Genomics (ACMG) in association with the Association for Molecular Pathology (AMP) and the College of American Pathologists [205], they propose standardized terminologies to define clinical significance, together with workflows and sets of rules for assessing the evidence for a case. The paramount goal is to avoid wrong clinical decisions based on hasty pathogenicity predictions.

In general, most of the best-established practices and criteria, including the ACMG/AMP guidelines, are only applicable in the particular case of monogenic or Mendelian disorders. Monogenic disorders are diseases caused by alterations on a single gene, hence there is a direct relationship between genotype (*i.e.*, the particular genetic sequence carried by an individual) and phenotype (*i.e.*, the detectable expression of the genotype, or the manifestation of symptoms in the context of a disease). Thus, in contrast to other more complex diseases, monogenic disorders constitute the simplest paradigm when it comes to diseases with genetic origins. Indeed, they present the opportunity to fully comprehend the molecular basis of the protein-disease relationship with minimized interferences of other known or unknown external factors. The clinical research on monogenic disorders has been greatly empowered by the rising applicability of NGS technologies and, in turn, it helps consolidate the framework of personalized medicine. Typically, diseases arising from monogenic disorders are caused by only one or two variants (frequently, SAVs) in a human individual. However, fewer than 50% of monogenic disorders are resolved after sequencing affected families [176]; this fact reminds us of the fundamental challenges that medical genetics still has to accomplish.

While the results of the efforts in the identification and characterization of pathogenic variants are usually published in research articles, the field has greatly benefited from the creation of publicly available databases of human variants with clinical annotations [183], [201]. The Online Mendelian Inheritance in Man (OMIM) [206] represents the most complete resource of curated genes related to monogenic disorders. OMIM includes comprehensive descriptions on genotype-phenotype relationships developed and the variants known to affect function, with references to relevant studies. Databases such as SwissVar [207] (currently discontinued; its data has been transferred to the UniProt Knowledgebase [208]), ClinVar [209], Leiden Open Variant Database (LOVD) [210], and the Human Gene Mutation Database (HGMD) [203] provide public portals to submit and query entries of variants with associated interpretations of their clinical significance. Depending on the database, the sources may include GWAS and/or individual submissions with experimental evidence that might subsequently be subject to different levels of data curation. In general, these repositories are mainly focused on germline mutations in monogenic disorders, and many entries correspond to SAVs.

Although we are still far from having a complete catalog of disease-causing mutations, the databases keep growing as clinicians and researchers around the world keep carrying out the arduous task of interpreting the clinical consequences of variants.

1.3.4 Pyruvate kinase deficiency: a monogenic disorder

Taking pyruvate kinase as an example, a metabolic disorder called pyruvate kinase deficiency (PKD; OMIM entry 266200) arises from the functional disruption of the PKR isoenzyme. Given the central role of PK in cellular metabolism, it is not surprising that health complications arise when the enzyme malfunctions or its regulatory systems fail. Indeed, altered PK activity and expression levels of the enzyme have been associated with various types of human cancers [55], [125], [128]. In comparison, PKD has a considerably simpler etiology and represents an example of a monogenic disorder.

PKD is the most frequent enzymopathy of the glycolytic pathway. It was first identified in the early 1960s and has been recognized as one of the most common causes of the disease called hereditary (or congenital) non-spherocytic hemolytic anemia, together with glucose-6-phosphate dehydrogenase deficiency [53], [55], [118], [199], [211].

PKD is caused by pathogenic variants in the *PKLR* gene. Although genetic variations of this gene likely affect both the PKL and PKR isoenzymes, the liver tissue is generally resistant to PKD because hepatocytes normally synthesize higher levels of the enzyme and/or have residual PKM2 activity. Conversely, erythrocytes are highly dependent on PKR because they lack a nucleus and mitochondria, thus relying on glycolysis for maintaining cell integrity and function. Erythrocytes deprived from sufficient ATP levels experience loss of membrane plasticity, cellular dehydration, and premature destruction in the spleen or liver [118], [211], [212].

The exact prevalence of PKD is unknown, although it has been estimated to be 1 to 8 per million, with a worldwide geographical distribution. Certain communities exhibit higher frequencies due to a founder effect or the potential protective effects from malaria [211], [212]. The disease is inherited in an autosomal recessive manner, thus being manifested only in individuals with compound heterozygous or homozygous genotypes for pathogenic mutations [53], [118], [199]. PKD is highly heterogeneous from biochemical and genetic points of view. Homozygotes generally retain <25% residual PK activity *in vitro*, whereas heterozygotes have 40% to 60% activity [211]. This variability has been related to possible differences in metabolic or proteolytic activity between individuals, or to the compensatory persistence of the PKM2 isoenzyme [118], [213]. Moreover, patients may present a broad spectrum of clinical symptoms with highly variable clinical severity, ranging from life-threatening neonatal anemia to mild symptoms or even fully compensated anemia [54], [118]. Patients may also face difficulties in accessing diagnostic testing, or may receive regular transfusions, thus complicating interpretation of diagnostic tests. Under all these circumstances, PKD is likely underdiagnosed and may have a prevalence higher than estimated [199], [211], [212]. For this reason, in the last years there have been collective efforts between international experts to study the current gaps in diagnosis of PKD and establish more robust diagnostic guidelines [199].

A range of symptoms have been associated with PKD, with possible additional complications that arise from chronic hemolysis, and variable manifestation at different ages. Some symptoms have a substantial impact on the everyday quality of life, whereas others are more bearable. The list includes: fetal hydrops, jaundice, scleral icterus, splenomegaly, anemia, low energy levels, irritability, fatigue or poor concentration, the formation of gallstones, increased risk for thrombotic complications, osteoporosis, iron overload, extramedullary hematopoiesis, and pulmonary hypertension [55], [211], [212]. An initiative from an international group of PKD patients maintains a website to provide comprehensible information about the disease, adapted for the general public, and to generate a community that can share everything that accompanies living with this disease (https://pyruvatekinasedeficiency.com).

Treatment approaches have been traditionally based on supportive measures, which include regular or intermittent blood transfusions, splenectomy, cholecystectomy, and iron chelation therapy to prevent organ damage from iron overload [54], [118], [211], [212]. More recently, other disease-modifying therapies have arisen or are in clinical development. Hematopoietic stem cell transplantation is a curative treatment option for PKD, although no clearly described indications exist and there is considerable associated risk. Gene therapy has also been explored with promising preclinical results (a Phase 1 clinical trial is currently ongoing). This treatment strategy has been successful in diseases caused by monogenic defects and other erythrocyte disorders [211], [212].

The recent incorporation of PK activation therapy with oral drugs has represented a significant step forward in the development of treatment strategies of PKD. In the last few years, the development of the synthetic compound known as mitapivat (or AG-348) has been the most notable. This small molecule is an allosteric activator of PKR that binds to a latent allosteric site different from the canonical FBP and amino-acid binding sites, located in a buried cavity between the N-terminal domain of a subunit and the A and C domains of the adjacent subunit across A-A' interface [53]–[55]. The design of such a compound was inspired by the previous discoveries [214], [215] of other small-molecule allosteric activators that bind to the equivalent site in PKM2. Mitapivat was recently approved by the Food and Drug Administration (FDA) and the European Medicines Agency, having demonstrated safe and effective treatment of PKD in adults through two Phase 3 clinical trials. Mitapivat is now commercially sold under the name PYRUKYND® (Agios Pharmaceuticals, Inc). A few other synthetic PKR activators are in clinical development. The review article of Van Dijk *et al*. [55] contains a summary of the main results from preclinical and clinical studies of subjects treated with PK activators in hereditary hemolytic anemia.

To date, more than 370 mutations associated with PKD have been identified in the causative gene *PKLR* [54], [199]. Most of these correspond to missense mutations encoding SAVs that affect the structure of the enzyme, its stability, or its catalytic function. Only a subset of PKD genotypes occur relatively frequently, while the majority of patients harbor a unique combination of mutations. Common missense variants include Arg510Gln, which is found in Northern Europe and the USA, and Arg486Trp in the Southern European population. There is relatively little predictive value between genotype and phenotype and the severity of the clinical course [212]. However, the more severe phenotypes are generally associated with less frequent and more disruptive mutations, such as premature stop codons, frameshifts, or large deletions [54], [211], [212].

Biochemical characterization of the protein product of genetic variants may be a valuable tool to assist with diagnosis and genetic counseling [118]. Most of the pathogenic missense mutations cluster in specific regions within the PKR structure, including subunit or domain interfaces and functional sites, affecting to a different extent thermostability, catalytic efficiency and response to the allosteric effector [121]. Identifying the nature, location, and type of molecular perturbations of the replaced amino acid in missense variants may facilitate predicting the functional consequences of other variants of similar nature. However, in general, direct predictions should not be considered solely based on analysis of the molecular properties of the altered molecule, since the clinical manifestations of a genetic disease reflect the interactions of a variety of physiological and environmental factors [118].

1.3.5 Computational techniques and tools in pathogenicity prediction

The common approaches to test the molecular impact of sequence variation and its relation with disease are functional assays, *i.e.*, either *in vivo* or *in vitro* tests of the stability and/or biological function of the RNA or proteins involved. In order for functional assays to be as helpful as possible, experiments should closely reflect the biological environment. For example, assaying enzymatic function directly from biopsied tissue from the patient or an animal model provides stronger evidence than expressing the protein *in vitro* [205]. Remarkably, the method called deep mutational scanning (DMS) is becoming an invaluable tool for experimental evaluation of SAVs by enabling the systematic assessment of the effects of hundreds or thousands of variations on a given target property such as cell growth or ligand binding [175]. Naturally, these kinds of efforts need complex equipment and are both expensive and time-consuming [172], [175], [181]. These drawbacks make the experimental analysis of the sheer volume of variants that are annotated simply impractical [179]. In this context, computational tools for predicting variant impact have emerged as a promising alternative. The use of these tools as a way to complement and strengthen the quality of clinical assertion can help bridge the gap between the vast amount of genomic data generated and the limited known genetic evidence [172], [183], [205].

In the last twenty years, more than 50 computational predictive tools that can aid in the interpretation of sequence variants have been released, both publicly and commercially available [172], [175], [200], [201], [204], [205], [216], [217]. The computational approach used by each tool may differ depending on the type of variant that is being assessed. The majority of them are specifically focused on addressing the functional effects of SAVs, which is not accidental given their already stated relevance both as a frequent cause of diseases and as an object of study for their effects in protein structure and function. Other tools specialize in the effect of splicing variants and non-coding mutations [205].

In general, all computational tools that predict the consequences of SAVs follow a similar procedure, regardless of their particular method. Firstly, the properties of the variant are evaluated to get insight into its potential impact on the protein structure or function. Each predictive tool takes into account a different biochemical or biological basis to perform the forecast. Then, at the light of the resulting feature set, the variant is labeled as pathogenic or not [173], [183]. The outcome of these approaches must be taken just as an assistance to clinical diagnosis. As it happens, if a variant is predicted to be “damaging” or “pathogenic”, it does not necessarily lead to a specific phenotype or disease condition. Experimental validation is still indispensable [176], [183], [202].

In a recent review by Liu *et al*. [183], they classify the existing computational predictive tools in six classes, based on the characteristics and included features of each type. The first and predominant class comprises homologous sequence-based predictive tools that base their outcomes in the evidence provided by comparative genomics. Indeed, the degree of evolutionary conservation of each position in genetic sequences is informative about the tolerance to certain changes. In general, the more conserved the position, the more likely for an amino-acid substitution to be deleterious. This class of predictive tool typically relies on the construction of a multiple sequence alignment of several similar protein sequences and the subsequent extraction of scores and metrics, hidden Markov Models or machine learning models. The tools SIFT [218], PANTHER [219], MutationAssessor [220], PROVEAN [221], and FATHMM [222] are among the most popular tools of this class.

The second class comprises structure-based predictive tools, which assess the impact of SAVs in proteins in the context of the macromolecular structure. Most tools calculate the change in free energy ($\Delta\Delta G$) upon an amino-acid replacement, which gives an idea of the change in conformational stability by approximating structural energy via physics-based and/or knowledge-based terms [175], [183], [204]. This metric is not exclusive to the field of pathogenicity prediction; it is rather commonly used to assess folding or protein interaction changes upon mutation [204]. In comparison to benign variants, disease-causing mutations predominantly impact the core of the protein, making structural destabilization the major cause for pathogenicity in monogenic disorders [175], [183]. A large volume of pathogenic SAVs are also found in structural and functionally essential regions, such as protein-protein interfaces or binding sites. For this reason, structure-based predictive tools often aim to improve predictions by taking into account structural annotations to assess whether the mutation is occurring in a hot spot or a functionally-relevant site [181], [201]. However, structure-based predictive tools have been hampered by the limited availability of structural data in comparison to sequence data, which is available for the entire human proteome [172], [183]. The days of this major drawback may now be coming to an end thanks to the newest deep learning approaches such as the AlphaFold project [223], [224], which is dramatically expanding the availability of considerably reliable structural models.

While structure-based approaches may give greater insights into the mutation effect than homologous sequence-based approaches, the former have not led to clearly significant increases in prediction accuracy so far [175] and may still be outperformed by the latter under certain circumstances [183]. Thus, combining both strategies into an integrated approach may aid in improving prediction capacity [172], [173], [179], [183]. In this regard, the third class of predictive tools are the sequence and structure combination-based tools, with MutPred2 [225], SNPs&GO [226], PolyPhen-2 [227], PMut [216], and MutationTaster [228] being among the popular ones. In a similar spirit, the fourth class of predictive tools are the so-called meta-predictors, which make predictions by integrating results of pre-existing tools. Meta-predictors aim to leverage on potential complementary performance of selected predictors in classifying variants. They are able to improve prediction performance and avoid some biases attributed to single tools [183]. The tools CONDEL [229], CADD [230], and REVEL [231] are among the most popular tools of this class. A fifth class of predictors classify variants according to population-frequency data, given that the ACMG/AMP guidelines state that a variant with more than 5% of frequency is considered as a stand-alone support for benign interpretation for rare monogenic disorders [183], [205]. For instance, the tool ClinPred [232] trains on clinically curated pathogenic and benign datasets, and also acts as a meta-predictor by including feature scores from 16 other pre-existing tools. The last class of predictive tools overlaps with several of the above but include disease-, phenotype-, and gene-specific features. They are able to specialize by selecting and refining the training and validation datasets, constructing sub-models for each gene, re-constructing the multiple sequence alignments and phylogenetic trees of the targeted genes, and employing additional rule-based classification systems [183].

Yet with all the progress achieved, there are still many challenges that need to be overcome in order for these methodologies to reach sufficient reliability in practical applications. Indeed, the available predictive tools are bringing us closer to personalized medicine: they can be successfully used for variant prioritization [204]. The methodologies that have been reviewed mostly refer to general-purpose predictive tools, which are trained with large datasets of a wide spectrum of sequence variation annotations in order to be applied to any human protein. Nevertheless, in general, the maximum accuracy achieved by general-purpose predictive tools so far is around 80–85%, which strongly limits their usage in clinical diagnosis [173], [175], [233]. Most of these tools also tend to have low specificity: they can overpredict SAVs as deleterious and fail to correctly identify pathogenic variants with a milder effect [205]. Moreover, they have been found to consistently perform badly for some protein families [216]. As an alternative, the implementation of specific predictive tools has also been explored. In this case, tools are trained only with variants from the gene, protein, or protein family of interest, especially when considerable experimental data is available, and entail a subsequent stage of thorough data curation. Specific predictive tools can potentially mitigate some of the limitations of general-purpose predictive tools and outperform them in some circumstances, although combining both approaches is the best strategy for achieving the highest success rates [233].

1.3.6 The use of dynamical features in pathogenicity prediction

Incorporating a dynamics-based rationale is the logical next step to enhance the accuracy of predictions of the impact of SAVs in proteins. Despite the continuous development of computational predictive tools, dynamical features have been largely neglected in the field of pathogenicity prediction. In the last years, the subject has begun to be raised by the community, with the increasing expectation that the integration of dynamical features will be a valuable complement to improve the state of the art [172], [175], [184], [201], [234]. Recently, Galano-Frutos *et al*. [175] have covered the subject in depth and provide insightful considerations for pursuing the goal of interpreting human genetic variations at large scale through dynamical data (and specifically, with MD simulations).

The major obstacle is the high computational cost of the task [175]. As with structure-based pathogenicity prediction, but perhaps more seriously, this issue results in the lack of availability of data. Of course, the assessment of the functional effects of protein variants via dynamics-based methods at the level of an individual study is possible; many instances are commonly found in the literature [167], [235]–[239]. However, the integration of such approaches in tools that can systematically evaluate any variant (especially in a massive way) is still a pending challenge. The problem is also related to the lack of robust and accurate analysis techniques that can quantitatively determine the impact of any mutation in conformational sampling and protein stability [175].

Only a handful of initiatives have released tools for predicting the impact of mutations on protein stability by including assessment of protein dynamics. The ENCoM [240] and DynaMut [241] web servers implement NMA simulations to inspect the set of possible conformational changes from a static equilibrium structure. Although they have not been designed specifically for clinical assistance, the outcome of the predicted change in protein stability upon a given SAV can be useful to speculate about its pathogenicity. However, they do not seem to surpass other predictive tools in terms of maximum performance. Possible reasons may be the limited ability of NMA simulations to sample the native conformational space and the inability to describe solvation interactions [175].

On the other hand, MD simulations provide a more accurate and comprehensive modeling of the dynamical behavior of biomolecular systems (see section 1.4.3). MD simulations have proved to be a powerful approach to detect the minor changes associated with protein variants [172], [184]. They can derive a mechanistic understanding of the impact of SAVs in protein structure and dynamics from realistic simulations (*e.g.*, explicit solvent, physiological temperature and pH). Given the structure of the wild-type protein, a starting model of the mutant structure can be generated by modeling the corresponding amino-acid substitution. The web-based tool ANGDelMut [242] relies on a classical atomistic MD approach (albeit in implicit solvent) to capture certain structural and dynamical features of SAVs and predict mechanisms of functional loss. ANGDelMut, however, is not a general computational predictive tool but rather specific for only the angiogenin protein, which is associated with amyotrophic lateral sclerosis. Another remarkable effort is the work of Fleming *et al*. [234], who implemented a neural network approach that predicts protein thermostability upon mutation using a set of features extracted from MD time-series data.

At present, no other dynamics-based predictive tools seem to have been released. Still, we must not neglect the contribution of independent studies that have aimed to take into consideration the role of dynamics as a source of insight for predicting the implications of mutations, mostly via MD-based approaches. A significant volume of this kind of studies can be found in the literature [167], [201], [204], [235]–[239], [243]–[248]. Remarkably, the Dynameomics project [248] gathered a rich collection of MD simulations of 29 different wild-type proteins and 200 associated SAVs, which provided insight into the molecular basis for structural disruption and destabilization mechanisms. Although the pertinence of modeling SAVs and performing subsequent MD simulations may be arguable, many of these works showed quite good correlations between their predictions and experimental data [175]. All in all, MD simulations are establishing a promising future for personalized medicine. Provided that the trends in both the optimization of MD algorithms and the power of computational hardware keep increasing, and counting with international coordination efforts, the predictive MD analysis of the entire set of possible SAVs of the human proteome could become a reality in the next decades [175].

1.4 The study of conformational ensembles

From the perspective of promoting dynamics-based knowledge to solve biomedical problems, we require techniques that can capture the dynamics of the involved biomolecular systems of study. Dynamical data is obtained when we can gather several instances of the structure of a system that bear a defined time relationship. Such a collection of data is also called a conformational ensemble. Conformational ensembles not only capture the intrinsic dynamics of a molecule, but also account for the additional variability that may arise from uncertainties and statistical noise during experiments and structure determination [249]. Being able to generate reliable and rich conformational ensembles is one of the challenges of contemporary structural biology and bioinformatics [5]. A variety of experimental and computational techniques have been developed to provide insights into protein dynamics by building complete 3D pictures of biomolecules at atomic or near-atomic resolution, and at different time intervals. Each technique has its own advantages and limitations and generally comes with particular spatial or temporal resolution constraints (Figure 1.13). Thus, the choice of method may depend on factors such as the size and the properties of the system under study, or the desired resolution of the final model.

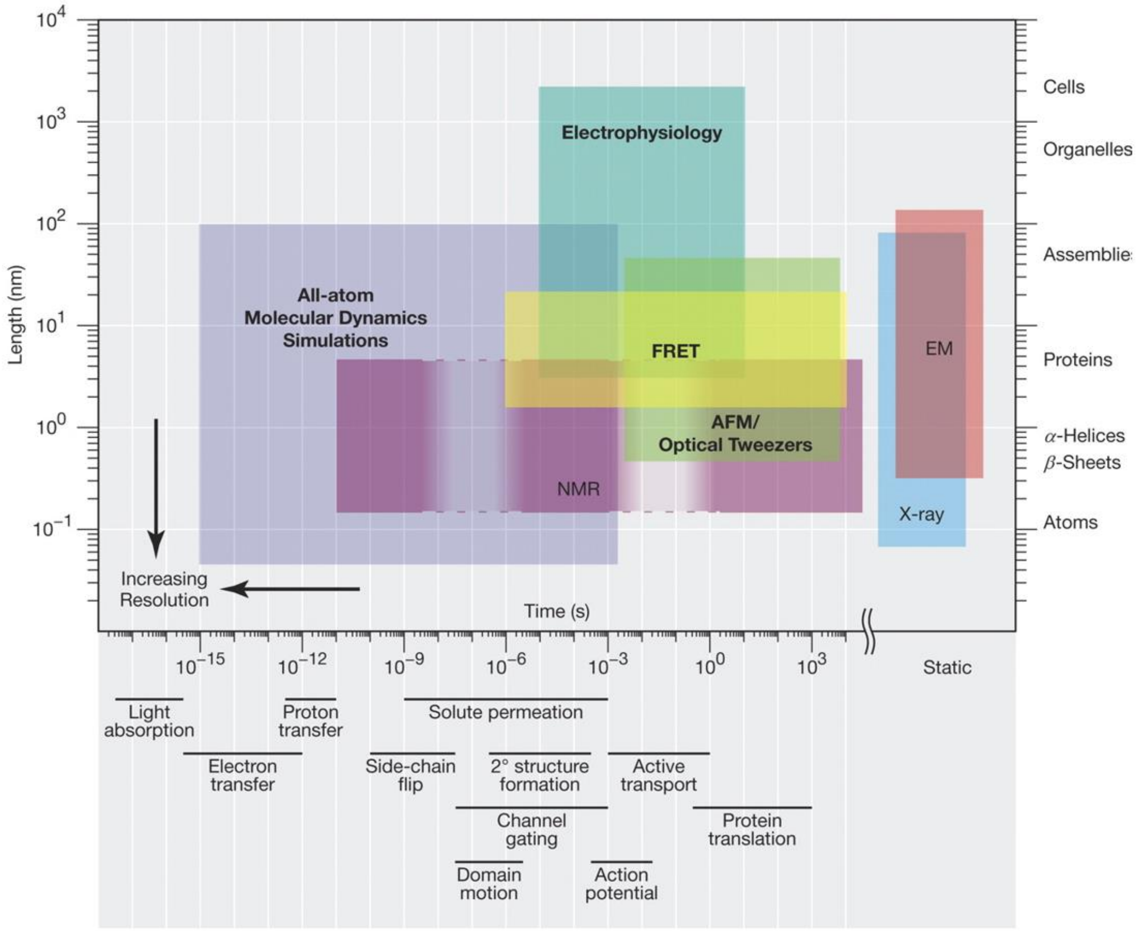


Figure 1.13. Spatiotemporal resolution of various biophysical techniques. Several common biological objects and processes are included next to the axes to compare their relative time and size scales. Techniques capable of yielding data on single molecules (as opposed to only on ensembles) are shown in bold. NMR methods can probe a wide range of timescales, but they provide limited information on motion at certain intermediate timescales, as indicated by the lighter shading and dashed lines. NOTE. Extracted from [250]. AFM, atomic force microscopy; EM, electron microscopy; FRET, Förster resonance energy transfer; NMR, nuclear magnetic resonance.

1.4.1 Experimental methods

The gold-standard experimental techniques for providing starter structures for dynamical analysis are X-ray crystallography, cryogenic electron microscopy (cryo-EM), and Nuclear Magnetic Resonance (NMR) spectroscopy. Over the years, we have accumulated a vast amount of resolved structures from these sources. The majority of these structures are deposited in the Protein Data Bank (PDB; www.wwpdb.org) [131], the reference repository for 3D structure data for large biological molecules (proteins, DNA, and RNA). At the time of writing this thesis, the PDB has more than 195,000 structures, with roughly 86.4% of them resolved by X-ray crystallography, 7.1% by NMR, 6.4% by cryo-EM, and 0.1% by other methods. The PDB has an invaluable impact on structural biology, as it allows researchers to use the available structures for conducting subsequent analyses.

The size of the present PDB contains a substantial degree of redundancy, since the continuous uploading not only encompasses new systems but also additional instances of already deposited molecules. Furthermore, one may often find structures of the same biomacromolecule resolved in different biological conditions (for instance, ligand-bound and unbound forms, or wild-type and mutant forms). These facts are interesting because the different static structures of a system provide a general idea of its conformational heterogeneity, thus revealing some initial clues on how to approach a study of dynamics. In general, the higher the found structural variability between the copies, the broader the array of techniques/studies that will be required to obtain sufficient coverage of the spatiotemporal scale of the functionally relevant dynamical events. Of course, the weaknesses of the PDB are directly influenced by the inherent limitations of the techniques that feed it. For instance, the composition of the PDB is biased, with systems presenting significant flexibility (*e.g.*, IDRs and IDPs) being especially absent.

In X-ray crystallography, also known as X-ray diffraction, a beam of X-ray light is projected towards a crystal, generating a diffraction pattern. The angle and intensity of the diffracted beams can be measured to deduce the 3D structural arrangement of the atoms within the crystal based on electron density signals. The major bottleneck of this procedure resides in obtaining a crystal of the highest possible quality [251], [252]. The crystallization of the protein or nucleic acid under study requires bringing a supersaturated solution of the compound to very specific conditions of ionic strength and co-solutes that ultimately induce the molecules to arrange themselves in a repeating series of unit cells by adopting a uniform orientation. The higher the degree of order within the crystal, the better the quality (and the higher the resolution of the resulting structural model). Such a procedure is often challenging, and it directly prevents flexible regions to be resolved because they cannot achieve sufficient crystal organization. Other hardly crystallizable systems are macromolecules of large molecular weight and some membrane proteins due to their poor solubilization [253]. Moreover, crystallization conditions differ from the conditions of the cell environment, therefore, they can also give rise to technical artifacts. Crystal-packing effects can induce some molecular contacts and interactions that are not present in physiological conditions and are rather a product of the imposed symmetric arrangement [254]. The co-solutes required for fulfilling crystallization can also interact with the structure and disrupt the native local conformation of some sites [93], [252]. Lastly, crystal structures represent the static information from dynamical entities expressed as the average atomic positions of the stabilized conformation within the crystal. When more than one electron density signal is obtained for particular atoms (*e.g.*, side chains of amino acids), crystallographic data sometimes includes the alternative positions with their fraction of occupancy. A rough measure of the atomic mean fluctuation is also reported in the B-factor metric, which provides an estimation of the flexibility of each region of the molecule [5].

Despite the challenges of the technique, X-ray crystallography has undoubtedly been the most important source for structural data of biomolecules, as demonstrated by the composition of the PDB [252]. Improved versions of the technique are focusing on attenuating the constraints of both the crystallization stage and the possible radiation damage of the structure due to long-term exposure to X-rays. The X-ray free-electron lasers (XFELs) are a promising solution, although they can only be produced in highly specialized facilities. When projected in femtosecond-long pulses, usable diffraction patterns are generated before any radiation damage can compromise the sample (serial femtosecond crystallography, or SFX). XFELs can also enable detection of fast dynamical information such as the intermediate states of an enzymatic reaction (time-resolved crystallography) and circumvent the need of cryo-cooling protein crystals which can perturb the native conformational heterogeneity (room-temperature and multi-temperature crystallography) [27], [104], [105], [252], [255].

Cryo-EM is another increasingly powerful method that since the last decade can reach a spatial resolution comparable to that of X-ray crystallography. This technique can be considered an advanced version of electron microscopy, where electron beams pass through a purified sample to yield a reconstructed 3D image. In cryo-EM, prior to electron irradiation, the sample is brought to cryogenic temperatures to reduce radiation damage. An advantage is that the crystallization stage is not needed. Instead, the sample forms an amorphous frozen solid that even allows single particle analysis and eludes the problems of crystal-packing and the need of averaging the signals of high symmetric lattices. This makes it particularly suitable for screening large macromolecules and macromolecular assemblies, while a disadvantage is that smaller systems pose more problems due to possible signal noise. Time-resolved experiments are possible by fast freezing multiple states of a system, however cryo-EM possesses the same limitations to study particularly flexible systems and fragments, with such regions suffering from too low spatial resolution [27], [256].

On the other hand, NMR methods are more proficient in providing both structural and dynamical data since it can quantitatively describe conformer populations and their exchange rates of interconversion. The experimental procedure involved is fundamentally different from that of X-ray crystallography or cryo-EM. The structural information of a given molecule is obtained by measuring the resonance frequencies of atomic nuclei (particularly ^{1}H and ^{15}N atoms) under a homogeneous external magnetic field. The exposure to the magnetic field generates some observables such as the Nuclear Overhauser Effect (NOE) and the J-coupling, which read signals related to the probable distances and interactions between the atoms involved. If the sample contains multiple conformers, they can be revealed by detecting different signal peaks. A major advantage of NMR is that the sample can be directly probed in solution, without the need of crystallization or cooling stages, therefore substantially closer to its native conditions and with the potential to describe intermolecular interactions. NMR has indeed helped in identifying regions involved in ligand binding, protein–protein interaction, and protein–nucleic-acid interaction [252]. However, the size of systems that can be studied through NMR is limited, as the spectral profile gets too difficult to process and interpret for large macromolecules. Moreover, this technique requires relatively large amounts of pure samples, and systems with poor solubility such as membrane proteins are harder to prepare. Contrary to X-ray crystallography and cryo-EM, NMR can resolve highly flexible systems, therefore it is a principal means of study of IDRs and IDPs. Thanks to being able to capture different conformers, it can be employed to capture the transient conformations involved in dynamical processes belonging to the picosecond-second timescale. Nevertheless, there are a few blind spots in the timescale that may be more challenging to resolve [107], [257], [258].

In addition to the gold-standard methods, other experimental techniques can also afford structural and dynamical data and have been useful to discover or confirm functional mechanisms of both proteins and nucleic acids. Hydrogen/deuterium exchange mass spectroscopy (HDX-MS) quantifies the exchange of hydrogen with deuterium in solution when heavy water (deuterium oxide, ^2H_2O or D_2O) is present. The exchange is produced rapidly in amide groups that belong to disordered regions that lack stable hydrogen-bonding. The detection of the macromolecular regions with different exchange rates provides information about the tertiary structure of proteins, and can help elucidate conformational changes upon ligand binding and during allosteric regulation [252], [259]. Cross-linking mass spectroscopy (XL-MS or CL-MS) relies on creating covalent bonds between spatially close regions of protein complexes, and afterwards inducing enzymatic digestion to analyze the resulting peptides in search of the regions enriched in cross-linked peptides. Thus, XL-MS can identify structural regions (domains or even single helices or loops) that are responsible for protein-protein interactions [260]. In single-molecule Förster resonance energy transfer (smFRET) two fluorescent dyes are attached to the biomolecule of study and then a laser beam is employed to produce an energy transfer between the excited donor fluorophore and the acceptor fluorophore. This technique can measure intramolecular distances within proteins and, in addition, can also characterize the times and amplitudes of their modulation during function [261]. Small-angle and wide-angle X-ray scattering (SAXS and WAXS) are X-ray diffraction methods that can be applied to samples in solution or less crystalline than in X-ray crystallography. SAXS can be applied to samples in solution and provides quantitative information about the macromolecular shape and weight by a rapid determination of the radius of gyration (Rg) metric. It can help determine the oligomeric states of proteins at low resolution, as well as the structure and dynamics of IDPs and integral membrane proteins. WAXS can probe smaller length scales and can sense small structural changes in proteins, characterize the breadth of the structural ensemble in solution, and identify proteins with similar folds [252], [262]. Infrared spectroscopy uses infrared radiation to excite vibrational transitions of molecules and determine various levels of their chemical composition and architecture, potentially also in a time-resolved manner. It is a valuable tool for the investigation of protein structure, of the molecular mechanism of protein reactions and of protein folding, unfolding and misfolding [263]. Finally, atomic force microscopy (AFM) relies on a mechanical probe that scans the sample either by contact, oscillating, or force approaches. This technique can provide overall topographical and mechanical data, to some extent time-resolved, and identify assembly patterns, signs of dynamical behavior and the ability to interact with other molecules [264].

All in all, when manipulated properly, experimental techniques are powerful tools to deliver reliable structural data, especially taking into account that they can complement each other to cover systems of different chemical and dynamical natures. However, they are laborious and hard to execute in a systematic way. For instance, in order to accumulate a substantial collection of structural and/or dynamical data of variations of the same system, experimental techniques are simply too costly and time-consuming. Indeed, in addition to the efforts related with performing the technique itself, one must also previously acquire the sample of the actual protein or nucleic acid under study, possibly facing stages of mutagenesis, transfection, purification or even artificial synthesis. Therefore, for certain research scenarios and in general as a suitable complement to experimental techniques, we resort to computational methods, which are nowadays very well established and reliable.

1.4.2 Computational methods: molecular modeling

1.4.2.1 Structure prediction

The 3D structure of biomacromolecules can be predicted and modeled by computational approaches to some extent. Structure prediction covers knowledge-based approaches that derive information from the already available structural data to determine as yet unknown structures. Given their condition of predictive tools, their purpose is to produce models as accurate as possible albeit with approximations. Of course, the existence of this kind of methods could not be possible without a prior robust background of already known structures. In the case of proteins, even though the number of resolved structures has increased exponentially in the last decades, the number of known protein sequences for which structural data still lacks is hundreds of times larger and grows at an even faster rate. Thus, structure prediction methods play a crucial role in filling this widening gap. For this reason, these modeling tools are bringing about substantial advances in clarifying protein interactions and making drug discovery faster, easier, cheaper, and more practical [265].

The main techniques in this field are called comparative modeling techniques, of which homology modeling is considered to be the most accurate. Besides, it is a fast approach with very well defined steps and low cost [265], [266]. Homology modeling operates under the central assumption that the 3D structure of a protein is mainly determined by the inherent properties of its amino-acid sequence. In turn, protein folds tend to be more conserved than sequence during evolution, therefore similar sequences entail very similar structures [8], [9], [265], [267]. In fact, the currently available data suggests that, although the number of possible proteins is essentially infinite, the number of actual folds is limited and can be classified into approximately 2000 different classes. Given two proteins that share a 70% of sequence identity, if we already know one structure, we can expect that the accuracy of performing homology modeling to build the missing structure is comparable to that of resolving its crystal structure. At 25% of sequence identity, the expected accuracy is still reasonably reliable [267].

The procedure of homology modeling starts by searching a database of structures (normally the PDB) for sequences that are homologous to the query sequence. The search is usually driven by sequence alignments, and also by profile-profile alignments and hidden Markov Models. The found sequences with the highest similarity will be the best template candidates for the structure prediction. Among the eligible templates, it is also useful to consider other factors such as phylogenetic similarity, the resolution of the experimental structures and the degree of similarity between their environmental conditions. Then, after a careful adjustment of the alignment between the query and the template sequences, a software of homology modeling is employed to build the backbone of the model and later add the amino-acid side chains. There are several approaches for model building, each with their advantages and limitations. Some of the more popular tools and servers are MODELLER [268], I-TASSER [269], and SWISS-MODEL [270]. The modeling of loops is a particularly delicate part of the process that can be reinforced by other specialized software; loops longer than ten amino acids yield poor accuracy. Finally, the models can be optimized and validated according to several parameters in order to choose the best quality template(s) [265], [267].

The major limitation of homology modeling is, of course, the lack of experimentally resolved structures of the homologous sequences [266]. When no direct homologous sequences or templates are found, the secondary structure must be predicted from the best-matching recognized templates of a database, in what is called threading or 3D-1D fold recognition. Also, the sequence may be segmented, modeled separately and combined back into a full structure. *De novo* prediction is also possible with *ab initio* approaches. In this case, an iterative process generates a large number of potential three-dimensional models with a Monte Carlo type of algorithm and possible refinements via short simulations. Models are either discarded or improved in each iteration. However, both threading and especially *de novo* prediction come with a high degree of uncertainty of the quality of the built models [265], [267]. Contact prediction tools constitute another subfamily of methods that attempt to reconstruct 3D models by building contact maps with predictions of residue-residue distances when no templates are available [271].

More recently, deep learning approaches were introduced in this field with unprecedented improvements in the accuracy of predictions. These employ artificial neural networks to derive hidden features and patterns from large collections of data that can cast predictions on new data [272]. Among the developed tools, the extraordinary success of the AlphaFold project [223], [224] was able to produce reliable models for the whole human proteome, and in its latest release the collection has expanded to cover nearly all cataloged proteins known to science (over 200 million structures). This groundbreaking feat challenges the so far settled notion that only experimental methods were a reliable source of 3D models to perform subsequent analyses.

The modeling of biomacromolecules through structure prediction approaches not only allows to circumvent the cumbersome procedures of the experimental techniques, but also provides direct means to model variants of the same system with little additional effort. This is a powerful advantage. For instance, amino-acid substitutions can be modeled from existing protein structures, allowing parallel study of a wildtype protein and its missense variants, as performed in this thesis. For this purpose, side-chain prediction software such as SCWRL [273] or the BioBB Structure Checking tool [274] can be employed to replace the atoms of a side chain with those of any other amino acid and accommodate them in a proper orientation based on the information from refined rotamer libraries and/or other scoring functions. It is also possible to model protein variants that are more complex than substitutions, such as insertions or deletions. However, the larger the altered fragment, the less reliable the resulting model. In fact, even amino-acid substitutions may render inaccurate models because the possible effects of misfolding cannot be taken into account. For example, it is possible to model a polar residue in a hydrophobic environment by amino-acid replacement, however whether the actual mutated protein *in vivo* would adopt an equivalent fold or fail to do so is rather unpredictable.

1.4.2.2 Conformational sampling: biomolecular simulations

While the methods covered in the last sections primarily provide structural models, with possible insights on the flexibility profile of the structures (*e.g.*, with NMR), dynamical information can be directly modeled with conformational sampling methods. These are computational simulations based on algorithms that are designed to sample the conformational landscape with different strategies and levels of resolution. Given a starting structural model, biomolecular simulations aim to represent the physical properties of the molecular system, quantify its potential energy, and predict how the structure will move. This family of methods are theoretical approaches with varying degrees of accuracy. Although no model can achieve a perfect description of reality, the current implementations of physics-based terms to describe molecular forces and interactions can indeed yield strong and reliable predictive power. However, coping with the highest levels of complexity generally comes with disappointingly impractical computational costs. Thus, every technique needs to introduce certain approximations in a balance between resolution, accuracy and computational feasibility. Among the variety of conformational sampling techniques, perhaps the best known are molecular dynamics (MD) simulations. In MD, a conformational ensemble is generated in an iterative process that integrates equations of motion for all the components of the system at successive time points. MD techniques and software can be implemented to operate in the different existing levels of complexity of molecular modeling.

Atomistic-level simulations constitute the models with the highest spatial resolution since they treat the system with atomic detail. The movement of every single atom is explicitly modeled to analyze molecular flexibility [5]. Within this category of simulations, the most rigorous quantitative description is the one provided by pure quantum mechanics (QM) models. An ideal prediction of dynamics in the atomistic level should be achieved by taking into account atomic nuclei and electrons and solving the corresponding time-dependent Schrödinger equation. In practice, an exact solution to the equation cannot be obtained for systems with large degrees of freedom (*i.e.*, no more than a few particles). Therefore, QM approaches usually rely on the theoretical approximations that have been developed in the last decades [5], [275]. Some methods benefit from the widely used Born-Oppenheimer approximation that allows to disconnect nuclear and electronic movements, given that nuclei are much heavier and slower than electrons. This approximation serves as a basis for treating many-body electronic systems with the Hartree-Fock (HF) method (also known as the self-consistent field method), which introduces molecular orbitals to represent electron movements in an averaged way. The HF method, in turn, is taken as a reference state in the so-called post-HF methods to correct some of its inaccuracies by taking into account electron correlations. In contrast, Density Functional Theory (DFT) methods generally offer a positive trade-off between accuracy and computational efficiency. They have become a preferred framework for modeling complex chemical systems like metalloproteins and studying reaction mechanisms at the active site of enzymes [252], [276], [277]. In DFT, molecular energy is expressed purely in terms of electron density. This simplification provides reliable energy estimates with a moderately low impact on accuracy and allows to simulate systems with sizes of 100-1000 atoms at a more affordable computational cost [275], [278]. Both in HF and DFT, molecular orbitals are constructed as linear combinations of a set of atom-centered functions, known as the basis set, that represent the individual atomic orbitals [279]. Given that the valence region of the atom carries more importance in chemical bonding, the so-called split-valence basis sets have become a popular option. Split-valence basis sets represent valence orbitals with several basis functions and leave core orbitals represented with just a single function, thus providing increased accuracy in the former. Even better descriptions of the electron distribution are achieved by adding polarization and diffuse functions to the basis set. The former are functions with higher angular momentum that help to describe how the electron cloud of an atom polarizes (distorts) under the influence of the other atoms in the molecule, while the latter are functions that spread out further from the nucleus that allow a better description of loosely bound electrons [276], [279], [280]. The MD techniques that operate in the QM regime are Born-Oppenheimer MD and Car-Parrinello MD. The former has been used to study protein dynamics; however, without further simplifications this technique suffers from a prohibitive computational cost. The latter allows to perform calculations with DFT methods, with established applications in material sciences and increasingly in proteins [5], [281].

On the other hand, the molecular mechanics (MM) description further increases computational efficiency by neglecting quantum effects and relying on classical mechanics. In MM methods, atoms, forces, and interactions are described with a set of parameters known as the force field. Force-field parameters are typically determined by deriving the data obtained from high quality QM calculations and/or experimental studies [5], [282]. Atoms are represented as spherical particles with associated values of charge, mass and volume. The degrees of freedom of the electronic configuration are no longer treated, but such information is implicitly present within the covalent bonds of the structure, which are represented as spring-like strings that follow Hooke's law. Angle and dihedral (torsion) parameters complete the set of classical potentials that describe the interactions between consecutively covalently-bonded atoms. Non-covalent interactions are described by the non-bonded terms: the electrostatic and van der Waals forces, modeled through Coulomb and Lennard-Jones potentials. Some force fields are especially accurate for describing particular molecular systems depending on their chemical composition (nucleic acids, proteins, carbohydrates and lipids). Through the years, the quality of the force fields has been consistently improving thanks to the continuous refinement of the parameters, although they still possess limitations [282]–[284]. With MM approaches, larger systems can be simulated in generally shorter computation times, thus providing the best option for simulating most biological systems. Nevertheless, given that they cannot describe changes in the electronic distribution, chemical rearrangement events such as enzymatic reactions cannot be modeled. Classical MD simulations lie within the MM level of conformational sampling. In this case, atomic movements are calculated using Newton's laws of motion by integrating the forces acting on each atom at each timestep [285]. Classical MD (widely known as just "MD") is the main source of dynamical information used in this project. Throughout the thesis, further considerations will be discussed on this subject, starting with a more extensive review at section 1.4.3.

The hybrid QM/MM methods aim to benefit from the high accuracy of QM and the computer efficiency of MM. The QM treatment allows modeling of the electronic rearrangements involved in the breaking and making of chemical bonds, while the MM treatment allows for the efficient inclusion of the wider environment and its effects on the reaction energetics [286]. Thus, in the QM/MM framework, the structure is divided into two parts: a small one (the QM part) that covers the essential region where the chemical reaction is expected to occur, and a large one (the rest of the structure; the MM part) that is included to complete the modeling of the whole macromolecule where a quantum level of treatment is not needed. The choice of the level of theory to treat the QM part is crucial and should be suitable for the reaction of interest, while atomic interactions at the boundary between both parts can be treated with different strategies [275], [286]. QM/MM calculations have been a fundamental tool in computational enzymology for several years. They can provide details of transition state structures that are otherwise impossible to pinpoint in experimental studies. These methods have also delivered insightful contributions in drug design, drug metabolism and biocatalyst design [275], [286].

With a lower spatial resolution, we can find the coarse-grained models. In the coarse-graining strategy, the degrees of freedom of the system are further reduced by grouping together several atoms in a single particle, pseudo-atom or "bead". Of course, the aim is to benefit from the dramatic increase of computational efficiency when the associated loss of accuracy is not an issue in order to tackle the scientific question of interest. In this sense, coarse-grained particles can represent different atomic formations. For example, in proteins, a residue can be represented with a few beads covering the most important chemical moieties, or rather by a single bead [287]. The fewer number of particles, the lower the complexity of the model. For obvious reasons, coarse graining is not compatible with QM or MM methods as described above. The approaches to describe the potential energy of the system fall into two categories. On the one hand, structure-independent models treat the system with knowledge-based potentials and are especially useful in applications of protein structure prediction. On the other hand, structure-based models are closer to MM methods in the sense that they employ force fields of non-physical statistical potentials that have been carefully calibrated in order to reproduce the structural and flexibility properties of macromolecules [5], [287]. Within the latter category, elastic network models (ENM) [288] represent the connections between particles with harmonic springs with reference distances extracted from experimental structures. Gō models [289] incorporate terms for non-bonded interactions. The MARTINI force field [290] adds torsional restraints aimed to specifically maintain native secondary structure. Some of these approaches can be integrated in very computationally efficient MD simulations. Coarse-grained MD has been successfully applied to study protein–protein interactions, biomolecular motors and particularly large systems of cellular biology such as ribosomes and membrane proteins [287].

Besides MD simulations, other sampling algorithms and strategies are widely used in the study of macromolecular dynamics. Normal mode analysis (NMA) is a technique that studies the vibrational harmonic oscillating motions of a mechanical system. NMA takes an experimental (or reference) structure, assumes that it corresponds to a local minimum conformation, and derives a set of normal modes of motion along the corresponding potential energy basin [291]. While standard all-atom NMA relies on classical force fields, the technique is frequently integrated with the ENM regime, thus replacing the force field with Hookean potentials (ENM-NMA), often also employing a coarse-grained resolution [5], [291]. Unlike in MD, the conformations sampled by NMA are not time-correlated. However, this method provides very relevant insight on the flexibility profile of the (native) structure. Moreover, the vibrational normal modes exhibiting the lowest frequencies (the so-called soft modes) in proteins describe large motions that are considered to be biologically functionally relevant. These soft modes can also be studied in a comparable manner with MD simulations followed by subsequent essential dynamics analysis [292], [293].

Monte Carlo (MC) simulations use stochastic methods to generate new conformations of a system and increasingly sample diverse regions of the conformational space. From a starting structure, the MC algorithm induces potential random movements to propose a new conformation. Then, the generated conformation is accepted or rejected based on the relative energy of the previous structure. Again, this sampling method does not deliver time-correlated dynamical data. Instead, an iterative algorithm is implemented with the aim of covering a statistically meaningful region of the conformational space in no particular order. MC simulations can be used with rationally selected sampling variables and constraints that can improve its sampling efficiency by avoiding certain regions in the Cartesian space that will likely yield rejections [5], [294].

Although not a conformational sampling tool *per se* [295], the energy minimization (EM; also known as geometry optimization) methodology aims to find a set of coordinates representing the minimum energy conformation for the given structure [296]. Given an initial structure, EM algorithms iteratively calculate the potential energy of the atomic arrangement and progressively move the atoms towards positions where the net interatomic forces are closer to zero. The produced intermediate snapshots do not have a physical meaning nor a time-dependent relationship. The identification of a low-energy conformer is a valuable means to study biomolecules since bioactive conformers often correspond to (near) energy-minimum states. EM is also widely used to correct possible structural anomalies in a molecular system in order to prepare it for subsequent simulations, such as MD or MC, or to refine models coming from structure prediction methods. EM algorithms can be classified into two categories: those that rely on the mathematical calculation of derivatives (*e.g.*, the steepest descent, conjugate gradient or the Newton-Raphson methods), and those that do not (*e.g.*, the simplex or the sequential univariate search methods). In the MM regime, EM methods commonly work with Cartesian coordinates, while in the QM regime internal coordinates are more frequently used. In practice, the exact true energy minimum is not reachable, therefore an approximation is needed to terminate the calculation when a reasonably low-energy model is obtained. A possible strategy is to end the simulation when the difference in energy between iterations gets smaller than a certain threshold. Usually, the technique advances from the starting point towards the closest local minimum of potential energy, without crossing energy barriers, however some EM methods can also visit adjacent basins of the energy surface in search of a more global energy minimum [295], [296].

In addition to all the strategies summarized above, which concern the modeling of the macromolecule of interest, biomolecular simulations also need reliable approaches to represent the chemical environment, *i.e.*, the solvent. Indeed, water is the medium where all biological activity takes place. Water molecules constantly interact with proteins: i) they facilitate the folding process by reinforcing hydrophobic interactions, ii) they form a network of transient hydrogen bonds with the solvent-exposed amino acids, iii) they provide an electrostatic screening effect, and iv) they can be involved in chemical reaction mechanisms. Furthermore, the aqueous medium in physiological conditions also contains dissolved ions that exert electrostatic interactions [297]. Therefore, solvent effects are generally considered in biomolecular simulations either by explicit or implicit solvent models. Explicit-solvent methods incorporate atomistic-level or coarse-grained solvent molecules in the simulation, thus including the calculations of the corresponding interatomic interactions. Through the years, hundreds of different potential models have been developed to treat water molecules with different levels of accuracy [298]. In contrast, implicit-solvent methods omit the solvent degrees of freedom by approximating the discrete solvent as a simulated continuous medium. Implicit models are considerably cheaper in terms of computational expense since the number of particles is reduced by several factors, however the specific solvent-solute interactions are lost [299]. While in classical MD studies the solvent is almost always represented with explicit models, in other conformational sampling methods the situation varies depending on the scientific scenario. For instance, in NMA and MC it is common to see implicit models to avoid the possible damping of large-scale motions. In the particular case of MC, explicit methods complicate the simulations because the chances of rejection of the tried conformations increase due to atomic overlaps being more likely to occur [291], [294].

1.4.3 Molecular dynamics simulations

Molecular dynamics (MD) simulations are commonly referred to as a veritable “computational microscope”, capable of valuably complementing many experimental methodologies and facilitating discovery in spatial and temporal domains that would otherwise be inaccessible [250], [285], [300]. With MD we can simulate the motion of a given atomic system and store such information in a collection of snapshots, called a trajectory. For the sake of clarity, we must take into account that “MD” is actually the name of a family of methods. As seen in the section above, the algorithms behind MD techniques can be based on either QM or MM approaches. However, the term “MD” is generally employed to refer to all-atom simulations powered by classical force fields, at least when no further specifications are given. In this thesis I also adopt this convention.

1.4.3.1 Applications and breakthroughs in structural biology

The first implementation of MD simulations in the fields of biochemistry and biophysics dates back to the late 1970s when McCammon and colleagues produced a 9.2 ps trajectory of the bovine pancreatic trypsin inhibitor protein in vacuum [301]. That simulation opened the door to a whole new era of structural biology with dynamics in the spotlight. In the words of the authors, “the results were instrumental in replacing our view of proteins as relatively rigid structures with the realization that they were dynamic systems, whose internal motions play a functional role” [302]. Since then, the popularity of this method has done nothing but grow [303], [304] (Figure 1.14). Nowadays, it remains the prevalent choice when it comes to deciphering functional traits of biomacromolecules like proteins or nucleic acids in terms of their structural conformations and dynamical behavior [5], [305], [306].

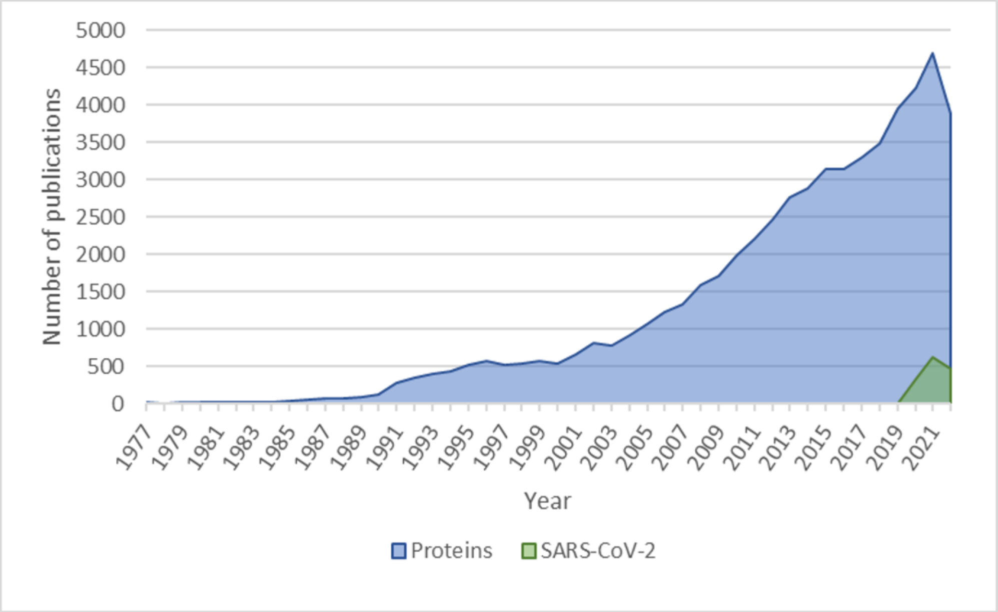


Figure 1.14. Trend in MD simulations of proteins. The popularity of MD for the study of proteins has followed an exponential growth, as evidenced by the number of publications per year that are related to the field. Recently, with the pandemic outbreak of the COVID-19 disease, a considerable piece of such studies has been devoted to aid in the design of a cure by deciphering the structural basis of the SARS-CoV-2 infection. The data was extracted from the Clarivate Analytics' Web of Science database, at the time of writing this thesis, for the time span between 1977 and 2022. The search was refined to include only research and review articles, for the topics "molecular dynamics" and "proteins" together (blue) and later incorporating the topic "SARS-CoV-2" (in green).

After decades of fruitful research, MD simulations have proven to be a quintessential tool for the fields of biochemistry, biomedicine and pharmacology. MD not only allows the simulation of native biomolecular interactions and motions in full atomic detail and with high temporal resolution, but also the introduction of carefully controlled perturbations (*e.g.*, ligands, external forces…) to predict their effect on structure and dynamics [108]. With this potential, MD has achieved remarkable milestones in describing biomacromolecular phenomena. Although it is impossible to provide a comprehensive list of such contributions, we can mention a few examples of interest that cover the study of biomolecular systems of diverse biochemical nature.

The process of protein folding has been probed with extensive MD simulations as in the cases of the villin protein headpiece subdomain [307] or the N-terminal domain of ribosomal protein L9 [308], where the timescales of reaching the native state agreed well with the experimentally observed folding rates.

MD simulations are especially useful to identify flexibility patterns and monitor conformational changes associated with biological function. The short peptide linker between the SH2 and SH3 domains of tyrosine kinases of the Src protein family exhibits a flexible behavior that governs the activation of such proteins [309]. Targeted MD simulations were able to determine some of the progressive intermediate conformations that populate the transition pathway between the open and closed conformations of the GroEL chaperonin protein, estimated to occur in the millisecond timescale [310]. More recently, such conformations have been further sampled by impressively long unbiased simulations, and even enabling the direct study of the folding of the small protein HP35 inside the GroEL cavity [311].

Large macromolecular complexes have been available for study via astonishing MD simulations achieved at the most cutting-edge supercomputing facilities [312]. Bock *et al*. [313] monitored the intermediate states and the underlying molecular forces in simulations of time-resolved translocation events in complete ribosome structures. Chandler *et al*. [314] simulated a membrane patch of 23 million atoms from a bacterial photosynthetic system, which provided insight into the packing of the proteins and the diffusion of the quinone molecules through the narrow lipid phase between the proteins of the complex.

Water and ion transport across the cell membrane has been measured in simulation for proteins such as aquaporins and potassium channels [315]. Similarly, the intermittent opening and closing of the long channel from the surface of acetylcholinesterase to its buried active site has been observed in simulation, thus providing a correspondence between the expected rate of diffusion of acetylcholine through the channel and the kinetic measurements of the enzymatic reaction [316].

When it comes to allostery, in section 1.1.2.3 we already have had the chance to highlight the capital importance of MD tools in the study of the allosteric mechanism. The allosteric nature of a biomolecular system can be confirmed by detecting the structural mechanisms that match what kinetic and thermodynamic data suggest. In the review by Hertig *et al*. [108], the authors list some examples of successful case studies of allosteric phenomena. In one of those, MD simulations were able to reveal the mechanism by which the binding of a G protein to a receptor protein (GPCR) causes a subtle conformational change in the G protein that accelerates the release of the ligand GDP molecule from it. This allosteric mechanism was subsequently validated by experimental procedures [317].

Nucleic acids have also been well studied with MD simulations to determine their functional roles in many biological processes and diseases. For instance, the destabilization of the structure of the rigid tetrad core of G-quadruplexes of DNA in the absence of coordinated ions was described more than twenty years ago [318], and later the possible folding intermediates of the structure were proposed [319]. Protein-DNA interactions have also been elucidated, such as the binding modes of the tetrameric formation of protein p53 when recognizing and accommodating DNA [320].

Free ligand binding simulations can provide mechanistic information about the process of binding to a macromolecule. Although they are not common due to their generally high computational cost, we can find some examples in the literature such as the study by Buch *et al*. [321] where 187 out of 495 MD simulations managed to capture the binding event of benzamidine to the protein trypsin, allowing to quantitatively reconstruct the complete binding process. Similarly, Sneha *et al*. [184] list several examples where MD simulations can aid in drug discovery and design and personalized medicine. MD simulations can detect and explore the minor changes in the interactions and binding patterns of drugs associated with protein variants. For instance, a local conformational change in variant His274Tyr of neuraminidase from the influenza A virus was revealed to be responsible for conferring resistance to the inhibitor drug oseltamivir [322].

Last but not least, the relevance of MD as a valuable asset in the science of today has been evidenced in the global community's response to the COVID-19 disease. The pandemic outbreak in 2020 shook all social strata, and countless research groups around the globe embarked on the effort of aiding in the design of a cure by deciphering the structural basis of the SARS-CoV-2 infection. As soon as structural data of the viral proteome became available, MD simulations, in conjunction with other *in* *silico* techniques, began to be employed at an unprecedented pace to study its functional dynamics [323] (Figure 1.14). In this joint response, the need to work cooperatively was highlighted and shared among the community [324], eventually resulting in the release of several databases of SARS-CoV-2 MD trajectories, such as BioExcel-CV19 (https://bioexcel-cv19.bsc.es/) and others [323], [325].

1.4.3.2 The role of force fields

The principles of the MD technique rely on a simple algorithm. Given a starting structure, Newton's classical equations of motion are calculated for all particles of the system to provide the evolution of their positions and velocities, iteratively, at successive time points [285]. The force fields that provide the molecular mechanics potentials for representing the physical properties of atom types and their chemical connectivity must be accurate enough to provide meaningful simulations. For that purpose, force-field parameters generally must be derived from reliable sources such as *ab initio* QM data and/or experimental data (NMR, structural databases) [282], [284].

Force fields are inherently imperfect because they contain approximations to reduce the computational cost and complexity of the simulations. When obtaining parameters from samples of biomolecular systems, the representation of the broad diversity of biochemical environments is inevitably limited. In the case of generic protein force fields, for instance, the simulation of IDPs has been revealed to be imprecise because they are underrepresented in the samples of (folded) proteins employed to build parameters [284], [326]. A continuous exchange between experimental biophysical techniques and MD is needed, and force fields should be validated against proteins with diverse folds. Indeed, there exists a certain risk to obtain misleading models from MD simulations as a consequence of using a set of parameters that does not describe the system of study with enough accuracy [283].

The most popular families of protein force fields in academia (AMBER [327], CHARMM [328], OPLS [329]) have been subjected to continuous refinements over the years. Comprehensive studies that compare simulation results with experimental data [282], [330] have shown that the current state-of-the-art force fields perform satisfactorily well when describing the majority of structural and dynamical properties of proteins. The different available force fields share similar mathematical functional forms, and they mainly differ in how they have refined a subset of relevant parameters for torsional angles. These are key components of force fields that serve as a bridge between bonded and non-bonded interactions of both backbone and side-chain atoms [282]–[284].

Despite successful advances, there is still margin for improvement to achieve more optimal force fields that can truly be transferable to proteins of diverse constitution. Any upcoming refinements will target the most prominent deficiencies in accuracy that have been found. The current force fields still overestimate the persistence of salt bridges and fail to accurately describe interactions that need polarizable effects to be explicitly taken into account, such as in metal-binding proteins [283]. As pointed out above, proteins in unfolded states or IDPs cannot be simulated with confidence, mainly because of the unbalance in the sampling of secondary structural elements and the inherent propensity to generate over-compacted conformations [326]. The solvation free energies of many amino-acid analogues in water are more unfavorable than experimental values, suggesting that most current protein force fields are generally too hydrophobic and also that the relative strengths of protein–water *versus* protein–protein interactions are incorrect [284]. Finally, protein folding simulations still present deficiencies in describing folding equilibria and their dependence on temperature, as well as in correctly identifying protein folding pathways/intermediates [284]. The use of machine-learning tools for deriving new potentials is a promising solution that is currently being explored, and we may soon see a growing tendency in resorting to these kinds of approaches [331].

1.4.3.3 Technical improvements in hardware

Besides the efforts invested on refining the force fields, the rest of improvements performed on classical MD as such are essentially technical. Remarkably, the emergence of modern computational architectures that are capable of running MD algorithms more efficiently has substantially boosted the performance of MD simulations. A couple of decades ago, the achievable simulation time would be of 10 ns or less, comprising hundreds or a few thousands of atoms [302]. Nowadays, in contrast, considerably larger simulations are affordable. An average MD simulation of our era easily comprises a million atoms and reaches multi-nanosecond to microsecond timescales [5], [304]. Should this trend continue, the relevance and reliability of MD simulations in both academia and industry will even keep increasing.

Indeed, computational cost is one of the main technical restraints of MD. It has been widely stated that for an MD trajectory to describe functionally relevant dynamics events it must span a sufficient amount of simulation time, comparable to the scale at which biological phenomena occur [5]. Although there is simply no threshold that ensures sufficient sampling, and thus this matter is subject to expert judgment, the appropriate time window is certainly beyond the 100 ns for an average-sized protein [302], [332]–[335]. Being able to perform such simulations routinely is computationally expensive, especially taking into account that this kind of studies often comprise not just a single simulation but an ensemble of trajectories of different system conditions or replicates. Storage capacity is also often a constraint; large disk space is required to store the produced trajectories (normally several GB per trajectory).

Fortunately, simulation engines have evolved in several ways to overcome the setbacks of computational cost. First and foremost, it has been possible for several years now to run MD computations distributed among hundreds or thousands of processors, running in parallel at a supercomputer in an efficient way. This fact has been crucial to enhance our capability of studying bigger, more realistic systems [5]. Besides, high-performance computing (HPC) facilities are growing and are increasingly accessible. With the continuous increases in computational power, we can expect that performance will be boosted even further by any forthcoming computing architectures [304]. Secondly, algorithms have been reorganized and optimized to run efficiently in graphical processing units (GPUs). Thus, even when having access to a supercomputer is not an option, the computation of MD trajectories in a local machine with GPUs is a moderately cheap alternative [305].

Over the last years, we have been witnessing a handful of all-atom MD simulations of colossal biomolecular systems [312]. Of course, these are often achieved only at state-of-the-art HPC facilities. For instance, two simulations of 100 ns of the whole capsid of the virus HIV-1 (64 million atoms) were performed on the supercomputer Blue Waters (National Center for Supercomputing Applications), using 4,000 Cray XE6 nodes [336]. Simulations of 40 ns and 150 ns of membrane patches of 20 million and 23 million atoms from the bacterium *Rhodospirillum photometricum*, respectively, were obtained by using the computational resources from the following facilities: Jaguar and Titan (Oak Ridge National Laboratory), Tsubame (Tokyo Institute of Technology), and Blue Waters [314]. More recently, an outstanding simulation of 500 ns of a chromatophore of 100 million atoms (136 million atoms with explicit solvent) was produced at the supercomputer Titan using up to 8,192 of its computation nodes [337].

As a side note, we can point out two initiatives that perhaps are the highest expressions of the phenomenon of advanced hardware. The first is the existence of HPC infrastructure specifically designed to run MD simulations efficiently. Only two instances of such initiative exist at the moment: the supercomputers Anton [338] and MDGRAPE [339]. Anton is famous for having produced the first protein simulations that reached the millisecond timescale [340], together with several other high-impact simulations. Currently, Anton is in its third incarnation which is 100-fold faster than any other supercomputer and is estimated to achieve the simulation of a million atoms at over 100 microseconds per day [341]. The second initiative is the Folding@home project (https://foldingathome.org/), which allows volunteer users around the world to lend their own personal computers as a way of sharing a distributed computing system. This means that MD simulations can be separated into work pieces, distributed across the network of volunteered machines, run, and finally returned to the project's database servers where they are compiled into an overall simulation. The philosophy behind this initiative is that supercomputers are often too busy to let an MD simulation scale to all their processors, whereas there are hundreds of millions of personal computer processors of comparable speeds setting idle around the world that can host such calculations in an orchestrated manner. In March 2020, as a result of the joint COVID-19 research project, Folding@home became the first computing system to break the exaFLOPS (10^{18} floating point operations per second) barrier, a prestigious feat of computer performance that was expected to be first achieved by non-distributed computing systems (*i.e.*, the “regular” supercomputers).

1.4.3.4 Technical improvements in software

Hand in hand with hardware, biomolecular software usability has also improved. For several years, there has existed fair criticism about the impractical use of MD simulations for non-experts. Indeed, the process of setting up a system for simulation often requires overcoming a steep learning curve. Depending on the complexity of the experiment, the user can face a plethora of technical decisions and parameter adjustments. A poor choice of parameters not only may compromise the quality of the study, but also lead to the degradations of computing performance. These handicaps are not caused, *a priori*, by a lack of sufficient simulation software, but rather by the intricacies of the configuration and the control of the simulation conditions. In fact, there is a considerable amount of different software packages, each with different features and merits, that are available freely or commercially. For a comprehensive list, see [342].

Embarking on MD simulations is nowadays more possible than ever because simulation software has adapted conveniently to reach non-expert users [304], [305]. The implementation of Graphical User Interfaces (GUIs) that work on top of traditional software is substantially helping by bringing more user-friendly environments that enhance interactivity, clarity, and reproducibility. Such is the case of interfaces like CHARMM-GUI [343], MDWeb [344], or QwikMD [345]. Continuing with this endeavor, modern software tools facilitate creating pipelines and following state-of-the-art MD protocols straightforwardly. The programmable environment HTMD [346] enables preparation, simulation, visualization and analysis of molecular systems in scripts written in Python programming language. The BioBB software library [274] allows building workflows by assembling a collection of prepared modules of common biomolecular simulation tools. More recently, the tool has been incorporated in a web-based GUI, BioBB-Wfs [347], that offers complex pipelines like those used in medicinal chemistry, biophysics, or computational biology pipelines, including automatic small molecule parameterization, protein-ligand docking or protein MD analyses. Other tools like MDBenchmark enable individual users to run benchmarks and find the optimal run parameters and settings for any simulation and hardware stack [303]. Last but not least, neither should we disregard the efforts of software developers to keep providing documentation and tutorials [348] or the distribution of comprehensive protocols by independent authors or research groups [344], [349]–[351].

1.4.3.5 FAIR principles and MD data sharing

Together with the exponential growth of MD studies, lately we are witnessing a growing concern to implement bioinformatics frameworks that can enable efficient storage and exchange of trajectory data [304], [352], [353]. Indeed, the data derived from MD simulations is potentially reusable, since trajectories often contain a vast amount of information from which typically only a subset is analyzed in the context of a single study. The massive quantities of expended computational resources and accumulated simulation time should somehow be further exploitable by the community to accelerate research.

Historically, the implementation of databases of MD trajectories has been hindered by several challenges. First and foremost, infrastructures should cope with the immense demand of disk space. Then, the deposited files should comply with pre-established types, formats, documentation… Best-practice guidelines on how to share MD simulations are still being defined [352], [354]. Finally, once deployed, resources must ensure a proper maintenance plan regarding economical sustainability, user support, and curation of the uploaded data. General-purpose data repositories like Zenodo (https://zenodo.org), FigShare (https://figshare.com), and Open Science Framework (https://osf.io) provide some advantages (global public access, wide file format acceptance…) and are being used. However, they have limited storage quotas and cannot help standardize and harmonize the rules of MD data sharing. A modest number of initiatives have developed special-purpose MD sharing platforms, such as BioSimGrid [355] (halted in 2009), MoDEL [356] and MoDEL_CNS (https://mmb.irbbarcelona.org/MoDEL-CNS/), BIGNASim [357], TMB-iBIOMES [358], GPCRmd [359], BioExcel-CV19 (https://bioexcel-cv19.bsc.es/), and SCoV2-MD [325].

This subject also raises questions about whether we as a scientific community are using all our potential to follow the best practices of research and data sharing, which are very well synthesized in the FAIR principles: Findable, Accessible, Interoperable, and Reusable [360]. Indeed, pursuing an effective MD data sharing model does not only aim to preserve trajectory data, but also to improve reproducibility and scientific transparency. Repositories can become powerful platforms of collaborative research, for instance, by providing tools for the interactive visualization and analysis of trajectories. In turn, these implementations are potentially beneficial for improving the quality of research dissemination. Usually, a publication of an MD study describes its results with the usual means of a manuscript (text, tables, plots, and figures) and occasionally encloses supplementary videos. However, the clarity of such explanations can be substantially enhanced if the described dynamical events can be followed directly with a web-based visualization tool [353]. Thus, journal articles (online editions) and web portals dedicated to scientific outreach should start incorporating standardized means of showing trajectories. The number of graphical interfaces optimized for trajectory display is conveniently growing [304], [325], [353]. Tools like Molywood [361] and 3dRS [362] aim to facilitate the design, generation, and sharing of videos and interactive biomolecular representations. These initiatives can encourage researchers to provide visual material to highlight structural and dynamical information.

1.4.3.6 Sampling quality

One of the main disadvantages of MD is the limited extent of conformational space that can be sampled at the accessible simulation timescales. Indeed, most important biomolecular processes often involve dynamical events that take place in slow timescales, possibly of dozens of microseconds or beyond the millisecond [305]. This is the case for diffusive ligand binding events, folding-unfolding processes, or even the conformational change cycles that comprise the mechanical part of biomolecular function (*e.g.*, allosteric transitions). Even though the current computational force enables the generation of long simulations that can span such timescales, this is usually not enough for ensuring more extensive conformational sampling.

The complex shape of the FEL makes most of the simulations explore just a small region around the energy minimum closest to the initial conformation. Due to the stochastic nature of MD, there is no *a priori* way of telling whether a (sufficiently long) simulation will be walking over energy barriers and thus exploring different basins of the conformational space, or rather stay trapped in a single basin most of the time (an effect known as conformational trapping) [17], [283], [363], [364]. For this reason, several simulation replicates are generally required to observe dynamical events at their entire spans. Moreover, even when a particular simulation manages to capture such an event by chance, a single observation is not statistically robust. The same event must be observed multiple times to be able to provide descriptions of its dynamical mechanism and the relative conformational populations with confidence [109], [335], [365]. From this well-known issue, an interesting discussion can be brought up concerning how we can handle trajectory ensembles. This subject is further introduced in section 1.4.5 and takes a relevant place throughout the thesis.

With the aim of overcoming the sampling problem, a wide variety of enhanced sampling methods have been developed over the years. These techniques employ diverse strategies that introduce controlled biases to MD simulations to favor a given transition or to capture longer-timescale events [5], [305]. Some examples are listed here below. In replica exchange methods [366], parallel MD simulations are launched in different conditions (typically temperature) and their structures are periodically interchanged according to a Metropolis acceptance algorithm. The resulting simulation benefits from the fact that the sampling ability increases with temperature, and thus usually achieves a larger overall sampling of the conformational space. Weighted ensemble methods [367] look out for rare sampling events of the conformational space and launch more simulations from those points, generating a branched ensemble of trajectories in the process. In metadynamics [368], a biasing term penalizes the system to re-visit regions previously sampled, therefore helping the simulation escape the initial local minimum. Umbrella sampling techniques employ a biasing potential to force the system to move in small steps along a transition from a predefined pair of initial and final states [5]. Targeted MD [369] implements a variant of umbrella sampling that drives the transition by monitoring and slowly reducing a metric of structure dissimilarity. Similarly, steered MD [370] pulls the system towards the final state via a steering force. Temperature-accelerated MD [371] was inspired by metadynamics and enhances the sampling by using an artificially high temperature associated with certain degrees of freedom [283]. In accelerated MD [372], the height of energy barriers is lowered by adding a bias potential to the true potential. Finally, NMR parameters have been also used to restrain MD simulations [283]. The application of enhanced sampling methods is subject to the specificity of the information that one wants to obtain; for instance, when certain conformations of the biomolecular system are known beforehand. However, for a study that aims to explore the native dynamical behavior without prior knowledge, regular MD is probably best suited [349].

1.4.3.7 Trajectory analysis

MD simulations generate an overwhelming amount of data. A trajectory file contains the atomic coordinates and velocity values that were sequentially calculated during the simulation. The typical MD run employs an integration time step of 1-2 femtoseconds [351], but usually the snapshots of the trajectory are stored with less frequency; probably three orders of magnitude lower, *i.e.*, every few picoseconds [348]. Otherwise, a single trajectory that spans a few dozens of nanoseconds would account for unreasonable memory usage, especially considering that dynamical events faster than the picosecond mainly correspond to atomic vibrations that are negligible for studying biomolecular function [4]. Aside from atomic coordinates and velocities, energy values are often also saved at the same time intervals [348].

Given the format of trajectory data, how can we treat it to describe the captured dynamical events? The most straightforward approach consists in visualizing the 3D model of the system along the simulation. Many of the gold-standard molecular rendering software applications, such as VMD [12] and PyMOL [373], can read trajectory files and display them as animations. The simplest 3D representation consists in drawing atoms as spheres that are connected by sticks to indicate covalent bonds. There are other more sophisticated representations that for instance draw ribbons or cylinders to highlight protein secondary structure segments. By visually inspecting trajectory animations one can look for any relevant motions occurring at the molecular sites of interest.

While this (common) practice can provide qualitative insight of the evolution and flexibility of a particular system, it does not suffice to derive definitive conclusions. We should be able to validate the nature of any spotted dynamical event in the light of questions such as the following. How often does it occur? Does it always occur with a similar duration and/or motion amplitude? Is it conditioned (preceded, accompanied or succeeded) by another event? Could it be just an artifact or a misleading perception? Moreover, relying on visual inspection gets impossible as the amount of data increases. It would take too long and be dangerously imprecise to thoroughly inspect trajectories of large systems and/or long simulations; let alone entire ensembles of trajectories. All in all, even though visual inspection is a crucial initial component of trajectory analysis [24], researchers must face the challenge to find suitable metrics to extract, quantify and present the relevant information depending on the target of the study.

Various metrics have been extensively applied for such purposes [184], [239], [248], [342], [374]. By means of basic geometric analyses we can inspect the evolution of the system along the trajectory by focusing on structural traits. Some of the most widely used metrics within this category comprise the measurement of the root-mean-square deviation (RMSD), the root-mean-square fluctuation (RMSF), interatomic distances and angles (for instance, to describe hydrogen bonds), the radius of gyration, the solvent-accessible surface area (SASA), or the secondary structure content. On the other hand, energy analyses comprise the calculation of binding free energies, for example with the molecular mechanics Poisson-Boltzmann or generalized Born combined with surface areas (MM/PBSA and MM/GBSA), or conformational free energies by constructing a FEL in terms of a given set of metrics that serve as reaction coordinates [239], [375].

Other specialized analysis techniques further characterize the dynamical behavior of the system by fitting the trajectory data into a mathematical or statistical model. Among them, dimensionality reduction methods aim to find collective variables that can capture underlying key features. Many approaches of dimensionality reduction have been proposed over the years to work with trajectory data, including recent implementations with machine learning methods [376], [377]. However, the foundational and still most prominent technique of this family of methods is the Principal Component Analysis (PCA) [377]–[380]. PCA constitutes the analytical framework in which this thesis develops new propositions for extracting meaningful features and comparing trajectory ensembles. Finally, other popular trajectory analysis strategies that are relevant to mention are those derived from network analysis [283], [381]–[383] and Markov state models [384], [385].

1.4.4 Dimensionality reduction of trajectory data

The structure of a biological macromolecule, such as a protein, possesses many degrees of freedom that allow it to adopt a myriad of different conformations. Accordingly, the variability contained in a given conformational ensemble needs to be described, *a priori*, by a set of multidimensional variables, such as atomic Cartesian coordinates, interatomic distances, or dihedral angles. Thus, in the case of an MD trajectory, dynamical information is obtained as a time series of high-dimensional data. Such amount of information, even if valuable, is not easily interpretable. Therefore, extracting key features from raw trajectory data is crucial to understand the dynamics of biological macromolecules. For this purpose, dimensionality reduction approaches began to be proposed only a few years after the first MD studies with proteins and have since been widely applied for over 30 years to elucidate macromolecular dynamics [386].

Dimensionality reduction methods operate under the following assumption: the high-dimensional space of the input data structure contains a lower-dimensional subspace that preserves most of the original variability of the data. In other words, the aim is to identify a set of collective variables (CVs) that effectively reduce the dimensionality where the problem is formulated [376], [377]. The procedure is rooted in the possibility that large data samples in the high-dimensional space hold redundancies and correlations. If this is the case, the interpretability of data generally improves. However, such simplification often always comes with an inevitable degree of information loss, therefore an optimal tradeoff must be found for it to be favorable [377].

In protein dynamics, extensive studies have proven that the significant, functional motions are expected to take place precisely in a low-dimensional subspace. Dimensionality reduction approaches have been exploited with broad success in protein functional and folding studies [378]. CVs capture the underlying collective atomic motions that can potentially provide mechanistic insight into biological function such as enzymatic catalysis or signal transduction [377]. On the other hand, the detection of characteristic features of the conformational ensemble via the CVs enables the construction of FELs. Accordingly, the structures can be classified into those that represent metastable states (the maxima of the conformational distribution) and the connection paths between them (yielding the energy barriers between the states) [375], [376], [378].

1.4.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate statistical technique that has been largely applied to analyze multidimensional datasets in many kinds of scientific and technical fields [378], [379].

Nowadays it is also often classified as an unsupervised machine learning technique [386]. In the earliest stages of MD studies with proteins, several researchers implemented PCA-based methods to analyze trajectories [387]–[394]. This strategy has been referred to as "quasi-harmonic analysis" [387], "molecule optimal dynamic coordinate analysis" [393], or "essential dynamics analysis" (EDA) [391], the latter being the common term used nowadays [378], [379] and thus also adopted in this thesis.

PCA, as the central technique in EDA, aims to identify a few CVs that correspond to the "essential" degrees of freedom that can describe the functionally relevant motions [391], [392], [395], [396]. Technically speaking, the standard PCA procedure applied with trajectory data consists in performing an eigendecomposition of either the covariance or the correlation matrix of the Cartesian coordinates of the atoms involved in the analysis [379] (see section 3.5.1 from the Methods chapter for further details). This mathematical operation enables expression of the original trajectory data points in terms of a new basis set of orthonormal vectors that are not linearly correlated and indicate the directions of the dataset that capture the maximum amount of variance. The projection of the data onto such new coordinates, called the principal components (PCs), enables inspection of the underlying key collective features of the dataset. In the particular context of trajectory analysis, PCs describe collective atomic displacements [334], [379], [386], [392], [395], [397]. Consequently, PCs allow for the exploration of the collective motions that best describe the conformational variability that was sampled during the simulation. Roughly speaking, PCA normally achieves a tremendous reduction of dimensions in the average protein trajectory, being 20 PCs usually more than enough to define an "essential subspace" that captures the motions governing biological function [379].

PCA is a useful asset both for the study of conformational mechanics and dynamics of biomacromolecules and the construction of folding FELs; many examples of successful studies can be found in the literature [378]. These have revealed and confirmed functional features of countless proteins (regulatory proteins, enzymes, channels…) not only by applying PCA to a single system, but also to different conditions of the same system for their comparison (for instance, ligand-bound and unbound states, or wild-type and mutated forms). PCA benefits from great simplicity since it can be applied in a straightforward manner without any *a priori* information or phenomenological parameter. However, a deeper understanding of the technique will allow one to exploit its full potential. For instance, a plain application of PCA to the whole system in the trajectory may not be sensitive enough for detecting the more localized, small-amplitude but functionally important motions; however, the resolution can be increased by applying it to a subregion of interest or excluding noisy coordinates that may be overshadowing them [376], [379], [386], [398]. Additionally, PCA has been used with trajectory data for other purposes, for instance, to feed enhanced sampling techniques [378], [399], in trajectory compression [400], small molecule docking [401] and determining the redox potential of proteins [402].

A frequently used variation of standard PCA relies on the use of internal coordinates instead of atomic Cartesian coordinates to perform the analysis. Internal coordinates such as intramolecular distances and angles are not affected by the overall motion of the macromolecule and thus manage to escape the possible interferences of such degrees of freedom. However, the direct correspondence of the data with definite conformations is lost, thus entailing a decrease in the interpretability of the collective features [375], [376], [379], [386], [403]. Some approaches of PCA that employ interatomic contacts and distances as input data are: internal distance pair coordinates (dpPCA) [404], contact distance-based (conPCA), reciprocal distance-based (iconPCA), and inter-C_α distances (C_αPCA) [405].

Since the number of pairwise atomic distances scales quadratically with the number of considered atoms, this makes the analysis increasingly expensive. Thus, dihedral angles are a most common alternative. Backbone dihedral angles are valuable descriptors of the secondary structure, whereas side-chain dihedral angles report inter-residue contacts [376]. Dihedral angle PCA (dPCA) uses backbone dihedral angles and converts them to sines and cosines to obtain a linear coordinate space with the usual Euclidean distance [403], [406]. An improved version called dPCA+ performs the analysis directly on the dihedral angles by transforming the data such that the maximal gap of the sampling is shifted to the periodic boundary of a dihedral angle [376]. In GeoPCA, dihedral angular data are projected on a sphere composed of the first two principal component geodesics [407]. In Torus-PCA (T-PCA), dihedral dynamics are characterized by taking advantage of the fact that the *n*-dimensional torus is a product space of *n* circles [408].

Since Cartesian PCA applies a linear transformation from the original variables to the new CVs, it detects only linear correlations, while trajectory data can also include nonlinear correlations [386], [409], [410]. Nonlinear correlated behavior in protein dynamics is more prominent on folding and unfolding processes, given the divergence of the sampled conformations [411], but other simpler phenomena can display a nonlinear correlation such as the collective rotations of two methyl groups [378]. The PCA approaches with dihedral angular data can already give insight into the possibly significant nonlinear correlated motions of the protein atoms [378], [386], [406]. For instance, the application of dPCA was able to describe more accurately the rugged nature of the FEL of protein folding [375]. In order to account for nonlinear correlations, the Kernel PCA approaches [412] are often a choice. Kernel PCA (kPCA) can be thought of as a generalization of PCA, where firstly a nonlinear transformation (the so-called kernel function) maps the input coordinates to a feature space of higher dimensionality with the aim that the transformed data becomes approximately linear. Then, a simple linear dimensionality reduction is performed in this space. Thus, the original nonlinear correlations are captured through the definition of the kernel itself [376], [379], [386], [412]. Kernels with sigmoidal, exponential and polynomial functions have been employed, although the most widely used is the Gaussian kernel. A linear kernel would make kPCA equivalent to Cartesian PCA [377], [380]. The choice of a proper kernel is not obvious because it is problem dependent and faces the impracticality of defining function parameters. Although kPCA may prove especially useful to study specific cases, the regular Cartesian PCA is a validated method to describe the dominant correlations present in atomic motions found in proteins [379].

1.4.4.2 Other dimensionality reduction approaches

Besides the PCA family of methods, a wide variety of other analytic techniques have been explored to achieve a meaningful dimensionality reduction of trajectory data. Two popular linear transformation approaches are multidimensional scaling (MDS) [413], [414] and time-lagged independent components analysis (tICA) [415], [416]. In MDS, the aim is to find a low-dimensional space that best preserves the pairwise distances between the original high-dimensional points by minimizing a cost function. MDS can be equivalent to PCA under certain conditions, however an advantage is that it can be used directly with a matrix of pairwise distances, thus allowing it to work in a more general way with non-Euclidean high-dimensional spaces and to address nonlinear correlations [377], [380]. The approach of tICA is similar to PCA but the produced CVs aim to maximize the autocorrelation time of the degrees of freedom rather than their variance. It accomplishes it by incorporating information on time dependency among the extracted eigenvectors (time-lagged covariance matrix). The produced

CVs are not orthogonal and show collective atomic motions that possess a rough estimate of their associated characteristic timescales, therefore are interesting to describe the slowest-relaxing modes of motion of the trajectory. tICA has become a popular tool for extracting kinetically relevant CVs and applying them for subsequent Markov state models analysis. However, the robustness of the technique is dependent on the ambiguous choice of a central parameter, the lag time [376], [377], [416].

A handful of approaches have been proposed for addressing nonlinear correlations. Nonlinear dimensionality reduction methods can indeed determine more intricate and elusive higher-order correlations and classify conformations from a set of trajectories, at low risk of loss of information on the protein dynamics [417]. Notwithstanding, they have characteristic disadvantages that may limit their applicability in practice: i) the reconstruction of data is difficult to interpret because the mapping loses the direct geometric interpretation, ii) the number of CVs has to be determined in advance, iii) the order of relevance of the obtained CVs is unknown or difficult to determine, and iv) some key parameters often have to be defined *a priori* [378], [379], [411].

Isometric feature mapping (Isomap) [418] generates a low-dimensional representation that best preserves the geodesic distance (the distance along a straight line in a curved manifold) by finding the shortest path through a network analysis performed on the high-dimensional space. The so-called scalable Isomap (ScIMAP) [419] is a variant of the Isomap algorithm specifically designed for big datasets such as the output of MD simulations. Stochastic Proximity Embedding (SPE) [249], [420] takes as input the structural similarity between all pairs of conformations, and uses an iterative method to obtain a low-dimensional projection in which pairwise distances are approximately preserved locally. Diffusion maps [421], [422] employ a Gaussian kernel (as in kPCA) to preserve the dynamical proximity between conformations visited in the high-dimensional space. Several specialized variants of diffusion maps have been proposed over the years to specifically improve diverse particularities of the technique [377]. Full correlation analysis (FCA) [423] minimizes the Shannon's mutual information metric to obtain maximally uncoupled collective coordinates. The Sketch-map algorithm [424] solves a highly nonconvex optimization problem to preserve the distances falling within a specific window which is assumed to characterize the important models of the system under study. Provided that a proper selection of parameters is made, Sketch-map may find a relevant low-dimensional representation of the data even when other simpler methods fail; however, this requires extensive trial-and-error operations with no guarantee of success [380]. The t-distributed stochastic neighbor embedding (t-SNE) [409] estimates, from the distances in the high-dimensional space, the probability of each point to be a neighbor of each other point. Such a procedure is dependent on a free parameter called "perplexity", which roughly represents the number of nearest neighbors whose probabilities are preserved by the projection. Uniform Manifold Approximation and Projection (UMAP) [410] is a fuzzy topology-based dimensionality reduction method that is similar to t-SNE but employs a different pairwise local distance metric between the points and a cross-entropy loss function.

Most of the current dimensionality reduction strategies that aim to account for high-order correlations apply deep learning approaches with artificial neural networks. In fact, it has been several years since the first applications of such technology with trajectory data, remarkably with the development of nonlinear Principal Component Analysis (NLPCA) [411] approaches, which perform nonlinear mapping with hierarchically arranged neural networks. Nevertheless, with the improvements of deep learning algorithms, these methods are currently gaining more relevance both in MD studies and in general in the majority of technical and scientific fields [377], [380], [425]. The increasing richness of data (up to the regime of big data) is prompting a shift in the methodologies employed in the field of molecular dynamics in favor of more automatized and less human-dependent procedures [377].

Of particular interest are the auto-associative neural networks or autoencoders (AEs). AEs rely on the sequential use of two neural networks, namely the encoder and the decoder. Firstly, the encoder maps the input features to a low-dimensional embedding called the latent space. Then, the decoder is capable of mapping back the compressed information to the original space. The process is iteratively optimized until the error between the original data and the reconstructed data points is minimized. Provided that the operation succeeds (*i.e.*, there is little loss of information), the CVs at the latent space serve as good descriptors of some underlying key collective features of the dataset. As it occurs with other nonlinear dimensionality reduction approaches, if a linear map was used in the AE, the procedure would be equivalent to PCA. Many trajectory analyses that employ AEs [110], [417], [426]–[428] and more specialized versions such as the variational autoencoders (VAEs) [425], [429]–[431] can be found in the recent literature.

1.4.5 The comparability of trajectory ensembles: a challenge

From the perspective of gathering rich conformational ensembles to achieve meaningful functional descriptions of biomolecules [5], [249], comparing biomolecular structures or states now becomes a question of comparing conformational distributions rather than individual conformations [432]. In general, conformational ensembles can be compared by spotting prominent differences in the scalar metrics computed for each individual ensemble, for instance, employing basic geometric analyses [248], [374], [398], [433], [434]. This comparative approach is straightforward and enables interpretation of the similarity both in a qualitative and quantitative manner, depending on the level of detail of the analysis.

In the case of MD trajectories, comparisons are less trivial because structural properties evolve as a function of time. Indeed, we are facing an extension of the challenges posed at section 1.4.3.7. In addition to the need of finding suitable “property spaces” to better interpret dynamical information, the next question is whether such representations and metrics are also powerful enough to provide a measure of similarity between different trajectories. If two trajectories yield similar time-dependent properties, are they similar? Or, if two trajectories exhibit very different time-dependent properties are they necessarily different? Kazmirski *et al*. [374] provided a conceptual example of the complexity of the problem by illustrating a hypothetical scenario of three trajectories that follow different pathways in their transition from an initial state to a final state (Figure 1.15). Conceptually, similar trajectories are those that span close regions of the multidimensional topological space (*i.e.*, they have sampled similar conformations). However, even when such regions overlap, the rate at which each trajectory has sampled the shared conformations may still differ. This entails time-dependent differences that one may want to take into account or not, depending on the purpose of the study. Similarity between two trajectories may be apparent or not, when examined from different angles. For this reason, in general, when analyzing multiple trajectories, it is important to perform a detailed comparison using multiple methods [374]. Nevertheless, in contrast to the amount of metrics that can be applied to compare individual structures, there has remained a lack of specific methods to compare conformational ensembles [398], [432], [435].

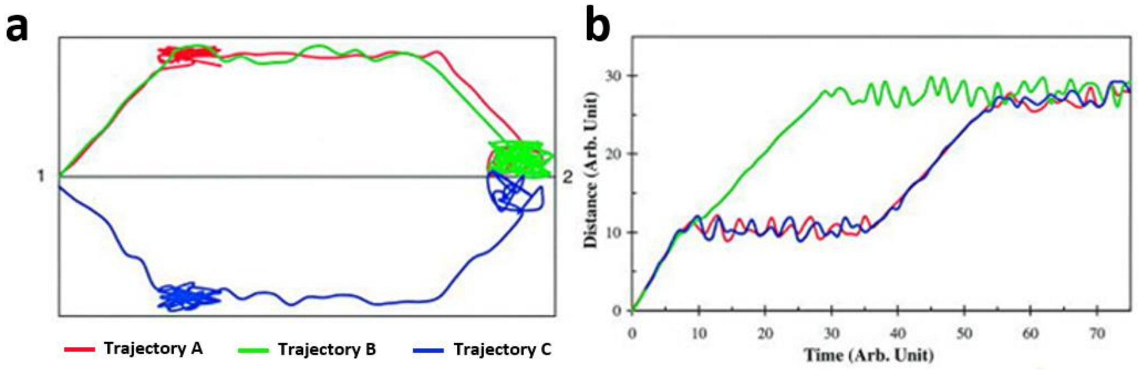


Figure 1.15. A conceptual example of the complexity of comparing trajectories. Three hypothetical trajectories of the dynamical process from a state 1 to a state 2 are represented. (**a**) The pathway of the three trajectories transitioning from state 1 to state 2. Trajectories A and B walk along the same pathway with A taking a longer amount of time to reach state 2 as it becomes trapped in a substate for a certain period of time. Trajectory C takes an alternate pathway, and like trajectory A becomes stuck in a substate along the way for a certain period of time. (**b**) Measuring the distance from the starting point against time for each of the three trajectories, it may seem that trajectory A and C follow the same path, while trajectory B walks a separate path until reaching state 2. Monitoring only this single property *versus* time leads to a wrong interpretation. NOTE. Adapted from [374].

A handful of studies have proposed a few comparative methods that aim to provide a (dis)similarity score between trajectories. A group of approaches are based on the idea that two ensembles can be compared by firstly estimating their underlying probability distributions and then quantifying their similarity using distance measures from probability and information theory. Quantitative comparisons can also be complemented by using the information derived from the correlation of motions between pairs of residues or atoms. One of the earliest proposals was the inter-ensemble RMSD (eRMSD) [433], which combines expressions of RMSD and the isotropically distributed covariance matrix of atomic positions. The harmonic ensemble similarity (HES) improved the eRMSD and gives weight to both differences in the mean conformation as well as differences in the fluctuations away from this mean. The clustering ensemble similarity (CES) applies a clustering algorithm followed by the calculation of the Jensen-Shannon divergence [249], [432]. Other studies have further proposed similar approaches [435]–[437]. The covariance overlap [438] as well as the dissimilarity index of Dynamical Cross-Correlation Matrices (DCCM) and Linear Mutual Information (LMI) [398] are examples of approaches that purely compare the information from correlation. Additionally, other methods implement comparisons on the residue-residue contacts [374] or use network theory metrics to perform the so-called difference contact network analysis (dCNA) [439].

On the other hand, the comparability between trajectories can be explored by dimensionality reduction methods. Among such strategies, EDA-based approaches constitute common ground to address this analytic problem. Indeed, in the same way as EDA provides insight into the underlying collective features of an individual trajectory, it can also reveal differential behavior between trajectories. With this purpose, a usual strategy involves performing PCA on each trajectory as an isolated experiment and then describing the resemblances and divergences of the corresponding essential spaces of each analyzed trajectory. The literature contains numerous instances of studies where trajectories are compared following this strategy: between replicates of the same MD conditions [293], [332], [334] and/or between different conditions such as the ligand-bound or unbound states of a protein [440]–[442] or mutant variants [235]–[239].

It should be noted, however, that such an approach practically only provides qualitative information. Each set of PCs originates from a separate space (*i.e.*, it has its own origin of coordinates). Therefore, a quantitative comparison of the PC values from different trajectories is meaningless. Instead, comparisons can be made by describing the differences in the observed structural traits manifested in the PCs, that is to say, by visually inspecting the captured collective motions in each case. Comparing the amount of conformational heterogeneity (variance) captured by each PC is also informative. For instance, depending on the divergence of the sampled conformations, a different number of PCs will be required to account for 50% of the variance (normally between 2 and 20) [379]. Additionally, the degree of overlap between the essential subspaces (*e.g.*, considering the first 10 or 20 PCs of each trajectory) can be quantified with the root-mean-square inner product (RMSIP). The RMSIP metric provides a general idea about the similarity between the regions of the conformational space sampled by individual trajectories. It ranges from 0 (*i.e.*, completely orthogonal spaces sampled by the essential subspaces of the two trajectories) to 1 (*i.e.*, a perfect overlap) [378], [391], [398]. Likewise, the similarity between pairwise PCs can also be assessed by calculating the corresponding inner product. A matrix of inner products can detect possible shifts in the order of relevance of similar PCs between trajectories [441].

As an alternative, the strategy known as combined-PCA [443] is an EDA-based approach that is more oriented towards trajectory comparison. Combined-PCA consists in concatenating the involved trajectories into a “multi-trajectory” (also called a “meta-trajectory”) and performing a single PCA on it. The main feature of combined-PCA is that it provides a single reference PC space for the whole ensemble of trajectories, enabling the extraction of the features of the average dynamical behavior of the ensemble. As such, the interpretability of combined-PCA has been discussed by several authors; depending on the degree of conformational space overlap of the different subsets of trajectories of the analyzed ensemble, each scenario prompts different considerations that should be taken into account [36], [109], [333], [365], [392], [443]–[446]. Mainly, this approach enables comparison by analyzing the properties of the different trajectories when projected onto the same PCs, namely the differences in the average and deviation values [36]. This strategy is usually implemented in studies that compare the trajectories of different biological conditions [214], [447], [448].

Several studies have explored other interesting implementations that allow envisioning new creative ways to exploit EDA-based approaches as a tool for effective trajectory comparison. For instance, Lindorff-Larsen *et al.* [249], [432] proposed the dimensionality reduction ensemble similarity (DRES) metric, which builds on the HES and CES methods proposed by the same authors, but employing a nonlinear dimensionality reduction approach to estimate the probability density of each conformational ensemble in terms of the generalized coordinates. Grosso *et al.* [440] implemented an unbiased method to quantify the similarity between different essential dynamics subspaces using the information provided by the angles between PCs. Their method was able to detect an increase in flexibility in the structure of glutaminase interacting protein (GIP) upon ligand binding that is otherwise hardly perceptible with standard EDA. In another study, Angarica *et al.* [237] developed a strategy for the detection of conformational instability based on combined-PCA and a subsequent clustering approach of the trajectory data projected onto the common PC subspace. They were able to obtain a classification of the severity of structural destabilization of the entire SNP mutational space

(227 missense variants) of the low-density lipoproteins receptor (LDL-r) LA5 domain. Finally, Ahmad *et al.* [449] proposed the Relative Principal Component Analysis (RPCA), a method that focuses on extracting CVs that are most informative of the changes between two trajectories that represent different sampled states of a system.

Chapter 2 Objectives

The main goals of the thesis are to explore the coupling between protein function and dynamical behavior, and to assess the performance of metrics derived from molecular dynamics (MD) simulations aimed at describing and comparing the nature of distinctive dynamical events. With these goals, specific objectives were established as follows:

1. To define a protocol for the analysis of differential dynamical behavior among MD trajectories of a protein system, grounded on the methodology of essential dynamics analysis (EDA) approaches.
2. To study the dynamical behavior of erythrocyte pyruvate kinase (PKR): i) by identifying and characterizing the distinctive motions involved in the allosteric events in the protein and its functional sites, and ii) by integrating the obtained insight with the previous knowledge of the structural and functional traits of this enzyme.
3. To detect characteristic dynamical alterations of a set of missense variants of PKR and assess their functional significance. Subsequently, to evaluate the capabilities of the proposed analytical strategy to discriminate between damaging and benign missense variants in proteins.

Chapter 3 Methods

The methodology employed in this study pivots around a central computational technique: classical molecular dynamics (MD). On the basis of the established goals (see the Objectives chapter), a comprehensive collection of MD simulations of the human erythrocyte pyruvate kinase protein (PKR) in multiple biological conditions was generated.

Such an effort entailed the need to assemble an advanced MD protocol, suitable for treating a large protein system with bound metals and small molecules. The protocol also sought to be reproducible in order to launch massive amounts of simulations. A substantial volume of work was accordingly destined to preparation procedures, including the setup of the initial structures and the development of specialized topology parameters. The subsequent analysis stage comprised both standard techniques and a strategy originally developed in this work that has been termed Consensus Essential Dynamics Analysis (CEDA). The approach of CEDA involves integrating the information from Principal Component Analysis (PCA) applied independently to each equivalent trajectory and deriving a consensus set of vectors that enable trajectory comparison in a common framework. The steps of CEDA are presented both separately and collectively as a proposed protocol for analyzing other similar biological systems. Additionally, the project also involved a stage of collecting a comprehensive dataset of pathogenic and potentially neutral missense variants of PKR.

This chapter describes the set of materials, computational techniques, software, and resources employed throughout the project for all the aforementioned purposes. It is worth noting that a fraction of the methodology followed in this thesis comprises techniques and practices that are nowadays strongly consolidated within its scientific field. Given the broadness of the background that underpins such methods, it has not been possible to include here a full report on all the theoretical and technical considerations. Instead, on the basis of the theory and the research framework established in the Introduction of the thesis, we now focus on the particular decisions that were conducted to achieve the goals of the study. Thus, the following sections aim to provide the primary practical details that facilitate following up each experiment, while the particular setups and implications of the employed procedure are further expanded and discussed at the Results and Discussion chapters.

Additionally, a set of relevant files involved in diverse procedures of this thesis has been uploaded to the online repository Zenodo (https://zenodo.org/) to facilitate reproducibility and further exploration of technical aspects of the methodology [450]. A complete detail of the contents can be found in Appendix A. Furthermore, the complete set of simulations and standard trajectory analyses have been made available at https://pklr.mddbr.eu.

3.1 Setup of structures

3.1.1 Protein structure (apoenzyme)

The initial model of PKR was obtained from the Protein Data Bank (PDB) entry 2VGB [121], which corresponds to an X-ray crystallographic structure of the protein, with a resolution of 2.73 Å, and co-crystallized in a complex with phosphoglycolate (PGA; a structural analog of the native substrate phospho*enol*pyruvate (PEP) and also a potent PKR inhibitor [126]), fructose 1,6-bisphosphate (FBP),

K^+, and Mn^{2+} (Mn^{2+} fulfills the role of the putative cofactor Mg^{2+} and equivalently occupies its binding site). The fact that the allosteric activator FBP is bound to the enzyme implies that it is found in the active conformation or R state [121]. All ligands were removed from the model in order to proceed with the setup of the protein (apoenzyme condition).

Due to the restraints imposed by the crystallization procedure [121], the 2VGB model exhibits a few structural gaps. Firstly, the first 56 N-terminal residues, as well as the last C-terminal residue, are absent from the model as it usually happens with flanking regions that are too flexible to achieve sufficient crystal organization. These gaps are equivalently found in each monomer of the structure. Additionally, a few other internal segments of the protein (located at domain B) could not either be resolved. These correspond to positions 167-174, 187-196, and 229-236 from subunit B, and position 167-171 from subunit D. In order to prepare the protein model for subsequent simulations, the internal gaps of subunits B and D were modeled after their respective symmetrical crystal coordinates from subunits A and C. This procedure was done manually with the tools of the software PyMOL (v2.3.2) [373].

When preparing the protein structure of PKR missense variants, mutated amino acids were modeled at each subunit of the protein with the *structure checking* utility from the BioBB library [274]. Afterwards, each site of amino-acid substitution was manually inspected to check for possible atomic clashes and overlaps after the replacement. The only variant that needed readjustment was Ser120Phe due to a structural overlap with the nearby side chain of Glu161. The side chain of Phe120 was subjected to the *Mutagenesis Wizard* function of the software PyMOL, which contains a collection of side-chain rotamers of frequent occurrence in proteins and enables selection of the one that better fits the needs of the structure. In this case, the selection of a suitable rotamer allowed the side chain of Phe120 at every protein subunit to accommodate among the nearby residues and minimize atomic clashes (see Figure 4.68 of the Results chapter).

Later, a set of quality check operations were carried out with the *structure checking* tool. Such a procedure provided confirmation that none of the following concerns are present in the structure: incorrect side-chain chirality, unusual peptide bond dihedrals, and severe atomic clashes in general. On the other hand, an assessment of the amide group orientation in the side chains of glutamine and asparagine residues was performed. The interpretation of the electron density signal obtained from X-ray diffraction experiments may face ambiguities when resolving amide groups specifically, since the data can be compatible with two rotamers related by a 2-fold symmetry axis, with the oxygen and nitrogen atoms interchanging positions. Since amides can act simultaneously as hydrogen-bond donors and acceptors, the chemical context of the vicinity may be useful to refine the assignment of their orientation, so that the number of favorable contacts is maximized. Following this rationale, the *structure checking* tool optimized the orientation of 17 amide groups that were involved in unusual contacts. Finally, oxygen atoms were added to the C-termini to complete their backbone carboxyl groups.

The software propKa (v3.1.8) [451] and PDB2PQR (v2.1.1) [452] were applied to predict the formal charge of all protein residues at physiologic pH 7.4 and the protonation configuration of histidine residues. All predicted protonation states were assessed by manual inspection and rationally reconsidered where stable hydrogen-bond or electrostatic contacts are known to occur, for instance, at the binding site of ligands and at protein interfaces. No charged states different from the usual ones at neutral pH were assigned to any amino-acid side chain. All N- and C-termini were set to their charged forms. This protonation configuration scheme was applied to all simulation conditions of the project, and the corresponding hydrogen atoms were modeled at each protein position with the *pdb2gmx* tool from GROMACS (v2018.3) [453].

3.1.2 Structures of the holoenzyme complexes

This section gives details of the sources and procedures that were applied to model the holoenzyme complexes that were studied in this project. Each holoenzyme condition incorporates a particular combination of small molecules. Besides the ligands of the protein, crystallographic water molecules (oxygen atoms) located at the coordination spheres of the metals were also included in every complex. The breakdown of this information is compiled in Table 3.1.

The structure and binding orientation of all the involved ligands were extracted from either the original model 2VGB or other PDB entries of pyruvate kinase. The PDB entry 4HYW [137] corresponds to an X-ray crystallographic structure (resolution of 2.35 Å) of PK from *Trypanosoma brucei* in complex with K^+, Mg^{2+} and fructose 2,6-bisphosphate. The PDB entry 4HYV [137] is similar to model 4HYW but incorporates PEP in a crystal-soaking experiment (resolution of 2.30 Å). The PDB entry 4FXF [149] corresponds to an X-ray crystallographic structure (resolution of 2.55 Å) of human PKM2 in complex with K^+, Mg^{2+}, oxalate (PEP analog), and MgATP.

Hydrogen atoms were added to complete the structures of the corresponding small molecules. The hydrogen atoms of the ligands were added at their corresponding sites at physiologic pH with the tool *reduce* [454] from the AmberTools suite (v17.3) [455]. The hydrogen atoms of the water molecules were firstly modeled at random orientations with the *pdb2gmx* tool from GROMACS; then, the instances with the most suitable orientations were rationally selected.

3.2 Parameterization

3.2.1 Protein parameters

The chosen force field to run MD simulations was AMBER99SB-ILDN [456], which is one of the gold-standard force fields that is widely recognized to provide an accurate description of many structural and dynamical properties of proteins [282], [284]. The AMBER99SB-ILDN force field is natively supported by the software GROMACS.

3.2.2 Ligand parameters

The inclusion of small organic molecules in the simulation system comes with the corresponding need to provide a given set of force-field parameters that are able to represent with enough reliability the physical properties of their atom types and their chemical connectivity in MD. This is not the case with common biopolymers such as proteins and nucleic acids, which are already well covered by the main force fields and therefore directly available within MD software. The missing parameters need to be either determined via quantum mechanics (QM) calculations and/or experimental data [282], [284], imported from complementary force fields or libraries, or extracted from the literature.

Table 3.1
Complexes of the holoenzyme conditions of the project

Condition name	Components	Source	Procedural notes
K-holo	Cofactor K^+	2VGB	
	2 water molecules coordinated to K^+	4HYW (oxygen atoms of residue IDs 693 and 796 from subunit B)	Superimposition of subunit B of 4HYW to each subunit of 2VGB using the backbone atoms of Glu241, Asp265, and Phe287 (PKR numbering) as the fitting group. [a]
K-Mg-holo	Cofactor K^+	Same as in K-holo	
	Cofactor Mg^{2+}	4HYW (residue ID 501 from subunit B)	Superimposition of subunit B of 4HYW to each subunit of 2VGB using the backbone atoms of Glu241, Asp265, and Phe287 (PKR numbering) as the fitting group. [a] After importing the involved molecules, new rotamers for the side chains of residues Phe287, Glu315, and Asp 339 of every subunit were chosen in order to resemble the binding geometry of Mg^{2+} in 4HYW as much as possible. [b]
	2 water molecules coordinated to K^+	4HYW (oxygen atoms with residue IDs 693, 747, 748, 756, 796, and 826 from subunit B)	
	4 water molecules coordinated to Mg^{2+}		
PEP-holo	Cofactor K^+	Same as in K-holo	
	Cofactor Mg^{2+}	2VGB (originally a Mn^{2+})	
	Substrate PEP	4HYV (residue ID 1003 from subunit A)	
	1 water molecule coordinated to K^+	4HYV (oxygen atoms with residue IDs 1168, 1259, and 1260 from subunit A)	Superimposition of PEP and the involved water molecules from subunit A of 4HYV to each subunit of 2VGB. The atoms of PEP that are analogously found in PGA were used as the fitting group. [a]
	1 extra water molecule surrounding K^+		
	1 water molecule coordinated to Mg^{2+}		
ADP-holo	Cofactor K^+	Same as in K-holo	
	2 water molecules coordinated to K^+	Same as in K-holo	
	Substrate MgADP (ADP and a complexed ion Mg^{2+})	4FXF (originally MgATP with residue IDs 604 and 605)	Superimposition of subunit D of 4FXF to each subunit of 2VGB directed by sequence alignment. [c] The γ-phosphate groups of ATP molecules were manually removed to obtain ADP molecules.
	4 water molecules coordinated to the ADP-bound Mg^{2+}	4FXF (oxygen atoms with residue IDs 701, 702, and 748 from subunit A) and rationally designed	The water molecule that was not imported from 4FXF was manually placed to complete the octahedral coordination sphere of the ADP-complexed Mg^{2+}. The placement was rationally designed to match the site where the atom O3P of PEP would be when both substrates are present.
PEP-ADP-holo	Cofactor K^+	Same as in K-holo	
	Cofactor Mg^{2+}	Same as in PEP-holo	
	Substrate PEP		

Table 3.1 (Continued)

	Substrate MgADP (ADP and a complexed ion Mg^{2+})	Same as in ADP-holo
	1 water molecule coordinated to K^+	Same as in PEP-holo
	1 water molecule coordinated to Mg^{2+}	
	3 water molecules coordinated to the ADP-bound Mg^{2+}	Same as in ADP-holo
FBP-holo	Cofactor K^+	Same as in K-holo
	Cofactor Mg^{2+}	Same as in K-Mg-holo
	Allosteric activator FBP	2VGB
	2 water molecules coordinated to K^+	Same as in K-Mg-holo
	4 water molecules coordinated to Mg^{2+}	
Full-holo	Cofactor K^+	Same as in K-holo
	Cofactor Mg^{2+}	Same as in PEP-holo
	Substrate PEP	
	Substrate MgADP (ADP and a complexed ion Mg^{2+})	Same as in ADP-holo
	Allosteric activator FBP	Same as in FBP-holo
	1 water molecule coordinated to K^+	Same as in PEP-holo
	1 water molecule coordinated to Mg^{2+}	
	3 water molecules coordinated to the ADP-bound Mg^{2+}	Same as in ADP-holo

[a] Performed with the software VMD [12].

[b] Performed with the *Mutagenesis Wizard* function of the software PyMOL [373].

[C] Performed with the *align* command of the software PyMOL.

The bond, angle and dihedral parameters for ligands PEP and FBP were obtained from the General Amber Force Field (GAFF) [457]. GAFF is a force field that was specifically designed to provide parameters for arbitrary small molecules. It serves as a library of parameters derived from comprehensive empirical and QM data that covers almost all the organic chemical space that is made up of C, N, O, S, P, H, F, Cl, Br and I. An advantage of GAFF is that it was designed to be compatible with the AMBER macromolecular force fields, allowing to combine parameters from both sources in the same simulation. Thus, GAFF is frequently used to fill in the missing parameters of complexes between a macromolecule and a ligand [455], [457]. The program ACPYPE (v2020.10.24.12.16) [458] was used to select the GAFF parameters that best suit ligands PEP and FBP and build the topology files in GROMACS format.

The determination of the atomic charges for both ligands was accomplished by reproducing the protocol that is consistent with the use of GAFF and other force fields of the AMBER family (and specifically AMBER99SB-ILDN) [459], namely, the Restrained Electrostatic Potential (RESP) [460]. Prior to the charge derivation, a geometry optimization of the structure is computed at the QM level, treating the system with the second-order Møller-Plesset perturbation theory (MP2) [461] approach. MP2 is a post-Hartree-Fock method that adds a second-order perturbation to provide an estimation of electron correlation effects on the energy and on the wave function. The calculation uses the polarized split-valence double-zeta basis set 6-31G(d) [462], [463], where core orbitals are built as a contracted Gaussian-type orbital of 6 functions and valence orbitals are built as the combination of two contracted Gaussian-type orbital (with 3 and 1 functions respectively), and *d*-type polarization functions are applied to non-hydrogen atoms. Geometry optimizations were carried out using redundant internal coordinates, with the threshold of convergence for the root-mean-square force set to 3×10^{-4} Hartree·Bohr^{-1} (1 Hartree = 627.5095 kcal·mol^{-1}, 1 Bohr = 0.52917721092 Å).

After the geometry optimization, atomic charges are determined at the QM level by performing a single-point energy calculation, treating the system with the Hartree-Fock method and with the basis set 6-31G(d). In this stage, a molecular electrostatic potential (MEP) is calculated with the Merz-Singh-Kollman (MK) scheme. The points of the MEP are defined with the original method [464], [465], with 4 layers per atom, the first one located at 1.4 times the van der Waals radii and with an increment of the scaling factor 0.2 for the rest; with the value of 1 for the density of grid points per Å^2. Afterwards, the calculated atomic charges are finally redistributed according to the RESP procedure, in a two-stage fit [466] where rotationally degenerate atoms receive equivalent charge values. The resulting charge distribution, even if generated in the gas phase, is able to reproduce solution phase interactions [459], [460]. All QM calculations were carried out with the software Gaussian16 (Rev. B.01) [467], at the supercomputing facilities of the Molecular Modeling and Bioinformatics (MMB) group of the Institute for Research in Biomedicine of Barcelona (IRB). Each calculation was parallelized in 16 cores of the *hpcluster* supercomputer. The input, log, and checkpoint files related to the QM calculations with Gaussian16 are available as part of the online supplementary material of this thesis (see Appendix A). The subsequent RESP fitting workflow was carried out with the programs *antechamber*, *espgen*, *respgen*, and *resp* from the AmberTools suite.

On the other hand, the parameters for the ligand ADP were extracted from the work of Meagher, Redman, and Carlson [468], available at the AMBER parameter database (http://amber.manchester.ac.uk/). Their parameterization procedure was equivalent to that followed in this work for ligands PEP and FBP, also intended to be compatible with the force fields of the

AMBER99 set. The obtained topology files were adapted to the GROMACS format with the program ACPYPE.

3.2.3 Metal-center parameters

The holoenzyme complexes contain the metal ligands K^+ and Mg^{2+}. These occupy defined sites of the metalloprotein (metal centers) and form coordination complexes with amino acids, the substrates PEP and ADP, and water molecules. The simulation of metal centers requires specific sets of force-field parameters, capable of both modeling the involved coordination bonds and maintaining the stability of the binding geometry. This section details the procedures followed to incorporate suitable parameters for treating the metal centers with a bonded/non-bonded hybrid model.

3.2.3.1 Van der Waals parameters

The van der Waals parameters allow the modeling of the attractive and repulsive forces that arise between two atoms based on their relative proximity and polarizability. The van der Waals parameters for K^+ and Mg^{2+} were obtained from the works of Merz, Li and colleagues [469], [470], and more specifically from their ion-oxygen distance (IOD) parameter sets for the 12-6 Lennard-Jones non-bonded model.

Depending on the implementation of the equation of the Lennard-Jones potential, the van der Waals parameters can be expressed in different terms. In this case, such parameters were obtained in the form of $R_{min}/2$ (Å) and ε (kcal·mol^{-1}), which is the format in which they are typically given when working with the force fields of the AMBER family. However, the implementation of the MD algorithm in GROMACS expresses the parameters in the form of σ (nm) and ε (kJ·mol^{-1}). The conversion of the first parameter requires multiplying by a factor of 2 to obtain R_{min}, then applying Equation 3.1,

$$\sigma = \frac{R_{min}}{\sqrt[6]{2}} \tag{3.1}$$

and finally multiplying by a factor of 10 to express the value in nanometers. The conversion of the second parameter only requires applying the conversion of units (1 kcal = 4.184 kJ). Table 3.2 shows the original values from the literature and after the conversion.

Table 3.2

Van der Waals parameters of the metal ligands in the holoenzyme complexes

Ionic species	Van der Waals parameters				References
	in AMBER format		in GROMACS format		
	$R_{min}/2$ (Å)	ε $\left(kcal/mol\right)$	σ (nm)	ε $\left(kJ/mol\right)$	
K^+	1.745	1.70181×10^{-1}	3.10924×10^{-1}	7.12036×10^{-1}	IOD parameter set from [469]
Mg^{2+}	1.395	1.49170×10^{-2}	2.48561×10^{-1}	6.24127×10^{-2}	IOD parameter set from [470]

3.2.3.2 Bonded parameters and atomic charges

The bonded parameters and the atomic charge distributions at the metal centers were generated with the protocol suggested by Li and Merz with their software Metal Center Parameter Builder (Python version 4.0; MCPB.py) [471], [472], from the AmberTools suite. The approach of MCPB.py was complemented with a number of modifications based on the modeling standards for metalloproteins that are available in the literature and that apply to the complexity of the metal centers of the PKR structure. The steps of the followed strategy are listed hereafter and further presented and discussed in the Results and Discussion chapters.

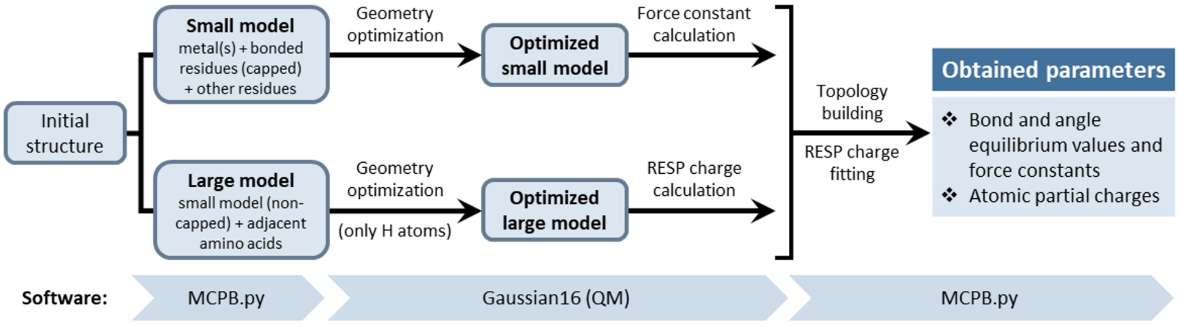


Figure 3.1. Main steps of the metal-center parameterization approach.

For each metal center, the parameterization began with the definition of the so-called cluster model, which corresponds to the portion of the system that accounts for the metal and the surrounding residues (amino acids, ligands, water molecules) that are relevant to describe the chemical environment of the coordination complex. Two different representations of the cluster model are employed in the standard methodology of MCPB.py. In the “small model”, the involved amino acids are represented with only the fraction of the structure that is essential to describe the interactions with the metal center. In the “large model”, the involved amino acids are fully represented and in addition the adjacent or a few intermediate residues are included. MCPB.py builds each model automatically, based on the input residue specifications of the user, according to the integrated capping scheme [471]. The different partial amino-acid structures can be represented either as acetyl (ACE) and/or N-methyl amide (NME) groups in the case of backbone moieties, as CH_3-R where R represents a side-chain group, or as glycine residues in the case of the intermediate amino acids of the chain. All non–amino-acid residues are always fully represented. Table 3.3 shows the amino acids included in the cluster model of each holoenzyme condition together with their representation in the small and large models. Two QM calculations were run in parallel with the small and the large model to derive, respectively, the bonded parameters and the atomic charges. All QM computations were performed with the software Gaussian16 (Rev. B.01), at the supercomputing facilities of the Molecular Modeling and Bioinformatics (MMB) group of the Institute for Research in Biomedicine of Barcelona (IRB). Each calculation was parallelized in 16 cores of the *hpcluster* supercomputer.

The first QM computation comprised a geometry optimization of the small model. The system was treated with the hybrid functional B3LYP, which includes a mixture of Hartree-Fock exchange with Density Functional Theory exchange-correlation. Specifically, B3LYP combines the B3 exchange functional [473], with the LYP correlation functional [474], with a 20% of exact exchange [475]. The Grimme D3 model of empirical dispersion correction with Becke-Johnson damping (GD3BJ) [476] was incorporated to properly take into account the long-range electron correlation effects in DFT. The chosen basis set was 6-31+G(d,p) [477], which is like 6-31G(d) but including *p*-type polarization functions on hydrogen atoms and *s*-type and *p*-type diffuse functions on non-hydrogen atoms. Solvent effects were introduced with a conductor-like Polarizable Continuum Model (C-PCM) [478], using the empirical value of 20 for the dielectric constant (ε) to simulate the average effect of both the protein and the water medium surrounding the metal centers [479], [480].

Table 3.3
PKR amino acids included in the cluster model of each holoenzyme condition.

Residue	Role	Representation in the small / large model				
		K-holo	**K-Mg-holo and FBP-holo**	**PEP-holo**	**ADP-holo**	**PEP-ADP-holo and Full-holo**
Ala115		- / ACE	- / ACE	- / ACE	- / ACE	- / ACE
Arg116	Ligand of K$^+$	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
Leu117		- / Gly	- / Gly	- / Gly	- / Gly	- / Gly
Asn118	Ligand of K$^+$	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
Phe119		- / Gly	- / Gly	- / Gly	- / Gly	- / Gly
Ser120	Ligand of K$^+$	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
His121	Chemical context	- / NME	- / NME	- / NME	CH_3-R / normal	CH_3-R / norm
Gly122		- / -	- / -	- / -	- / NME	- / NME
Leu155		- / ACE	- / ACE	- / ACE	- / ACE	- / ACE
Asp156	Ligand of K$^+$	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
Thr157	Ligand of K$^+$	ACE / normal	ACE / normal	ACE / normal	ACE / normal	ACE / norm
Lys158		NME / Gly	NME / Gly	NME / Gly	NME / Gly	NME / Gly
Gly159		- / Gly	- / Gly	- / Gly	- / Gly	- / Gly
Pro160		- / Gly	- / Gly	- / Gly	- / Gly	- / Gly
Glu161	Chemical context	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
Ile162		- / NME	- / NME	- / NME	- / NME	- / NME
Ala285		- / ACE	- / ACE	- / ACE	- / ACE	- / ACE
Ser286	Chemical context	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
Phe287		- / NME	- / NME	- / NME	- / NME	- / NME
Ser312		- / ACE	- / ACE	- / ACE	- / ACE	- / ACE
Lys313	Chemical context	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / normal	CH_3-R / norm
Ile314		- / NME	- / Gly	- / Gly	- / NME	- / Gly
Glu315	Ligand of Mg^{2+}	- / -	CH_3-R / normal	CH_3-R / normal	- / -	CH_3-R / norm
Asn316		- / -	- / NME	- / NME	- / -	- / NME
Gly338		- / -	- / ACE	- / ACE	- / -	- / ACE
Asp339	Ligand of Mg^{2+}	- / -	CH_3-R / normal	CH_3-R / normal	- / -	CH_3-R / norm
Leu340		- / -	- / NME	- / NME	- / -	- / NME
Ala370		- / -	- / -	- / -	- / ACE	- / ACE
Thr371	Chemical context	- / -	- / -	- / -	CH_3-R / normal	CH_3-R / norm
Gln372		- / -	- / -	- / -	- / NME	- / NME
Leu404		- / -	- / -	- / -	- / ACE	- / ACE
Ser405	Chemical context	- / -	- / -	- / -	CH_3-R / normal	CH_3-R / norm
Gly406		- / -	- / -	- / -	- / NME	- / NME

NOTE. All residues with annotation on their role correspond to the input given to MCPB.py to build the cluster model. The rest were automatically included by the software in the large mode adjacent or intermediate residues, according to the capping scheme. Abbreviations: ACE, acetyl group; NME, N-methyl amide group.

The geometry optimization was carried out using redundant internal coordinates, with the threshold of convergence for the root-mean-square force set to 3×10^{-4} Hartree·Bohr^{-1}. A rational selection of frozen atoms was applied, comprising all backbone atoms, all hydrogen atoms of the capping groups, and several individual atoms that account for the structural context of the boundaries of the cluster model. With the optimized geometry, a step of force-constant calculation was performed at the same QM level of theory to derive the bond and angle parameters of the coordination complex from the corresponding Hessian matrix with the Seminario method [481]. Dihedral parameters were not generated, as is usually recommended for the bonded model of metal centers [482]–[484].

The second QM computation comprised a geometry optimization of only the hydrogen atoms of the large model. The simulation settings were the same as in the geometry optimization of the small model, except for the frozen atoms which of course were all non-hydrogen atoms. With the geometry optimized, a single-point energy calculation was performed to determine atomic charges, this time with the basis set 6-311++G(3df,3pd). This is a polarized split-valence triple-zeta basis set where core orbitals are built as a contracted Gaussian-type orbital of 6 functions and valence orbitals are built as the combination of three contracted Gaussian-type orbital (with 3, 1, and 1 functions, respectively). The basis set also contains several functions: 3 sets of *d*-type and a set of *f*-type polarization functions (applied to non-hydrogen atoms), 3 sets of *p*-type and a set of *d*-type polarization functions (applied to hydrogen atoms), a set of *s*-type and *p*-type diffuse functions (applied to non-hydrogen atoms), and a set of *s*-type diffuse functions (applied to hydrogen atoms). In this stage, a MEP was calculated with the MK scheme and the final atomic charges were determined with the RESP procedure. The ChgModB approach of MCPB.py was followed for the RESP fitting, in which: i) all amino acids that bind the metal via a side-chain group have the atomic charges of their backbone heavy atoms (CA, N, C and O) fully restrained to their force-field values (*i.e.*, AMBER99SB-ILDN), and ii) all amino acids that bind the metal via a backbone atom have all their atomic charges flexible. Additionally, the rest of amino acids of the cluster model (those that do not bind the metal), as well as all water molecules, had their atomic charges fully restrained to their force-field values.

The input, log, and checkpoint files related to the QM calculations with Gaussian16 are available as part of the online supplementary material of this thesis (see Appendix A). After the QM calculations, the program *tleap* from the AmberTools suite was used to produce the corresponding topology files, which were converted to the GROMACS format with the program ACPYPE.

The parameterization up until this point corresponds to the implementation of a bonded model to treat the metal centers. However, after a stage of testing the parameters in MD simulation, the models were fine-tuned with an empirical approach that combines properties of both the bonded and the non-bonded schemes. Such a strategy consists in selectively treating coordination bonds as simple harmonic restraints by employing the corresponding generated bond parameters. As a result, the van der Waals and electrostatic interactions of the neighboring atoms up to 3 bonds away are accounted for as dictated by the force field, as opposed to the treatment of regular chemical bonds in which these interactions are abolished or scaled down. This was accomplished by manually establishing such bonds in the GROMACS topology as “type 6” bonds (harmonic potentials) and deleting the associated 1-4 interaction scaling factors. The implementation of this methodology may serve to mitigate structural distortion in the vicinity of specific bonds, which may arise due to the absence of 1-3 terms in the conventional bonded model [484], [485]. This approach was applied to all coordination bonds between: i) metals and amino acids, ii) PEP and K^+, and iii) ADP and its complexed Mg^{2+}.

In respect of the angle parameters, they were retained since they proved to be beneficial to maintain the experimental binding geometry of the complexes. The only angle parameter that was discarded was that between K^+ and the hydroxyl group of Ser120 because otherwise it constrains the needed mobility of the group to find stable transient hydrogen bonds with either Glu161 or any other water molecule nearby. Finally, all water molecules were fully treated with the non-bonded model. Both their bond and angle parameters with the metals were discarded in order to let water flow freely and occupy the coordination binding sites transitionally according to their stability in simulation.

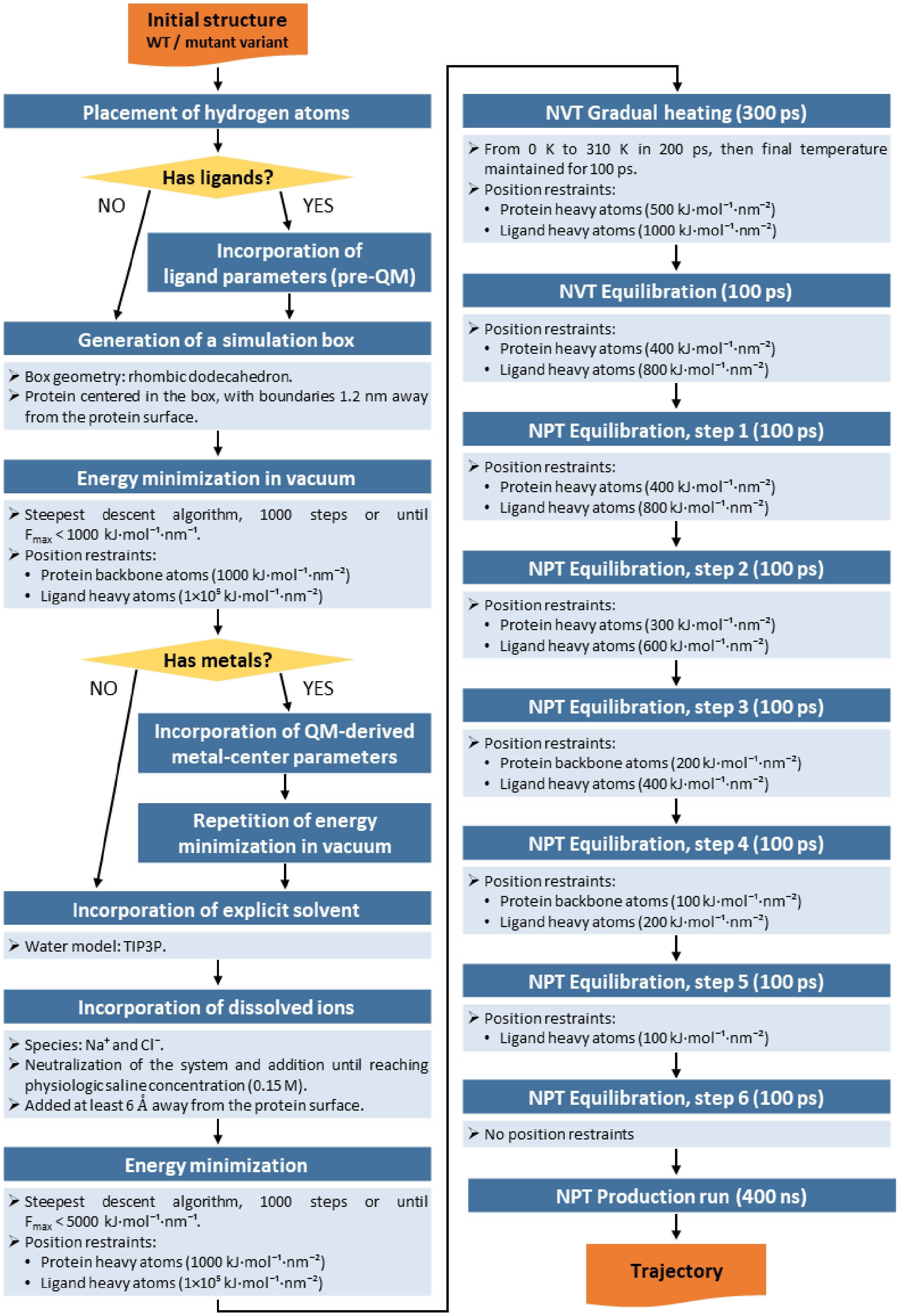
3.2.3.3 Metal centers of the mutant protein systems

The procedure described above for the parameterization of the metal centers was carried out with the wild-type (WT) protein system. Later, these metal-center parameters were transferred to the mutant protein systems to economize the computational cost of this stage of the project. The majority of the missense variants taken into account in this project fall outside of the metal binding sites, therefore the metal-center parameters are fully transferable. However, some of the modeled amino-acid substitutions do affect a few positions of the cluster model that are listed in Table 3.3. For those mutant protein systems, the transfer of the metal-center parameters was addressed according to the following criteria.

- If the mutated position is not part of a coordination complex (the original amino acid does not bind to any metal):
    - The bonded parameters remain unchanged from the original WT parameterization.
    - Atomic charges also preserve their original values, except for the mutant amino acid which is given its corresponding force-field atomic charge values (in the same way as the original amino acid was treated).
- If the mutated position belongs to a coordination complex (the original amino acid binds to a metal):
    - The bonded parameters that used to apply to this position are removed while the rest are retained.
    - Atomic charges undergo a mild redistribution to compensate for the differences caused by the methods employed to assign the total charge between the original and the mutant amino acids. The original amino acid had QM-derived atomic charges, and therefore its total charge does not necessarily correspond to an integer value (but is usually close). In contrast, the mutant amino acid adopts its corresponding force-field atomic charge values, which add up to an integer value. This generates a deviation of a small decimal amount that is automatically redistributed by the program ACPYPE. Such cases were manually assessed to rule out any undesired imbalance.

3.3 Molecular dynamics protocol

The MD protocol consisted of several concatenated operations and simulations, starting with the setup structure as the first input for the workflow until producing the last simulation, called the production run, which accounts for the actual analyzable trajectory data. Figure 3.2 shows a comprehensive flow chart of the process. The present section further elaborates on the technical details of each stage, including the generation of the simulation box, the generation of the physiologic aqueous medium, the energy minimizations, and the equilibration and production phases.



Initial structure
WT / mutant variant
Placement of hydrogen atoms
Has ligands?
NO
YES
Incorporation of ligand parameters (pre-QM)
Generation of a simulation box

- Box geometry: rhombic dodecahedron.
- Protein centered in the box, with boundaries 1.2 nm away from the protein surface.

Energy minimization in vacuum

- Steepest descent algorithm, 1000 steps or until $F_{max} < 1000$ kJ·mol^{-1}·nm^{-1}.
- Position restraints:
  - Protein backbone atoms (1000 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (1×10^5 kJ·mol^{-1}·nm^{-2})

Has metals?
NO
YES
Incorporation of QM-derived metal-center parameters
Repetition of energy minimization in vacuum
Incorporation of explicit solvent

- Water model: TIP3P.

Incorporation of dissolved ions

- Species: Na$^+$ and Cl$^-$.
- Neutralization of the system and addition until reaching physiologic saline concentration (0.15 M).
- Added at least 6 Å away from the protein surface.

- Steepest descent algorithm, 1000 steps or until $F_{max} < 5000$ kJ·mol^{-1}·nm^{-1}.
- Position restraints:
  - Protein heavy atoms (1000 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (1×10^5 kJ·mol^{-1}·nm^{-2})

- From 0 K to 310 K in 200 ps, then final temperature maintained for 100 ps.
- Position restraints:
  - Protein heavy atoms (500 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (1000 kJ·mol^{-1}·nm^{-2})

- Position restraints:
  - Protein heavy atoms (400 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (800 kJ·mol^{-1}·nm^{-2})

- Position restraints:
  - Protein heavy atoms (400 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (800 kJ·mol^{-1}·nm^{-2})

- Position restraints:
  - Protein heavy atoms (300 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (600 kJ·mol^{-1}·nm^{-2})

- Position restraints:
  - Protein backbone atoms (200 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (400 kJ·mol^{-1}·nm^{-2})

- Position restraints:
  - Protein backbone atoms (100 kJ·mol^{-1}·nm^{-2})
  - Ligand heavy atoms (200 kJ·mol^{-1}·nm^{-2})

- Position restraints:
  - Ligand heavy atoms (100 kJ·mol^{-1}·nm^{-2})

- No position restraints

Figure 3.2. MD protocol and steps of the workflow.

Additionally, the flow chart facilitates identification of the points of the process where the ligand and metal-center parameters were incorporated. The original files of the AMBER99SB-ILDN force field, distributed in the native installation of GROMACS, were modified to include new entries with the data of the ligands of this project (custom residue and atom names and types, non-bonded terms). Details of this process are available as part of the online supplementary material of this thesis (see Appendix A).

The workflow was designed following the standard procedures and the recommended guidelines applied to protein MD simulations [344], [348]–[351], and further tuned to treat a large protein system such as PKR. The assembly of the steps of the workflow was carried out via a combination of in-house scripts (Python and Bash programming languages) and calls to diverse GROMACS tools. The workflow could be accurately reproduced at present with the equivalent modules from the BioBB library [274].

The aqueous medium was modeled with explicit solvent to account for water-mediated interatomic interactions. The water model TIP3P [486] was employed. TIP3P is a three-point water model, which means that point charges are defined at the oxygen and the two hydrogen atoms. It is a standardly used explicit model in all-atom MD of proteins with the AMBER force fields [356].

Each system was simulated in Periodic Boundary Conditions (PBC), whereby the entire system is placed in a unit cell that is infinitely replicated in the three spatial dimensions. This construction eliminates surface effects by following the minimum image convention: particles and interactions that cross the boundary of the unit cell simply emerge back from the opposite side of the box thanks to periodic images. Consequently, the use of PBC enables conservation of the total number of particles in the simulation and allows simulating bulk solid and liquid properties.

The lattice of periodic unit cells can be constructed with different box geometries. A large enough box is needed to avoid the interaction of the protein with its own periodic images [348]. However, in simulations with explicit solvent, the greater the amount of water molecules within the system the greater the computational cost. Therefore, it was important to find an appropriate balance. In this study, a rhombic dodecahedron box was used to optimize computational impact, as fewer solvent molecules are required to fill the box around the protein [170], [453]. The protein was placed in the center of the box, with the boundaries at a distance of 1.2 nm of the outermost atom, which accounted for sufficient size to elude self-interaction of the protein across PBC.

After filling the box with water solvent, physiological conditions were mimicked by incorporating dissolved ions. Accordingly, several solvent water molecules were replaced by the standard monoatomic ions, Na^+ and Cl^-, firstly to neutralize the system and then subsequently until reaching an ion concentration of 0.15 M. The parameters used to treat these ions were the standard ones provided by GROMACS.

A customized restriction in ionic placement was added at this stage. Normally, with the algorithm of the *genion* tool of GROMACS, water molecules are randomly replaced by the corresponding number of ions of both species. However, an implication of this method is the fact that an ion can be artificially placed in energetically unfavorable sites in contact with the protein by chance, thus potentially introducing local anomalies especially if a problematic ion is sterically trapped for a significant amount of simulation time. Therefore, a restriction was incorporated to the algorithm to add ions at least 6 Å away from the protein surface. Of course, during simulation, this restrain was no longer active; ions diffused and eventually made favorable contacts with the protein surface.

Two energy minimization (EM) runs were applied at different points of the workflow. The first EM was applied before the addition of the solvent (EM in vacuum), to relax the possible bad contacts involving side-chain atoms that may arise both from inaccuracies in the experimental data of the crystallographic structure and the addition of hydrogen atoms. In the case of the holo simulations, this EM calculation was run twice: before and after the incorporation of the QM-derived parameters of the metal centers. For this EM stage, the steepest descent algorithm was used until reaching a maximum force of 1000 kJ·mol^{-1}·nm^{-1} with a maximum of 1000 steps. Strong position restraints were applied on the protein backbone atoms using a force constant of 1000 kJ·mol^{-1}·nm^{-2}. In the holo simulations, the heavy atoms of ligands, including crystallographic water molecules, were fully restrained (represented with a force constant of 1×10^{5} kJ·mol^{-1}·nm^{-2}). The second EM was applied after solvation and incorporation of dissolved ions. The aim of this EM stage was to optimize the orientation of the added solvent molecules. The steepest descent algorithm was used until reaching a maximum force of 1000 kJ·mol^{-1}·nm^{-1} with a maximum of 5000 steps. Strong position restraints were applied on the protein heavy atoms (1000 kJ·mol^{-1}·nm^{-2}). In holo simulations, the heavy atoms of ligands were fully restrained (1×10^{5} kJ·mol^{-1}·nm^{-2}).

After the EM stages, the MD simulation formally begins. The first stages of a standard MD simulation of a protein constitute the equilibration phase, typically aimed at relaxing the system from the non-equilibrium initial conditions and bringing it to a stable state in the desired conditions (temperature and pressure). The equilibration phase generally comprises two stages in different thermodynamic conditions, referred to as ensembles in statistical mechanics. First, the NVT ensemble (or canonical ensemble) is used, whereby the number of particles (N), volume (V), and temperature (T) are kept constant. After temperature is stabilized, the NPT ensemble (or isothermal-isobaric ensemble) is used to keep both temperature and pressure (P) constant, allowing the volume to equilibrate to provide the correct density.

The equilibration phase was divided into multiple steps to conduct a progressive relaxation of the system, starting with strong position restraints and gradually releasing them. Figure 3.2 contains a detailed breakdown of the applied values of force constant and the corresponding recipient groups of atoms. The NVT ensemble was applied in the first two stages of the equilibration phase, first in a gradual heating of the system from 0 K to 310 K (physiologic temperature) and then maintaining the final temperature. Subsequently, the NPT ensemble was applied in six equilibration stages. The total simulation time of the equilibration phase was 1 ns. Finally, after the equilibration phase, the production phase comprised a simulation time of 400 ns, from which the first 25 ns were discarded as an additional extension of the equilibration based on the stability analysis of the trajectories (see section 4.1.2.1 of the Results chapter).

In NVT ensembles, the average temperature was maintained at 310 K via the velocity-rescaling thermostat [487] in two separate baths: one for the protein (plus ligands, if present) and another for the solvent and dissolved ions. In NPT ensembles, pressure was maintained constant via the Berendsen [488] or the Parrinello-Rahman [489] barostats in restrained and free simulations, respectively. Moreover, restrained simulations were treated with a scaling of the center of mass of the reference coordinates with the scaling matrix of the pressure coupling.

Intramolecular interactions were treated with the AMBER99SB-ILDN force field [456]. Long-range electrostatic interactions were treated with Particle Mesh Ewald (PME) [490], for full-system periodic electrostatics. The cut-off distance for both electrostatic and van der Waals pairwise interactions was set to 1 nm. The leap-frog algorithm [491] was used to calculate the atomic velocities and coordinates with an integration time step of 2 fs. The LINCS algorithm (Linear Constraint Solver) [492] was used to maintain bonds involving any hydrogen atom at their equilibrium values, as the frequency of vibration of such bonds is in the order of 1 fs (thus, below the integration time step).

Table 3.4 shows a breakdown of the total collection of MD simulations of this study specifying the simulation replicates produced per PKR condition. Each replicate was started with different sets of initial atomic velocities generated for the same initial structure. For the WT apo condition, two different batches of trajectory ensembles were produced. The main analyses of the project were conducted with the first batch (5 simulation replicates), whereas the second batch (5 additional simulation replicates) was procured with the sole purpose of assessing the replicability of the experiments with respect to the first batch. For each WT holo condition, 5 simulation replicates were performed. The simulation of the PKR mutant variants covered the apo condition and a single holo condition, either K-Mg-holo, PEP-holo, PEP-ADP-holo, or FBP-holo, rationally selected according to the expected type of dysfunction that may be manifested in dynamics, on the basis of both the available clinical annotations and the location of the particular amino-acid substitution. Given the high number of PKR variants studied, the simulation replicates of these systems were reduced to 3 per variant and condition, to make the computational cost more feasible.

Table 3.4

MD simulations per PKR condition

Condition	Trajectory ensembles	Simulation replicates per ensemble	Total simulations
WT			
Apo	2 [a]	5	**10**
K-holo	1	5	**5**
K-Mg-holo	1	5	**5**
PEP-holo	1	5	**5**
ADP-holo	1	5	**5**
PEP-ADP-holo	1	5	**5**
FBP-holo	1	5	**5**
Full-holo	1	5	**5**
Mutant variants			
Apo	61 [b]	3	**183**
Holo [c]	61 [b]	3	**183**
Total	**131**		**411**

[a] The main analyses of the project were conducted with the first batch of trajectory ensembles of the WT apo condition. The second batch was procured with the sole purpose of assessing the replicability of the experiments with respect to the first batch.

[b] A trajectory ensemble for each of the 61 simulated missense variants of PKR.

[c] A single holo condition, either K-Mg-holo, PEP-holo, PEP-ADP-holo, or FBP-holo, was simulated for each PKR variant.

The collection of trajectories of this research project have been made publicly available for the sake of reproducibility and to facilitate reutilization for further analyses. The complete set of simulations, as well as standard trajectory analyses can be found at the online database PKLR from the Molecular Dynamics Data Bank (https://pklr.mddbr.eu), a project funded by the European Union’s Horizon Europe programme under grant agreement 101094651. In addition, the structure and topology files of all the systems subjected to MD are available as part of the online supplementary material of this thesis (see Appendix A).

All simulations were carried out at the supercomputing facilities of Barcelona Supercomputing Center (BSC), using the computational resources of the supercomputer MareNostrum4, and with the GROMACS (v2018.3) software. Each execution of the workflow was parallelized in 192 MareNostrum4 cores, requiring approximately 40,000 CPU hours (over 8 days of execution) to complete the simulations and a disk storage of approximately 280 GB. The total computation time and disk storage of the project comprised over 16 million CPU hours and 120 TB, respectively. The project activities BCV-2019-1-0004, BCV-2019-3-0006, BCV-2020-3-0005, and BCV-2021-3-0002 of the Red Española de Supercomputación (RES) provided the needed support to cope with the expended computational resources.

3.4 Standard trajectory analysis

3.4.1 Removal of PBC

All trajectories were processed to remove periodicity conditions prior to conducting further analysis. This step is essential for two reasons: firstly, it facilitates a more efficient visualization of the trajectories; secondly, it eliminates any ambiguity about the atomic positions that should be considered when a molecule crosses the boundaries of the PBC box at given time points. Consequently, this ensures that each analytical technique can interpret the data adequately.

The removal of PBC was achieved with consecutive operations conducted with the *trjconv* tool from GROMACS, following the workflow that is suggested at the documentation of the software. First, the trajectory data was re-written so that all molecules that were broken across PBC became whole (*i.e.*, the broken fragments reunited at the box boundaries so that no molecules remained split). Secondly, a similar step was applied to maintain the tetrameric assembly (and the bound ligands, when applicable) unseparated. Lastly, the protein was centered in the box, and its rotational and translational components were disregarded by applying a structural superposition (least-squares fitting) with the initial structure of the production run.

3.4.2 Root-mean-square deviation

The root-mean-square deviation (RMSD) of atomic positions is a measure of structural similarity between two structures of a molecule, expressed as the average distance between two sets of atomic coordinates. Given two structures v and w each with N atoms under analysis, the RMSD follows the equation:

$$RMSD\ (v, w) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(v_{i,x} - w_{i,x}\right)^2 + \left(v_{i,y} - w_{i\ ,y}\right)^2 + \ \left(v_{i,z} - w_{i,z}\right)^2\right)} \tag{3.2}$$

where v_i and w_i refer to the i-th atom of each corresponding structure, with the subscripts x, y, and z designating their Cartesian coordinates in the 3-dimensional space. RMSD values are expressed in length units, typically in angstroms or nanometers. RMSD is standardly used as a quantitative measure to express the structural deviation of one or several structures with respect to a reference conformation. The lower the value, the more similar the conformations. RMSD may also be computed by mass-weighting the different particles, which incorporates atomic masses as terms of the equation.

RMSD is generally computed after optimal superposition between the structures to remove the effects of global translation and rotation and thus focus only on the internal rearrangements of the molecule. Accordingly, in the interpretation of RMSD, it is crucial to identify the regions of the molecule that exhibit the highest degree of flexibility or mobility. Perturbations confined to a specific region, such as surface loops or hinged domains, have the potential to create large RMSD values. This could lead to the interpretation that the two conformations are highly dissimilar even when the overall structure yields substantial overlap. For this reason, RMSD may not only be calculated over all atoms of the structure, but also on a subset either to focus on a region of interest (single domains, active site…) or to exclude highly variable elements. When analyzing global conformational changes between protein structures, only backbone or α-carbon atoms are typically considered. Using such subsets of atoms also allows for comparison between protein structures with sequence variants, thanks to the equivalencies between the pairs of backbone atoms regardless of the particular amino acids [493].

The frequent uses of RMSD comprise characterization of the quality of biomolecular simulations, clustering of related conformations, and definition of free-energy landscapes employing it as a reaction coordinate [494]. In this study, RMSD was used to assess the structural stability during the MD simulations, by comparing the structure at each frame with respect to the initial structure, and to evaluate and compare the modeled coordination complexes along geometry optimizations and MD simulations.

3.4.3 Root-mean-square fluctuation

The atom positional root-mean-square fluctuation (RMSF) quantifies the fluctuation of an atomic position about its average value in an ensemble of conformations of a molecule. This measure is related to RMSD, but instead of expressing structural divergence between structures, it reveals which fragments of the system are the most mobile. Given an ensemble of T structures, the RMSF of an atom i is calculated as:

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{j=1}^{T}\left(\vec{v}_{i,j} - \langle\vec{v}_i\rangle\right)^2} \tag{3.3}$$

Where $\vec{v}_{i,j}$ is its position (set of Cartesian coordinates) in the conformation j, and $\langle\vec{v}_i\rangle$ is its average position over the whole ensemble. RMSF values are expressed in length units, typically in angstroms or nanometers. The higher the RMSF value, the greater the mobility of the atom within the ensemble.

The most common use of RMSF with protein structures is to analyze the flexibility of each residue along the trajectory of an MD simulation. As with RMSD, the RMSF analysis is typically preceded by structural superposition of the structures and restricted to backbone or α-carbon atoms, except when the localized fluctuations of side chains are also of interest [494]. By plotting the residue number *vs.* RMSF per residue, the regions of the system with higher flexibility (higher contribution to molecular motion) can be identified. In this study, RMSF was employed for such purposes, to identify the most mobile regions of the protein in the simulations and, more specifically, to assess the structural variability of each amino acid included in the cluster models of the metal centers of the protein.

3.4.4 Free-energy difference between states from simulation trajectories

The free-energy changes associated with a transition of the molecular system in one state or the other (conformational changes, interactions…) along a trajectory can be determined by measuring the probability of finding the system in each of the defined states. An interconversion between states can happen spontaneously if the associated free-energy barrier is sufficiently low to allow the transition with a certain probability. In the absence of an external source of energy, the system will tend to sample the lowest free-energy states more often. Free-energy changes can be described with multiple levels of complexity by means of different approaches. By measuring this thermodynamic quantity with computational techniques, we can inquire about the fundamental properties of the system and compare them with the experimental data.

With MD simulations, we use the formalisms of statistical mechanics whereby the difference in free energy ($\Delta G_{A\to B}$ in J·mol^{-1}) between two states A and B can be expressed through the probability of observation of those states as:

$$\Delta G_{A\to B} = -RT\ln\frac{P_B}{P_A} \tag{3.4}$$

where R is the gas constant (8.314 J·K^{-1}·mol^{-1}), T is the temperature of the system (310 K in the simulations of this study), and P_A and P_B are the probabilities of observing states A and B, respectively. With this straightforward approach, the difference in free energy can be estimated just by determining the proportions of each corresponding state in the ensemble generated in the simulation.

The states can be defined as those configurations that satisfy a particular objective criterion such as the RMSD from a particular folded conformation. It is important to note that this technique is only valid under the assumption that the simulation time is enough to capture an equilibrium between states, *i.e.*, where the reversible transition has occurred with sufficient frequency in the ensemble to obtain reliable statistics [333], [495]–[497]. In this study, this method was applied to report the relative free-energy differences between the different coordination-complex configurations of the metal centers of the protein that occurred along the simulations.

3.5 Consensus Essential Dynamics Analysis

This section describes the methodological components that were integrated into the central analytic strategy that was designed in this thesis, namely, Consensus Essential Dynamics Analysis (CEDA). CEDA was applied to identify the most representative collective motions from a trajectory ensemble and subsequently compare the extracted dynamical traits between alternative trajectory ensembles in a unified framework. The following subsections present the theoretical basis of the techniques and metrics related to CEDA, while the rationale of their implementation in the actual analytical procedures is progressively presented with full detail throughout the Results chapter in parallel with the reports on the corresponding analyses (sections 4.1.3 and 4.2). In addition, a point-by-point summary of the full developed protocol has been included at the end of that chapter (section 4.3) to provide a unified recapitulation of the resulting approach.

3.5.1 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that is widely used in diverse scientific and technical fields to capture correlations among variables from high-dimensional data, thus enabling extraction of underlying key features. The application of PCA in trajectory analysis allows for the decomposition of the trajectory into simple linearly uncorrelated collective motions that account for most of the conformational variability that was sampled during the simulation. This procedure is also known as Essential Dynamics Analysis (EDA) because it achieves representation of dynamics in terms of a reduced set of variables that bear the "essential" dynamical information, often associated with the biologically relevant behavior of proteins [379], [391].

In this study, PCA was applied with the standard procedure using as input the atomic Cartesian coordinates of the protein along the trajectory. Usually, only backbone or α-carbon atoms are employed for the analysis because they suffice to account for the overall correlated motions and, in addition, the computational cost is dramatically reduced [498]. However, the analysis can be applied to any subset of atoms and subregions of the macromolecule when small-scale motions are of interest [379].

The first step in Cartesian PCA involves removing the global effects of rotational and translational motion to consider only the atomic fluctuations related to the internal rearrangements of the molecule. Accordingly, each frame in the trajectory is subjected to structural superposition on a reference structure. Importantly, this structure should be representative of the whole ensemble of conformations to yield optimal equivalent orientation of all the structures [391], [397].

A given conformation of a molecule can be mathematically represented as a vector $\mathbf{x}$ of $3N$ components, where N is the number of atoms, each with its set of three-dimensional Cartesian coordinates. Accordingly, an MD trajectory is a set of points scattered in this $3N$-dimensional space, such that each point $\mathbf{x}(t)$ corresponds to a given conformation sampled at a given time value t [363], [389], [438]. Despite the high dimensionality of this space, most degrees of freedom are constrained due to bonded interactions that of course prevent any pair of bonded atoms from fluctuating in an independent manner with respect to each other [438]. Further constraints are imposed by angle and dihedral energetic restrictions, non-bonded interactions, and steric hindrance. In folded proteins, the concept also scales up, with the different levels of structural organization narrowing down the available subspace. Regions outside the constrained subspace would imply deformation or unfolding of the protein and, thus, will never be visited. This fact, in turn, implies that the available conformational space is characterized by a substantial degree of correlation between the variables (*i.e.*, the Cartesian coordinates) because the structurally coupled groups of atoms move collectively [419].

The pairwise correlations are accounted for by calculating the covariance matrix $\mathbf{C}$ of the trajectory data. Alternatively, the correlation matrix can also be employed, which corresponds to the normalized version of the covariance matrix and is useful when the objective is to identify correlated motions without necessarily large amplitudes of motion, disregarding the skewing effect towards the largest atomic displacements [379], [404]. The covariance c_{ij} between two coordinates x_i and x_j is calculated as follows:

$$c_{ij} = \frac{1}{S}\sum_{t=1}^{S}\left(\left(x_i(t) - x_{ref,i}\right)\left(x_j(t) - x_{ref,j}\right)\right) \tag{3.5}$$

where $x_{ref,i}$ and $x_{ref,j}$ are the equivalent coordinate values of a reference structure, S is the total number of frames or conformations, and t denotes the time iteration from the first to the last frame of the trajectory data (although data can be provided in any order, as this analysis can be applied, for instance, to a set of non-temporal experimental structures). Typically, the average structure of the trajectory is selected as the reference structure, so that the deviations of each conformation are calculated with respect to the center of mass of the data. Nevertheless, in purpose-specific scenarios, a structure other than the average can be used to express deviation with respect to alternative reference points [237]. In the computation of the covariance matrix, the trajectory data is implicitly centered around the chosen reference structure.

Subsequently, the covariance matrix $\mathbf{C}$ is diagonalized via a mathematical procedure known as eigendecomposition, which produces a set of $3N$ pairs of eigenvalues (λ_1, λ_2, …, λ_{3N}) and eigenvectors ($\boldsymbol{r}_1$, $\boldsymbol{r}_2$, …, $\boldsymbol{r}_{3N}$), arranged as:

$$\mathbf{R}^{\mathrm{T}}\mathbf{CR} = \boldsymbol{\Lambda} \tag{3.6}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues diag(λ_1, λ_2, …, λ_{3N}), and $\mathbf{R}$ is an orthogonal coordinate transformation matrix with its columns being the eigenvectors ($\mathbf{R}^{\mathrm{T}}$ is its transpose). The set of eigenvectors constitutes a new basis set of orthonormal vectors that point in the directions of the dataset along which maximum variance is captured. Each eigenvector $\boldsymbol{r}_i$ is associated with the eigenvalue λ_i that expresses a magnitude of the corresponding variance along its direction. Eigenvectors originate from the location of the reference structure that was employed in the construction of the covariance matrix. Thus, the original (centered) trajectory data now can be expressed in terms of the new coordinate system given by the orthogonal eigenvectors. For this reason, the process can be understood as a rotation of the original axes of Cartesian coordinates to the new orientation according to the set of eigenvectors [378], [438].

In trajectory analysis, the valuable feature of this procedure is that the directions of the new system of coordinates, or collective variables, reflect the collective atomic displacements of the molecular structure that best describe the conformational variability that was sampled during the simulation. Exploration of the dynamics along the i-th collective variable is accomplished by projecting the original trajectory data points $\mathbf{x}(t)$ onto the eigenvector $\boldsymbol{r}_i$, which results in what is called a Principal Component (PC), that is, the transformed trajectory data expressed in coordinates $\boldsymbol{p}_i(t)$ [379], [397], [446]:

$$\boldsymbol{p}_i(t) = \sum_{j=1}^{3N}\left(\boldsymbol{r}_{ij}\left(x_j(t) - x_{ref,j}\right)\right) \tag{3.7}$$

PCs can then be visualized as structures $\mathbf{x}'(t)$ by applying a transformation back to the three-dimensional representation (atomic Cartesian coordinates). This operation enables inspection of the captured concerted motions along each PC.

$$\boldsymbol{x}'(t) = \boldsymbol{r}_i \cdot \boldsymbol{p}_i(t) + \boldsymbol{x}_{ref} \tag{3.8}$$

Dimensionality reduction is accomplished by retaining only a subset of PCs that describe the collective motions of larger spatial scales. Since eigenvalues represent the weights of the different collective motions in the total atomic fluctuation, they are sorted in descending order to identify their related eigenvectors. The relevance of PCs is often expressed in terms of percentage of variance, which is calculated as the ratio of each eigenvalue to the sum of all eigenvalues. Normally, in the average protein trajectory, less than 5 PCs already account for 50% of cumulative variance, and 20 PCs are usually more than enough to capture most of the behavior of the system. The selection of the desirable number of PCs may be assessed either by establishing a threshold of cumulative variance, by determining the “elbow” point of the corresponding “scree plot”, or by detecting the PCs with non-Gaussian probability density distributions [379], [446].

Provided that there are at least $3N$ different conformations under analysis, which is advisable [379], the number of eigenvalues (and eigenvectors) is also $3N$, from which 6 should equal zero as their corresponding eigenvectors represent the overall rotation and translation that was removed via structural superposition. With a number of conformations S lower than $3N$, the total number of nonzero eigenvalues is at most S-1 because this is the number of directions required to account for their internal degrees of freedom [237], [292], [334], [391], [446].

On another note, when the analysis involves different atomic species (*e.g.*, protein backbone atoms), every coordinate in Equations 3.5, 3.7, and 3.8 must be weighted with the square root of the atomic mass in order to obtain physically relevant dynamic PCs compliant with the laws of Newton [397]. Finally, the equivalent information can be provided with an alternative mathematical procedure called singular value decomposition. This method is usually chosen when computational cost should be minimized, as it eludes calculation of the covariance matrix and directly takes the $3N$-dimensional trajectory data matrix as input. However, the data should be explicitly centered around the reference structure before applying the procedure [386].

3.5.2 Cosine content

The cosine content is a qualitative indicator for determining whether sufficient sampling time has been achieved to ensure a reliable physical significance of PCs. This measure is calculated for the PCs that bear the higher variance values of atomic fluctuations, for instance, the first five PCs. It was proposed by Hess [397], [438] when he demonstrated that a system of particles undergoing a process of random diffusion (also termed Brownian motion) generates cosine-shaped projections of the first few dominant PCs, with cosine periods equal to half the PC index along the total simulation time window.

MD trajectories with a statistically insufficient number of samples produce the same pattern of cosine-shaped PCs, which is indicative of the simulation being too short so that the captured protein dynamics cannot be distinguished from multidimensional random diffusive behavior. The interpretation of this observation was further clarified by Palese [499], who noted that cosine-shaped PCs in protein dynamics do not exactly reflect pure Brownian motion, albeit the two phenomena share some features. In short timescales, proteins explore a reduced flat portion of the energy landscape, with shallow minima, such that the system does not have time to encounter the more significant kinetic barriers of the true underlying shape of the potential. This barrier-free diffusion is Brownian-mimetic. However, even in very short simulations, coherent protein motions (*e.g.*, α-helix bending or stretching motions) may take place where one would only expect purely random events [378], [386], [398], [446], [499], [500].

Accordingly, the cosine content measures the resemblance of the variation of PC values along the trajectories to the corresponding cosine curves. The cosine content CC_i of the i-th PC $\boldsymbol{p_i}(t)$ is calculated as:

$$CC_i = \frac{2}{T}\left(\int_0^T \cos\left(\frac{i\pi t}{T}\right)\boldsymbol{p_i}(t)dt\right)^2 \left(\int_0^T \boldsymbol{p_i^2}(t)\,dt\right)^{-1} \tag{3.9}$$

where T is the total simulation time. Its value ranges from 0 (no cosine shape) to 1 (perfect cosine shape). A very high cosine content along the first PCs is, thus, an indicator of not sufficient sampling or sampling that can be improved with regard to the atomic fluctuations of interest and in the context of the timescale and local sampling enabled by a classical MD simulation [398].

In practice, this measure cannot be used for quantitative assessment of sampling because it is affected by statistically high uncertainty. For this reason, it is employed just as an indicator for poor sampling when the cosine content is high, but no further interpretation can be drawn from values close to zero [378], [446]. The cutoff value indicative of inadequate sampling is arbitrary. Based on empirical determination, cosine-content values not higher than 0.2 in small peptides and 0.5 in proteins are considered reasonable to justify acceptable sampling [334], [397], [398], [500].

3.5.3 Cosine similarity and distance

The cosine similarity (also known as Orchini similarity, angular similarity, or normalized dot product) is a measure of the similarity between two vectors of an inner product space. It expresses the similarity in the direction or orientation of the vectors, disregarding differences in their magnitude or scale. Mathematically, the cosine similarity S_c is equivalent to the cosine of the angle θ between the vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, which is calculated as:

$$S_c(\boldsymbol{v},\boldsymbol{w}) \coloneqq \cos(\theta) = \frac{\boldsymbol{v}\cdot\boldsymbol{w}}{\|\boldsymbol{v}\|\|\boldsymbol{w}\|} = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2}\cdot\sqrt{\sum_{i=1}^{n} w_i^2}} \tag{3.10}$$

where $\boldsymbol{v}\cdot\boldsymbol{w}$ is the dot product of the vectors, and $\|\boldsymbol{v}\|$ and $\|\boldsymbol{w}\|$ are their respective magnitudes, also known as Euclidean norms or ℓ^2-norms [501], [502].

The cosine similarity is bounded between -1 and 1, according to the range of the cosine function. A value of 0 means that the two vectors are orthogonal (with an angle of 90°) and thus uncorrelated. The closer the value to 1, the smaller the angle, meaning that the vectors are more collinear and correlated (*i.e.*, similar). Values lower than 0 indicate angles greater than 90°, with -1 meaning that the two vectors are collinear but point in opposite directions (they are inversely or negatively correlated; with an angle of 180°) [441], [502].

The cosine similarity may be subtracted from 1 to express the complementary measure known as the cosine distance D_c [501]:

$$D_c(\boldsymbol{v},\boldsymbol{w}) \coloneqq 1 - S_c(\boldsymbol{v},\boldsymbol{w}) \tag{3.11}$$

The cosine similarity and distance are formally considered to be non-metric measures as they do not obey all the formal mathematical properties of metrics. They are commonly used in information retrieval, text document clustering, biological taxonomy, and gene feature mapping. In comparison to other distance measures, the cosine similarity performs particularly well with sparse numerical data, as it only considers the non-zero vector coordinates to provide the measure of proximity [502], [503]. Depending on the application, the cosine similarity may be expressed bounded in the interval [0, 1]. For instance, in this thesis, the cosine similarity was applied to express the similarity between the eigenvectors derived from PCA of equivalent trajectories. Since eigenvectors define the directions of PCs, and PCs reflect collective atomic displacements of the structure, collinear eigenvectors with opposite directions capture the same collective motions albeit reversed. In CEDA, such collective motions are equivalent in dynamical nature. Hence, cosine similarity was expressed in absolute value to disregard the differences due to opposite directions. The resulting measure ranges from 0 (totally dissimilar; orthogonal directions) to 1 (totally similar; either with equivalent or opposite directions). Consequently, the corresponding cosine distance was also expressed in the interval [0, 1].

3.5.4 Agglomerative hierarchical clustering

Clustering is a methodology aimed at grouping data objects based on the similarity between them. These objects are numerical measurements on a set of variables or attributes, and can thus be represented as points or vectors in a multidimensional space. A plethora of different clustering algorithms and techniques have been developed [502], [503].

Hierarchical clustering is a family of connectivity-based algorithms that categorize data points into a hierarchical set of clusters, organized in a tree structure. Agglomerative approaches, in particular, construct the hierarchy in a "bottom-up" manner. Each object is initially considered a single-element cluster (or singleton cluster). In each iteration of the algorithm, the pair of clusters with the highest similarity are combined into a new larger cluster. This process continues until all data points are members of a single cluster, encompassing the entire collection [502], [504], [505].

The clustering results are depicted in a dendrogram, a binary tree diagram that presents the original data objects as leaves and the clusters as inner nodes at the various hierarchy levels. The root of the tree represents the single cluster of the entire collection. The height of the link at which two clusters are first joined is proportional to the distance between them, termed the cophenetic distance. Dendrograms display data objects along one axis and cophenetic distance along the other axis (Figure 3.3). The visualization of a dendrogram facilitates choosing a particular cutting point to yield the most optimal number of separate clusters, depending on the purpose of the application. Among the possible strategies, a given cophenetic distance value may be selected to split the dendrogram and obtain clusters characterized by minimum similarity value among their members [504].

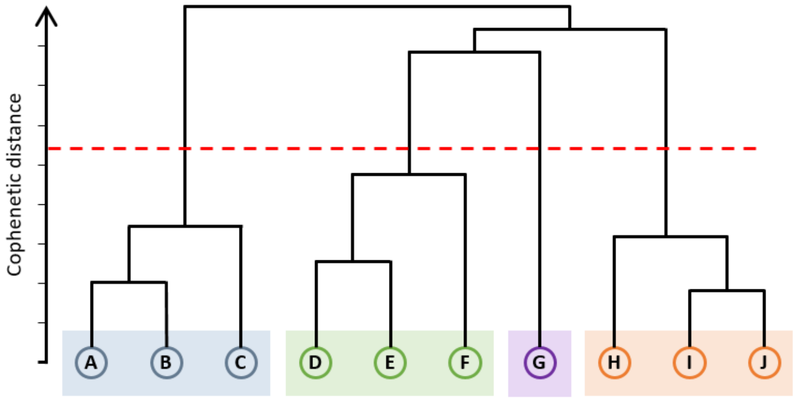


Figure 3.3. Hierarchical clustering represented in a dendrogram. In this illustrative example, there are ten data objects A–J that have been clustered according to a given measure of distance. The red dashed line represents the chosen cophenetic distance value selected to split the dendrogram and obtain four clusters (shown in different colors).

A measure of distance d must be defined to calculate the dissimilarity values between individual data points. Hierarchical clustering methods accept a wide variety of distance measures, as long as they can define a magnitude of dissimilarity and satisfy the mathematical properties of symmetry and positive definiteness [503]. The corresponding dissimilarity matrix between pairwise objects is taken as the input data.

Subsequently, cluster similarities are then inferred as a function of d via a particular linkage criterion. A popular choice, especially in computational biology, is the formulation of the average-linkage algorithm [506], also known as UPGMA (unweighted pair group method with arithmetic mean) or group-average clustering. In UPGMA, the distance between two clusters is defined as the average d of all pairwise data points between the first and the second clusters. On the basis of this definition, at each step of the algorithm, the two clusters with the smallest average-linkage distance d_{UPGMA} are combined [502], [505]. The d_{UPGMA} between two clusters $A = \{a_1, a_2, \ldots, a_N\}$ and $B = \{b_1, b_2, \ldots, b_M\}$, with N and M being their respective number of members (also called the cardinalities), is [501]:

$$d_{UPGMA}(A,B) = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} d(a_n\,, b_m) \tag{3.12}$$

In this thesis, as part of the CEDA strategy, the UPGMA algorithm was applied to perform clustering of eigenvectors and construct the corresponding dendrogram, using the cosine distance (bounded in the interval [0, 1] as derived from cosine similarity expressed in absolute value) between pairwise vectors as input data. The corresponding implementation was carried out with the Python package SciPy [507].

3.5.5 Kernel Density Estimation

A density estimator is an algorithm that models an estimate of the underlying probability distribution that generates a given N-dimensional dataset. Kernel density estimation (KDE), also known as the Parzen-Rosenblatt window [508], [509], is a widely used density estimator which relies in a nonparametric approach, *i.e.*, it does not require the assumption that data is drawn from a known distribution from a parametric family. The nonparametric nature of KDE makes it a very flexible approach to model random or complicated data distributions. The approach of KDE is comparable to that of histograms, but with the advantage of removing the dependence on the location of the sub-intervals or bins along the domain of values.

The conceptual idea behind KDE consists in fitting instances of a kernel function $K(x)$ at the location of each data point. Then, the estimate of the probability distribution is constructed as the sum of all fitted kernels (Figure 3.4). Given a set of observations $\{x_1, x_2, \ldots, x_n\}$ sampled from an unknown distribution with density function p, the estimate $\widehat{p}_n$ at any given point x is defined as:

$$\widehat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.13}$$

The particular kernel function may be chosen from a range of non-negative symmetric functions and, importantly, determines the shape of the final distribution. Frequently, smooth kernel functions are chosen, such as the Gaussian function, which facilitate representation of the estimate with a continuous curve. At regions with many observations, there will be many instances of the kernel function that will yield a large total value of the KDE function. On the other hand, regions with sparse observations will have a lower contribution to the density estimate.

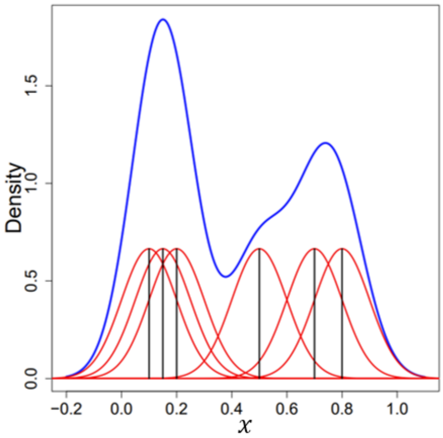


Figure 3.4. An illustration of how KDE is constructed with one-dimensional data. In this example, there are six observations, located at the positions indicated by black lines. An instance of a Gaussian function (kernel) is fitted at the location of each observation (red curves), which are summed to obtain the density estimate (blue curve). NOTE. Adapted from [510].

The parameter h is called the bandwidth or the smoothing parameter. As its name suggests, this value controls the smoothness of the generated curve. Changes in the bandwidth modify the shape of the employed kernel, such that it is wider with higher values of h and narrower with lower values of h. The bandwidth is a free parameter, meaning that its value is not pre-defined by the model and must be chosen empirically in any scenario. Importantly, it has a strong impact on the quality of the results of the estimate by controlling the bias-variance trade-off [510].

When h is too small, the model does not provide much more information than the raw data, even reflecting randomness rather than the true underlying density. Too narrow kernels lead to a high-variance estimate or over-fitting, where the presence or absence of a single point makes a large difference, resulting in a curve that contains too many spurious artifacts (spiky surface). This scenario is called "under-smoothing". In contrast, when h is too large, the model fails to capture important features of the data. Too wide and shallow kernels lead to a high-bias estimate or under-fitting, where the informative variations in density of the underlying structure have been obscured by a large smooth curve that is too unspecific. This scenario is called "over-smoothing". When h is correctly adjusted, the underlying density of the dataset is revealed more clearly. The dependence on the bandwidth in KDE may be compared with the selection of the bin width in a histogram. In nonparametric statistics, bandwidth selection is a classical research topic. Several approaches have been developed to find optimized values of the parameter, although one may adjust it empirically [510]. Figure 3.5 shows the three cases with an example of KDE with different values of the smoothing parameter, performed on a single dataset.

The Gaussian kernel is one of the most common choices of kernel. It is expressed as a Gaussian (or normal) distribution rescaled to have a unit area under the curve, with zero mean and unit variance:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{3.14}$$

In the context of the KDE equation (Equation 3.13), the Gaussian kernel becomes centered around the value of the given observation x_i (*i.e.*, this value becomes the mean of the distribution) with a standard deviation equal to h [511].

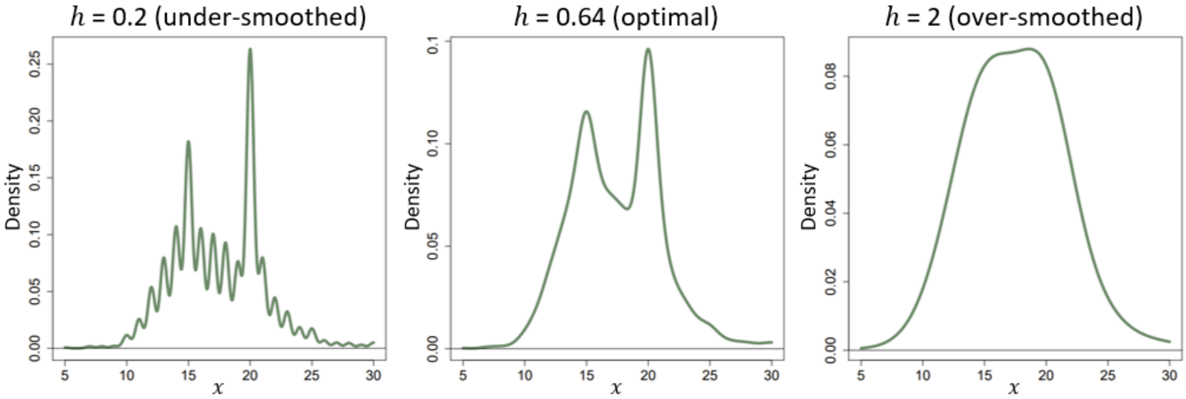


Figure 3.5. Influence of the bandwidth in KDE. The three panels display different KDE performed on the same hypothetical one-dimensional dataset (not shown), using three different values of the smoothing parameter h. The left panel is the case of under-smoothing, with an excessively small bandwidth h. The center panel is the case of the correct value of smoothing, which can be found, for instance, using a bandwidth-selection algorithm. The right panel is the case of over-smoothing: the chosen h is too large. NOTE. Adapted from [510].

KDE can be extended to estimate multivariate densities. In this thesis, KDE was applied to generate the density distributions of trajectory data projected onto the consensus vectors of CEDA. The KDE algorithm was applied through its implementation in the Python package scikit-learn [512], with a Gaussian kernel. A bandwidth value of 0.3 was empirically chosen because it showed the best smoothness trade-off when applied to the datasets of this study. The corresponding KDE curves (univariate data), surfaces (bivariate data), and hypersurfaces (trivariate data) were constructed with 100 points along the domain of each dimension. To cope with performance degradation with high dimensional data, a parallelization scheme was implemented via in-house Python scripts to split the KDE calculation to run in multiple cores.

3.5.6 Bhattacharyya coefficient

The Bhattacharyya coefficient [513] (also known as fidelity similarity or Hellinger affinity) is a measure that quantifies the similarity between two probability distributions. It is also often described as a measure of the degree of overlap between statistical samples or populations. Given two discrete probability distributions P and Q on the same domain with N classes or bins, their Bhattacharyya coefficient BC is calculated as:

$$BC(P,Q) = \sum_{i=1}^{N} \sqrt{p_i q_i} \tag{3.15}$$

where p_i and q_i are the probabilities of occurrence of event or sample i in each respective distribution such that $\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} q_i = 1$. Given such properties, the geometric interpretation of the BC is the cosine of the angle between the N-dimensional vectors $\left(\sqrt{p_1}, \sqrt{p_2}, \dots, \sqrt{p_N}\right)$ and $\left(\sqrt{q_1}, \sqrt{q_2}, \dots, \sqrt{q_N}\right)$ [513], [514].

The Bhattacharyya coefficient ranges from 0 to 1, where 1 indicates maximum similarity (identical distributions) and 0 indicates total divergence (no overlap between the distributions). Figure 3.6 shows the visual representation of the Bhattacharyya coefficient between two examples of probability density distributions. This measure has been suggested as a powerful measure for comparing histogram data [514], and can be applied both to unimodal and multimodal distributions, whether they are Gaussian or not. It is commonly used in statistics, pattern recognition, image processing, document comparison, information retrieval, data analysis with machine learning and, in general, in data science tasks of clustering and classification. Its expression (Equation 3.15) can be generalized to calculate the similarity between multivariate distributions [515]. In this thesis, it was used to express the similarity between the density distributions of trajectory data projected onto the consensus vectors of CEDA.

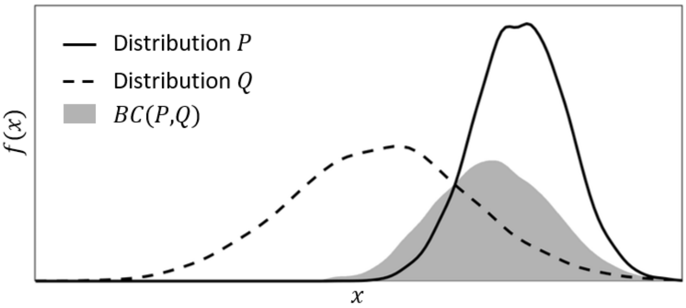


Figure 3.6. Visual representation of the Bhattacharyya coefficient. The dashed and solid lines are examples of probability density functions P and Q from Equation 3.15. The corresponding visual representation of the Bhattacharyya coefficient is shown as a curve filled in gray. NOTE. Adapted from [511].

3.6 Repositories of protein variants

This section elaborates on the set of actions that were conducted to collect a dataset of known missense variants of PKR. The landscape of the repositories of human genetic variants is continuously evolving; there is an increasing volume of initiatives that publish new data or that gather and harmonize the information from multiple databases to display it in single data hubs. The search actions conducted in this thesis were performed between the end of 2019 and the beginning of 2020. Therefore, the available methods to retrieve this kind of data may have been optimized since then, and the databases may now contain a higher volume of downloadable content.

On the one hand, most of the queried repositories comprised databases of human protein variants that include annotations and interpretations of their clinical significance. On the other hand, queries were also made on data portals of large-scale sequencing projects. From these sources, only genetic variants causing a single amino-acid variant (SAV; equivalent expression for a missense variant) were retrieved, whether from single nucleotide variants (SNVs) or small in-frame indels. Data was filtered to display only the variants affecting gene *PKLR*, and more specifically, the transcript of PKR (or, equivalently, the UniProt ID P30613). The collected data principally comprised: i) the associated amino-acid substitution (original and alternative amino acids and the affected position in the protein), ii) the corresponding change in the cDNA (when available), and iii) the enclosed literature references.

The collected information was integrated in an internal database as a catalog of the known missense variants of PKR from diverse sources. Table 3.5 lists the employed repositories, including a brief description of their goals and the nature of their contents, as well as the specific filters that were applied in the queries and the additional data fields retrieved.

Table 3.5
Repositories of human genetic/protein variants queried in this project

Name	Description	Retrieved data (missense variants)	References
SwissVar	A portal of a comprehensive collection of SNPs and diseases, including annotations on the underlying protein products, the molecular details of each variant, the genotype-phenotype relationship of each specific variant based on literature, as well as pre-computed information (*e.g.*, conservation scores, structural features) to help assess the effect of the variant. The database has been discontinued since August 12th, 2020; the data is now available as part of the UniProt Knowledgebase (UniProtKB).	Filtered by: - Gene: *PKLR* - Disease: “pyruvate kinase deficiency of red cells”	[207], [208]
Humsavar	A downloadable text file with all missense variants annotated in UniProtKB/Swiss-Prot human entries. It provides a variant classification which is intended for research purposes only, not for clinical and diagnostic use. Variants used to be classified into the categories “Disease”, “Polymorphism” and “Unclassified”, based on the variant annotation curated from literature reports. Labels may change over time and must not be considered as a definitive statement about the pathogenic role of a variant. In the current version, these categories were renamed to follow the terminology recommended by the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) [205]: “LP/P” (likely pathogenic or pathogenic), “LB/B” (likely benign or benign), “US” (uncertain significance).	Filtered by: - Gene: *PKLR* - Disease: “pyruvate kinase deficiency of red cells” / none.	[208]
ClinVar	A public archive of reports of the relationships among human genomic variations and phenotypes, with supporting evidence. It processes submissions reporting variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data. Assertion of clinical significance comprises several methods, including clinical testing, research, and reports from the literature. It does not include uncurated sets of data from GWAS studies, although variants that were identified through GWAS and have been individually curated to provide an interpretation of clinical significance are in scope.	Filtered by: - Gene: *PKLR* (PKR transcript) - Condition: “pyruvate kinase deficiency of red cells” / “Not provided” / “Not specified”. Additional data collected: - Clinical significance (last reviewed report)	[209]
Leiden Open Variation Database (LOVD)	A web-based open source database designed to collect and display genomic variants, together with annotations about potential causal connections with diseases. Submissions consist of sequence variants found in individuals, derived from clinical sequencing studies. Entries are subsequently subject to data curation by experts. LOVD is used by clinical geneticists to store and share information, as well as to diagnose and provide advice to patients carrying a genetic disease.	Filtered by: - Gene: *PKLR* Additional data collected: - Clinical classification	[210]
Human Gene Mutation Database (HGMD)	An initiative to collate known (published) gene lesions responsible for human inherited disease. Entries comprise various types of mutation (coding regions, splicing and regulatory regions of human nuclear genes) associated with human inherited disease, plus disease-associated/functional polymorphisms reported in the literature. Somatic mutations and mutations in the mitochondrial genome are excluded. Mutations lacking obvious phenotypic consequences are not usually included. HGMD provides information of practical diagnostic importance to: i) researchers and diagnosticians in human molecular genetics, ii) physicians interested in a particular inherited condition in a given patient or family, and iii) genetic counselors.	Filtered by: - Gene: *PKLR* (PKR transcript) - Phenotype: “pyruvate kinase deficiency” / “Haemolytic anaemia”	[203]

Table 3.5 (Continued)

Exome Aggregation Consortium (ExAC)	A coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community. The ExAC launched a browser framework that enabled display of exome sequence data from 60 706 unrelated individuals. The browser was intended for use by clinical geneticists researching variants of interest for patients as well as biologists exploring variation in specific genes. Nowadays, the browser is no longer available and the dataset has been transferred to the gnomAD initiative.	Filtered by: - Gene: *PKLR* (PKR transcript) - All ExAC quality filters passed Additional data collected: - Allele frequency - Number of homozygotes	[188]
Genome Aggregation Database (gnomAD)	A resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community. In its first release, which contained exclusively exome data, it was known as ExAC. The project comprises two datasets, namely, gnomAD v2 and v3. They contain largely non-overlapping samples and both datasets must be used to capture the full set of variation across gnomAD. The v2 set contains fewer whole genomes than v3, but also contains a very large number of exomes that substantially increase its power as a reference in coding regions. On the other hand, the v3 set represents a very large increase in the number of genomes and contains richer ethnic representation. Both sets were considered in this thesis.	Filtered by: - Gene: *PKLR* (PKR transcript) - All gnomAD quality filters passed Additional data collected: - Allele frequency - Number of homozygotes	[189]
Single Nucleotide Polymorphism database (dbSNP)	A public-domain archive for a broad collection of simple genetic polymorphisms. Includes human single nucleotide variations, microsatellites, and small-scale insertions and deletions. Each entry includes the sequence context of the variant, the occurrence frequency (by population or individual), the molecular consequence, the experimental method(s), protocols, and conditions used to assay the variation, and genomic and RefSeq mapping information for both common variations and clinical mutations. The data was retrieved via the Ensembl Variation database, which is an infrastructure that integrates genetic variation data from several other sources and provides a set of interconnected components for data visualization and analysis.	Filtered by: - Gene: *PKLR* (PKR transcript)	[193]- [195]
Catalogue Of Somatic Mutations In Cancer (COSMIC)	A comprehensive resource for exploring the impact of somatic mutations in human cancer. The COSMIC database gathers data from individual studies of mutations in known cancers (from the literature) and from large-scale genome screening studies of cancer samples, such as The Cancer Genome Atlas (TCGA) [197] and the International Cancer Genome Consortium (ICGC) [198]. The contents of the database undergo manual curation by experts. The goal of the project is to document and catalogue somatic mutations that are implicated in the development of cancer.	Filtered by: - Gene: *PKLR* Additional data collected: - Somatic status	[516]
BioMuta	A database of cancer-associated non-synonymous single-nucleotide variations. The project aims to aggregate data across many studies into a single source with a unified representation and annotation of functional attributes. Genetic variations are mapped to genomes and RefSeq nucleotide entries, and unified through UniProtKB/Swiss-Prot positional coordinates. Mutation data is retrieved from other sources such as COSMIC, TCGA, and ICGC. The database also features a small number of variants added following manual literature review of PubMed abstracts.	Filtered by: - UniProt ID: P30613	[517]

On the basis of the gathered data, the search was complemented with a comprehensive exploration of the available literature related to the reported clinical and experimental conditions of each variant. The following features were included as annotations of the internal database: phenotype and pathogenicity state, genotype (zygosity) of the patients, ethnic origin, enzymatic activity, kinetic parameters, thermostability, provided rationale of structural abnormalities and other evidence of function impairment. Furthermore, Dr. Richard van Wijk (Department of Clinical Chemistry and Haematology, University Medical Center Utrecht) kindly provided further variant annotations from personal clinical research.

Finally, the published data from the “pyruvate kinase” challenge of the fourth edition of the Critical Assessment of Genome Interpretation (CAGI) [518] was also incorporated. CAGI is a global community experiment to objectively assess computational methods for predicting phenotypic impacts of genomic variation. CAGI proposes challenges whereby participants are provided genetic variants and make predictions of the resulting phenotypes. Subsequently, these predictions are evaluated against gold-standard experimental or clinical data by independent assessors, and the ranking of the best predictions submitted to the challenge is provided along with the answer key. The “pyruvate kinase” challenge [519] revolved around the prediction of the effects of missense mutations in human liver pyruvate kinase (PKL) on its activity and allosteric regulation. As recombinant protein expression of PKL in *Escherichia coli* has more rate of success than that of PKR, the activity and the regulatory properties of the enzyme are typically studied with this isoform. However, mutation effects are assumed to be equivalent in PKR, as both isoforms share virtually the same kinetic properties despite PKR having 31 extra N-terminal amino acids.

The published data of this CAGI challenge contains experimental kinetic values related to 113 different amino-acid replacements at 9 different positions of PKL, mostly near the active site, as well as all possible amino-acid replacements with alanine at 430 positions of PKL. The kinetic values comprise: i) the enzymatic activity in binary values, with 0 for no detected activity or 1 for detected activity; ii) the allosteric coupling constant (Q_{ax}) for the negative effector alanine, with values between 0 (total negative allosteric effect) and 1 (no allosteric effect); and iii) the Q_{ax} for the positive effector FBP, with the value of 1 indicating no allosteric effect and values greater than 1 indicating a positive allosteric effect.

Chapter 4 Results

The results of this thesis have been divided into two main blocks. The first block presents the study of the dynamical traits of human erythrocyte pyruvate kinase (PKR) using molecular dynamics (MD) trajectories. The collection of trajectories covers different biological conditions of the wild-type (WT) protein complexed with its natural ligands. In the initial part of the chapter, the modeling strategy and the performance of the developed parameters of the simulated systems are evaluated. This is followed by the application of traditional techniques of trajectory analysis aimed at assessing the principal features of stability and flexibility of the structures along the trajectories.

The final part of the first block of results constitutes a central part of the thesis, related to the design and implementation of the Consensus Essential Dynamics Analysis (CEDA), a method devised to identify the distinctive collective motions of the protein from the most representative (consensus) behavior of the ensemble of trajectories. The corresponding sections report the most significant findings about the conformational changes of key regions of the enzyme, manifested in the trajectories, along with their possible implication in enzymatic function and allosteric communication mechanisms.

The second block revolves around a set of comparative analyses of missense variants of PKR with respect to the WT behavior. In the corresponding sections, the process for selecting variants for subsequent MD simulation is described. This is followed by a report on the results derived from the application of the framework of CEDA to identify signs of differential behavior in the trajectories of PKR variants. The potential dynamical alterations are reported in terms of the similarity with the conformational profiles of the WT enzyme. The analyses conducted in this segment represent the culmination of the study based on the formulated objectives.

Finally, this chapter concludes with a summary of the final protocol of the CEDA strategy, which provides a global view of the methodology developed and proposed in this research project, and after which this thesis is entitled.

4.1 Study of the WT protein

The study of PKR in its WT form comprised both the apoenzyme and also several of the possible holoenzyme states according to different combinations of the ligands that bind to the active and allosteric sites. The components included in each holo condition studied in this project can be found in detail in section 3.1.2 from the Methods chapter. Nonetheless, an overview of this information (cofactors, substrates, and allosteric regulator) is provided here in Table 4.1 to keep it on hand for the present chapter.

Table 4.1

Components of the PKR enzymatic complex per condition

Condition	Cofactors		Substrates		Allosteric effector
	K^+	Mg^{2+}	PEP	MgADP	FBP
Apo					
K-holo	X				
K-Mg-holo	X	X			
PEP-holo	X	X	X		
ADP-holo	X			X	
PEP-ADP-holo	X	X	X	X	
FBP-holo	X	X			X
Full-holo	X	X	X	X	X

NOTE. Abbreviations: ADP, adenosine diphosphate; FBP, fructose 1,6-bisphosphate; PEP, phosphoenolpyruvate.

4.1.1 Evaluation of the parameterization procedures

This section presents the series of actions that were applied with regards to the evaluation of the parameterization of the ligand molecules and the metal centers required to model the holo conditions of PKR in MD simulation. It is worth noting that the holo conditions that only differ in the absence or presence of FBP in the allosteric site of the protein did not require an individual parameterization. Accordingly, the FBP-holo and Full-holo systems reused the metal-center parameters from the K-Mg-holo and ADP-PEP-holo systems, respectively. The results of the parameterization procedure are presented mainly in a descriptive and qualitative manner, with the support of a series of figures that facilitate the tracing of the main points of the process. The structure and topology files of all the parameterized systems are available as part of the online supplementary material of this thesis (see Appendix A).

4.1.1.1 Ligands

The first operation regarding the parameterization stage of the project consisted in the determination of the internal bonded parameters and atomic charges for the ligands PEP and FBP. The set of bonded parameters for each ligand were extracted from the GAFF force field; the program ACPYPE was used to build the corresponding topology files. Atomic charges were determined with the RESP methodology after a geometry optimization at the QM level. The geometry optimization stage for both ligands was straightforward, reaching convergence in 21 steps in the case of PEP and 24 steps in the case of FBP. Figure 4.1 shows the chemical structure of both ligands together with the obtained atomic-charge values. The RESP methodology allowed the assignment of equivalent charges in the cases of rotationally degenerate atoms, which is particularly important when it comes to modeling the resonant forms of the phosphate and carboxylic groups.

Given that PEP is one of the coordination ligands of the cofactor Mg^{2+}, these atomic charges were only used in the first minimization in vacuum of the system. Afterwards, new atomic charges for the whole coordination complex were derived at the metal-center parameterization stage, replacing these ones. The same occurs with the ligand ADP, except that the original parameters of this molecule were not generated in this project but extracted from [468]. Conversely, the atomic charges of FBP were retained for the whole MD simulation workflow, given that this ligand does not belong to any metal center.

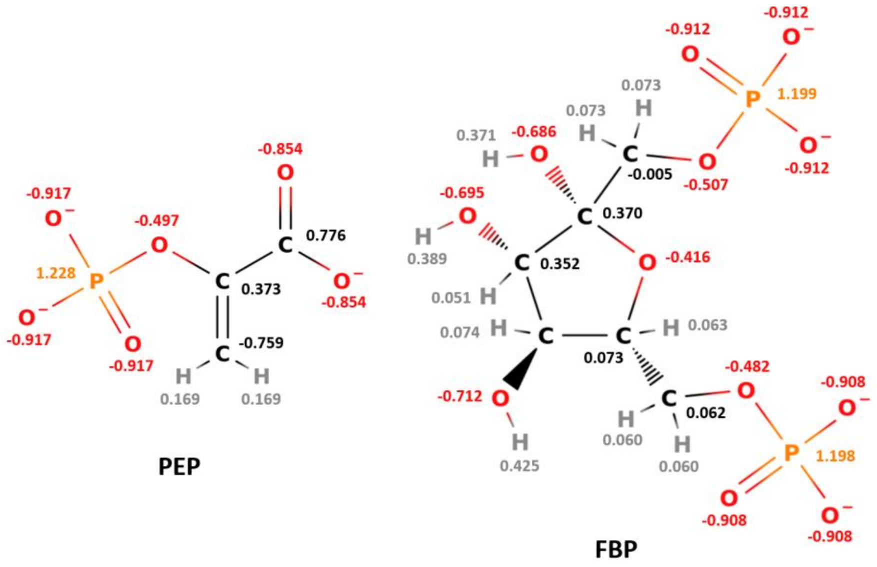


Figure 4.1. QM-derived atomic charges of ligands PEP and FBP. Atom labels, bonds, and charges are colored by atomic species. NOTE. The images were generated with the software MolView (http://molview.org).

4.1.1.2 Metal centers

The incorporation of metal ligands in the holo systems required deriving a set of parameters that can properly model the involved coordination complexes, as well as maintain their binding geometry stable during MD simulations. The parameterization strategy was mainly built on the guidelines and protocols of the software MCPB.py, which employs a cluster model to represent the metals and the coordination ligands [471], [472]. From there, a number of modifications were introduced to take into account further modeling standards and good practices as found in the literature for similar metalloproteins. The process resulted in a thorough empirical procedure aimed to determine the set of decisions and settings (mostly concerning the QM calculations) that are required to satisfactorily mimic the chemical environment of the metal centers of PKR. An overview of the employed methodology can be found in the Methods chapter. The present section therefore focuses on the specifications that help understand how the protocol was designed, what it pursues and what has been accomplished.

To assemble the initial structure of each holo condition, the chosen model of the PKR was the PDB structure 2VGB [121], which is also the protein structure employed in the MD simulations in the apo condition. According to the contents of each holo condition, the needed ligands were imported from other PDB structures of pyruvate kinases co-crystallized with such molecules (structures 4HYW, 4HYV, and 4FXF). Combining these sources is possible thanks to the consistent similarities in the spatial arrangement of the active site components among the pyruvate kinase family, which are shown in detail in Figure 4.2. The figure also highlights which molecules have been imported from each of the complementary structures. This information coincides with that provided in Table 3.1 of the Methods chapter and Table 4.1 of the present chapter.

The only notable difference with regard to the spatial arrangement of the metal centers corresponds to the case of the cofactor Mg^{2+} imported from the structure 4HYW (Figure 4.2c). This Mg^{2+} binding site, called the Mg-3 site, has been found to be occupied when the substrate PEP is unbound and would potentially work as a priming mechanism to attract the next PEP molecule and maintain the enzyme in the active R-state conformation [137]. Conversely, the other canonically described Mg^{2+}

binding site, called the Mg-1 site, can be seen in 2VGB since the substrate analog PGA is bound. The Mg-1 and Mg-3 sites involve the same amino-acid chains, namely Glu315 and Asp339 (in PKR numbering), but in altered orientations. The side chain of Phe287 also rotates in correlation with the positions of Mg^{2+}, Glu315, and Asp339. The side-chain orientations that are characteristic of the Mg-3 site have been obtained by generating rotamers in 2VGB and selecting the suitable ones. In the Mg-3 site, three water molecules complete the positions of the coordination sphere of Mg^{2+} left vacant by the unbound substrate. In this project, the Mg-3 site has been chosen to model conditions K-Mg-holo and FBP-holo, while the Mg-1 site is used in conditions PEP-holo, PEP-ADP-holo and Full-holo.

In connection with the above, it is also interesting to note the differences and similarities in the protein conformation depending on which ligands were included in the crystallization. The structural overlays shown in Figure 4.2 were obtained by a structural superposition (least-squares fitting) using as a fitting group the backbone atoms of the amino acids that were used to build the cluster model and that are located in the A domain of the protein. These are depicted with a licorice representation in the images and can also be found listed in Table 3.3 of the Methods chapter with the roles "ligand of K^+/Mg^{2+}" and "chemical context". The structural superpositions show that there is almost an exact match between the A domains of the structures and that the C domains adopt an overall equivalent conformation. Conversely, the B domains show different degrees of closure upon substrate binding that have been described in the literature [119], [124], [135], [137], [139], [147]: open in the absence of substrate (structure 4HYW), partially closed when PEP is bound (structures 2VGB and 4HYV), and fully closed when ADP/ATP is bound.

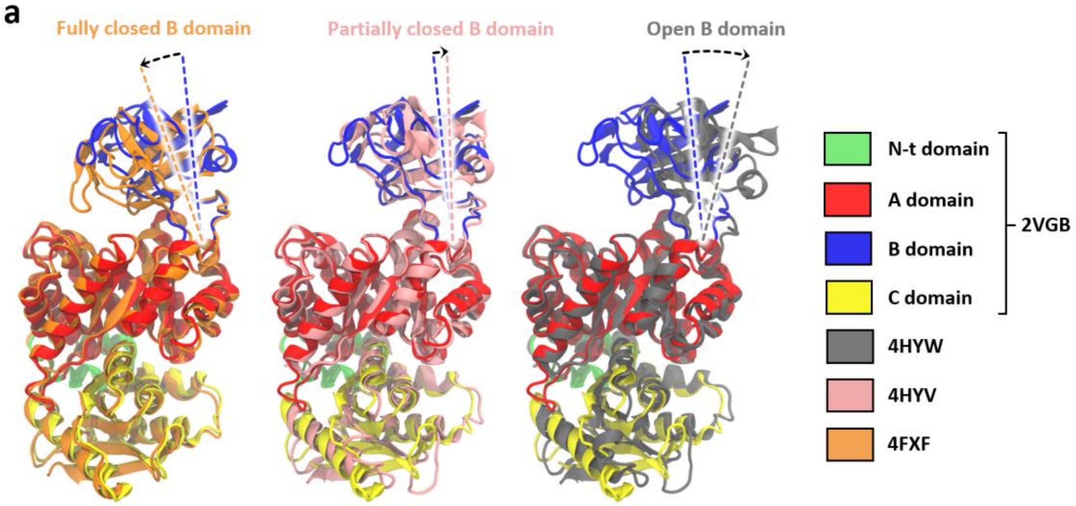


Figure 4.2. The PDB structures employed in the modeling of the holoenzyme conditions of PKR. A structural superposition between the main structure 2VGB (subunit A) and the structures 4HYW (subunit B), 4HYV (subunit A) and 4FXF (subunit D) is shown. (**a**) Pairwise comparison of the overall conformation of the monomers. The different degrees of closure of the B domains are highlighted with illustrative dashed lines and arrows. (**b, c, d**) Close-up of the metal centers of the pairwise superposed structures. The ligands that were imported to model the holoenzyme systems of PKR are marked with a black pointing-finger symbol and highlighted with black thick edges. The backbones of the structures are depicted with a ribbon representation. The amino acids and ligands employed in the cluster model are depicted with licorice or spherical representations. Structure 2VGB is colored in cyan and by atomic species, while the rest of structures are colored as in (a). Coordination bonds are shown as black thick dashed lines, only in structures 4HYW, 4HYV and 4FXF. Labels are colored in black when they refer to a residue or region that is present in both structures; otherwise they are colored with the corresponding color of the structure. The numbering of amino acids corresponds to 2VGB. In (b), the Mg-1 and Mg-3 sites are encircled with black dashed lines. NOTE. Abbreviations: N-t, N-terminal; OXL, oxalate; PGA, phosphoglycolate; Wat, water. The images of the protein structure were generated with the software VMD.

Figure 4.2 (Continued)

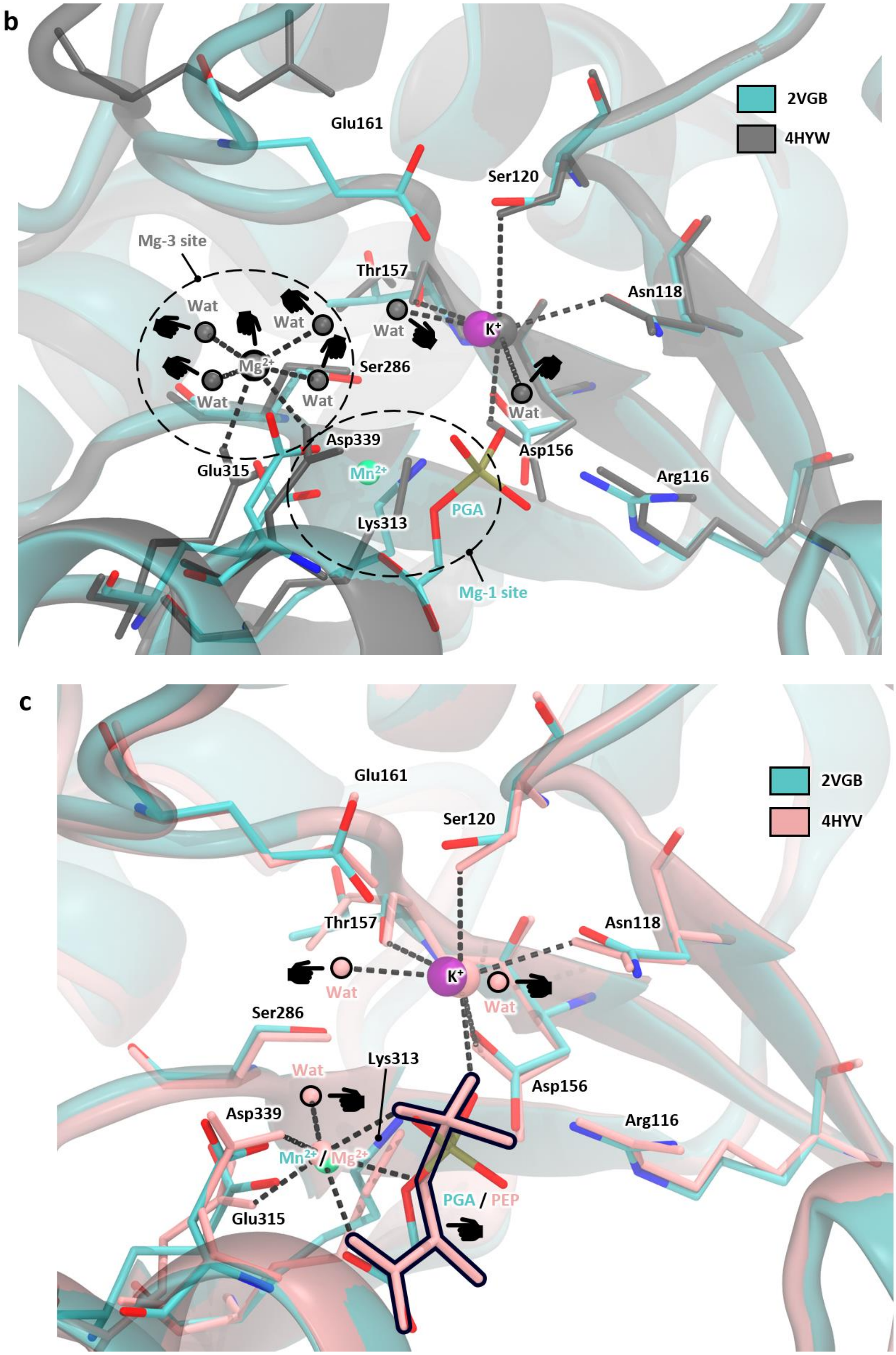
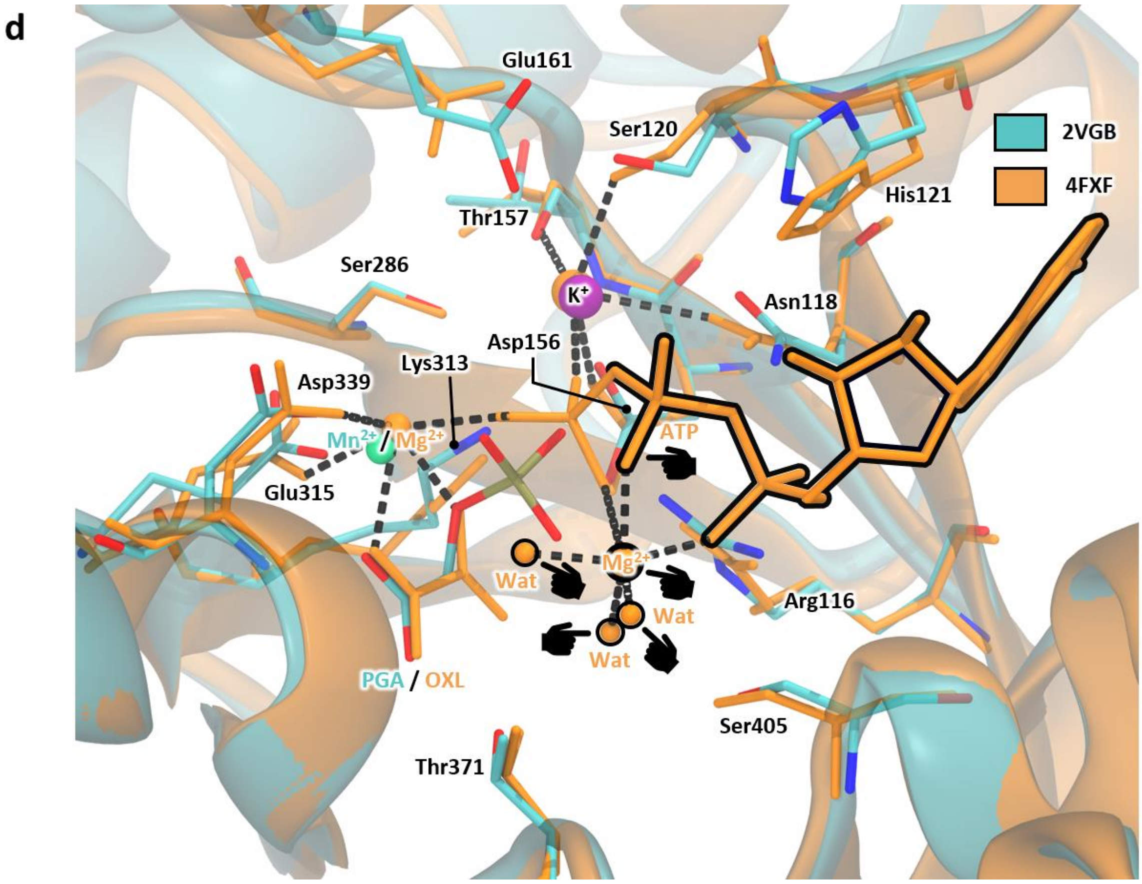


Figure 4.2 (Continued)



The high degree of structural conservation of the active site in the pyruvate kinase family as well as the availability of crystallographic structures with ligands have provided valuable criteria to guide and assess the parameterization of the metal centers. Thus, it has been possible to monitor the procedure to ensure that the experimental binding geometries are qualitatively well reproduced. In this sense, the QM geometry optimizations of the cluster model, which were carried out prior to the calculation of the bond and angle parameters and atomic charges, were satisfactory. Several attempts were needed until finding an optimal combination of components of the cluster model and simulation settings. Here, the details of the selected configurations are provided.

The primary components of the cluster model are the metals and the ligands of their first coordination spheres. By exploring the environment of the coordination complexes, more residues were rationally included in each cluster model. These help specify a chemical context where steric hindrances and relevant interactions with the primary components are taken into account, both in structural and energetic terms. Figure 4.3 (subfigures a, d, g, j, and m) shows 2D schemes of the components and their interactions in each cluster model built, where the rationally included residues are depicted with a faded representation to distinguish them from the coordination complexes.

The delimitation of a cluster model implies that there will always exist a number of interactions that are not represented and occur at the boundary between the included and the excluded components, thus affecting the freedom of movement of the former. Some of these can be taken into account by providing a scheme of frozen atoms that will not be able to move during the simulation. These are marked in Figure 4.3 with an “F” or with a black sphere representation. Firstly, all amino-acid backbone atoms were frozen so that the cluster model would not decouple from the overall conformation of the protein. Therefore, *a priori* only the amino-acid side chains and the non-protein small molecules were free to move. According to the protocol of MCPB.py, a full geometry optimization is applied only to the "small model" of the cluster model, whereby the majority of amino acids are represented in the form CH_3-R (with R corresponding to the side-chain group). Thus, all these methyl groups were frozen to meet the constraints of the protein backbone. In addition, a few other atoms of the model were also frozen to represent further interactions and regions across the boundary:

- The hydrogen atom HH11 of Arg116 was frozen because it makes a H-bond beyond the boundary with the hydroxyl group of Thr93.
- The non-coordinated oxygen atom of the carboxylate of Glu315 was frozen because it makes a H-bond with the hydrogen atom on the backbone nitrogen atom of Phe287. This only applies to the Mg-1 site models.
- Both oxygen atoms of the carboxylate of PEP were frozen because they make H-bonds with the hydrogen atoms on the backbone nitrogen atoms of both Gly338 and Asp339.
- All atoms that comprise the rings of adenine and ribose in the ADP molecule were frozen because this region of the molecule is inserted in a pocket that lies beyond the cluster model.
- All heavy atoms of Ser286 were frozen. This residue was included in the model to account for the steric hindrance that prevents Lys313 from approaching Asp156. Furthermore, it maintains the network of H-bonds between itself and Lys313 and Thr157.
- All heavy atoms of Glu161 were frozen. This residue is located in the hinge of the B domain of the protein, being around the metal center of K^+ only when the B domain is its partially or fully closed conformation; otherwise, it moves away (see, for instance, structure 4HYW in Figure 4.2c). When present, the carboxylate of Glu161 potentially makes H-bonds with the hydroxyl group of Ser120 and one water molecule coordinated to K^+. When absent, its role can be fulfilled by the water medium. Glu161 was included in the cluster model for consistency with the conformation of structure 2VGB and to take into account its electrostatic effects without the need to optimize its geometry in particular.

Figure 4.3 (subfigures b, e, h, k, and n) shows the initial and final states of each geometry optimization of the "small model". The simulations involved moderate rearrangements. Table 4.2 reports the corresponding RMSD values as well as the number of steps that each geometry optimization needed to reach convergence. The inclusion of a polarizable continuum model (PCM) improved the physicochemical properties of the model by accounting for the electrostatic influence of its surroundings (*i.e.*, a protein environment with partial exposure to water). Without the PCM, residues such as Asn118 and Asp156 that are coordination ligands of K^+ and are also at the boundary of the cluster model underwent reorientations and failed to properly bond to the metal.

The widest movements usually corresponded to the water molecules, since their initial placement and orientation were the most approximate of the ensembles. The inclusion of the water molecules was decisive to complete and give consistency to the coordination complexes. Indeed, the QM simulations effectively recognized the majority of the included water molecules as coordination ligands. Only in the case of K-holo, one of the water molecules that was intended to coordinate K^+ moved away to interact with Asn118, Ser120 and another water molecule (Figure 4.3a-c), leaving K^+ with just 5 coordination ligands. Interestingly, as it will be shown in section 4.1.2.3.3, the K^+ of K-holo is also mostly found in a 5-coordinate state in MD simulations, closely followed by the 6-coordinate state.

Therefore, this metal center in this condition seems to be stable with either 5 or 6 coordination ligands. Moreover, this displaced water molecule ended up fulfilling a role analogous to that of a structural water molecule in PEP-holo (Figures 4.2c and 4.3g-i). This water molecule in PEP-holo makes H-bonds with Asn118, Ser120, and the phosphate group of PEP, and was decisive to stabilize the geometry optimization of this metal center, specifically the orientations of Asn118 and PEP. All in all, strong consistencies have been found between the experimental geometries and the QM simulations.

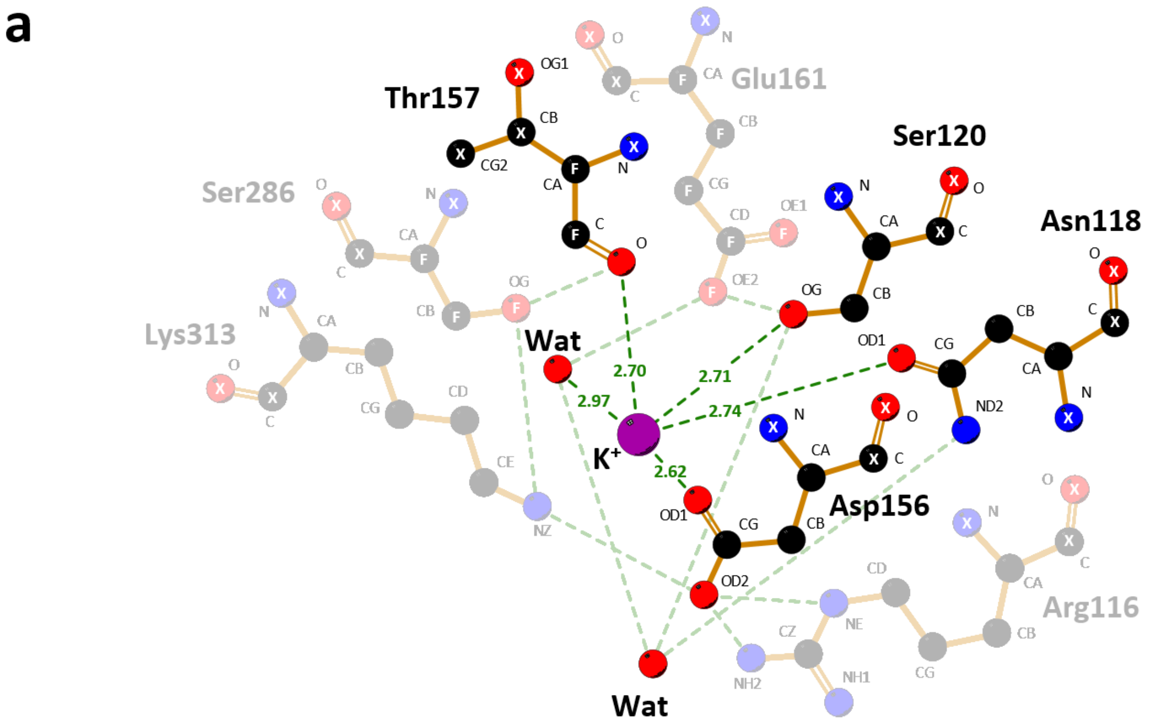


Figure 4.3. Schematics and results of the geometry optimizations of the cluster models. The purpose of this figure is to show detailed representations of the residues included in each cluster model and their interactions, as well as the course of the geometry optimizations performed on their MCPB.py "small model" versions. Five cluster models were employed to model the metal centers of the holo conditions of PKR: one for condition K-holo (a-c), one for conditions K-Mg-holo and FBP-holo (d-f), one for condition PEP-holo (g-i), one for condition ADP-holo (j-l), and one for conditions PEP-ADP-holo and Full-holo (m-o). (**a, d, g, j, m**) 2D schemes of the components of each cluster model and their interactions. Two kinds of molecules comprise each cluster model: the primary components are the metals and the ligands of their first coordination spheres, whereas the rest were rationally included to represent the relevant chemical context of the vicinity. To distinguish between the two kinds, the latter are depicted with a faded representation. Atoms are colored by species. Non-bonded interactions after geometry optimization are depicted with green dashed lines. The distance values (in Å) of the coordination bonds are included. Labels of each residue and atom name are included. All hydrogen atoms have been omitted for the sake of clarity. The atoms marked with an "X" were absent or replaced in the "small model" according to the capping scheme of MCPB.py. The atoms marked with an "F" were frozen during the geometry optimizations. (**b, e, h, k, n**) Initial and final states of each geometry optimization of the "small model". The former are shown in a faded representation. Atoms are colored by species. Coordination bonds in the final state are depicted with black dashed lines. A black sphere representation is included to show the atoms that were frozen during the geometry optimizations. Labels of each residue are included. (**c, f, i, l, o**) 3D structures of the cluster models in their structural context within PKR after geometry optimization of the "small model". Atoms are colored by species. The backbone of the protein is shown in ribbon representation, colored according to the domains of the protein: A domain in red, B domain in blue, and C domain in yellow. Coordination bonds are depicted with black dashed lines, with their distance values (in Å). Labels of each residue are included. NOTE. Abbreviations: Wat, water. The images were generated with the softwares LigPlot+ (v2.2.5) [520] and VMD.

Figure 4.3 (Continued)

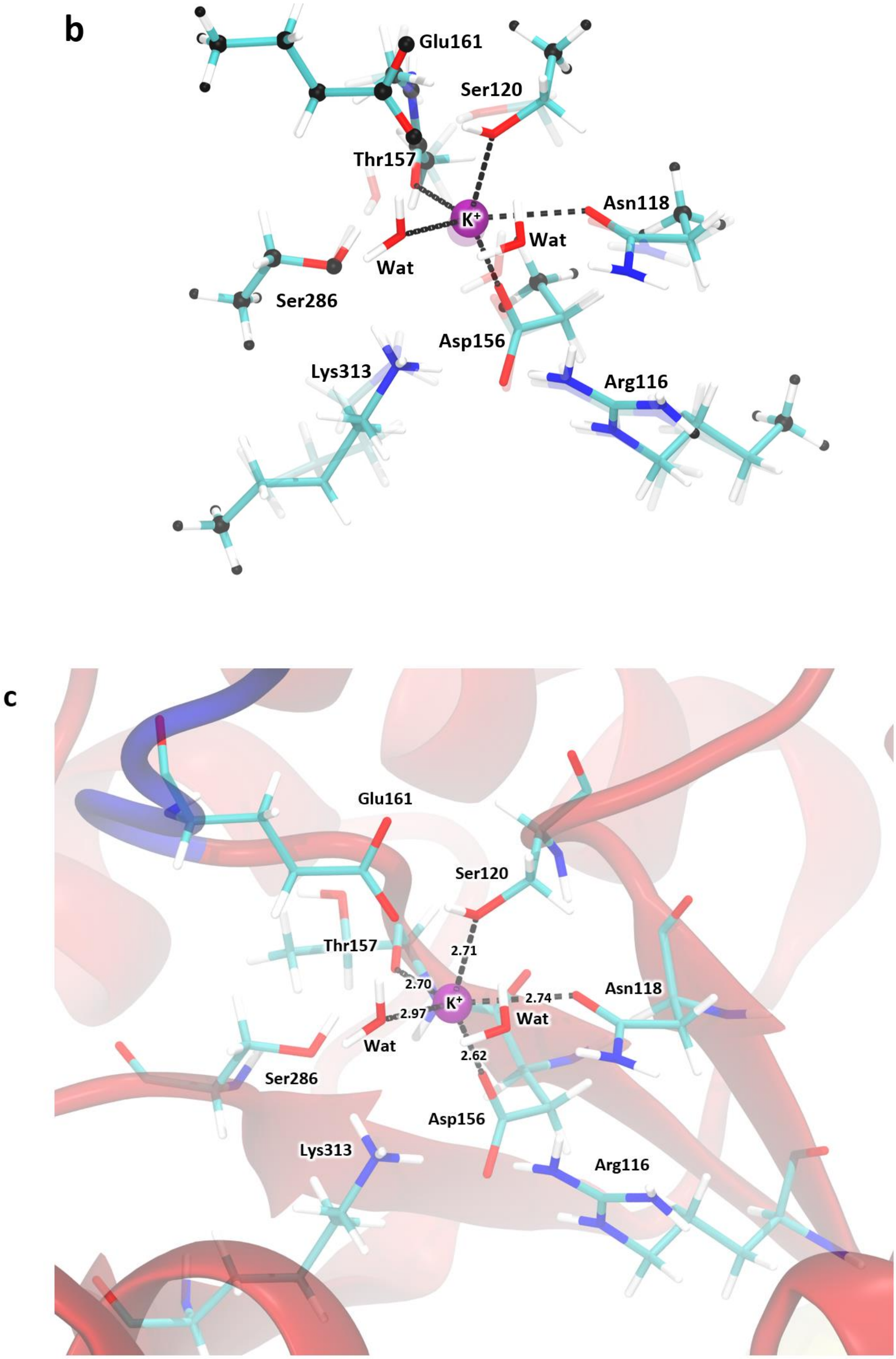


Figure 4.3 (Continued)

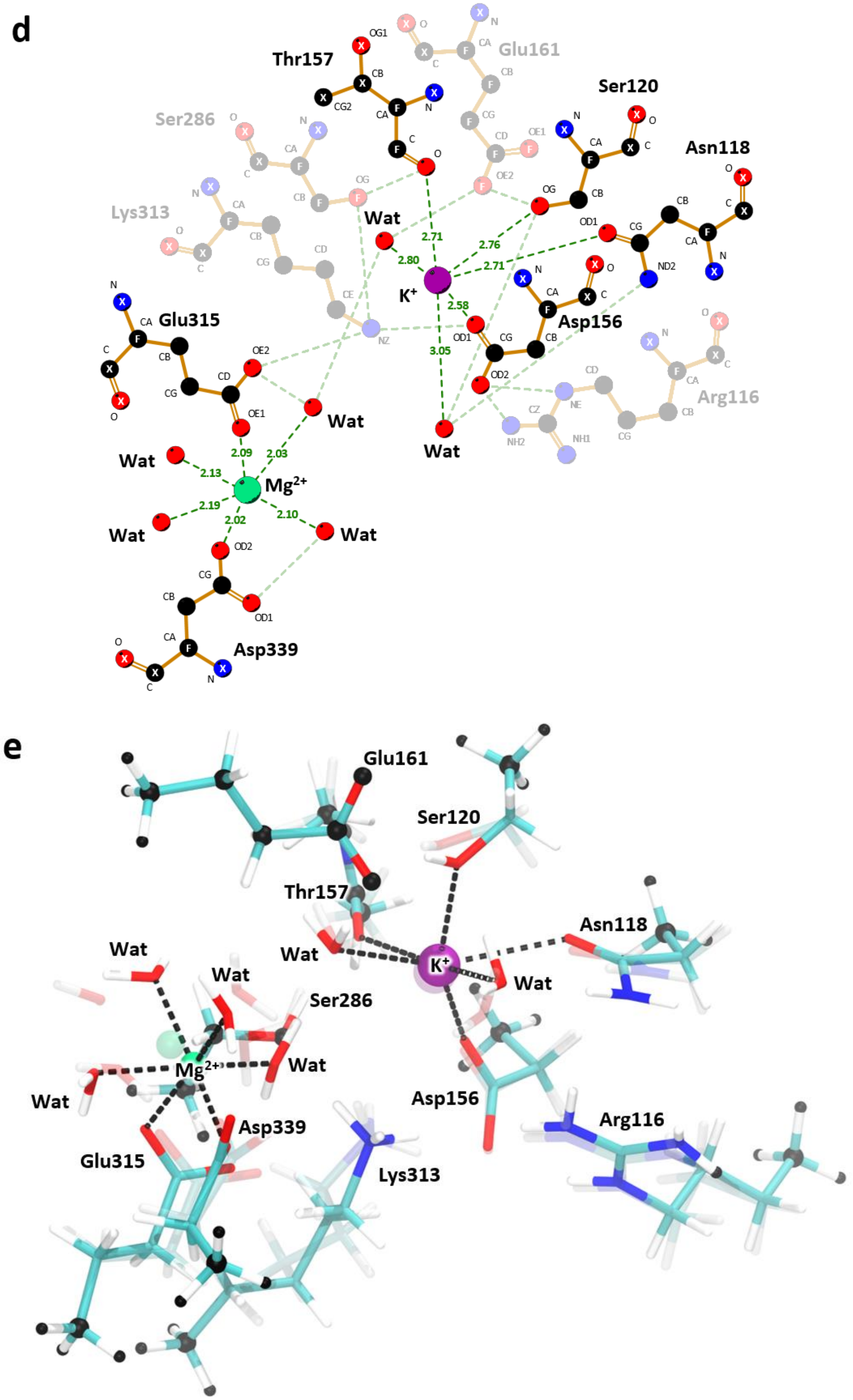


Figure 4.3 (Continued)

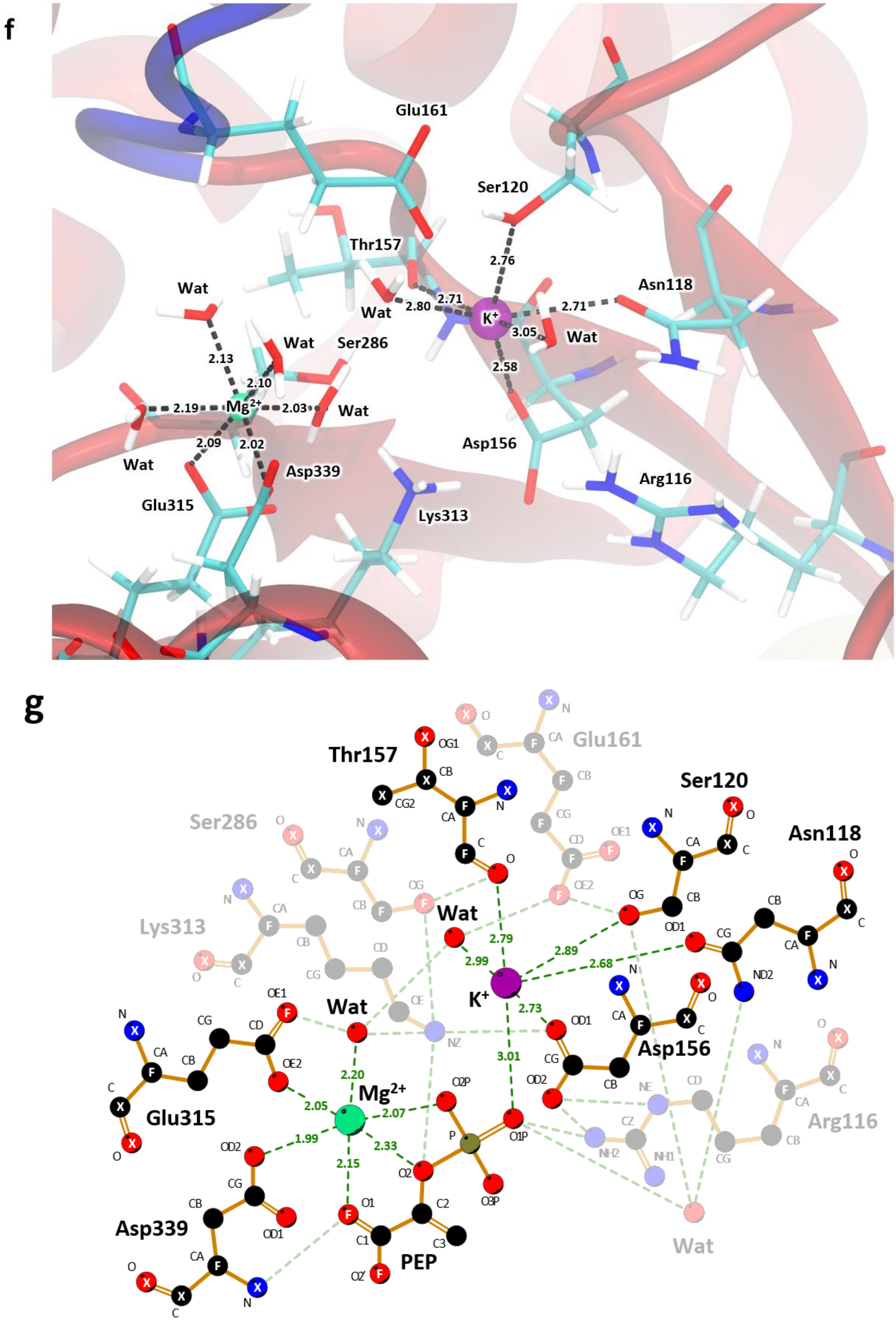


Figure 4.3 (Continued)

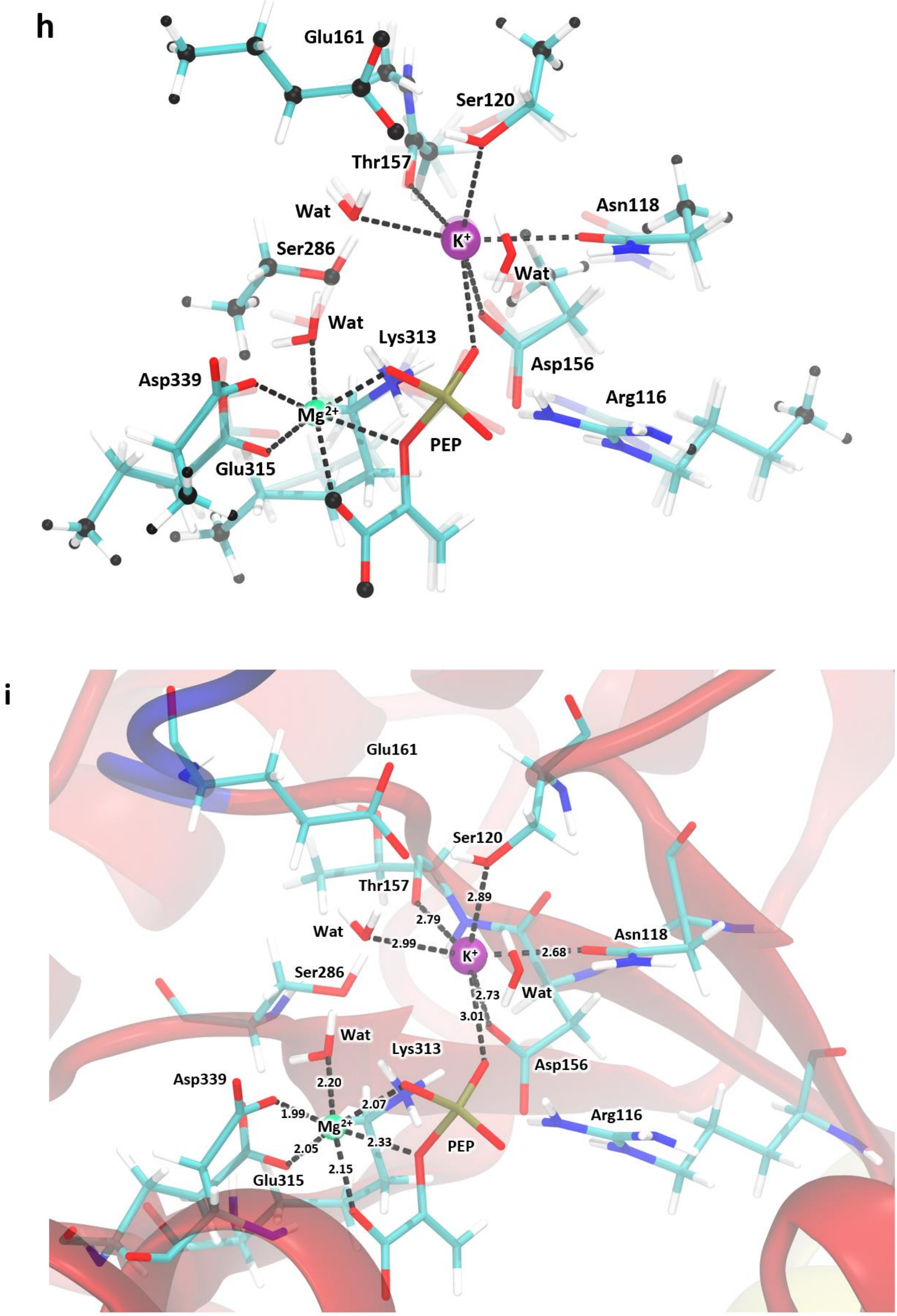


Figure 4.3 (Continued)

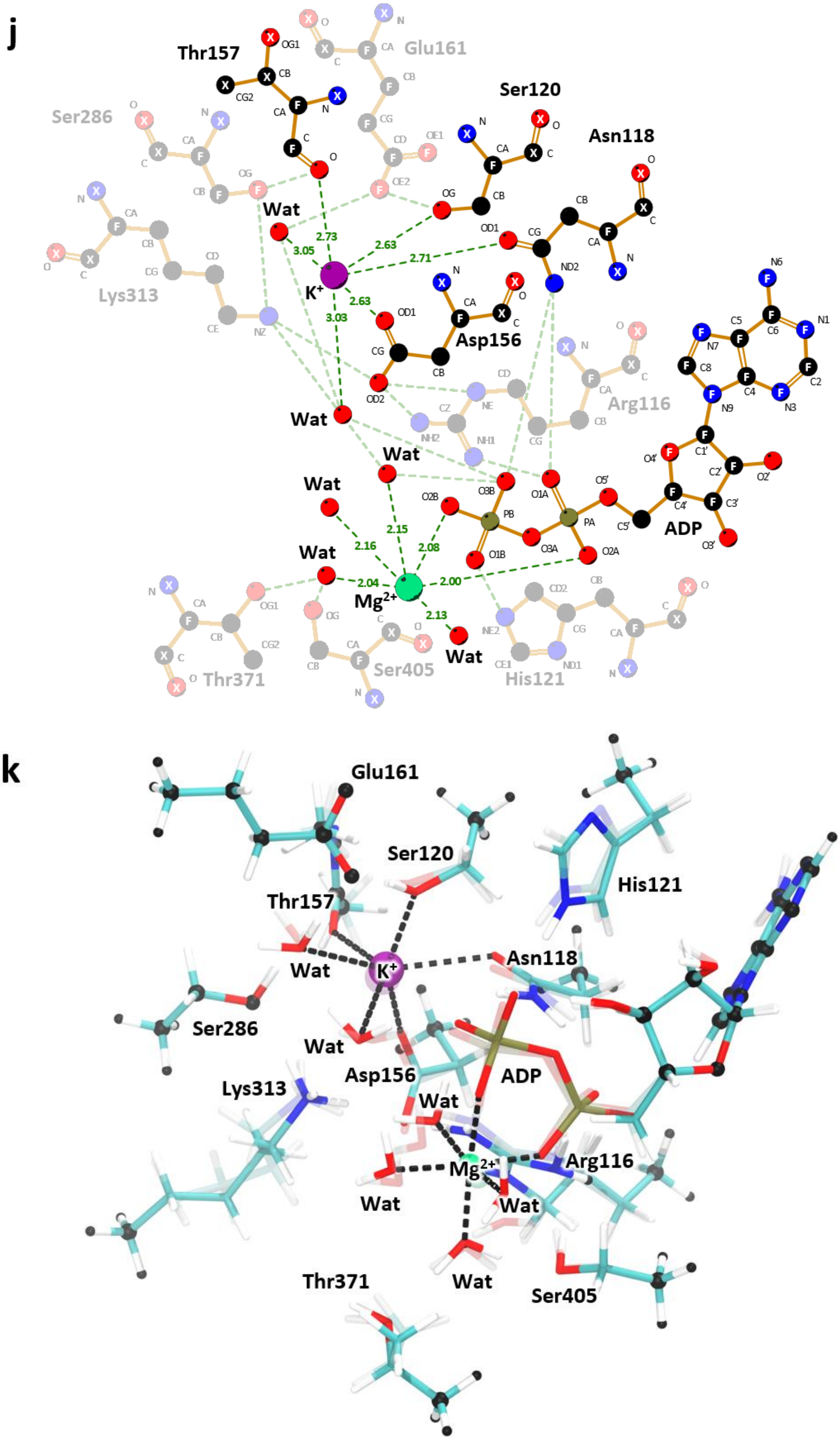


Figure 4.3 (Continued)

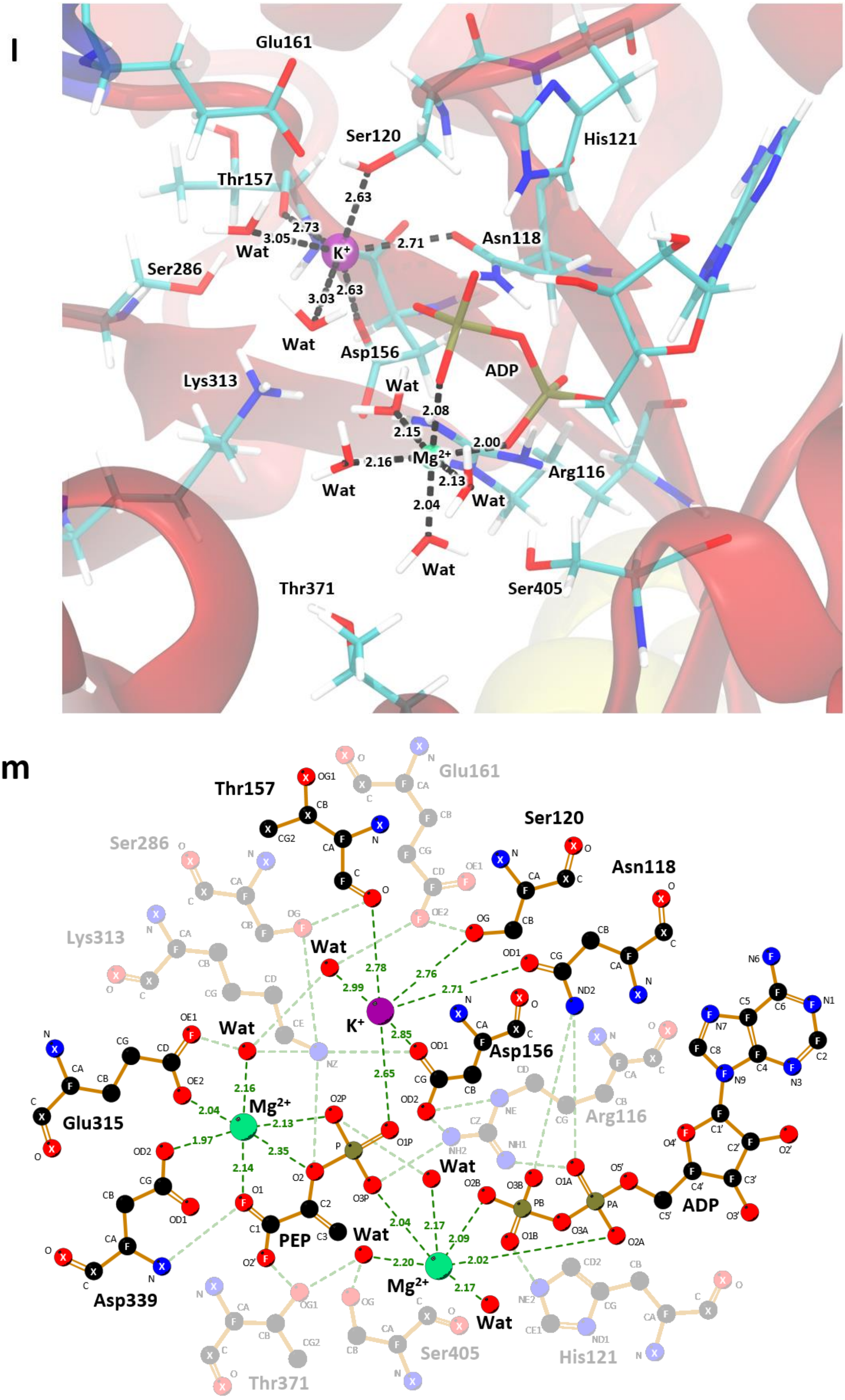
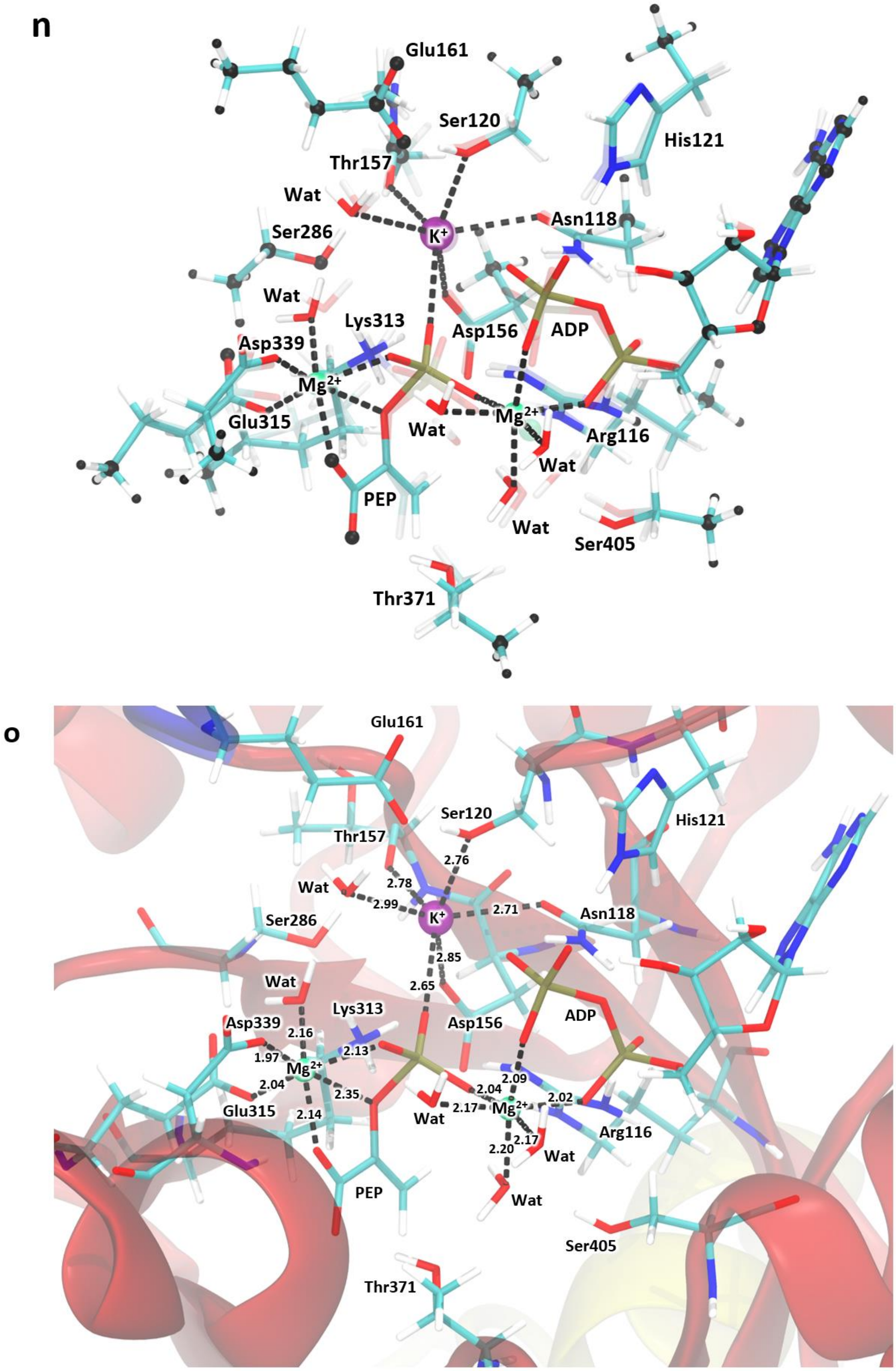


Figure 4.3 (Continued)



From the optimized structures of the “small model”, bond and angle parameters were calculated and imported to the corresponding MD topologies. Regarding the “large model” representations of the cluster models, they were subjected to hydrogen-only geometry optimizations to subsequently derive atomic charges. The response of the resulting parameter sets was tested in several trials of MD simulations and refined iteratively until observing a stable behavior and well-reproduced geometries over the course of the simulations. Section 4.1.2.3 further elaborates on the geometrical and structural assessment of the metal centers during MD.

Table 4.2

Structural divergence between the initial and final states of the geometry optimizations

Modeled conditions	Small model		Large model
	RMSD (Å)	Number of steps	Number of steps
K-holo	0.563	98	177
K-Mg-holo and FBP-holo	0.634	132	93
PEP-holo	0.423	129	53
ADP-holo	0.388	107	85
PEP-ADP-holo and Full-holo	0.365	95	39

Before parameter refinement, occasional simulation crashes were experienced due to inaccuracies in the treatment of the K^+ metal center with the standard bonded model. Prematurely terminated simulations were accompanied by LINCS warning messages. The LINCS algorithm (Linear Constraint Solver) [492] maintains control of bond lengths according to their equilibrium values in simulation. When anomalies are detected in the corresponding values at a specific time step, an incident report is generated. Specifically, the bond of the hydroxyl group of Ser120 was found to exceed the default maximum value of 30 degrees of rotation per time step. This moiety coordinates K^+ and simultaneously acts as a H-bond donor with other nearby chemical groups. The H-bond is usually formed with the side chain of Glu161. This interaction is structurally dependent on the conformational transition of the B domain, based on the evidence from crystallographic data [137]. The interaction can be observed when the B domain is in a range of partially or totally closed conformations, whereas it may be absent when the open conformation is sampled. A careful examination of the problematic simulations revealed that the incident was precisely produced in instances of the latter case, where the eventual interruption of the H-bond with Glu161 caused the side chain of Ser120 to undergo unphysical behavior due to an imbalance of the forces acting on it. The stability of the region substantially improved by applying a variation of the bonded model that treats the coordination bonds as harmonic restraints and considers the full set of non-bonded terms between consecutively connected atoms. Such an approach achieved the correct modeling of the interchange of H-bond acceptors of Ser120.

All coordinated water molecules were fully treated with the non-bonded model (*i.e.*, their QM-derived parameters were excluded). Table 4.3 presents a qualitative validation of the employed metal-center parameters, using the reference values recommended by the experts of MCPB.py. According to their guidelines, appropriate metal-center parameters should generally adhere to the following criteria (although exceptions may occur): i) the bond force constants between a metal ion and its coordinated atoms are lower than 200 $\text{kcal·mol}^{-1}\text{·Å}^{-2}$; ii) the equilibrium bond distances between a metal ion and its coordinated atoms are lower than 2.8 Å; iii) the angle force constants related to the metal ion are lower than 100 $\text{kcal·mol}^{-1}\text{·rad}^{-2}$; iv) the equilibrium angle values related to the metal ion are greater than 100 degrees; and v) the RESP charge of a metal ion is lower than its oxidation state.

Table 4.3

Validation of the employed metal-center parameters according to the reference criteria of MCPB.py

Criteria	Modeled conditions				
	K-holo	K-Mg-holo and FBP-holo	PEP-holo	ADP-holo	PEP-ADP-holo and Full-holo
Bond force constants < 200 kcal·mol^{-1}·Å^{-2}	4 / 4	6 / 6	10 / 10	6 / 6	13 / 13
Equilibrium bond distances < 2.8 Å	4 / 4	6 / 6	8 / 10	6 / 6	12 / 13
Angle force constants < 100 kcal·mol^{-1}·rad^{-2}	10 / 10	13 / 13	31 / 31	13 / 13	37 / 37
Equilibrium angle values > 100°	7 / 10	9 / 13	19 / 31	9 / 13	23 / 37
RESP charge of metal ion less than oxidation state	1 / 1	2 / 2	2 / 2	2 / 2	3 / 3

Most of the obtained parameter sets are consistent with the criteria listed above. The most significant exception is the case of angle parameters, where several equilibrium values are below 100°. For instance, this occurs between some of the coordination ligands of the K$^+$ metal center, where the local conformation of the protein and thus the corresponding orientation of the side chains promote a distorted trigonal prismatic geometry. Another example is the molecule of PEP, which is a tridentate ligand of Mg^{2+} and therefore adopts a very specific conformation.

4.1.2 Trajectory analysis

The following section describes the results of the analyses performed to examine the stability and flexibility of the systems along the trajectory. Standard methods for structural and geometric analysis were applied to address two main objectives: i) to monitor the structural divergence of the trajectories with respect to the initial structure, and ii) to characterize the flexibility profile of the protein, *i.e.*, to identify the regions with high or low conformational variability.

The insight gained at this primary level allowed us to subsequently guide further advanced analyses more effectively, targeting the regions that contain the most relevant dynamical events. Results are presented in an orderly manner, starting with the properties of the WT apo system, and then comparing them qualitatively with those of the holo conditions and inferring the differential effects of ligand binding (cofactors, substrates and allosteric activator) on the enzyme behavior.

Finally, a more specific analysis (and complementary to the previous one) on the region of the metal binding sites was performed to evaluate the performance of the incorporated parameter sets. The aim was to assess whether the stability and flexibility of this region were consistent with the expectations derived from the parameterization procedure that was selected for this project.

4.1.2.1 Stability

The evolution of the stability of the systems along the MD simulations was examined by computing the root-mean-square deviation (RMSD) of each snapshot with respect to the initial structure of the trajectory. This widely employed analysis in MD studies offers a general indication of the extent of conformational divergence explored throughout the simulation time.

Figure 4.4a (top panel) shows the plot corresponding to the calculation of the RMSD of the WT apo tetramer (for each of the 5 trajectory replicates separately). For both the calculation of the metric and the prior structural superposition (least-squares fitting), only the protein backbone atoms were considered. Alongside the original data from each time series, a two-sided moving average (darker line) is also shown, which comprises values up to 5 ns on either side of each point and allows for a clearer interpretation of data progression. As a first remark, during the first nanoseconds of simulation the characteristic relaxation curve can be noted, in which the structure rapidly diverges from its initial conformation. RMSD values exhibit an abrupt transition from zero to a range between 0.27 and 0.32 nm at approximately 25 ns. Subsequently, values remain relatively stable until 60 ns. This progression is common to all 5 replicates, perhaps with the exception of replicate #2, which displays a more pronounced initial rise and reaches slightly higher RMSD values before decreasing to the aforementioned range.

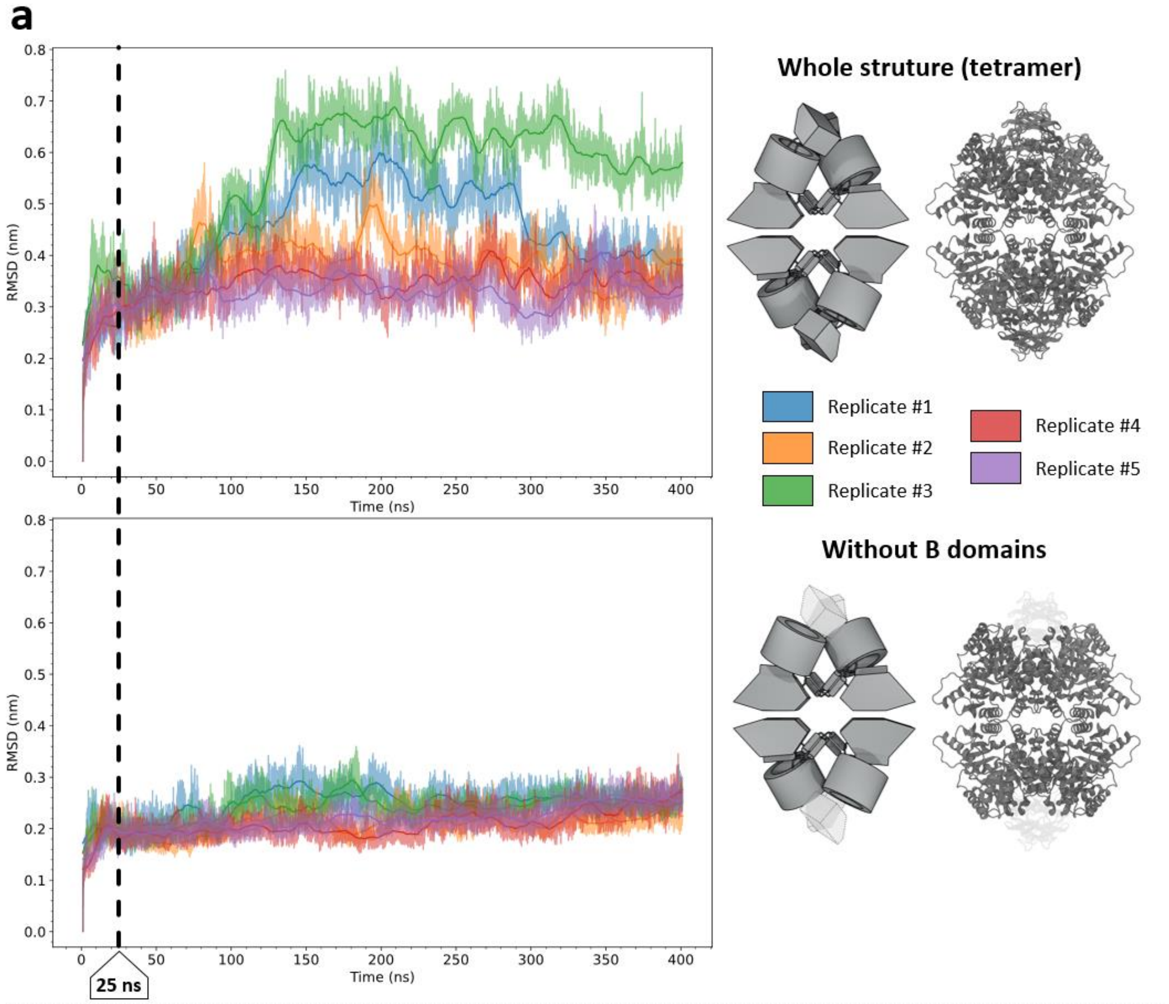
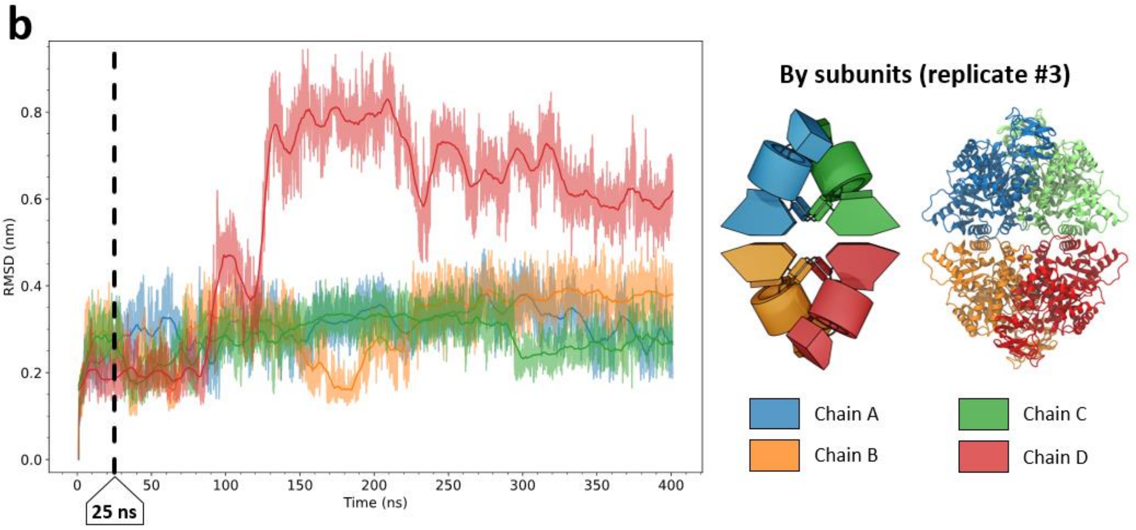


Figure 4.4. Time-series RMSD of the WT apo trajectories. The analysis was applied to the protein backbone atoms. The darker line plotted alongside each time series represents the two-sided moving average of the data, encompassing values up to 5 ns on either side of each point. Schematic representations of the analyzed structure in each panel are included with a diagram of the arrangement of the subunits and domains of the protein and the corresponding ribbon representation. A vertical black dashed line indicates the cutoff of 25 ns that marks the end of the structural relaxation phase. (**a**) RMSD values of the whole tetramers (top panel) and disregarding the B domains (bottom panel). (**b**) RMSD values by monomers of the trajectory replicate #3. A separate structural superposition was performed for each monomer before computing the corresponding RMSD values. NOTE. The images of the protein structure were generated with the software VMD. The 3D schematic models were built with the software Blender [136].

Figure 4.4 (Continued)



Beyond this time window, two distinct behaviors can be observed. On the one hand, some trajectories remained stable for the rest of the simulation, without major deviations from the average RMSD value. On the other hand, other trajectories exhibited significant structural divergence, exploring a range of conformations characterized by higher and diverse RMSD values. Replicates #4 and #5 exemplify the former behavior, while replicates #1 and #3 are representative of the latter. Replicate #2 displays an intermediate behavior, characterized by moderate stability with a broader range of RMSD fluctuation.

The observed disparity in behavior among the replicates raises an ambiguity regarding the overall stability profile of the system over the course of 400 ns of MD simulation. Nevertheless, it is well-established that the B domain of pyruvate kinases is a particularly mobile region of the protein and is capable of adopting multiple conformations, as evidenced by both MD simulations and X-ray crystallography [121], [132], [138], [164]. The flexibility of the B domain can be attributed to the two small linkers that covalently connect it to the A domain and serve as a hinge mechanism. Accordingly, the RMSD analysis was repeated but excluding the B domains of the tetramer to assess the stability of the protein core (N-terminal, A, and C domains; ~80% of the structure). The corresponding results (Figure 4.4a, bottom panel) revealed that the tetramer core remained strongly stable after the initial time of structural relaxation. All structures fluctuated close to their average RMSD values, with replicates #1 and #3 now being only marginally higher than the rest. Therefore, our simulations confirm that the main source of structural divergence of PKR corresponds to the sampling of different conformations of the B domain. In contrast, the protein core exhibits considerable stability, as consistently demonstrated among trajectory replicates.

This analysis not only facilitates the understanding of the stability profile of the system, but also serves as an approach to discriminate between equilibration and production phases of the trajectories. Despite the inclusion of equilibration stages within the MD protocol, satisfactory equilibration is not achieved until structural relaxation has been completed. By identifying the endpoint of the initial RMSD curve, it is possible to determine the portion of trajectory that should be discarded from subsequent analyses to avoid the inclusion of potential artifacts derived from structural relaxation. Based on the observations from the analysis conducted, this cutoff can be established at 25 ns. This suggestion becomes particularly evident when examining the RMSD plot of the tetramer cores, in which the RMSD values of all trajectory replicates have already reached a plateau at 25 ns. The inclusion of the B domains in the analysis obscures the presence of a distinct cutoff, as the conformational divergence may reduce the clarity of the plateau.

In addition, the RMSD analysis was conducted at the level of individual enzyme monomers to investigate whether the stability profile of the tetramer is proportionally reflected in each of its four subunits. Figure 4.4b illustrates the case of the trajectory replicate #3, which was shown to exhibit the greatest structural divergence among replicates (Figure 4.4a, top panel). Notably, the behavior of each subunit within the tetramer was not equivalent. Chain D stands out as being particularly unstable, and its RMSD profile is comparable to that displayed by the whole tetramer in the respective plot. On the other hand, chains A, B, and C were comparatively more stable, balancing between two close RMSD states in an independent manner. Incidentally, the RMSD profile of each monomer conforms to the post-relaxation cutoff of 25 ns. These observations suggest that the behavior of the B domains within the tetramer is not symmetrical, in agreement with the previous studies of pyruvate kinases [121], [132], [138], [164]. The four B domains may undergo smaller or larger conformational changes independently between different simulation replicates. Moreover, the apparent degree of stability of the entire system is strongly influenced by the behavior of the individual B domains in a basic analysis such as RMSD.

Finally, the trajectories of the different holo conditions were also analyzed, both including and excluding the B domains (Figure 4.5). In general, the results indicate that the B domains contribute significantly to the overall fluctuation of the protein, similarly to the case of the apo trajectories. However, some exceptions occur in which the RMSD profiles of the whole tetramer and the tetramer core are qualitatively equivalent. This behavior is displayed by certain trajectory replicates of the PEP-holo and PEP-ADP-holo conditions, suggesting that the presence of PEP might induce distinctive dynamical events with the B domain and the subunit core coupled with each other. In addition, a positive correlation between the stability of the protein and the number of bound ligands is apparent. This trend can be attributed mainly to the ligands that bind to the active site. Remarkably, this implies that while the constraints of cofactor and substrate binding were imposed on the A domain, the B domain also exhibited rigidity, possibly both by propagation of dynamical effects and interactions with the ligands. On the other hand, the effects of the allosteric activator FBP on the stability of the B domains and the tetramer core are not clear from this analysis. The incorporation of FBP to the protein with cofactors and substrate (Full-holo *vs.* PEP-ADP-holo) suggests that FBP may have a stabilizing effect on the tetramer core, but further evidence is needed to confirm this hypothesis.

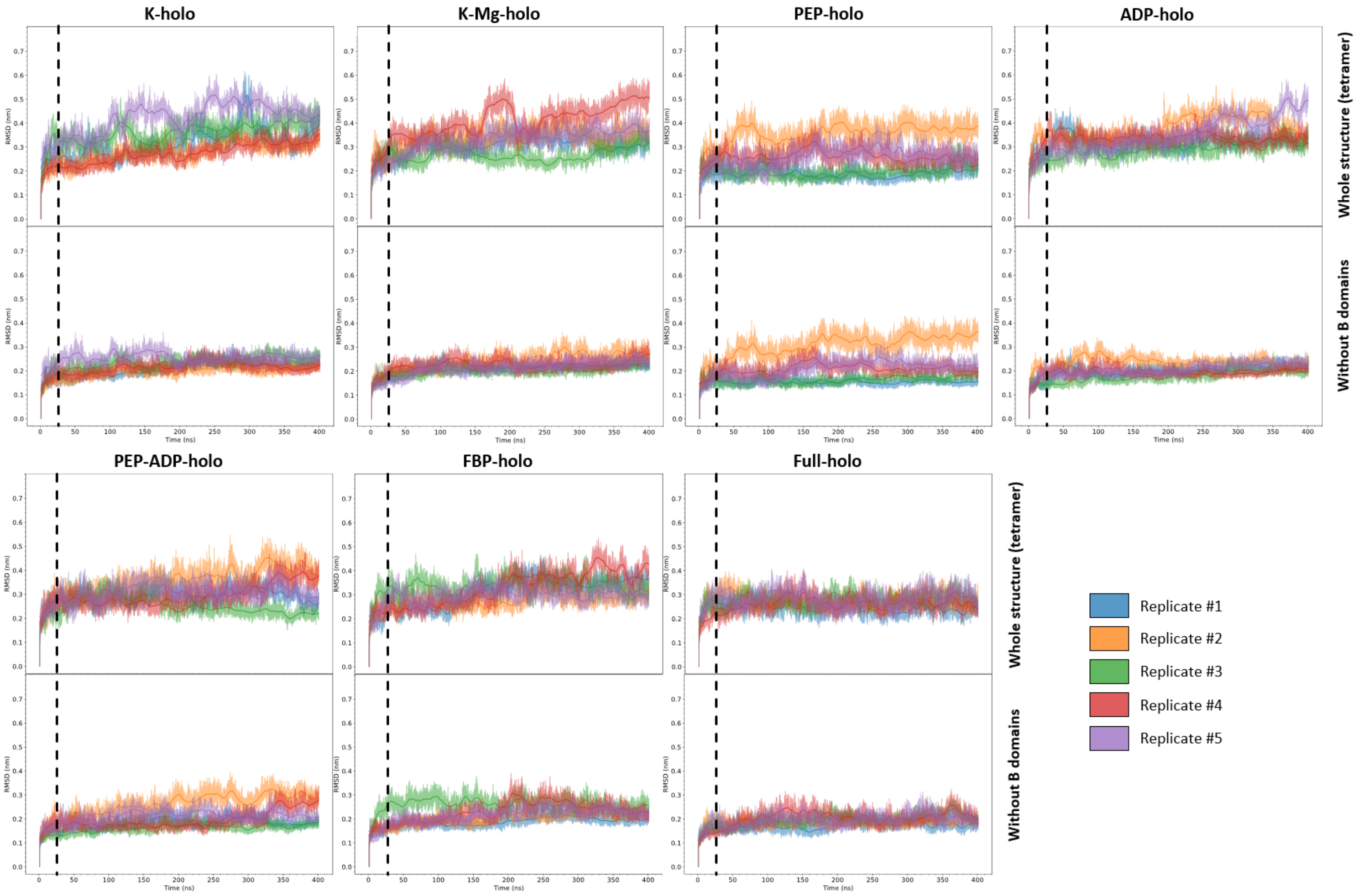


Figure 4.5. Time-series RMSD of the WT holo trajectories. The analysis was applied to the protein backbone atoms. The darker line plotted alongside each time series represents the two-sided moving average of the data, encompassing values up to 5 ns on either side of each point. For each holo condition, the top and bottom panels correspond to RMSD values of the tetramers including and excluding the B domains, respectively. A vertical black dashed line indicates the cutoff of 25 ns that marks the end of the structural relaxation phase.

4.1.2.2 Flexibility

Following the RMSD analysis, the information was complemented with a root-mean-square fluctuation (RMSF) analysis. This metric quantifies the divergence from the average position of each region of the protein, facilitating the identification of flexible fragments. The RMSF was computed for the backbone atoms of different protein regions (after a corresponding structural superposition) and expressed as a mass-weighted average per residue. Figure 4.6 shows the RMSF profile of a WT apo tetramer, using the trajectory replicate #3 as an example. In accordance with the insights gained from the RMSD analysis, the B domains stand out as the most flexible regions, especially in chain D.

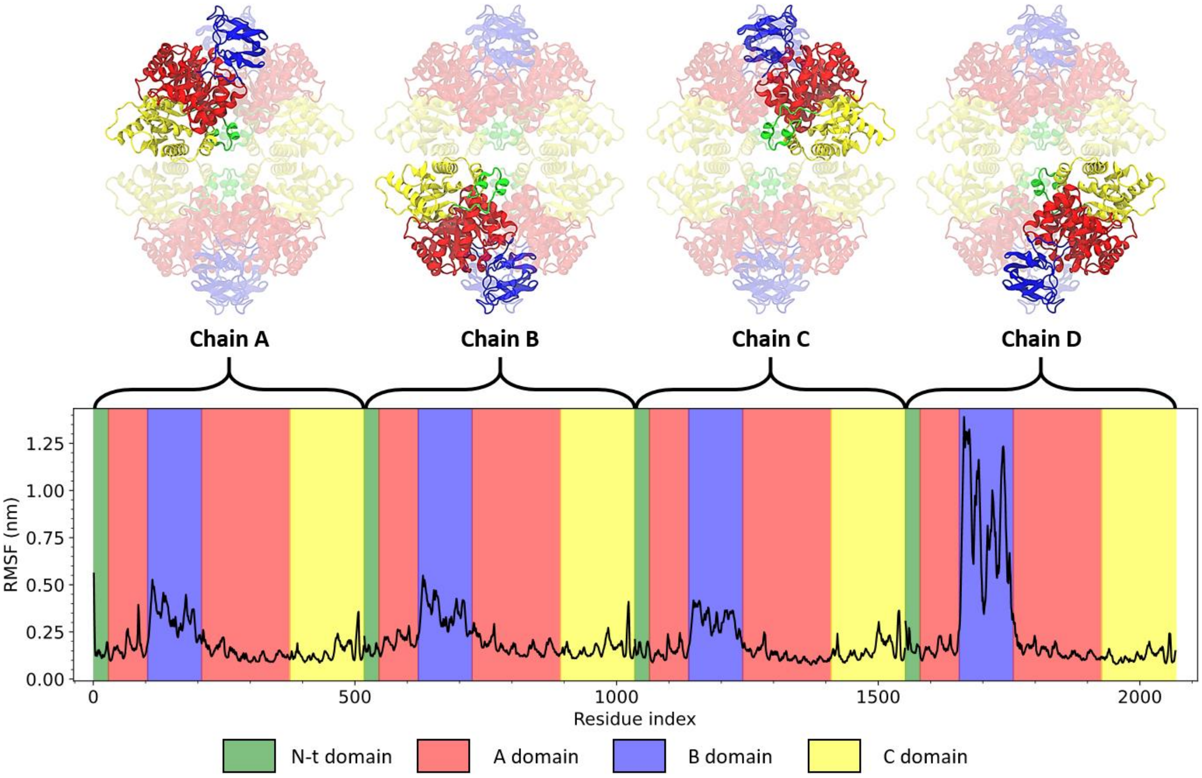


Figure 4.6. RMSF per residue of a WT apo tetramer. Data corresponds to trajectory replicate #3. The regions corresponding to each subunit of the tetramer are indicated with labels and accompanied by schematic representations of the structure with a ribbon representation with the respective subunit highlighted. NOTE. The images of the protein structure were generated with the software VMD

The flexibility profile at the level of the enzyme monomer was characterized in greater detail. Firstly, all subunits from each trajectory replicate were subjected to individual RMSF analyses. Figure 4.7 shows the corresponding results with the averaged data for all monomers. Subsequently, a multi-trajectory of all subunits was generated by concatenating their respective trajectory data. RMSF was calculated for each protein domain on the multi-trajectory to examine the most flexible fragments. Such results are presented in Figure 4.8.

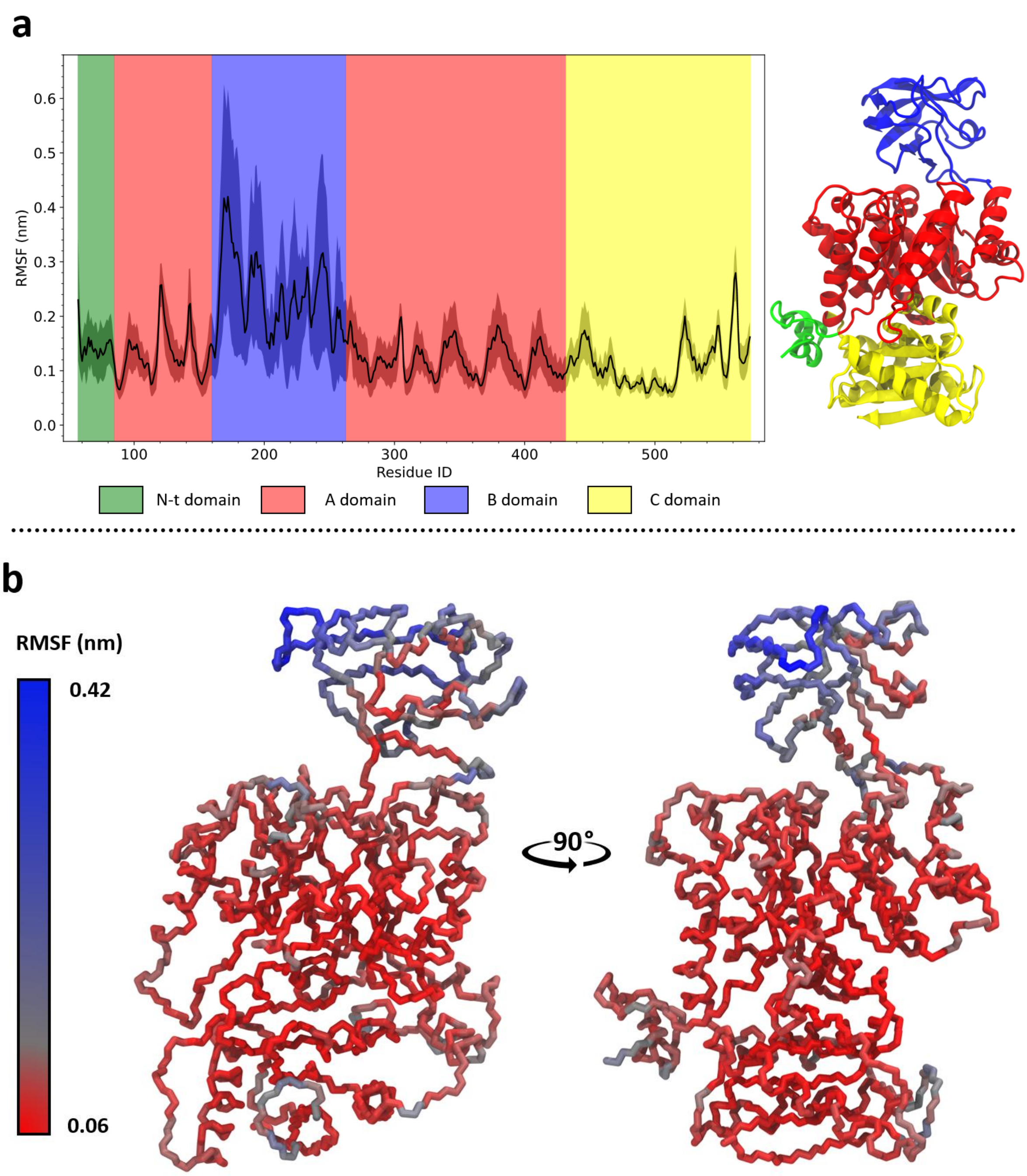


Figure 4.7. Average RMSF per residue of WT apo monomers. The analysis comprises the data from all trajectory replicates. (**a**) RMSF plot of the average values with standard deviation (depicted with a darker area around the line). A schematic representation of the structure of the monomer with a ribbon representation is included. (**b**) Backbone structure of the monomer colored by average RMSF value. RMSF values range from 0.06 to 0.28 nm and correspondingly map to a red-gray-blue color transition, as indicated in the color bar. Two views of the structure depicted with the licorice representation are provided, with a horizontal rotation of 90° with respect to each other. NOTE. The images of the protein structure were generated with the software VMD.

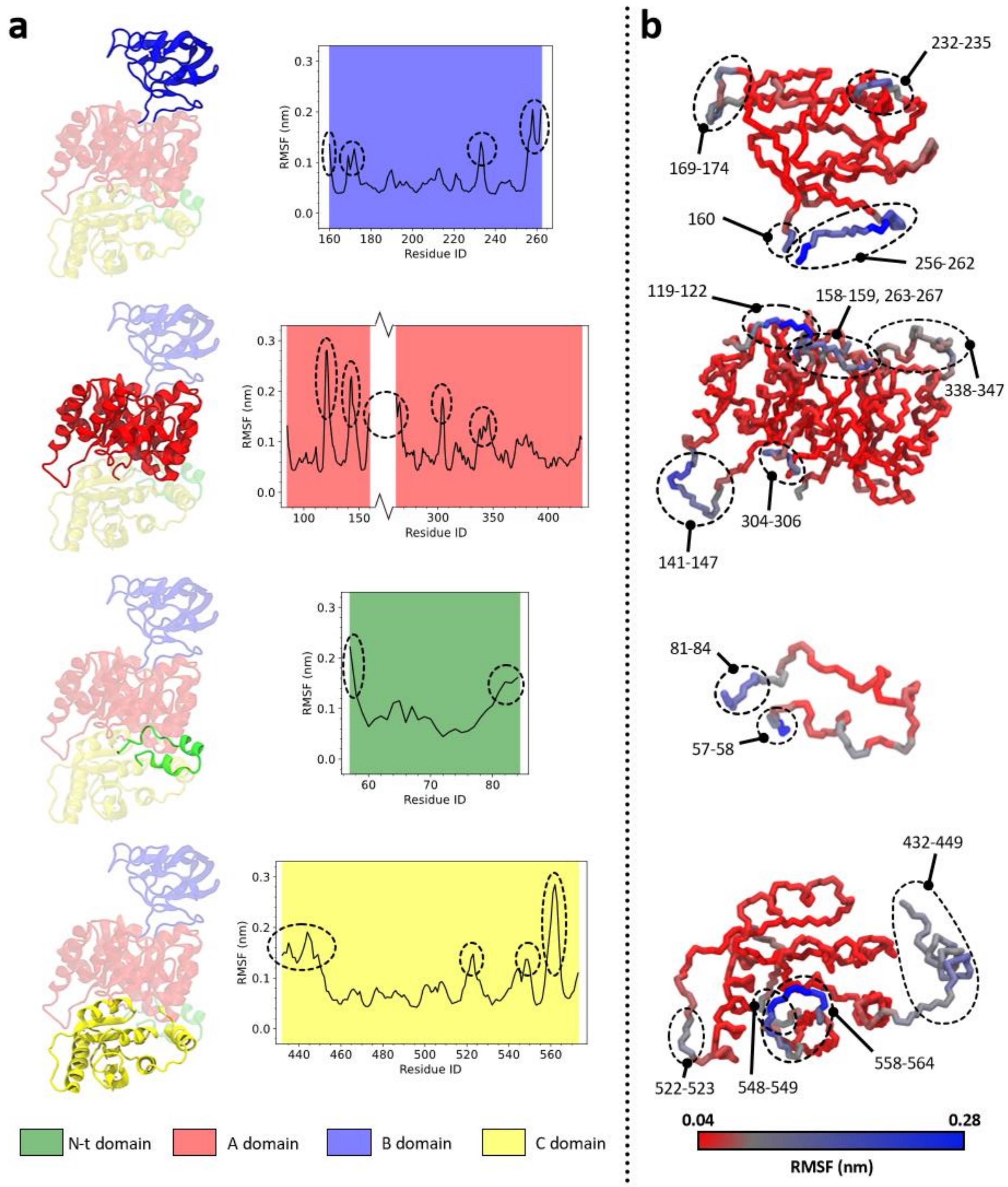


Figure 4.8. RMSF per residue of the WT apo domains. The analysis was performed on the concatenated version of the trajectory data of all subunits from each trajectory replicate. (**a**) RMSF plot of each domain. Each plot is accompanied by a schematic representation of the structure of the monomer with a ribbon representation and with the respective domain highlighted. The structural fragments with particularly high RMSF values are encircled with dashed lines. (**b**) Backbone structures of the domains colored by average RMSF value. RMSF values range from 0.04 to 0.28 nm and correspondingly map to a red-gray-blue color transition, as indicated in the color bar. The structure of each domain is presented in the same order as in (a), namely (from top to bottom), B, A, N-terminal, and C domains. Structures are depicted with the licorice representation. The structural fragments with particularly high RMSF values are encircled with dashed lines, matching those in (a) and incorporating their residue ID labels to facilitate identification of the most flexible regions. NOTE. The images of the protein structure were generated with the software VMD.

The flexibility profiles show that all domains have stable tertiary structure folds. Distinctive local fluctuations occurred mainly at the loop fragments between the secondary structure elements and far from the core of the structure. All linker fragments that connect the different domains of the monomer exhibited significant flexibility, which suggests that the conformational divergence of the monomer largely arises from rigid-body motions between domains.

The structural mobility of the B domain with respect to the protein monomer was high and variable. It displays the highest average RMSF values of the structure, with broad standard deviation intervals that reflect how this domain may undergo smaller or larger conformational changes independently among subunits and trajectory replicates. The B domain possesses two internal loops with significant flexibility, namely, at residues 169-174 and 232-235. The hinge region that connects it with the A domain is composed of two linker fragments located at both endpoints of its sequence. Remarkably, the final linker is longer and displays the highest flexibility among all inter-domain linkers.

The A domain is characterized by a rigid core, as reflected in the eight distinctive valleys of the RMSF profile that correspond to the β-strands of the barrel fold of this domain. Conversely, the loops that interconnect the secondary structure elements exhibited prominent peaks, in particular, L-Aβ2-Aα2 (residues 119-122), L-Aα2-Aβ3 (residues 141-147), L-Aα4-Aβ5 (residues 304-306), and the fragment 338-347, which forms a small α-helix (Aα6') that connects Aβ6 and Aα6. Notably, the L-Aβ2-Aα2 loop contains residue Ser120 which binds to the cofactor K^+. The L-Aα2-Aβ3 loop and the Aα6' helix are particular protrusions of the A domain that extend towards the C and B domains, respectively, and serve as contact points with these. Consequently, their flexibility could be coupled to the motion of these two domains. Finally, the eight α-helices of the barrel adopt intermediate values between the loops and the β-strands.

The C domain displays three loops of significant flexibility, namely, at residues 522-523, 548-549, and 558-564. Remarkably, the latter corresponds to the L-Cβ4-Cβ5 loop that accommodates the allosteric activator FBP, absent in the apo condition. The region of the C domain at the interface with the A domain exhibited the lowest RMSF values of the monomer, together with the β-barrel of the A domain. Lastly, the N-terminal domain displays marked flexibility at its terminus fragment, as expected from a free protein tail. The missing initial fragment (residues 1-56) is hypothesized to be disordered given its common absence in crystallographic structures.

The flexibility profile of the different holo conditions were analyzed using the same approach as in Figure 4.7, *i.e.*, by inspecting the average RMSF values of all monomers from each trajectory replicate of the same condition. Figure 4.9 shows the corresponding analyses. The comparison between the RMSF profiles of each condition reinforces the idea that the bound ligands at the active site exerted a clear rigidification effect on the B domain, in accordance with the observations from the RMSD analyses. This effect is already noticeable at the K-holo condition, which only features the cofactor K^+ as a ligand. Additionally, the binding of this cofactor abolished the fluctuation of the L-Aβ2-Aα2 loop given the constraints of Ser120 as a coordination ligand of the metal. The binding of the PEP strongly reduced the variability of flexibility of this domain, as can be observed by the narrowed standard deviation of conditions PEP-holo, PEP-ADP-holo, and Full-holo. In other words, monomers with PEP bound behaved more uniformly both within the same tetramer and between trajectory replicates. Lastly, the L-Cβ4-Cβ5 loop that stabilizes the binding of FBP to the allosteric site (positions 558-564) was markedly rigidified when the ligand was present, *i.e.*, in the FBP-holo and Full-holo conditions.

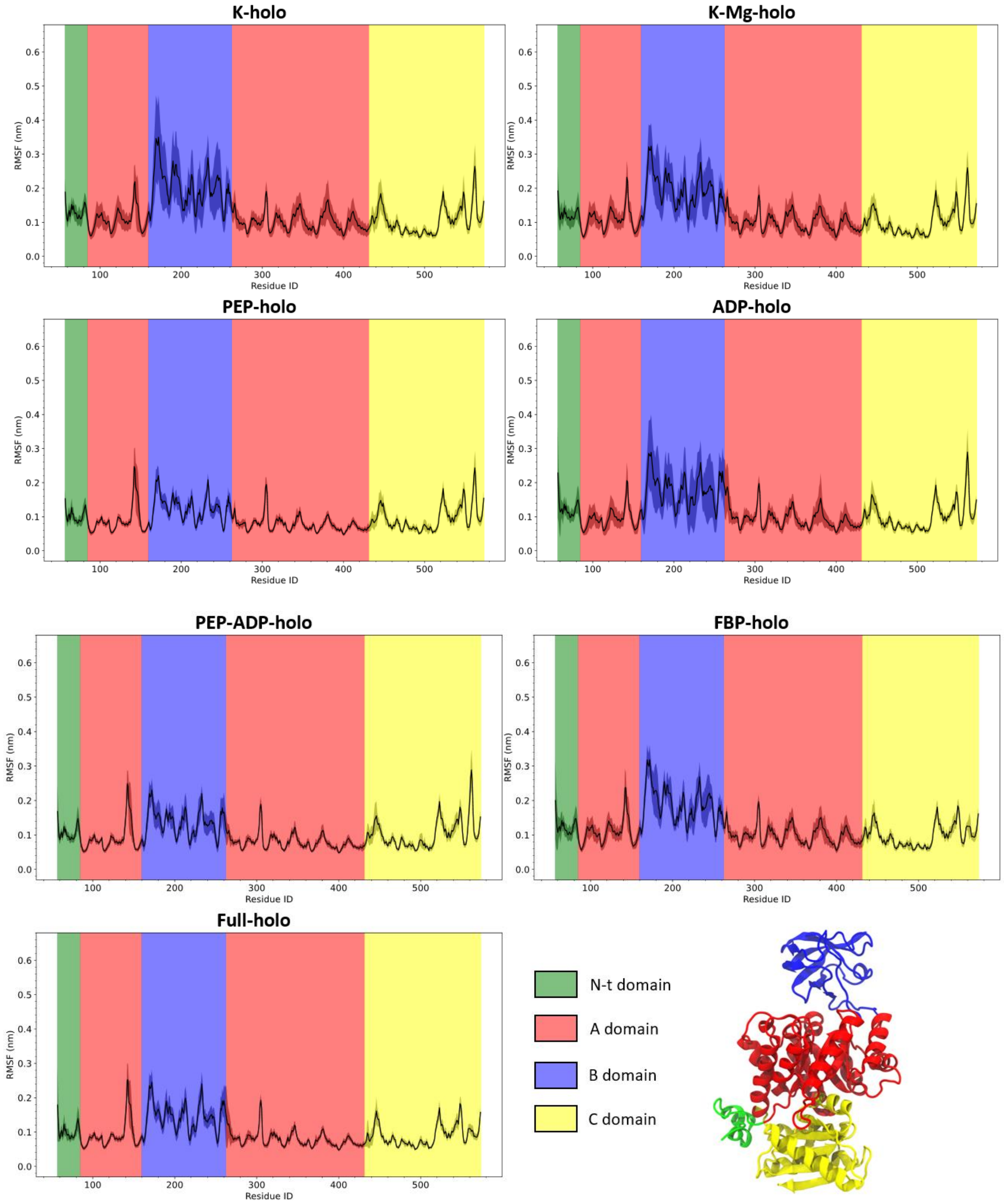


Figure 4.9. Average RMSF per residue of WT monomers of each holo condition. The analysis comprises the data from all trajectory replicates of the same condition. Standard deviations are depicted with the darker areas around the average. A schematic representation of the structure of the monomer with a ribbon representation is included. NOTE. The image of the protein structure was generated with the software VMD.

4.1.2.3 Analysis of the metal centers

While in section 4.1.1.2 descriptive and qualitative assessments of the metal-center parameterization were given, the present section elaborates on a quantitative analysis on the geometrical and structural features of the metal centers along the MD simulations. The conducted analyses consisted in: i) inspecting the variability of the distance values of the coordination bonds, ii) assessing the overall flexibility of the region, and iii) measuring the relative abundance of the sampled coordination numbers (by the solvent).

4.1.2.3.1 Distance values of the coordination bonds

Firstly, the distance values of all the modeled coordination bonds were collected to inspect their variability in MD with respect to their parameterized equilibrium values. Figure 4.10 shows the corresponding analysis, which was carried out on all trajectories of each holo condition. Each box plot in the figure contains the aggregated data of each PKR subunit in the 5 trajectory replicates of the corresponding bond and condition. Where applicable, the original distance values from the source crystallographic structures are included as an additional reference. The analysis does not comprise the coordination bonds with water molecules, since those were not parameterized and were rather allowed to be established transiently with different water molecules as they freely occupy the coordination binding sites.

All distance values were consistently maintained during MD with an acceptable range, although as a general tendency they were marginally greater than the parameterized equilibrium values. Value distributions of the K^+ metal center feature moderately larger upper tails with more outlier occurrences, whereas distributions of the Mg^{2+} metal centers are narrower and more symmetrical. When it comes to the comparison with the crystallographic values, metal-ligand distances were not reproduced in general, except in some instances such as the bonds of ADP with its complexed Mg^{2+} ion.

The parameterized coordination bonds are either lower or greater with respect to their crystallographic counterparts. In general, the coordination bonds involving K^+ fall in the first case, with the simulation values lying between the crystallographic and the parameterized values. In contrast, the bonds involving Glu315 and Asp339 with the cofactor Mg^{2+} in the Mg-1 site are instances of the second case. These crystallographic values are especially low in the Mg-1 site (conditions PEP-holo, PEP-ADP-holo, and Full-holo), compared to those of the Mg-3 site (conditions K-Mg-holo and FBP-holo). It should be noted that the crystallographic model of the Mg-1 site featured Mn^{2+} instead of Mg^{2+}. When it comes to the parameterization of these bonds, the Mg-1 and Mg-3 sites share very similar equilibrium values, which were also very stable in MD.

The only bond lacking a crystallographic reference was that between PEP (atom O3P) and the ADP-bound Mg^{2+}. It is not possible to obtain a crystallographic structure naturally with these ligands since the chemical reaction takes place. In our MD simulations, the distance values for this bond were greater than the corresponding parameterized equilibrium value. This behavior was expected, since the simulation is requested to sustain the repulsion between the phosphate groups of PEP and ADP while being incapable of modeling the phosphoryl-transfer reaction that would eventually occur.

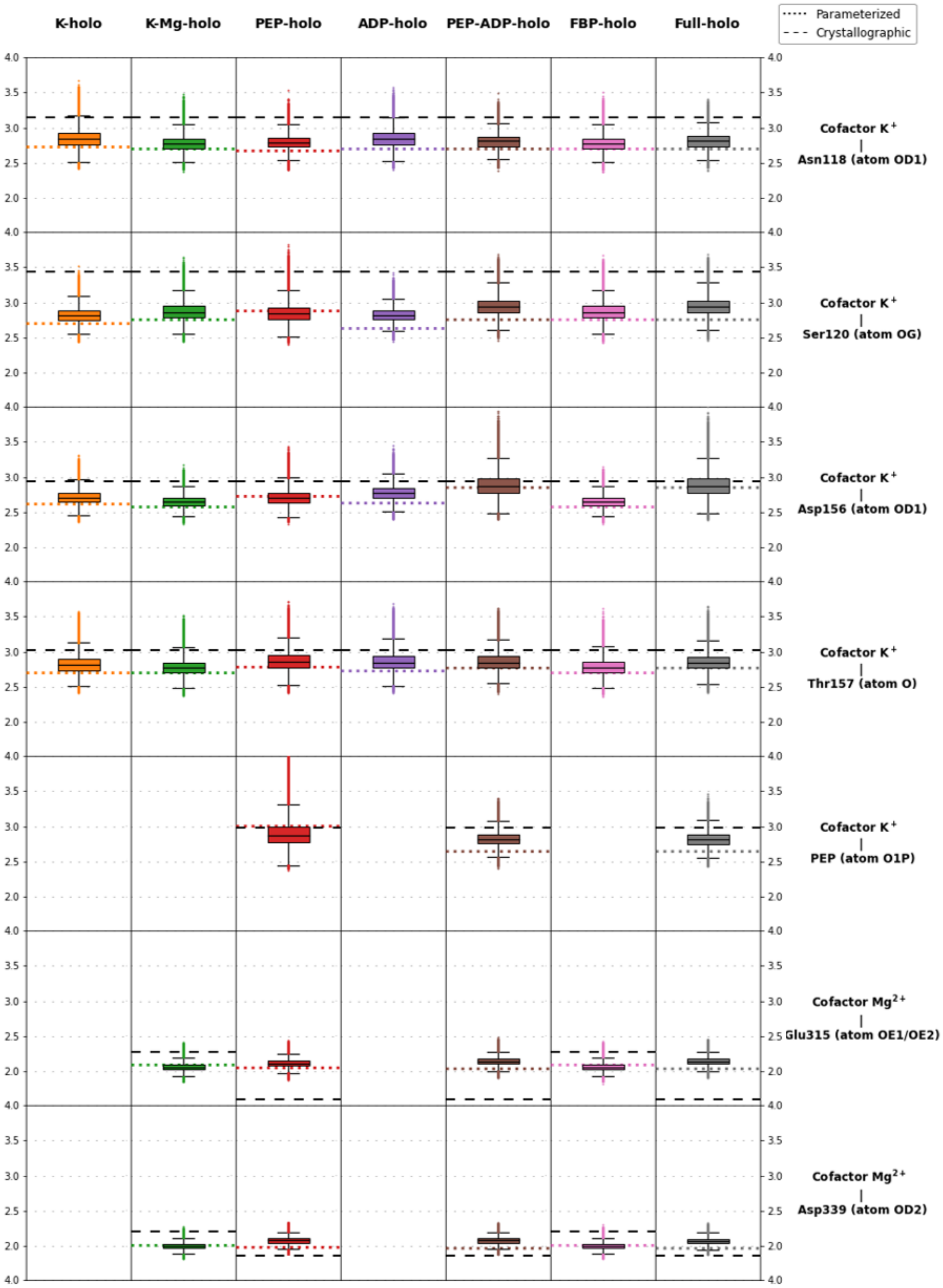
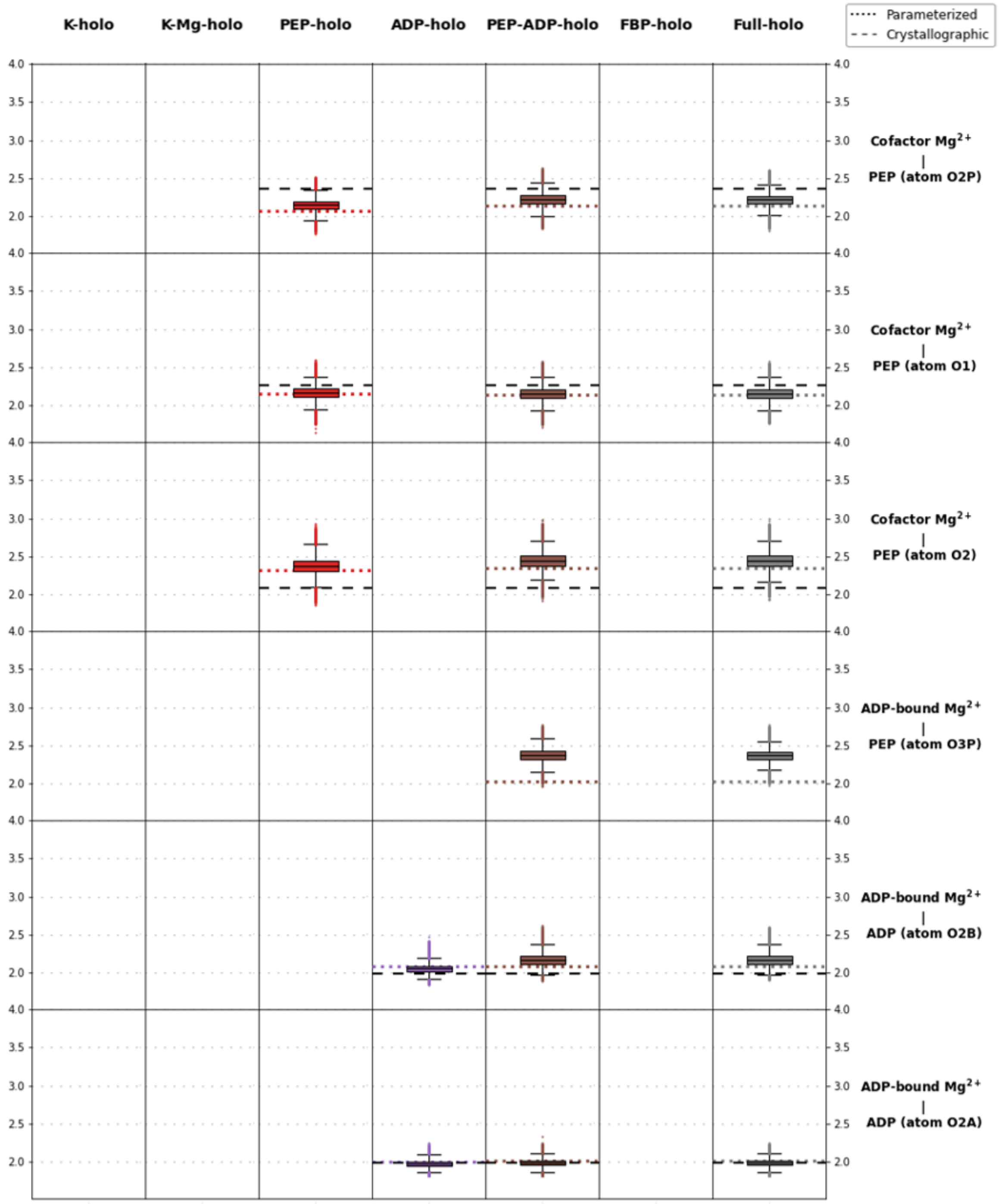


Figure 4.10. Distance values (Å) of the modeled coordination bonds along the MD simulations. Data is presented as a grid of box plots where rows correspond to each coordination bond and columns correspond to each holo condition. Each box plot contains the aggregated data of each PKR subunit in the 5 corresponding trajectory replicates. The parameterized equilibrium distance value of each bond is included as a reference with a dotted line in each box plot. Where applicable, the original distance values from the source crystallographic structures are included as an additional reference, with a dashed line. The data of each holo condition is colored as follows: K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray.

Figure 4.10 (Continued)



4.1.2.3.2 Stability and flexibility of the region

After having confirmed the stability of the modeled coordination complexes in terms of bond distances, the next analysis aimed to assess the overall flexibility of the region in simulation. This issue was explored with an RMSD and RMSF analysis of the protein local region of the metal centers. Figure 4.11 shows the RMSD values (nm) of each holo simulation, using as reference the corresponding structures of the geometry optimization of the “small model” (Figure 4.3; subfigures c, f, i, l, and o) in each case. In addition, the data is compared with the same analysis carried out on the apo simulations (with the same reference structures, respectively). Each box plot in the figure contains the aggregated data of each PKR subunit in the 5 trajectory replicates of the corresponding conditions. The RMSD analysis was performed only on the amino acids coordinated to K^+ or Mg^{2+}, using the backbone atoms as the fitting group, and calculating the RMSD concerning all their atoms. The ligands were not included in the analysis.

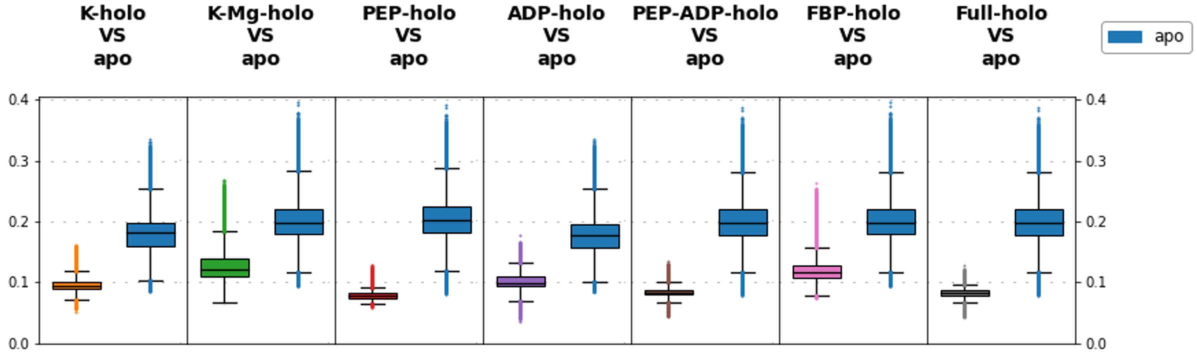


Figure 4.11. RMSD values (nm) of the coordinated amino acids along the MD simulations. Data is presented as a row of box plots where columns correspond to the comparison of each holo condition with the apo condition. Each box plot contains the aggregated data of each PKR subunit in the 5 corresponding trajectory replicates of each condition. RMSD was calculated taking into account all atoms of the amino acids coordinated to K^+ or Mg^{2+}, and using the backbone atoms as the fitting group. The reference structure for the calculation in each box plot was the geometry-optimized "small model" of the corresponding holo condition. The data of each MD condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray.

The analysis reveals that the metal centers of the holo conditions maintained a stable local conformation during the simulations. In general, RMSD values fluctuated around 1 Å when the protein region was constrained only by the K^+ metal center (K-holo, ADP-holo). When the model included the Mg-1 site and the PEP molecule (PEP-holo, PEP-ADP-holo, Full-holo) the region became stiffer with values below 1 Å. Conversely, the models with the Mg-3 site (K-Mg-holo, FBP-holo) reported moderately higher values, while still below 2 Å (disregarding outliers). The corresponding structural divergence for this range of RMSD values consists only in minor side-chain fluctuations without backbone conformational changes (not shown). When it comes to the apo simulations, the analysis reveals that the region was twice as flexible, in general, given that there is a broader range of RMSD values centered around 2 Å. An exploration of the structures with RMSD values of 2 Å or higher revealed backbone conformational changes (not shown). All in all, the protein region of the metal centers had an overall higher conformational divergence in the apo condition than in the holo conditions.

The structural variability of each metal-center amino acid was further explored by calculating the RMSF of the side chains. In this case, the analysis was performed not only on the coordinated amino acids, but also on the rest of amino acids included in the cluster models. The results are shown in Figure 4.12, where the RMSF values (nm) are compared between conditions. Once again, each box plot in the figure contains the aggregated data of each PKR subunit in the 5 trajectory replicates of the corresponding conditions. For each analyzed residue, the backbone atoms were used as the fitting group, and the RMSF was calculated for the side-chain atoms.

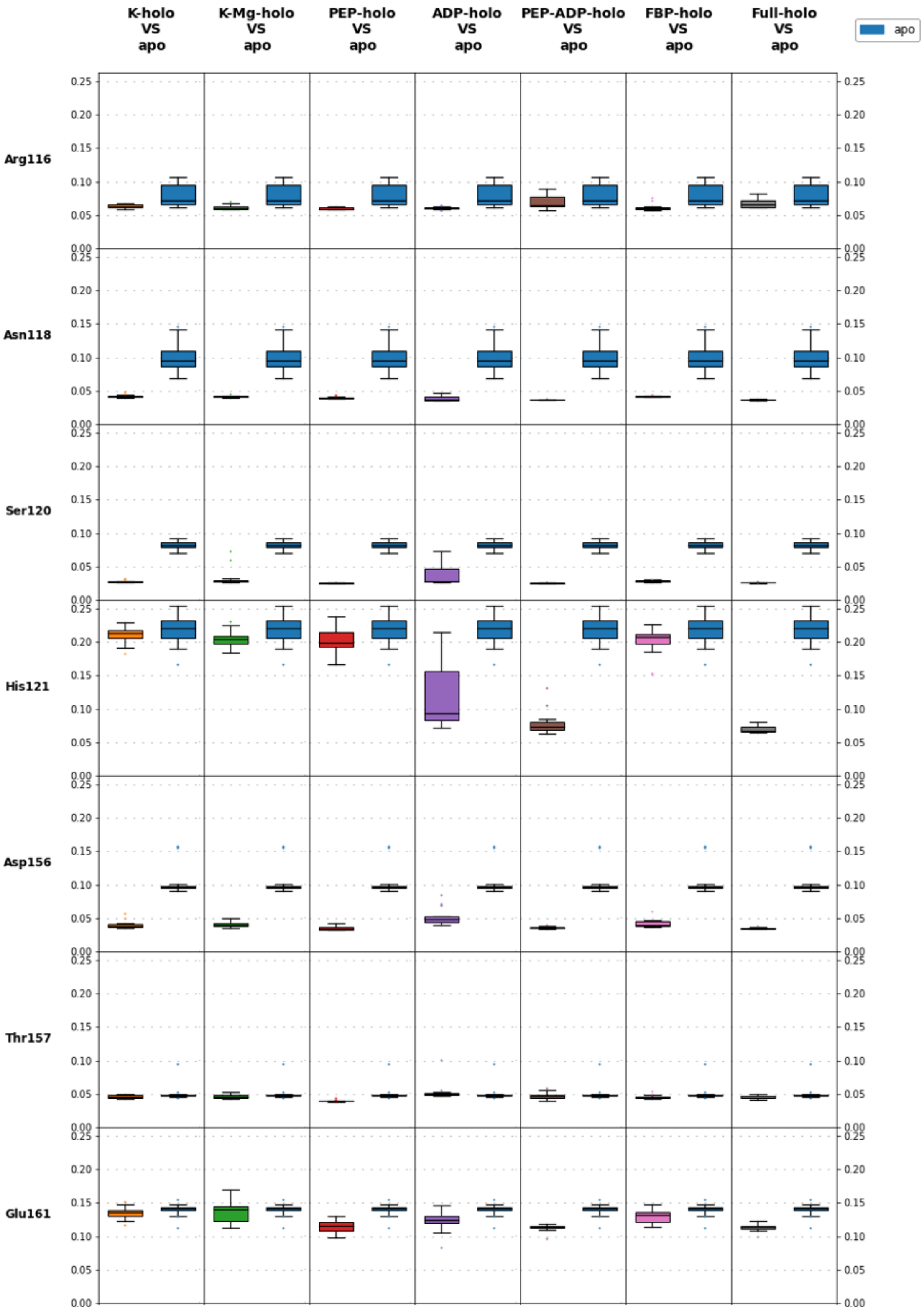
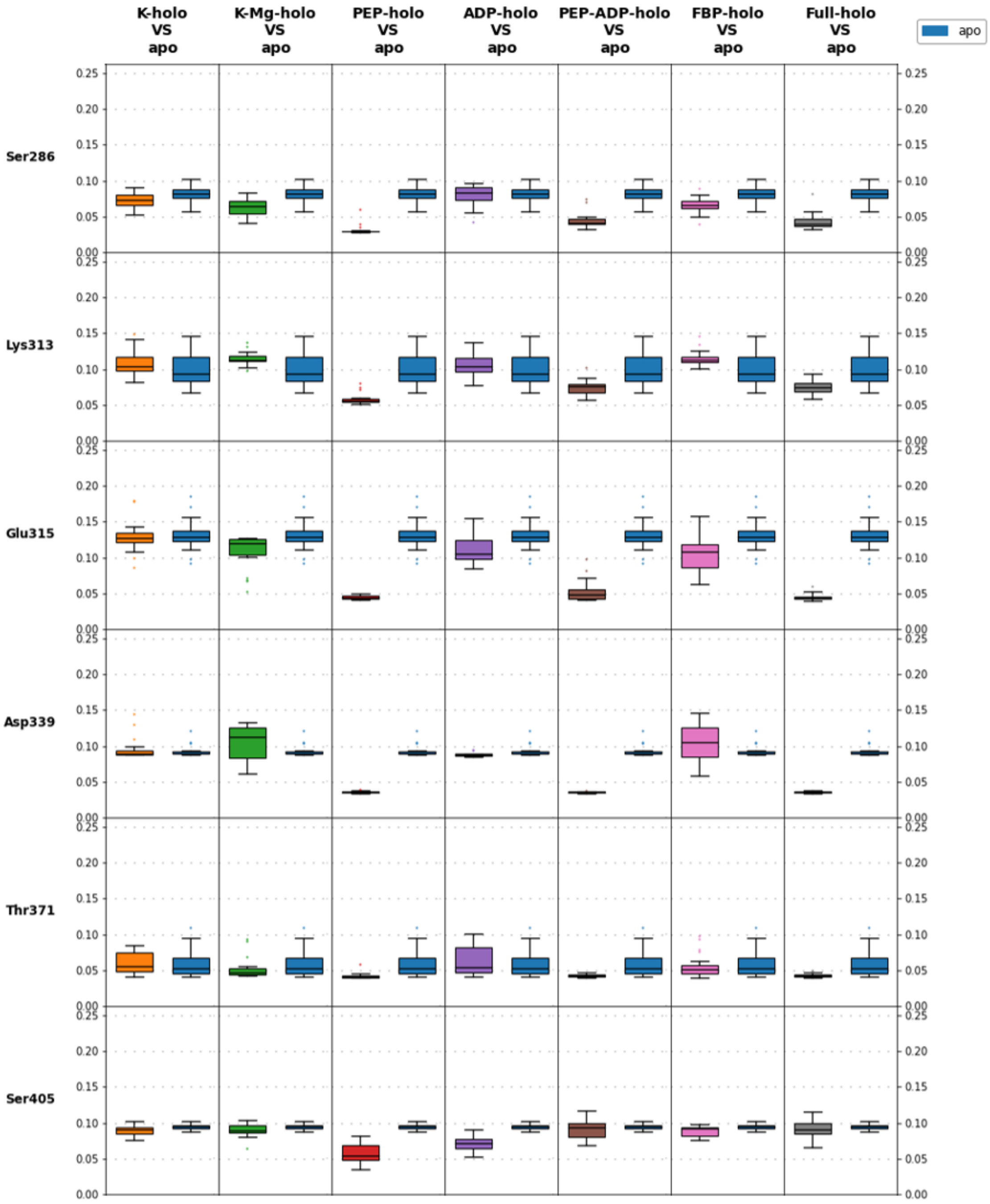


Figure 4.12. RMSF values (nm) of the amino acids of the cluster models along the MD simulations. Data is presented as a grid of box plots where rows correspond to each amino acid and columns correspond to the comparison of each holo condition with the apo condition. Each box plot contains the aggregated data of each PKR subunit in the 5 corresponding trajectory replicates of each condition. RMSF was calculated taking into account all side-chain atoms of the corresponding amino acid, using the backbone atoms as the fitting group. The data of each MD condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray.

Figure 4.12 (Continued)



In accordance with the RMSD analysis, in general, each residue in the apo simulations had higher fluctuation than in the holo simulations. This is especially true for the amino acids that coordinate K^+ with their side chain, namely Asn118, Ser120, and Asp156. Thr157 also coordinates K^+ but with the oxygen atom of its backbone. Therefore, it maintained equivalent RMSF values both in apo and holo conditions. As for the amino acids that coordinate Mg^{2+}, namely Glu315 and Asp339, they showed different behaviors depending on which Mg^{2+} binding site was modeled. With the Mg-1 site, they drastically reduced fluctuation, whereas with the Mg-3 site they had more variability, even displaying more fluctuation than in the apo condition in the case of Asp339. This observation correlates with the fact that these models showed a moderately higher flexibility in the RMSD analysis.

The rest of the amino acids of the cluster model (*i.e.*, not coordinated to the metals) display equivalent RMSF values between apo and holo conditions. As previously observed in the RMSD analysis, when the model contains the PEP molecule, the majority of amino acids become more rigid and thus adopt the lowest RMSF values with the smallest data variance. His121 behaves differently: it only decreases fluctuation when ADP is included in the active site, which proves that this amino acid is indeed involved in ADP binding.

4.1.2.3.3 Relative abundance of coordination numbers

Up to this point, the trajectory analyses have shown that the developed bonded parameters for the metal centers succeed in modeling stable structures that represent the holoenzyme conditions of the system. To supplement these results, the last analysis focused on evaluating the non-bonded portion of the metal-center parameterization approach. Specifically, the surroundings of the metal centers in simulation were explored to determine whether the coordination spheres are spontaneously completed or not, in accordance with the expected coordination sites. With this aim, a count was made of the number of instances in which other molecules can be found closer than a distance threshold with respect to the metals. From these counts, it was possible to determine which coordination-complex configurations have occurred along the simulations, along with their relative abundance.

Water molecules are mainly the expected coordination ligands, in accordance with the experimental evidence provided by the crystallographic structures. However, the simulation also contains ions in solution, thus they have been also considered as coordination ligand candidates. Finally, Glu161, even if not a putative coordination ligand of K^+, was confirmed to be able to interact with the metal for short spans of time, being close enough to consider this interaction as possible.

The chosen distance cutoff depended on the analyzed pair of metal and coordination ligand. Starting from the tentative guess of 3.5 Å, each type of cutoff was empirically adjusted until the number of reported false positives was minimized without compromising the true positives. A false positive could occur, for instance, when the oxygen atom of a water molecule is indeed found within the distance cutoff but in an improper orientation to interact with the metal. For the K^+ metal center the final cutoff values were the following: 3.1 Å for water (oxygen atom), 3.5 Å for Cl^-, and 3.1 Å for Glu161 (either of the oxygen atoms of its side-chain carboxylate group). For the Mg^{2+} metal centers: 2.8 Å for water (oxygen atom), and 3.5 Å for Cl^-. The *trjorder* tool from GROMACS was used to provide the occurrences in each trajectory snapshot.

By integrating the data from each type of probed coordination ligand, the different possible coordination-complex configurations or states were identified. Figure 4.13 shows the relative abundance of each state found in simulation, for each metal center in each holo condition. In the case of K^+, a few states with less than 0.1 % of abundance were dismissed because they corresponded to negligible convoluted configurations (for instance, the coincidental presence of water molecules, Cl^- and Glu161 at the same time, or similar variations) and are not included in the figure. The featured states are defined in the bottom table of each bar plot (12 for cofactor K^+, 5 for cofactor Mg^{2+}, and 5 for ADP-bound Mg^{2+}). The interaction of the side chain of Glu161 (symbolized $RCOO^-$) with K^+ was simplified as a single contact regardless of whether one or both oxygen atoms of the carboxylate group participate in the interaction. The figure contains the aggregated data of each PKR subunit in the 5 corresponding trajectory replicates of each condition.

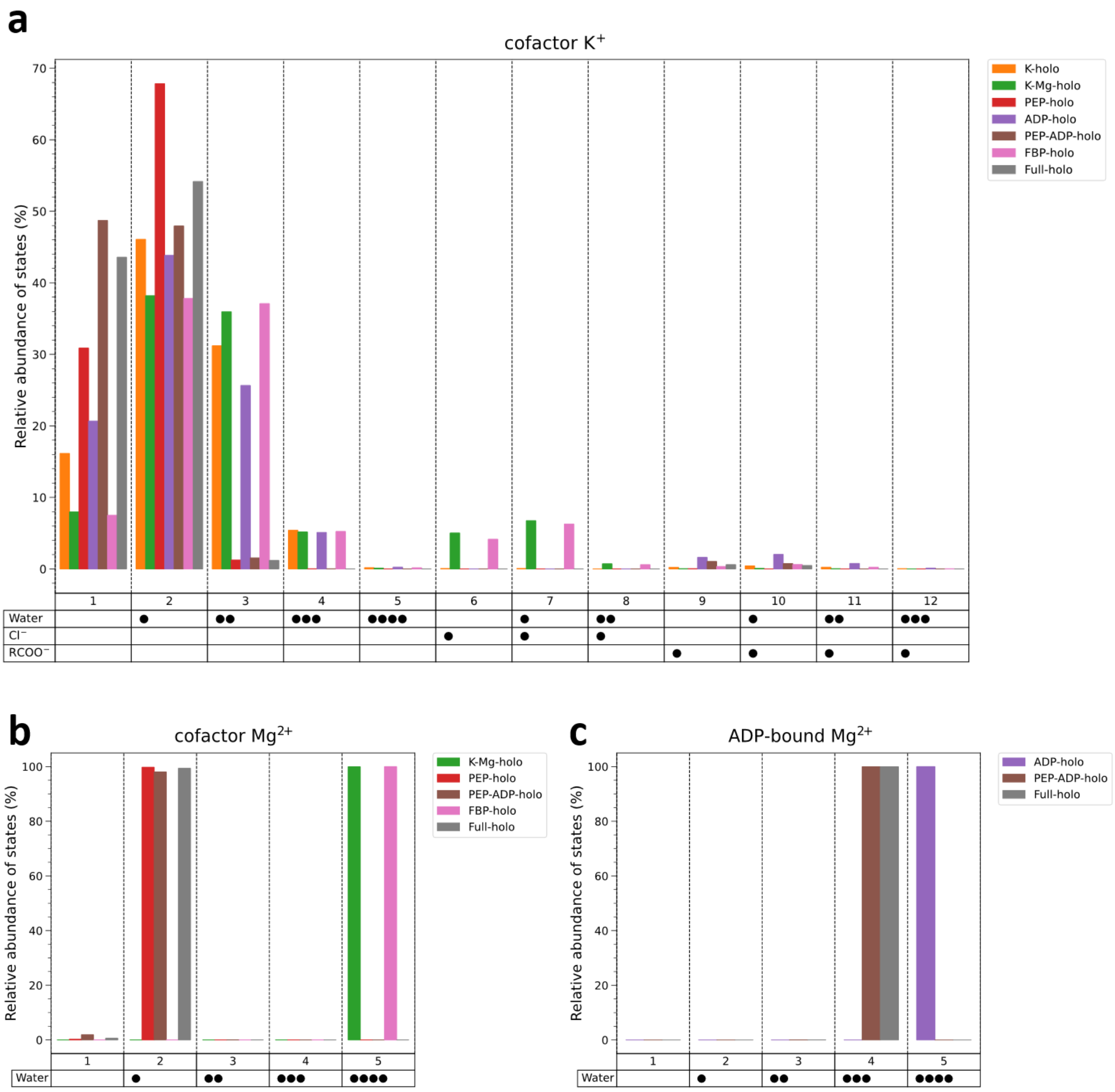


Figure 4.13. Relative abundance (%) of the states of transient coordination ligands of each metal center. The breakdown of the possible states is given in a tabular format, where rows specify the molecule species and columns contain the aggregate contents of each state. The number of molecule species (water, Cl^- ions or carboxylate groups ($RCOO^-$)) is indicated by the number of black circles in each cell. Then, each column defines the region of the horizontal axis that corresponds to each state in the bar plot. Each bar plot contains the aggregated data of each PKR subunit in the 5 corresponding trajectory replicates of each condition. (**a**) States of the metal center of cofactor K^+. (**b**) States of the metal center of cofactor Mg^{2+}. (**c**) States of the metal center of the ADP-bound Mg^{2+}.

Regarding the K^+ metal center (Figure 4.13a), in the absence of PEP the most abundant state was that with a single coordinated water molecule (38-46%). This observation matches the results of the geometry optimization of the “small model” of K-holo, in which one of the water molecules that was intended to coordinate K^+ moved away, leaving K^+ in a 5-coordinate state. However, this configuration was closely followed by the 6-coordinate state with 2 water molecules (25-37%). The state without water molecules was the third most abundant (8-21%), followed by the state with 3 water molecules (5%). The state with 4 water molecules is negligible.

Remarkably, the only conditions where Cl^- ions in solution were detected to occasionally bind to K^+ (4-7%) were those when the Mg-3 site was occupied (K-Mg-holo and FBP-holo). This observation could be related to the hypothesis of Zhong *et al*. [137] whereby the Mg-3 site could work as a priming mechanism to attract the next PEP molecule, which is negatively charged like Cl^-. Figure 4.14a shows an instance of the 6-coordinate K^+ metal center with a coordinated water molecule and Cl^-. When the model included PEP without ADP, the most abundant state (68%) was that with one coordinated water molecule (therefore, a 6-coordinate state). However, the state without water also frequently occurred (30%). With both PEP and ADP, both states occurred with a similar frequency (44-54%). All in all, the K^+ metal center has been found to be stable both with 5 or 6 coordination ligands.

In addition to the coordinated water molecules, the exploration of the structures in this analysis also allowed capturing in simulation the stabilizing water molecule near the K^+ metal center that was modeled in the PEP-holo condition following the experimental evidence of structure 4HYV (Figures 4.2c and 4.3g-i). This water molecule makes H-bonds with Asn118, Ser120, and the phosphate group of PEP, and can be seen in Figure 4.14b.

In the presence of ADP, the states of Glu161 coordinated to K^+ become non-negligible (1-3%). This could be related to the fact that the B domain of the protein tends to adopt the fully closed conformation when ADP/ATP is bound, therefore Glu161 would have more chances to be found close to K^+. Figure 4.15a-b shows a sequence of two close snapshots in simulation where firstly Glu161 is coordinated to K^+ and later it abandons its coordination site as two water molecules nearby establish new coordination bonds with K^+. Figure 4.15c shows an instance of the K^+ metal center with one coordinated water molecule and a bidentate coordination of Glu161.

Regarding the Mg^{2+} metal centers (Figure 4.13b-c), no coordinated Cl^- were ever detected; just water molecules. There is almost no variability of the configurations, with the 6-coordinate states being the predominant (other configurations with coordination number less than 6 are negligible). The Mg-3 site always had 4 coordinated water molecules and adopted an octahedral geometry. An instance of this metal center can be seen in Figure 4.14a. The Mg-1 site nearly always was found with 1 coordinated water molecule and adopted an octahedral geometry, except in a few snapshots where such coordination site was empty (1%). An instance of the predominant state of this metal center can be seen in Figure 4.14b. Incidentally, the figure also features a Na^+ ion that is interacting with PEP beside the Mg-1 site, thus illustrating that the parameterization approach is compatible with the modeling of spontaneous interactions with free ions in solution. Lastly, the ADP-bound Mg^{2+} always had 3 or 4 coordinated water molecules in the presence or absence of PEP, respectively, and adopted an octahedral geometry.

Finally, these results are alternatively reported in the form of the relative ΔG (kJ·mol^{-1}) between the states of each metal center in each holo condition. To perform the calculations, Equation 3.4 was employed, setting the most abundant state in each case as the reference state. Such results are shown in Figure 4.16. An interesting remark is that in the K^+ metal center a similar pattern of relative stabilities can be seen between the states according to both the number of coordinated water molecules and the number of coordinated Cl^- or $RCOO^-$.

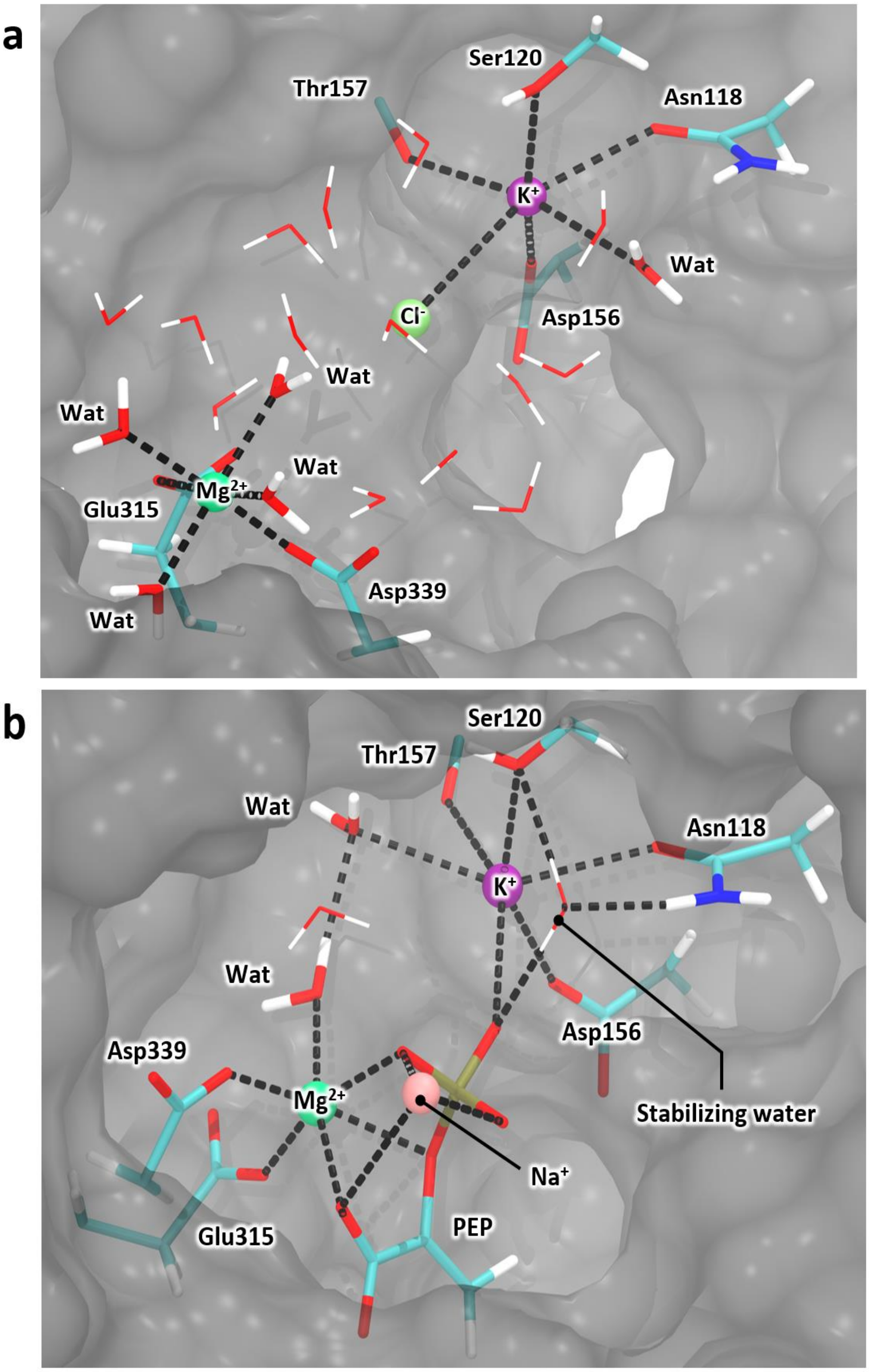


Figure 4.14. Instances of 6-coordinate states of the metal centers of cofactors K^+ and Mg^{2+} in MD simulation. The figure shows the components of the coordination complexes (in licorice or sphere representations) as well as other water molecules and ions that surround the metal centers (in line or sphere representations). The relevant residues are labeled, and their components are colored according to atomic species. The rest of the local region of the protein is included with a gray surface representation. All coordination bonds and H-bonds are shown as black dashed lines. (**a**) A Cl^- ion in solution coordinates K^+. The Mg^{2+} metal center is located at the Mg-3 site. (**b**) The stabilizing water molecule of the K^+ metal center was captured in action. The Mg^{2+} metal center is located at the Mg-1 site and features a Na^+ ion in solution interacting with PEP. NOTE. Abbreviations: Wat, water. The images were generated with the software VMD.

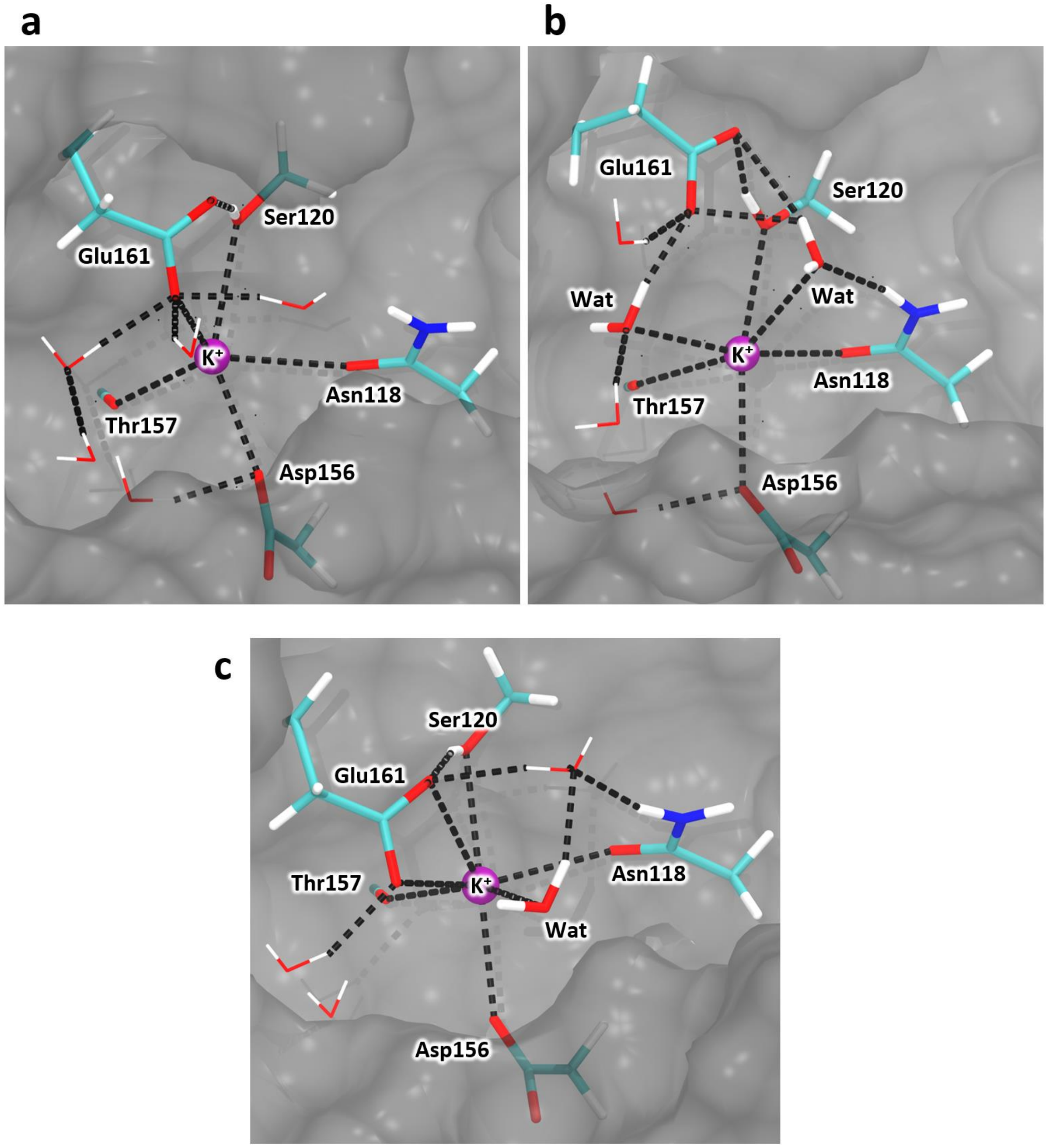


Figure 4.15. Instances of different states of transient coordination ligands of the metal center of cofactor K^+ in MD simulation. The figure shows the components of the coordination complex (in licorice or sphere representations) as well as other water molecules that surround the metal center (in a line representation). The relevant residues are labeled, and their components are colored according to atomic species. The rest of the local region of the protein is included with a gray surface representation. All coordination bonds and H-bonds are shown as black dashed lines. (**a-b**) A sequence of two close snapshots in simulation where firstly Glu161 is coordinated to K^+ and later it abandons its coordination site as two water molecules nearby establish new coordination bonds with K^+. (**c**) K^+ is seen with one coordinated water molecule and a bidentate coordination of Glu161. NOTE. Abbreviations: Wat, water. The images were generated with the software VMD.

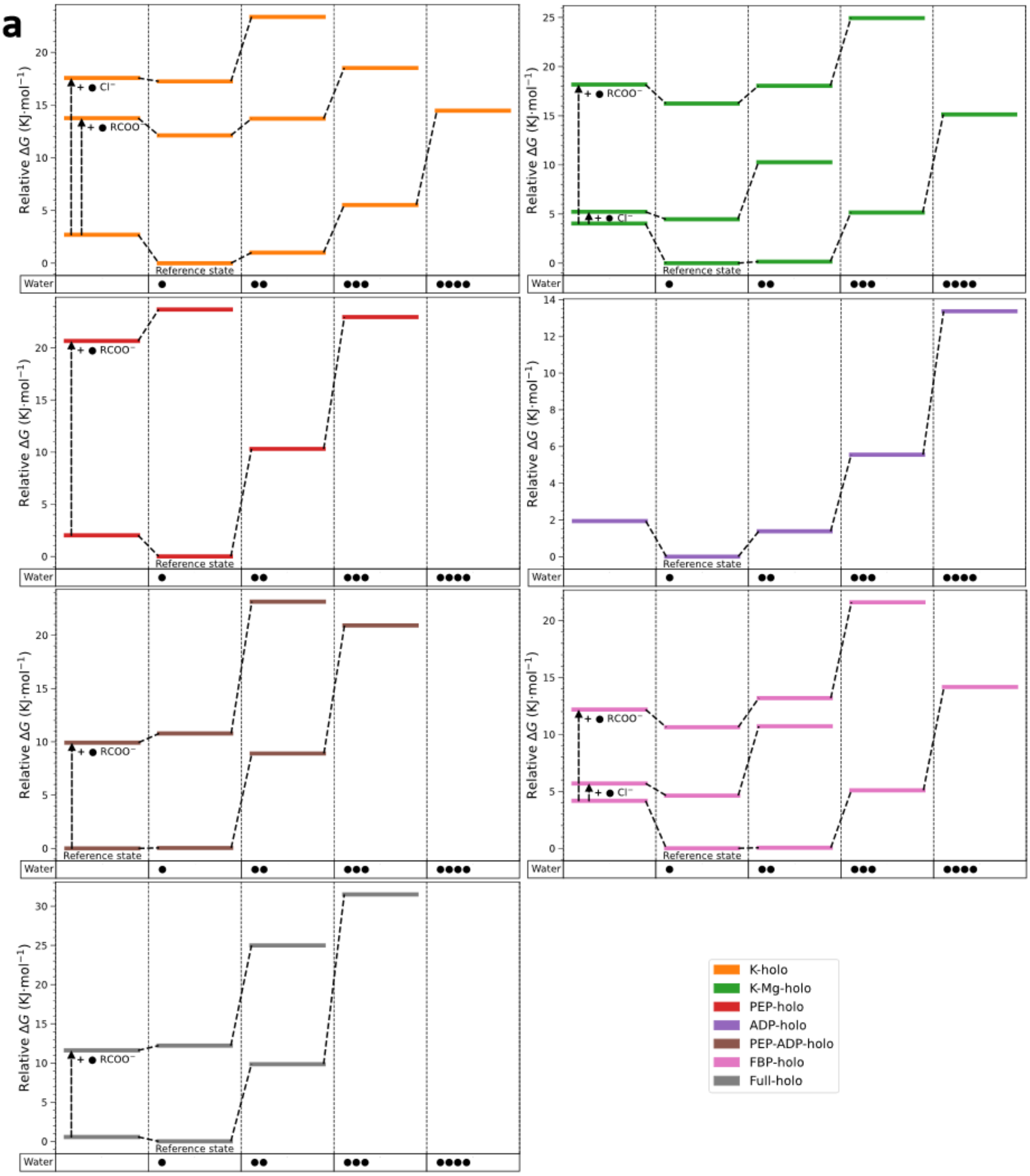
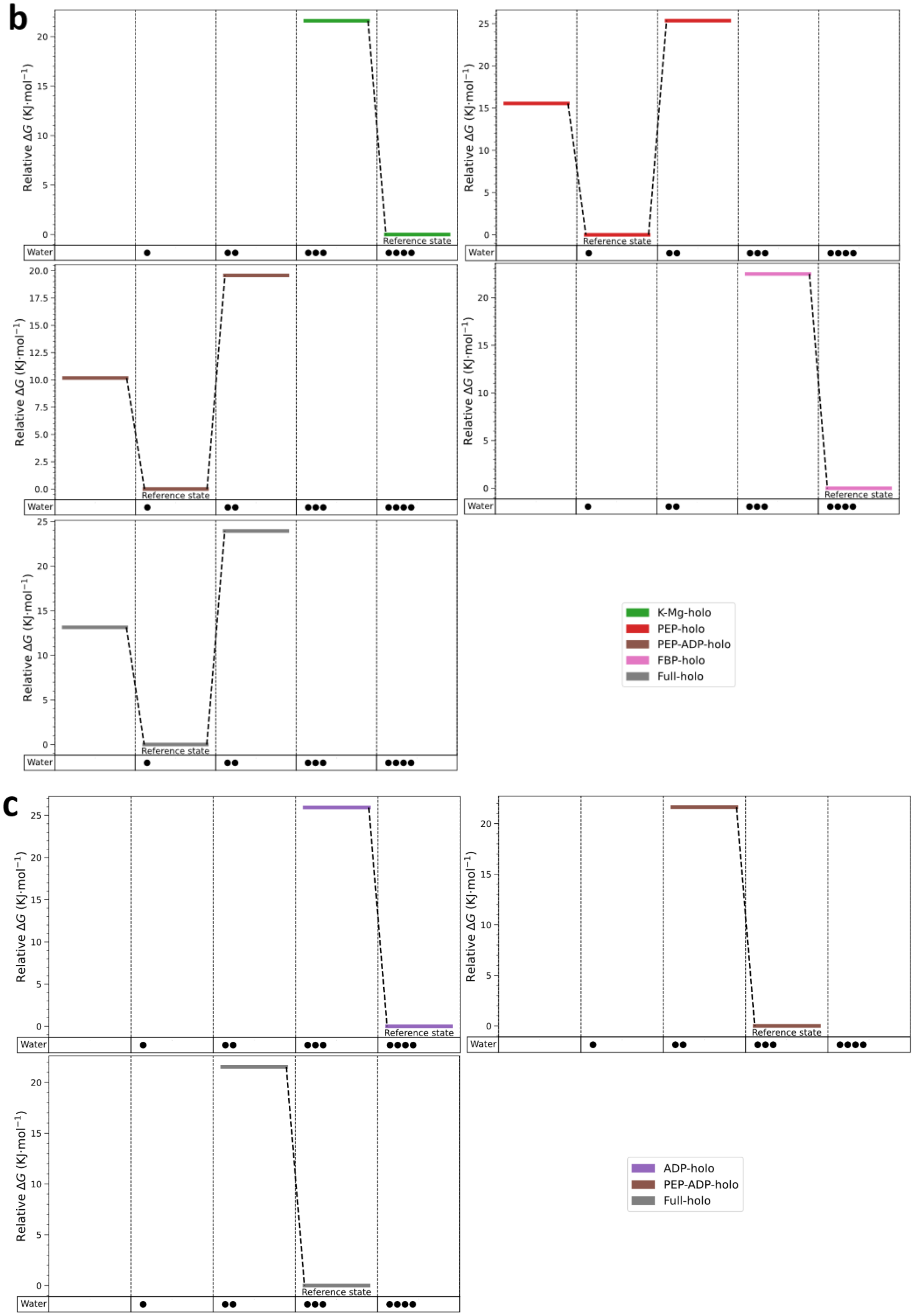


Figure 4.16. Relative free-energy differences between the states of transient coordination ligands of each metal center. The horizontal axis shows the possible states according to the number of coordinated water molecules (as indicated by the number of black circles), whereas the states that incorporate either Cl^- ions or carboxylate groups ($RCOO^-$) are indicated by the vertical dashed arrows. Each plot contains the aggregated data of each PKR subunit in the 5 corresponding trajectory replicates of each condition. (**a**) States of the metal center of cofactor K^+ in separate plots per holo condition. (**b**) States of the metal center of cofactor Mg^{2+} in separate plots per holo condition. (**c**) States of the metal center of the ADP-bound Mg^{2+} in separate plots per holo condition.

Figure 4.16 (Continued)



4.1.3 Consensus Essential Dynamics Analysis

Following the structural and geometric analyses of the trajectories, we now delve into the extraction and analysis of actual dynamical features. The results presented in this section constitute the outcome of the central endeavor of the thesis. The conducted experiments are based on the traditional methodology of trajectory analysis known as Essential Dynamics Analysis (EDA). In the paradigm of EDA, the underlying collective motions of the trajectory are explored through the application of a dimensionality reduction technique, namely, Principal Component Analysis (PCA). The aim of this procedure is to ascertain potential mechanistic interpretations of the observed motions on the basis of the known biological function of the protein.

As detailed in the Methods chapter, PCA can be employed to determine a subset of collective variables (CVs) that capture the maximum amount of variance of the dataset. These CVs extend along directions that are mathematically characterized by an orthonormal basis of eigenvectors. In turn, each eigenvector comes with its respective eigenvalue, which quantifies the relative portion of variance captured along the direction of the vector. The new coordinates are called the Principal Components (PCs). The projection of the original data of the trajectory onto a reduced number of PCs reveals the relevant correlated atomic displacements along the directions of the dataset.

One of the main motivations of this thesis is to investigate the utilization of EDA as the basis for a novel methodology that facilitates the integration of the PCA output from independent trajectories of a system within a unified framework. This approach enables the characterization of the dynamical traits of the biomacromolecule under examination that are broadly manifested among the replicates of a single reference condition. In turn, it provides quantitative indicators for comparison of such reference traits with alternative conditions (*e.g.*, apo *vs.* holo, or WT *vs.* mutant variants). The method has been termed Consensus Essential Dynamics Analysis (CEDA).

The CEDA strategy consists in deriving a set of Consensus Principal Components (CPCs) by applying a clustering algorithm on the most relevant eigenvectors obtained from the standard PCA of each individual trajectory. The centroid vectors of the resulting clusters bear the CPCs of the trajectory ensemble. CPCs accentuate the common qualities of the collective motions described by the members of the cluster, such that: i) the predominant fraction of the shared collective motion is maintained or emphasized, and ii) the minor variations displayed only by sporadic cases are filtered out or dampened. The analysis of the projections of the involved trajectories to the single consensus set facilitates the comparison of the different simulation conditions in a consistent way.

The implementation of this method is progressively presented with more detail throughout this section as the corresponding results of the analysis of PKR are reported. With the following experiments, the aim was to design and test the protocol based on its conceptual idea while simultaneously conducting the study of the dynamic behavior of the protein in a manner similar to traditional EDA. The corresponding analyses were applied on the portion of each trajectory that excludes the initial relaxation phase of 25 ns that was determined in section 4.1.2.1. Thus, the analyzed trajectories have a length of 375 ns each. To illustrate and discuss the collective motions of interest, schematic representations of the analyzed structures are provided, depicting several of the relevant conformations along the path of the motions. To facilitate a more detailed inspection of the motions, supplementary videos have been generated and are available as part of the supplementary material of this thesis (see Appendix A).

4.1.3.1 Collective motions of domains within the PKR monomer: the A and B domains

PKR is a modular enzyme that does not only fold into distinct domains but also oligomerizes into a symmetric tetramer. This fact implies that the function of PKR does not solely rely on the local arrangement of catalytic residues at the active site, but also on the orchestrated intercommunication between modules and functional sites. This degree of structural and dynamical complexity is what precisely confers the enzyme with its full range of allosteric capabilities.

The speculated conformational transitions between the T (inactive) and the R (active) states of the enzyme (Figure 1.12 from the Introduction chapter) embody this behavior. The structural rearrangements traditionally described in the proposed models have been found by comparison between static structures of pyruvate kinases, and are a summarized version of the actual dynamical events taking place. Assessing the mechanism of transmission of dynamical information at the level of each decomposed motion is critical to having a complete view of the function of the enzyme. Domains in contact, via either flexible hinges or interfaces, undergo the motions of individual interest that can be studied to reveal different features of the dynamical behavior of the protein (symmetrical *vs.* independent transitions, rigid-body motions *vs.* internal rearrangements).

In particular, within the monomeric structure of PKR, domains are inherently able to move relatively to each other as suggested by the high flexibility of the joint regions according to the RMSF analysis (section 4.1.2.2). The monomer represents the basic unit of the architecture of the protein. Therefore, the characterization of such motions is a crucial starting point towards the general understanding of the function of the enzyme coupled to dynamics.

The collective motions between the A and B domains are of high relevance as they imply direct changes in the environment of the active site. Trajectory analyses via RMSD and RMSF have shown that the B domain tends to exhibit the highest structural mobility of the monomer, with its dynamical behavior being potentially uncoupled between subunits. These features make this region a suitable candidate to begin extracting dynamical features and establishing a comparative framework to assess the persistence or absence of collective motions between trajectory replicates and conditions.

4.1.3.1.1 Derivation of CPCs of the WT apo condition

The reference condition for extracting the set of CPCs was the WT apoenzyme. Accordingly, PCA was applied to each trajectory of this condition. The analysis was enriched by treating all individual monomeric instances of the A and B domains as separate trajectories of the same system, due to their symmetrical locations within the tetramer and their analogous role from a dynamical point of view. Therefore, from the 5 simulations of the WT apo tetramer, 20 actual trajectory replicates of the A and B domains (one per monomer) were available and employed. Throughout this section, each trajectory replicate will be designated by the number of the simulation (#1 to #5) and the chain ID of the monomer (A, B, C, or D) (for example, replicate #3-D refers to the trajectory of the subunit with chain ID “D” of the third simulation of the tetramer).

Prior to the covariance calculation, an indispensable step in CEDA involves aligning the structures of all the analyzed trajectories to a single reference structure. This ensures that the structural coordinates of all the sampled conformations can be equivalently expressed in terms of either the CVs or the original three-dimensional Cartesian coordinates with a common mathematical transformation, applicable to all trajectories. The average structure of all monomeric trajectories of PKR was generated and used as the general reference for the structural superposition (least-squares fitting). The A domain was chosen as the fitting group for the removal of the rotational and translational components. Due to its rigidness, the A domain constitutes a stable reference region to orient all trajectories and facilitate inspection of the relative motions of the B domain with respect to the A domain. Only the Cα atoms of the structure (347 atoms) were considered in the analysis, therefore each PCA yielded a set of 1041 eigenvectors.

Subsequently, a clustering analysis was conducted with the pool of eigenvectors generated from the multiple PCAs. A primary decision at this stage involved determining the quantity of eigenvectors to incorporate into the clustering analysis. It is crucial to import a minimum threshold of eigenvectors from each trajectory that ensures the acquisition of a significant fraction of data variance. Thus, the chances of detecting similar collective motions between trajectories will be maximized, irrespective of whether they appear in the same PC indices. On the other hand, the inclusion of an excessive number of eigenvectors may introduce noise and unnecessarily slow down the calculation. To address this question, the percentage of variance explained by the PCs was explored. Figure 4.17 shows the mean values of variance (individual and cumulative) explained by PCs #1 to #30 of the 20 trajectory replicates.

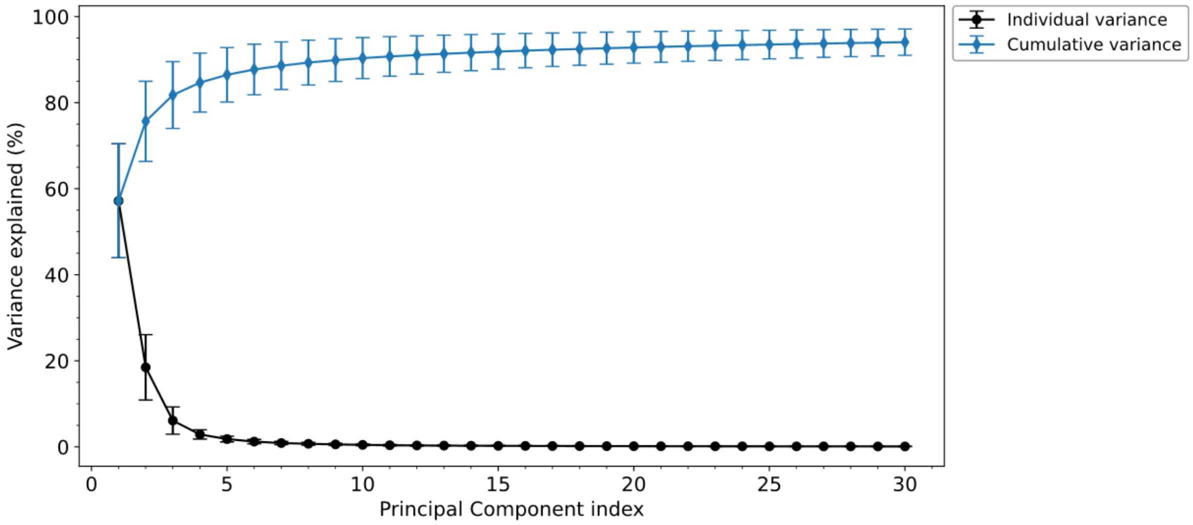


Figure 4.17. Percentage of variance explained by the PCs of the A and B domains in the WT apo condition. Data corresponds to the mean values from the first 30 PCs of the 20 trajectory replicates of the experiment. Standard deviation intervals are included as error bars.

The figure suggests that the third PC may mark the "elbow" point beyond which the subsequent PCs only provide marginal information. PCs #1 to #3 accounted for around 80% of cumulative variance, which is typically an acceptable cutoff in EDA studies. However, a greater number of eigenvectors were incorporated in the clustering analysis to determine whether analogous minor collective fluctuations among replicates occur and to assess the sensitivity of CEDA in detecting them. The following two criteria were established. First, a minimum of 6 eigenvectors per replicate were included, given that PC #6 is the last PC that individually captures more than 1% of variance and the first that accumulates more than 85% of variance on average. Second, the remaining eigenvectors were also included until either gathering a 95% of cumulative variance or reaching a maximum of 20 eigenvectors per replicate.

The cosine content of the PCs of all the selected eigenvectors was calculated. This measure provides a qualitative indicator for determining whether sufficient sampling time has been achieved to ensure a reliable physical significance of PCs. The cosine content value expresses the resemblance of the variation of PC values along the trajectories to cosine-shaped curves, ranging from 0 (no cosine shape) to 1 (perfect cosine shape). High cosine contents are indicative of PCs describing protein dynamics in barrier-free diffusive events less meaningful than the biologically relevant atomic displacements. In protein MD simulations, a cosine content lower than 0.5 is considered to be acceptable to proceed with EDA [334], [397], [398], [500].

Figure 4.18 shows that the vast majority of PCs exhibit cosine content values below 0.5, suggesting that reasonable local conformational sampling of the A and B domains was achieved and, thus, PCs reflect biologically relevant dynamical features. Only two instances of high cosine content were found. These correspond to the PC #1 of trajectory replicates #4-B and #5-D, with values 0.77 and 0.57 respectively. This fact does not entail their automatic exclusion from further analysis, since even when sufficient sampling quality is achieved, PCs may display cosine-like shapes by chance if they describe an actual transition of the system from one state to another. This possibility should not be considered if high cosine content values had been obtained in a generalized manner, which would indicate poor sampling and therefore require caution when continuing the analyses. The significance of the two PCs of high cosine content shall be assessed in the light of the clustering analysis, which may clarify whether they represent meaningful collective motions comparable to those observed in other trajectory replicates.

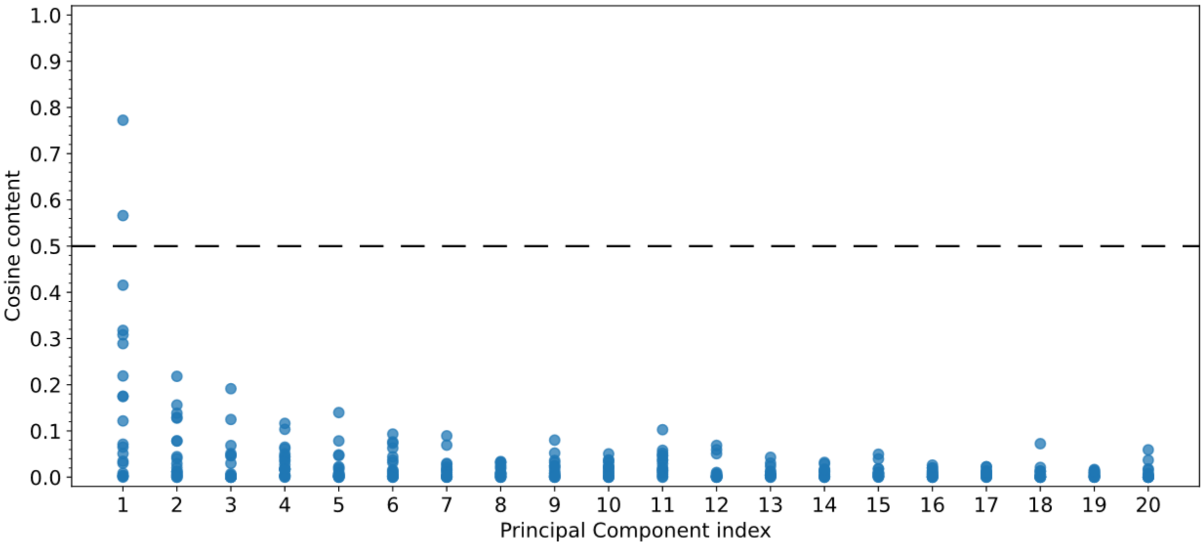


Figure 4.18. Cosine content values of the PCs considered in the CEDA of the A and B domains in the WT apo condition. The cutoff value of 0.5 below which the sampling is considered to be acceptable in protein dynamics is shown with a horizontal dashed black line.

To perform the clustering, a dissimilarity matrix between pairwise eigenvectors of the analysis was generated employing the cosine distance (derived from the calculation of the cosine similarity in absolute value; thus, bounded in the interval [0, 1]). An agglomerative hierarchical clustering algorithm was applied to the dissimilarity matrix, using the average-linkage method to calculate the distance between clusters in each combination step and construct a dendrogram. A cophenetic distance cutoff of 0.4 was applied to split the dendrogram and obtain clusters. Finally, from the resulting clusters, those with at least 20% of coverage (*i.e.*, with a representation of at least 4 out of the 20 trajectory replicates that participate in the clustering) were retained, yielding a total of 10 relevant clusters that describe a consensus behavior. The other minor clusters were discarded. Figure 4.19 shows the corresponding dendrogram, highlighting the important elements of the process to facilitate visualization of the results.

A primary exploration of the dendrogram reveals two distinctive main families of clusters. The first family, located in the lower section of the dendrogram, comprises 3 large clusters (clusters #2 to #4) and a few eigenvectors that do not form relevant clusters. These 3 major clusters are of particular significance, as they include representation of all or most trajectory replicates and exhibit high similarity values. They emerge prominently at low values of cophenetic distance, ranging from 0.02 to 0.15, in the form of smaller clusters that rapidly expand and merge into larger ones with increasing cophenetic distance.

Clusters #3 and #4 hit their maximum sizes, consisting of 20 and 19 members, at values of approximately 0.24 and 0.27 respectively. In terms of coverage, the former is the only cluster that achieves a 100% coverage (*i.e.*, it is composed of exactly one eigenvector instance from each of the 20 trajectory replicates), suggesting the presence of an ubiquitous collective motion. The latter reaches 95% of coverage, meaning that the corresponding collective motion is also predominant but potentially absent in one trajectory, namely, replicate #1-A. Cluster #2 is less compact and consists of 17 members. It starts forming at 0.15 and continues to incorporate members until approximately 0.37, at which point it has an 85% of coverage and lacks participation from trajectory replicates #1-B, #2-A, and #3-D. Notably, cluster #2 could gain three additional members by merging with cluster #1 (a minor cluster with 15% coverage) at approximately 0.44. However, the newcomers would only incorporate representation from the first two absent trajectories, while a trajectory already present in the cluster would double its representation. For this reason, the merged version of the cluster would only reach 95% of coverage instead of 100%. The repeated trajectory is replicate #1-A, which incidentally is the absent trajectory at cluster #4.

The second family of clusters of the dendrogram is characterized by scattered eigenvectors that only form smaller clusters of both lower similarity and coverage values. Cluster #23 has 35% of coverage, while clusters #24, #26, #28, #29, #31, and #34 have exactly 20% of coverage. This section of the dendrogram is sensitive to readjustments in the criteria for selecting the relevant clusters. For instance, increasing the cophenetic distance cutoff to 0.45 would result in several changes: i) the minor clusters #11 and #12 would merge and comply with the minimum size and coverage to become relevant; ii) cluster #23 would gain a new member, increasing its coverage from 35% to 40%; and iii) clusters #28 and #29 would merge, combining their individual 20% coverage into 40%. Conversely, by decreasing the cutoff to 0.35, clusters #24 and #34 would no longer be relevant, while cluster #23 would lose a member that would reduce its coverage from 35% to 30%. A more restrictive cutoff of 0.3 would filter out all clusters except #23, which would still retain a 20% coverage. Due to such instability in the number of relevant clusters in this section of the dendrogram, the criterion was simply to establish the cophenetic distance cutoff that best elucidates the 3 main clusters of the first family. As described above, 0.4 is a cutoff that optimizes the maximum coverage of cluster #2 without incorporating redundant provenances.

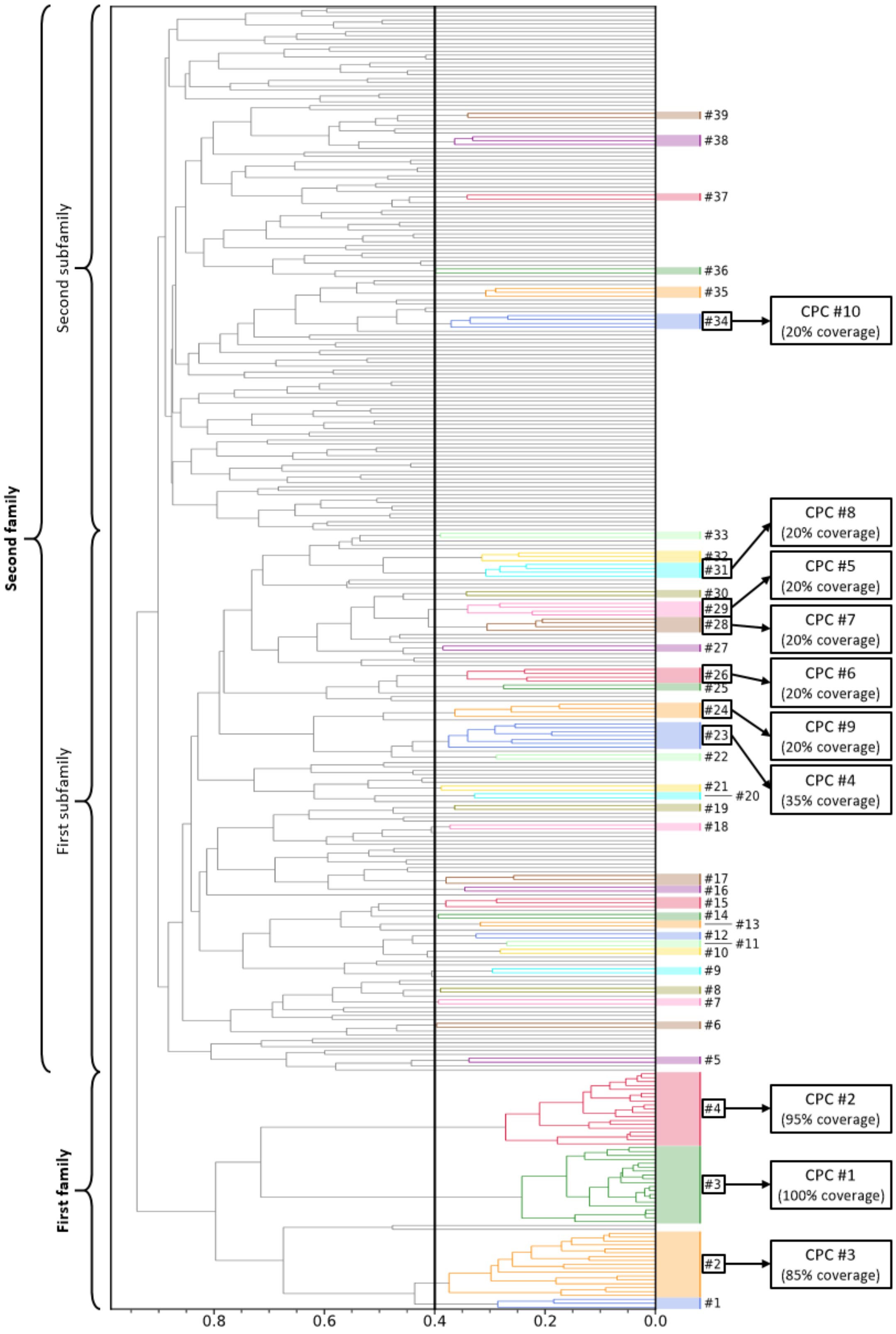


Figure 4.19. Dendrogram of the CEDA of the A and B domains in the WT apo condition. The cophenetic distance cutoff of 0.4 used to split the dendrogram and obtain clusters is shown with a vertical black line. Non-singleton clusters are shown in various colors to facilitate visual identification. The 10 clusters with at least 20% of coverage were selected to acquire the Consensus Principal Components (CPCs) of the experiment. CPCs were numbered from #1 to #10 in decreasing order of coverage and average percentage of variance of their cluster members. The figure indicates which clusters yielded each CPC, together with their achieved coverage. At the left margin of the dendrogram, the span of the detected families and subfamilies of clusters is indicated with labeled curly brackets.

On the other hand, the minimum percentage of coverage is also a subjective parameter that highly influences the number of retained clusters of the second family. With a coverage greater than 20% as a requirement for consensus behavior, 6 clusters would be discarded. In contrast, by lowering the minimum coverage to 15%, 5 additional clusters would be retained. In this case, the cutoff was finally set to 20%, which allowed the exploration of a reasonable number of consensus collective motions. The only trajectories that had no representation in any cluster of the second family were replicates #1-A and #4-B.

The next step in CEDA involved computing the centroid (average) vector of each selected cluster. Vectors within the same cluster may point in opposite directions, describing similar collective atomic displacements albeit reversed. Therefore, this step required choosing a reference direction in each cluster and then flipping all opposite vectors before computing the corresponding centroids. The 10 centroid vectors resulting from this experiment represent the set of CPCs of the 20 trajectory replicates of the A and B domains of the WT apo condition. CPCs were numbered from #1 to #10 in decreasing order of coverage and average percentage of variance of their cluster members.

The collective motion described along each CPC was examined by projecting trajectory data onto the centroid vectors. Besides the structural superposition to the global average structure of the A and B domains with the former as the fitting group (applied previously, in the PCA stage), the inspection of the CPCs also requires a data centering of the aligned trajectories around the very same average structure so that it represents the origin of the system of coordinates.

Characterizing and conveying the observed motions requires establishing a supporting terminology to unambiguously refer to: i) the involved parts/regions of the structure, ii) the referential points of views from which they are observed, and iii) the nature and path of the movement. For that purpose, certain terms and conventions from descriptive geometry (reference system of 3D views) and mechanics (mechanical degrees of freedom of movement of rigid bodies) were adopted. This terminology was complemented with analogies to real-world objects or situations, or references to the actual function of the motion in the context of the protein, which serve to summarize the set of technical terms into a plainer language. Figure 4.20 shows schematic diagrams of the A and B domains with the terminology that will be employed in this section. The mechanical degrees of freedom of the B domain as a rigid body describe its primary rotational (yaw, roll, and pitch) and translational (heave, surge, and sway) directions of motion. Three main points of reference (top, side, and front views) will be used to orient the protein region and show images of its structure. On the basis of this terminology, Figure 4.21 shows schematic representations of the path and extreme conformations of the collective motions from CPCs #1 to #10. Finally, these motions can also be viewed in the Supplementary Videos S4.1 to S4.10.

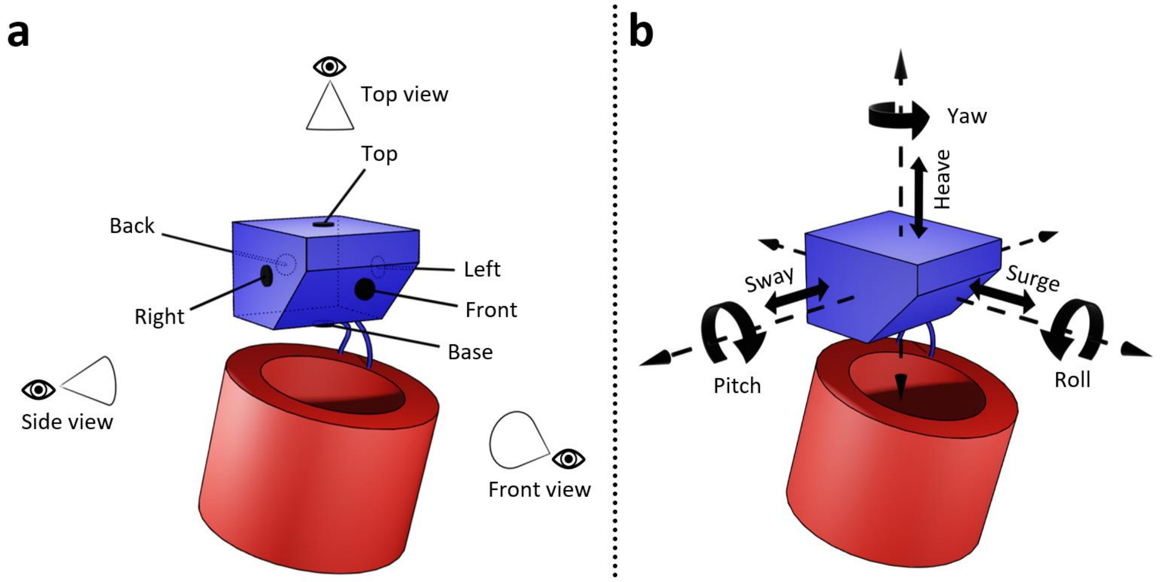


Figure 4.20. Schematic diagrams of the A and B domains of PKR. Various markers indicate the terminology to refer to (**a**) the regions, the reference views and (**b**) the directions of motion along the six mechanical degrees of freedom of the B domain as a rigid body. The A and B domains are colored in red and blue, respectively. NOTE. The 3D schematic models were built with the software Blender [136].

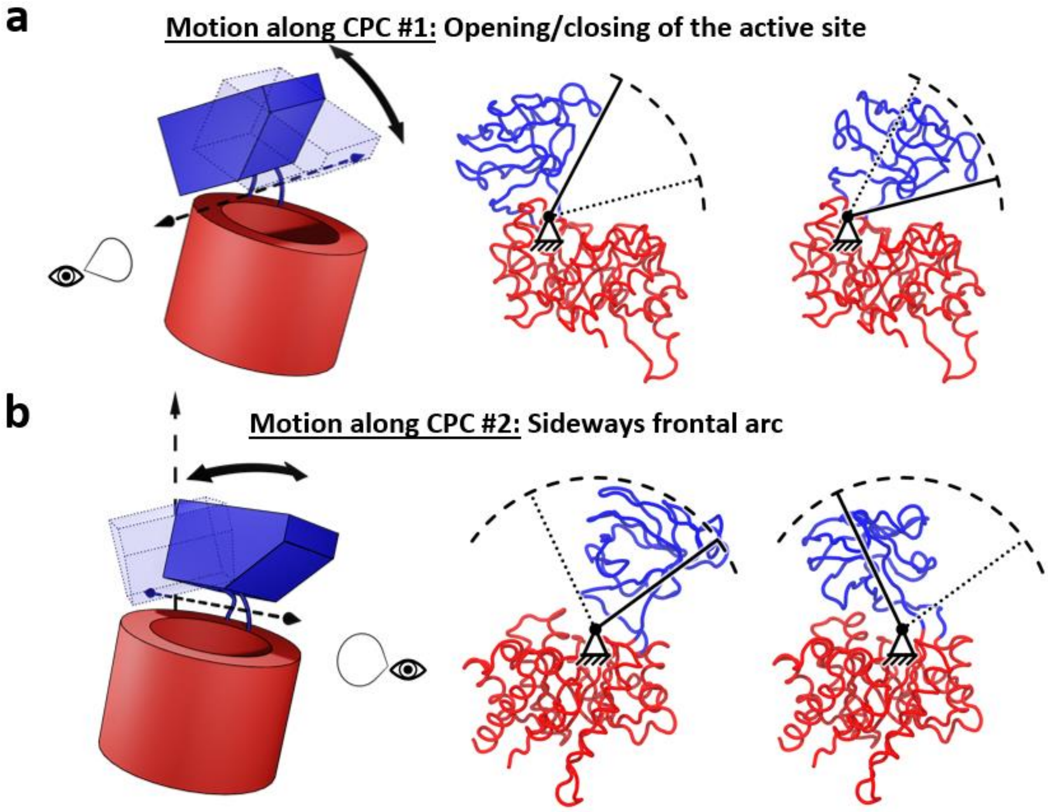
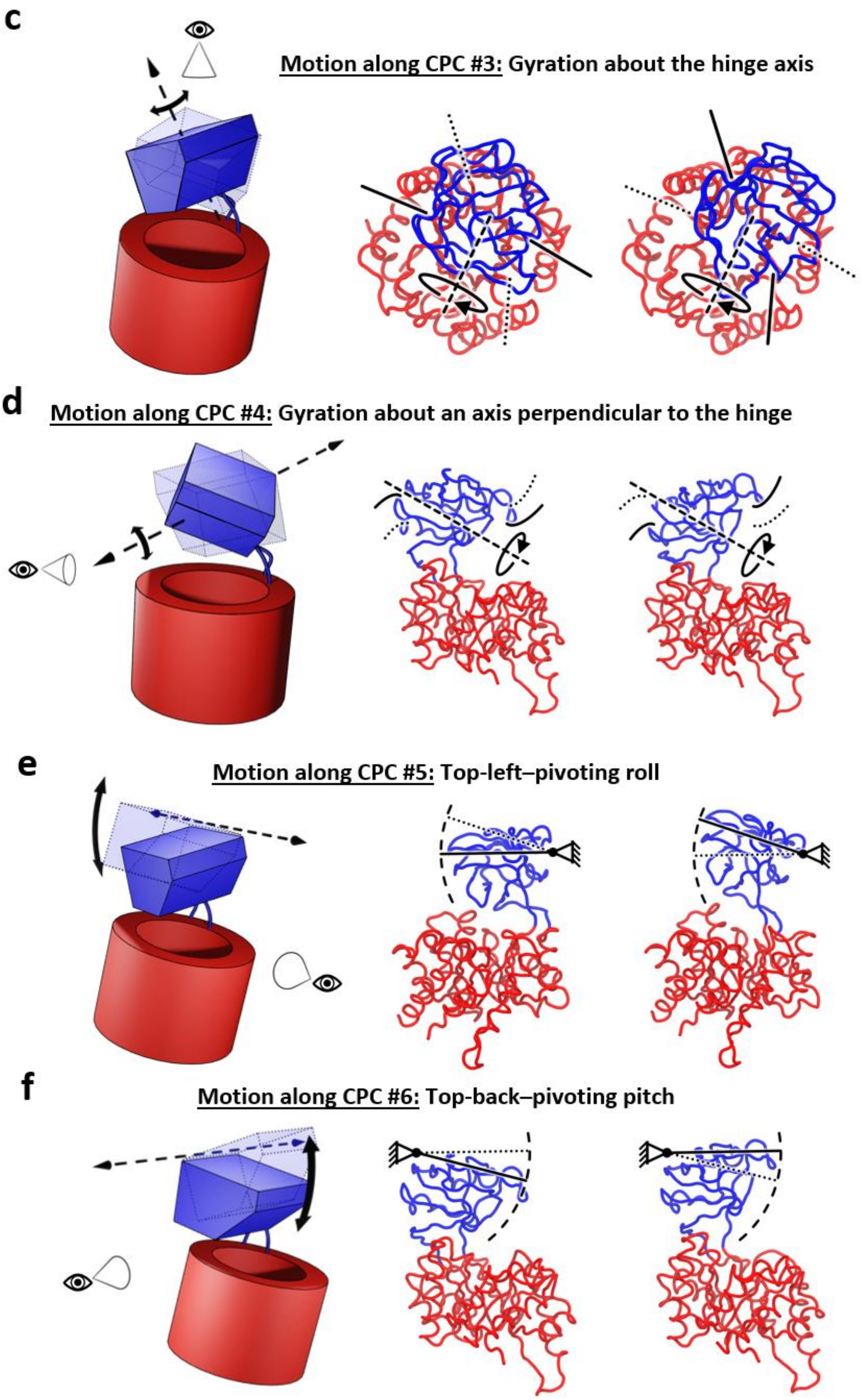

a
Motion along CPC #1: Opening/closing of the active site
b
Motion along CPC #2: Sideways frontal arc

Figure 4.21. Consensus collective motions of the A and B domains in the WT apo condition. (**a-j**) Schematic representations of the path and extreme conformations of the motion captured in each CPC from #1 to #10. In each panel, the protein domains are depicted with the diagrams of Figure 4.20 (left) and with the trace representation between Cα atoms (center and right). The diagram representation and markers of motion are omitted in (j), which instead shows a few intermediate conformations of the L-Bβ5-Bβ6 loop in transparent gray. The A and B domains are colored in red and blue, respectively. These motions can also be viewed in the Supplementary Videos S4.1 to S4.10. NOTE. The images of the protein structure were generated with the software VMD. The 3D schematic models were built with the software Blender.

Figure 4.21 (Continued)



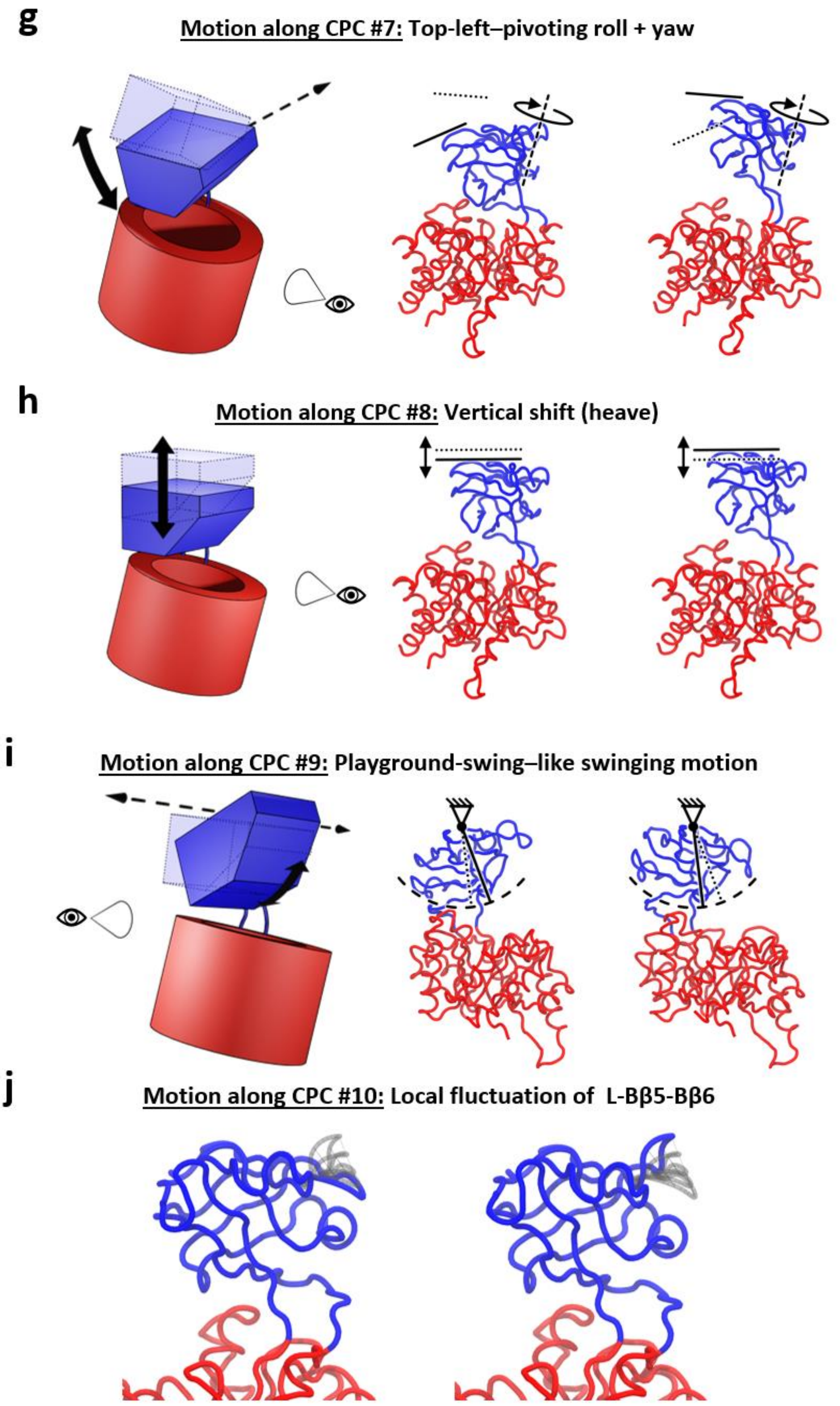
Motion along CPC #3: Gyration about the hinge axis

Motion along CPC #4: Gyration about an axis perpendicular to the hinge

Motion along CPC #5: Top-left–pivoting roll

Motion along CPC #6: Top-back–pivoting pitch

Figure 4.21 (Continued)



Motion along CPC #7: Top-left–pivoting roll + yaw

Motion along CPC #8: Vertical shift (heave)

Motion along CPC #9: Playground-swing–like swinging motion

Motion along CPC #10: Local fluctuation of L-Bβ5-Bβ6

CPCs #1, #2, and #3 arose from the three major clusters of the experiment, located in the first family of clusters of the dendrogram (see Figure 4.19). These clusters agglomerated the eigenvectors with higher eigenvalues, mostly corresponding to PCs with indices #1 to #3. The collective motions captured in these CPCs have large amplitudes of motion and reveal the space of highest conformational freedom of the B-domain oscillations. The collective motion captured by CPC #1 (Figure 4.21a, Supplementary Video S4.1) shows the B domain undergoing a combination of pitch and surge that results in a hammer-like swinging motion. This movement coincides with the classically described transition between the ligand-bound and -unbound states in crystallographic structures, namely, the opening/closing of the active site with the B domain acting as a lid to cover or uncover the top of the A domain. The collective motion captured by CPC #2 (Figure 4.21b, Supplementary Video S4.2) shows a combination of yaw, roll, and sway that results in the B domain describing an arc from side to side with its frontal region, also ascending and descending along the first and second halves of the path of motion, respectively. This complex motion can also be understood as the combination of a metronome-like and a camera-panning–like swinging motions. Finally, the collective motion captured by CPC #3 (Figure 4.21c, Supplementary Video S4.3) shows the B domain undergoing a gyration about an axis that coincides with the overall direction of the hinge between domains.

Remarkably, the two PCs that displayed high cosine content (Figure 4.18), namely, the PC #1 of trajectory replicates #4-B and #5-D, were found to be perfectly integrated in CPCs #1 and #3, respectively. This observation confirms that the collective displacement of coordinates captured by these PCs corresponded to biologically relevant dynamical features, equivalent in nature to those of other trajectory replicates, despite their PC projection values having stronger resemblance to cosine shapes.

On the other hand, CPCs #4 to #10 arose from the smaller clusters of the experiment, located in the second family of clusters of the dendrogram. The collective motions of these CPCs correspond to B-domain oscillations with considerably smaller amplitudes of motion. Interestingly, within this family of clusters, an additional subdivision allows us to distinguish between two subfamilies (Figure 4.19). The two subfamilies correlate with the nature of the captured collective motions. CPCs #4 to #9 belong to the first subfamily, which comprises rigid-body motions of the B domain. CPC #10 belongs to the second subfamily, which instead is characterized by local rearrangements of the structure of the B domain. Such a distinction was confirmed by inspecting the collective motions of both the involved CPCs and several other minor clusters or individual PCs.

Moreover, there is a correlation between the two subfamilies and the range of PC indices that they agglomerated. The CPCs from the first subfamily comprise PCs with indices mainly between #4 and #6, with only a few instances of #7 and #8. Conversely, the CPC from the second subfamily resulted from the consensus between PCs with indices #13, #14, and #20. This observation suggests that the consensus dynamics among trajectory replicates of this protein region occurred within the 85% of cumulative variance on average (Figure 4.17), and that PCs that captured less than 1% of the variance had low or null similarity.

The collective motions captured by CPCs #4 to #10 (Figures 4.21d-j, Supplementary Videos S4.4-S4.10) can be described as follows. CPC #4: a combination of roll, pitch, and yaw that results in the gyration about an axis perpendicular to the hinge between domains (a gyration perpendicular to that of CPC #3). CPC #5: a combination of roll, heave, and sway that results in a variation of roll with the top-left region acting as the pivot point. CPC #6: a combination of pitch, heave, and surge that results in a variation of pitch with the top-back region acting as the pivot point. CPC #7: a variation of CPC #5 that adds a component of yaw. CPC #8: a vertical shift (heave) that makes the hinge between domains stretch or compress. CPC #9: a combination of pitch and surge that results in a playground-swing–like swinging motion. CPC #10: the local fluctuation of the L-Bβ5-Bβ6 loop.

It is important to point out that the collective motions of the B domain were also accompanied by certain local fluctuations of the A domain. However, these were not distinctive of particular CPCs but repeatedly appeared in most of them, irrespective of the nature of the corresponding B-domain motion. Such local fluctuations mainly involve the most flexible loops of the A domain that were previously detected via the RMSF profiles (Figure 4.8), namely, L-Aβ2-Aα2 (residues 119-122), L-Aα2-Aβ3 (residues 141-147), and L-Aα4-Aβ5 (residues 304-306). What is the significance of such ubiquitous fluctuations in the context of CEDA? Remarkably, the incidence of these loop fluctuations was substantially lower, or almost null, in CPCs #1 to #3. This last observation is indicative of them potentially occurring by chance as background noise, rather than in a concerted manner with the diverse B-domain oscillations. This is because the lower incidence of the fluctuations in CPCs of high coverage (*i.e.*, CPCs #1 to #3) can be explained as the result of a more representative average behavior in contrast to that of CPCs with lower coverage.

This question was elucidated by examining the variance of the components of the clustered eigenvectors of each CPC. Vector components of eigenvectors represent the relative contribution of each atom to the displacement in the X, Y, and Z directions along the correlated motion of the corresponding PC. For instance, Figure 4.22 shows the plots of the X-, Y-, and Z-components of the centroid vector (mean values ± standard deviation) of the cluster that produces CPC #2. Reminiscent of RMSF profiles, the plots show that the B domain undergoes a correlated displacement as a block in the three spatial directions along CPC #2. The standard deviation interval illustrates the differences in orientation that the B domain adopts among replicates for the same overall motion. In other words, the collective motion does not occur along a unique clean path; the CPC represents the average path among the possible variations of the motion. Conversely, the vector components that relate to the A-domain flexible loops consist of a set of opposite values that cancel out upon averaging. Thus, CPC #2 displayed almost no fluctuation of these regions, in contrast to the majority of the individual trajectories that belong to the cluster, as shown in Figure 4.23. In CPCs with lower coverage, the attenuation was less perceptible due to the lower amount of structural divergence at these positions.

In conclusion, the fluctuations of the A-domain flexible loops appeared correlated with other motions by chance due to their high dynamical activity, although they oscillated in divergent directions with lack of consensus when accompanying the larger displacements of the B domain. This denoising effect achieved in CEDA precisely accords with the rationale behind the design of this strategy. When different correlated motions are truly concerted, CPCs provide the consensus paths of motion. Otherwise, random correlated fluctuations become attenuated or neutralized. As a result, CPCs render a denoised version of the predominant collective motion. As argued in the Discussion chapter, the performance of CEDA depends on the chosen method and parameters for the calculation of eigenvector similarity.

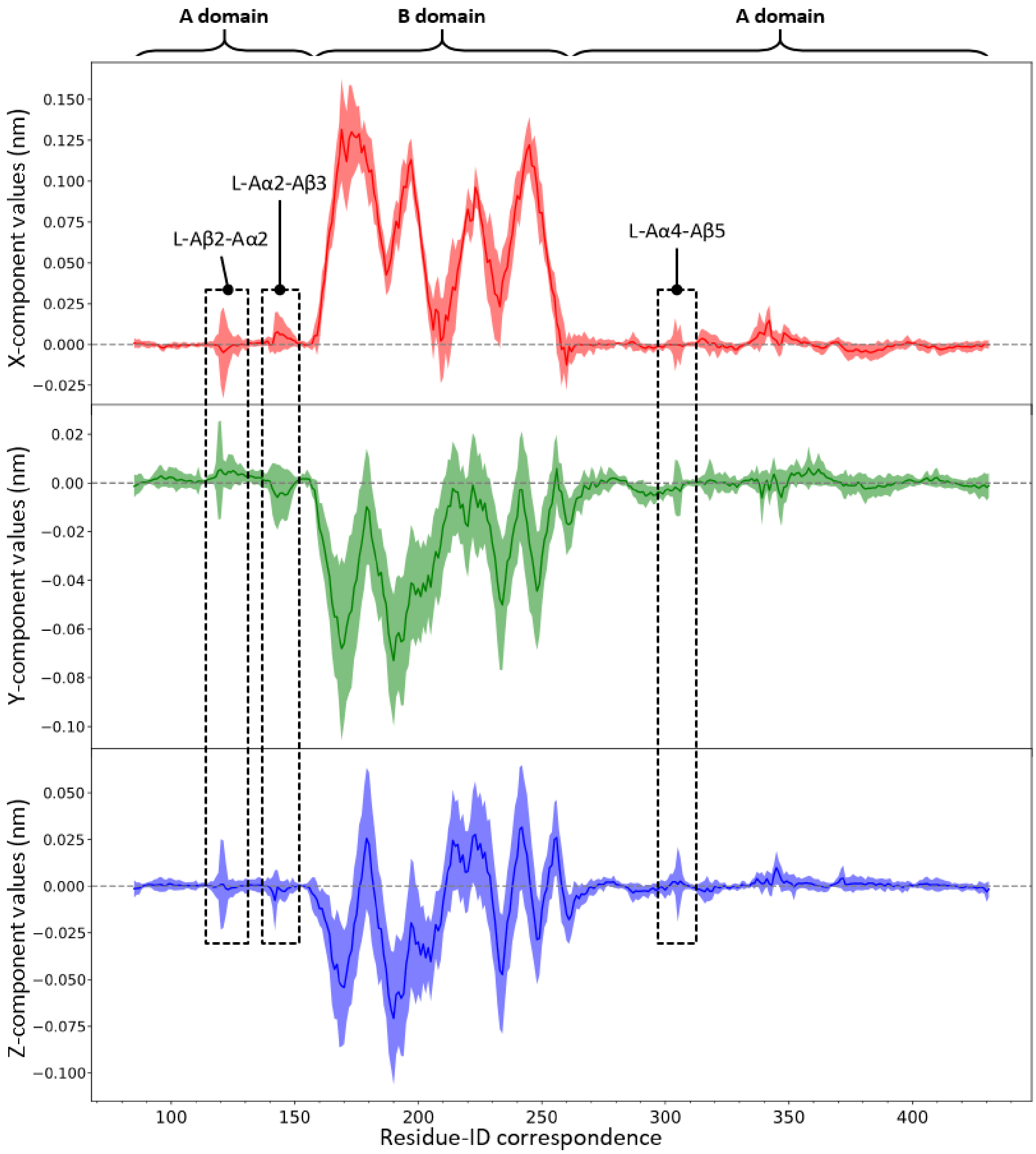


Figure 4.22. Vector components of the centroid vector of CPC #2 of the A and B domains in the WT apo condition. The X-, Y-, and Z-components are colored in red, green, and blue, respectively. Data corresponds to the mean values of the corresponding cluster of eigenvectors ± their standard deviation (shaded intervals). The A domain comprises residue IDs 85 to 159 and 163 to 431, while the B domain comprises residue IDs 160 to 262. The location of the A-domain loops L-Aβ2-Aα2 (residues 119-122), L-Aα2-Aβ3 (residues 141-147), and L-Aα4-Aβ5 (residues 304-306) is highlighted with the dashed black rectangles.

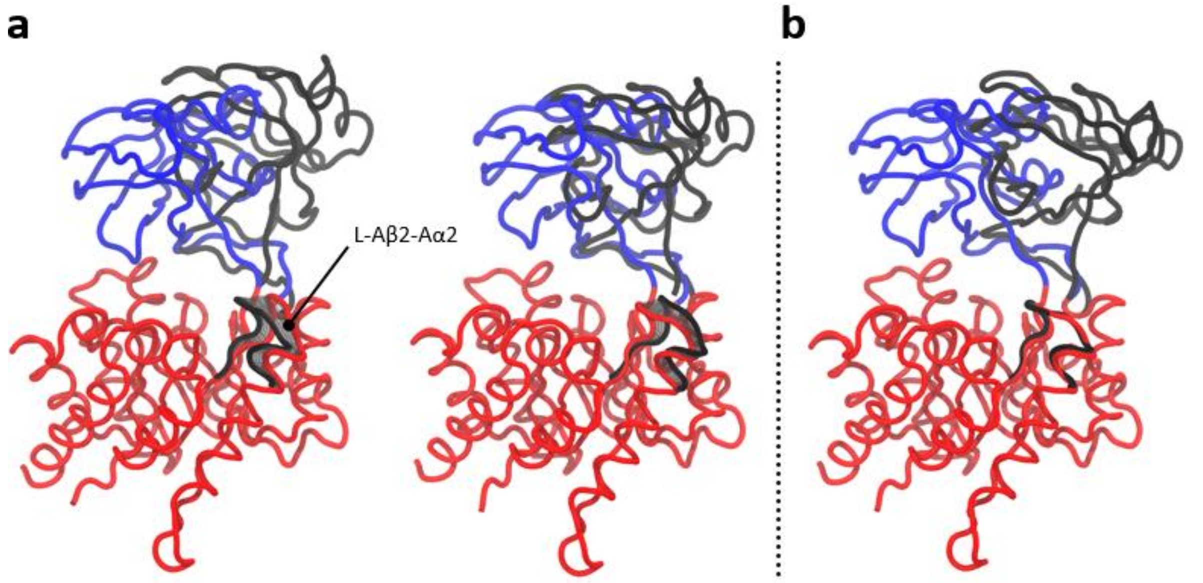


Figure 4.23. Local fluctuations of L-Aβ2-Aα2 along CPC #2 of the A and B domains in the WT apo condition. Structures are depicted with the trace representation between Cα atoms. In each panel, the initial conformations of the A and B domains are colored in red and blue, respectively. Structures colored in black correspond to the final conformations of the B domain and the L-Aβ2-Aα2 loop. Several intermediate conformations of the loop are also shown in transparent gray. (**a**) Example of dynamical divergence in two of the clustered eigenvectors. The tilting motion of the B domain from its rightward to its leftward orientation is accompanied by the fluctuation of the L-Aβ2-Aα2 loop in opposite directions between eigenvectors. (**b**) Attenuated fluctuation represented in the centroid vector (average behavior). The L-Aβ2-Aα2 loop barely fluctuates, with no discernible intermediate conformations. NOTE. The images were generated with the software VMD.

4.1.3.1.2 Density distribution of projection values of CPCs in the WT apo condition

After deriving the CPCs of the WT apo condition and revealing the corresponding consensus collective motions, the next stage revolved around a deeper characterization of the conformational distribution intrinsic to such motions. The distribution of projection values along each CPC was examined to determine whether certain conformations are more abundant within the span of the motions. This allows for ascertaining the conformational diversity among equivalent trajectories of the system, as well as studying the possible functional implications of both the motions and their most distinctive conformations.

For that purpose, the Kernel Density Estimation (KDE) method was employed to estimate the probability density function of the range of projection values between the extreme positions of the motion. KDE facilitates representation of random data distributions in a similar manner as histograms, but with the advantage of providing a smooth continuous curve along the domain of values. The estimation requires choosing a value for a free parameter, called the bandwidth, which has an influence on the smoothness of the generated curve. Suboptimal bandwidth values may produce undesired effects: “under-smoothed” curves contain too many spurious artifacts (spiky surface), whereas “over-smoothed” curves obscure the informative variations in density of the underlying structure. The trials of KDE with our data showed marked tolerance (strongly conserved curve shapes) to diverse values of bandwidth. Since data was abundant (187501 analyzed frames per trajectory), the estimate robustly reproduced the true underlying densities. Ultimately, the empirical value of 0.3 was selected because it showed the best smoothness trade-off. KDE curves were represented with 100 points along the domain of the data.

Figure 4.24 shows the density distributions along CPC #1 for each trajectory replicate and for the aggregated projection data of the whole trajectory ensemble. Representative structures of the A and B domains at different projection values are included to inform about their correspondence with the range of conformations. Noticeably, both the shape and span of the distributions are diverse among replicates. Therefore, the collective motion along CPC #1 (the opening/closing of the active site) was manifested with different degrees of closure between equivalent trajectories. This was the case despite the motion being present in all 20 replicates with strong consensus features. In other words, this observation indicates that the fact that a consensus collective motion is detected in an ensemble of trajectories does not imply that the sampled amplitudes of motion are the same between the individual trajectories.

Despite the wide conformational heterogeneity, certain regions of the spectrum concentrate more density and correspond to the conformations with higher sampling. These are reflected in the final distribution of the aggregated projection data, which is trimodal. The most frequent conformation comprises projection values around 1.6 and corresponds to a partially closed form of the B domain. This state is a local maximum of the distribution. Around the value of 4.5, although not manifested as a local maximum, another density concentration corresponds to a more open form of the B domain. Finally, the local maximum of less size around the value of -7.2 corresponds to a closed form of the B domain. The low levels of density at the extreme values of the spectrum suggest residual sampling of conformations with highly closed (negative extreme values) and highly open (positive extreme values) forms of the B domain.

While some simulations individually covered a range of conformations of the spectrum, others mainly remained in a single conformation and its surroundings. The largest volume of sampled conformations differed from that of the original crystallographic structure of the simulations, the PDB entry 2VGB [121]. The projection values of 2VGB were determined by projecting its structural data onto the vector of CPC #1, with a procedure equivalent to that applied to trajectory data. The structure 2VGB is a WT holo PKR co-crystallized with both cofactors K^+ and Mg^{2+}, the substrate analog phosphoglycolate and the allosteric effector FBP. All four subunits of 2VGB exhibit closed B domains, with projection values around -6 (marked with black vertical dashed lines in Figure 4.24b), near the local maximum of the closed forms in simulation. Therefore, it appears that the removal of ligands for simulation in the apo condition promoted the sampling of the more open forms of the B domain, as has been shown to occur frequently in apo PKs [119], [124], [135], [137], [139], [147], [160], [164].

In agreement with this idea, a more thorough exploration of the projection values as time-series data along the course of each simulation showed that transitions occurred predominantly from closed to open forms (see the example of replicate #2-A in Figure 4.25a). Nevertheless, the reverse transition (open-to-closed transition) was also detected, albeit in fewer instances (see the example of replicate #3-B in Figure 4.25b), thus confirming that the apo protein is able to oscillate between both forms within the time span of these simulations. Moreover, one replicate even remained stable most of the time in the closed form (replicate #5-D). On another note, the conformational profiles of the different replicates show further evidence of the lack of symmetry in the B-domain conformations between the subunits of the tetramer, reinforcing the idea suggested in previous results of both this study and others [121], [132], [138], [164]. Otherwise, we should see roughly equivalent profiles between subunits of the same replicate in a more consistent way.

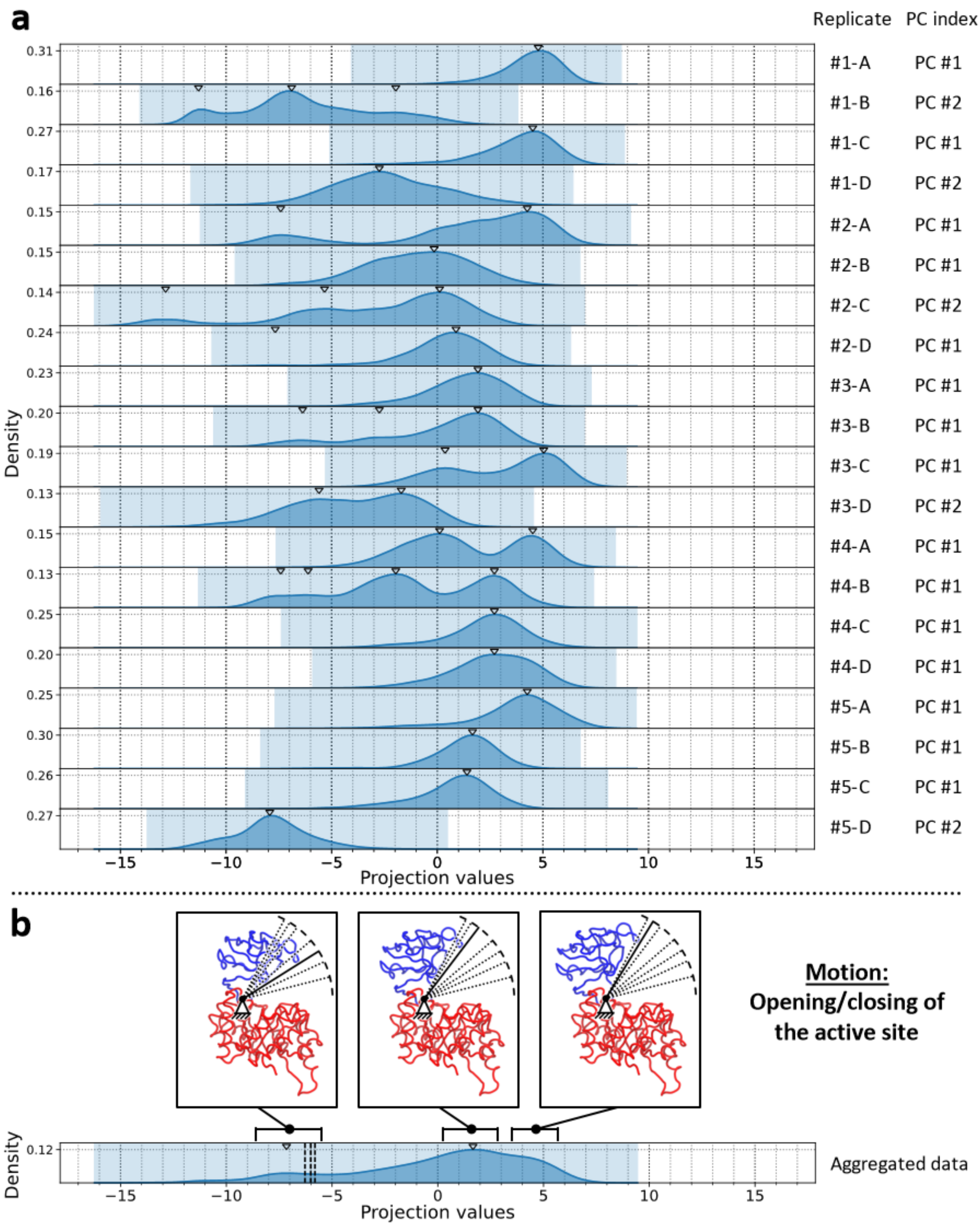


Figure 4.24. Density distributions along CPC #1 of the A and B domains in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the A and B domains at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and B domains are colored in red and blue, respectively. NOTE. The images of the protein structure were generated with the software VMD.

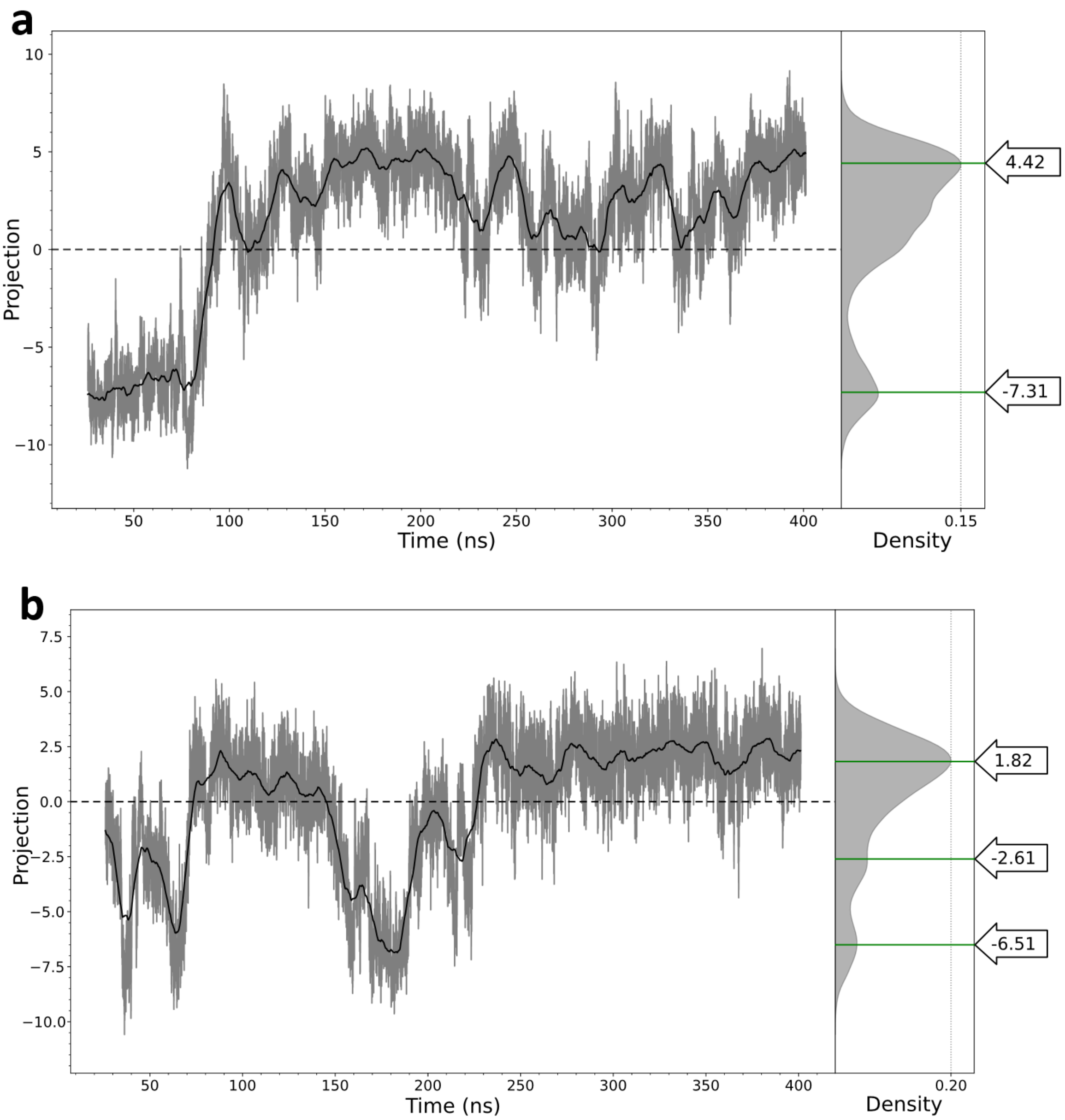


Figure 4.25. Examples of time-series projection values along CPC #1 of the A and B domains in the WT apo condition. Data from the first 25 ns of simulation was excluded from the analysis as it accounted for the initial relaxation phase. The darker line plotted alongside each time series represents the two-sided moving average of the data, encompassing values up to 5 ns on either side of each point. The subplots at the right margin show the density of projection values as estimated with a KDE with bandwidth 0.3 (the distributions are equivalent to those shown in Figure 4.24 for the same trajectory replicates). The local maxima of the distributions are indicated with green lines and labeled arrows with the corresponding projection values. (**a**) Example of replicate #2-A that underwent the predominant conformational transition of the trajectory ensemble, from a closed conformation to more open conformations. The simulation started in the closed conformation. Then, at the time interval between 80 and 90 ns, it shifted to the open conformations. (**b**) Example of replicate #3-B that underwent both the forward and the reverse conformational transitions. The simulation balanced between the closed and open conformations, with the corresponding shifts at the time intervals between 70 and 80 ns, 150 and 160 ns, and 190 and 230 ns.

The density distributions of projection values along CPCs #2 and #3 are shown in Figures 4.26 and 4.27, respectively. The clusters of these two CPCs did not have full coverage of all replicates. However, the few absent trajectories (labeled as “None” in the figures) do manifest certain degrees of conformational variance when projected along the directions of CPCs #2 and #3, especially the latter.

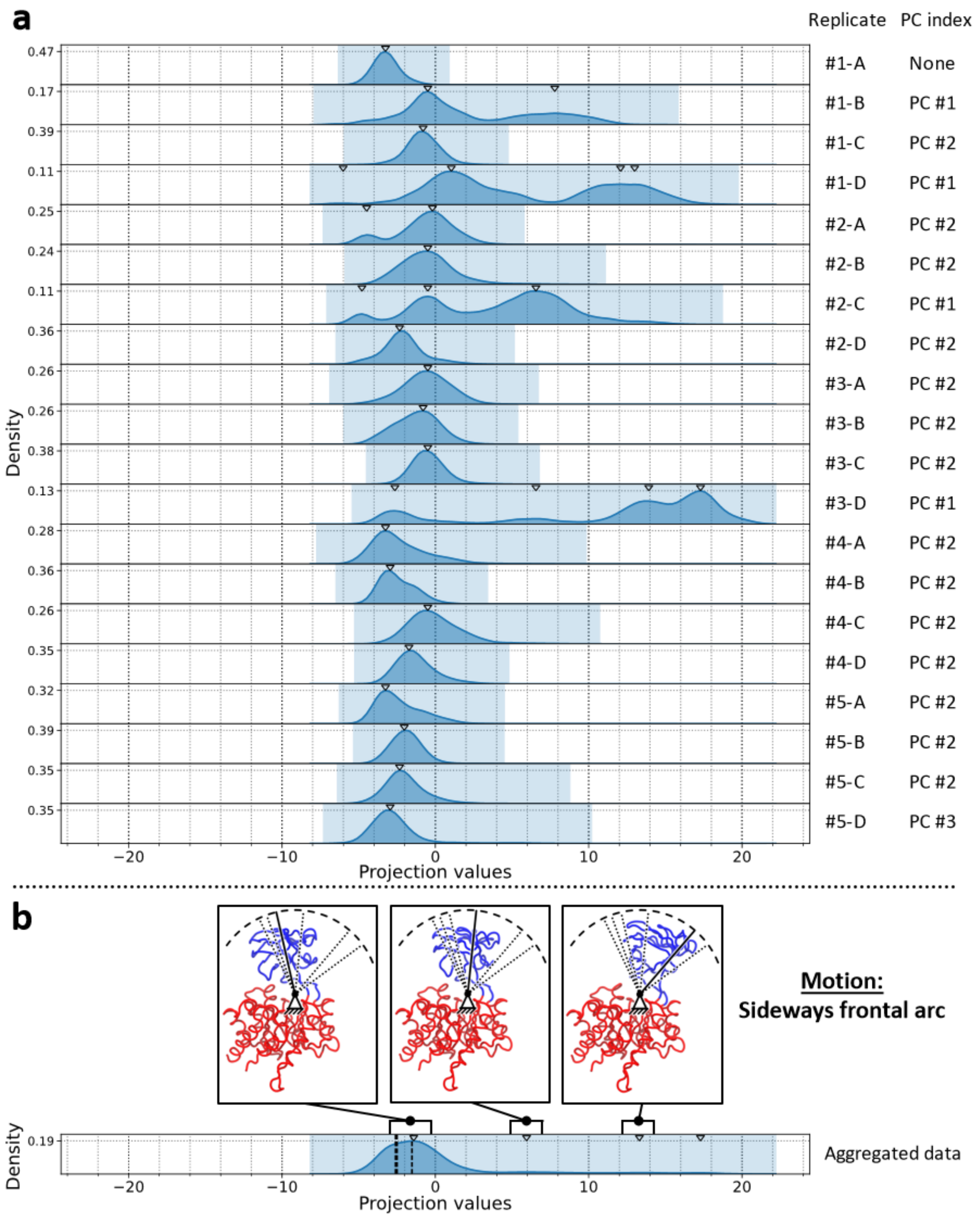


Figure 4.26. Density distributions along CPC #2 of the A and B domains in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier (labeled “None” when the replicate is absent from the cluster). (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the A and B domains at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and B domains are colored in red and blue, respectively. NOTE. The images of the protein structure were generated with the software VMD.

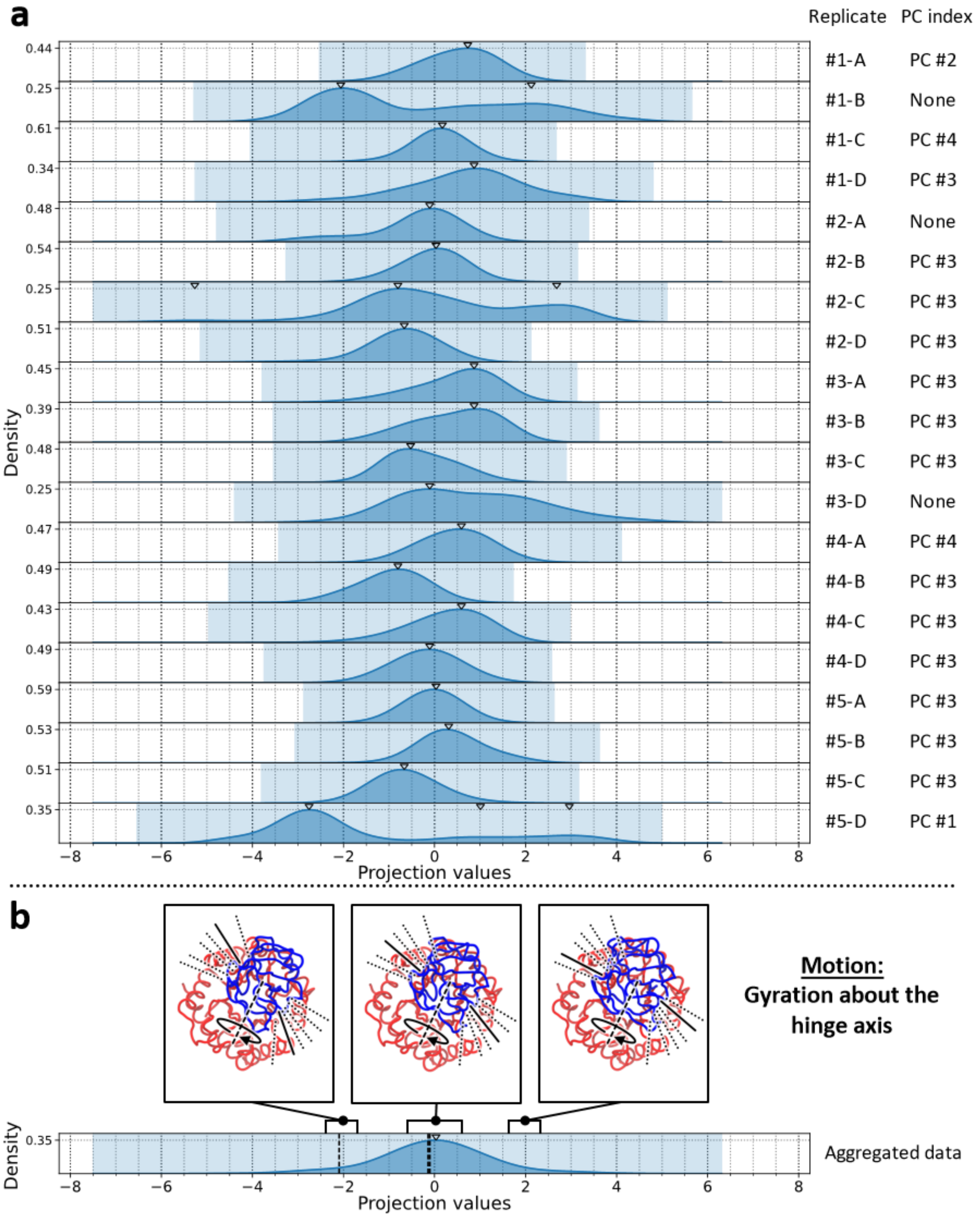


Figure 4.27. Density distributions along CPC #3 of the A and B domains in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier (labeled “None” when the replicate is absent from the cluster). (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the A and B domains at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and B domains are colored in red and blue, respectively. NOTE. The images of the protein structure were generated with the software VMD.

The distribution of CPC #2 is mainly unimodal, with a local maximum around the value of -1.6. However, it displays minor volumes of density at the region of positive values beyond the value of 4. Such a region was practically only sampled by four distinct trajectory replicates, namely, replicates #1-B, #1-D, #2-C, and #3-D, and corresponds to conformations with the B domain tilted to its left. Not surprisingly, the eigenvectors of these replicates have the highest eigenvalues of the cluster and bear the PC index #1, in contrast to the rest of the cluster members which have PC indices #2 and #3 (only one instance of the latter).

The distribution of CPC #3 is unimodal and features heavy tails. Unlike the former two distributions, it is centered around the global average structure (*i.e.*, the zero value). Interestingly, the structure 2VGB exhibits conformational heterogeneity among its subunits. While three subunits display projection values near 0, the remaining subunit displays a value near -2 that corresponds to a conformation of B domain with its frontal region slightly more oriented to its right.

Figure 4.28 shows the density distributions along CPCs #4 to #10. This time, for the sake of brevity, the plots corresponding to each individual trajectory replicate have been omitted to show only the density of the aggregated projection data. The distributions of these CPCs mainly are unimodal, with different degrees of skewness, and are mostly centered around the global average structure. Only CPC #8 appears to exhibit a minor volume of density further from the main region, at positive projection values beyond 1.5. The total span of projection values of the majority of these distributions is considerably narrower than that of the first three CPCs. All in all, these features suggest that these collective motions describe fluctuations around a single energy minimum per CPC. The conformational heterogeneity between the subunits of structure 2VGB is repeatedly exhibited in most CPCs.

Density distributions may also be examined along several CPCs at once by arranging the corresponding projection data and applying KDE. Such multidimensional distributions provide more detailed topologies of the underlying conformational heterogeneity. For instance, an apparent concentration of density along one CPC may correspond to several distinct regions upon consideration of a second CPC. The interpretation of these topologies is similar to that of a free-energy landscape, although inverted: conformational stability is represented with local maxima of frequency instead of local minima of energy. Figure 4.29 shows the pairwise two-dimensional plots of CPCs #1 to #3, which are the most relevant.

The comparison between CPCs #1 and #2 (Figure 4.29a) provides new insights in the interpretation of the characteristic conformations. While the distribution of CPC #2 alone may be described as unimodal (disregarding the scattered minor volumes of density), now two characteristic values can be distinguished within the major volume of density. One value coincides with the one-dimensional peak (around the value of -1.6), while the other (around the value of -3) becomes more perceptible. Remarkably, the two-dimensional distribution facilitates additional distinction of the open and partially closed forms of the B domain (values of CPC #1 around 4.8 and 2; now both manifested as local maxima) by these two values of CPC #2. Furthermore, the closed form of the B domain (value of CPC #1 around -7.7) mainly adopts the newly acknowledged characteristic value of CPC #2, which corresponds to a more rightward-tilted B domain. In contrast, bimodality of CPC #2 becomes irrelevant in its comparison with CPC #3 (Figure 4.29c), whereby both motions are clearly describing the fluctuation around a single and wide conformational ensemble. Finally, the comparison between CPCs #1 and #3 (Figure 4.29b) reveals that the open and partially closed forms of the B domain are entirely circumscribed in the central region of density of CPC #3, while the closed form defines a separate narrower conformational ensemble at the more negative side of the spectrum.

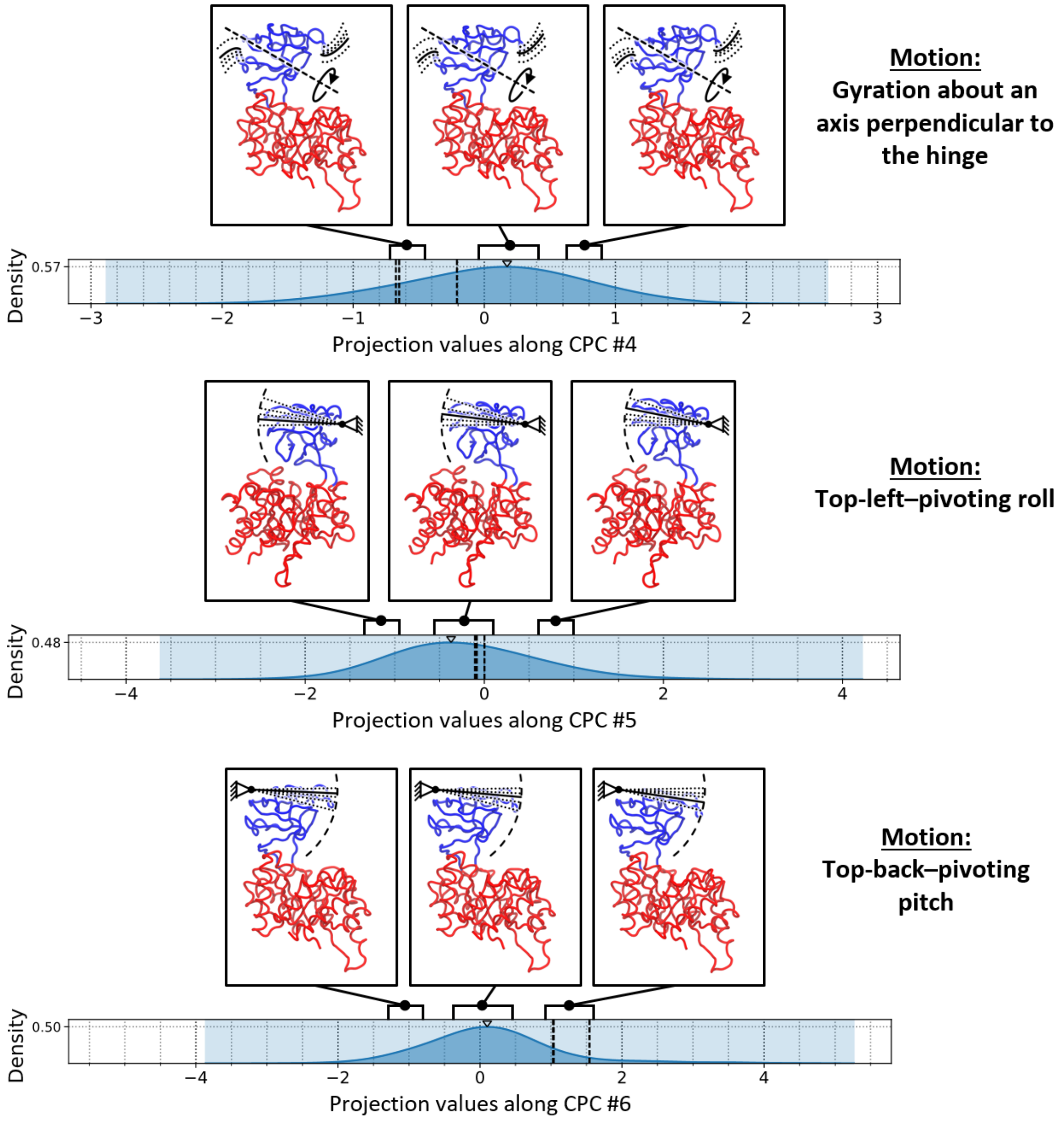
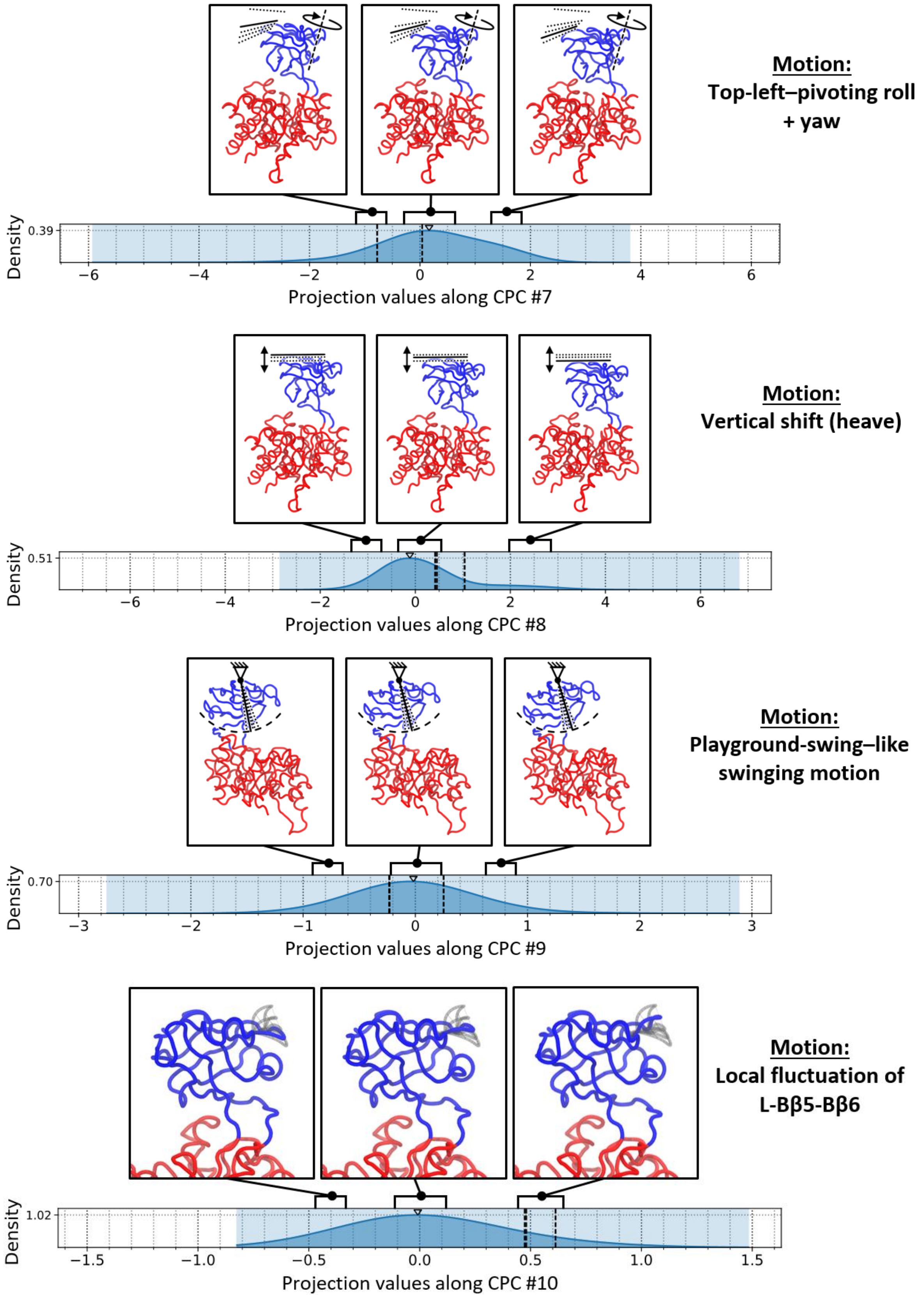
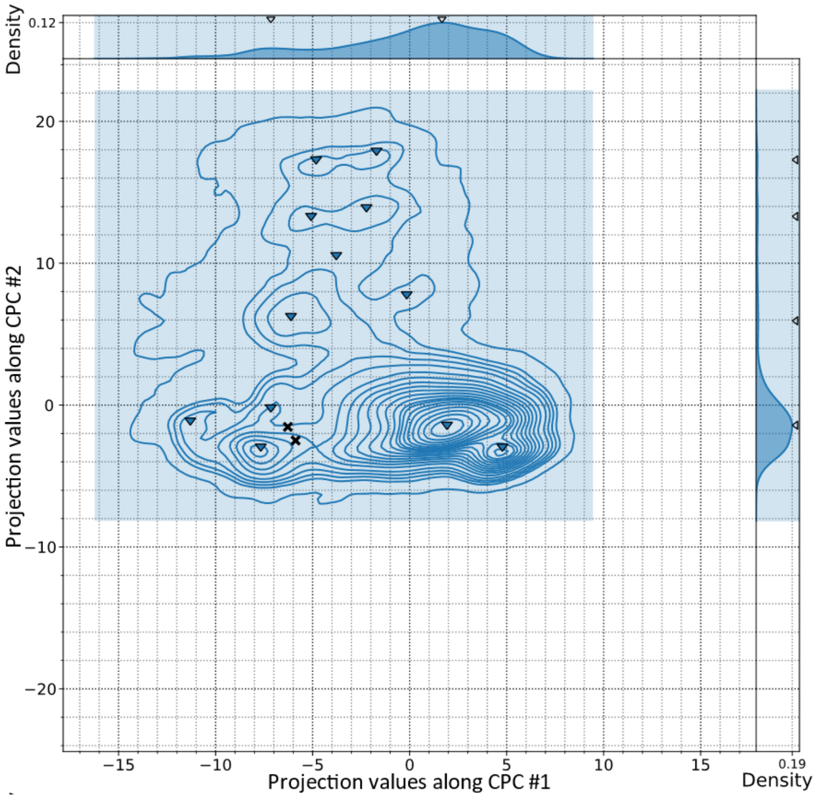


Figure 4.28. Density distributions along CPCs #4 to #10 of the A and B domains in the WT apo condition. KDE curves of the aggregated projection data with representative structures of the conformation of the A and B domains at different projection values. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and B domains are colored in red and blue, respectively. NOTE. The images of the protein structure were generated with the software VMD.

Figure 4.28 (Continued)





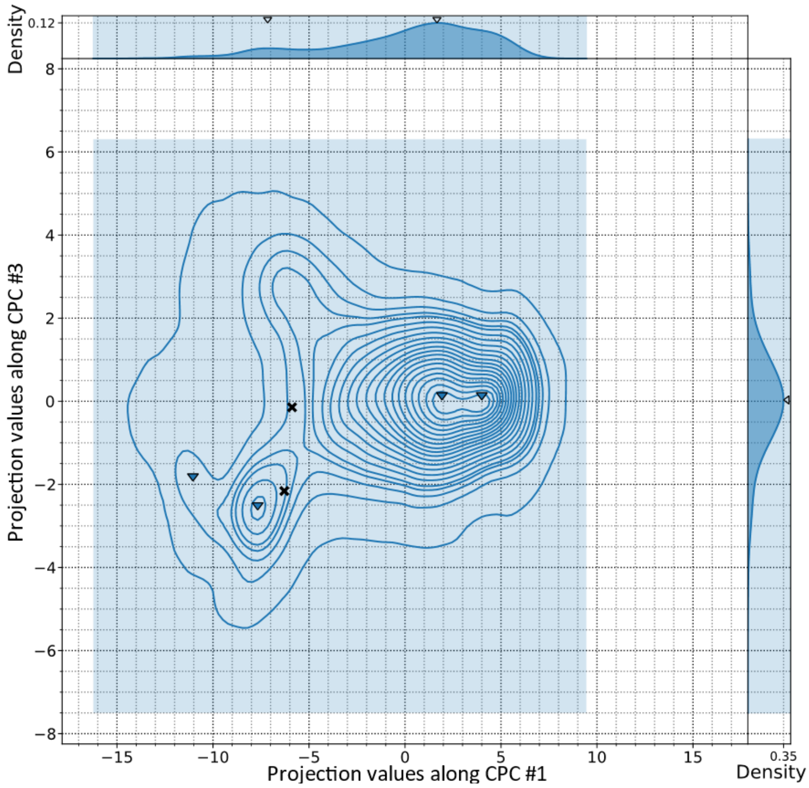
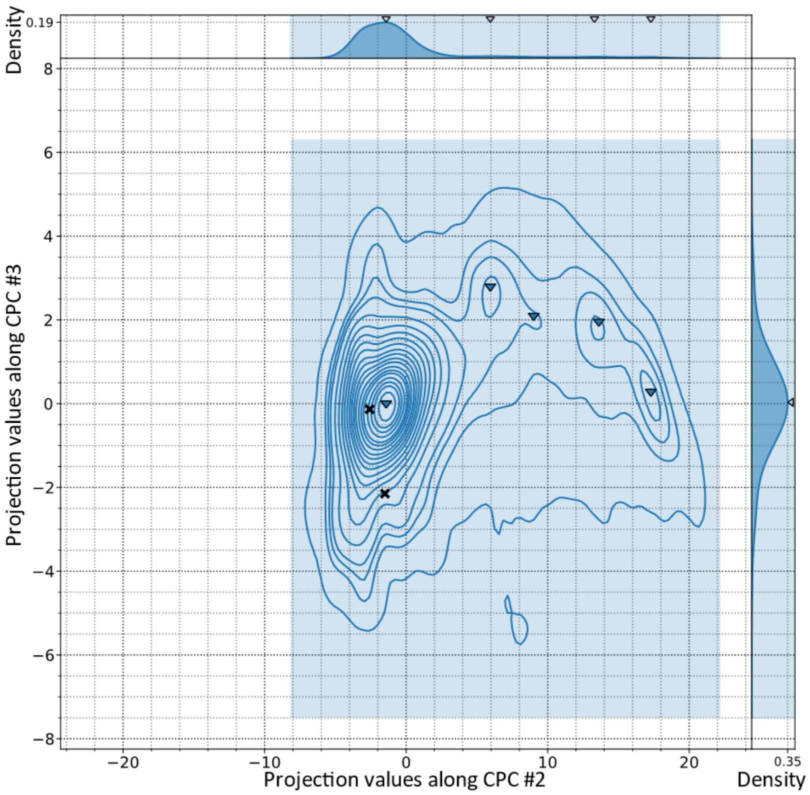


Figure 4.29. Pairwise 2D density distributions along CPCs #1 to #3 of the A and B domains in the WT apo condition. The highlighted rectangular regions indicate the total span of projection values. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Triangle markers indicate the locations of local maxima. The projection values of the subunits of the structure 2VGB are indicated with "X" markers. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins.

Figure 4.29 (Continued)



4.1.3.1.3 Comparison with an additional trajectory ensemble of the WT apo condition

The reliability of the obtained CPCs and conformational profile of the A and B domains in the WT apo condition is conditional on the robustness of its replicability with new equivalent trajectories. For this reason, 5 additional simulations (#6 to #10) of the WT apo tetramer were executed, thus providing a new batch of 20 trajectory replicates of the A and B domains in this condition. Importantly, these additional simulations were procured with the sole purpose of assessing the replicability of the experiments. Therefore, their data was neither mixed with the original simulations nor incorporated as input for subsequent experiments.

The first replicability assessment consisted in evaluating the similarity between the original set of CPCs and those derived with the second ensemble of trajectories. The procedure of CEDA was accordingly applied to these new replicates and the corresponding set of CPCs was obtained. The clustering of eigenvectors generated 9 relevant clusters with the same parameters as in the original experiment. The structure of the resulting dendrogram (Figure 4.30) coincided significantly with that of the original dendrogram, with three major clusters of high coverage values (95-100%) and with the rest bearing low coverage values near the cutoff (20-35%) (clusters are designated by their number and a prime symbol). The dendrogram noticeably repeated the subdivision into two main families of clusters. Interestingly, in this regard, the only difference is that cluster #5′ (the third in decreasing order of coverage) is now located in the second family of clusters.

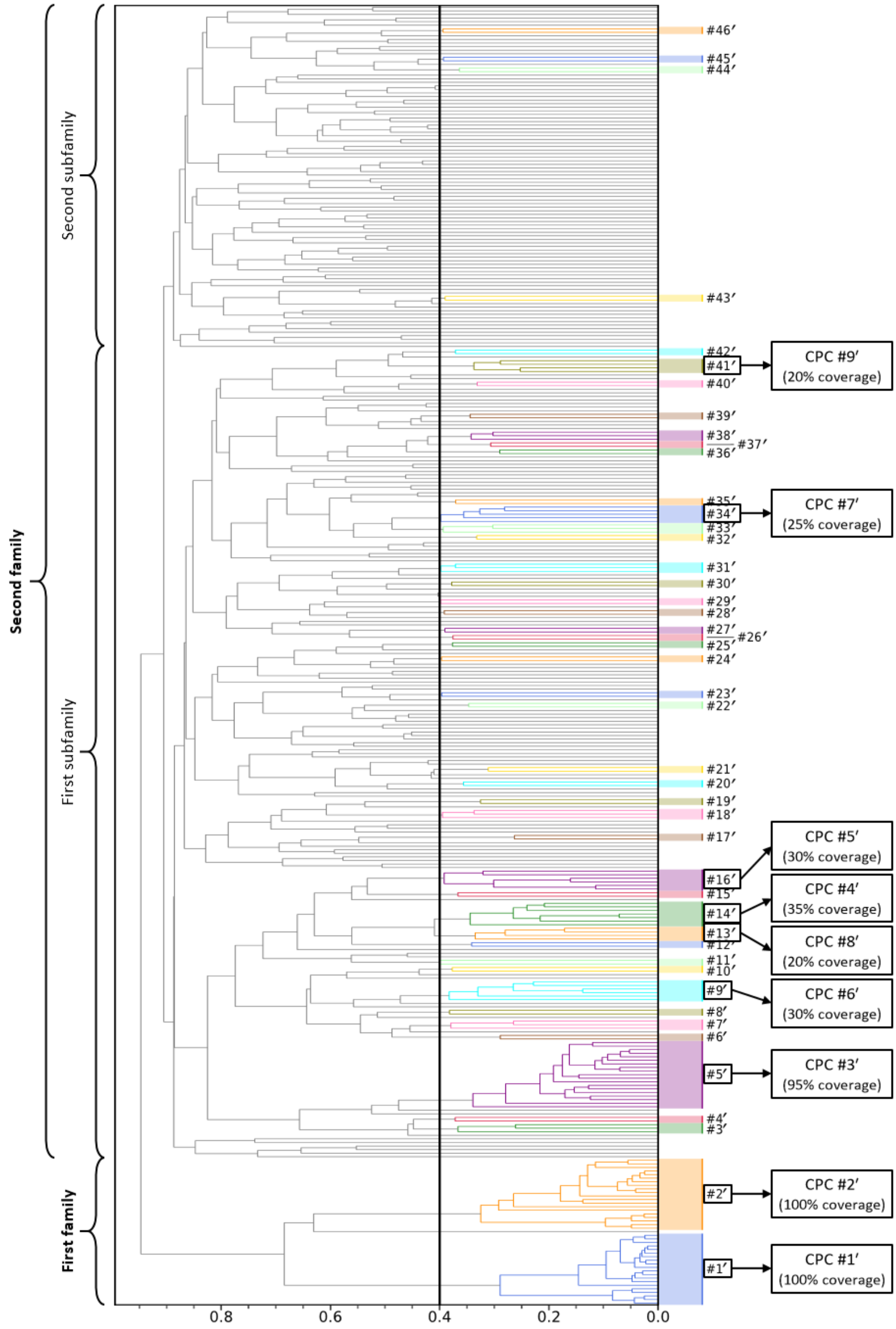


Figure 4.30. Dendrogram of the CEDA of the A and B domains in the WT apo condition (second batch of simulations). The cophenetic distance cutoff of 0.4 used to split the dendrogram and obtain clusters is shown with a vertical black line. Non-singleton clusters are shown in various colors to facilitate visual identification. The 9 clusters with at least 20% of coverage were selected to acquire the CPCs of the experiment. CPCs were numbered from #1′ to #9′ in decreasing order of coverage and average percentage of variance of their cluster members. The figure indicates which clusters yielded each CPC, together with their achieved coverage. At the left margin of the dendrogram, the span of the detected families and subfamilies of clusters is indicated with labeled curly brackets.

The centroid vector of each cluster was computed to obtain the 9 corresponding CPCs. Then, the similarity between the first and the second sets of CPCs was measured in terms of the cosine of the angle between vectors (cosine similarity), expressed in absolute value to disregard differences due to opposite directions. The results of the comparison are shown in Figure 4.31, where each radar chart displays the pairwise similarity values between one of the 10 original CPCs and each of the 9 new CPCs. CPCs from the new set are designated by their index and a prime symbol.

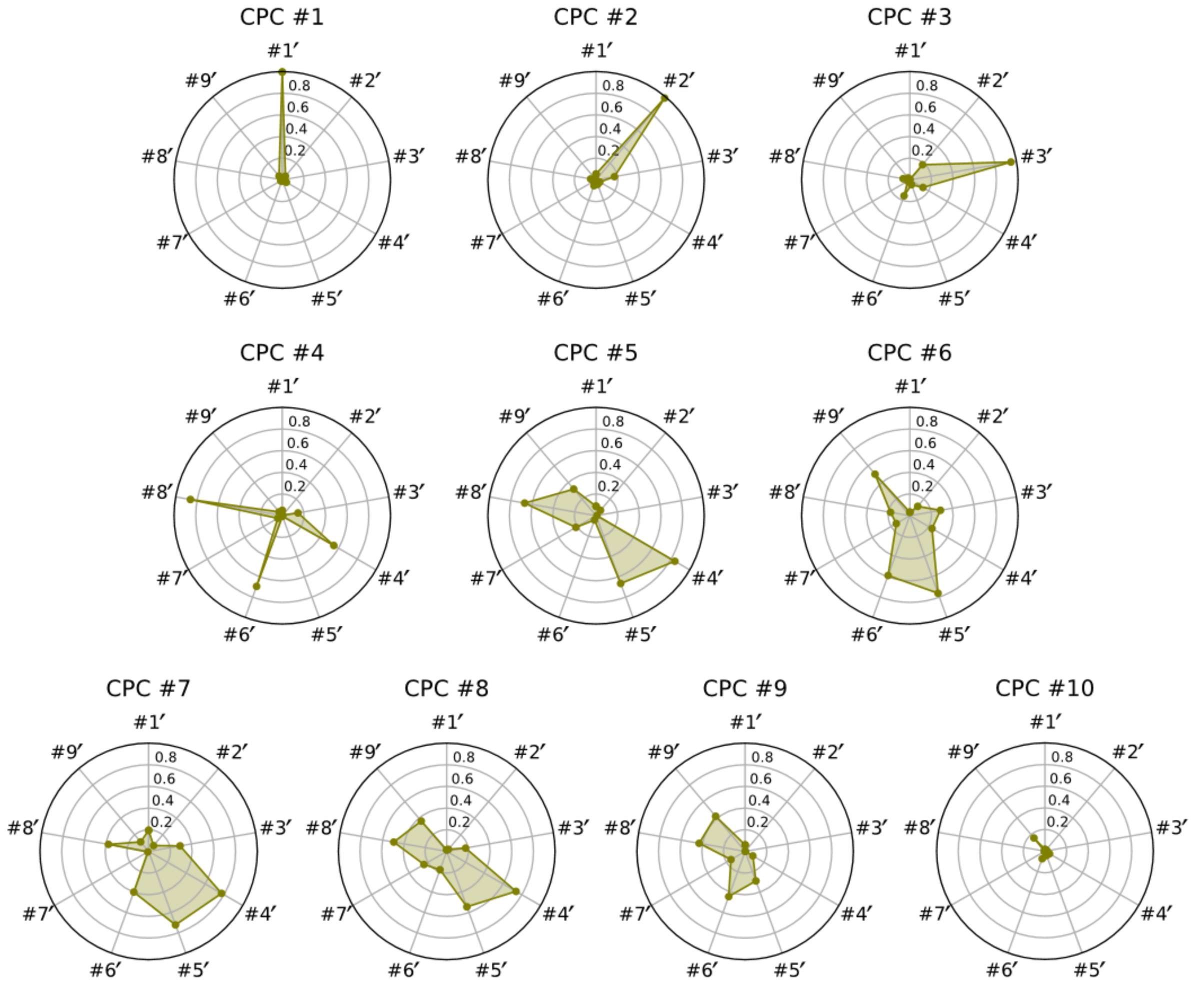


Figure 4.31. Similarity between the original and additional sets of CPCs of the A and B domains in the WT apo condition. Each radar chart displays the pairwise similarity values between one of the 10 CPCs from the original set (indicated at the caption of each chart) and each of the 9 CPCs from the additional set (distributed along the angular axis and designated by a prime symbol). Similarity is expressed in absolute value of cosine similarity, bounded between 0 (no similarity) and 1 (full similarity), along the radial axis.

The correspondence between both sets is remarkably high at the first three CPCs (#1 with #1′; #2 with #2′; #3 with #3′), as evidenced by the respective sharp edges that indicate a cosine similarity value close to 1. Thus, both trajectory ensembles were strongly characterized by these main collective motions. In contrast, the rest of the original CPCs do not exhibit a clear correspondence with individual CPCs of the new set, but rather mild or moderate levels of similarity with a wider number of vectors. This observation confirms that CPCs with low coverage values are also less represented among trajectory ensembles, and suggests that they traverse conformational subspaces of mixed collective features. Interestingly, CPCs #7′ and #9′ do not appear to be similar to any of the original CPCs. Finally, CPC #10 is not represented in the second set of CPCs, which is coherent with the lack of clusters at the corresponding region of the new dendrogram.

On the other hand, a second replicability assessment was conducted to evaluate the similarity between the conformational profiles of both trajectory ensembles in terms of the original set of CPCs. Accordingly, the trajectories from the new batch of replicates were subjected to structural superposition and data centering around the original global average structure. Then, the processed trajectory data was projected onto the vectors of the original CPCs. Subsequently, the corresponding density distributions of projection data were generated. Figure 4.32 shows the pairwise two-dimensional plots of CPCs #1 to #3 with overlays of the density distributions of both trajectory ensembles to facilitate comparison. Filled areas in the plots represent the intervals of highest 95% density of each distribution.

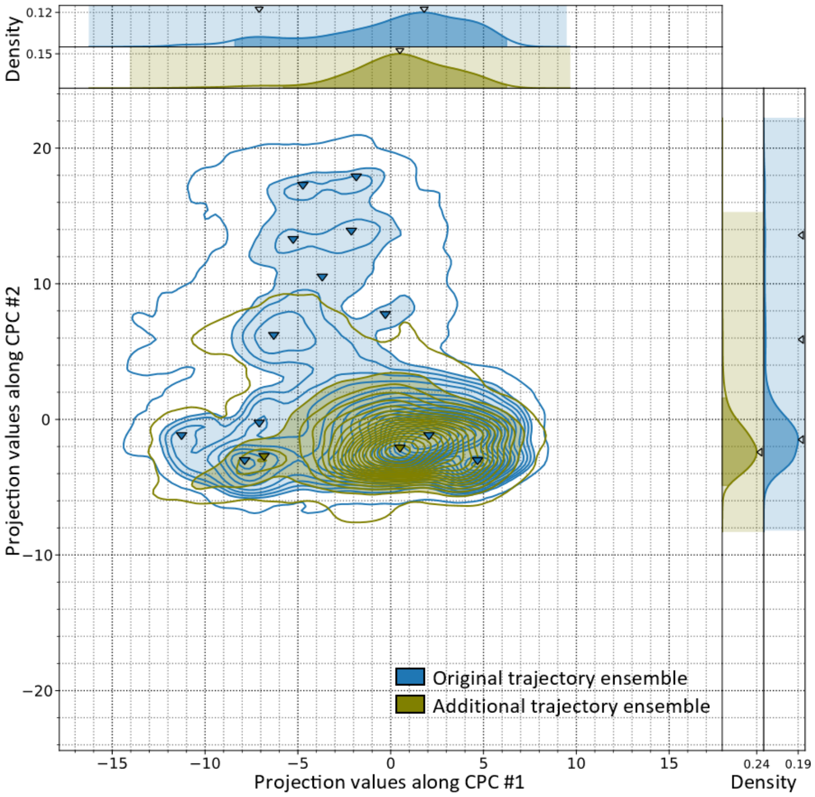
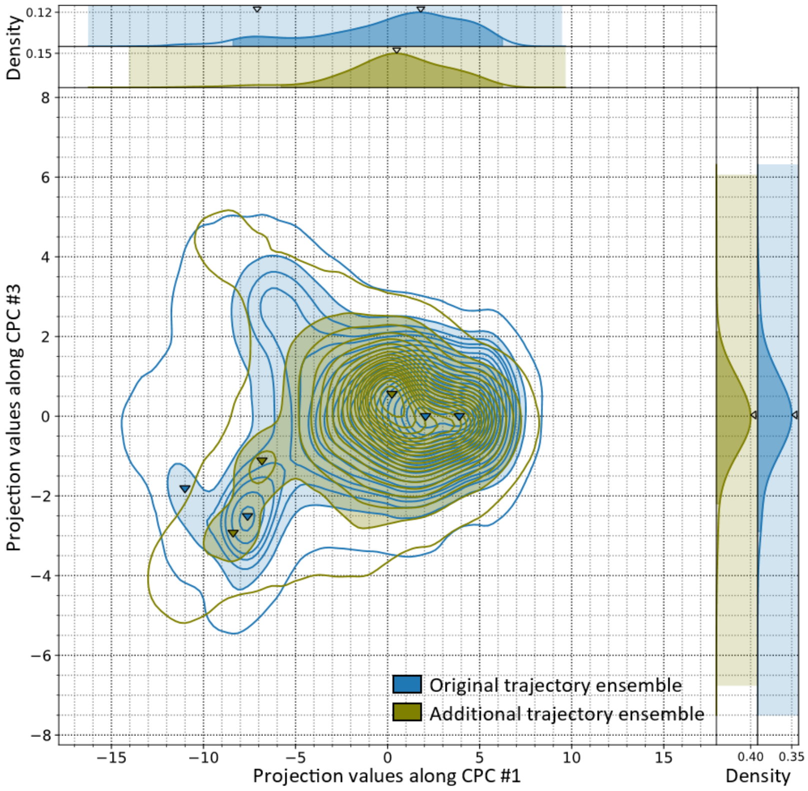
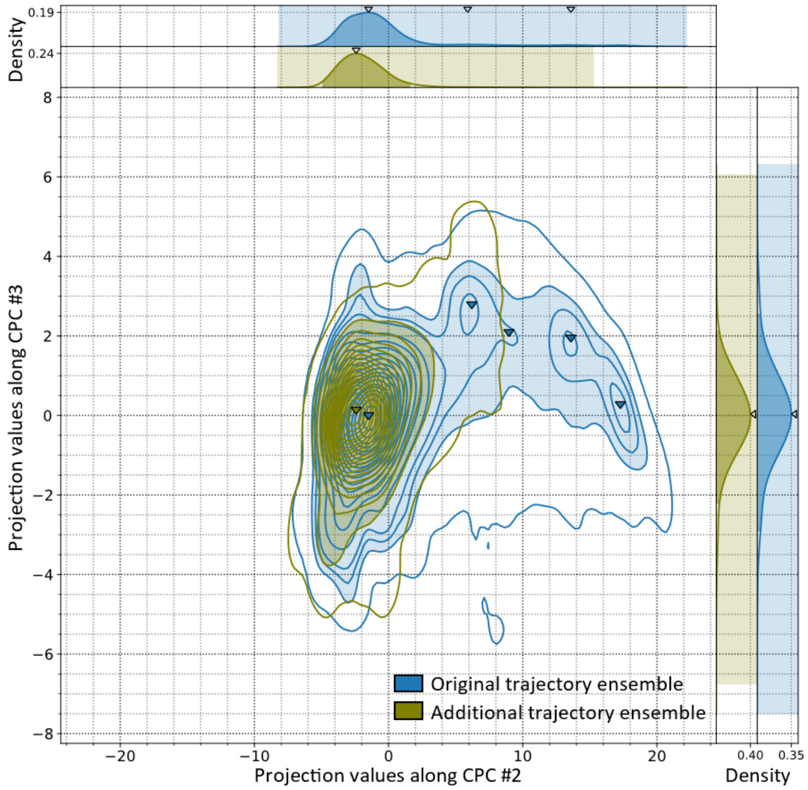


Figure 4.32. Comparative analysis of the original *vs*. additional WT trajectory ensembles (A and B domains) in the apo condition: conformational profiles (2 CPCs). The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 (top), #1 *vs*. #3 (center), and #2 *vs*. #3 (bottom) from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figure 4.32 (Continued)





Both trajectory ensembles share a substantial fraction of their principal concentrations of density, which is indicative of an overall equivalent exploration of the same conformational space. However, the distributions also exhibit significant differences. Most importantly, the second ensemble is less heterogeneous. Regarding CPC #1, the open and partially closed forms of the B domain exhibit equivalent conformational profiles, although the second distribution displays a slight shift towards the negative side of the spectrum. However, the closed form of the B domain was considerably less sampled; it is only detected as a local maximum in the two-dimensional distributions. Thus, the second trajectory ensemble is again characterized by a predominant transition from closed to open forms of the B domain, with only a marginal sampling of the open-to-close transition by a few replicates (not shown). Along the direction of CPC #2, the second trajectory ensemble scarcely explored the leftward-tilted forms of the B domain. Moreover, there is no appreciable distinction between subpopulations within the major volume of density of the distribution. The distributions along CPC #3 do not display notorious changes between trajectory ensembles, although, when inspected in combination with CPCs #1, the major conformations are shifted towards the positive side of the spectrum of CPC #3. On another note, the density distributions along CPCs #4 to #10 (not shown) exhibited equivalent profiles between the two ensembles.

The similarities and differences found between the two trajectory ensembles have been described in terms of a visual inspection between the corresponding data distributions. In order to complement this assessment with quantitative indicators, three metrics were employed to summarize the comparison between the main features of the reference and target distributions. Below is a brief explanation of each metric accompanied by Figure 4.33, which shows a conceptual example of their application to two hypothetical reference and target distributions.

The first metric, termed the *overlap*, was defined as the percentage of the span of the target distribution, weighted by its density values, that is within the span of the reference distribution. This metric represents the extent of conformational space sampled by the target condition that coincides with that of the reference condition. In the example of Figure 4.33b, it accounts for about 67% of the target distribution, meaning that 33% of the trajectory frames of the target condition correspond to conformations outside of the reference conformational space.

The second metric, termed the *coverage*, is related to the definition of the overlap but interchanging the reference and target distributions. More specifically, it was defined as the percentage of the span of the reference distribution, weighted by its density values, that is covered by the span of the target distribution. In this case, it represents how well the conformational space sampled by the target condition covered that of the reference condition. In the example of Figure 4.33c, it accounts for about 80% of the reference distribution, meaning that 20% of the reference conformational space was never sampled by the simulations in the target condition.

Finally, the last metric is the Bhattacharyya coefficient (BC), which is a statistical measure that quantifies the similarity between two statistical samples or populations (Equation 3.15). The BC is bounded between 0 and 1, whereby a value of 0 denotes total dissimilarity (distributions do not overlap), while a value of 1 corresponds to identical distributions. The BC was applied to compare the full reference distribution with the overlapping fraction of the target distribution in the span of the former, as shown in the example of Figure 4.33d, with both distributions being normalized such that the total density of each distribution equals 1.

In the quantitative comparison between the two trajectory ensembles of the WT apo condition, the reference and target distributions refer to those of the original and additional simulations, respectively. Rather than calculating the similarity metrics between the one- or two-dimensional pairs of distributions that were already visually inspected, the quantitative comparison was conducted on the three-dimensional density distributions along CPCs #1, #2, and #3 of both trajectory ensembles. While such three-dimensional distributions cannot be conveniently shown in a plot, they account for the combined distributional features along each dimension. Importantly, only the intervals of highest 95% density of each distribution were considered to disregard outlier regions before computing the metrics.

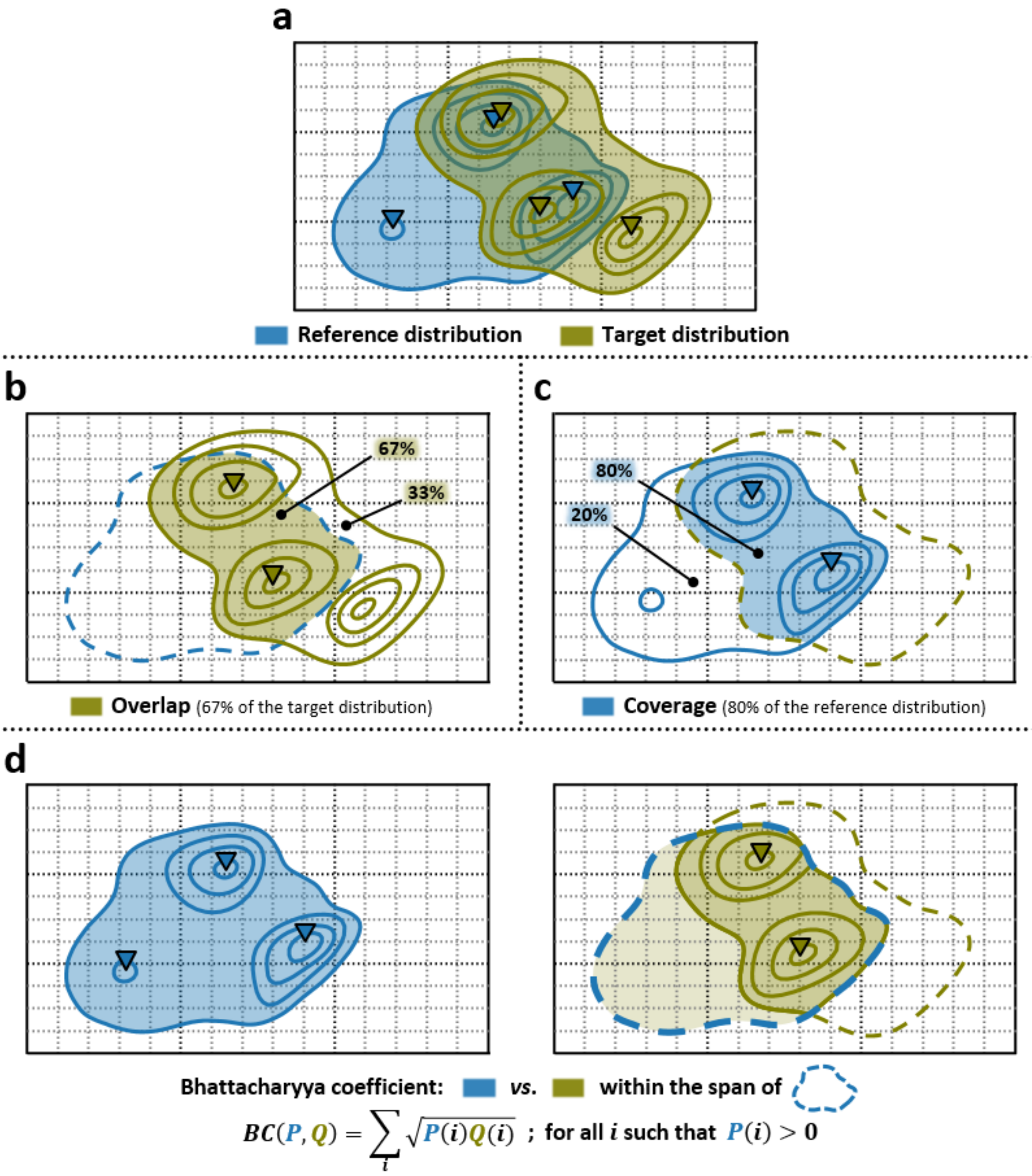

a
Reference distribution
Target distribution
b
67%
33%
Overlap (67% of the target distribution)
c
80%
20%
Coverage (80% of the reference distribution)
d
Bhattacharyya coefficient: vs. within the span of
$BC(P,Q) = \sum_i \sqrt{P(i)Q(i)}$; for all i such that $P(i) > 0$

Figure 4.33. Conceptual example of the application of the similarity metrics to two hypothetical reference and target distributions. The density distributions have been depicted with a representation that resembles that of Figure 4.32. The intervals of density percentage delineated by contour lines are arbitrary. Triangle markers indicate the locations of local maxima, only for the fractions of the distributions that are considered in every panel. (**a**) The reference and target distributions in the same system of coordinates. (**b**) Visualization of the overlap: the fraction of the target distribution that is within the span of the reference distribution. (**c**) Visualization of the coverage: the fraction of the reference distribution that is within the span of the target distribution. (**d**) The Bhattacharyya coefficient is calculated between the full reference distribution and the overlapping fraction of the target distribution in the span of the former.

The overlap of the target distribution with the reference distribution was 99.29%. Such a high overlap can already be appreciated in the two-dimensional plots (Figure 4.32), whereby the distributions of the second ensemble exist almost entirely within the span of the distributions of the original ensemble. The measurement of the coverage served to provide a more complete interpretation of the extent of similarity between the distributions. Even though the second ensemble appears to leave uncovered a wide region of the span of the original ensemble, the achieved coverage was as high as 83%. Therefore, the uncovered region only accounted for 17% of the conformational population of the reference distribution. Finally, the BC was 0.86. This score summarized the joint effect of the previously described observations concerning the changes in relative proportions: i) the undersampling of the closed form of the B domain, ii) the absence of the extreme regions along CPC #2, and iii) the various mild shifts in shape or position of the local maxima and other characteristic sites of the topology.

4.1.3.1.4 Comparison between the trajectory ensembles of the WT apo and holo conditions

In the preceding section, the comparative capabilities of CEDA were exemplified by contrasting the two trajectory ensembles of the WT apo condition. Subsequently, the comparison was extended to alternative conditions. The reference apo condition was contrasted with the seven trajectory ensembles of different holo conditions, thus constituting a first major comparative block of central importance for understanding the dynamical traits of PKR and their functional significance. Based on the insights of the dynamical behavior of the A and B domains in the apo condition, the objective was to ascertain whether the different holo conditions exhibited similar or distinct conformational profiles with respect to the reference set of CPCs.

The comparison was conducted by assessing the similarities and differences between the conformational profiles along CPCs #1, #2, and #3 from the apo condition. As shown previously (Figures 4.31 and 4.32), these three CPCs captured the most representative dimensions of the conformational space sampled in equivalent WT apo trajectory ensembles. Accordingly, the holo trajectories (A and B domains) were subjected to structural superposition and data centering around the reference (apo) global average structure. Then, the processed trajectory data was projected onto the vectors of the aforementioned CPCs, and the corresponding density distributions of projection data were generated. The comparison was based on the visual examination of the one- and two-dimensional distributions, shown in Figures 4.34 and 4.35, and the calculation of the similarity metrics (the overlap, coverage, and BC metrics) between the three-dimensional distributions, which have been compiled in Figure 4.36. The two-dimensional plots of CPCs #2 *vs*. #3 have been omitted for the sake of brevity and because their visual inspection did not provide any additional significant information beyond what is already discernible from the other plots.

As a general remark, all holo conditions explored a conformational space that partially or totally overlaps with that of the apo condition at the most characteristic regions. Thus, the different conditions principally differed in the relative proportions between conformational populations that share similar spaces. This observation holds significant value as it validates the use of a single framework of reference CPCs to measure conformational variance among different trajectory ensembles. As a result, from the detected similarities and differences we may infer potential correlations between the distinct conformational profiles of the A and B domains and their enzymatic function. The projection values of the subunits of 2VGB are also displayed to enable comparison with each condition. The corresponding locations substantially align with the regions of higher density of several holo conditions, especially concerning CPCs #1 and #2.

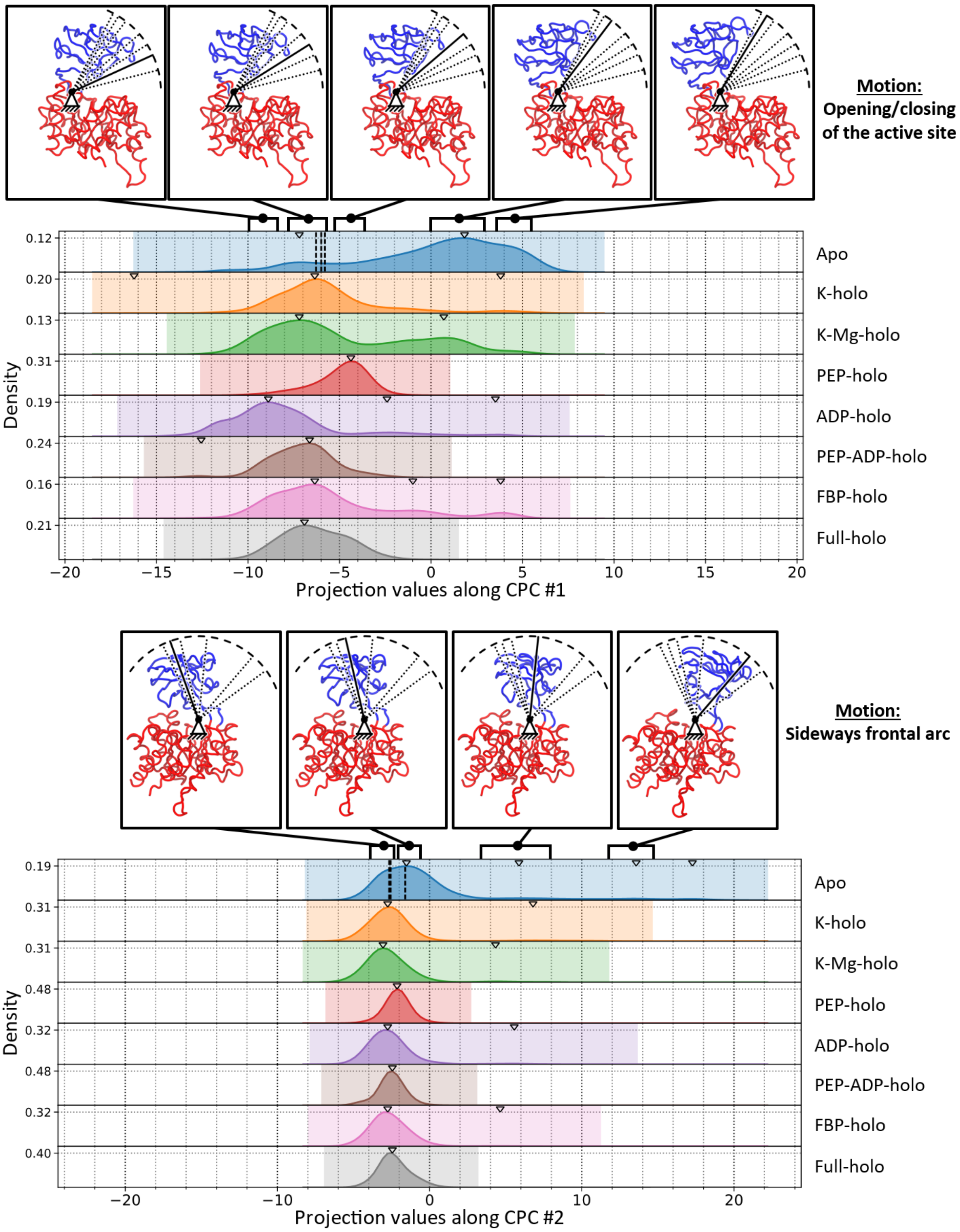
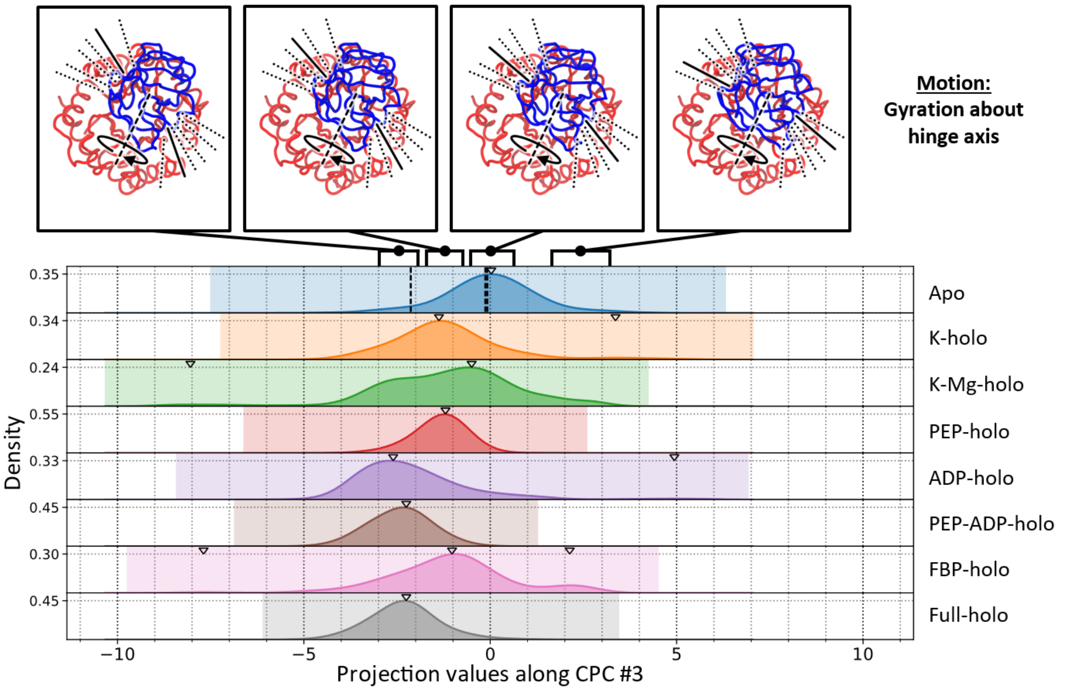


Figure 4.34. Comparative analysis of the WT trajectories (A and B domains) in apo *vs*. holo conditions: conformational profiles (1 CPC). The 1D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 (top), #2 (center), or #3 (bottom) from the reference (WT apo) condition. The data of each condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray. The plots show the KDE curves of the aggregated projection data along with representative structures of the conformation of the A and B domains at different projection values. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and B domains are colored in red and blue, respectively. NOTE. The images of the protein structure were generated with the software VMD.

Figure 4.34 (Continued)



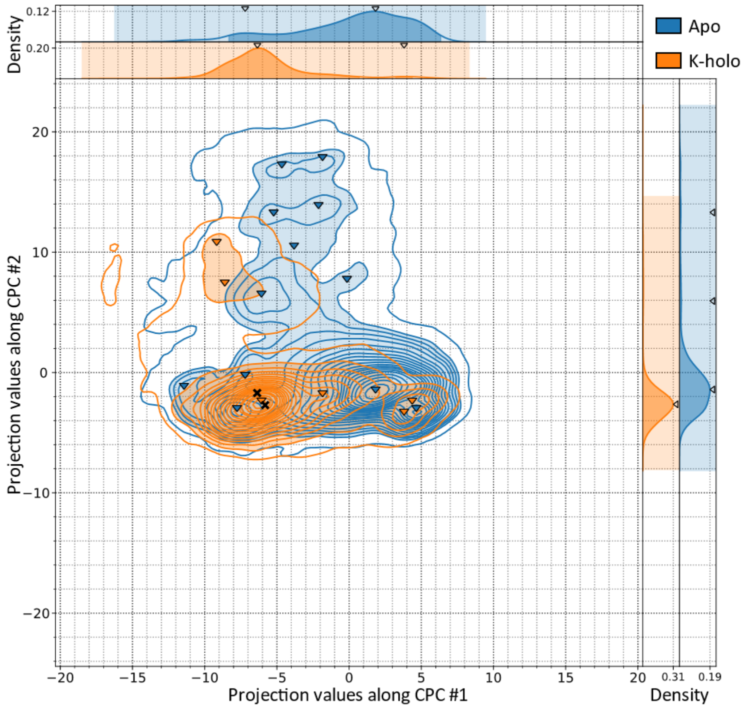


Figure 4.35. Comparative analysis of the WT trajectories (A and B domains) in apo *vs*. holo conditions: conformational profiles (2 CPCs). The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2, #1 *vs*. #3, and #2 *vs*. #3 from the reference (WT apo) condition. The data of each condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. The projection values of the subunits of the structure 2VGB are indicated with “X” markers. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figure 4.35 (Continued)

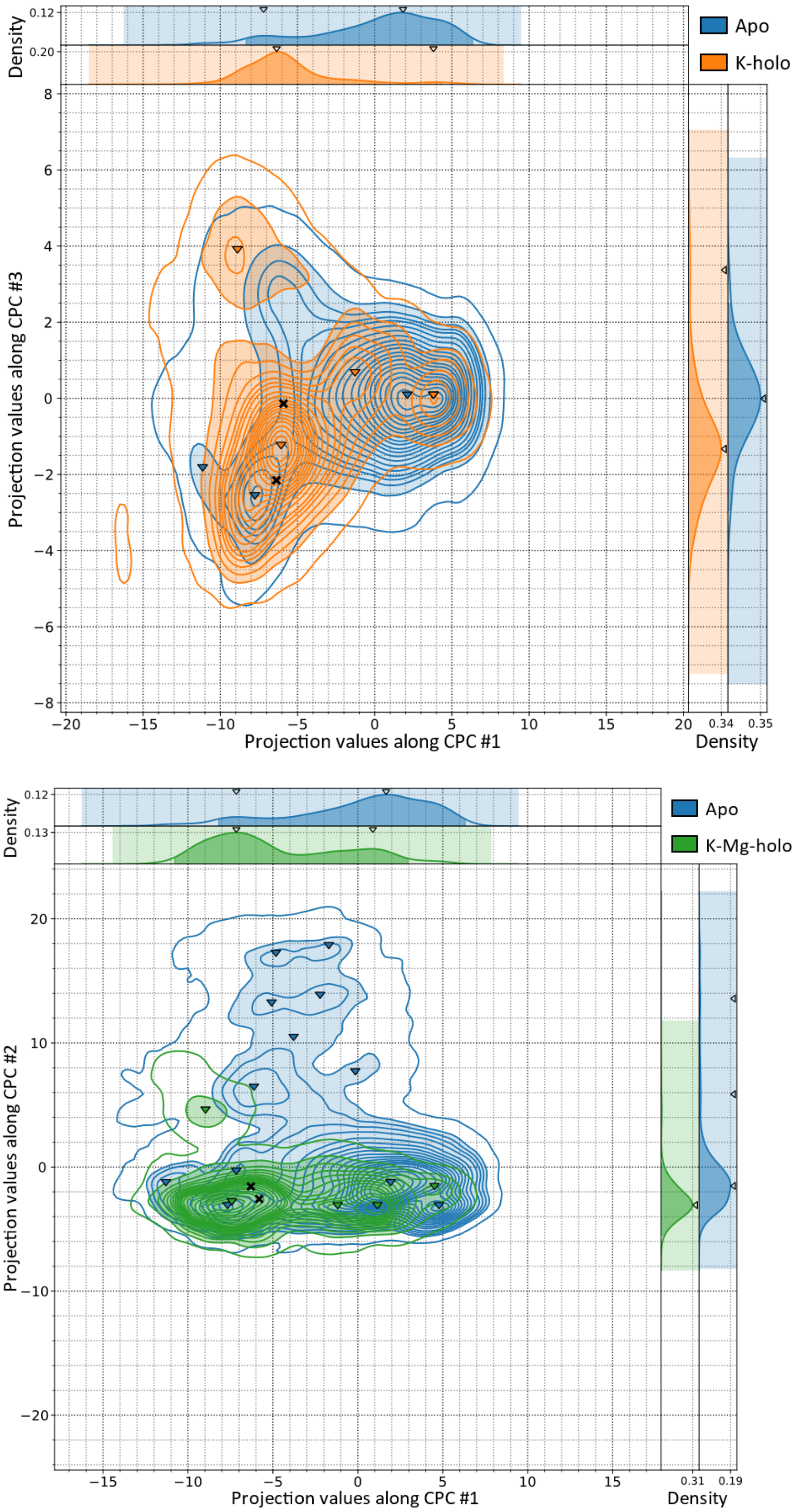
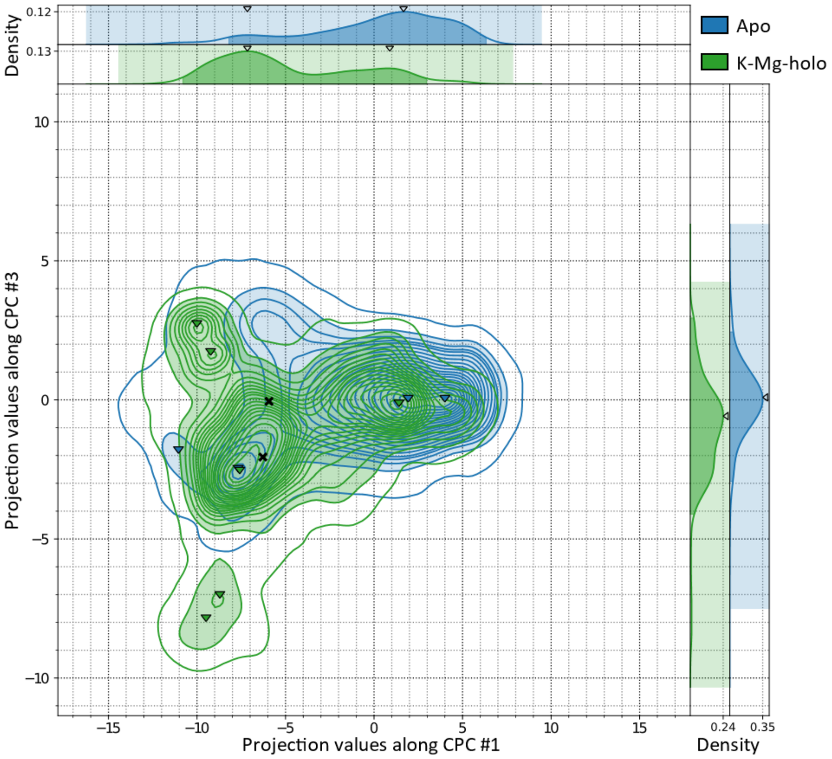


Figure 4.35 (Continued)



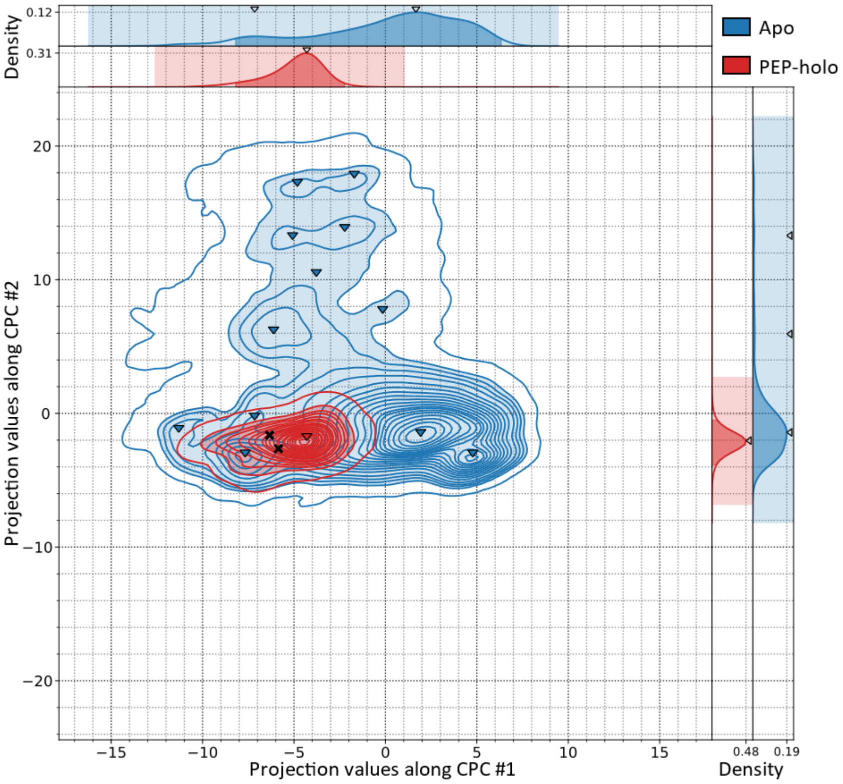


Figure 4.35 (Continued)

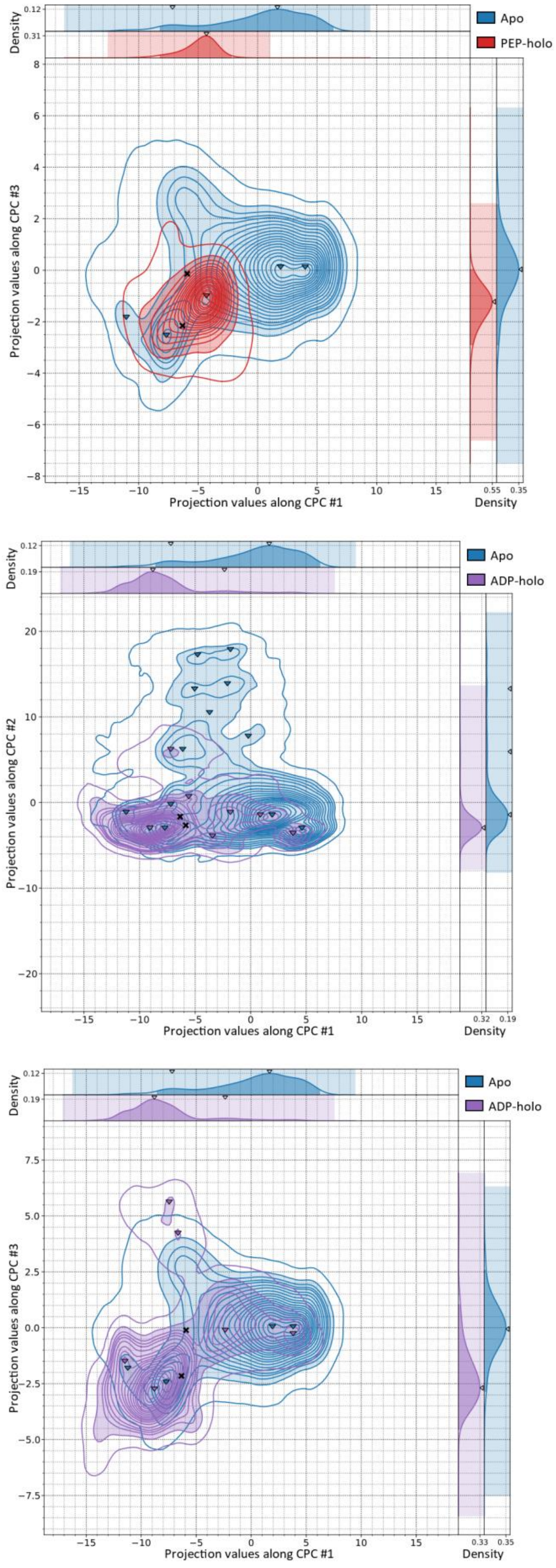


Figure 4.35 (Continued)

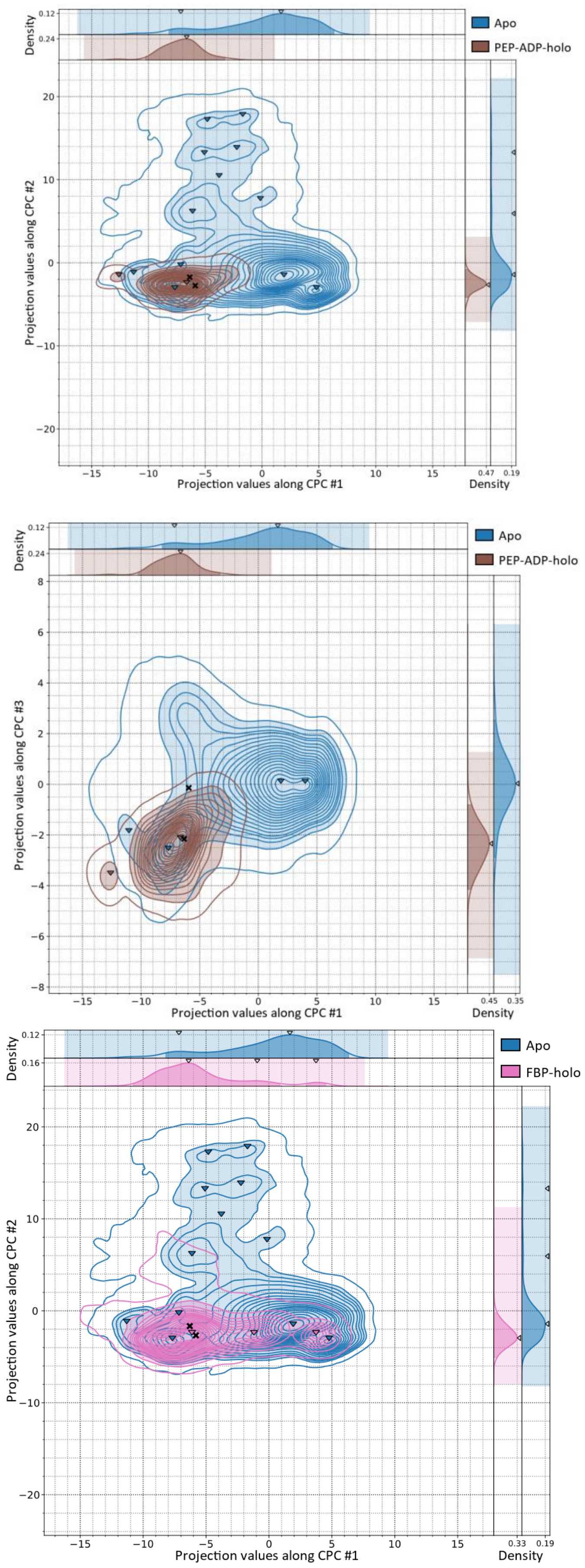
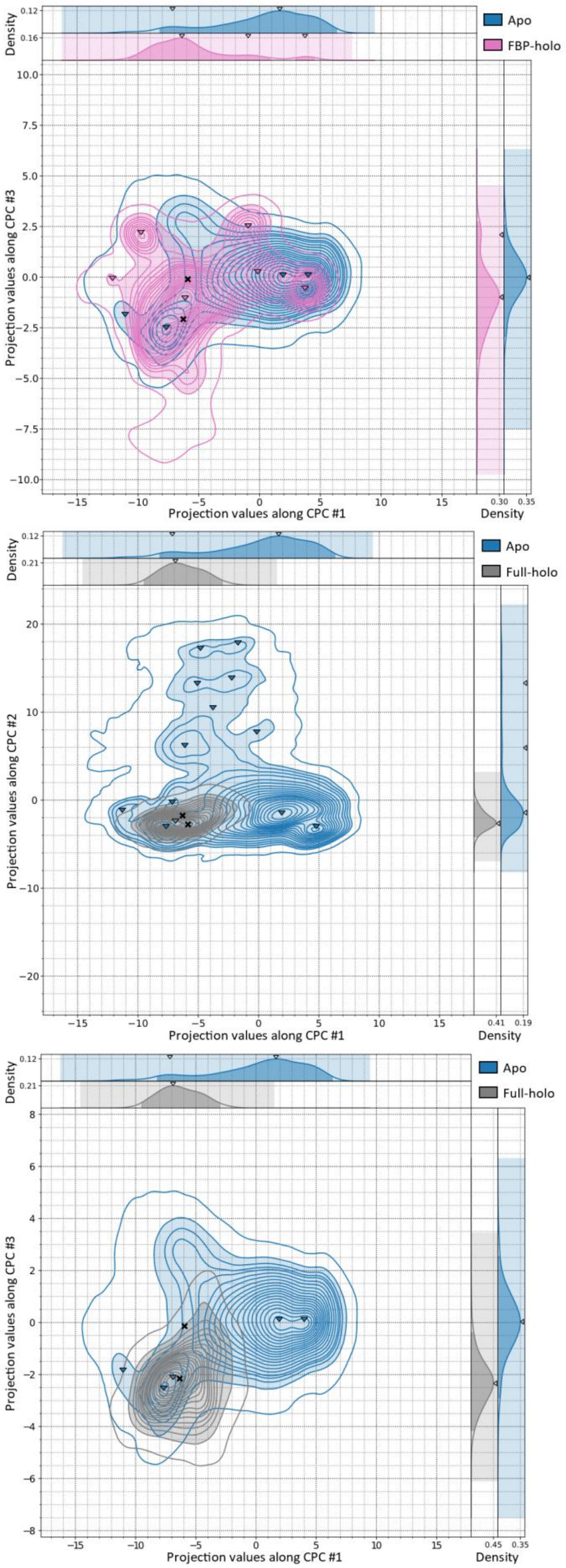


Figure 4.35 (Continued)



Regarding CPC #1, all holo conditions exhibit an overall preference for the closed forms of the B domain. Therefore, the presence of ligands at the active site appears to have exerted an influence to retain a closed conformation. However, sampling of the more open forms of the B domain persisted, implying that such an effect could be described as a shift in the conformational equilibrium. The presence of cofactor K^+ alone was sufficient to induce this effect, as evidenced by the conformational profile of the K-holo condition. The K-Mg-holo condition exhibits the least attenuation of sampling of the open forms of the B domain, despite having both K^+ and Mg^{2+}. The same observation applies to the FBP-holo condition, which also contains both cofactors and exhibits comparable levels of relative conformational proportions. In both of these conditions, Mg^{2+} was modeled at the Mg-3 binding site, based on the structural evidence presented by Zhong *et al*. [137] (see section 4.1.1.2). The higher sampling of the open forms of the B domain in these conditions possibly correlates with the higher dynamical activity of the Mg-3 site hinted at in the stability and flexibility analyses of the metal centers (section 4.1.2.3).

On the other hand, PEP (complexed with Mg^{2+} at the Mg-1 site) is the ligand that imposed the highest degree of conformational restriction, as indicated by the complete permanence of the PEP-holo, PEP-ADP-holo, and Full-holo conditions at the region of the conformational spectrum that mainly corresponds to closed forms. This effect is consistent with the higher stability and lower flexibility of the B domain in these conditions reported in the RMSD and RMSF analyses (sections 4.1.2.1 and 4.1.2.2). The predominant form of the closed B domain in the majority of holo conditions is analogous to that found in the apo condition, with local maxima values between -6 and -8. This characteristic region of the spectrum was also distinctively sampled in the PEP-holo and ADP-holo conditions, however their distributions exhibit shifted local maxima. In PEP-holo, the shift occurred towards a slightly more open form of the B domain, with a local maximum around the value of -4.5. This characteristic form is also observed in the distributions of the other two conditions with PEP, although in lower proportions: residually in PEP-ADP-holo and as a second predominant form in Full-holo. In ADP-holo, the shift occurred towards a more closed form of the B domain, with a local maximum around the value of -9. The corresponding region of this form is also detected with various density levels in other holo conditions. Finally, the ADP-holo condition also significantly sampled a more extreme closed form of the B domain around the value of -11.

Regarding CPC #2, all holo conditions exhibit a major volume of density comparable to that of the apo condition albeit narrower, with distinctive local maxima values between -2 and -4. Regions with projection values greater than 3 were only marginally explored by the conditions that lack PEP. Finally, regarding CPC #3, the holo distributions are centered around different negative values of projection, unlike the apo distribution which is centered around the zero value. The details of the topology of the conformational space along this CPC can be more clearly discerned when inspected as two-dimensional distributions with CPC #1. The presence of ADP (*i.e.*, ADP-holo, PEP-ADP-holo, and Full-holo conditions) appears to have particularly favored the same rotary orientation as that adopted by the B domain in its characteristic closed form in the apo condition. In the K-Mg-holo condition, there is also significant coverage of the predominant conformations of the apo condition. In addition, this condition sampled a distinctive region that corresponds to closed and leftward-oriented forms of the B domain. This very same region was also sampled by the FBP-holo condition, as well as another one that corresponds to partially closed and leftward-oriented forms.

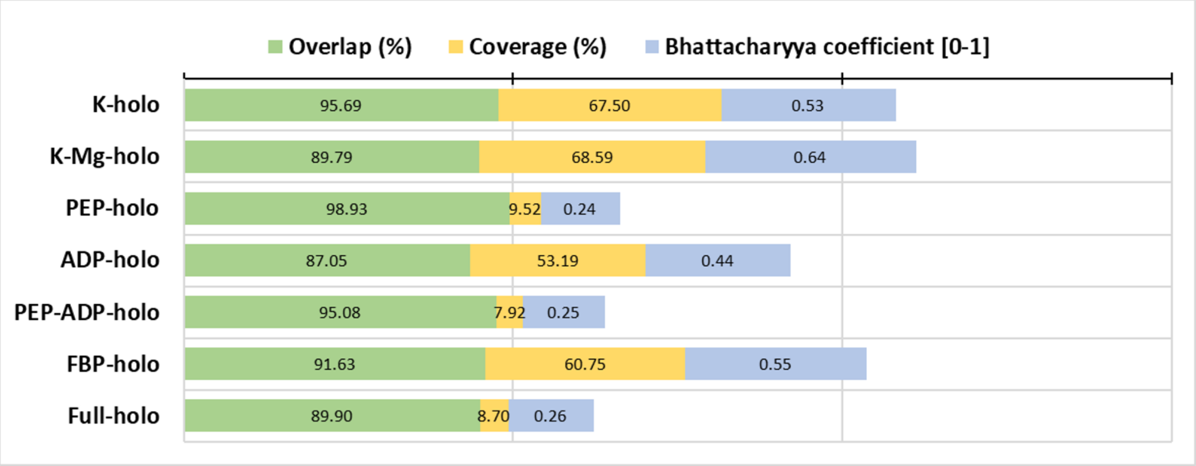


Figure 4.36. Comparative analysis of the WT trajectories (A and B domains) in apo *vs*. holo conditions: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. The bar sizes of the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

The reported similarity metrics (Figure 4.36) allow establishing a single comparative framework to summarize the extent of deviation relative to the reference condition. The three metrics provide complementary insights, enabling the identification of the particular features that contribute to the similarity or dissimilarity between two trajectory ensembles. From the perspective of the overlap metric, all holo conditions manifest a reasonable degree of similarity with the apo condition, as evidenced by the range of high values between 87% and 99%. This can be attributed to the fact that, in all instances, the majority of the sampled conformational space aligns with that of the apo condition. This feature is readily observable through visual analysis of the density distributions. Distributions that exhibit some deviations from the reference region are those that have obtained lower scores within this range of overlap values. These findings suggest that the primary differences between the apo and holo conditions do not arise from the emergence of new significant conformations. Instead, they potentially result from changes in the relative proportions of the conformational equilibrium.

In respect of the coverage metric, the observed wider range of values denotes differential behavior between conditions. The PEP-holo, PEP-ADP-holo, and Full-holo conditions are notable for their exceptionally low values. In these conditions, the B domain was stabilized in its closed conformation, close to the original crystallographic holo structure. Accordingly, the conformational ensemble of the open forms of the B domain, which comprises a considerable portion of the apo condition, was not covered. Conversely, the other holo conditions exhibit intermediate coverage values that are indicative of an increased degree of accessibility to the apo conformational space, albeit significantly incomplete.

Finally, the BC metric displays a range of moderate and low values. The BC was applied to compare the reference distribution with exclusively the overlapping fraction of the target distribution. Thus, the BC manifests the extent of similarity between conditions disregarding the divergent conformational space of the target condition, if any. In this sense, this metric is mainly dependent on the quality of the coverage, and not the overlap. For instance, the density distributions of the PEP-holo, PEP-ADP-holo, and Full-holo conditions are characterized by very localized topologies in comparison to the more extensive distribution of the apo condition. Consequently, low coverage correlates with low BC. Since the location and shape of the distribution is equivalent in all three conditions, the BC score resulted in very similar values between 0.24 and 0.26. In contrast, the other holo conditions are found in a range between 0.44 and 0.64 that accounts for both the partial coverage and the conformational shift between the open and closed conformations.

To complement the comparison of the density distributions, an additional assessment was conducted to determine the consistency with which the reference CPCs accurately represent the intrinsic consensus behavior of the trajectory ensemble of each holo condition. In other words, the aim was to assess whether each trajectory ensemble independently generates CPCs that are analogous to those of the reference CPCs #1, #2, and #3. This enables verification on whether the conformational variations that have been examined in terms of the density distributions result from equivalent collective motions in each case.

The analysis was conducted with the same methodology as in the first half of section 4.1.3.1.3. Accordingly, the procedure of CEDA was applied to the trajectory ensemble of each holo condition, using the same parameters of the original experiment. The resulting sets of CPCs contained between 7 and 10 CPCs. The similarity between these and the reference CPCs #1, #2, and #3 was measured via the cosine similarity metric, expressed in absolute value to disregard differences due to opposite directions. The corresponding results are shown in Figure 4.37, where each radar chart displays the pairwise similarity values between one of the reference CPCs and each of the CPCs of each holo condition.

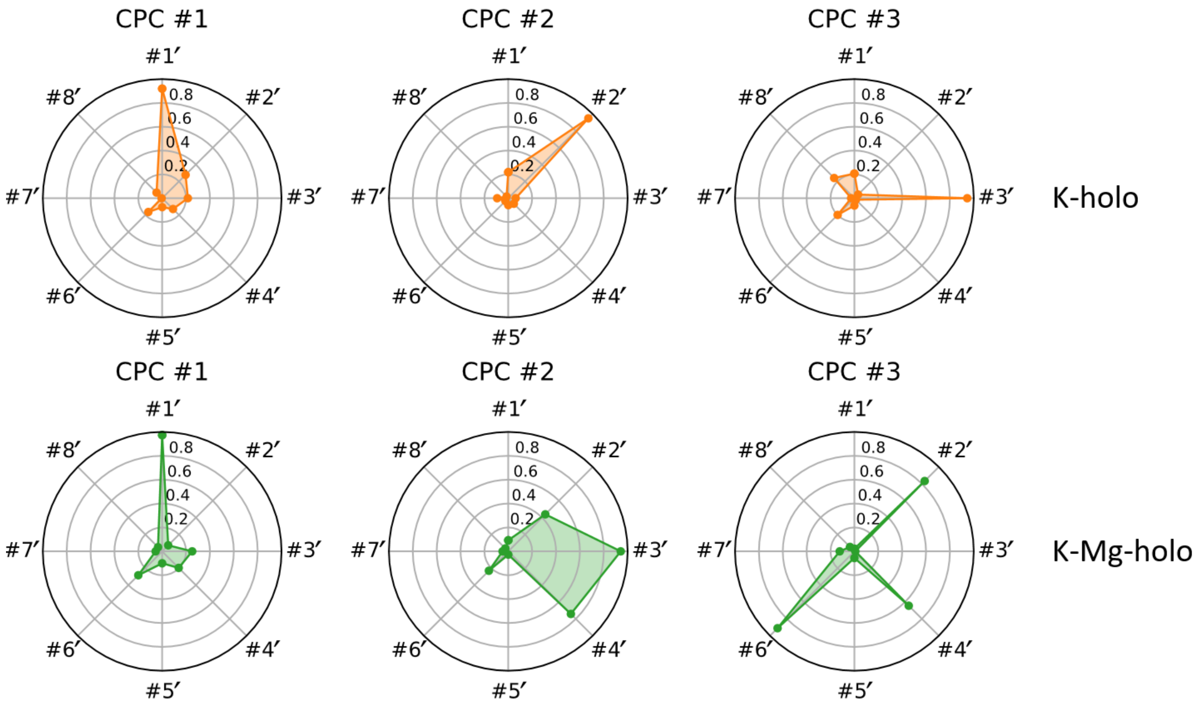
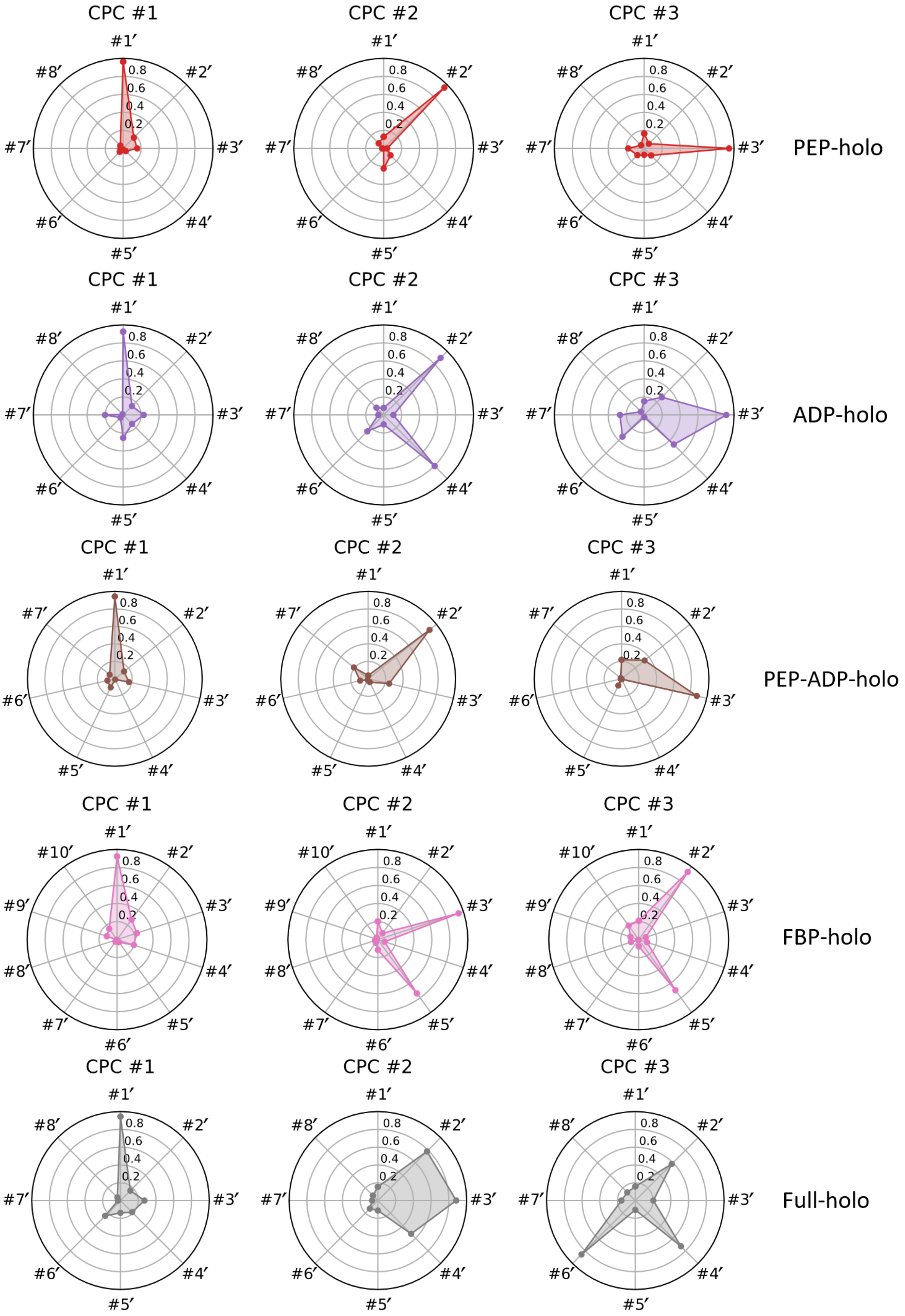


Figure 4.37. Similarity between the WT apo and holo sets of CPCs in the A and B domains. Each radar chart displays the pairwise similarity values between one of the first three CPCs from the apo set (indicated at the caption of each chart) and each of the CPCs from the corresponding holo set (distributed along the angular axis and designated by a prime symbol). Similarity is expressed in absolute value of cosine similarity, bounded between 0 (no similarity) and 1 (full similarity), along the radial axis. The data of each holo set is colored as follows: K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray.

Figure 4.37 (Continued)



Results show a total correspondence of CPC #1 among all conditions, even in those that exhibited less conformational heterogeneity (narrower density distributions), namely, PEP-holo, PEP-ADP-holo, and Full-holo. This confirms that the opening-and-closing motion of the B domain is an ubiquitous dynamical behavior both in the absence and the presence of ligands at the active site, with different degrees of motion amplitude. In respect of CPCs #2 and #3, they are significantly represented in all conditions, displaying high values of similarity. However, a one-to-one correspondence is only detected in K-holo, PEP-holo, and PEP-ADP-holo, whereas the other conditions exhibit both high similarity with one CPC and also moderate similarity with one or two more CPCs. This suggests that the collective motions of CPCs #2 and #3 are broadly manifested among the trajectory ensembles, albeit with less uniformity.

4.1.3.2 Collective motions of domains and subunits across the PKR tetramer

The homotetrameric structure of PKR is characterized by a symmetrical arrangement in a dimer-of-dimers formation. Each subunit possesses two primary points of contact with the adjacent subunits. Depending on the reference point of contact, two distinct pairs of subunits, or dimers, can be identified. The A-A′ dimer is formed through the A-A′ interface between adjacent A domains, while the C-C′ dimer is formed through the C-C′ interface between adjacent C domains. In such an assembly, the collective motions that involve the A and C domains are intricately coupled across the subunits and have been suggested to drive the integral conformational changes identified in the crystallographic studies of pyruvate kinases.

Two independent CEDAs were performed to analyze the dynamical events at each of the two types of intersubunit contacts. The methodology employed is analogous to that of the experiment on the collective motions of the A and B domains, which was reported in detail along with several complementary assessments of the robustness of the results. For these two subsequent analyses, a more concise version of the experimental section is provided to emphasize the most relevant results.

4.1.3.2.1 The A-A′ dimers

The analysis of the A-A′ dimers was aimed at examining the collective motions between the adjacent A domains. The A-A′ interface is the closest contact point between the active sites of the tetramer. Consequently, the structural rearrangements in this region are potentially involved in a transmission of information between subunits, such that the binding of PEP in one subunit leads to higher substrate affinity in the neighboring subunit. Accordingly, the identification of distinctive dynamical events may provide crucial insights into understanding the phenomenon of cooperativity in this enzyme.

The CEDA procedure was applied to the WT apo trajectories to derive the corresponding set of CPCs. Each tetramer comprises two pairs of adjacent A domains, resulting in 10 analyzable A-A′ pairs of domains from the 5 simulations of the WT apo tetramer. The global average structure of all A-A′ pairs of the trajectory ensemble was generated and used as the general reference for structural superpositions. For each A-A′ pair, two complementary PCAs were performed, alternating the role of the fitting group between both domains (*i.e.*, A-A′ with A as the fitting group and A′-A with A′ as the fitting group; in both cases superposed to A of the global average structure). This approach provided a total of 20 analyzable trajectory replicates, designated by the number of the simulation (#1 to #5) and the pair of chain IDs of the involved subunits, with the first ID corresponding to the superposed structure. This strategy was adopted to determine whether the analysis of dynamical data of the adjacent A domains, when conducted from each respective point of reference, yields information that is identical or complementary in nature.

Only the Cα atoms of the region (488 atoms) were considered in the analysis, therefore each PCA yielded a set of 1464 eigenvectors. The first 20 eigenvectors of each trajectory replicate were included in the clustering analysis, accounting for approximately 85% of the cumulative variance on average (Figure 4.38). The assessment of the cosine content identified 6 instances of PCs with values exceeding the threshold of 0.5 (Figure 4.39). These instances correspond to the PC #1 of replicates #3-BD, #3-DB, #4-BD, #4-DB, #5-AC, and #5-CA. It is noteworthy that the same three A-A' pairs of domains are represented twice in these PCs from both of their points of reference. This observation is consistent with the expectation that, for a given relative collective displacement between a pair of domains, the respective PC values from each point of reference may display similar time-series progression patterns, regardless of whether the displayed motion is symmetrical or not. The degree of integration of these PC instances of high cosine content in the clusters of higher coverage will clarify the significance of the corresponding captured collective motions.

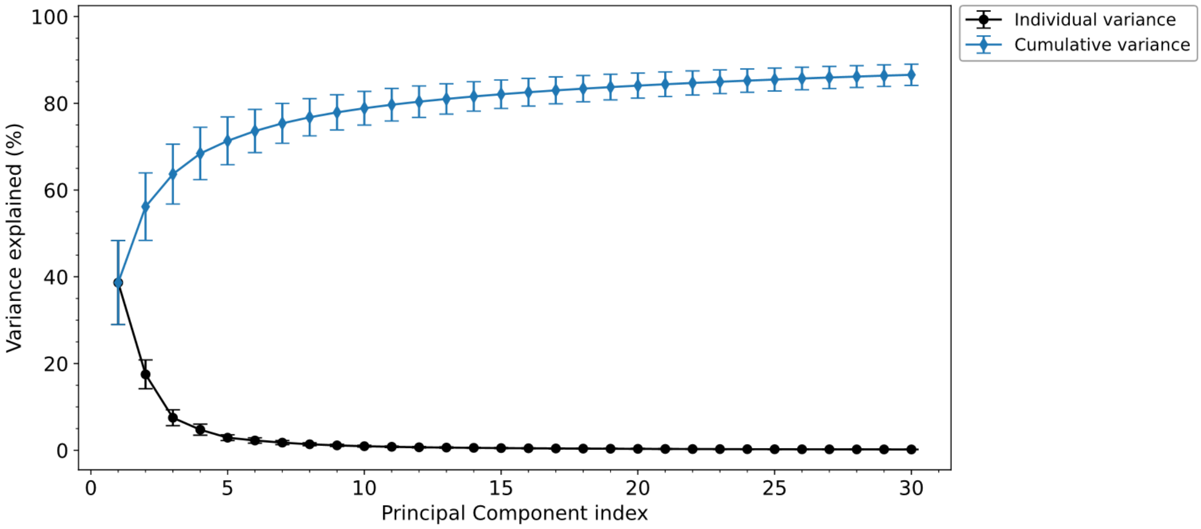


Figure 4.38. Percentage of variance explained by the PCs of the A-A' pairs of domains in the WT apo condition. Data corresponds to the mean values from the first 30 PCs of the 20 trajectory replicates of the experiment. Standard deviation intervals are included as error bars.

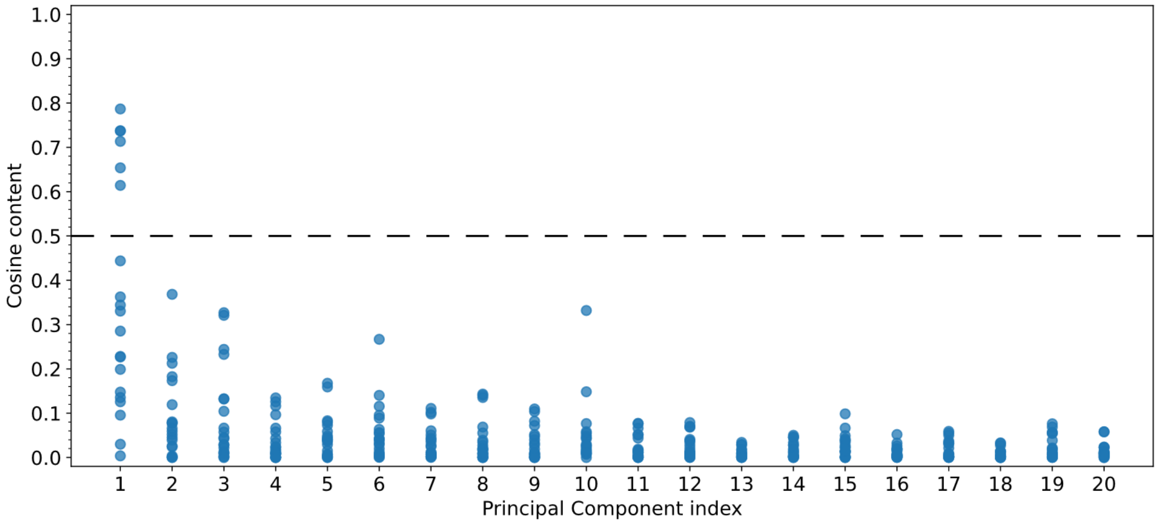


Figure 4.39. Cosine content values of the PCs considered in the CEDA of the A-A′ pairs of domains in the WT apo condition. The cutoff value of 0.5 below which the sampling is considered to be acceptable in protein dynamics is shown with a horizontal dashed black line.

A cophenetic distance cutoff of 0.45 was applied to split the dendrogram of the hierarchical clustering (Figure 4.40), and clusters with less than 20% of coverage were discarded, thus retaining a total of 5 relevant clusters. The primary division of the dendrogram allows identification of two distinct sections. The lower section contains 2 large clusters of full coverage, namely, clusters #1 and #2. The upper section of the dendrogram is characterized by scattered eigenvectors that only form smaller clusters of both lower similarity and coverage values. Clusters #14, #19, and #21 meet the established criteria to produce CPCs, with 65%, 20%, and 20% of coverage, respectively. The centroid vector of the 5 relevant clusters was computed to acquire the CPCs of the experiment, numbered from #1 to #5 in decreasing order of coverage and average percentage of variance of their cluster members.

Next, the collective motion captured in each CPC was visually examined. Figure 4.41 shows schematic diagrams of the A-A′ pair of domains along with the terminology that will be employed to refer to the observed motions. The structure was arranged in a reference orientation with the A′ domain (mobile region) piled on top of the A domain (anchored region; fitting group), thus enabling a clear visualization of the rotational and translational degrees of freedom of the former. In the front view of the structure, the top of the barrel of the A′ domain (active-site cleft) is oriented towards the observer. Figure 4.42 shows the schematic representations of the path and extreme conformations of the collective motions from CPCs #1 to #3.

The collective motion captured in CPC #1 (Figure 4.42a, Supplementary Video S4.11) shows a combination of roll and sway. The resulting motion can be described as the rolling of the barrel structure of the A′ domain along the contact interface with the A domain. The collective motion captured in CPC #2 (Figure 4.42b, Supplementary Video S4.12) shows a combination of pitch and surge that results in a seesaw-like swinging motion of the barrel structure of the A′ domain, with the contact interface with the A domain acting as the pivot point. Together, the clusters of these two CPCs agglomerated all instances of PCs with indices #1 and #2 of the trajectory ensemble. The index number did not determine the cluster membership of these PCs; rather, both clusters contained a mixture of the two types. Specifically, cluster #1 (CPC #1) agglomerated 12 instances of PC #1 and 8 instances of PC #2, while cluster #2 (CPC #2) contains the complementary combination.

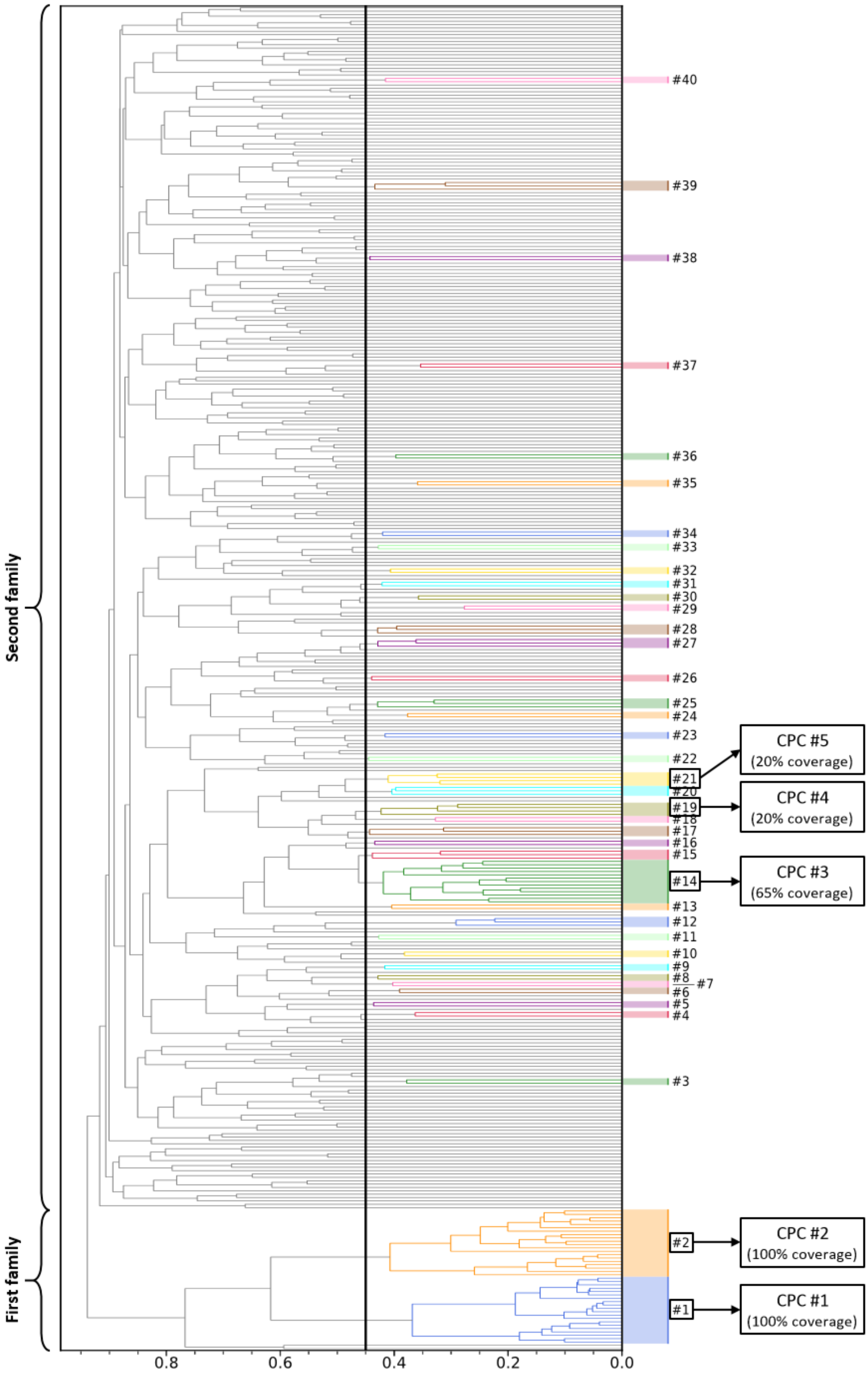


Figure 4.40. Dendrogram of the CEDA of the A-A' pairs of domains in the WT apo condition. The cophenetic distance cutoff of 0.45 used to split the dendrogram and obtain clusters is shown with a vertical black line. Non-singleton clusters are shown in various colors to facilitate visual identification. The 5 clusters with at least 20% of coverage were selected to acquire the CPCs of the experiment. CPCs were numbered from #1 to #5 in decreasing order of coverage and average percentage of variance of their cluster members. The figure indicates which clusters yielded each CPC, together with their achieved coverage. At the left margin of the dendrogram, the span of the detected families of clusters is indicated with labeled curly brackets.

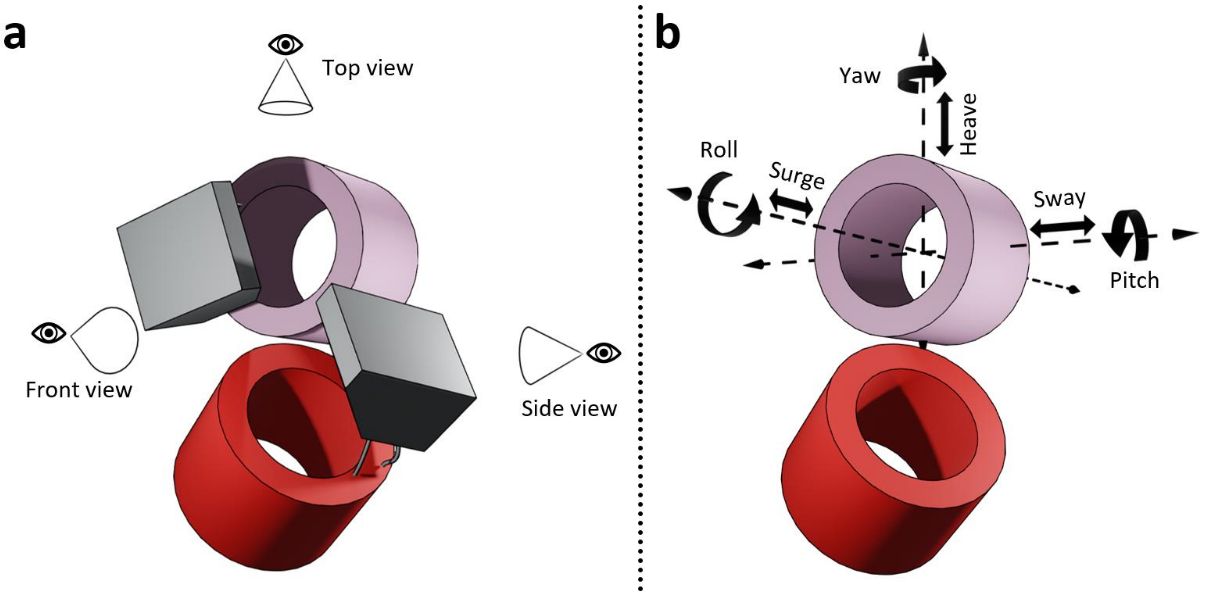
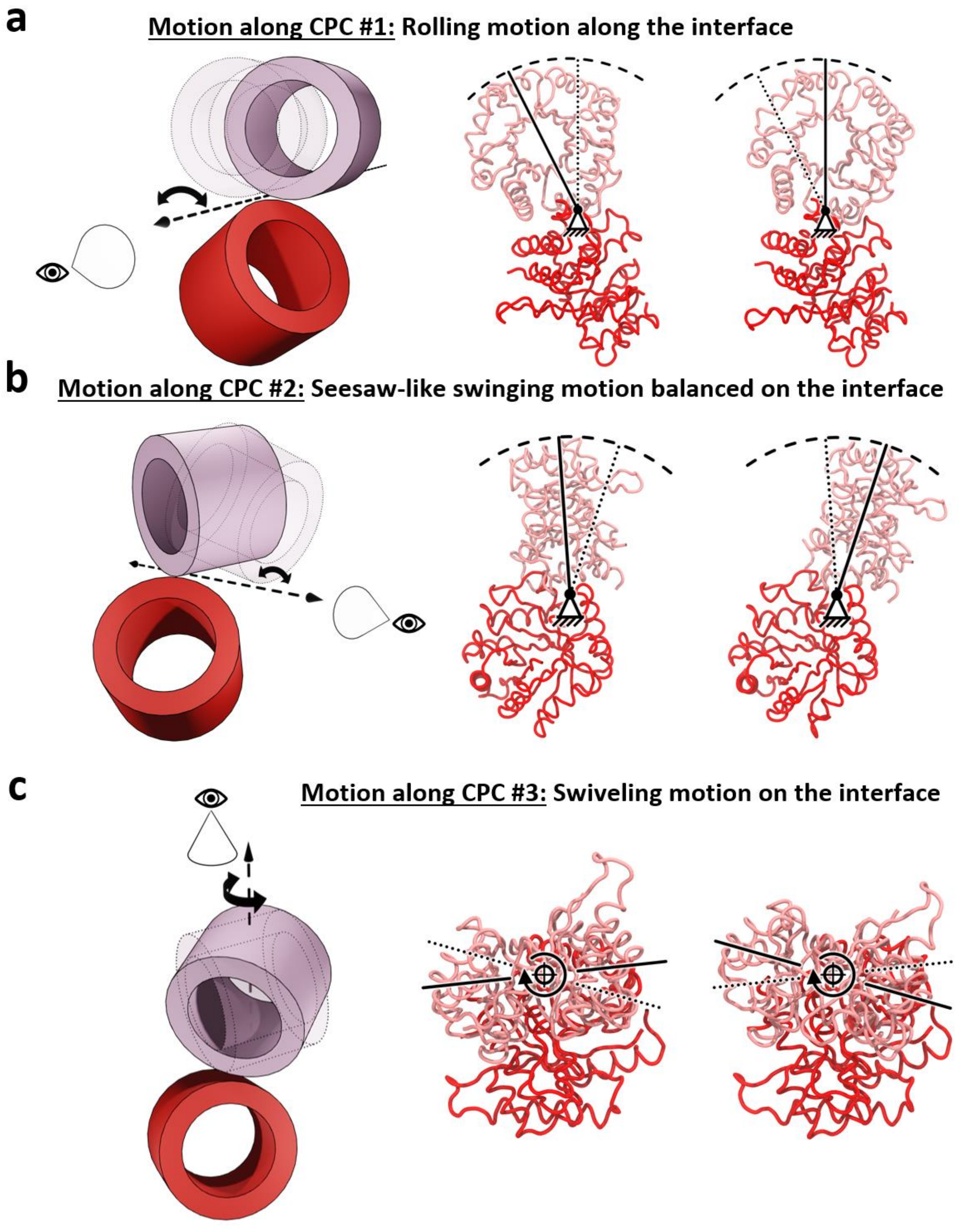


Figure 4.41. Schematic diagrams of the adjacent A domains across the A-A′ interface of PKR. Various markers indicate the terminology to refer to (**a**) the reference views and (**b**) the directions of motion along the six mechanical degrees of freedom of the A′ domain as a rigid body. The A and A′ domains are colored in red and pink, respectively. In (a), the B domains of each subunit are also depicted (gray) to show that the top of the barrel of the A′ domain is oriented towards the observer in the top view of the structure. NOTE. The 3D schematic models were built with the software Blender.

Moreover, the consideration of the two alternative points of reference (fitting groups) per pair of adjacent domains did not account for duplicated information but rather served to enrich the sample of the two types of consensus motions of the ensemble. Otherwise, the PC counterparts (equal indices) from each version of the same trajectory would have been consistently classified in alternative clusters as a consequence of having captured two distinct perspectives of a single collective displacement. For each pair of adjacent A domains, PCs of the same index from A-A′ and A′-A co-localize in the same cluster on 6 occasions, while the opposite case is observed on 4 occasions. Finally, in connection with this matter, the PCs with high cosine content are well integrated into these two clusters, thus confirming the significance of their captured collective displacement of coordinates despite their PC projection values having stronger resemblance to cosine shapes. All these observations suggest that the sampling of this protein region principally occurred within the conformational space defined by the dimensions of CPCs #1 and #2, which captured comparable degrees of conformational variance.

The collective motion captured in CPC #3 (Figure 4.42c, Supplementary Video S4.13) features the yaw rotation of the A′ domain. It can be described as a swiveling motion of the barrel structure of the A′ domain, with the contact interface with the A domain acting as the joint. This CPC comprises PCs with indices between #3 and #5. The motions along CPCs #4 and #5 mostly show variations of the former three and comprise PCs with various indices between #3 to #8. CPC #4 features a combination of heave and yaw that results in the swiveling motion of CPC #3 but with an additional displacement whereby the A′ domain approaches or separates from the A domain. CPC #5 features a combination of pitch and heave that results in a variation of pitch with the back region acting as the pivot point. Such motions can be viewed in the Supplementary Videos S4.14 and S4.15.



Motion along CPC #1: Rolling motion along the interface

Motion along CPC #2: Seesaw-like swinging motion balanced on the interface

Motion along CPC #3: Swiveling motion on the interface

Figure 4.42. Consensus collective motions of the adjacent A domains across the A-A′ interface in the WT apo condition. (**a-c**) Schematic representations of the path and extreme conformations of the motion captured in CPCs #1, #2, and #3. In each panel, the protein domains are depicted with the diagrams of Figure 4.41 (left) and with the trace representation between Cα atoms (center and right). The A and A′ domains are colored in red and pink, respectively. These motions can also be viewed in the Supplementary Videos S4.11 to S4.13. NOTE. The images of the protein structure were generated with the software VMD. The 3D schematic models were built with the software Blender.

The observed motions mainly involve rigid-body displacements of the adjacent A domains relative to one another. However, due to the structural arrangement of the interdomain joint, such motions additionally transmit local conformational changes at the top of the barrel structure, where the active site of the enzyme is located. Specifically, the width of the active-site cleft fluctuates in a correlated manner with the major motions of CPCs #1, #2, and #3. The corresponding local conformational changes are shown in detail in Figure 4.43 and the Supplementary Videos S4.16, S4.17, and S4.18 to enable identification of the involved structural elements. The active-site cleft of the A domain narrows in correlation with the rigid-body motions of the A′ domain, namely, the rolling motion towards its left (CPC #1; Figure 4.43a), the seesaw-like swinging motion towards its front (CPC #2; Figure 4.43b), and the clockwise swiveling motion (CPC #3, Figure 4.43c). The narrowing of the active-site cleft occurs via the mutual approach of the following elements at the top of the barrel: L-Aβ2-Aα2, L-Aβ3-Bβ1, L-Bβ8-Aα3, Aα3, Aα6′, Aα6, Aα7′, L-Aα7′-Aα7 and Aα7. These observations are consistent with previous crystallographic studies that report a potential cooperative mechanism whereby subunit rotation and the restructuring of Aα6′ and its vicinity are coupled to prime the adjacent PEP binding sites to bind substrate more efficiently [127], [128], [132], [149].

On another note, a few elements at the base of the A-domain barrel also undergo slight rearrangements in correlation with the rigid-body displacement of the A′ domain. These can be especially noted in CPC #1 (Figure 4.43a) and involve the following elements: L-Nα2-Aβ1 and Aα8, which are the linker fragments with the N-terminal and C domains), and L-Aα2-Aβ3 that is a loop that notably protrudes towards the C domain. Consequently, the relative reorientation of the adjacent A domains appears to have an impact in the relative conformations of both the N-terminal and C domains, with potential successive global rearrangements of the whole tetramer.

After the derivation of the CPCs and the examination of the captured motions, the experiment proceeded with the exploration of the corresponding conformational profiles. The probability density functions of the projection data along each CPC were estimated with KDE, using a bandwidth of 0.3 and representing the curves with 100 points along the domain of the data. Only the conformational profiles along CPCs #1, #2, and #3 were explored as they constitute the clusters of greatest coverage and conformational sampling variance of the experiment. Figures 4.44, 4.45, and 4.46 show the one-dimensional density distributions along each of these CPCs, both for each individual trajectory replicate and for the aggregated projection data. Representative structures of the A-A′ pair of domains at different projection values are included to inform about their correspondence with the range of conformations. The projection values of the subunits of the structure 2VGB are also indicated in the figures.

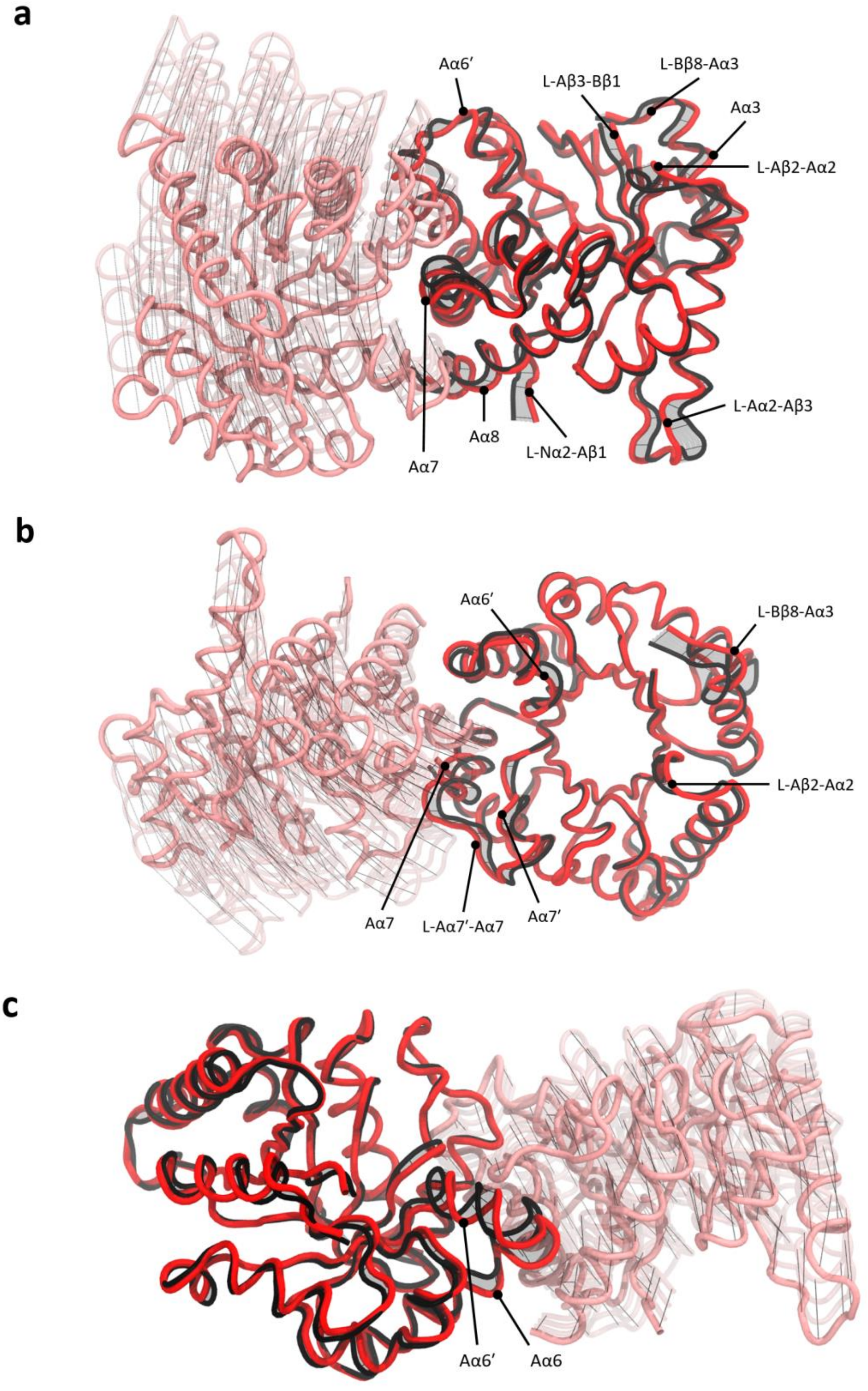


Figure 4.43. Local structural rearrangements of the A domain correlated with the collective motions of the adjacent A′ domain in the WT apo condition. (**a-c**) Schematic representations of the path and extreme conformations of the motion captured in CPCs #1, #2, and #3. In each panel, the protein is depicted with the trace representation between Cα atoms. The initial conformations of the A and A′ domains are colored in red and pink, respectively. Structures colored in black correspond to the final conformations of the A domain. Several intermediate conformations are also shown in transparent gray for the A domain and in pink for the A′ domain. The black lines connecting the intermediate structures indicate the path of motion of Cα atoms. These motions can also be viewed in the Supplementary Videos S4.16, S4.17, and S4.18. NOTE. The images were generated with the software VMD.

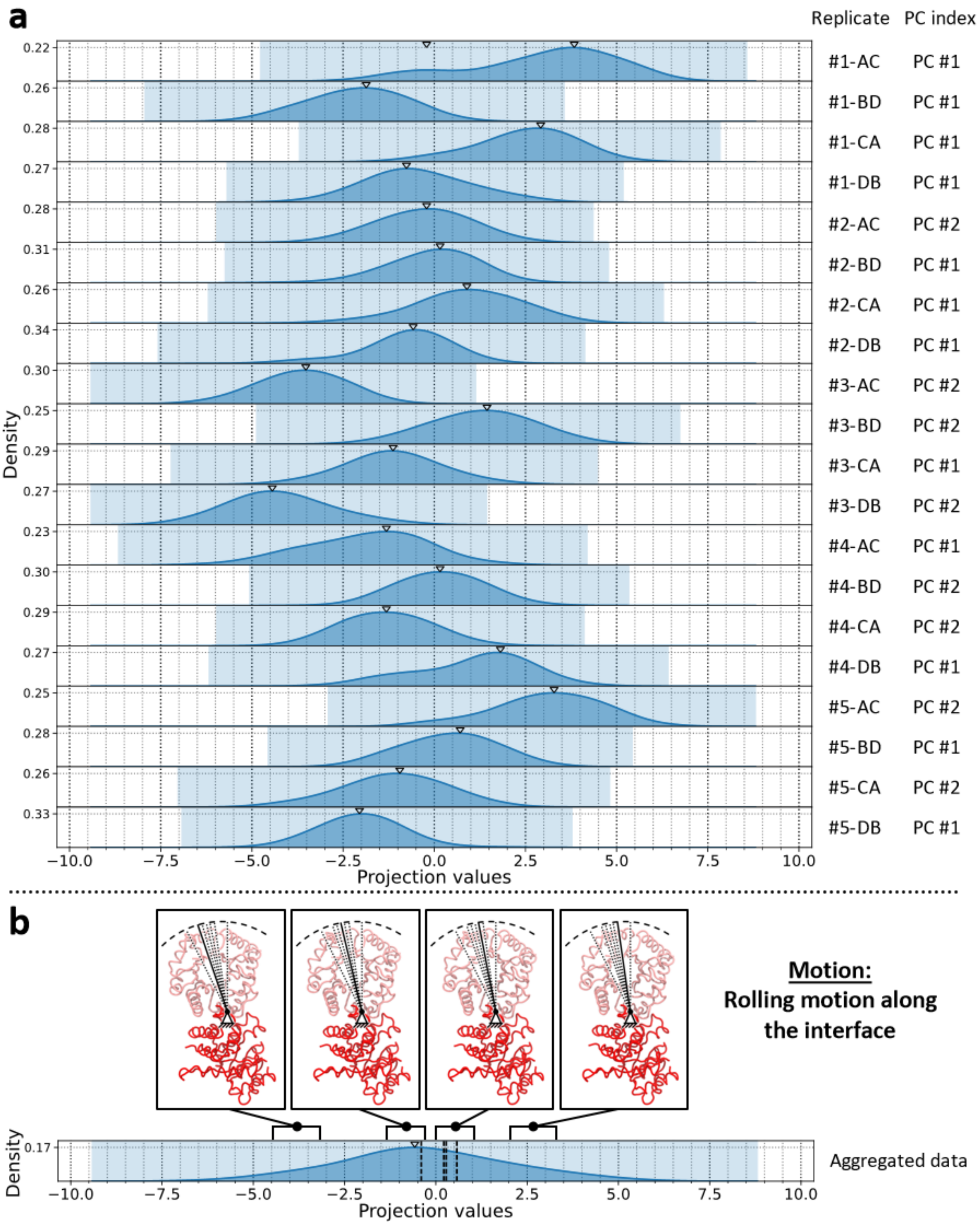


Figure 4.44. Density distributions along CPC #1 of the A-A′ pairs of domains in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the A-A′ pair of domains at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and A′ domains are colored in red and pink, respectively. NOTE. The images of the protein structure were generated with the software VMD.

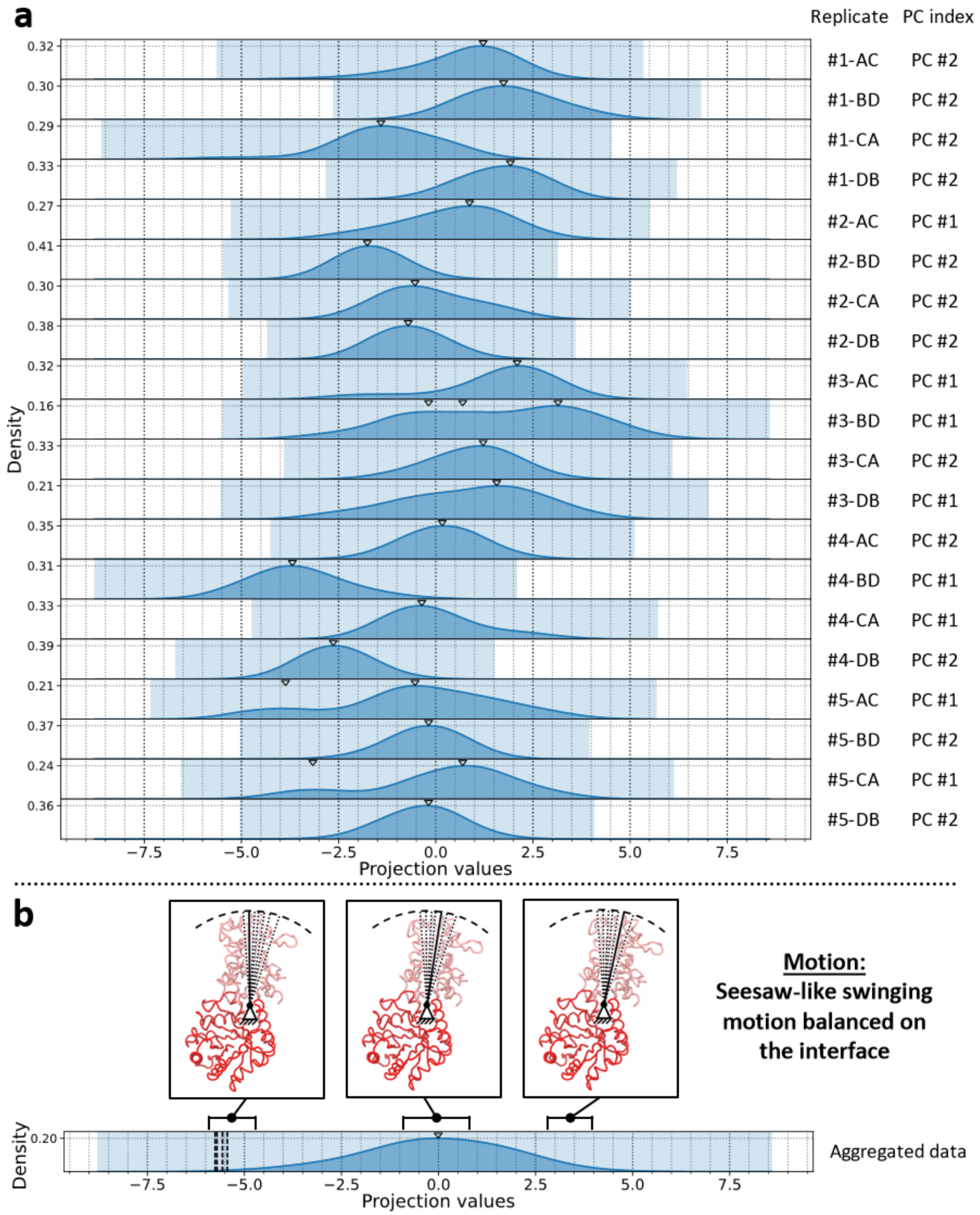


Figure 4.45. Density distributions along CPC #2 of the A-A′ pairs of domains in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the A-A′ pair of domains at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and A′ domains are colored in red and pink, respectively. NOTE. The images of the protein structure were generated with the software VMD.

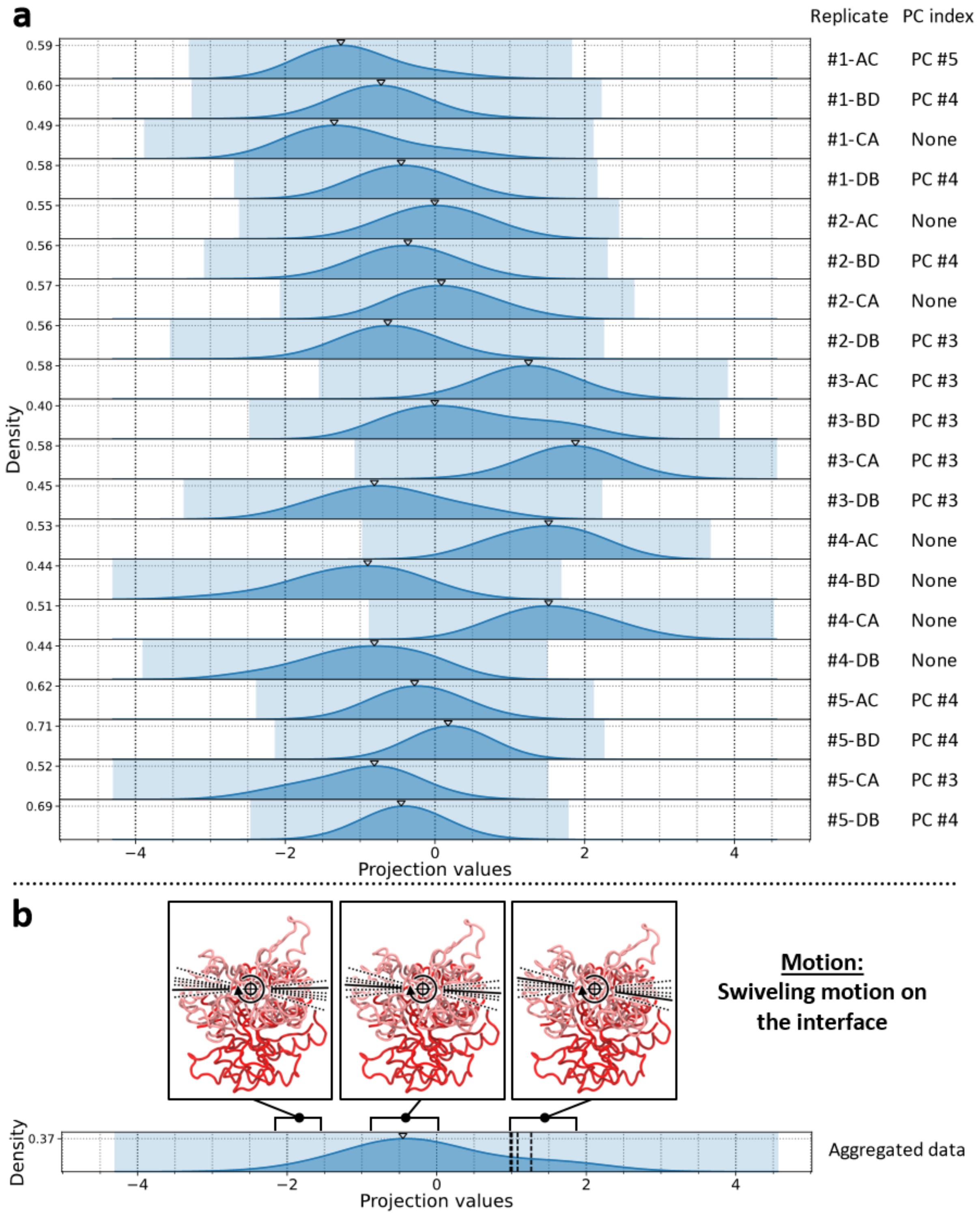


Figure 4.46. Density distributions along CPC #3 of the A-A′ pairs of domains in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the A-A′ pair of domains at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and A′ domains are colored in red and pink, respectively. NOTE. The images of the protein structure were generated with the software VMD.

The aggregated projection data along each CPC generate unimodal distributions with large dispersion. The total span of projection values diminishes as the CPC index increases. The distributions of the aggregated data are mostly centered around zero or nearby values, suggesting that the structures sampled conformations around an energy minimum close to the global average structure of the ensemble. The distributions of the individual trajectory replicates have diverse locations in the spectrum of values, although, in some instances, there is a certain correspondence in the conformational profiles between the two trajectory replicates of the same pair of domains with alternative fitting groups (*e.g.*, #5-AC with #5-CA, or #5-BD with #5-DB, along CPC #2).

The structure 2VGB displays different degrees of divergence with respect to the trajectory ensemble at each CPC. For CPC #1, both the local maximum and the 2VGB values have similar conformations with projection values near zero. For CPC #2, however, the local maximum matches the global average structure, while 2VGB exhibits a conformation at the negative extreme values of the spectrum, around -5.5. This conformation has the A′ domain tilted forward and the active-site cleft of the A domain narrowed, consistent with the presence of the substrates. This implies that the apo simulations generally left the initial holo conformation of the crystallographic structure and adopted a range of conformations around the center of the spectrum of CPC #2, which are potentially more stable without ligands. Finally, the conformational profile of CPC #3 suggests a conformational equilibrium between the most sampled form of the trajectory ensemble, with a local maximum around -0.5, and another form between 1 and 1.5 that is distinctive of 2VGB and persisted in a few replicates. This conformation has the frontal region of the A′ domain oriented to its right and, again, a narrowed active-site cleft of the A domain.

The exploration of the conformational profiles was complemented with the generation of the two-dimensional density distributions (Figure 4.47). In this case, the visual examination of the corresponding plots did not provide additional insight. No major topological irregularities or singularities can be discerned besides those implied in the one-dimensional distributions. In general, the three dimensions appear to define an energy minimum and its surroundings. The deviation of 2VGB from the sampling region of the trajectory ensemble is especially noted in the distributions of CPC #1 *vs*. #2 and #2 *vs*. #3.

In a similar manner as shown in the second half of section 4.1.3.1.3, the additional set of WT apo simulations (#6 to #10) was employed here to evaluate the replicability of the obtained conformational profiles with new equivalent trajectories of this condition. On the one hand, Figure 4.48 shows the overlays of the pairwise two-dimensional plots of CPCs #1 to #3, which were visually compared to assess the similarities and differences in the sampling along the reference collective motions. On the other hand, the quantitative indicators (the overlap, coverage, and BC metrics) were calculated for the comparison between the corresponding three-dimensional distributions.

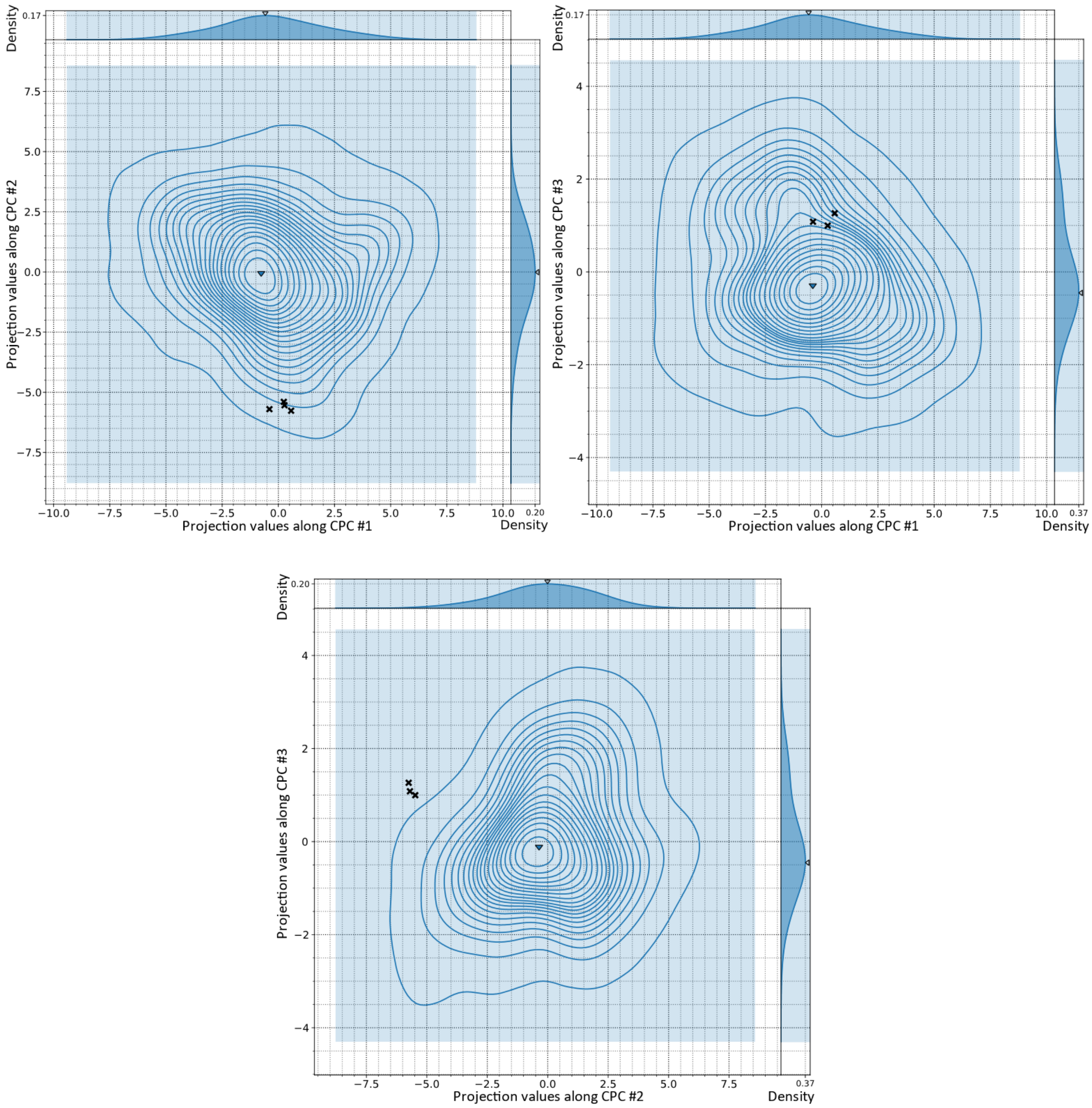


Figure 4.47. Pairwise 2D density distributions along CPCs #1 to #3 of the A-A′ pairs of domains in the WT apo condition. The highlighted rectangular regions indicate the total span of projection values. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Triangle markers indicate the locations of local maxima. The projection values of the subunits of the structure 2VGB are indicated with “X” markers. Each two-dimensional plot also features subplots of the one-dimensional KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins.

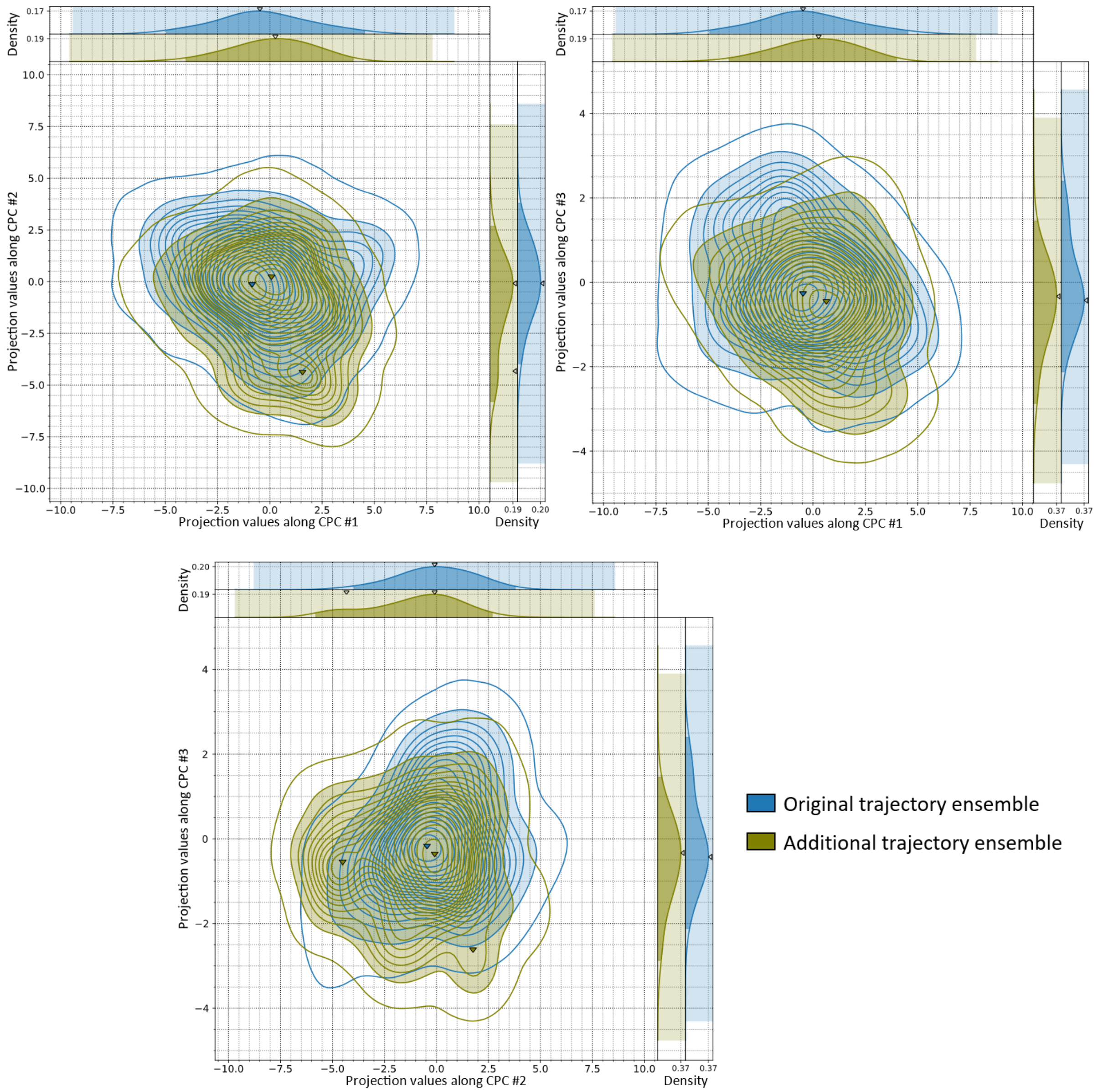


Figure 4.48. Comparative analysis of the original *vs*. additional WT trajectory ensembles (A-A' pairs of domains) in the apo condition: conformational profiles (2 CPCs). The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 (top-left), #1 *vs*. #3 (top-right), and #2 *vs*. #3 (bottom) from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Among the noteworthy differences between the conformational profiles, the second trajectory ensemble exhibits a characteristic region of moderate density around the value of -4.5 of CPC #2, closer to the values of the structure 2VGB. The rest of the sampling appears to be substantially equivalent, albeit with small differences in the span at the peripheral regions of the distributions. The overlap between the two WT apo trajectory ensembles was 87.41%. This value reflects the existence of a fraction of density of the second trajectory ensemble (12.59%) that lies outside of the reference region of the original trajectory ensemble. The coverage was 85.38%, and mainly reflects the lower degree of dispersion of the second trajectory ensemble, particularly along CPCs #1 and #3. The BC was 0.88, and captures the combination of the aforementioned effects.

Following from the characterization of the results of the WT apo condition, the next stage of the analysis revolved around the comparison with the conformational profiles of the WT holo trajectory data when projected onto the reference set of CPCs. The comparison was based on the visual examination of the one- and two-dimensional distributions, shown in Figures 4.49 and 4.50, and the calculation of the similarity metrics between the three-dimensional distributions, which have been compiled in Figure 4.51. The two-dimensional plots with CPCs #1 and #3 have been omitted for the sake of brevity and because their visual inspection did not provide any additional significant information beyond what is already discernible from the other plots.

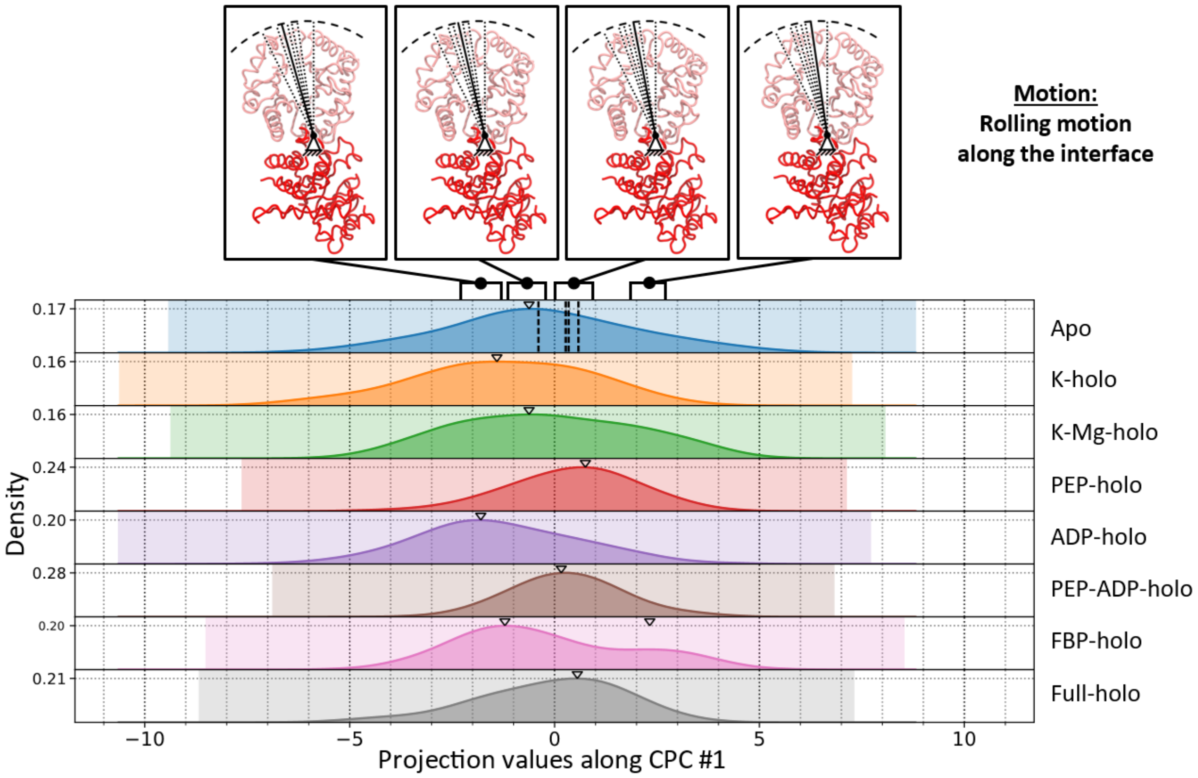
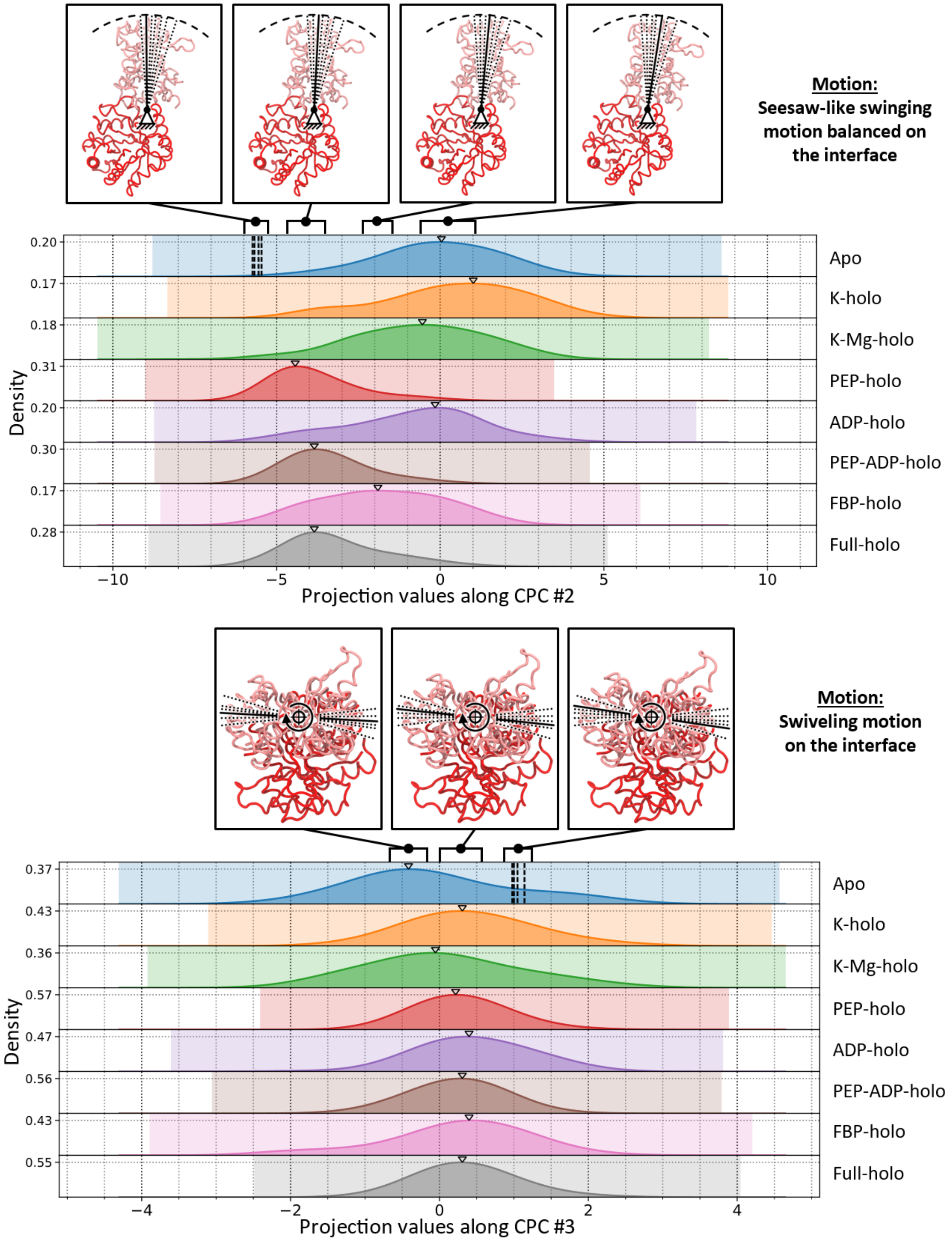


Figure 4.49. Comparative analysis of the WT trajectories (A-A′ pairs of domains) in apo *vs*. holo conditions: conformational profiles (1 CPC). The 1D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 (top), #2 (center), or #3 (bottom) from the reference (WT apo) condition. The data of each condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray. The plots show the KDE curves of the aggregated projection data along with representative structures of the conformation of the A-A′ pair of domains at different projection values. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. The A and A′ domains are colored in red and pink, respectively. NOTE. The images of the protein structure were generated with the software VMD.

Figure 4.49 (Continued)



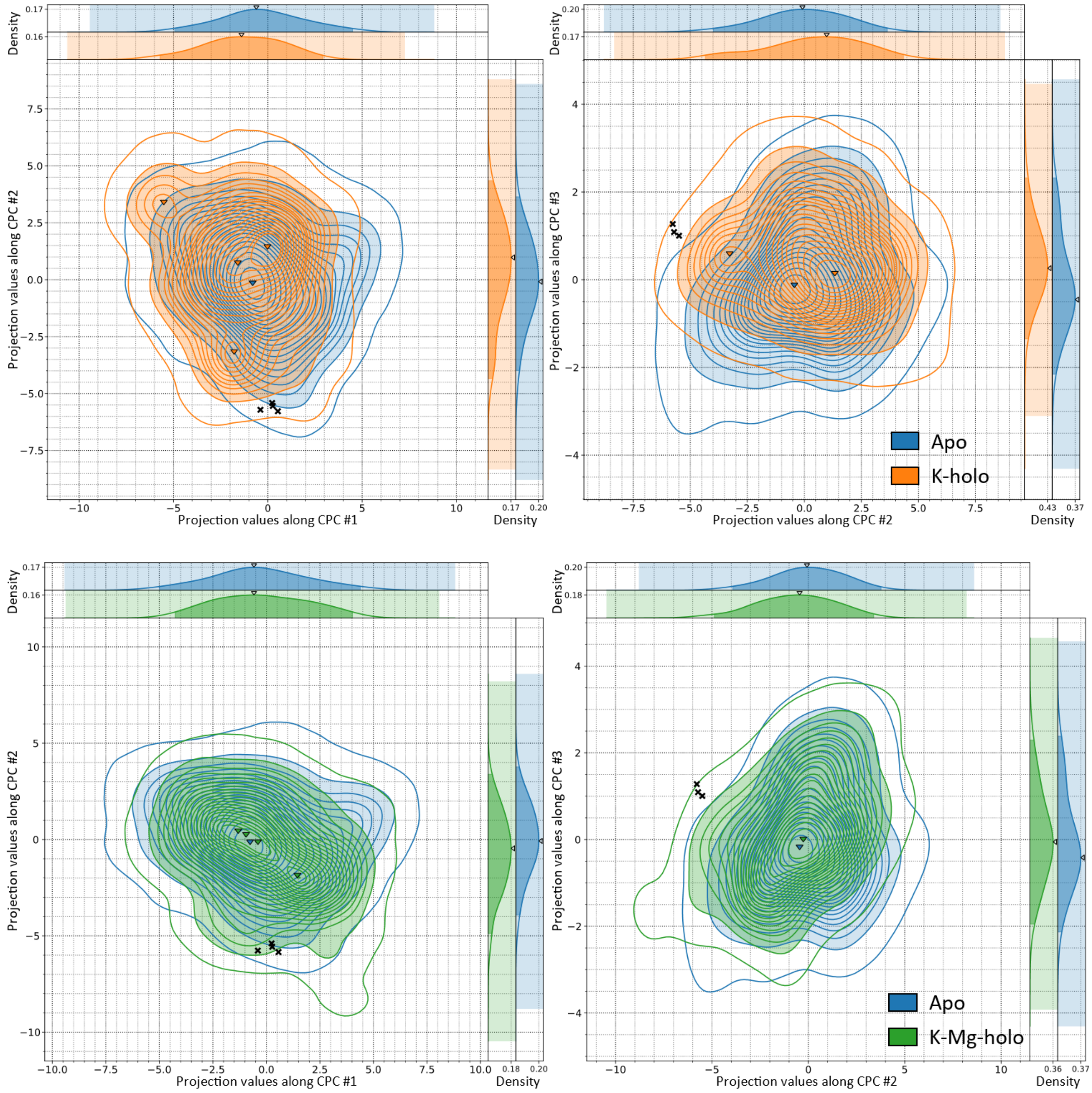


Figure 4.50. Comparative analysis of the WT trajectories (A-A’ pairs of domains) in apo *vs*. holo conditions: conformational profiles (2 CPCs). The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 (left panels) and #2 *vs*. #3 (right panels) from the reference (WT apo) condition. The data of each condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. The projection values of the subunits of the structure 2VGB are indicated with “X” markers. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figure 4.50 (Continued)

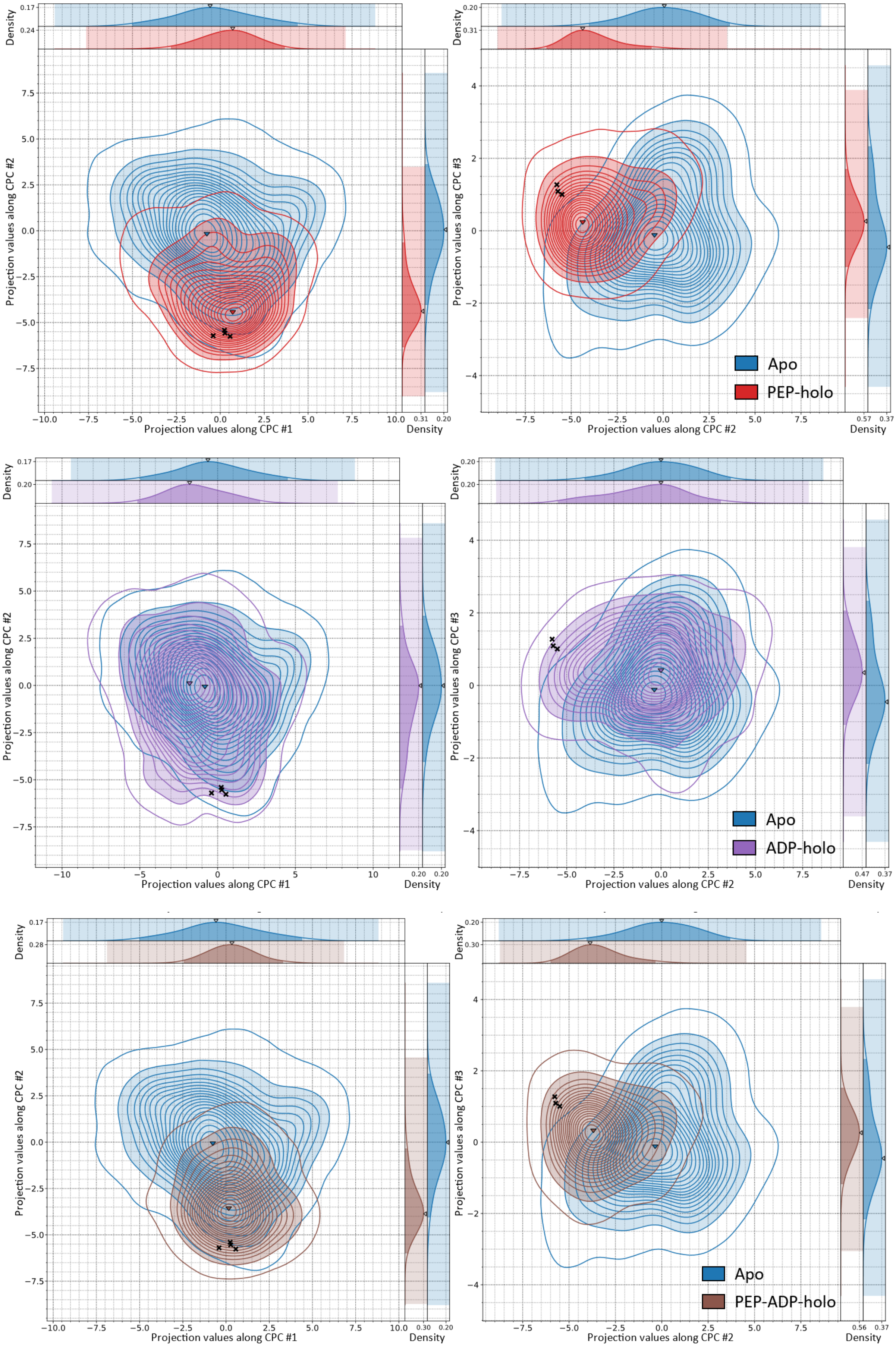
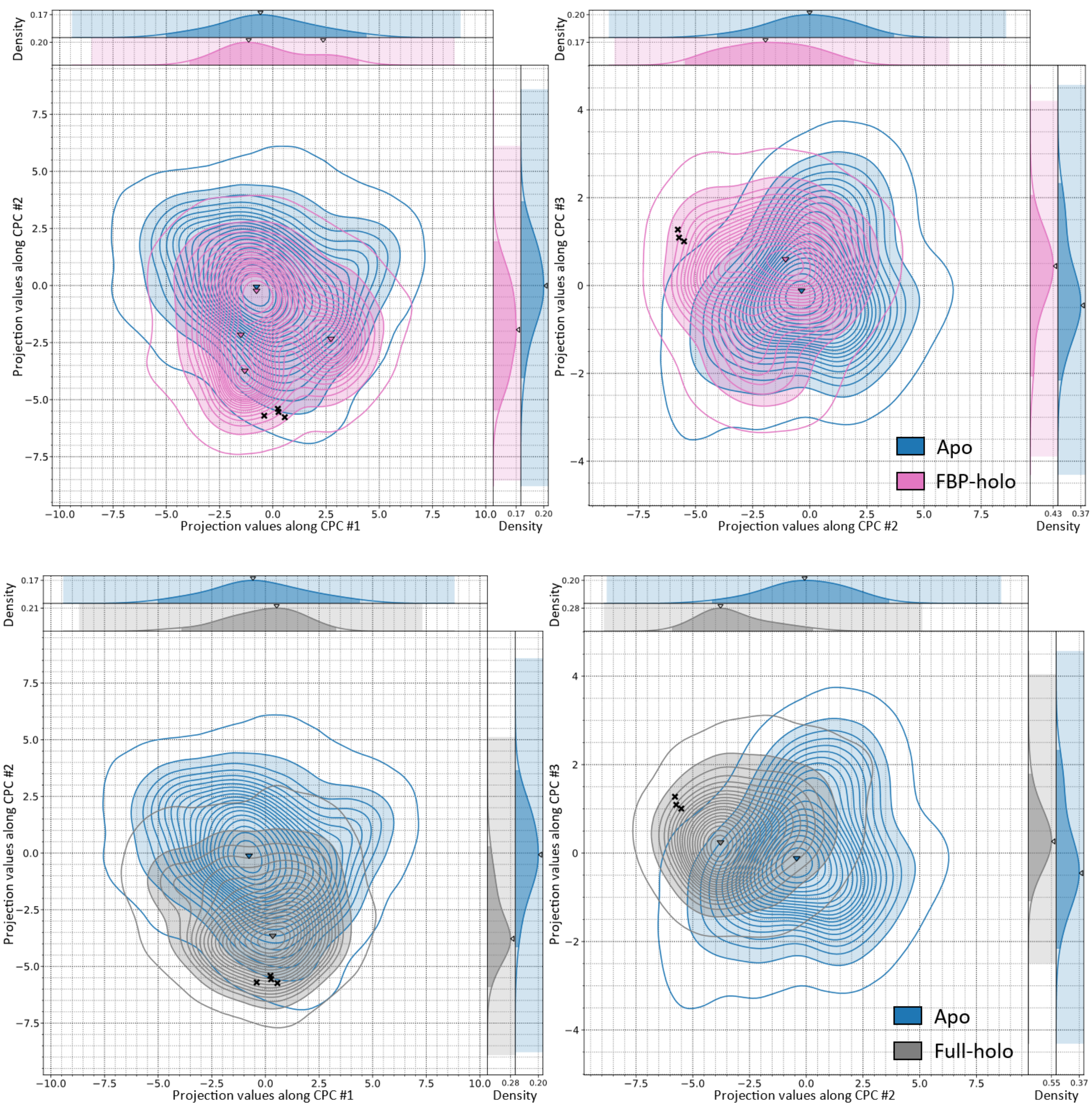


Figure 4.50 (Continued)



In general, the density distributions of the holo conditions preserve unimodality. The main differences compared to the apo condition involve moderate shifts in the distribution centers and small alterations in density dispersion. While the distributions along CPCs #1 and #2 display various shifts depending on the type of holo condition, all holo conditions align distinctively with similar distribution centers along CPC #3.

In the presence of PEP at the active site (conditions PEP-holo, PEP-ADP-holo, and Full-holo), the density distributions are narrower and consistently show shifts towards the projection values of 2VGB. Along CPC #2, these distributions display heavy tails that extend towards the spectrum region characteristic of the apo condition, thus indicating signs of an underlying conformational equilibrium. Consequently, PEP is the substrate that most significantly attenuated conformational diversity by constraining sampling around the conformations of the crystallographic structure. This effect was also strongly represented in the analysis of the A and B domains, thus indicating a consistent behavior. This observation provides additional evidence supporting the described mechanism of cooperativity between adjacent active sites (A-A' interface), whereby PEP binding induces relative motions and symmetrical rearrangements between the A domains [127], [128], [132], [149]. In line with this proposed model of cooperativity, simulations with all active sites occupied with PEP firmly stabilized the active conformation. Conversely, in the absence of PEP, the structure exhibited greater conformational freedom.

The holo conditions exhibiting minimal changes compared to the apo condition are those containing only cofactors at the active site, namely K-holo and K-Mg-holo. Particularly along CPCs #1 and #2, the density distributions are not significantly shifted but display higher dispersion. Specifically, the two-dimensional distributions of K-holo exhibit more local maxima scattered around the span of the distributions. The ADP-holo condition displays density distributions that resemble those of K-holo but with slightly less dispersion. Thus, the addition of MgADP to the K^+-bound active site does not seem to have caused significant sampling changes of the analyzed conformations and motions.

The FBP-holo condition displays density distributions with intermediate characteristics between those from the other groups of holo conditions. Of particular relevance are the possible differences with respect to the K-Mg-holo condition, as these conditions differ only by the fact that FBP-holo has FBP bound to the allosteric site. Along CPC #2, the FBP-holo density distribution is more shifted towards the characteristic region of the PEP-bound holo conditions, suggesting that FBP may have induced retention of sampling of conformations that are close to those induced by PEP binding. Along CPC #3, the FBP-holo density distribution is more aligned with the rest of holo conditions and exhibits less dispersion than that of K-Mg-holo. Along CPC #1, the FBP-holo density distribution has a similar span than that of K-Mg-holo but is more polarized in two opposing regions. The region with minor density does not particularly overlap with other holo conditions, although, when inspected in combination with CPC #2, it seems to correspond to conformations closer to those of the PEP-bound holo conditions.

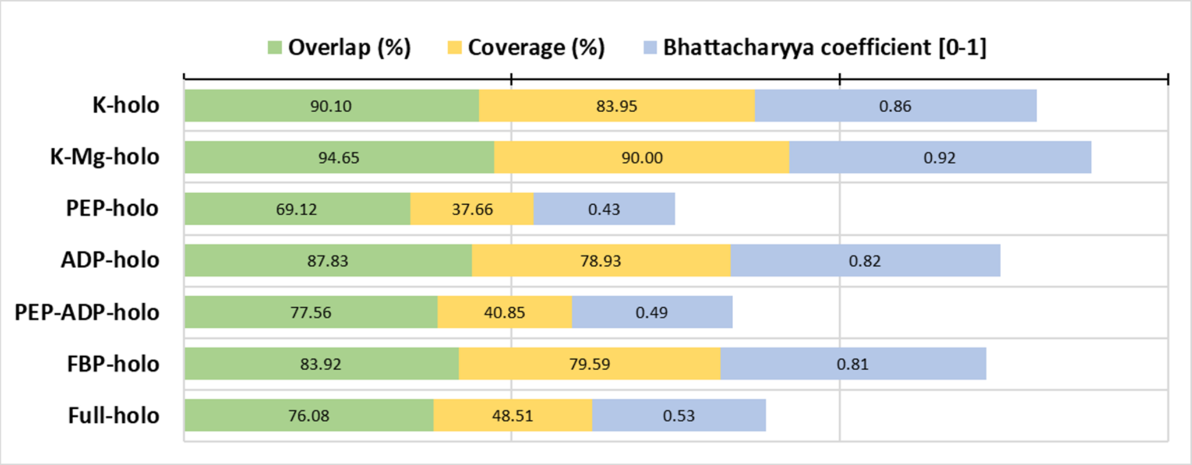


Figure 4.51. Comparative analysis of the WT trajectories (A-A' pairs of domains) in apo *vs*. holo conditions: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. The bar sizes of the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

The principal traits discerned from the visual examination of the density distributions are reflected in the ranges of values of the similarity metrics along CPCs #1 to #3. The three metrics, both individually and collectively, facilitate the classification of the different types of behavior of the holo conditions. On one side, the K-holo and K-Mg-holo conditions display the highest values of each metric that denote a substantial degree of similarity with the apo condition. In terms of these indicators, they are essentially within the range of intrinsic variability exhibited by different trajectory ensembles of the apo condition, as seen from the comparison between the original and the additional apo simulations. These are closely followed by a second group that includes the ADP-holo and FBP-holo conditions and displays slightly lower values. For ADP-holo, these values can be attributed to slightly lower sampling of the apo spectrum of conformations compared to K-holo. For FBP-holo, these values signify the intermediate behavior of this condition between the former and the following groups of holo conditions. The last group comprises the PEP-holo, PEP-ADP-holo, and Full-holo conditions, which display the lowest values as a result of more restrictive sampling.

4.1.3.2.2 The C-C′ dimers

The analysis of the C-C′ dimers was aimed at examining the collective motions of a larger region of the enzyme, namely, both the A and C domains of the adjacent subunits. The A and C domains constitute the major central region of each subunit, called the AC core. The study of the dynamical events between the adjacent AC cores offers the possibility to characterize the conformational changes across the C-C′ interface and, simultaneously, uncover potential coupling mechanisms with the A-A′ interface via the corresponding intrasubunit rearrangements. Importantly, the binding site of the allosteric activator FBP is located beside the C-C′ interface. Therefore, the identification of distinctive patterns of motion of the AC cores may provide insight on the mechanism of propagation of the allosteric effect to the active site.

The CEDA procedure was applied to the WT apo trajectories to derive the corresponding set of CPCs. Each tetramer comprises two pairs of adjacent AC cores across the C-C′ interface, resulting in 10 analyzable AC-C′A′ pairs of cores from the 5 simulations of the WT tetramer. The global average structure of all AC-C′A′ pairs of the trajectory ensemble was generated and used as the general reference for structural superpositions. In a similar way as in the experiment with the A-A′ dimers, an approach to obtain 20 analyzable trajectory replicates was implemented, whereby two complementary PCAs were performed for each AC-C′A′ pair, alternating the role of the fitting group between the respective A domains (*i.e.*, AC-C′A′ with A as the fitting group and A′C′-CA with A′ as the fitting group; in both cases superposed to A of the global average structure).

Only the Cα atoms of the region (772 atoms) were considered in the analysis, therefore each PCA yielded a set of 2316 eigenvectors. The first 20 eigenvectors of each trajectory replicate were included in the clustering analysis, accounting for approximately 90% of the cumulative variance on average (Figure 4.52). The assessment of the cosine content identified 4 instances of PCs with values exceeding the threshold of 0.5 (Figure 4.53). These instances correspond to the PC #1 of replicates #3-CD, #3-DC, #3-BA, and #5-CD. It is noteworthy that the trajectory replicates from simulation #3 are overrepresented. As in the previous experiments, the degree of integration of these PC instances of high cosine content in the clusters of higher coverage will clarify the nature of the corresponding captured collective motions.

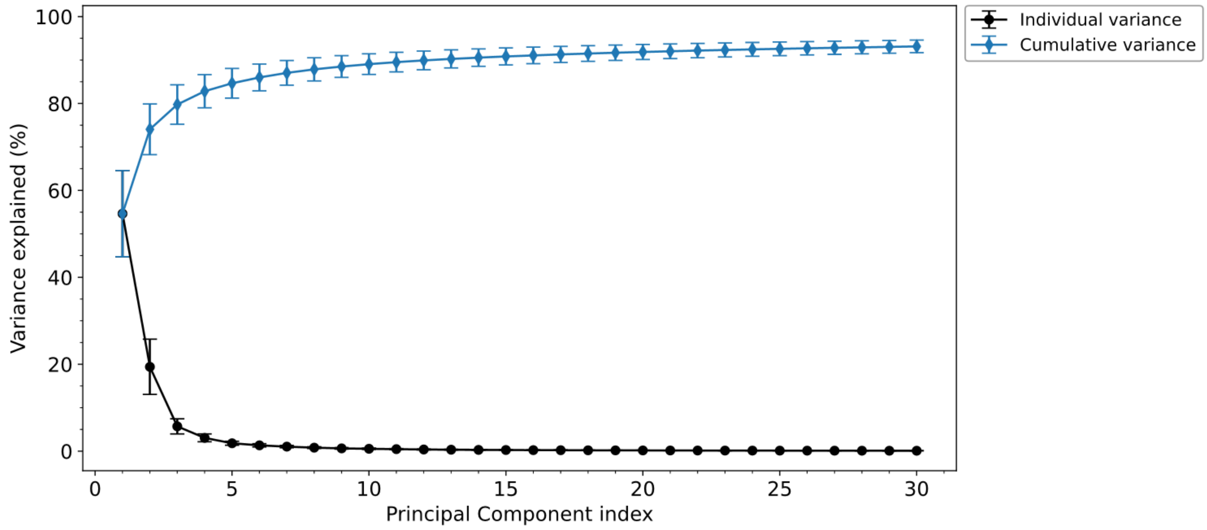


Figure 4.52. Percentage of variance explained by the PCs of the AC-C'A' pairs of cores in the WT apo condition. Data corresponds to the mean values from the first 30 PCs of the 20 trajectory replicates of the experiment. Standard deviation intervals are included as error bars.

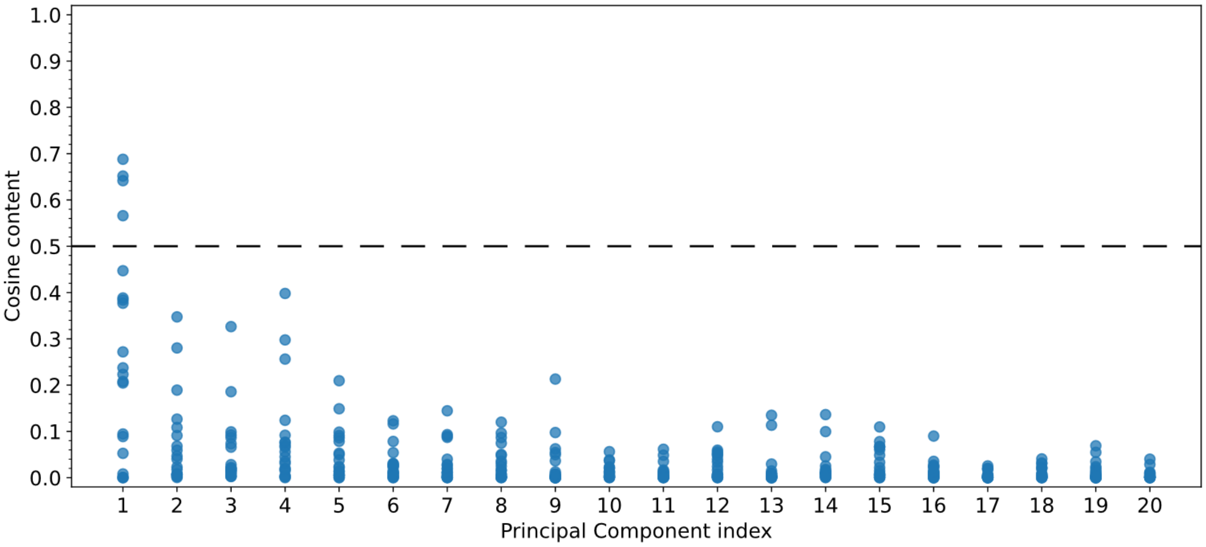


Figure 4.53. Cosine content values of the PCs considered in the CEDA of the AC-C'A' pairs of cores in the WT apo condition. The cutoff value of 0.5 below which the sampling is considered to be acceptable in protein dynamics is shown with a horizontal dashed black line.

A cophenetic distance cutoff of 0.4 was applied to split the dendrogram of the hierarchical clustering (Figure 4.54), and clusters with less than 20% of coverage were discarded, thus retaining a total of 9 relevant clusters. The primary division of the dendrogram allows identification of two distinct sections, similarly to the previous CEDA experiments. The lower section contains 2 large clusters of high coverage, namely, clusters #1 and #2 with 100% and 95% coverage values, respectively. The upper section of the dendrogram contains a large cluster (cluster #6 with 80% coverage) as well as scattered eigenvectors that form smaller clusters of both lower similarity and coverage values. Clusters #7, #10, #11, #13, #14, and #15 meet the established criteria to produce CPCs, with 30%, 25%, 25%, 35%, 20%, and 20% of coverage, respectively. The centroid vector of the 9 relevant clusters was computed to acquire the CPCs of the experiment, numbered from #1 to #9 in decreasing order of coverage and average percentage of variance of their cluster members.

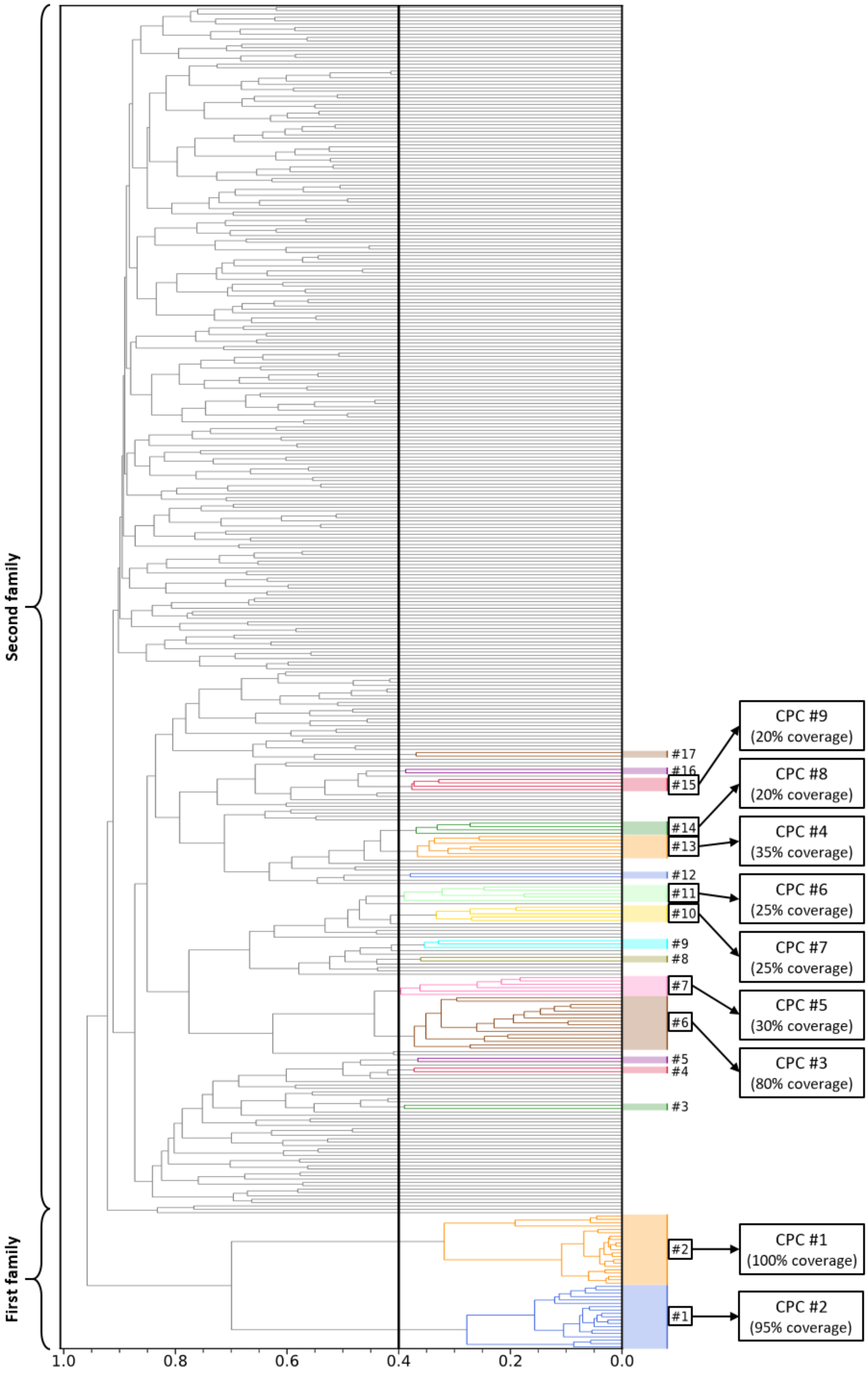


Figure 4.54. Dendrogram of the CEDA of the AC-C'A' pairs of cores in the WT apo condition. The cophenetic distance cutoff of 0.45 used to split the dendrogram and obtain clusters is shown with a vertical black line. Non-singleton clusters are shown in various colors to facilitate visual identification. The 9 clusters with at least 20% of coverage were selected to acquire the Consensus Principal Components (CPCs) of the experiment. CPCs were numbered from #1 to #9 in decreasing order of coverage and average percentage of variance of their cluster members. The figure indicates which clusters yielded each CPC, together with their achieved coverage. At the left margin of the dendrogram, the span of the detected families of clusters is indicated with labeled curly brackets.

Next, the collective motion captured in each CPC was visually examined. Figure 4.55 shows schematic diagrams of the AC-C′A′ pair of cores along with the terminology that will be employed to refer to the observed motions. The structure was arranged in a reference orientation with the A domain (anchored region; fitting group) on the top of the rest of domains. Three primary spatial directions were defined as the dimensions of a hypothetical rectangular box (width, height, and depth) containing the tetrameric structure. Three perpendicular sectional planes that intersect at the center of the box help retain further reference features of the AC-C′A′ pair of cores in the context of the whole tetramer. The horizontal plane coincides with the C-C′ interface, separating the upper and the lower subunits. A vertical plane coincides with the A-A′ interface, defining the positions where the absent adjacent subunits would be. Finally, the remaining vertical plane intersects with all four subunits and has been termed the "tetramer plane" because it covers its full extent. Figure 4.56 shows the schematic representations of the path and extreme conformations of the collective motions from CPCs #1 to #3.

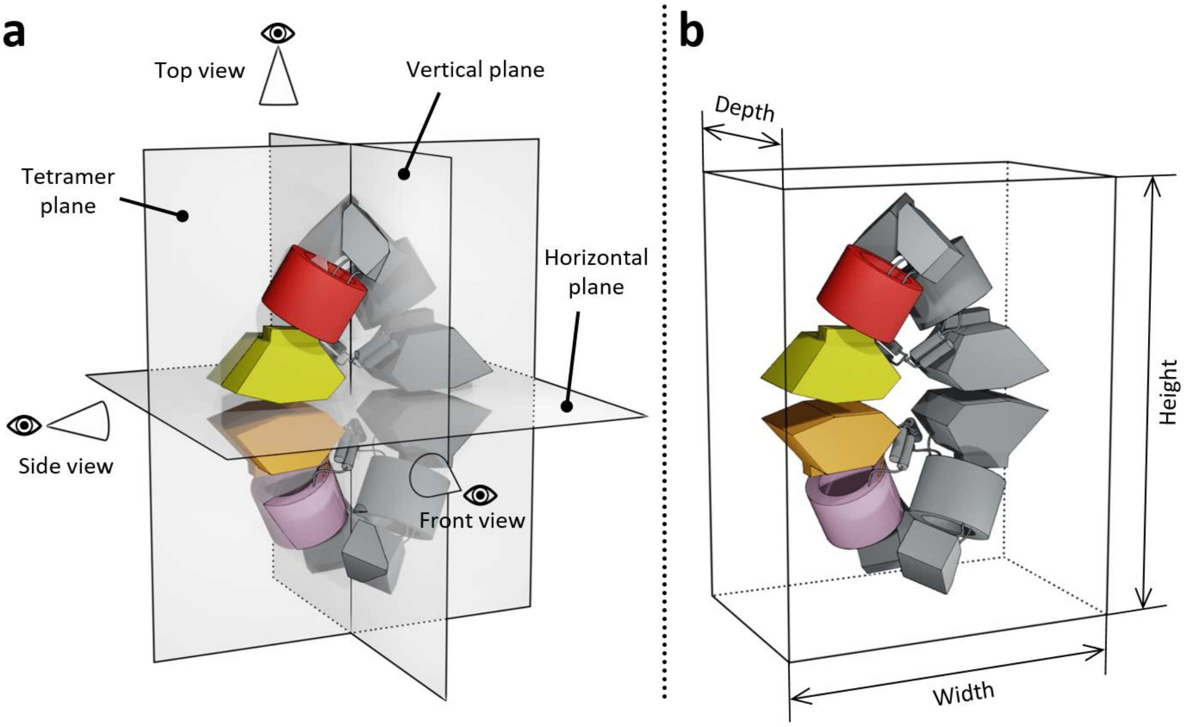


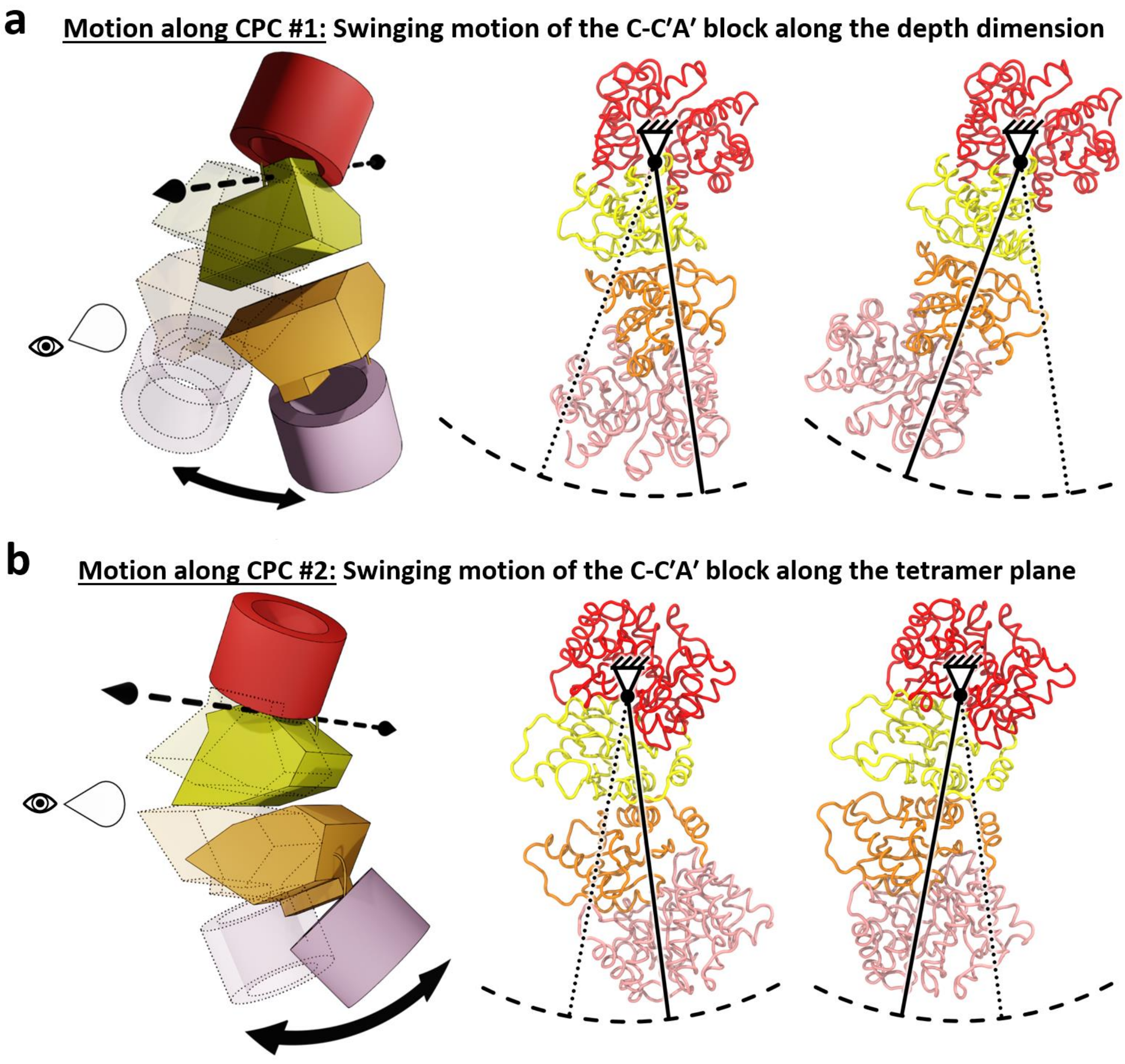
Figure 4.55. Schematic diagrams of the adjacent AC cores across the C-C′ interface of PKR. Various markers indicate the terminology to refer to (**a**) the reference views and sections of the structure, and (**b**) three primary spatial directions defined as the dimensions of a hypothetical rectangular box that contains the tetrameric structure. The structure under analysis is colored as follows: A domain in red, C domain in yellow, A′ domain in pink, and C′ domain in orange. The rest of the structure of the tetramer is also depicted (gray) to provide a reference of the relative positions where the absent domains and subunits would be. NOTE. The 3D schematic models were built with the software Blender.

In contrast with the previous analyses, the structure under investigation now encompasses four domains. With the A domain of the upper subunit being the fitting group of the experiment, the captured collective motions are characterized by collective displacements of the remaining three mobile domains. All observed motions feature a primary path of motion given by the displacement of the C domain with respect to the A domain (upper subunit), facilitated by the joint-like capabilities of the A-C interface, and with the Cα4 helix acting as a pivot point. In turn, the C′A′ core of the lower subunit follows the path described by the C domain and mainly moves accordingly with it as a single structural block, thanks to the union of the C-C′ interface. In some instances, the A′ domain also

]

exhibits separate paths of motion, as the conformational freedom of this domain depends on the constraints of its adjacent subunit, even though it is not included in the analysis.

The collective motion captured in CPC #1 (Figure 4.56a, Supplementary Video S4.19) shows a swinging motion of the C-C′A′ block along the depth dimension of the tetramer. The collective motion captured in CPC #2 (Figure 4.56b, Supplementary Video S4.20) shows a swinging motion of the C-C′A′ block along the plane of the tetramer. The collective motion captured in CPC #3 (Figure 4.56c, Supplementary Video S4.21) shows a horizontal rotational motion of the C-C′A′ block about a vertical axis.

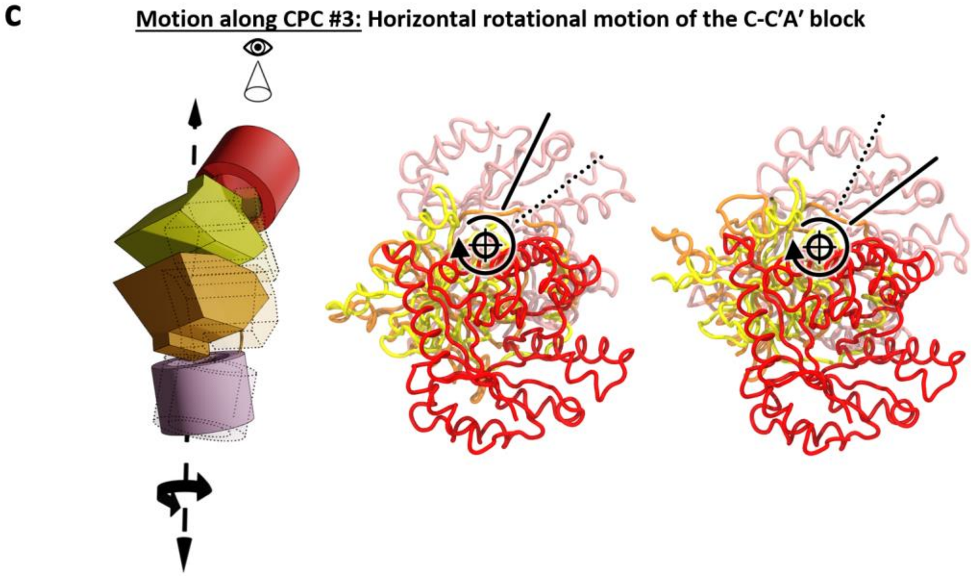


Motion along CPC #1: Swinging motion of the C-C′A′ block along the depth dimension

Motion along CPC #2: Swinging motion of the C-C′A′ block along the tetramer plane

Figure 4.56. Consensus collective motions of the adjacent AC cores across the C-C′ interface in the WT apo condition. (**a-c**) Schematic representations of the path and extreme conformations of the motion captured in CPCs #1, #2, and #3. In each panel, the protein domains are depicted with the diagrams of Figure 4.55 (left) and with the trace representation between Cα atoms (center and right). Domains are colored as follows: A in red, C in yellow, A′ in pink, and C′ in orange. These motions can also be viewed in the Supplementary Videos S4.19 to S4.21. NOTE. The images of the protein structure were generated with the software VMD. The 3D schematic models were built with the software Blender.

Figure 4.56 (Continued)


C
Motion along CPC #3: Horizontal rotational motion of the C-C'A' block

The composition of the corresponding clusters of these CPCs correlates with the index of the clustered PCs. CPC #1 agglomerated 17 instances of PC #1 and 3 of PC #2. All the PCs with high cosine content are well integrated into this cluster, thus confirming the significance of their captured collective displacement of coordinates despite their PC projection values having stronger resemblance to cosine shapes. CPC #2 agglomerated 16 instances of PC #2 and 3 of PC #1. CPC #3 agglomerated 14 instances of PC #3 and 2 of PC #4. Consequently, the conformational space defined by the dimensions of CPCs #1 to #3, in decreasing order of conformational variability, is representative of the sampling of the whole trajectory ensemble. In turn, this distribution of PCs implies that the relative importance of each collective motion was equivalent among all AC-C'A' pairs of cores, for either of the two alternative points of reference (fitting group) with respect to the global average structure. This resulted in the derivation of a more robust consensus representation of the motions.

CPCs #4 to #9 captured collective motions of lower amplitude, and comprise PCs with various indices between #3 to #7. The corresponding motions can be viewed in the Supplementary Videos S4.22 to S4.27, and can be described as follows. CPC #4: a vertical shift of the C-C'A' block. CPC #5: a variation of the rotational motion of CPC #3 with an alternative reorientation of the A' domain. CPCs #6 and #7: horizontal shifts along the depth of the tetramer, in combination with a vertical shift similar to that of CPC #4. CPC #8: a variation of the vertical shift of CPC #4. CPC #9: a combination of horizontal and vertical shifts with a distinctive rotation of the barrel of the A' domain.

After the derivation of the CPCs and the examination of the captured motions, the experiment proceeded with the exploration of the corresponding conformational profiles. The probability density functions of the projection data along each CPC were estimated with KDE, using a bandwidth of 0.3 and representing the curves with 100 points along the domain of the data. Only the conformational profiles along CPCs #1, #2, and #3 were explored as they constitute the clusters of greatest coverage and conformational sampling variance of the experiment. Figures 4.57, 4.58, and 4.59 show the one-dimensional density distributions along each of these CPCs, both for each individual trajectory replicate and for the aggregated projection data. Representative structures of the AC-C'A' pair of cores at different projection values are included to inform about their correspondence with the range of conformations. The projection values of the subunits of the structure 2VGB are also indicated in the figures.

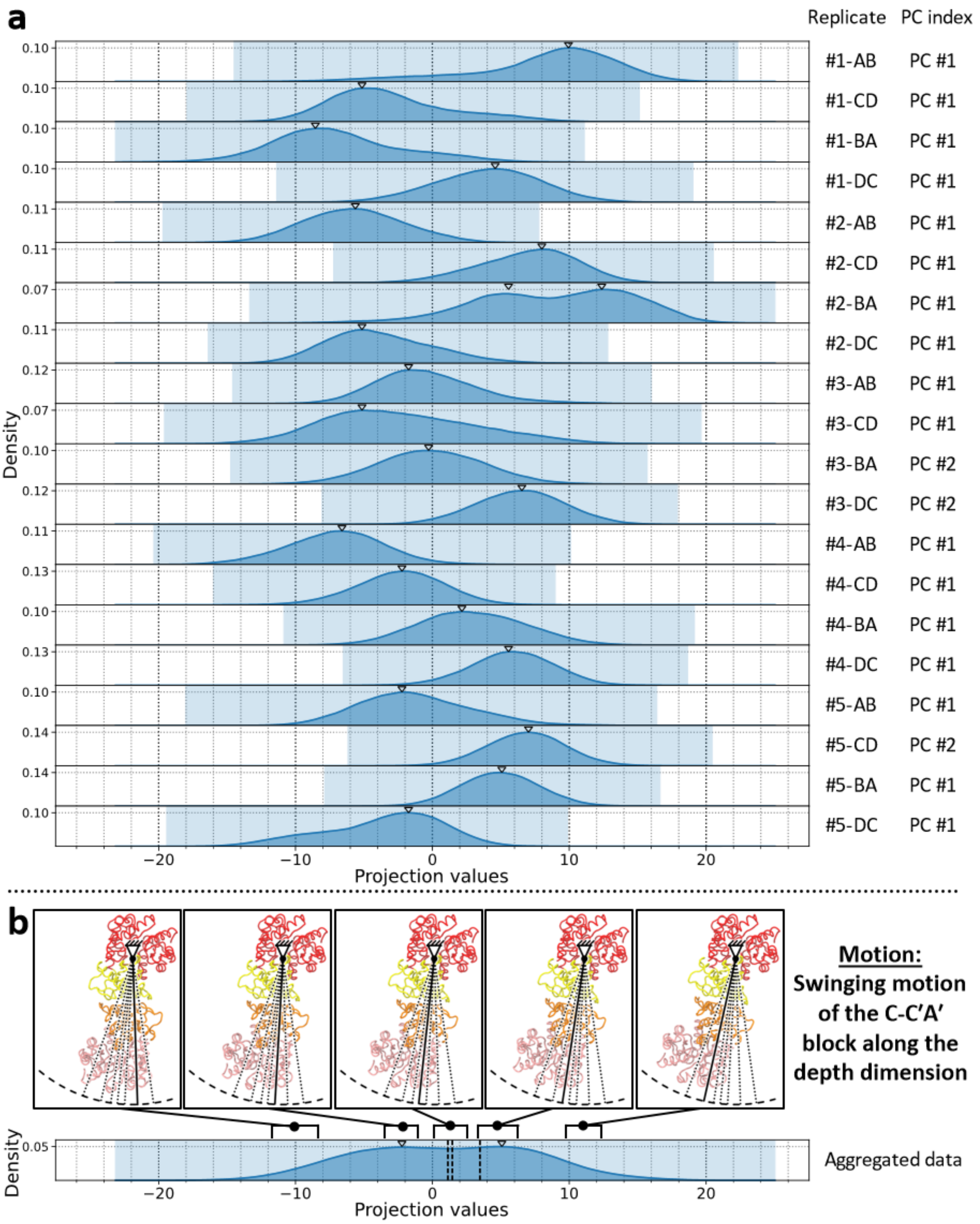


Figure 4.57. Density distributions along CPC #1 of the AC-C'A' pairs of cores in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the AC-C'A' pair of cores at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. Domains are colored as follows: A in red, C in yellow, A' in pink, and C' in orange. NOTE. The images of the protein structure were generated with the software VMD.

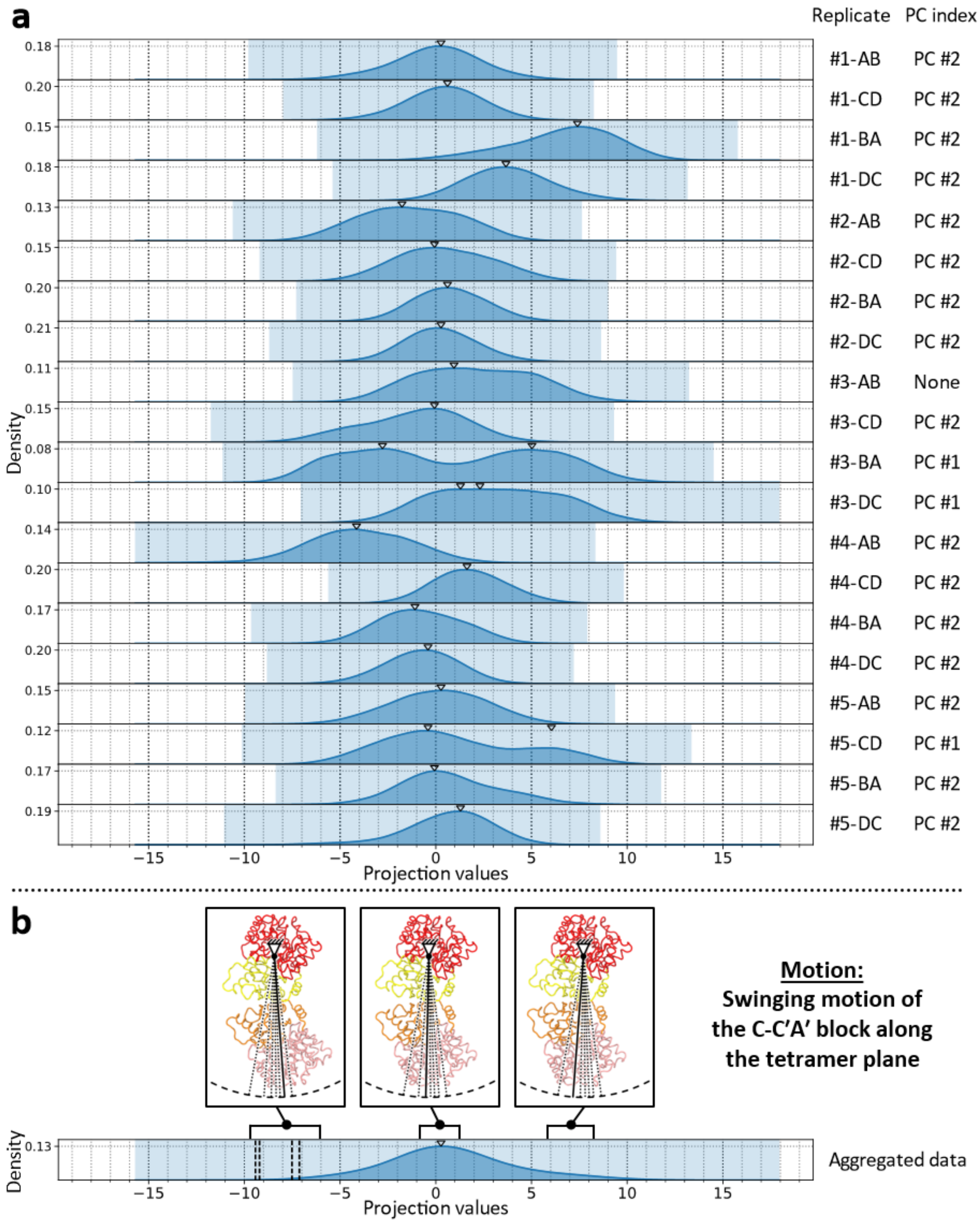


Figure 4.58. Density distributions along CPC #2 of the AC-C'A' pairs of cores in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the AC-C'A' pair of cores at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. Domains are colored as follows: A in red, C in yellow, A' in pink, and C' in orange. NOTE. The images of the protein structure were generated with the software VMD.

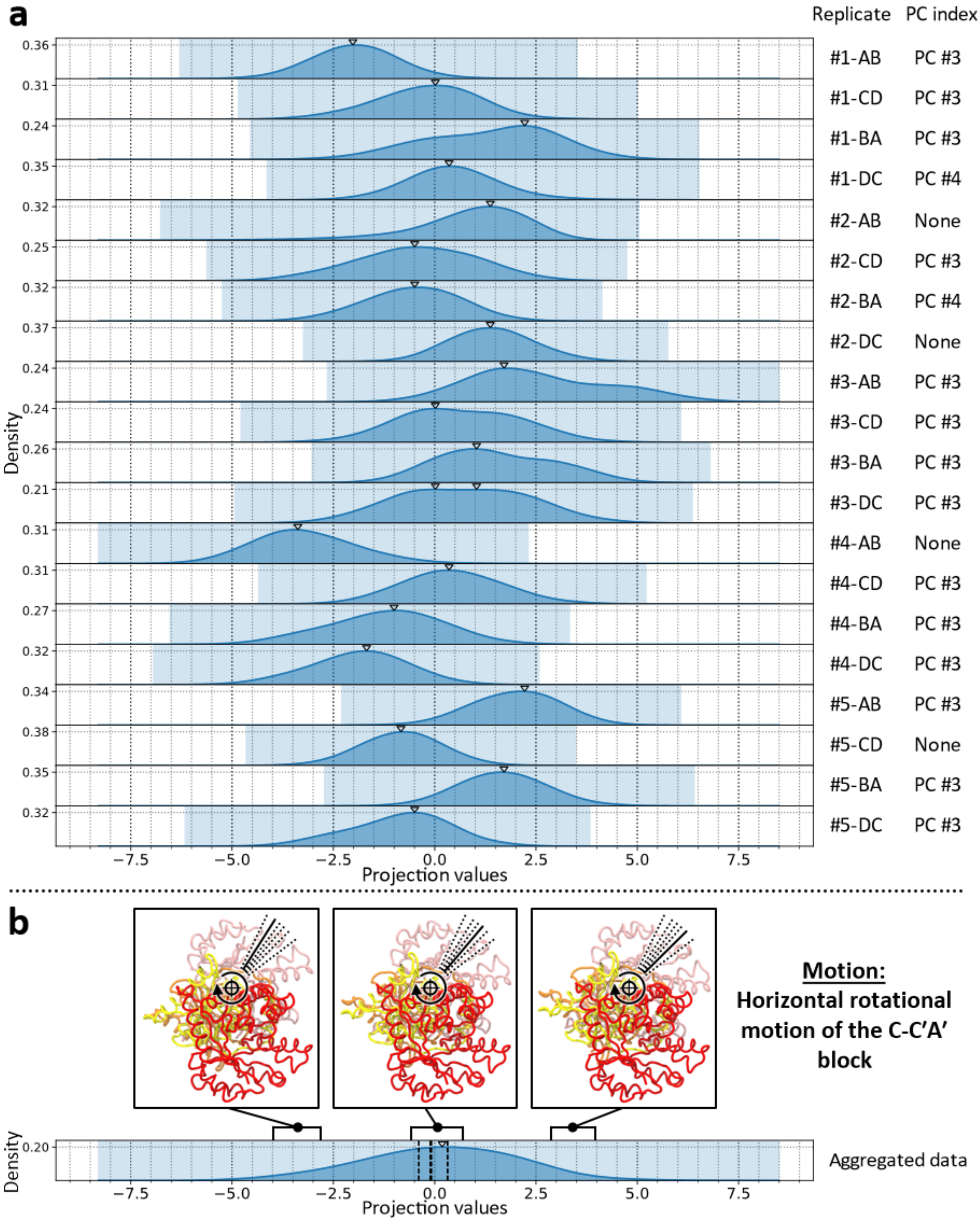


Figure 4.59. Density distributions along CPC #3 of the AC-C′A′ pairs of cores in the WT apo condition. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. (**a**) KDE curves of the projection data per trajectory replicate. The PC index of the clustered eigenvector of each replicate is indicated next to its identifier. (**b**) KDE curve of the aggregated projection data with representative structures of the approximate conformation of the AC-C′A′ pair of cores at different intervals of projection values. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. Domains are colored as follows: A in red, C in yellow, A′ in pink, and C′ in orange. NOTE. The images of the protein structure were generated with the software VMD.

The corresponding distributions of the aggregated projection data are mostly unimodal, with large dispersion, and centered around zero. This suggests that the structures sampled conformations around an energy minimum close to the global average structure of the ensemble. As usual, the total span of projection values diminishes as the CPC index increases. The distribution along CPC #1 exhibits a particularly wide plateau that spans between the values of -5 and 7 and is generated by the various localizations of the density distributions of the individual trajectory replicates around the central region of the conformational spectrum. Incidentally, there appears to exist a certain complementarity in the information provided in CPC #1 by the two trajectory replicates of the same pair of adjacent AC cores with alternative fitting groups. It can be noticed that when one exhibits positive projection values, the other tends to exhibit negative projection values, and *vice versa* (*e.g.*, #1-AB with #1-BA, or #1-CD with #1-DC). This implies that the conformational change captured by CPC #1 is described with opposite orientations depending on whether the point of reference is the upper or the lower subunit of the C-C′ dimer. This complementarity is not observed in the other CPCs.

The structure 2VGB displays projection values integrated within the central region of the conformational spectrums of CPCs #1 and #3, suggesting that there is no distinction in the preferred conformations between the apo and holo conditions of the enzyme along these directions of the sampling space. Conversely, in CPC #2, the projection values of 2VGB are located far from the center of the distribution, between the values of -7 and -10, corresponding to a conformation with the C-C′A′ block more tilted towards the center of the tetramer. This suggests that this conformation is not stable in the apo condition but rather dependent on the presence of one or several of the ligands of the crystallographic structure.

The exploration of the conformational profiles was complemented with the generation of the two-dimensional density distributions (Figure 4.60). No major topological irregularities or singularities can be discerned besides those implied in the one-dimensional distributions. In general, the three dimensions appear to define the basin of an energy minimum. The projection values of 2VGB lie at the center of the distribution of CPCs #1 *vs*. #3, whereas their deviation from the sampling region becomes evident when CPC #2 is inspected.

Next, as with the previous CEDAs, the obtained conformational profiles were compared to those of the additional set of WT apo simulations (#6 to #10) to evaluate replicability. On the one hand, Figure 4.61 shows the overlays of the pairwise two-dimensional plots of CPCs #1 to #3, which were visually compared to assess the similarities and differences in the sampling along the reference collective motions. On the other hand, the quantitative indicators (the overlap, coverage, and BC metrics) were calculated for the comparison between the corresponding three-dimensional distributions.

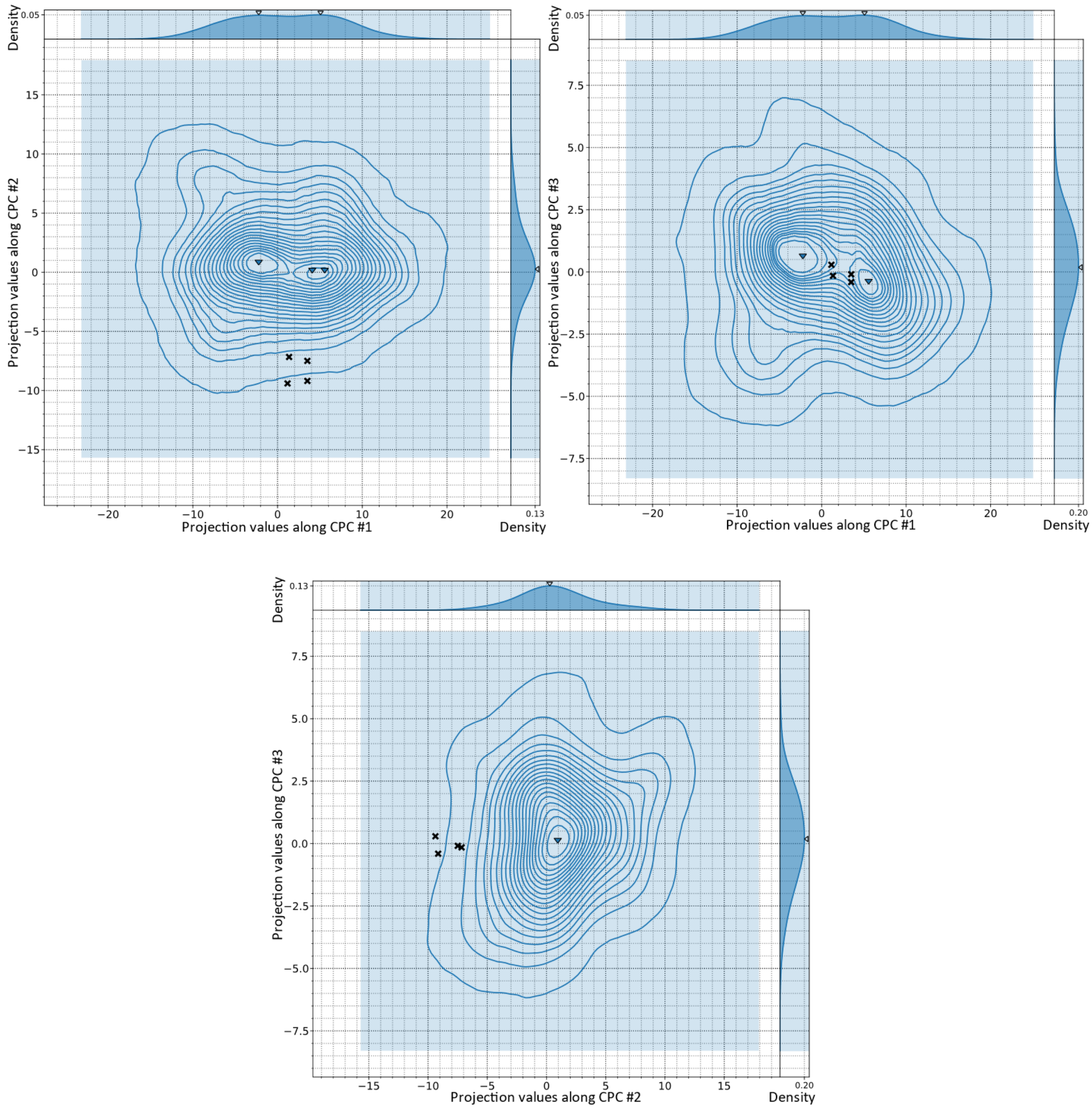


Figure 4.60. Pairwise 2D density distributions along CPCs #1 to #3 of the AC-C’A’ pairs of cores in the WT apo condition. The highlighted rectangular regions indicate the total span of projection values. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Triangle markers indicate the locations of local maxima. The projection values of the subunits of the structure 2VGB are indicated with “X” markers. Each two-dimensional plot also features subplots of the one-dimensional KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins.

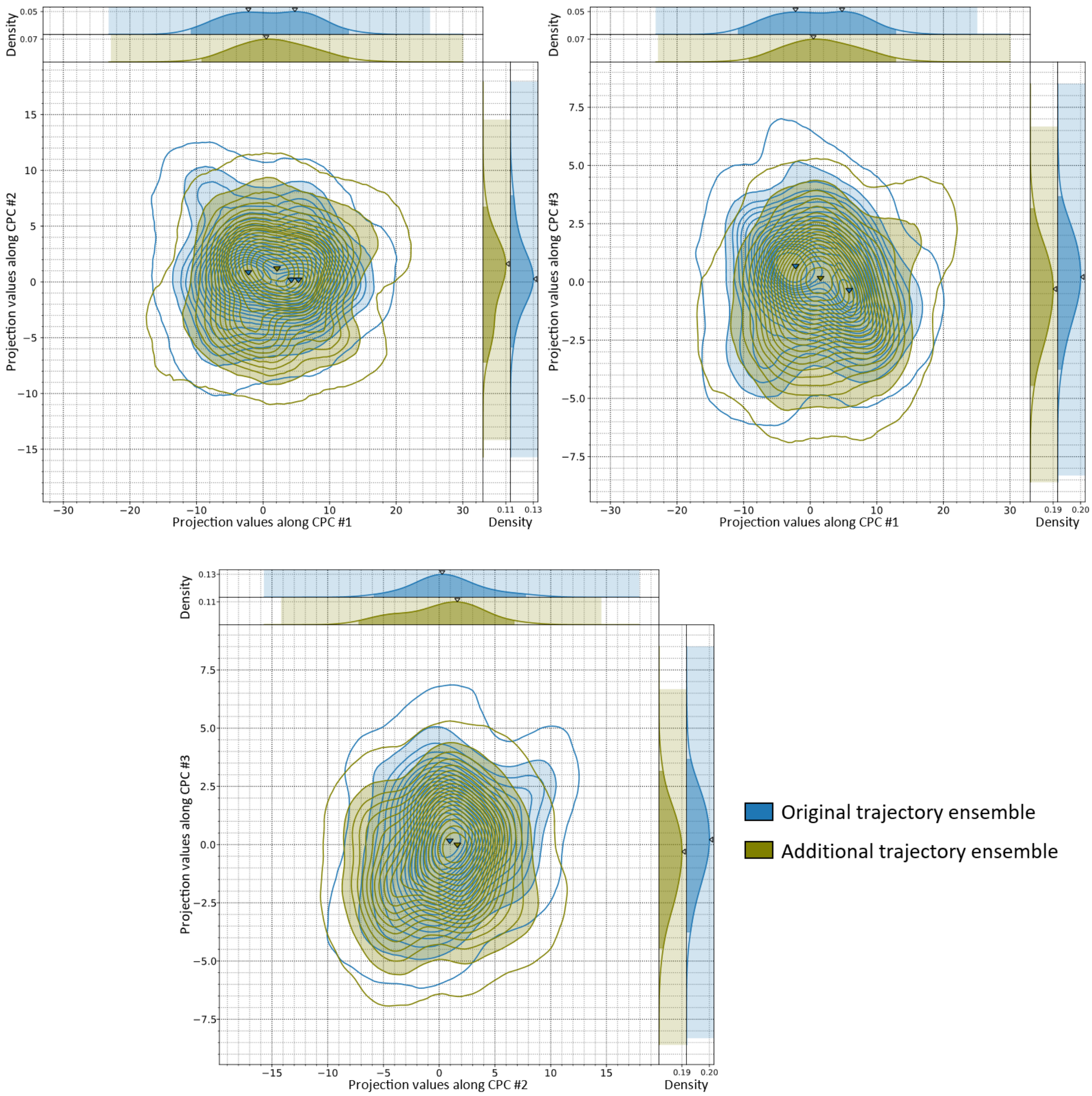


Figure 4.61. Comparative analysis of the original *vs*. additional WT trajectory ensembles (AC-C'A' pairs of cores) in the apo condition: conformational profiles (2 CPCs). The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 (top-left), #1 *vs*. #3 (top-right), and #2 *vs*. #3 (bottom) from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

The principal concentrations of density are shared between both trajectory ensembles, indicating an overall equivalent exploration of the same conformational space. The values of the quantitative indicators were overall high, with an overlap of 90.24%, a coverage of 93.90%, and a BC of 0.93. Among the noteworthy differences between the conformational profiles, the second trajectory ensemble does not exhibit the wide plateau at the central region of CPC #1 which is instead more focused in the form of a single central peak. In addition, it also displays a characteristic region of moderate density around the value of -5 of CPC #2, closer to the values of the structure 2VGB.

Following from the characterization of the results of the WT apo condition, the next stage of the analysis revolved around the comparison with the conformational profiles of the WT holo trajectory data when projected onto the reference set of CPCs. The comparison was based on the visual examination of the one- and two-dimensional distributions, shown in Figures 4.62 and 4.63, and the calculation of the similarity metrics between the three-dimensional distributions, which have been compiled in Figure 4.64. The two-dimensional plots with CPCs #1 and #3 have been omitted for the sake of brevity and because their visual inspection did not provide any additional significant information beyond what is already discernible from the other plots.

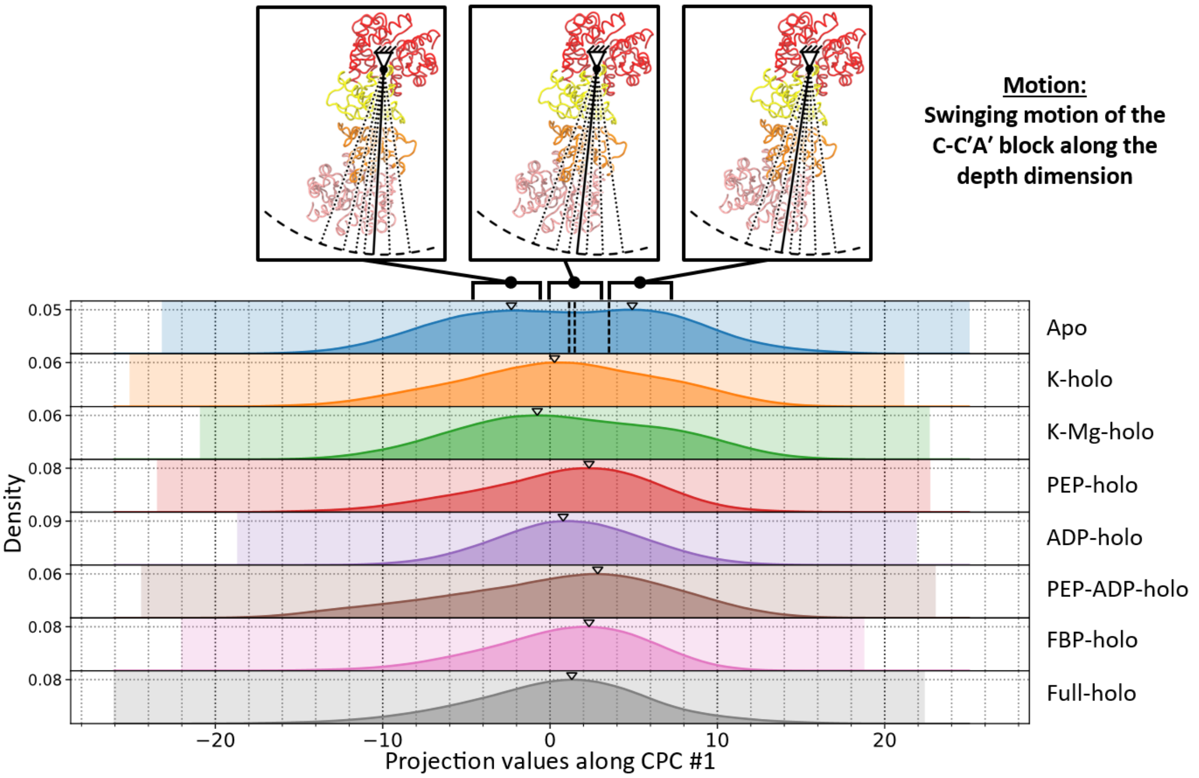
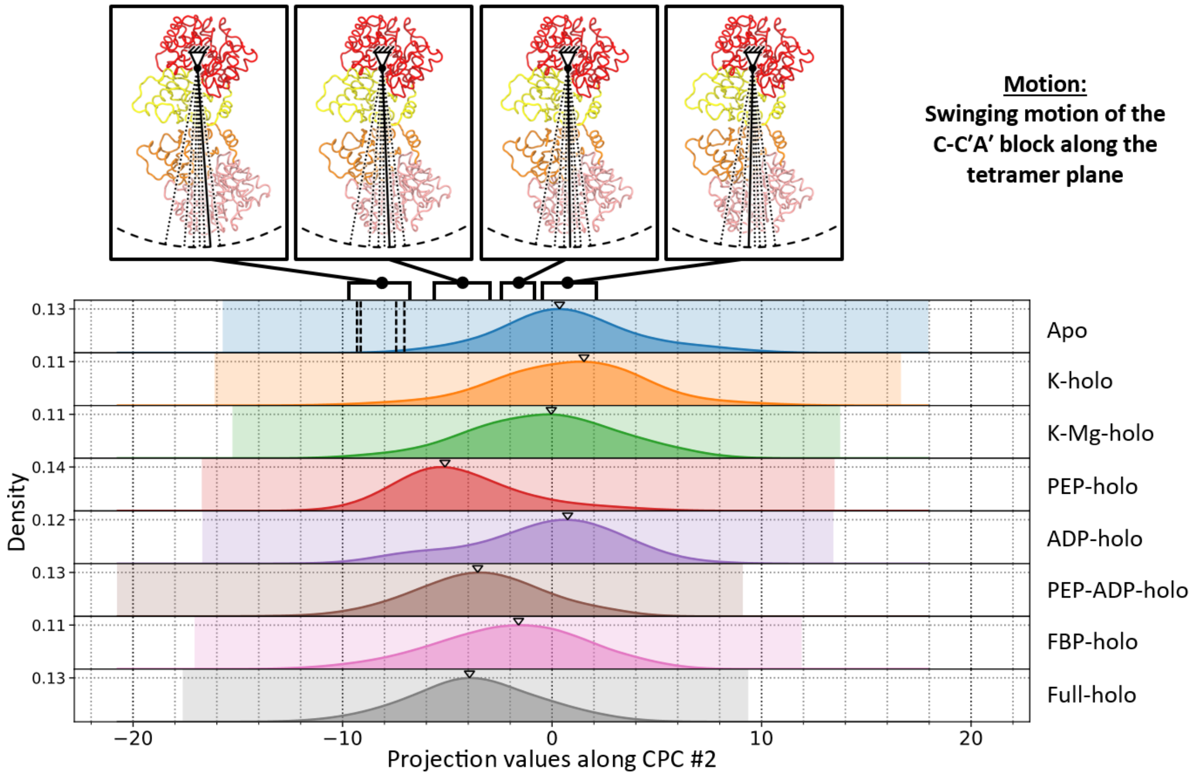
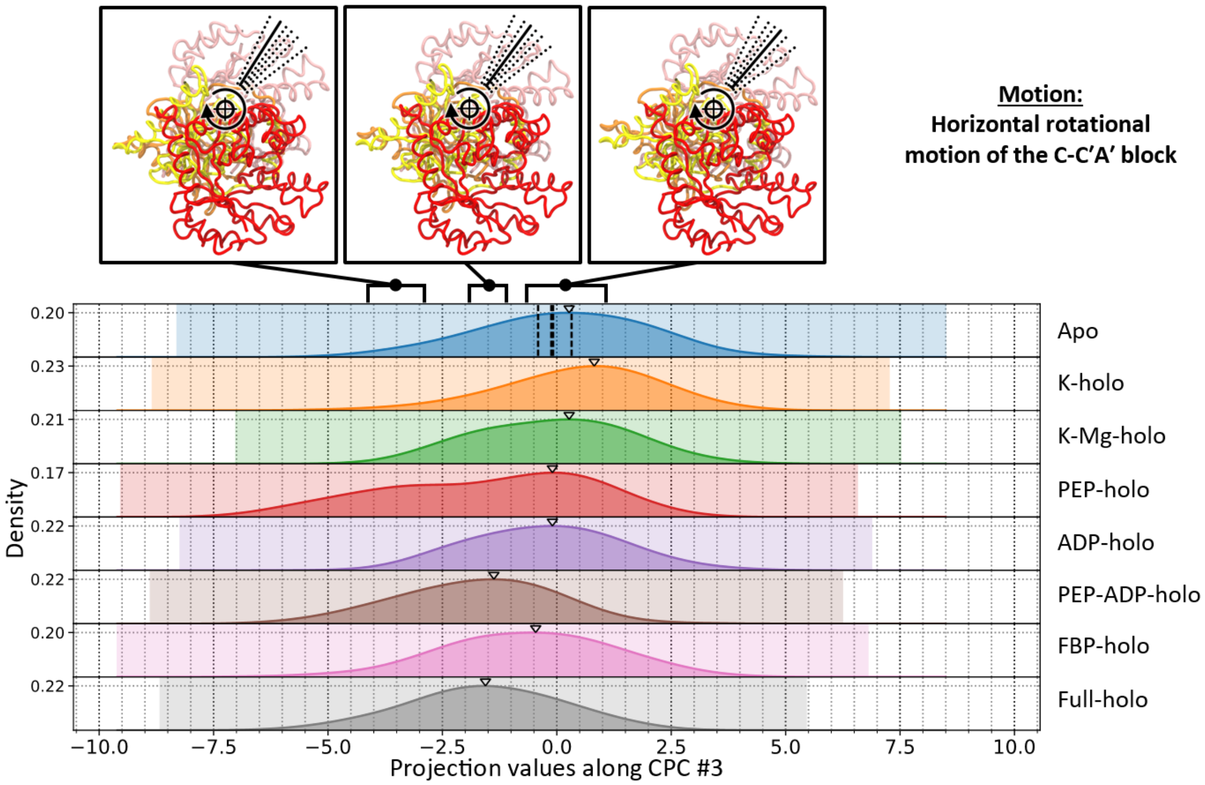


Figure 4.62. Comparative analysis of the WT trajectories (AC-C'A' pairs of cores) in apo *vs*. holo conditions: conformational profiles (1 CPC). The 1D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 (top), #2 (center), or #3 (bottom) from the reference (WT apo) condition. The data of each condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray. The plots show the KDE curves of the aggregated projection data along with representative structures of the conformation of the AC-C'A' pair of cores at different projection values. The highlighted region along the abscissa indicates the total span of projection values. Triangle markers at the top margin of each plot indicate the locations of local maxima. Vertical dashed lines indicate the projection values of the subunits of the structure 2VGB. Structures are depicted with the trace representation between Cα atoms. Domains are colored as follows: A in red, C in yellow, A' in pink, and C' in orange. NOTE. The images of the protein structure were generated with the software VMD.

Figure 4.62 (Continued)





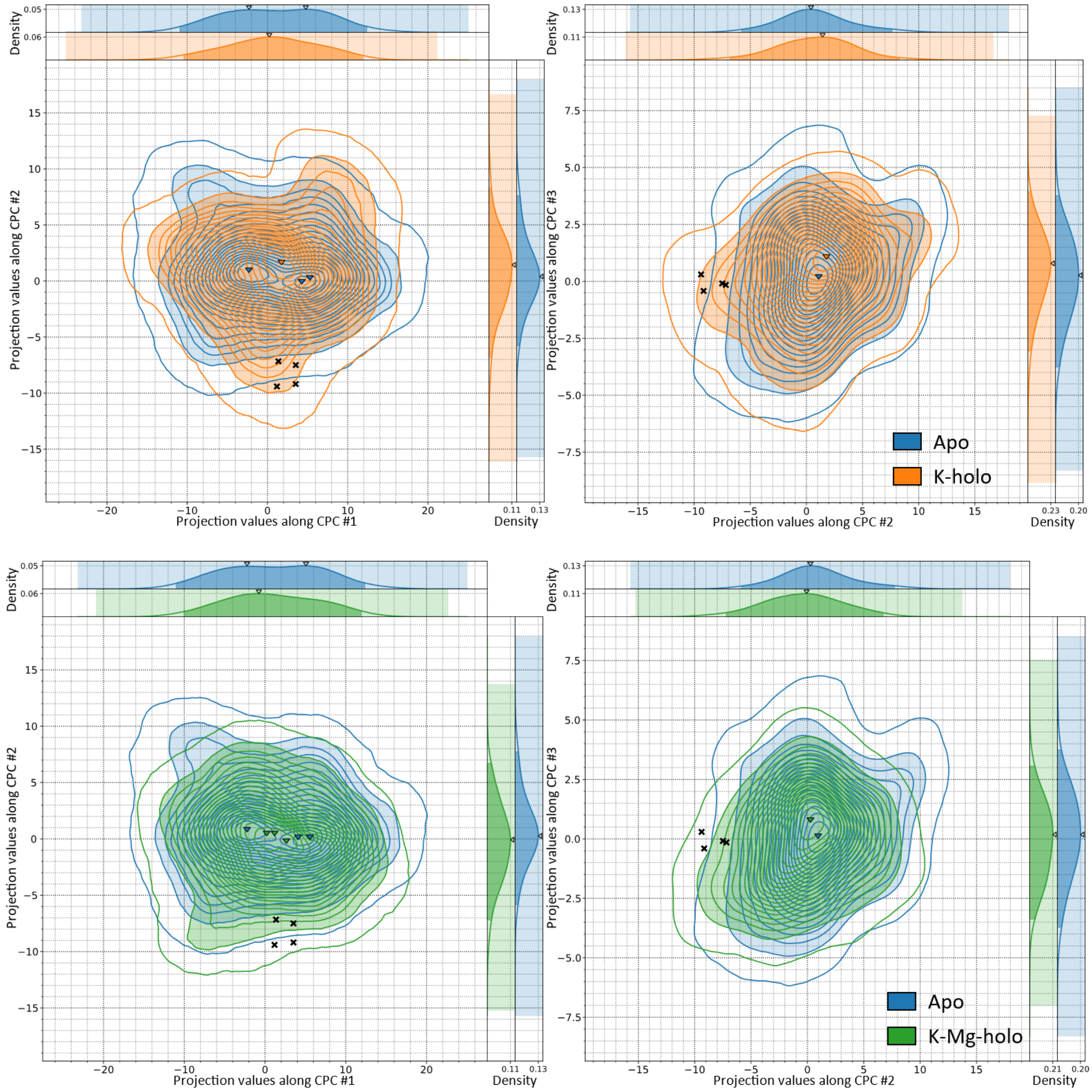


Figure 4.63. Comparative analysis of the WT trajectories (AC-C’A’ pairs of cores) in apo *vs*. holo conditions: conformational profiles (2 CPCs). The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 (left panels) and #2 *vs*. #3 (right panels) from the reference (WT apo) condition. The data of each condition is colored as follows: apo in blue, K-holo in orange, K-Mg-holo in green, PEP-holo in red, ADP-holo in purple, PEP-ADP-holo in brown, FBP-holo in pink, and Full-holo in gray. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. The projection values of the subunits of the structure 2VGB are indicated with “X” markers. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figure 4.63 (Continued)

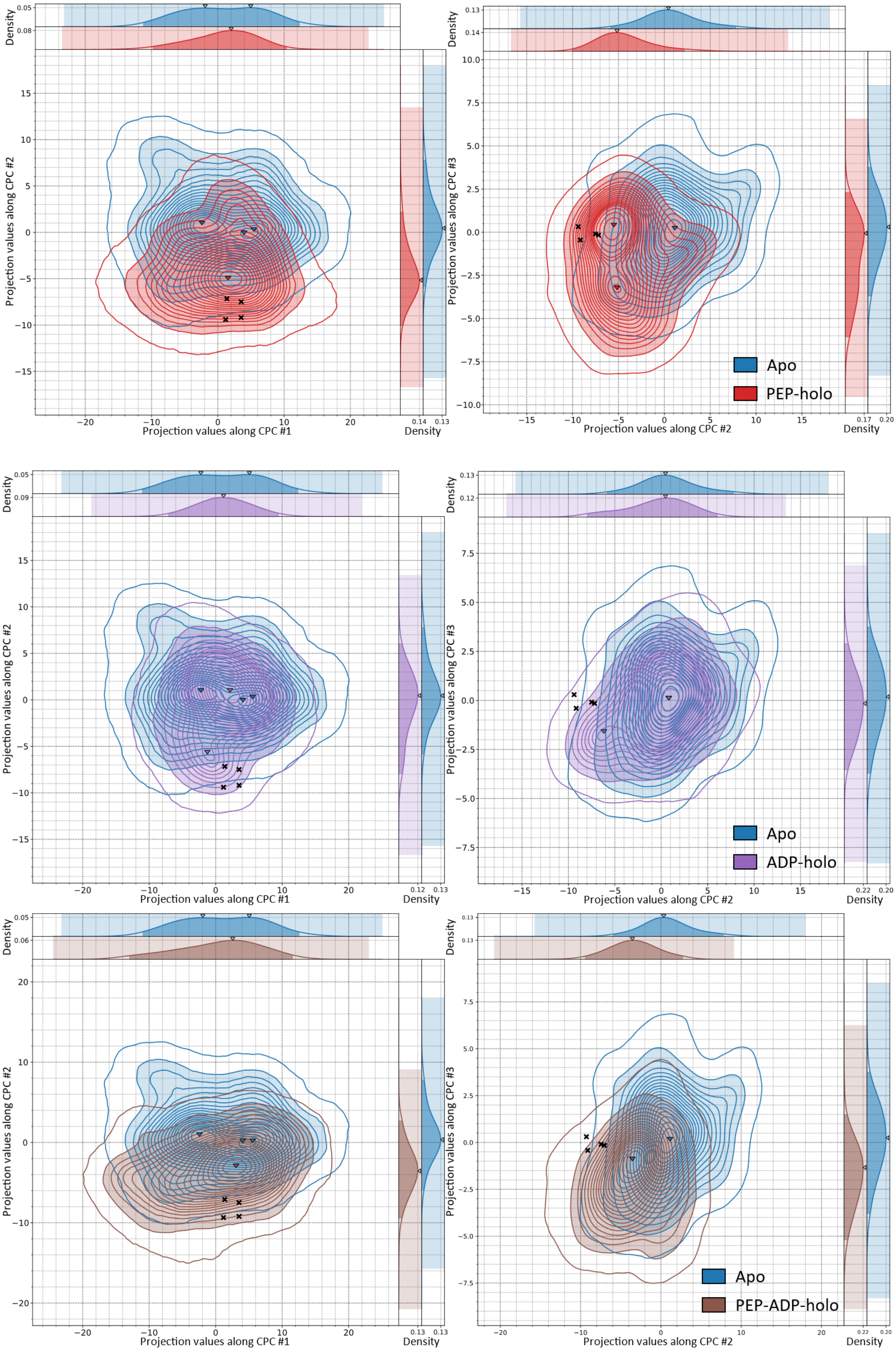
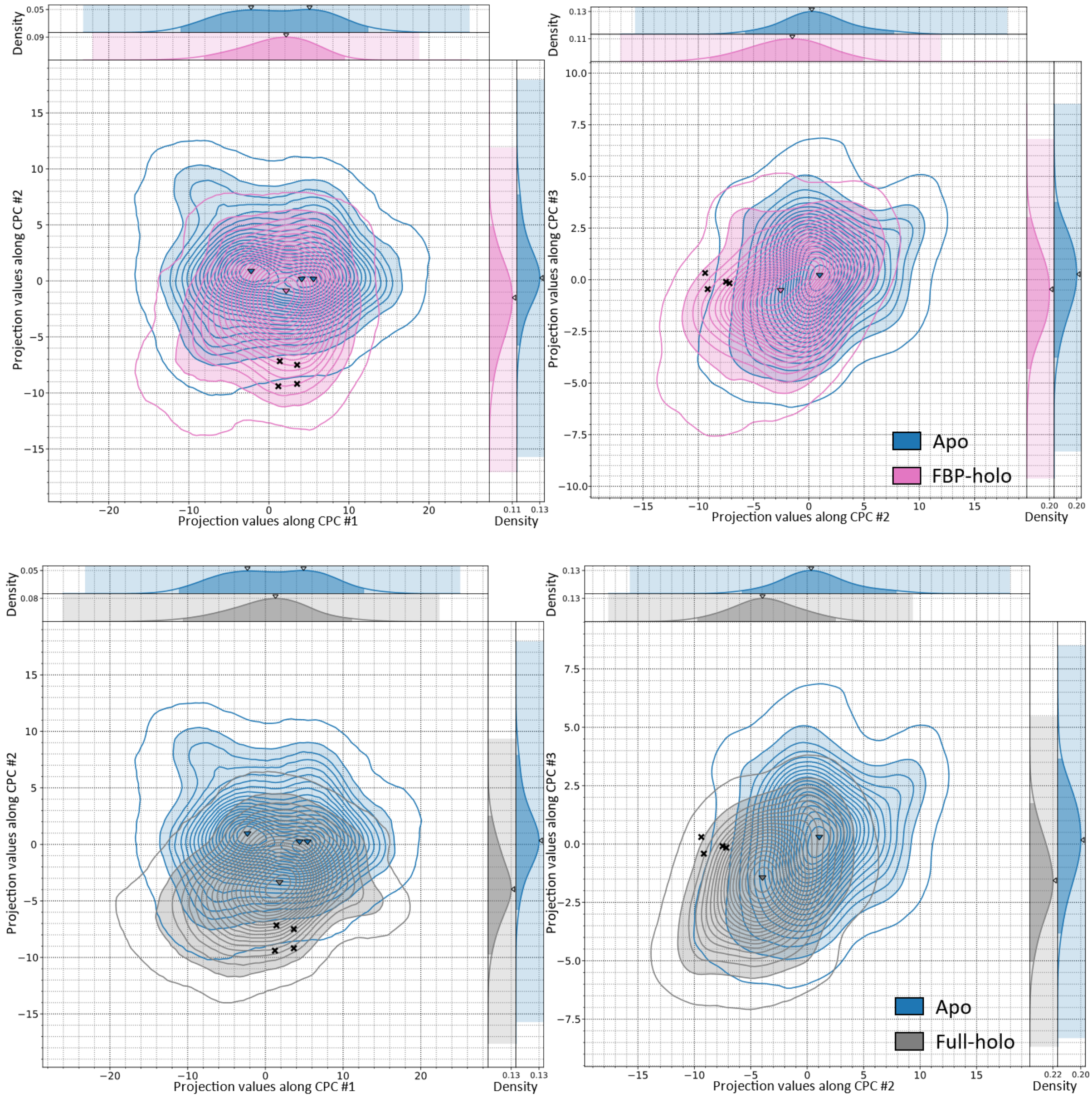


Figure 4.63 (Continued)



The changes in the conformational profiles of the holo simulations with respect to the apo simulations involve moderate shifts in the centers of the density distributions. The most significant differences occurred along the conformational spectrum of CPC #2, as anticipated by the previous observations of the localization of the projection values of structure 2VGB. The density distributions along CPC #3 also reveal mild differential behavior of certain trajectory ensembles. In contrast, all holo density distributions predominantly overlap with the reference distribution along CPC #1, which is indicative of an overall equivalent exploration of the same conformational space. In some instances, however, there is appreciably less dispersion of values with respect to the center of the distributions. The presence of cofactors at the active site (K-holo and K-Mg-holo conditions) did not induce significant differential sampling of the conformational space along the three analyzed CPCs, as shown by the high similarity of the density distributions with the apo condition.

Consistent with the previous CEDAs performed on the other protein regions, all holo conditions with bound PEP (PEP-holo, PEP-ADP-holo, and Full-holo) exhibit the most prominent differential behavior with respect to the apo condition. Along CPC #2, the corresponding distributions are shifted towards the distinctive localization of the structure 2VGB in the conformational spectrum, thus establishing consistency between crystallographic and dynamical data. Although the centers of such distributions are not fully aligned with the projection values of 2VGB, these characteristic shifts nonetheless suggest the existence of a conformational equilibrium between active and inactive states, whereby each state is stabilized by the presence or absence of PEP, respectively. The distribution of the PEP-holo condition is the one that exhibits the greatest overlap with the projection values of 2VGB, as shown by the two-dimensional plots. This condition also displays a heavy tail that extends towards the inactive state. When MgADP is the only bound substrate (ADP-holo condition), the conformational equilibrium is found in a state more reminiscent of that of the apo condition (especially the second apo trajectory ensemble; Figure 4.61), whereby the protein mainly remained in the inactive conformation albeit the sampling of the characteristic region of the PEP-bound conformation was retained in minor proportions.

Along CPC #3, the displayed shifts in the distributions are harder to interpret, although in general they represent conformational changes of lower significance than those of CPC #2, given that the distance between centers of the distributions differs in less units of projection value. PEP-holo exhibits bimodality: while the main region of density is aligned with the distribution of the apo condition, it displays distinctive sampling of the area around the projection value of -3.5. This region of the conformational spectrum of CPC #3 corresponds to a form of the C-C’A’ block rotated counterclockwise as seen from the top view. Such a conformation was less sampled in the other trajectory ensembles and deviates from the characteristic projection values of 2VGB. With the incorporation of MgADP at the active site (PEP-ADP-holo and Full-holo conditions), unimodality is regained and the corresponding distributions become centered around the projection value of 1.5, closer to the distributions of both the apo and ADP-holo conditions.

The differential behavior of the FBP-holo deserves separate attention. Notably, its density distribution along CPC #2 exhibits an intermediate location between those of the PEP-bound holo conditions and the apo condition. Such an observation can be especially noted in the light of the two-dimensional plots of CPCs #2 *vs*. #3, between conditions. A similar effect was observed earlier when studying the collective motions of the adjacent A domains (section 4.1.3.2.1). However, this observation holds particular relevance in the context of the present analysis, as the analyzed region includes the allosteric site and its structural environment. The similarities between the results of both experiments, the implications of the presence of FBP, and the possible connections between the conformational changes at the C-C′ and A-A′ interfaces are further explored in the next section (4.1.3.2.3).

The principal traits discerned from the visual examination of the density distributions are reflected in the ranges of values of the similarity metrics along CPCs #1 to #3. The total score provided by the three metrics facilitates distinction of the different behaviors of the holo conditions with respect to the apo condition. The high similarity of the conformational space of both the K-holo and K-Mg-holo conditions with that of the apo condition is expressed with high values of all metrics. The next condition in order of similarity is the ADP-holo condition, which only exhibits mildly lower values of coverage and BC that reflect the retention of a fraction of sampling in the characteristic region of the active conformation of the enzyme. Next is the FBP-holo condition, which displays coverage and BC values similar to those of ADP-holo; this time due to the mild shift in the center of the unimodal distribution towards the active conformation. For this reason, the value of the overlap is lower than in ADP-holo. Finally, the stronger shifts of the density distributions of the PEP-bound holo conditions are reflected in lower values of all three metrics, especially the PEP-holo condition that is the one with more volume of sampling farther from the apo conformational space.

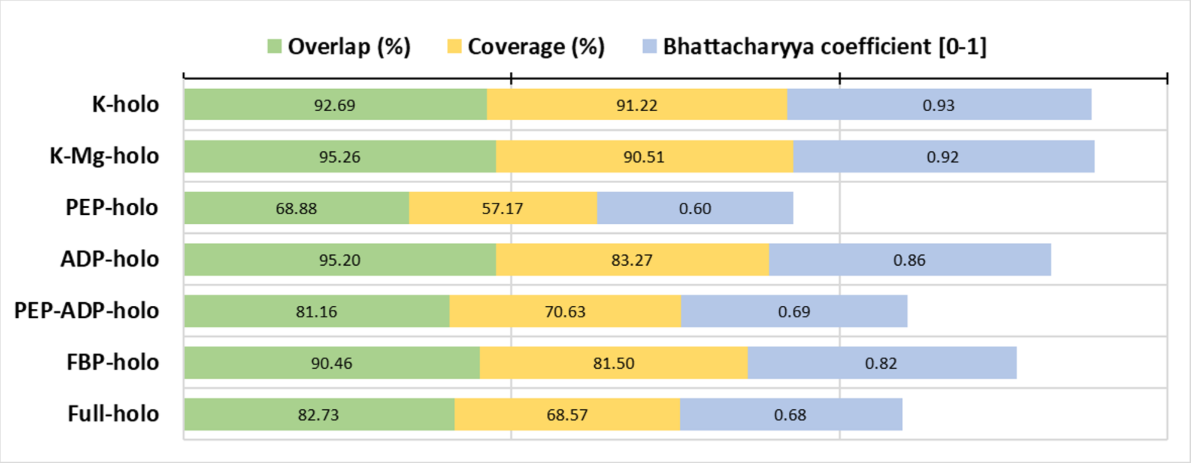


Figure 4.64. Comparative analysis of the WT trajectories (AC-C'A' pairs of cores) in apo *vs*. holo conditions: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. The bar sizes of the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

4.1.3.2.3 Correspondence between the conformational changes at the A-A′ and C-C′ interfaces

With the outcome from the independent CEDAs of each region of PKR, we can subsequently search for signs of correspondence between the detected conformational profiles of the alternative simulation conditions. Specifically, the results from the two previous sections provide a chance to identify potential couplings between the distinctive collective motions at the A-A′ and C-C′ interfaces of PKR.

Remarkably, significant correspondence has been detected between the information provided by the second CPC from both CEDAs (central panels of Figures 4.49 and 4.62). The two conformational spectrums display highly similar patterns of differential behavior among the apo and holo conditions. In both scenarios, two characteristic regions of the conformational spectrum are distinguished. The sampling of one or the other primarily depended on the presence of the substrate PEP at the active site, with the PEP-bound holo conditions being close to the location of the crystallographic structure 2VGB. Thus, from a functional point of view, these two regions supposedly correspond with the active and inactive states of the enzyme. This observation suggests that the collective motions captured in terms of each CPC potentially occurred in a concerted manner. Or, in other words, they account for complementary perspectives of a single transition between the states.

In order to test this hypothesis and to achieve further characterization of the nature of the observed collective motions along CPC #2 from both CEDAs, the corresponding transitions were re-examined by changing the point of reference through which the motion is visually inspected. To this end, the average tetrameric structure of the trajectory ensemble of the apo condition was generated and used as a common reference for structural superpositions (least-squares fitting). For each transition, a different fitting group was employed, superposing it to the analogous region of the reference tetrameric structure. For the transition of the AC-C’A’ pair of cores, the adjacent C domains were set as the fitting group. For the transition of the A-A’ pair of domains, the whole structure was set as the fitting group.

The alternative perspective of the transition of the AC-C’A’ pair of cores along CPC #2 no longer features rotational and translational components of the C domains, thus facilitating the inspection of the local conformational changes at the C-C’ interface. Figure 4.65 and the Supplementary Video S4.28 show the corresponding collective motion. From this perspective, the transition from the PEP-unbound to the PEP-bound conformation involves symmetrical local rearrangements across the C-C’ interface (Figure 4.65c). At the section furthest from the center of the tetramer, the gap of the interface widens with the mutual separation of the local region of the adjacent C domains. In particular, the Cα5 helix and the L-Cβ3-Cα5 loop from each domain are points of high amplitude of motion that determine the gap width of this section of the interface. At the opposite section of the interface, close to the center of the tetramer, the Cα1 helices from the adjacent C domains reorient and become more parallel between each other.

Importantly, the position of the loop of the allosteric site (L-Cβ4-Cβ5) also fluctuates coupled to these rearrangements. The role of this loop is to lock the FBP molecule bound at the allosteric site by covering it via stabilizing interactions. Accordingly, in the apo simulations, the removal of FBP from the initial crystallographic structure has enabled the unfolding motion of the loop. In contrast, the FBP-holo simulations should retain the folded state of the loop due to the presence of FBP. The conformational profile of the FBP-holo condition along CPC #2 is characterized by a wide density distribution that is centered around an intermediate position between the PEP-bound and -unbound states. Therefore, results suggest that the presence of FBP might be preventing the total shift of the equilibrium towards the inactive state, even in the absence of PEP. This observation, together with the fact that the unfolding motion of the loop (symmetrically in the adjacent C domains) is a strong consensus behavior among the apo trajectories in the transition towards the inactive state, suggests that the position of the loop might exert an influence in its direct surroundings, helping to break or form other interactions related with the observed rearrangements of the C-C’ interface. These functional implications have been suggested in other studies [124], [128], [148], [149], [165] and are described more in depth in the Discussion chapter.

This perspective of the transition also enables observation of the rigid-body motions of the A domains that are correlated with the rearrangements of the C-C’ interface. The A domain of the upper subunit used to be the fitting group of the transition. Now, it undergoes a swinging motion along the plane of the tetramer, analogous to that of the original transition but under the point of reference of the fitted C domains (Figure 4.65b). The motion of the A domain accompanies the local rearrangements of the structural elements of the C domain that have been previously described. On the other hand, in the alternative perspective of the transition of A-A’ pair of domains along CPC #2, both A domains display gyration motions that involve the mutual approach of the adjacent active sites (Figure 4.66 and Supplementary Video S4.29). The comparison of both transitions reveals that the A domains follow qualitatively comparable paths of motion. An overlay of both transitions, superposed to the average tetrameric structure, was generated and can be viewed in the Supplementary Video S4.30 for a clearer inspection of the motions.

.

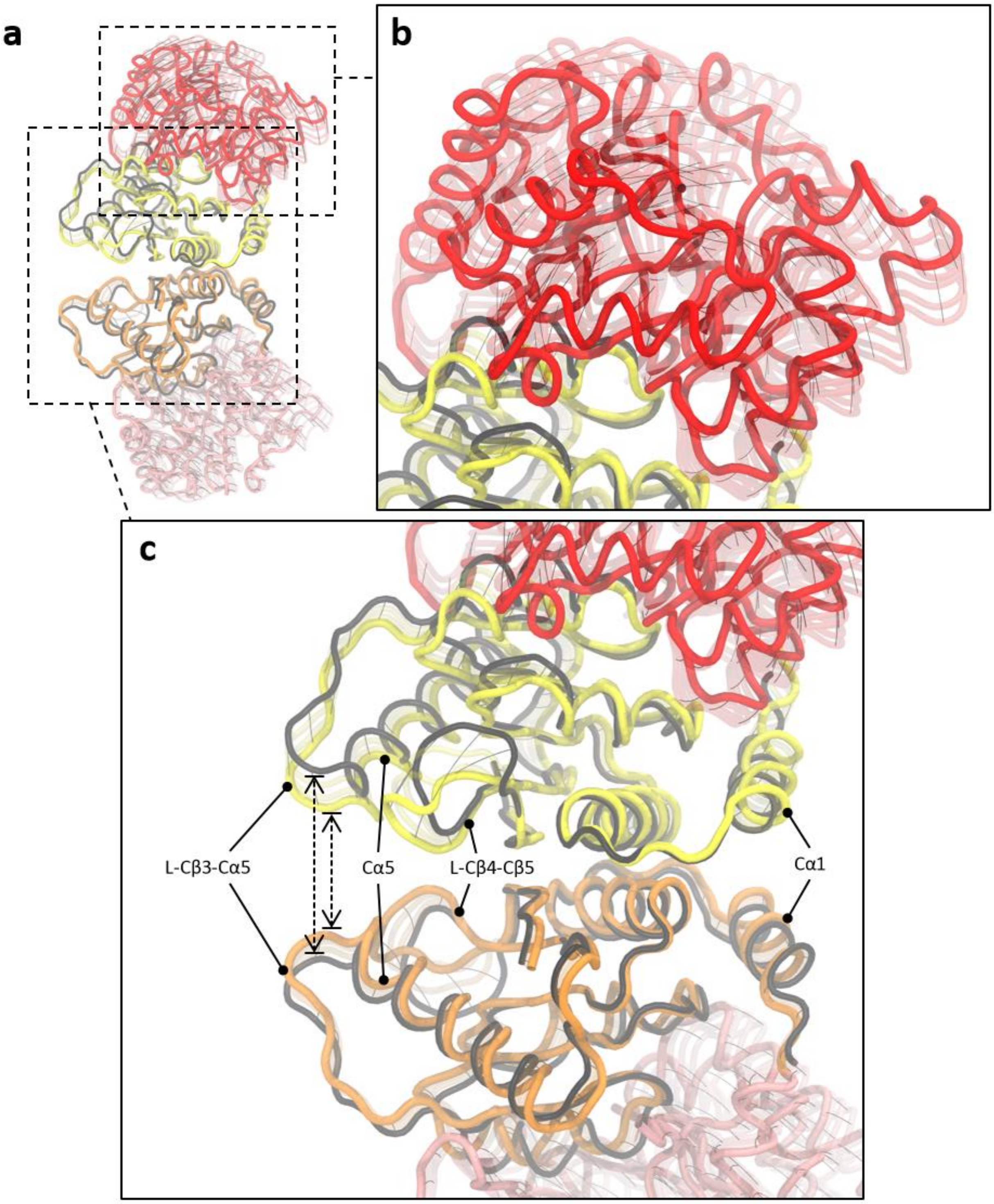


Figure 4.65. Alternative perspective of the collective motion of the adjacent AC cores across the C-C′ interface along CPC #2 in the WT apo condition. The protein is depicted with the trace representation between Cα atoms. Domains are colored as follows: A in red, C in yellow, A′ in pink, and C′ in orange. The opaque structure corresponds to the PEP-unbound conformation, and a few conformations of the transition to the PEP-bound conformation are represented with transparent structures. The PEP-bound conformation of the C domains is colored in black to better highlight the conformational differences of the C-C′ interface. The black lines connecting the intermediate structures indicate the path of motion of Cα atoms. This motion can also be viewed in the Supplementary Video S4.28. (**a**) General view of the whole structure. (**b**) Swinging motion of the A domain of the upper subunit. (**c**) Local structural rearrangements of the C-C′ interface. NOTE. The images were generated with the software VMD.

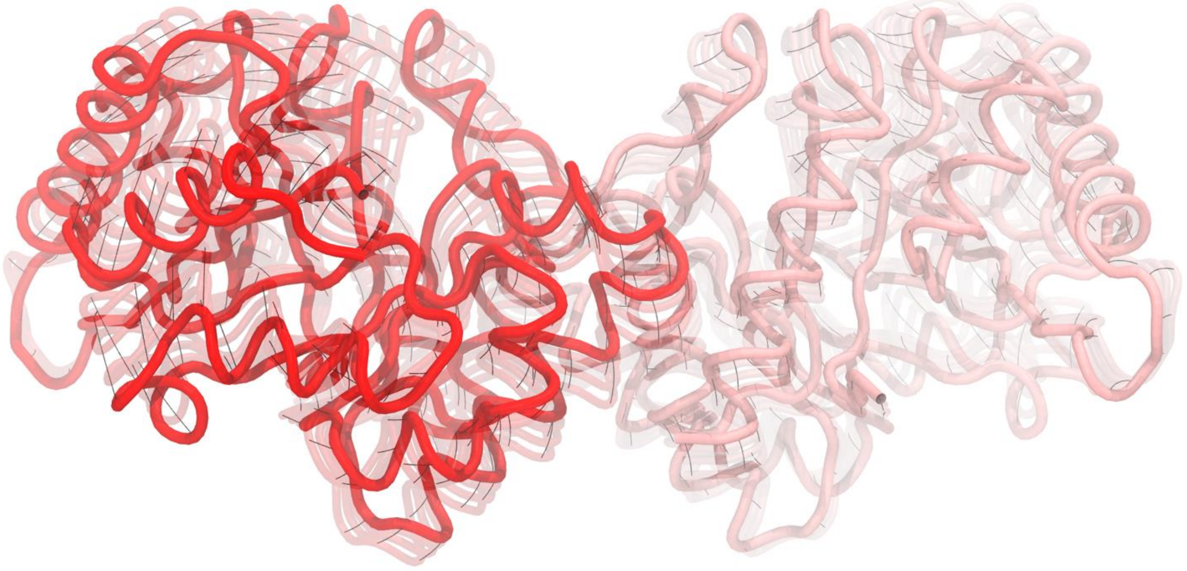


Figure 4.66. Alternative perspective of the collective motion of the adjacent A domains across the A-A′ interface along CPC #2 in the WT apo condition. The protein is depicted with the trace representation between Cα atoms. The A and A′ domains are colored in red and pink, respectively. The opaque structure corresponds to the PEP-unbound conformation, and a few conformations of the transition to the PEP-bound conformation are represented with transparent structures. The black lines connecting the intermediate structures indicate the path of motion of Cα atoms. This motion can also be viewed in the Supplementary Video S4.29. NOTE. The image was generated with the software VMD.

4.2 Analysis of missense variants of PKR

Following from the study of the WT PKR, the next experimental part of the project revolved around the characterization of the functional effects of missense variants of the enzyme. The following sections elaborate on the conducted set of actions aligned with this goal. First, the construction process of a catalog of the known missense variants of PKR is explained, along with the subsequent selection of a representative set of variants for simulation. Secondly, the modeling procedure of the corresponding amino-acid substitutions and the performed simulations in apo and holo conditions are reviewed, including a few notes about the stability and flexibility analyses of the trajectories. The last section represents the most important part of this block of results, in which the CEDA strategy was applied to detect potential dynamical alterations in the simulated variants in comparative terms with respect to the WT behavior.

4.2.1 Construction of a comprehensive database of missense mutations of PKR

Pyruvate kinase deficiency (PKD) is the most frequent glycolytic enzymopathy and one of the most common causes of the disease known as chronic hereditary non-spherocytic hemolytic anemia. Over the last decades, an extensive volume of genetic variations have been identified in the gene *PKLR* that cause PKD by either functional impairment of the enzyme or lack of protein production, thus representing a paradigmatic case of a monogenic disorder. A considerably large portion of the detected genetic mutations produce missense variants or single amino-acid variants (SAVs) of the protein [199].

In order to obtain a global view of the identified missense variants of this protein, a comprehensive exploration of public databases of genetic variants was performed, comprising repositories of clinical annotations and genome-sequencing data portals (see section 3.6 of the Methods chapter). The search was complemented with a subsequent extensive scan of the available literature to gather further clinical and functional annotations. This exercise culminated with the integration of the gathered data in a database that serves as a catalog of the known missense variants of PKR. The publication of the full database lies outside of the scope of this thesis, as reproduction and distribution of the contents is currently not possible in order to comply with copyright policies of several of the queried resources. Below is a detail of the dimensions of the database, together with an analysis of how the available types of data sources interrelate with each other.

The constructed database contains a total number of 789 unique missense variants from diverse sources. The first informative distinction is given between the missense variants characterized by a somatic *vs*. germline origin. The former correspond to variants that were retrieved from the COSMIC and BioMuta repositories (*i.e.*, mutations implicated in the development of cancer; putatively somatic) and that were not found in the rest of the queried sources. These are 130, thus representing 16.5% of the total. Conversely, the latter correspond to variants that have been detected either in large-scale genome-sequencing projects or in clinical studies of hereditary PKD. These are 564, thus representing 71.5% of the total. Finally, there is a fraction of variants that are only registered in dbSNP and correspond to entries submitted by diverse screening projects of smaller scale or independent laboratories, or that simply lack clear records on their provenance. It is hard to determine their origin without a case-by-case inspection and, therefore, they were left unclassified. These are 95, thus representing 12% of the total.

For this study, only the variants with a germline origin have been considered, as they account for the genetic variation in the human population and, thus, are the target of clinical studies of the incidence and transmission of pathogenic mutations associated with PKD. It is important to note that the aforementioned distinction between somatic and germline origins based solely in terms of database provenance is not entirely reliable. Overlaps between the two types of repositories can be found, thus creating ambiguities in the true nature of the variants.

The extent of contradictory information was examined with the measurement of the overlap of the data between the corresponding types of repositories. From the 408 entries retrieved from the ExAC and gnomAD datasets, 67 can also be found in COSMIC or BioMuta. The documentation from the COSMIC database does warn about the unavailability of definitive evidence for putative somatic mutations incorporated from the published literature, as well as the possible occurrence of annotation mistakes in laboratories and the intrinsic error rate of DNA sequencing methods. Consequently, entries include a “somatic status” label that states whether the mutation has been confirmed to have somatic origin or not. Even so, the number of entries in COSMIC with confirmed evidence of somatic origin that are also found in ExAC and gnomAD is 31.

A total of 258 entries of the catalog are associated with PKD. From those, 201 were retrieved from the queried repositories that include annotations on the phenotype or the clinical manifestations. While there is high overlap between the different sources, they did not contribute equally. The number of retrieved entries, in descending order, was the following: 185 from HGMD, 156 from LOVD, 101 from Humsavar/SwissVar, and only 24 from ClinVar. On the other hand, 55 of the PKD-associated missense variants of the catalog were acquired thanks to the generous collaboration of Dr. Richard van Wijk from his personal clinical research. In addition, 2 extra variants were retrieved from the exploration of the recent literature. These variants were not found in any of the queried repositories.

In respect of the overlap with other types of repositories, 36 of the PKD-associated variants can also be found in COSMIC or BioMuta, with 11 instances from COSMIC claiming evidence of somatic origin and thus holding contradictory information. On the other hand, 102 of the PKD-associated variants also appear in the ExAC and gnomAD datasets. Even though in such projects of large-scale genome sequencing the aim is to annotate DNA variation in the worldwide human population without interference of disease-associated genotypes, there is certainly a chance that some participants actually suffer from undiagnosed or neglected diseases. Moreover, PKD is an autosomal recessive disorder and, therefore, an individual can be a carrier of a harmful variant and still not develop the disease. Accordingly, the majority of instances correspond to detections in a few heterozygous individuals (*i.e.*, variant carriers in just one allele), normally less than ten and possibly up to a few dozens, with allele frequencies between $4×10^{-6}$ and $5×10^{-4}$, approximately. There are also a few isolated instances with a few hundreds of detections that may reach allele frequencies around $4×10^{-3}$. Some of these are notoriously known for being among the most commonly reported mutations in PKD patients, such as Arg486Trp and Arg510Gln [199]. Interestingly, 7 of the PKD-associated variants were detected in the homozygous state in ExAC or gnomAD, although only in one to five individuals at most.

Finally, a couple of additional remarks. The extent of overlap between the v2 and v3 datasets of gnomAD was examined. From the total of 398 unique missense variants coming from both datasets, 174 are exclusive of the v2 set and 65 of the v3 set. On another note, the data imported from the “pyruvate kinase” CAGI challenge was able to provide experimental kinetic values for 34 missense variants of the catalog, involving entries from diverse sources.

4.2.2 Selection, setup and simulation of the PKR mutant structures

In accordance with the objectives of the thesis, the next goal was to evaluate the capabilities of the framework of CEDA to detect dynamical alterations in missense variants of PKR and assess their functional significance. Since performing MD simulations of all the gathered variants was not feasible due to the prohibitive computational cost, a set of representative variants was selected.

The selection of variants was conducted from two angles. The first consideration was to incorporate a fraction of variants that hold validation or some evidence of pathogenicity and a fraction of variants that are potentially neutral or benign. Such a distinction allowed for testing of whether the strongest signs of dynamical alteration are more likely to be found in variants already classified as pathogenic via other methods. The candidate variants to be included in the pathogenic subset were those with a suggested implication in PKD, as reported in clinical studies or experimental assays. Typically, these correspond to the entries of the HGMD, LOVD, ClinVar, and Humsavar/SwissVar repositories or, alternatively, to independent variants found in the literature. Variants repeatedly reported as pathogenic by different sources were prioritized. On the other hand, the variants retrieved from the data portals of large-scale sequencing projects (and not overlapping with the pathogenic subset) were the candidates to be part of the potentially neutral subset. In this regard, variants with an explicitly declared somatic origin or with incomplete information were not considered for the selection.

On the basis of this primary classification, further criteria were introduced to achieve a diverse and balanced representation of features according to the position and the structural and functional contexts of the protein. The final set of variants satisfied all the following criteria, in order of priority: i) variants affecting the same (or close) positions both in the pathogenic and potentially neutral conditions, ii) representation of all four domains of the protein, iii) representation of the active and allosteric sites and possibly specific residues that participate in ligand binding or the catalytic mechanism, iv) representation of the interdomain and intersubunit interfaces, and v) representation of other protein regions without particular documentation as functional sites and comprising both solvent-exposed and buried positions. This last consideration is of particular interest from the perspective that structural perturbation at any site may cause population shift of the conformational ensemble, in accordance with the current understanding of the allosteric phenomenon [38], [50]. In other words, the aim was to broaden the search for dynamical alterations by considering amino-acid replacements at positions beyond the more obvious functional regions.

A total of 61 variants were selected for subsequent MD simulation, with 40 in the pathogenic subset and 21 in the potentially neutral subset. Table 4.4 shows the list of variants along with details on the features of their position in the protein and the public repositories where they can be found (except for COSMIC and BioMuta which have been omitted for the reasons given above). All variants from the pathogenic subset are reported in the literature; a full list of the corresponding bibliographic references can be found in Table 1 of Appendix B of this manuscript. Figure 4.67 shows the 2VGB structure highlighting the location of the mutated residues.

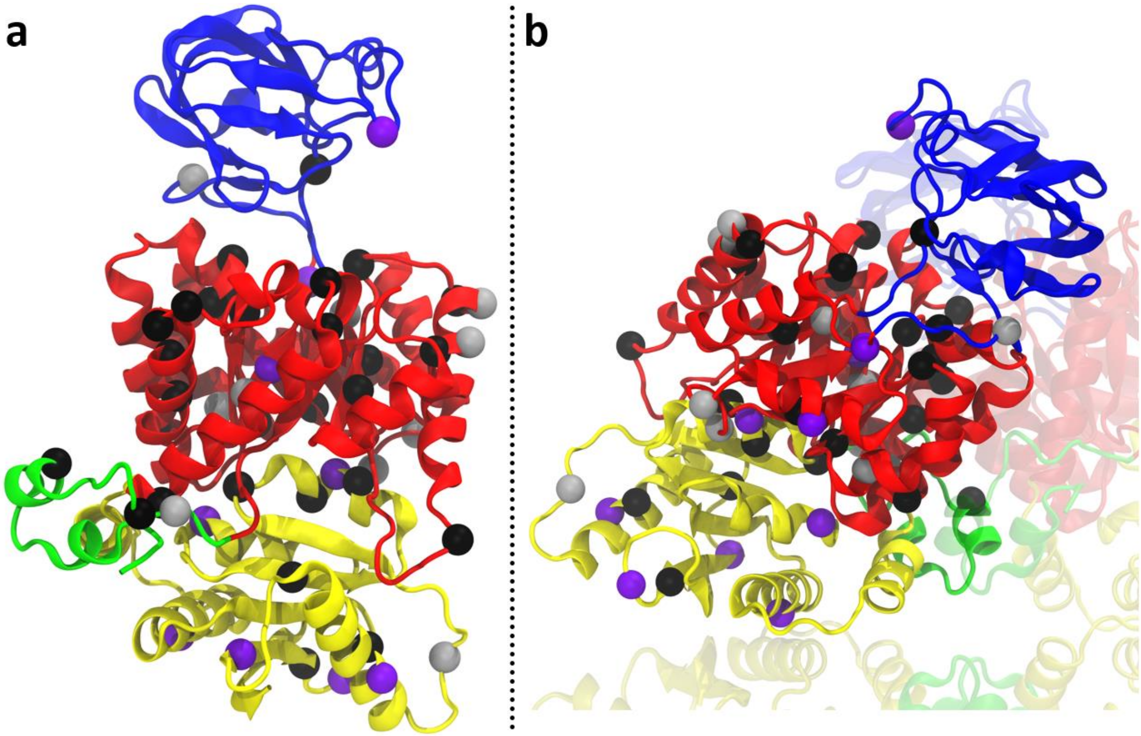


Figure 4.67. Location of the selected missense variants of PKR. Panels show the 2VGB structure depicted with a ribbon representation. The positions of the mutated residues are indicated with spheres. Purple spheres correspond to positions shared by variants in both the pathogenic and the potentially neutral subsets. Black and gray spheres correspond to positions of variants from the pathogenic and potentially neutral subsets, respectively. Domain regions are colored as follows: N-terminal domain in green, A domain in red, B domain in blue, and C domain in yellow. (**a**) Monomeric structure of PKR. (**b**) A subunit of PKR in the context of the tetrameric structure. The rest of subunits are represented with transparent structures. NOTE. The images were generated with the software VMD.

Table 4.4
Selected missense variants of PKR

Variant	Subset [a]	Protein region	Functional site	Ligand binding	Interface	Sources [b]
Leu73Pro	P	Nα2	-	Mitapivat [c]	Nt-A′	gnomAD (v3), HGMD, Humsavar, LOVD
Ser80Pro	P	L-Nα2-Aβ1	-	-	Nt-A′	HGMD, Humsavar, LOVD
Glu81Lys	N	L-Nα2-Aβ1	-	-	Nt-A′	dbSNP, ExAC, gnomAD (v2+v3)
Ala115Pro	P	Aβ2	-	-	-	HGMD, Humsavar, LOVD
Ser120Phe	P	L-Aβ2-Aα2	Active site	K^+	A-B	HGMD, Humsavar, LOVD
His124Gln	P	Aα2	-	-	-	dbSNP, gnomAD (v2)
Glu125Ala	N	Aα2	-	-	-	dbSNP, ExAC, gnomAD (v2)
Glu129Lys	N	Aα2	-	-	-	dbSNP, ExAC, gnomAD (v2+v3)
Ser130Tyr	P	Aα2	-	-	-	ClinVar, dbSNP, HGMD, Humsavar, LOVD
Gly143Ser	P	L-Aα2-Aβ3	-	-	-	HGMD, LOVD
Leu155Pro	P	Aβ3	-	-	-	HGMD, Humsavar, LOVD
Thr157Pro	N	Aβ3	Active site	K^+	-	dbSNP, ExAC, gnomAD (v2)
Arg163Cys	P	L-Aβ3-Bβ1	-	ADP	A-B	ClinVar, dbSNP, HGMD, Humsavar, LOVD
Glu172Gln	P	L-Aβ3-Bβ1	-	-	-	dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Glu172Gly	N	L-Aβ3-Bβ1	-	-	-	dbSNP, gnomAD (v2+v3)
Ala257Thr	N	L-Bβ8-Aα3	-	-	-	dbSNP, ExAC, gnomAD (v2)
Gly263Ala	N	L-Bβ8-Aα3	-	-	-	dbSNP, gnomAD (v2)
Gly263Trp	P	L-Bβ8-Aα3	-	-	-	HGMD, Humsavar, LOVD
Ala295Thr	N	Aα4	-	-	-	dbSNP, ExAC, gnomAD (v2+v3)
Ala295Val	P	Aα4	-	-	-	dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Pro303Leu	N	L-Aα4-Aβ5	-	-	-	dbSNP, ExAC, gnomAD (v2+v3)

Table 4.4 (Continued)

Gly307Ser	N	L-Aα4-Aβ5	-	-	A-C	dbSNP, ExAC, gnomAD (v2+v3)
Ile310Asn	P	Aβ5	-	-	-	HGMD, Humsavar, LOVD
Glu315Lys	P	L-Aβ5-Aα5	Active site	Mg^{2+}	-	ClinVar, dbSNP, HGMD, Humsavar, LOVD
Leu327Val	N	Aα5	-	-	-	dbSNP, ExAC, gnomAD (v2+v3)
Gly332Ser	P	Aβ6	-	-	-	ClinVar, dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Arg337Gln	P	Aβ6	Active site	-	A-A′	dbSNP, HGMD, Humsavar, LOVD
Asp339His	P	Aα6′	Active site	Mg^{2+}, PEP	A-B	dbSNP, gnomAD (v2), HGMD, Humsavar, LOVD
Arg359Cys	P	Aα6	-	-	A-Nt′	dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Thr371Ile	P	L-Aβ7-Aα7′	Active site	PEP	A-A′	HGMD, LOVD
Thr384Met	P	Aα7	-	-	A-A′	ClinVar, dbSNP, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Arg385Lys	P	Aα7	-	-	A-A′	HGMD, LOVD
Asp390Asn	P	Aα7	-	-	A-A′	dbSNP, HGMD, Humsavar, LOVD
Ala394Asp	P	Aα7	-	-	-	dbSNP, HGMD, Humsavar, LOVD
Ala394Val	P	Aα7	-	-	-	HGMD, Humsavar, LOVD
Ile402Val	N	Aβ8	-	-	-	dbSNP, gnomAD (v2+v3)
Met403Ile	P	Aβ8	Active site	-	-	dbSNP, ExAC, gnomAD (v2+v3), HGMD, LOVD
Met403Thr	N	Aβ8	Active site	-	-	dbSNP, ExAC, gnomAD (v2)
Thr408Ile	P	Aα8′	-	-	-	HGMD, Humsavar, LOVD
Gly411Ser	P	Aα8′	-	-	-	HGMD, LOVD
Ala430Thr	P	Aα8	-	-	Nt-A, A-C	gnomAD (v2), HGMD
Gly458Ala	N	Cα2	-	-	C-C′	dbSNP, gnomAD (v2)
Gly458Asp	P	Cα2	-	-	C-C′	dbSNP, ExAC, gnomAD (v2), HGMD, Humsavar, LOVD
Arg486Gln	N	Cα3	-	-	A-C	dbSNP, ExAC, gnomAD (v2+v3)
Arg486Trp	P	Cα3	-	-	A-C	ClinVar, dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD

Table 4.4 (Continued)

Ile494Thr	P	Cβ2	-	-	-	dbSNP, gnomAD (v2+v3), HGMD, LOVD
Arg504Leu	P	Cα4	-	-	A-C	dbSNP, ExAC, HGMD, Humsavar, LOVD
Gln505Arg	N	Cα4	-	-	A-C	dbSNP, gnomAD (v3)
Gln505Glu	P	Cα4	-	-	A-C	HGMD
Val506Ile	P	Cα4	Amino-acid allosteric site	-	-	ClinVar, dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Arg510Gln	P	L-Cα4-Cβ3	Amino-acid allosteric site	-	A-C	ClinVar, dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Arg518His	N	L-Cβ3-Cα5	-	-	A-C	dbSNP, ExAC, gnomAD (v2+v3)
Pro521Ser	N	L-Cβ3-Cα5	-	-	-	dbSNP, ExAC, gnomAD (v2+v3)
Arg531Cys	P	Cα5	-	-	-	HGMD, Humsavar
Arg531His	N	Cα5	-	-	-	dbSNP, gnomAD (v2+v3)
Arg532Trp	P	Cα5	FBP allosteric site	FBP	-	dbSNP, ExAC, gnomAD (v2+v3), HGMD, Humsavar, LOVD
Val552Ala	N	Cβ4	-	-	-	dbSNP, ExAC, gnomAD (v2+v3)
Val552Met	P	Cβ4	-	-	-	dbSNP, HGMD, Humsavar, LOVD
Gly557Ala	P	Cβ4	FBP allosteric site	-	-	HGMD, Humsavar, LOVD
Arg559Gln	N	L-Cβ4-Cβ5	FBP allosteric site	FBP	-	dbSNP, gnomAD (v3)
Arg559Gly	P	L-Cβ4-Cβ5	FBP allosteric site	FBP	-	HGMD, Humsavar, LOVD

NOTE. Abbreviations: ADP, adenosine diphosphate; FBP, fructose 1,6-bisphosphate; PEP, phosphoenolpyruvate.

[a] The variants from the pathogenic subset (labeled as “P”) have been suggested to be implicated in pyruvate kinase deficiency, as reported in clinical studies or experimental assays (either from public repositories or the literature). The variants from the potentially neutral (labeled as “N”) were retrieved from the data portals of large-scale genomic sequencing projects and have not been associated with pyruvate kinase deficiency.

[b] List of public repositories where the corresponding variants can be found (databases of variants with a somatic origin not included). In alphabetical order: ClinVar [209]; dbSNP, Single Nucleotide Polymorphism database [193]; ExAC, Exome Aggregation Consortium [188], gnomAD, Genome Aggregation Database (v2 and v3 datasets) [189]; HGMD, Human Gene Mutation Database [203]; Humsavar from the UniProt Knowledgebase [208]; LOVD, Leiden Open Variation Database [210].

[c] Mitapivat (or AG-348) is a synthetic compound that is an allosteric activator of PKR. It binds to a cryptic binding site of the protein, different from that of FBP. It has been recently approved to treat hemolytic anemia in patients with pyruvate kinase deficiency [53]-[55.]

Each amino-acid replacement was modeled at each subunit of 2VGB, as detailed in the Methods chapter (section 3.1.1). The corresponding new side chains were able to occupy the available local space without generating severe atomic clashes with their surroundings. The subsequent energy minimizations removed the steric strain of each model and achieved accommodation of the new side chains among the nearby residues. The only variant that required specific readjustment before the energy minimization was Ser120Phe. Specifically, the expansion from the hydroxyl group of serine to the bulkier phenylalanine led to an overlap with the side chain of Glu161. An alternative side-chain rotamer was selected from a library of the software PyMOL, enabling mitigation of the atomic clashes (Figure 4.68). Incidentally, this variant deprives the cofactor K$^+$ of one of its coordination ligands, which may be the molecular basis of its pathogenicity. In the initial model, the new side chain is blocking a significant portion of the binding site of K$^+$ (see Figure 4.3c), although it may move away during MD.

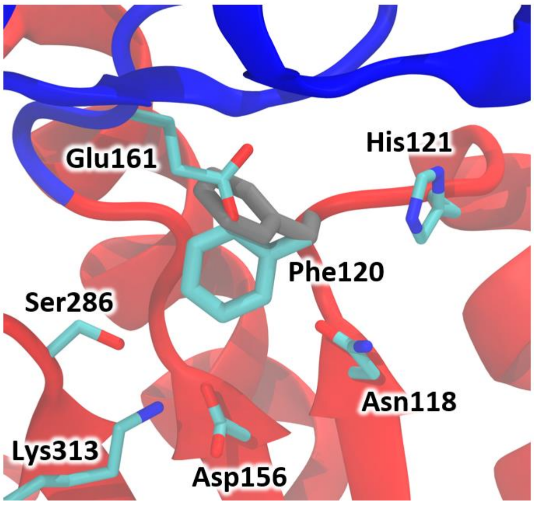


Figure 4.68. Modeling of the Ser120Phe variant of PKR. The replacement of serine by phenylalanine at position 120 in structure 2VGB caused structural overlap with the side chain of Glu161. The side chain of Phe120 is depicted with a thick licorice representation, colored in gray in the initial position and in cyan after selecting the suitable rotamer. The nearby side chains are depicted with a thinner licorice representation and colored by atomic species. Hydrogen atoms are not displayed as they had not been modeled yet. The backbone of the structure is depicted with a ribbon representation, with the A and B domains colored in red and blue, respectively. NOTE. The image was generated with the software VMD.

After the setup stage of the mutant systems, the structures were subjected to MD simulations employing the same protocol as in the WT systems. The whole set of PKR variants was simulated in the apo condition and, additionally, in one of the following holo conditions: K-Mg-holo, PEP-holo, PEP-ADP-holo, or FBP-holo. The particular holo condition for each variant was rationally selected according to the expected type of dysfunction that may be manifested in dynamics, on the basis of both the available clinical annotations and the location of the particular amino-acid substitution. For instance, the above-mentioned Ser120Phe system was simulated in the K-Mg-holo condition based on the hypothesis that the mutation may principally interfere with the sampling of the regular conformations of the K$^+$-bound enzyme (provided that the binding event actually takes place *in vivo*).

For that purpose, the following criteria were taken into consideration. First, variants affecting the same positions both in the pathogenic and potentially neutral conditions were evidently given the same holo condition to facilitate direct comparison between the two classes. Then, the PEP-holo and FBP-holo were the conditions with the most specific candidates. For the PEP-holo condition: i) variants within or near the active site, especially affecting PEP-binding residues; and ii) variants within or near the A-A′ interface. For the FBP-holo condition: i) variants within or near the A-C interface; ii) variants within or near the allosteric sites of the protein; iii) variants within or near the interfaces between N-terminal and A domains (either within or between subunits); and iv) variants scattered around the C domain. Next, the K-Mg-holo condition: i) variants within or near the active site, especially affecting cofactor-binding residues; ii) in general, variants scattered around the A domain and especially affecting the hydrophobic core of the barrel; and iii) a few variants miscellaneously at the C domain. Lastly, the PEP-ADP-holo condition was the least specific: i) variants affecting the B domain, especially at the hinge with the A domain; and ii) variants scattered around the A domain. Table 4.5 shows the corresponding final distribution.

Table 4.5

Simulated holo condition of each PKR variant

Variant	Subset [a]	Holo condition			
		K-Mg-holo	**PEP-holo**	**PEP-ADP-holo**	**FBP-holo**
Leu73Pro	P				X
Ser80Pro	P				X
Glu81Lys	N				X
Ala115Pro	P	X			
Ser120Phe	P	X			
His124Gln	P			X	
Glu125Ala	N			X	
Glu129Lys	N			X	
Ser130Tyr	P			X	
Gly143Ser	P			X	
Leu155Pro	P	X			
Thr157Pro	N	X			
Arg163Cys	P			X	
Glu172Gln	P			X	
Glu172Gly	N			X	
Ala257Thr	N			X	
Gly263Ala	N			X	
Gly263Trp	P			X	
Ala295Thr	N	X			
Ala295Val	P	X			
Pro303Leu	N	X			
Gly307Ser	N				X
Ile310Asn	P	X			
Glu315Lys	P	X			
Leu327Val	N	X			
Gly332Ser	P	X			
Arg337Gln	P		X		
Asp339His	P		X		
Arg359Cys	P				X
Thr371Ile	P		X		
Thr384Met	P			X	
Arg385Lys	P		X		
Asp390Asn	P		X		
Ala394Asp	P		X		

Table 4.5 (Continued)

Ala394Val	P		X		
Ile402Val	N		X		
Met403Ile	P		X		
Met403Thr	N		X		
Thr408Ile	P			X	
Gly411Ser	P			X	
Ala430Thr	P				X
Gly458Ala	N	X			
Gly458Asp	P	X			
Arg486Gln	N				X
Arg486Trp	P				X
Ile494Thr	P	X			
Arg504Leu	P	X			
Gln505Arg	N				X
Gln505Glu	P				X
Val506Ile	P				X
Arg510Gln	P				X
Arg518His	N				X
Pro521Ser	N				X
Arg531Cys	P				X
Arg531His	N				X
Arg532Trp	P				X
Val552Ala	N				X
Val552Met	P				X
Gly557Ala	P				X
Arg559Gln	N				X
Arg559Gly	P				X

[a] Same considerations as in Table 4.4.

After the MD stage, the stability and flexibility of the mutant systems along the trajectories were examined with the RMSD and RMSF analyses. Although a full report on the corresponding results for the whole collection of trajectories could not be included in this manuscript, the analyses provided equivalent information to that of the WT systems, in general. The main observations are summarized hereafter, along with a few examples.

In the apo trajectories, structural divergence was strongly influenced by the asymmetrical sampling of different conformations of the B domains of the tetramer, as indicated by the comparison of the RMSD profiles with and without such regions. The tetramer core (the N-terminal, A, and C domains) exhibits considerable stability. Figure 4.69 shows two examples of strongly stable tetramer cores with conformational changes of the B domains (top panels; variants Gly332Ser and Ala394Asp), and two examples where the tetramer cores exhibit structural divergence along the simulations and among the replicates of the same variant (bottom panels; variants Ala257Thr and Thr408Ile). Similarly to the WT simulations, the cutoff of 25 ns remains a suitable choice to select the endpoint of the initial relaxation of the systems in general. Again, the characteristic relaxation curves are better identified when the B domains are excluded from the analysis. Accordingly, all trajectories had the first 25 ns discarded from subsequent analyses as an extension of the equilibration phase.

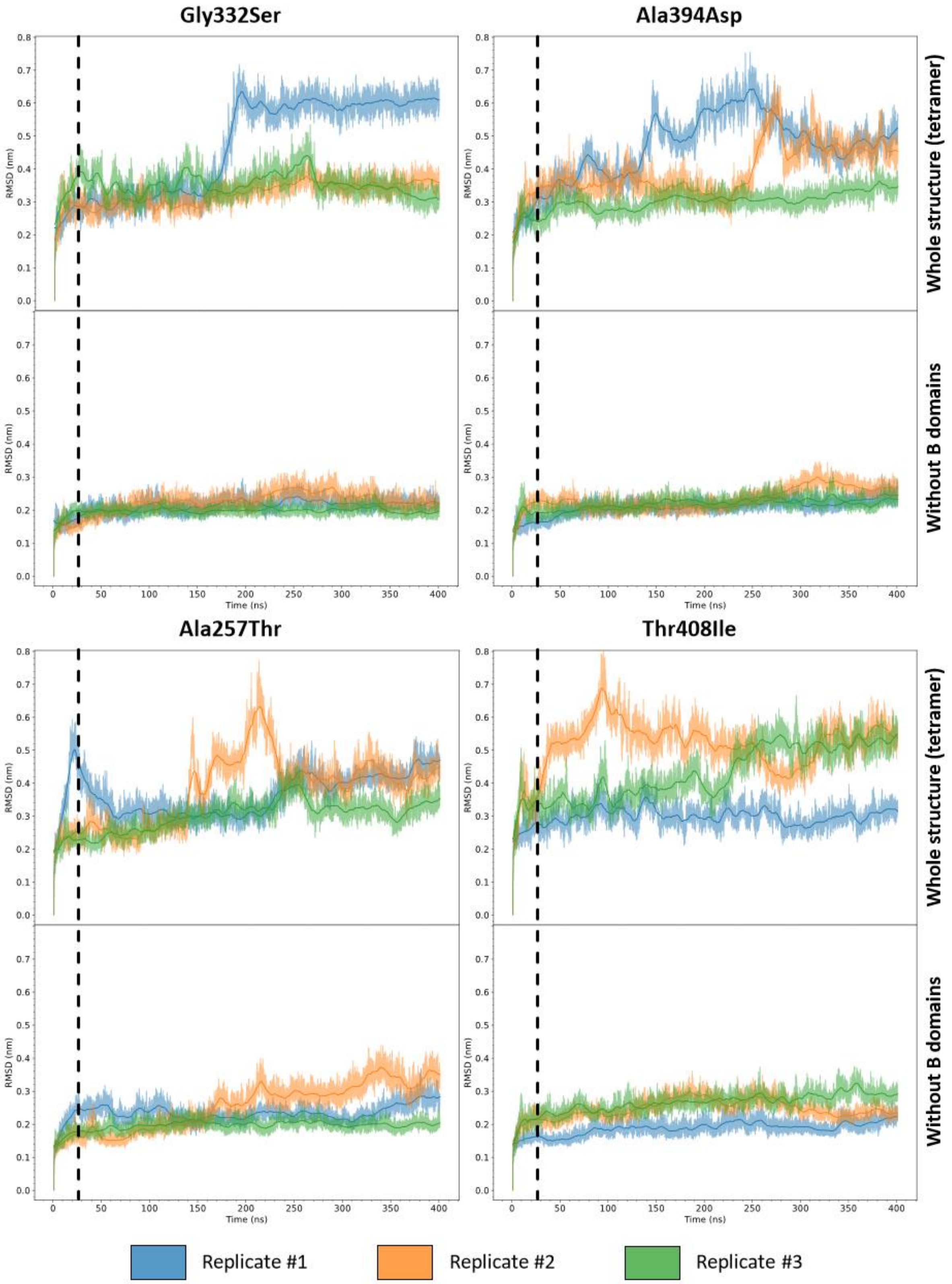


Figure 4.69. Time-series RMSD of the trajectories of four missense variants in the apo condition. The analysis was applied to the protein backbone atoms. The darker line plotted alongside each time series represents the two-sided moving average of the data, encompassing values up to 5 ns on either side of each point. For each variant, the top and bottom panels correspond to RMSD values of the tetramers including and excluding the B domains, respectively. A vertical black dashed line indicates the cutoff of 25 ns that marks the end of the structural relaxation phase.

On the other hand, the holo trajectories exhibit a tendency towards higher stability as the number of ligands bound to the active site increases, as was the case of the WT trajectory ensembles. In the presence of PEP (PEP-holo and PEP-ADP-holo conditions), qualitatively equivalent RMSD profiles can be observed with and without including the B domains in the analysis. Thus, this suggests that the B domain was more tightly coupled to the dynamical events of the tetramer core. Only a few variants may have diverged from this behavior, with the RMSD profile of the whole structure indicating less stability than without considering the B domain. For instance, replicate #1 of Glu172Gly simulated in the PEP-ADP-holo condition (Figure 4.70). This variant affects a position in the frontal region of the B domain.

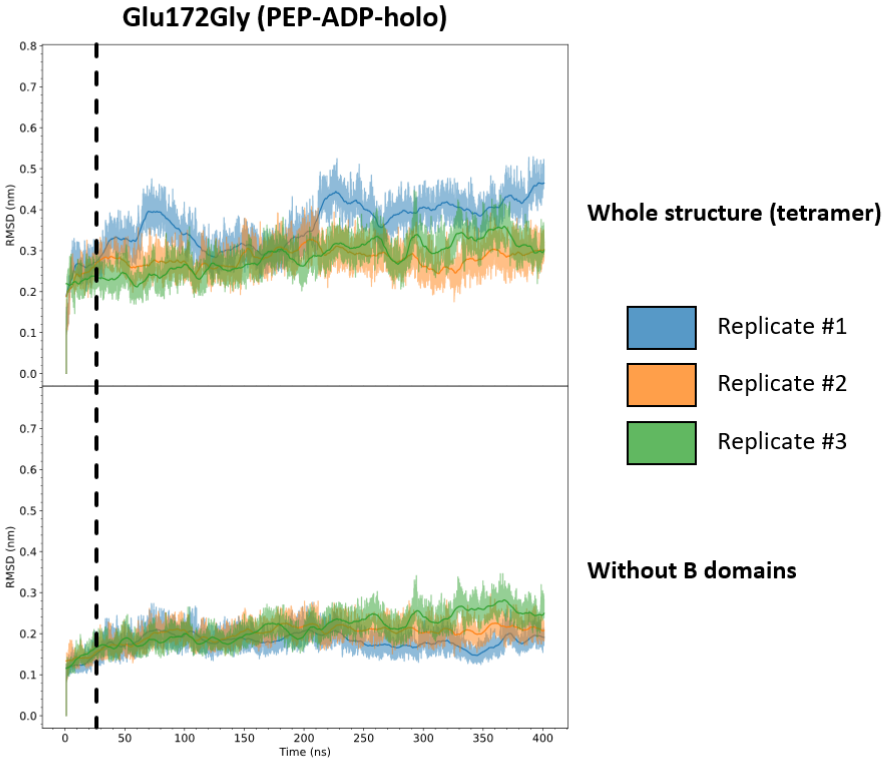


Figure 4.70. Time-series RMSD of the trajectories of the variant Glu172Gly in the PEP-ADP-holo condition. The analysis was applied to the protein backbone atoms. The darker line plotted alongside each time series represents the two-sided moving average of the data, encompassing values up to 5 ns on either side of each point. The top and bottom panels correspond to RMSD values of the tetramers including and excluding the B domains, respectively. A vertical black dashed line indicates the cutoff of 25 ns that marks the end of the structural relaxation phase.

The flexibility profiles obtained via RMSF enabled inspection of the mobile capabilities of the B domains from a different angle. Figure 6.71 shows the RMSF data of two of the examples mentioned above. For instance, in the trajectory replicate #1 of variant Gly332Ser simulated in the apo condition, the B domain of chain B was the specific region that accounted for most of the structural divergence exhibited in the RMSD plot (Figure 4.69, top-left panel). In respect of variant Glu172Gly simulated in the PEP-ADP-condition (replicate #1), although B-domain fluctuations are overall lower due to the presence of PEP and the rest of ligands of the active site, the B domain of chain A reached higher values (around 0.5) than the average value of the WT system for that region (around 0.2; Figure 4.9 of section 4.1.2.2). This particular mobility of this B domain suggests that it may have fluctuated with independence of the rest of the structure, in contrast to the most frequent behavior in the presence of PEP. On another note, besides the higher fluctuation of the B domains, the RMSF profiles of the mutant systems also display the peaks of the local fluctuations of the loop fragments that extend outwards from the main fold of the structure.

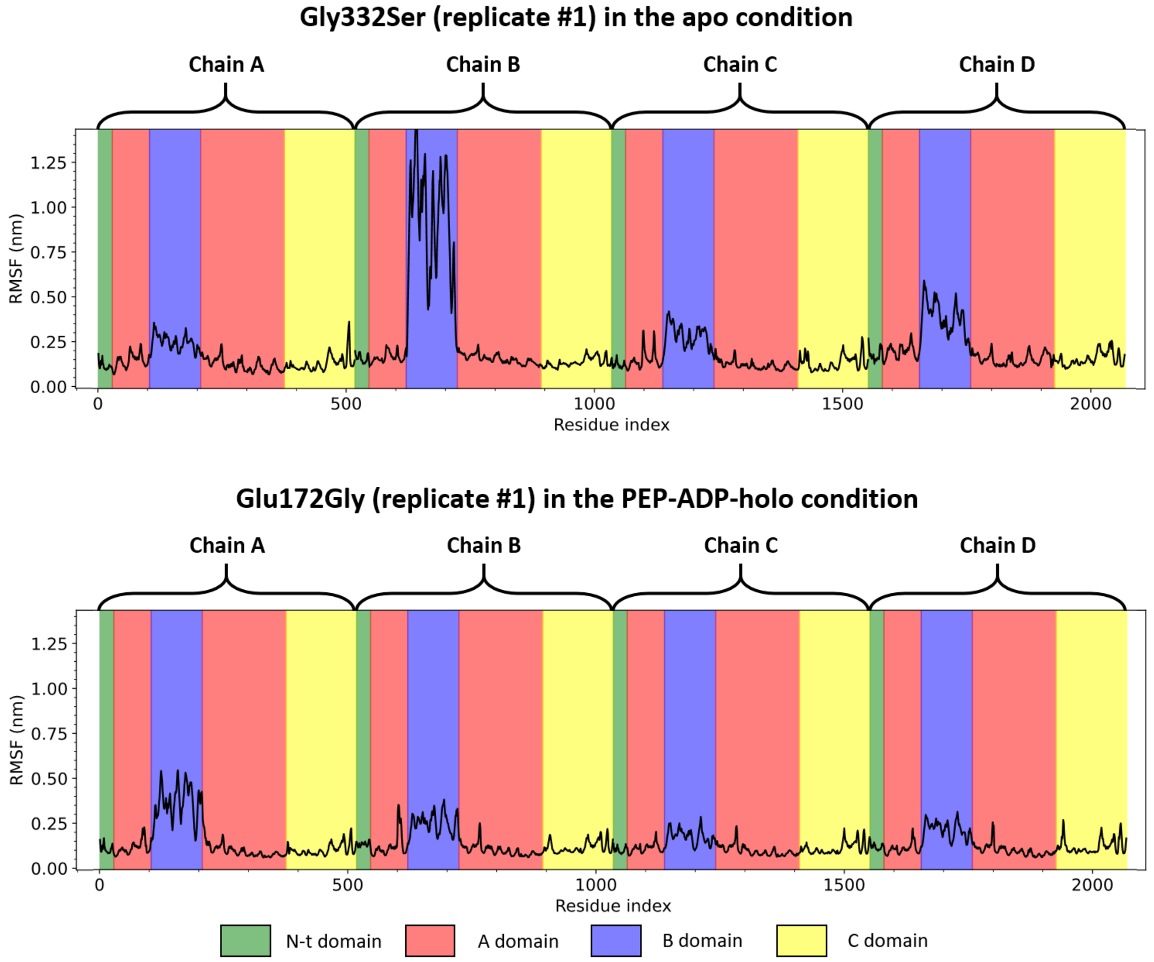


Figure 4.71. RMSF per residue of the trajectories of two missense variants. Data corresponds to replicate #1 of Gly332Ser in the apo condition (top) and replicate #1 of Glu172Gly in the PEP-ADP-holo condition (bottom). The regions corresponding to each subunit of the tetramer are indicated with labels.

4.2.3 Comparative analysis with CEDA

Finally, the trajectories of the mutant systems were analyzed with the framework of CEDA, with the aim of detecting signs of dynamical alteration with respect to the WT behavior. The impact of each mutation was evaluated in terms of the differences in the conformational profiles that had been studied in the previous sections for the WT systems. Thus, several comparative analyses were conducted, involving: i) each of the three examined regions of the enzyme, namely, the A and B domains, the adjacent A domains (A-A′ interface), and the adjacent AC cores (C-C′ interface); and ii) the different apo and holo simulation conditions.

In all the assessments, the WT apo condition served as the reference condition, thus providing the reference set of CPCs for each of the regions under analysis. Accordingly, the trajectories of the simulated variants were subjected to structural superposition and data centering around the reference (WT apo) global average structure of the corresponding regions. Then, the processed trajectory data was projected onto the corresponding set of CPCs. Subsequently, the density distributions of projection data were generated.

On the basis of this procedure, common for all trajectories, the subsequent comparisons between density distributions differed in the nature of the alternative conditions being compared. Figure 4.72 shows a schematic view of the scenario. The comparative analyses between the WT and mutant trajectories in the apo condition involved the density distributions of the reference condition (by definition; WT apo) and a target condition (mutant apo). Conversely, in the comparative analyses between the WT and mutant trajectories in holo conditions, the reference condition (WT apo) provided the common set of CPCs to then compare the density distribution of two target conditions (WT and mutant holo) between each other.

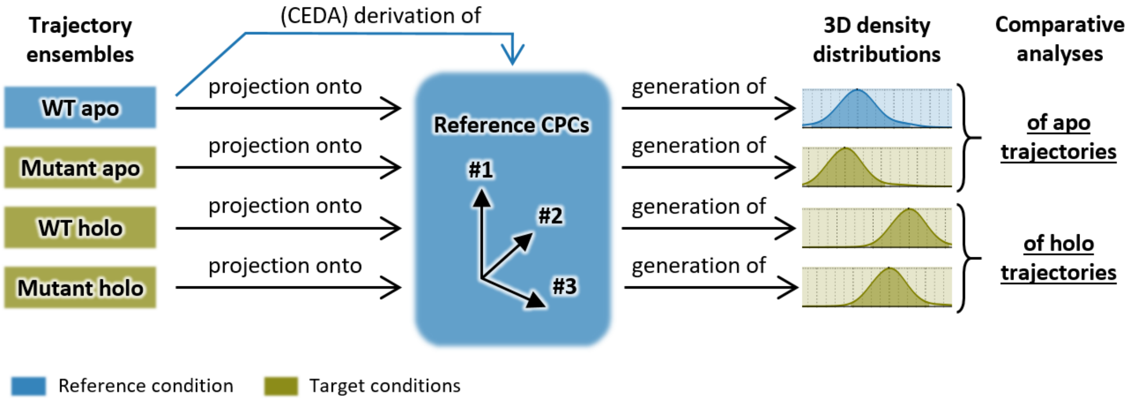


Figure 4.72. Schematic diagram of the process to compare the WT and mutant trajectories with the framework of CEDA.

Given the high number of WT *vs*. mutant comparisons, the assessments were principally made in terms of the set of quantitative indicators proposed in this work (introduced in section 4.1.3.1.3) to express the (dis)similarity between the CEDA-derived density distributions, namely, the overlap, the coverage, and the Bhattacharyya coefficient (BC). These metrics were systematically calculated for each compared WT-mutant pair of trajectory ensembles, providing a score for the similarity between their respective three-dimensional density distributions along CPCs #1, #2, and #3 of the corresponding region under analysis. Importantly, only the intervals of highest 95% density of each distribution were considered to disregard outlier regions before computing the metrics.

Thus, the main results from this section are presented in the form of rankings, from the variant that exhibits the greatest differential behavior to the one that exhibits the least. These sets of values also provide a global perspective to evaluate how effective the probed collective motions were in revealing differential behavior in the collection of mutant simulations as a whole. In addition, for some illustrative instances, images of the two-dimensional density distributions will be included to facilitate a more complete interpretation of the types of dynamical alterations that have been detected. It is important to highlight that the simulations of the WT systems comprised 5 replicates, whereas the number of simulation replicates per variant and condition was reduced to 3 to allow for a more feasible computational cost (Table 3.4 from the Methods chapter). Consequently, the comparisons conducted in this section were performed with a greater amount of trajectory data in the reference condition than in the target conditions. Thus, although the major differences in sampling are expected to be caused mainly by the impact of the mutations, the unequal amount of data might be a secondary factor. Nevertheless, dissimilarities due to the latter reason are expected to manifest mainly as deviations in the span of the peripheral regions of the density distributions, which account for low weight in the calculation of the similarity metrics.

4.2.3.1 Comparison of apo trajectories

The results related to the comparison between the WT and mutant trajectories in the apo condition are presented below, for each of the three examined regions of the enzyme. The distributions of values for each metric were examined by means of box plots, also differentiating between the variants from the two subsets, *i.e.*, pathogenic or potentially neutral. The plots also include (dashed horizontal lines) the values that were obtained in the analogous analyses involving the two batches of simulations of the WT apo condition. Such values afford an additional reference to evaluate the differential behavior of the PKR variants with respect to the WT behavior. That is, for variants bearing similarity values below these thresholds, the lower the values, the more significant their differential behavior with respect to the intrinsic variability of the WT apo condition.

In the case of the comparative analysis of the A and B domains, as seen from the box plots (Figure 4.73), the mutant trajectories exhibit values of the overlap metric consistently lower than the corresponding threshold (99.29%; *i.e.*, the two trajectory ensembles of the WT apo condition were characterized by a strongly conserved overlap). In contrast, the ranges of values for the coverage and BC metrics are more centered around their threshold values (83% and 86%, respectively). This implies that the overlap metric is the most insightful indicator in this comparative analysis as it allows for highlighting a higher number of variants with potentially strong dynamical alterations, divergent from the WT behavior. Accordingly, variants were sorted from the lowest to the greatest overlap values to generate the ranking (Figure 4.74). In general, the distributions of values do not reveal differences between the pathogenic and potentially neutral subsets.

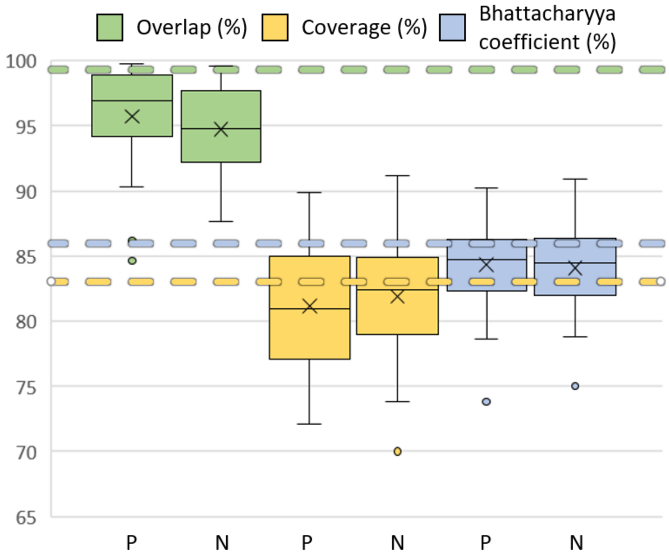


Figure 4.73. Comparative analysis of the WT *vs*. mutant trajectories (A and B domains) in the apo condition: distributions of the quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Data is presented in the form of box plots, distinguishing between the pathogenic (P) and potentially neutral (N) subsets of variants. The BC (values bounded between 0 and 1) is expressed in percentage values to enable a balanced interpretation of the three metrics. Dashed horizontal lines correspond to additional reference values as determined from the comparative analysis of the two trajectory ensembles of the WT apo condition.

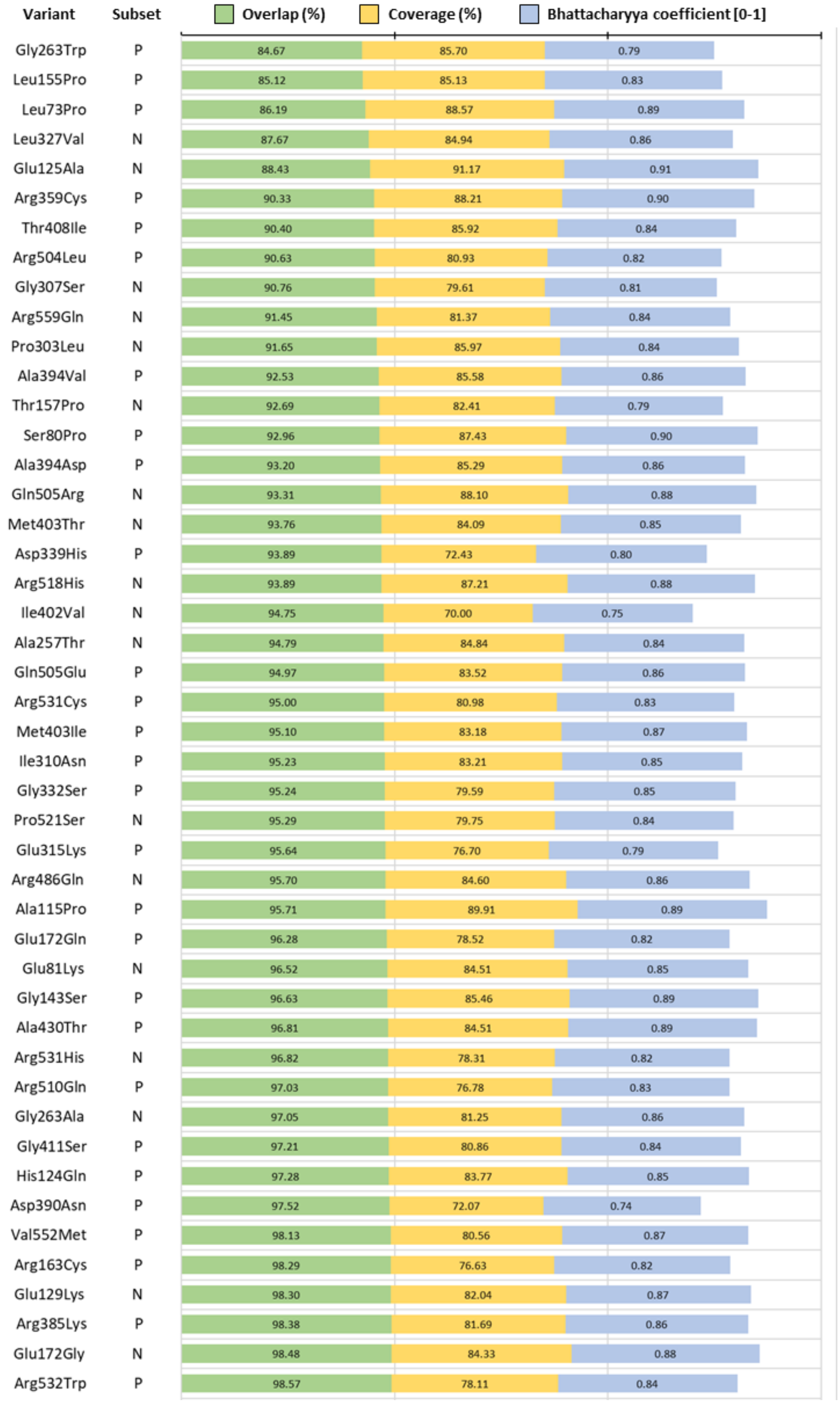
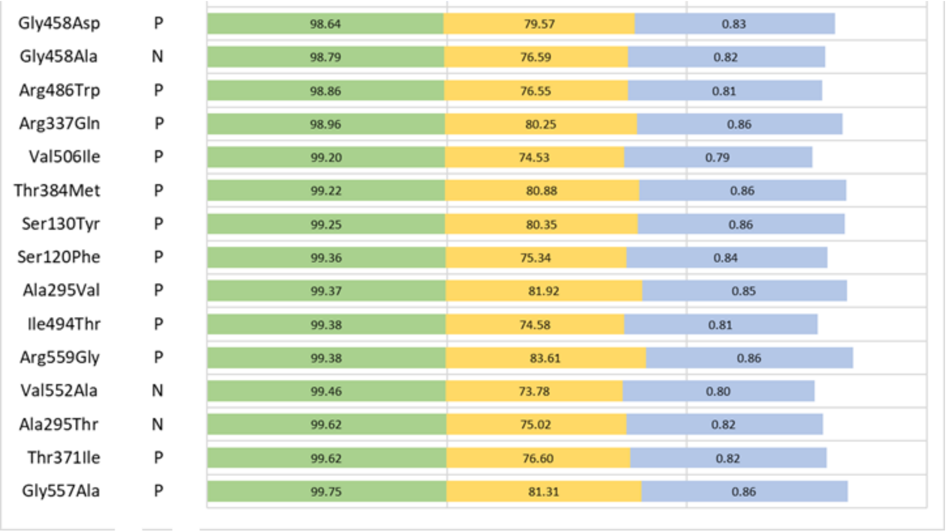


Figure 4.74. Comparative analysis of the WT *vs*. mutant trajectories (A and B domains) in the apo condition: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Next to the variant names, the labels "P" or "N" are included to indicate whether they belong to the pathogenic or the potentially neutral subsets, respectively. Data is sorted in ascending order according to the overlap metric. The bar sizes corresponding to the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

Figure 4.74 (Continued)



No specific correlation between the overlap and the other two metrics is suggested on the basis of the ranking. Figure 4.75 illustrates some examples of this comparative analysis by showing the density distributions of CPCs #1 *vs*. #2. For instance, Gly263Trp, a variant from the pathogenic subset, has the lowest overlap score (*i.e.*, it is the first in the ranking). This non-conservative amino-acid replacement affects a position at the hinge between the A and B domains, which likely influences the dynamical activity of the B domain. The conformational profile (Figure 4.75a) indicates that this variant sampled abnormal conformations outside the WT region, leading to a lower overlap than the reference value. In contrast, Gly263Ala, a variant from the potentially neutral subset, affects the same position with a more conservative amino-acid replacement and has a high overlap score (37th in the ranking). Its conformational profile (Figure 4.75b) displays no significant abnormal conformations (except for a small low-density region). Interestingly, this variant scarcely sampled the closed (active) conformation. However, the replicability test of the WT apo condition revealed specific intrinsic variability in sampling at that conformational region (Figure 4.32).

Leu73Pro is an example of a variant that affects a position distant from the region of analysis, specifically, the N-terminal domain. Results suggest that the amino-acid replacement induced structural alterations that propagated to the A and B domains and influenced the conformational sampling of the latter. It ranked third, with its conformational profile (Figure 4.75c) showing abnormal conformations that protrude from the WT region. However, this variant displays one of the highest BC scores because the overlapping region of the density distributions preserved the relative proportions between the major populations.

Regarding the coverage metric, in general, the variants with scores particularly lower than the threshold value of intrinsic WT variability sampled only at the characteristic regions of major density, with low conformational heterogeneity. Some of these variants only sampled the open forms of the B domain, such as Asp339His (pathogenic subset; 18th in the ranking; Figure 4.75d) or Ser120Phe (pathogenic subset; 54th in the ranking; Figure 4.75e). Importantly, these variants affect key residues involved in the cofactors/PEP binding and catalytic efficiency. On the other hand, other variants achieved sampling of both the open and closed forms of the B domain, such as Ile402Val (potentially neutral subset; 20th in the ranking; Figure 4.75f).

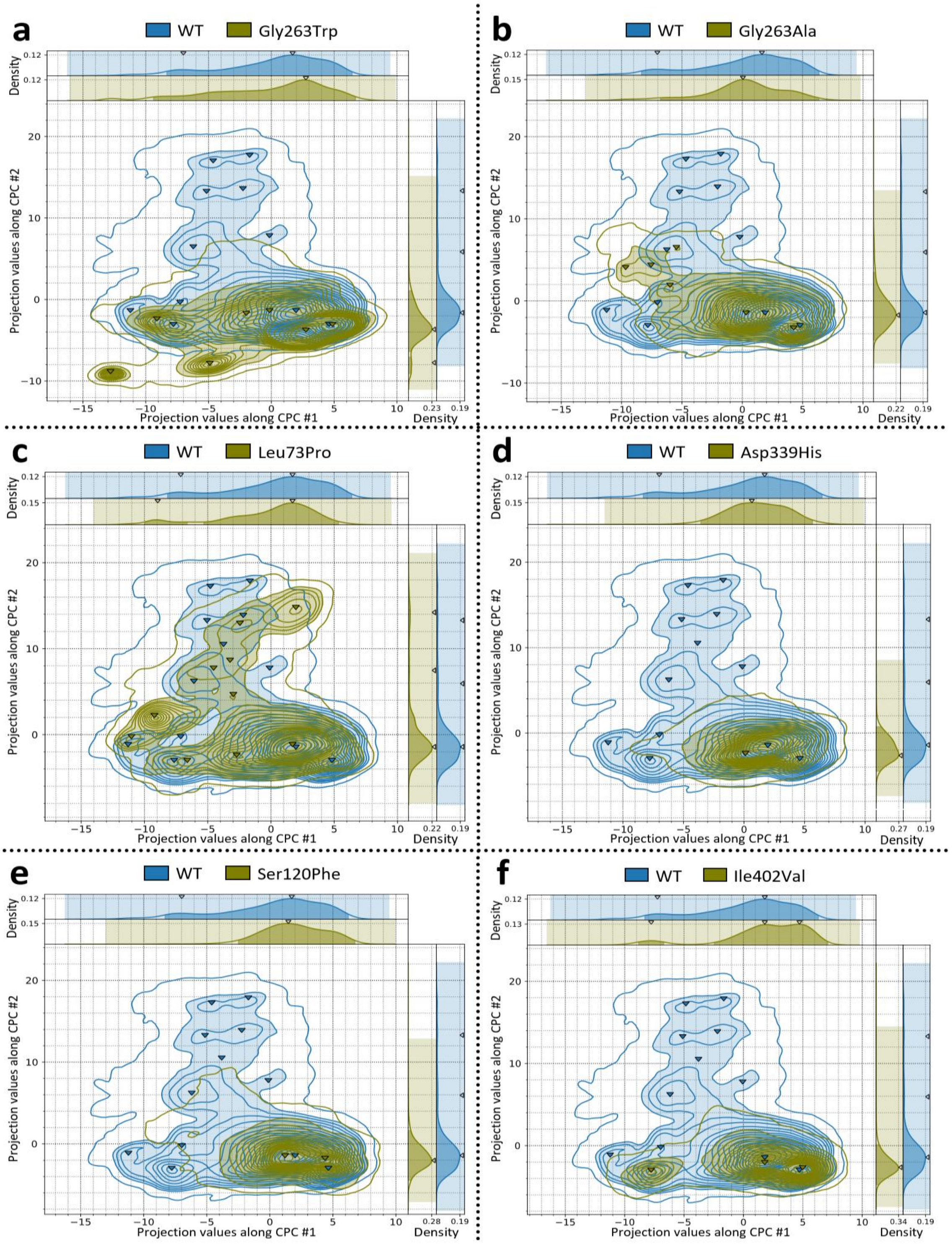


Figure 4.75. Comparative analysis of the WT *vs*. mutant trajectories (A and B domains) in the apo condition: conformational profiles (2 CPCs). This figure only shows the examples of six variants. (**a**) WT *vs*. Gly263Trp. (**b**) WT *vs*. Gly263Ala. (**c**) WT *vs*. Leu73Pro. (**d**) WT *vs*. Asp339His. (**e**) WT *vs*. Ser120Phe. (**f**) WT *vs*. Ile402Val. The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

In the case of the comparative analysis of the adjacent A domains (A-A' interface), the chosen indicator to rank the variants was the coverage metric, based on the corresponding box plots (Figure 4.76). The overlap metric was discarded, as the majority of mutant trajectories exhibit values above the threshold (87.41%) and, thus, hardly distinguishable from the intrinsic variability of the WT behavior. On the other hand, the coverage and the BC metrics allow for a clearer distinction of the most altered variants because there is a higher number of variants bearing values below the corresponding thresholds (85.38% and 88%, respectively). Between the two indicators, the former was selected because it offers a direct interpretation of the reported differential behavior, namely, the extent of reference conformational space sampled by the target condition. The resulting ranking is shown in Figure 4.77. In general, the distributions of values do not reveal differences between the pathogenic and potentially neutral subsets.

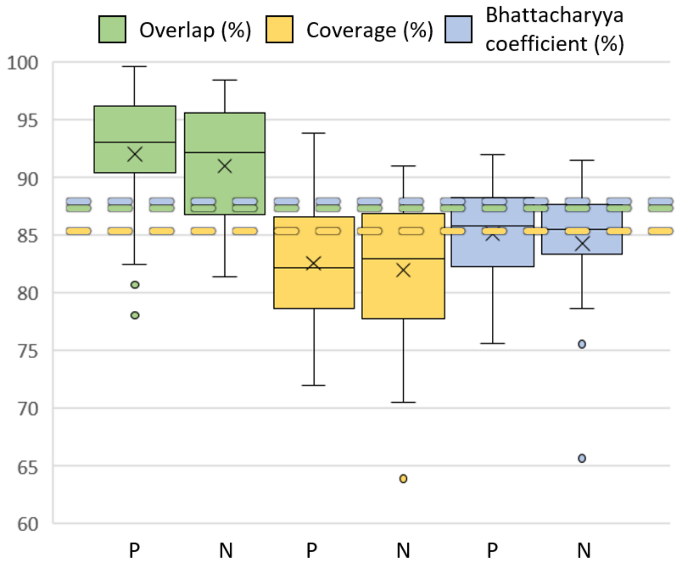


Figure 4.76. Comparative analysis of the WT *vs*. mutant trajectories (A-A' pairs of domains) in the apo condition: distributions of the quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Data is presented in the form of box plots, distinguishing between the pathogenic (P) and potentially neutral (N) subsets of variants. The BC (values bounded between 0 and 1) is expressed in percentage values to enable a balanced interpretation of the three metrics. Dashed horizontal lines correspond to additional reference values as determined from the comparative analysis of the two trajectory ensembles of the WT apo condition.

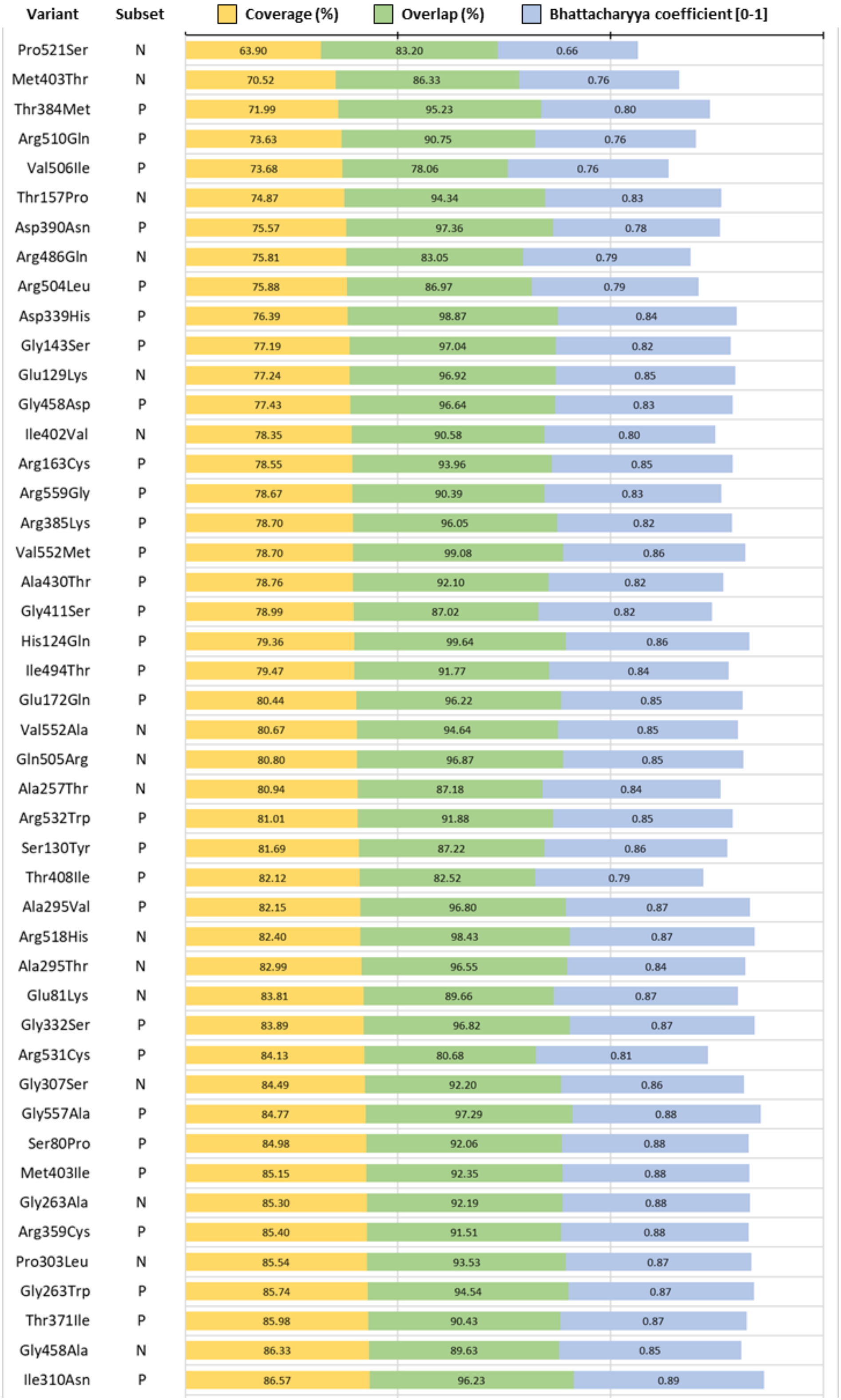


Figure 4.77. Comparative analysis of the WT *vs*. mutant trajectories (A-A′ pairs of domains) in the apo condition: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Next to the variant names, the labels "P" or "N" are included to indicate whether they belong to the pathogenic or the potentially neutral subsets, respectively. Data is sorted in ascending order according to the coverage metric. The bar sizes corresponding to the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

Figure 4.77 (Continued)

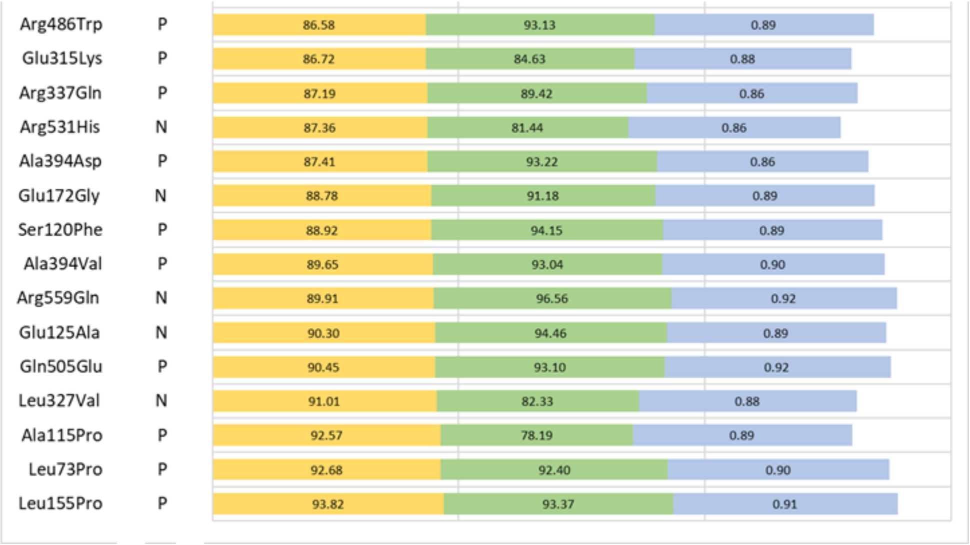


Figure 4.78 illustrates some examples of this comparative analysis by showing the density distributions of CPCs #2 *vs*. #3. The variant Pro521Ser belongs to the potentially neutral subset and, however, is the first of the ranking with the lowest coverage score. Its conformational profile (Figure 4.78a) exhibits a strong shift of the density distribution along CPC #2. Interestingly, this region of the conformational spectrum is close to the characteristic region of the PEP-bound trajectories (Figure 4.50), only differing in the fact that the latter is also characterized by higher values of CPC #3. The variant Val506Ile (pathogenic subset; 5th in the ranking) stands out as one of the variants with lowest scores on both the coverage and overlap metrics. Its conformational profile (Figure 4.78b) features a shift along CPC #2, albeit less pronounced than that of Pro521Ser, and bimodality along CPC #3. It shares a central region of major density with the WT system, however most of its conformational population lies at negative values of CPC #3, far from the characteristic region of the active conformation.

The variant Thr384Met (pathogenic subset; 3rd in the ranking) affects a position in helix Aα7, at the A-A′ interface. It has a low coverage score and, simultaneously, a high overlap score. It is characterized by a more restricted region of conformational sampling, almost entirely inside the WT region and with less dispersion of values (Figure 4.78c). The variants Arg385Lys and Asp390Asn also affect positions nearby in the A-A′ interface and display both similar scores and conformational profiles (not shown). Finally, the variant Ala115Pro is illustrative of the limitations in the information provided by just a single metric. This variant ranked high because it has one of the highest coverage scores (pathogenic subset; 59th in the ranking). However, simultaneously, it has one of the lowest overlap scores. Its conformational profile (Figure 4.78d) displays high dispersion of values in both CPCs #2 and #3, covering a wide extension of the WT region. Interestingly, the region of major density is closer to the PEP-bound trajectories than in the last two examples.

In the case of the comparative analysis of the adjacent AC cores (C-C′ interface), the whole set of mutant trajectories exhibit coverage values lower than the corresponding threshold (93.90%; Figure 4.79). Interestingly, in terms of the overlap metric, despite the fact that the majority of variants exhibit values above the threshold (90.24%), a potential difference between the pathogenic and the potentially neutral subsets is suggested, with the former containing more variants below the threshold. Thus, for this comparative analysis, the variants were ranked according to the sum of their overlap and coverage values (Figure 4.80).

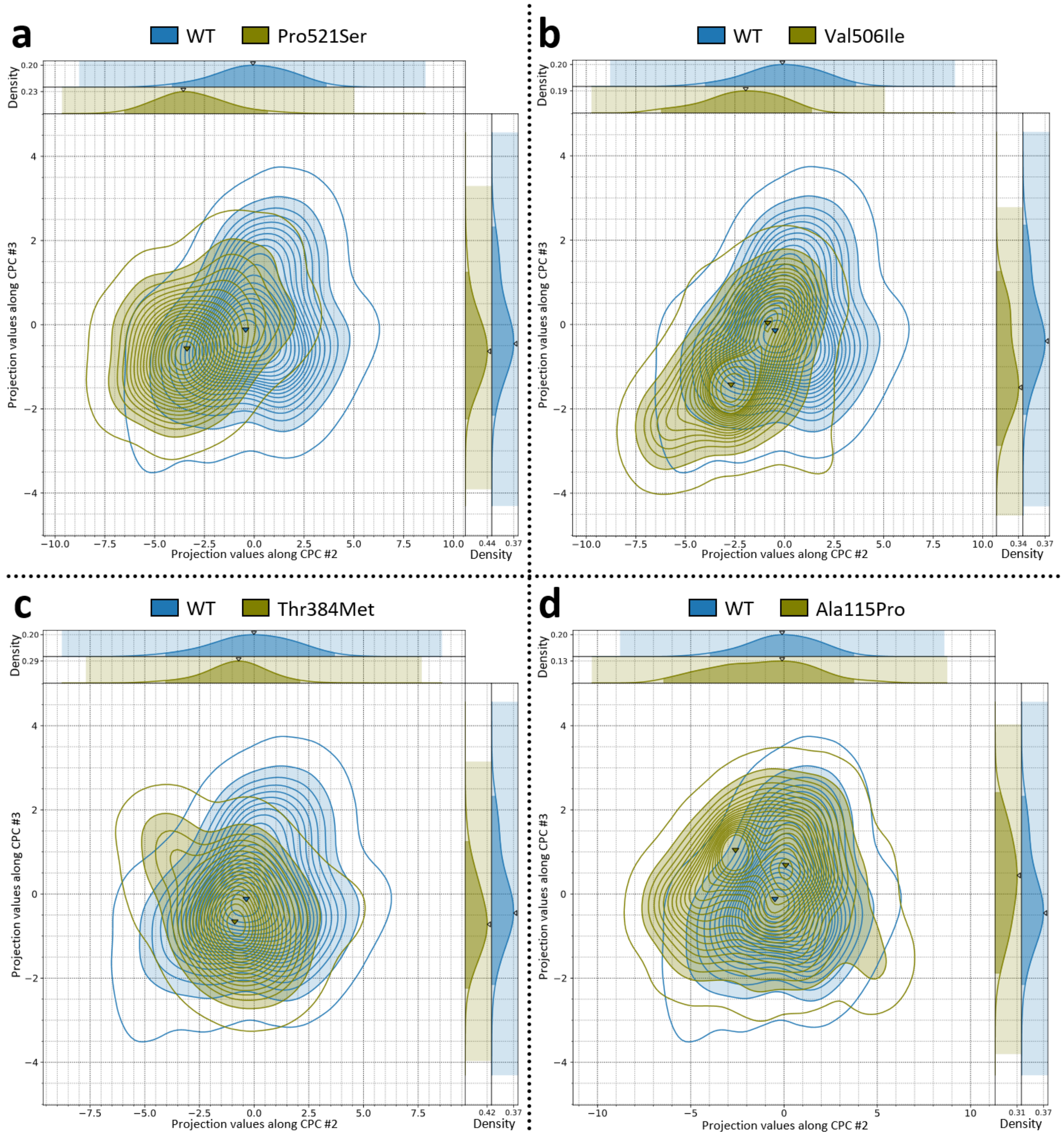


Figure 4.78. Comparative analysis of the WT *vs*. mutant trajectories (A-A′ pairs of domains) in the apo condition: conformational profiles (2 CPCs). This figure only shows the examples of four variants. (**a**) WT *vs*. Pro521Ser. (**b**) WT *vs*. Val506Ile. (**c**) WT *vs*. Thr384Met. (**d**) WT *vs*. Ala115Pro. The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #2 *vs*. #3 from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

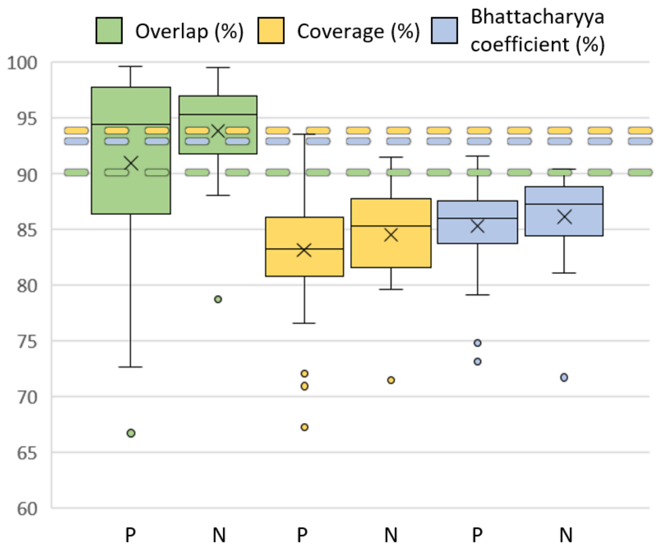


Figure 4.79. Comparative analysis of the WT *vs*. mutant trajectories (AC-C'A' pairs of cores) in the apo condition: distributions of the quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Data is presented in the form of box plots, distinguishing between the pathogenic (P) and potentially neutral (N) subsets of variants. The BC (values bounded between 0 and 1) is expressed in percentage values to enable a balanced interpretation of the three metrics. Dashed horizontal lines correspond to additional reference values as determined from the comparative analysis of the two trajectory ensembles of the WT apo condition.

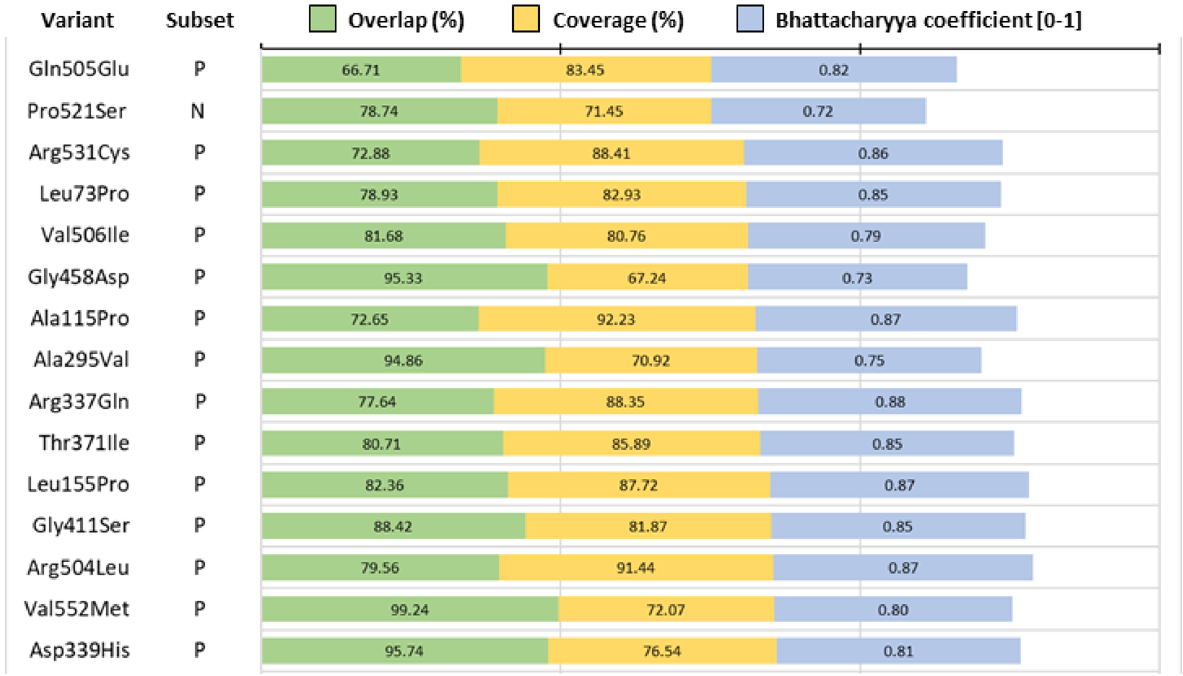


Figure 4.80. Comparative analysis of the WT *vs*. mutant trajectories (AC-C'A' pairs of cores) in the apo condition: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Next to the variant names, the labels "P" or "N" are included to indicate whether they belong to the pathogenic or the potentially neutral subsets, respectively. Data is sorted in ascending order according to the sum of the overlap and coverage metrics. The bar sizes corresponding to the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

Figure 4.80 (Continued)

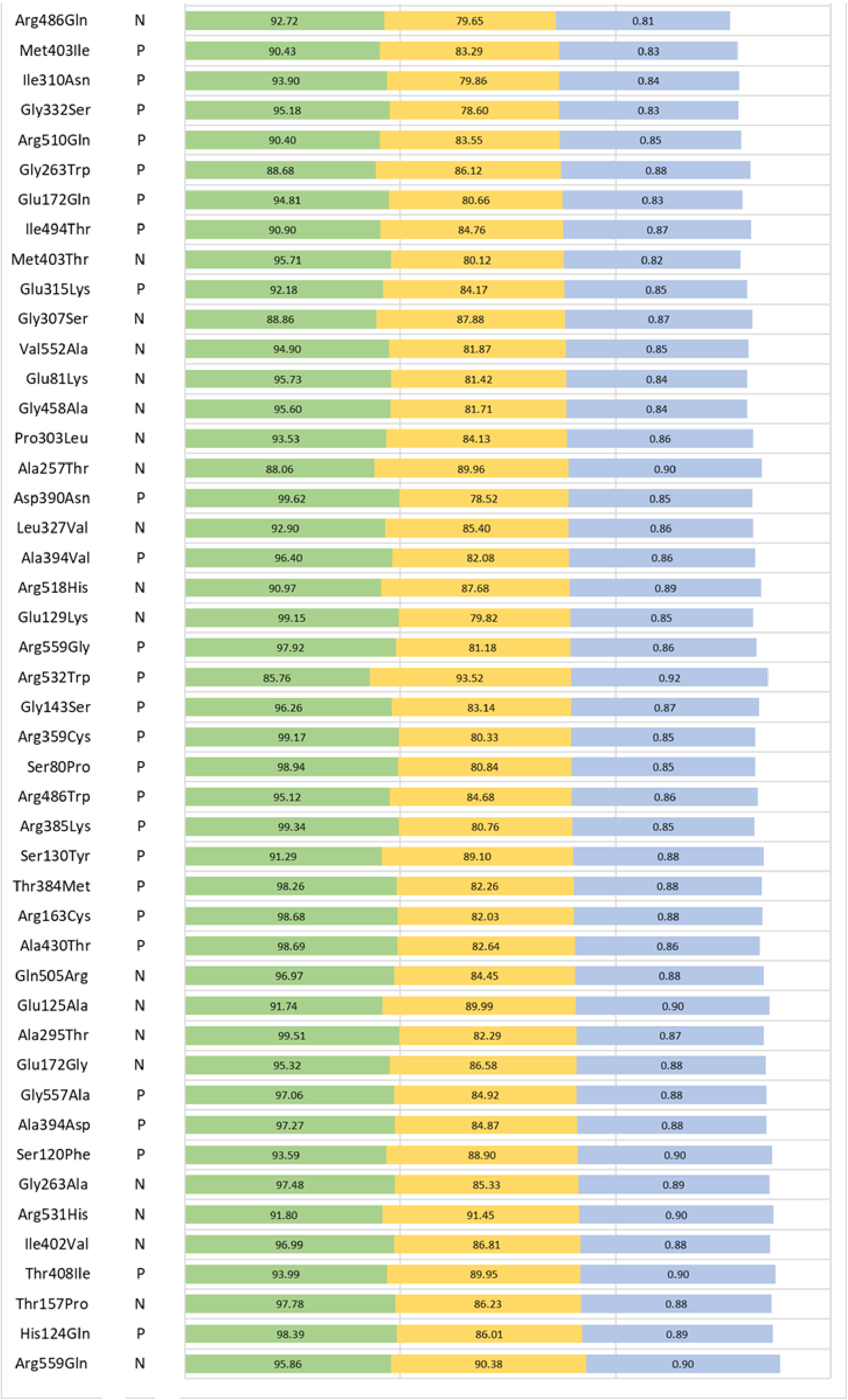


Figure 4.81 illustrates some examples of this comparative analysis by showing the density distributions of CPCs #2 *vs*. #3. The variant Gln505Glu (pathogenic subset; 1st in the ranking) has the lowest score. Its conformational profile (Figure 4.81a) is particularly shifted along CPC #3 towards the opposite direction of the PEP-bound trajectories. It is also characterized by higher dispersion along CPC #1 (not shown). This variant affects the Cα4 helix, at the A-C interface. This helix, specifically, constitutes a main pivot point of the rigid-body collective motions that were revealed in the CEDA of the WT apo condition. The amino-acid replacement introduces a negative charge. In contrast, variant Gln505Arg (potentially neutral subset; 48th in the ranking) affects the same position but with a different amino-acid replacement that introduces a positive charge. Despite the fact that both variants introduce electrostatic changes in the same position, Gln505Arg exhibited regular dynamical activity, as it displays no significant differential behavior (Figure 4.81b).

The variant Val552Met (pathogenic subset; 14th in the ranking) has a very high overlap score, but also one of the lowest coverage scores. This amino-acid replacement affects a position in the Cβ4 strand, at the C-C′ interface. Its conformational profile (Figure 4.81c) is characterized by being entirely confined within the WT region with substantially less dispersion. In contrast, the variant Val552Ala (potentially neutral subset; 27th in the ranking) affects the same position and exhibits an overall more equivalent exploration of the WT region of the conformational space (Figure 4.81d).

Several variants that were already highlighted in the previous comparative analysis due to their strong signs of differential behavior also exhibit low similarity scores in the present analysis. For instance, Leu73Pro (pathogenic subset; 4th in the ranking) displays dynamical alterations similar to those of Gln505Glu (not shown). Ala115Pro (pathogenic subset; 7th in the ranking) exhibits considerably higher dispersion of density with multiple peaks (not shown), which suggests instability or higher fluctuations. Two particularly interesting cases are Pro521Ser and Val506Ile. The former (potentially neutral subset; 2nd in the ranking) differs from the WT apo condition in that its conformational profile (Figure 4.81e) displays shifts along CPCs #2 and #3. However, this conformational profile is comparable to that of the WT FBP-holo condition (Figure 4.63). The set of dynamical alterations exhibited by this variant in both this and the previous analyses suggest that it might enhance sampling of the active conformation of the enzyme in the absence of ligands. Finally, the variant Val506Ile (pathogenic subset; 5th in the ranking) exhibits a shift along CPC #3 similar to that of Pro521Ser (not shown).

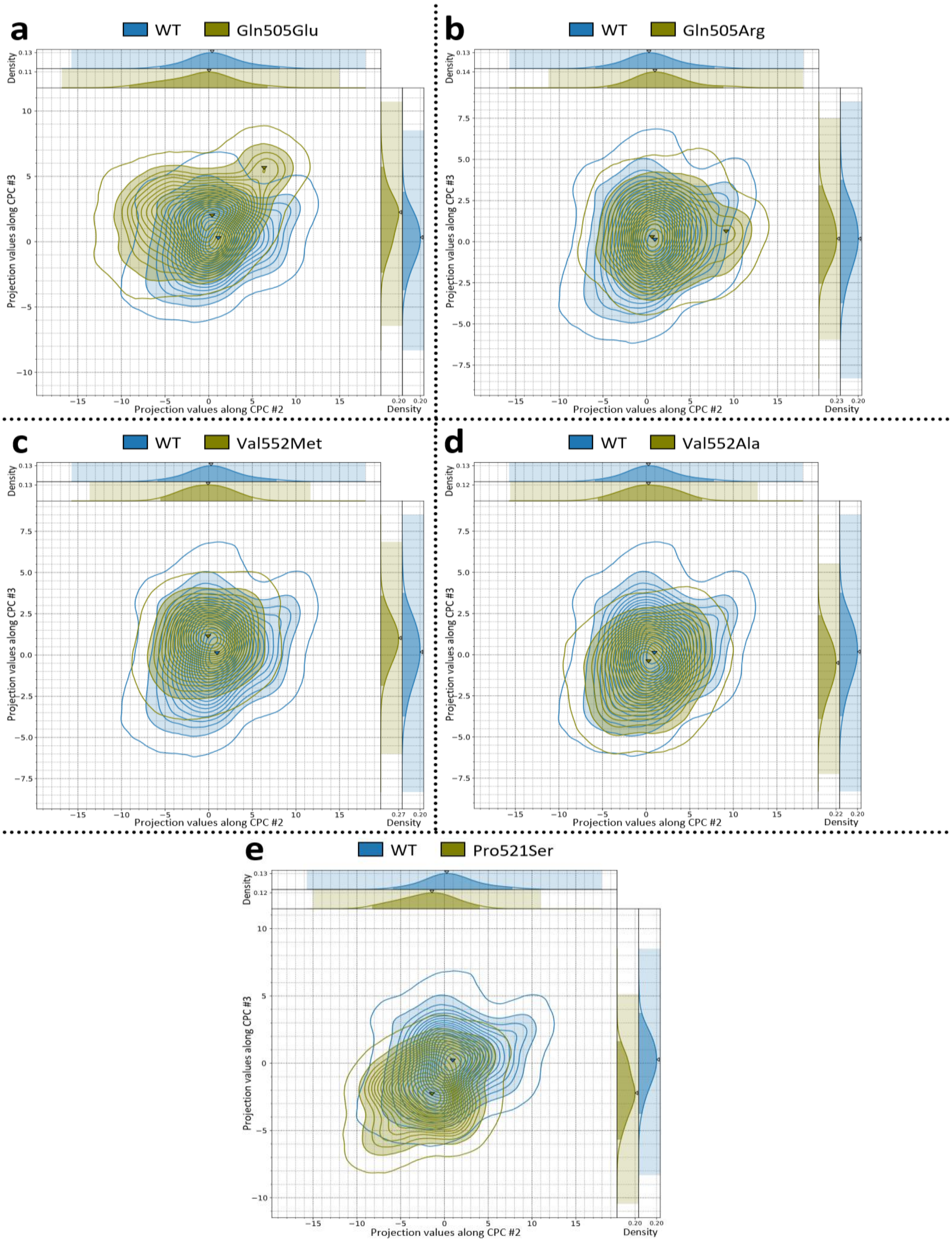


Figure 4.81. Comparative analysis of the WT *vs*. mutant trajectories (AC-C'A' pairs of cores) in the apo condition: conformational profiles (2 CPCs). This figure only shows the examples of five variants. (**a**) WT *vs*. Gln505Glu. (**b**) WT *vs*. Gln505Arg. (**c**) WT *vs*. Val552Met. (**d**) WT *vs*. Val552Ala. (**e**) WT *vs*. Pro521Ser. The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #2 *vs*. #3 from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

4.2.3.2 Comparison of holo trajectories

The results reported in this section relate to the comparative analyses conducted between the WT and mutant trajectories in holo conditions. As described above (Table 4.5), each variant was simulated in one of the following four holo conditions: K-Mg-holo, PEP-holo, PEP-ADP-holo, or FBP-holo. As shown schematically in Figure 4.72, for each holo condition, the corresponding WT *vs*. mutant comparisons were made between the density distributions that result from projecting the trajectory data onto the reference sets of CPCs from the WT apo condition.

In these assessments, the variants were systematically ranked based on the sum of values of the three similarity metrics. This approach was adopted due to the potential diversity of manifestations of differential behavior that may arise from dividing the collection of mutant simulations into four subsets of alternative holo conditions. A single metric may not capture this diversity in a meaningful way. Therefore, considering all three metrics simultaneously ensures more balanced rankings across the entire collection of mutant simulations. Furthermore, unlike the comparison of apo trajectories, there are no additional reference values available in this scenario that would aid in exploring the most relevant metric in each comparative analysis on the basis of the intrinsic variability of sampling of each WT holo condition.

Figures 4.82 and 4.83 show the box plots and the rankings of the results of the comparative analysis of the A and B domains. Figure 4.84 illustrates some examples of this comparative analysis by showing the density distributions of CPCs #1 *vs*. #2. The variant Ser120Phe (pathogenic subset; 1st in the ranking of K-Mg-holo) deprives the cofactor K^+ of one of its coordination ligands and introduces the bulky side chain of phenylalanine which partially blocks the K^+-binding site. The binding of the cofactor was modeled (K-Mg-holo condition) by incorporating the QM-derived parameters of the remaining three amino acids of the coordination complex (Asn118, Asp156, Thr157), thus yielding a hypothetical Ser120Phe K^+-bound model, which was stable in MD. The resulting conformational profile (Figure 4.84a) reveals that, although the sampling of the closed form of the B domain was retained in minor proportions, the conformational equilibrium was shifted towards the open forms, as opposed to the WT behavior in this condition. In addition, this variant did not sample the closed form in the apo condition (Figure 4.75e).

The variant Arg163Cys (pathogenic subset; 1st in the ranking of PEP-ADP-holo) affects a position in the hinge between the A and B domains and manifests dynamical alterations by having sampled more open forms of the B domain than in the WT behavior (Figure 4.84b). In the apo condition, this variant does not manifest significant alterations (not shown). On the other hand, the variant Gly263Trp (pathogenic subset; 2nd in the ranking of PEP-ADP-holo), which is also located in the hinge fragment, displays dynamical alterations both in the apo (Figure 4.75a) and the PEP-ADP-holo (Figure 4.84c) conditions. The variant Gly263Ala (potentially neutral subset; 12th in the ranking of PEP-ADP-holo) affects the same position as the previous variant and does not manifest dynamical alterations neither in the apo (Figure 4.75b) nor the PEP-ADP-holo (Figure 4.84d) conditions.

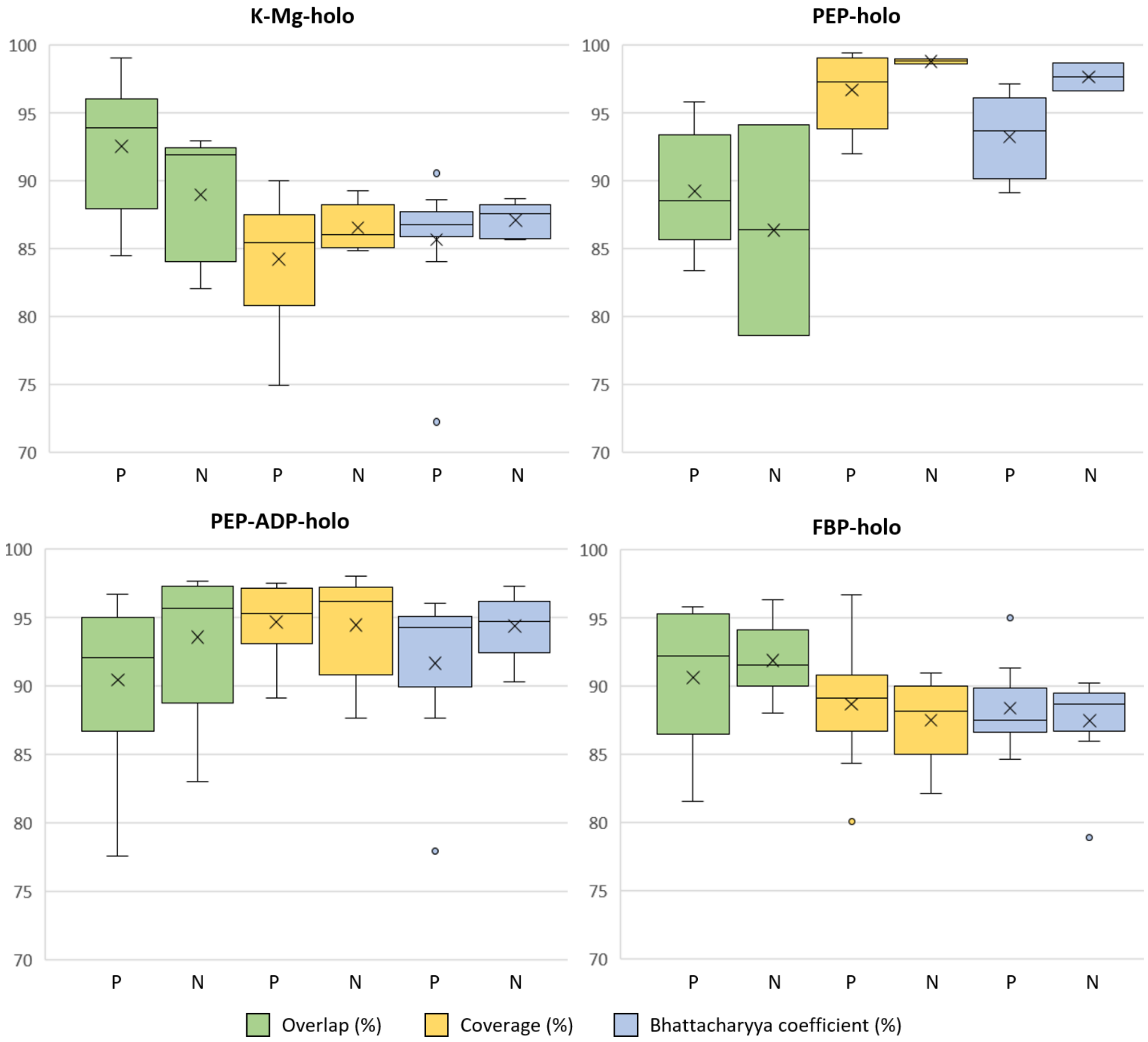


Figure 4.82. Comparative analysis of the WT *vs*. mutant trajectories (A and B domains) in holo conditions: distributions of the quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Data is presented in the form of box plots, distinguishing between the pathogenic (P) and potentially neutral (N) subsets of variants. The BC (values bounded between 0 and 1) is expressed in percentage values to enable a balanced interpretation of the three metrics.

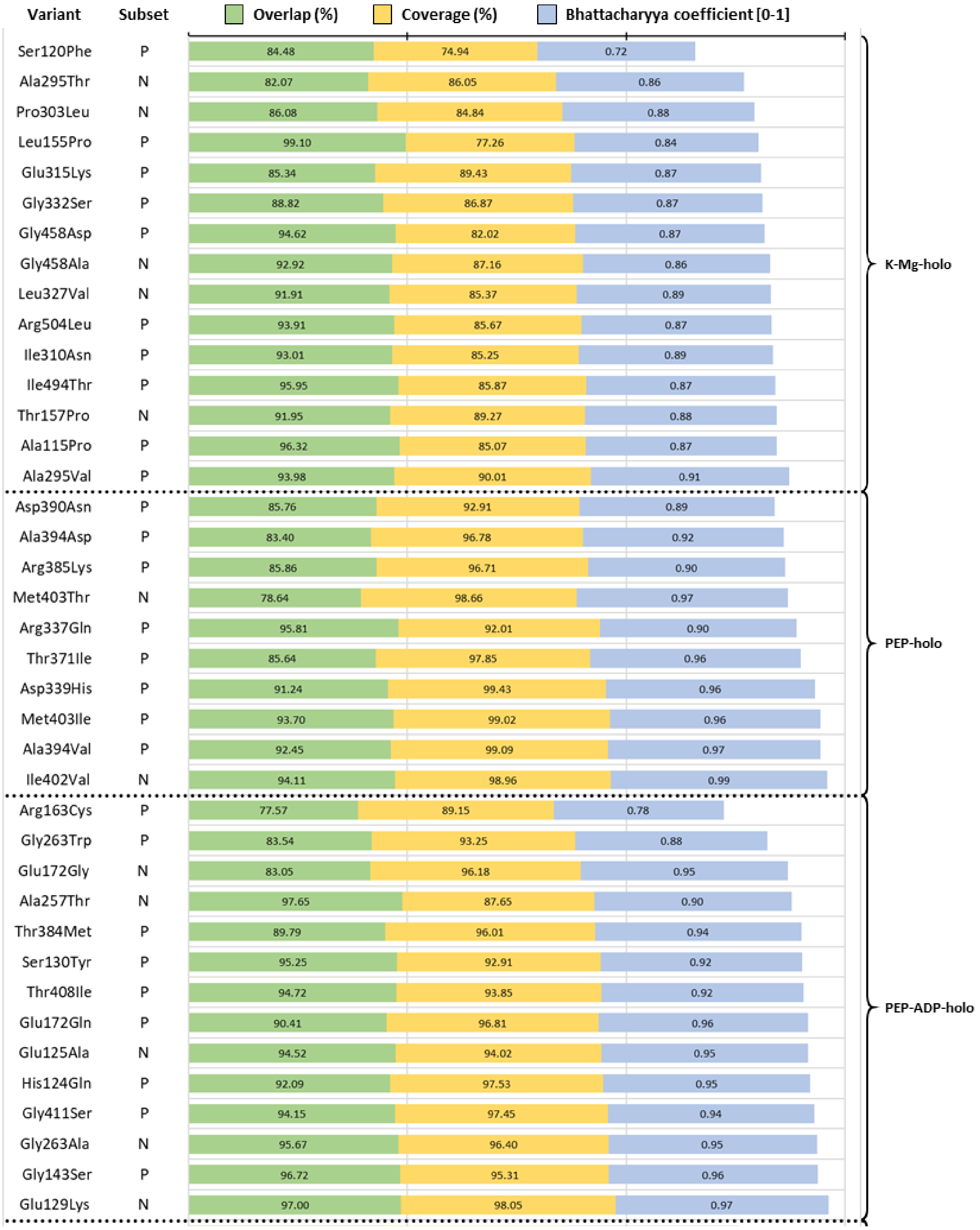
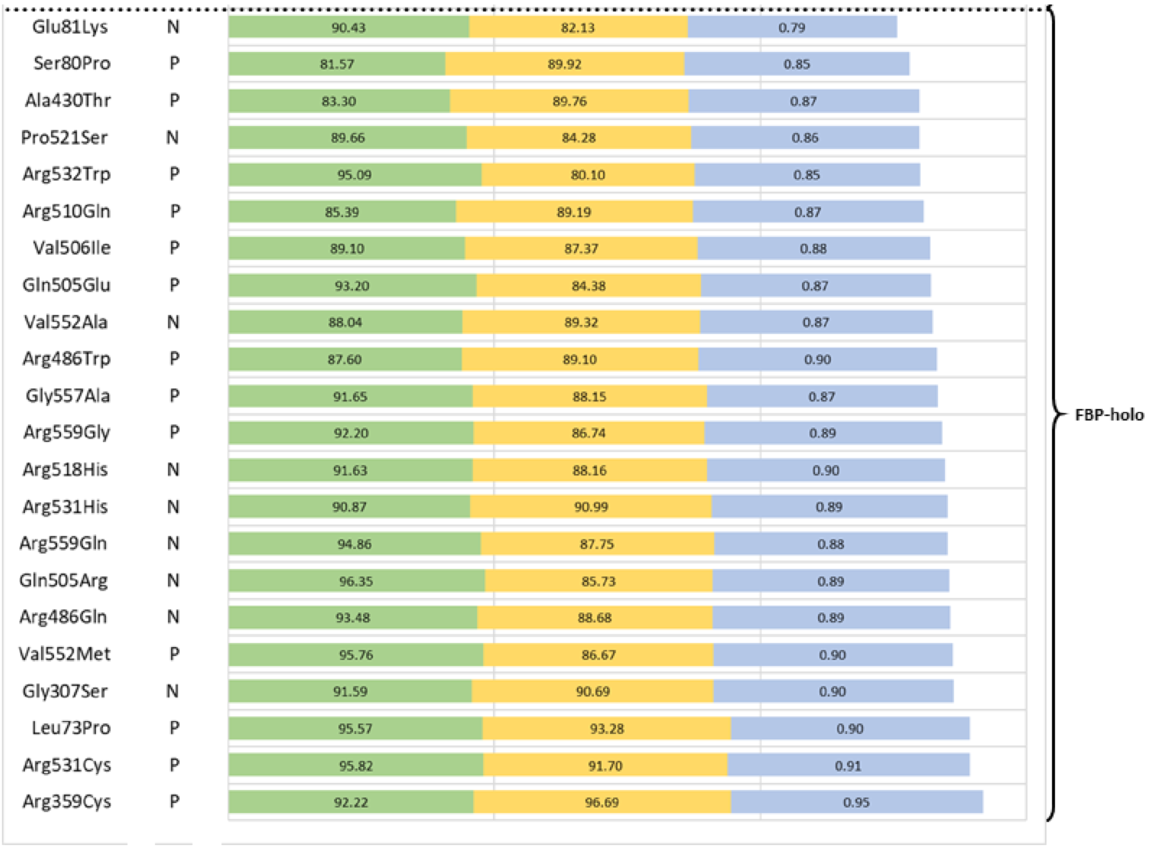


Figure 4.83. Comparative analysis of the WT *vs.* mutant trajectories (A and B domains) in holo conditions: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Next to the variant names, the labels "P" or "N" are included to indicate whether they belong to the pathogenic or the potentially neutral subsets, respectively. For each block of rows corresponding to each holo condition, data is sorted in ascending order according to the sum of the three metrics. The bar sizes corresponding to the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

Figure 4.83 (Continued)



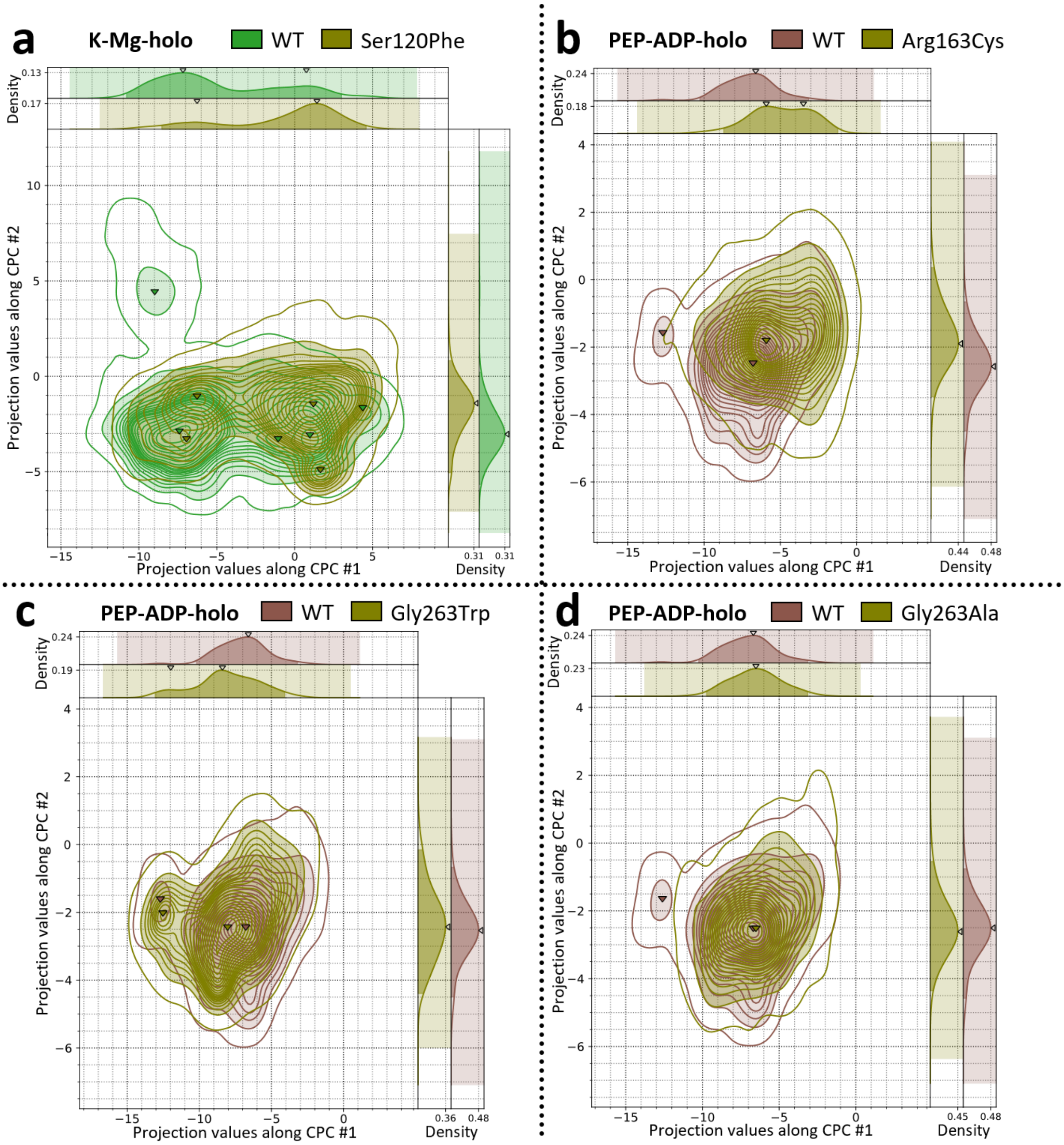


Figure 4.84. Comparative analysis of the WT *vs*. mutant trajectories (A and B domains) in holo conditions: conformational profiles (2 CPCs). This figure only shows the examples of four variants. (**a**) WT *vs*. Ser120Phe in the K-Mg-holo condition. (**b**) WT *vs*. Arg163Cys in the PEP-ADP-holo condition. (**c**) WT *vs*. Gly263Trp in the PEP-ADP-holo condition. (**d**) WT *vs*. Gly263Ala in the PEP-ADP-holo condition. The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #1 *vs*. #2 from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figures 4.85 and 4.86 show the box plots and the rankings of the results of the comparative analysis of the adjacent A domains (A-A' interface).

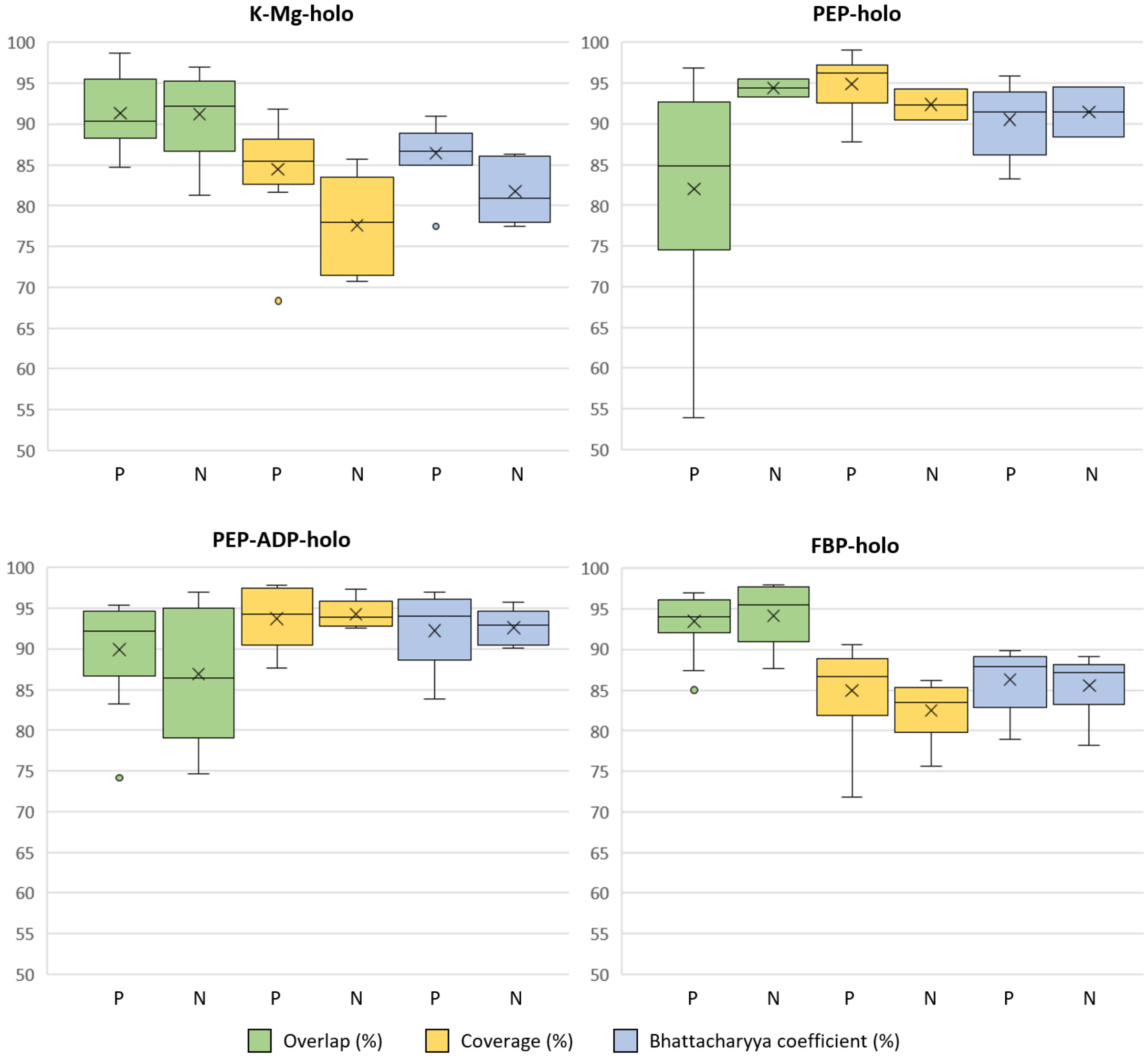


Figure 4.85. Comparative analysis of the WT *vs*. mutant trajectories (A-A' pairs of domains) in holo conditions: distributions of the quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Data is presented in the form of box plots, distinguishing between the pathogenic (P) and potentially neutral (N) subsets of variants. The BC (values bounded between 0 and 1) is expressed in percentage values to enable a balanced interpretation of the three metrics.

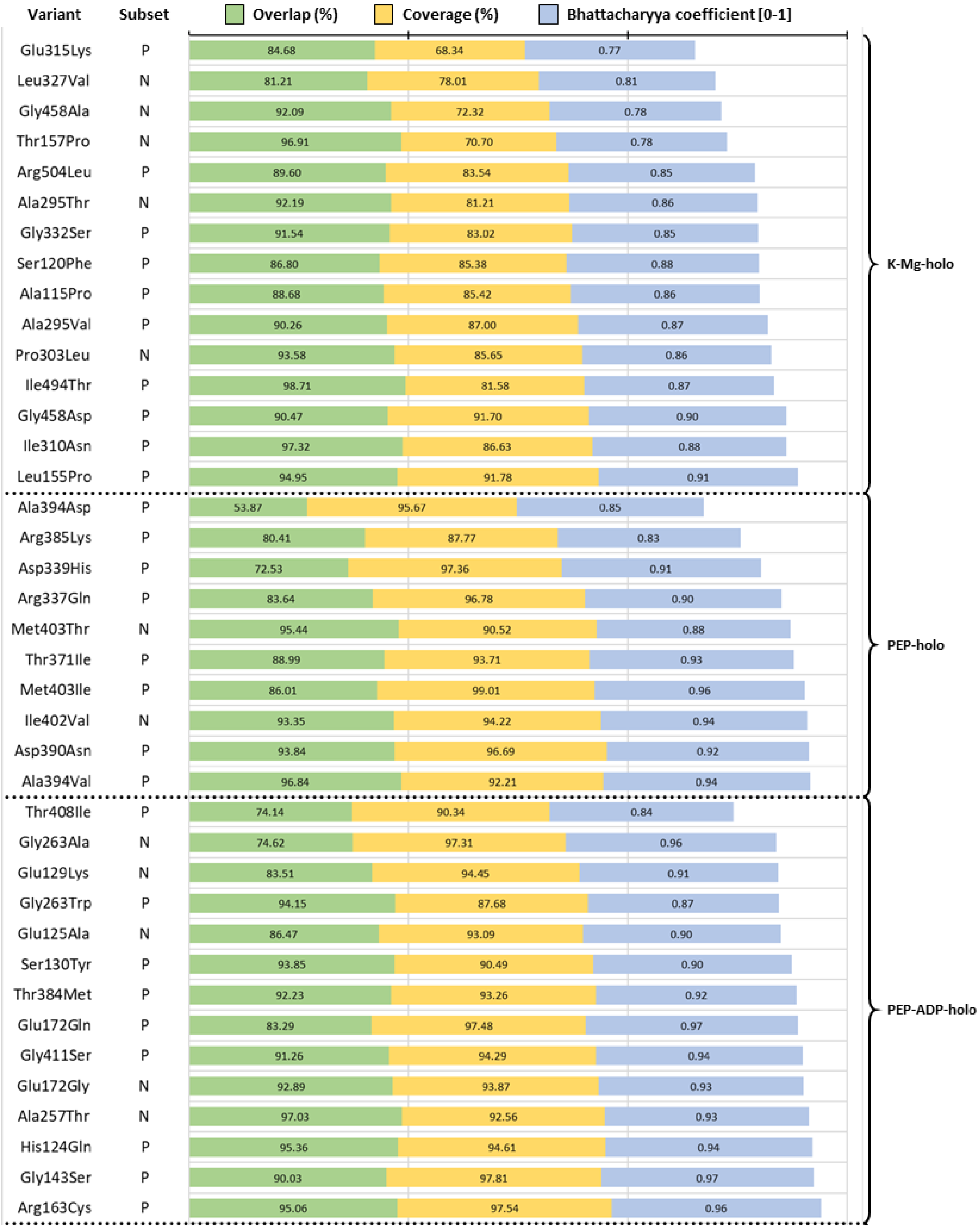


Figure 4.86. Comparative analysis of the WT *vs*. mutant trajectories (A-A' pairs of domains) in holo conditions: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Next to the variant names, the labels "P" or "N" are included to indicate whether they belong to the pathogenic or the potentially neutral subsets, respectively. For each block of rows corresponding to each holo condition, data is sorted in ascending order according to the sum of the three metrics. The bar sizes corresponding to the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

Figure 4.86 (Continued)

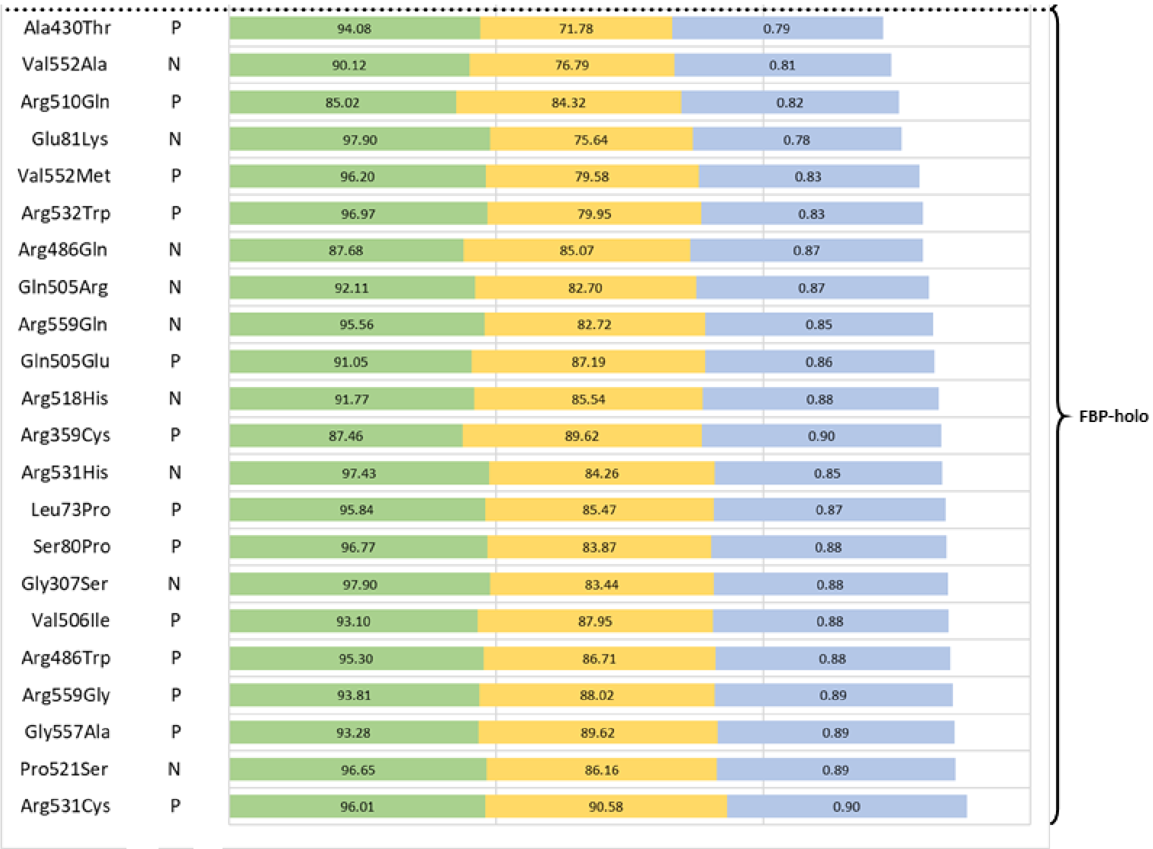


Figure 4.87 illustrates some examples of this comparative analysis by showing the density distributions of CPCs #2 *vs*. #3. The variant Ala394Asp (pathogenic subset; 1st in the ranking of PEP-holo) manifests strong dynamical alterations. Specifically, its conformational profile (Figure 4.87a) shows that approximately half of the conformational population remained in the region characteristic of the apo condition. These results suggest that this variant is unstable in the active conformation. The amino-acid replacement affects a position at helix Aα7, near the A-A' interface but with its side chain buried and facing the β barrel of the A domain. Conversely, the variant Ala394Val (10th in the ranking of PEP-holo), which affects the same position and also belongs to the pathogenic subset, did not introduce alterations (Figure 4.87b) and had the best score of its ranking.

The variant Ala430Thr (pathogenic subset; 1st in the ranking of FBP-holo) affects a position at helix Aα8, preceding the linker fragment with the C domain and at the small interface between the N-terminal and the A domains. Based on its conformational profile (Figure 4.87c), this variant achieved sampling of the characteristic region along CPC #2 in the presence of cofactors, whereas it failed to undergo a change in sampling towards the region of the active conformation along CPC #3 as the WT enzyme does in the presence of FBP. Finally, the variant Arg510Gln (pathogenic subset; 3rd in the ranking of FBP-holo) exhibits a conformational profile (Figure 4.87d) with somewhat opposite features as Ala430Thr.

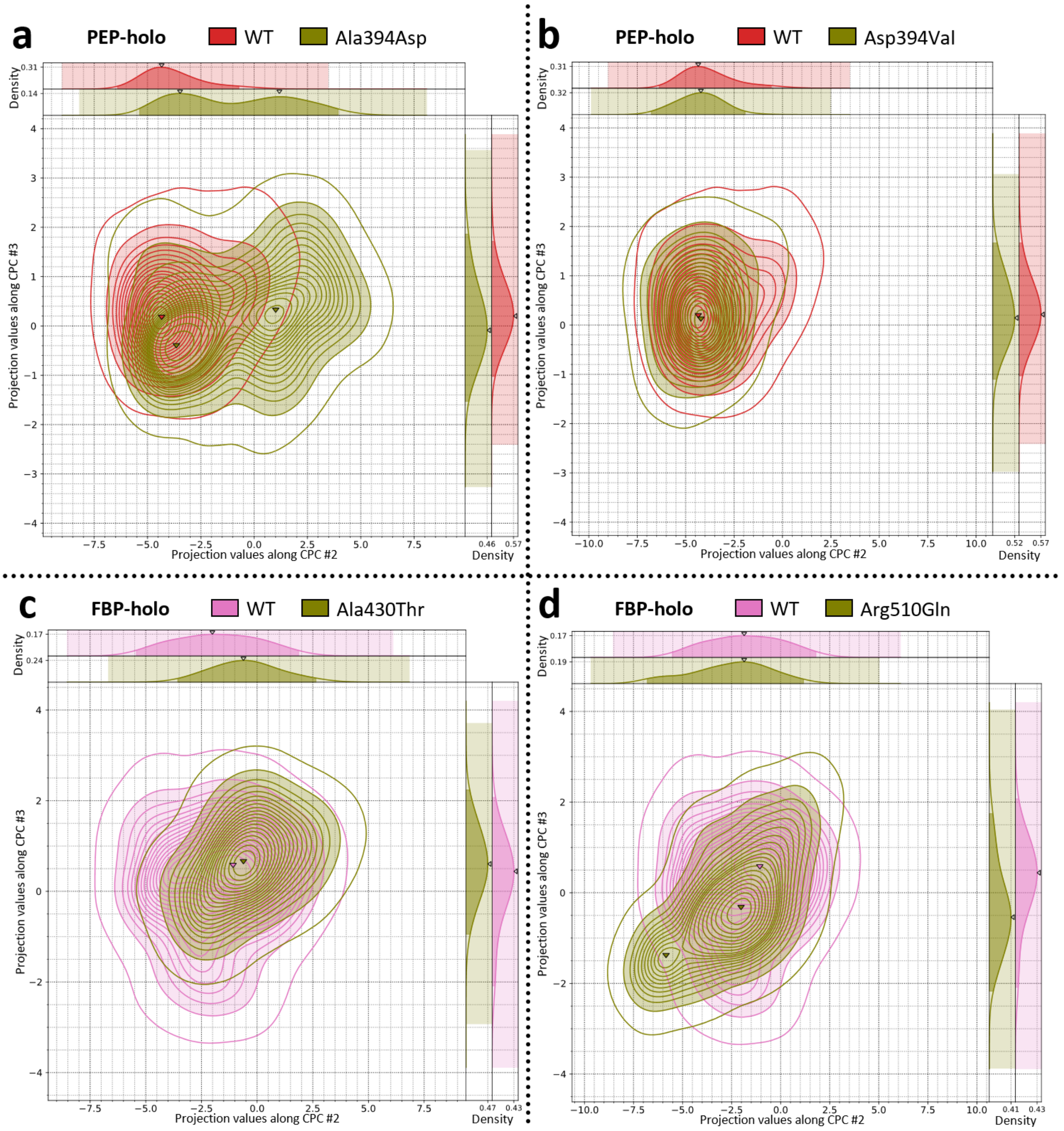


Figure 4.87. Comparative analysis of the WT *vs*. mutant trajectories (A-A' pairs of domains) in holo conditions: conformational profiles (2 CPCs). This figure only shows the examples of four variants. (**a**) WT *vs*. Ala394Asp in the PEP-holo condition. (**b**) WT *vs*. Ala394Val in the PEP-holo condition. (**c**) WT *vs*. Ala430Thr in the FBP-holo condition. (**d**) WT *vs*. Arg510Gln in the FBP-holo condition. The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #2 *vs*. #3 from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figures 4.88 and 4.89 show the box plots and the rankings of the results of the comparative analysis of the adjacent AC cores (C-C′ interface).

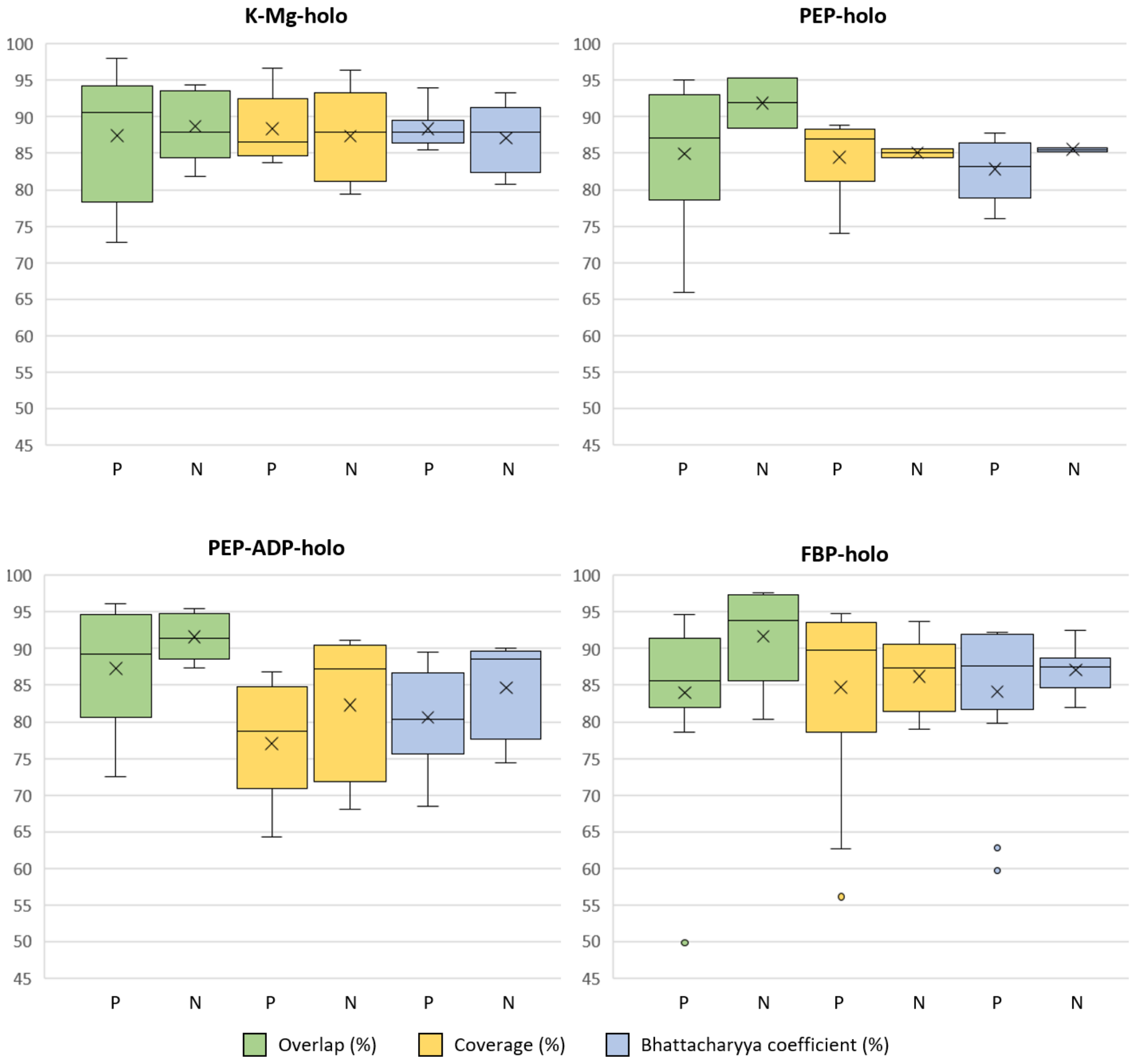


Figure 4.88. Comparative analysis of the WT *vs*. mutant trajectories (AC-C′A′ pairs of cores) in holo conditions: distributions of the quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Data is presented in the form of box plots, distinguishing between the pathogenic (P) and potentially neutral (N) subsets of variants. The BC (values bounded between 0 and 1) is expressed in percentage values to enable a balanced interpretation of the three metrics.

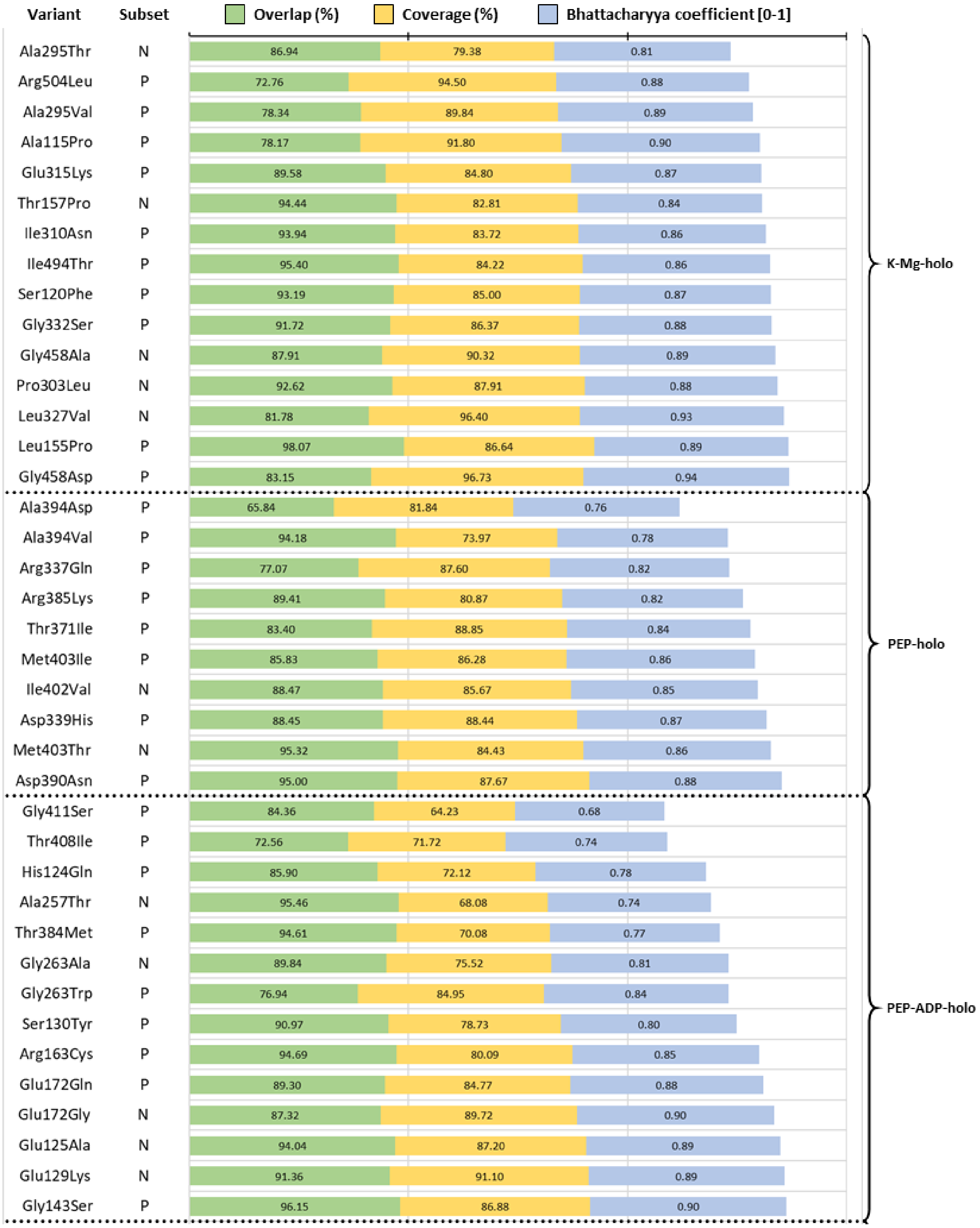


Figure 4.89. Comparative analysis of the WT *vs*. mutant trajectories (AC-C'A' pairs of cores) in holo conditions: quantitative indicators. The similarity metrics (overlap, coverage, and Bhattacharyya coefficient [BC]) were calculated between the 3D density distributions that correspond to the trajectory data projected onto CPCs #1 to #3 from the reference (WT apo) condition. Next to the variant names, the labels "P" or "N" are included to indicate whether they belong to the pathogenic or the potentially neutral subsets, respectively. For each block of rows corresponding to each holo condition, data is sorted in ascending order according to the sum of the three metrics. The bar sizes corresponding to the BC (values bounded between 0 and 1) are shown with a scaling proportional to the overlap and coverage metrics (expressed in percentage values) to enable a balanced interpretation of the three metrics.

Figure 4.89 (Continued)

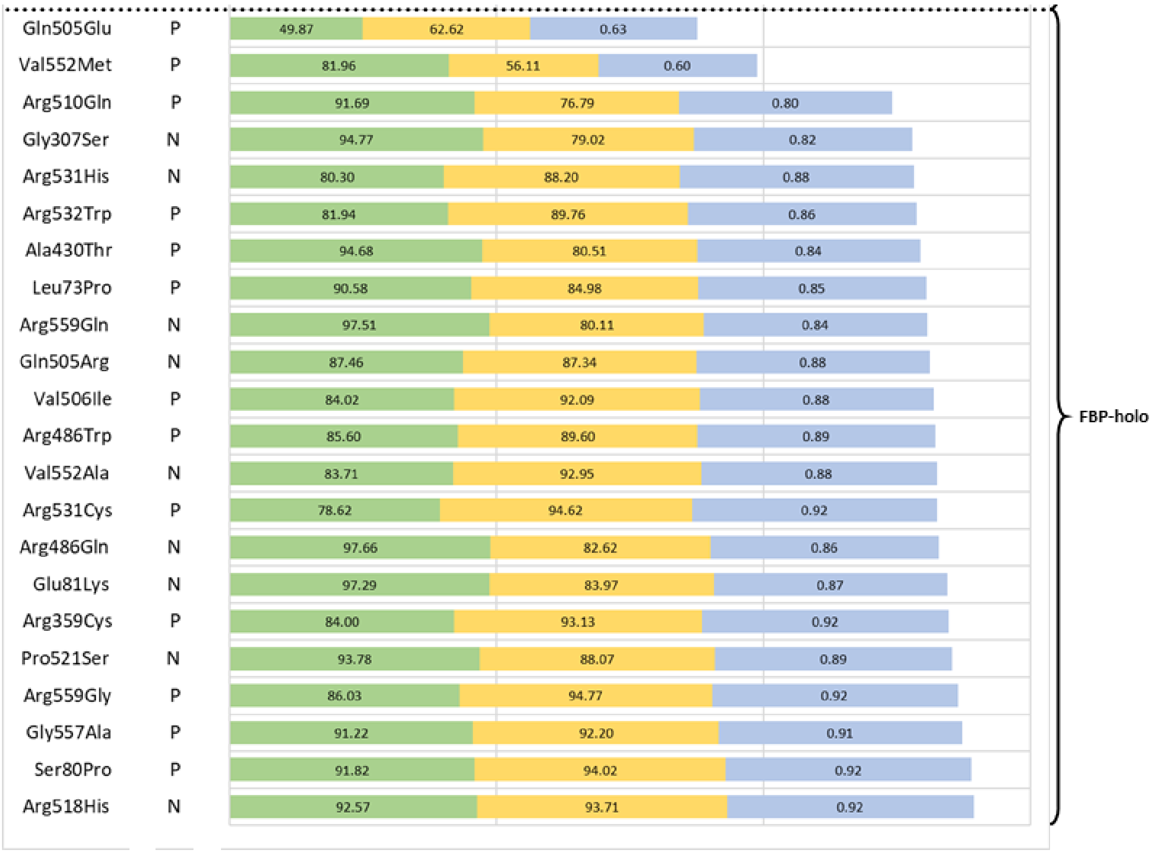


Figure 4.90 illustrates some examples of this comparative analysis by showing the density distributions of CPCs #2 *vs*. #3. Several variants that were already highlighted in the previous comparative analysis due to their strong signs of differential behavior also exhibit low similarity scores in the present analysis. Importantly, these results show further demonstration about the correspondence between the conformational changes at the A-A′ and the C-C′ interfaces (the second CPCs from both CEDAs) that was first highlighted in section 4.1.3.2.3, in the study of the WT system. For instance, the conformational profile of the variant Ala394Asp (pathogenic subset; 1st in the ranking of PEP-holo) again exhibits stability in the inactive conformation despite the presence of PEP (Figure 4.90a), thus leading to the probable impairment of allosteric communications at both types of subunit interfaces. The case of the variant Ala394Val (pathogenic subset; 2nd in the ranking of PEP-holo) is harder to interpret. While it did not manifest dynamical alterations in the previous comparative analysis, it now exhibits signs of differential behavior due to its low coverage with respect to the WT PEP-holo condition. However, its corresponding conformational profile (Figure 4.90b) is considerably similar to that of the WT PEP-ADP-holo or Full-holo conditions (Figure 4.63). Thus, the reported potential dynamical alterations do not suggest dysfunction and, therefore, its status as a pathogenic variant remains unclear.

The variant Gln505Glu (pathogenic subset; 1st in the ranking of FBP-holo) exhibits a strong shift of the density distribution along CPC #3 (Figure 4.90c). The same type of dynamical alteration was detected for this variant in the comparative analysis of the apo condition (Figure 4.81a). These observations account for strong evidence of the pathogenicity of this variant. On the other hand, the variant Gln505Arg (10th in the ranking of FBP-holo), which affects the same position and belongs to the potentially neutral subset, manifests complete regular behavior with respect to the WT enzyme in both analyses (Figures 4.81b and 4.90d). The pair of variants Val552Met (pathogenic subset; 2nd in the ranking of FBP-holo) and Val552Ala (potentially neutral subset; 13th in the ranking of FBP-holo) present a similar case (Figures 4.81c-d and 4.90e-f), except that the latter exhibits a skewed distribution along CPC #3 that suggests less neutrality.

The variants Gly411Ser (pathogenic subset; 1st in the ranking of PEP-ADP-holo) and Thr408Ile (pathogenic subset; 2nd in the ranking of PEP-ADP-holo) are examples of the PEP-ADP-holo condition that failed to undergo a change in sampling towards the region of the active conformation along CPC #3 as the WT enzyme does in presence of both substrates (Figure 4.90g-h). These amino-acid replacements affect the small helical fragment Aα8′ near the active site. Therefore, results suggest that the structural alterations induced by these mutations propagated to the rest of the structure of the core and influenced conformational sampling. Finally, the variant Arg504Leu (pathogenic subset; 2nd in the ranking of K-Mg-holo) exhibits higher dispersion than the WT behavior in the presence of cofactors, with multiple peaks of density scattered around the conformational spectrum of CPCs #2 and #3 (Figure 4.90i). The amino-acid replacement affects the Cα4 helix, at the A-C interface (a main pivot point of the rigid-body collective motions between the A and C domains). Therefore, the conformational profile suggests that the mutation induces instability at the region, leading to higher fluctuation.

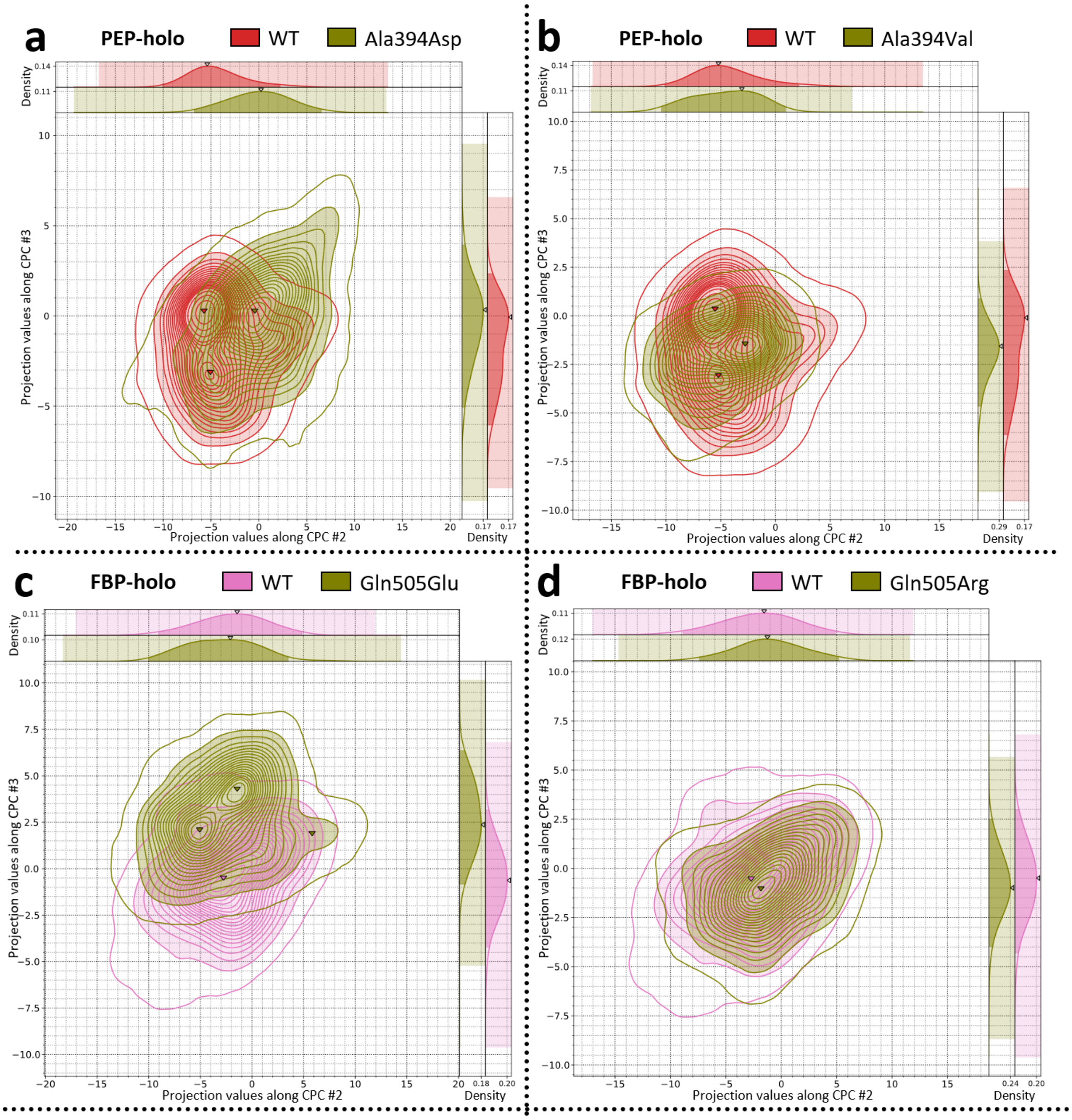
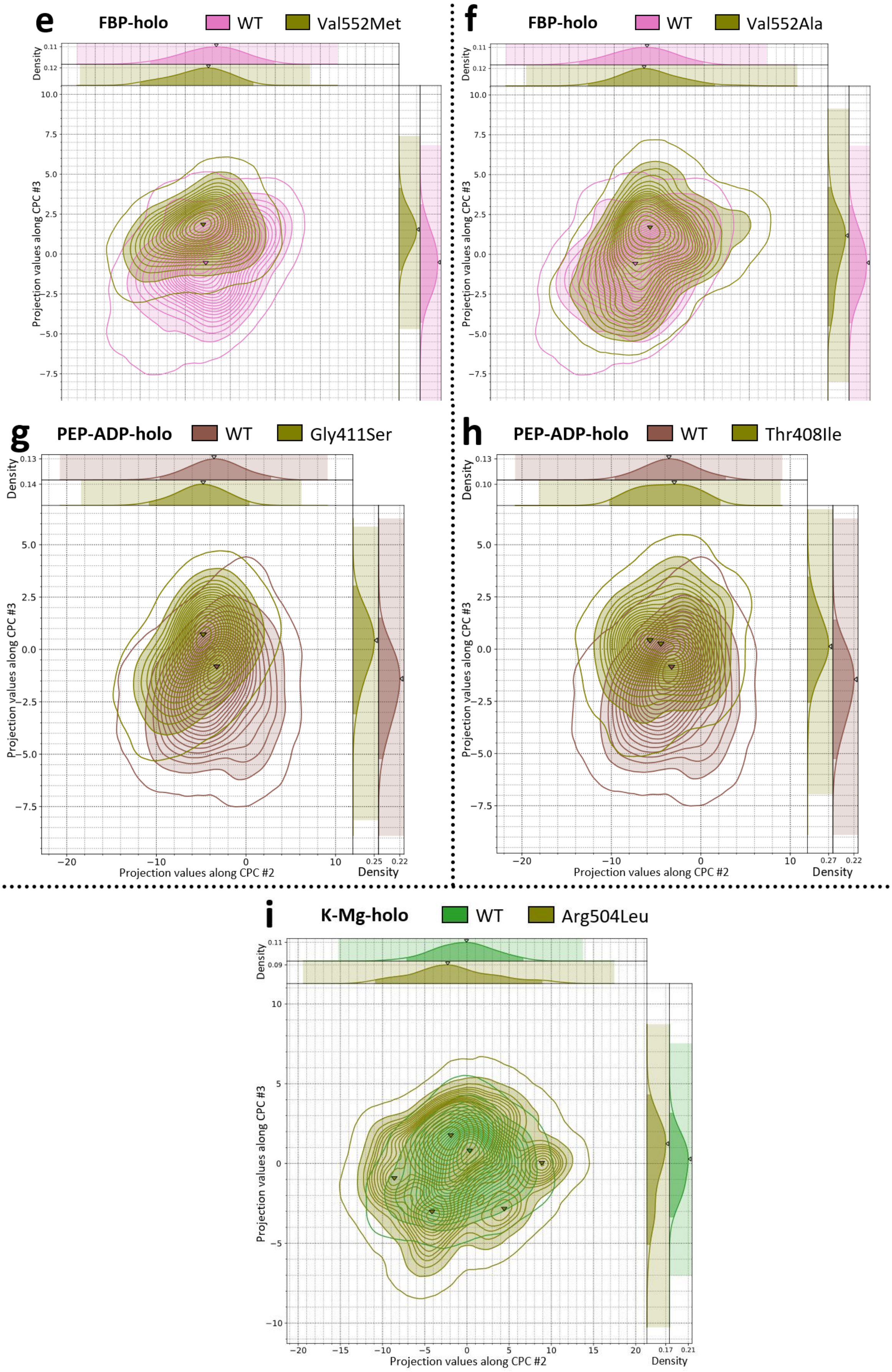


Figure 4.90. Comparative analysis of the WT *vs*. mutant trajectories (AC-C'A' pairs of cores) in holo conditions: conformational profiles (2 CPCs). This figure only shows the examples of nine variants. (**a**) WT *vs*. Ala394Asp in the PEP-holo condition. (**b**) WT *vs*. Ala394Val in the PEP-holo condition. (**c**) WT *vs*. Gln505Glu in the FBP-holo condition. (**d**) WT *vs*. Gln505Arg in the FBP-holo condition. (**e**) WT *vs*. Val552Met in the FBP-holo condition. (**f**) WT *vs*. Val552Ala in the FBP-holo condition. (**g**) WT *vs*. Gly411Ser in the PEP-ADP-holo condition. (**h**) WT *vs*. Thr408Ile in the PEP-ADP-holo condition. (**i**) WT *vs*. Arg504Leu in the K-Mg-holo condition. The 2D density distributions shown in this figure correspond to the trajectory data projected onto CPCs #2 *vs*. #3 from the reference (WT apo) condition. Contour lines delineate 21 levels of highest density percentage: the farthest encompasses the 99.5% of the highest density and serves to enhance perception of the boundaries of the distributions, while the next 20 levels encompass intervals from the 95% to the 0.5% of the highest density in steps of 0.5. Filled contour areas represent the interval of highest 95% density. Triangle markers indicate the locations of local maxima. Each 2D plot also features subplots of the 1D KDE curves along each individual CPC in the top (abscissa) and right (ordinate) margins. In the 1D subplots, the highlighted rectangular regions indicate the total span of projection values, while the filled area under the curve represents the intervals of highest 95% density.

Figure 4.90 (Continued)



4.3 Summary of the CEDA protocol

The implementation of CEDA has been progressively illustrated in the previous sections through the analysis of the trajectory ensembles of PKR. To conclude the Results chapter, this section provides a summary of the main steps of the protocol to facilitate a complete view of the whole procedure. The main part of the protocol relates to the derivation of the set of CPCs from the reference trajectory ensemble. Subsequently, the protocol closes with the proposition followed in this thesis to characterize the conformational distributions along CPCs and achieve a comparative analysis within and between alternative conditions.

Derivation of Consensus Principal Components (CPCs) from a trajectory ensemble

1. **Generate an ensemble of MD trajectories of a macromolecular system** simulated in equivalent conditions, such that the trajectories can be considered as replicates of a single reference condition.

2. Optionally, **choose a structural region that will be the focus of the analysis.** Filter the trajectories to retain only the involved subset of atoms.

    - If there are multiple copies of the chosen region in the topology of the system, and their role in the structure is equivalent (*e.g.*, a certain protein domain or modular region that is present multiple times in a symmetrical multimer), they can be extracted as separate trajectory replicates to enrich the analysis.

3. **Compute the average structure of all trajectory replicates.**

    - If the analysis concerns only a particular structural region of the system, apply this step only to the considered subset of atoms.

4. **Perform (Cartesian) PCA independently on each trajectory replicate.**

    - Prior to the covariance calculation, apply a structural superposition (least-squares fitting) of all trajectories to the reference structure from step 3. Optionally, choose a particular fitting group for the removal of the rotational and translational components in the structural superposition. The analysis will be sensitive to this choice, as it will determine the main point of reference to orient all trajectories and facilitate inspection of the captured collective motions of the structure relative to the position of the superposed region.

5. **Perform a clustering of the resulting eigenvectors.**

    - Include the desired number of eigenvectors per trajectory replicate.
    - Build a dissimilarity matrix using the cosine distance between pairwise eigenvectors. Importantly, cosine distance should be derived from cosine similarity expressed in absolute value, resulting in a measure bounded between 0 (full similarity) and 1 (no similarity).
    - Choose the desired clustering method (*e.g.*, agglomerative hierarchical clustering), feed it with the dissimilarity matrix, and define the criteria for retrieving the most relevant clusters.

- Assess the overall quality and meaningfulness of the obtained clusters. Explore possible biases or tendencies within the composition of the clusters (for instance, check for especially absent or predominant trajectory provenances of the eigenvectors, check for repeated provenances within a cluster…). If needed, retune the clustering parameters or redesign the analysis with alternative structural regions and fitting groups to adjust to the resolution of the dynamics of the system.

6. **Get the centroid (average) vector from each relevant cluster.** The set of centroid vectors will yield the reference CPCs of the experiment.

    - Due to the properties of the cosine distance bounded in the interval [0, 1], vectors within the same cluster may point in opposite directions, describing the same collective atomic displacements albeit reversed. Therefore, this step requires choosing a reference direction in each cluster and then flipping all opposite vectors (angle greater than 90°) before computing the corresponding centroids.

7. **Project the individual trajectories onto the desired CPCs and transform data back to atomic Cartesian coordinates to examine the captured collective motions.**

    - This step requires having applied the structural superposition from step 4 and applying a subsequent data centering of all trajectories around the reference structure from step 3, prior to the data projection. This procedure will establish a common origin of coordinates for all trajectories in terms of the set of CPCs.

<u>Comparative analysis within and between alternative conditions in terms of the reference set of CPCs</u>

8. **Obtain density distributions of the projection values** acquired in step 7, both of the individual trajectories and the aggregated data.

    - Choose the desired method for estimating the probability density functions (*e.g.*, Kernel Density Estimation).

9. **Compare the features of the resulting density distributions.**

    - This procedure allows for ascertaining the conformational diversity among the equivalent trajectories of the system, as well as studying the possible functional implications of both the motions and their most distinctive conformations.
    - Determine the similarities and differences between the distributions. Do they share relative maxima and minima? Does the span of projection values coincide? Are the shapes proportionally similar? Does the aggregated data exhibit distinctive regions of high/low density?
    - The assessment may be conducted by visual inspection of the plotted data and complemented with quantitative metrics that express measures of (dis)similarity between distributions.
    - Identify relevant values of the data distributions and characterize the distinctive structural conformations of each condition.

10. **Generate trajectory ensembles of the system simulated in alternative conditions.** Alternative conditions must hold atomic correspondence (number, identity, and order of atoms) with the reference system.

11. **Project the individual trajectories of the alternative conditions onto the desired CPCs of the reference condition** (apply the considerations from step 7).

12. **Perform a comparative analysis between alternative conditions** with the procedure and the insight from steps 8 and 9.

Chapter 5 Discussion

5.1 Models and simulations of PKR

The proposed research project required, in the first place, being able to model and simulate the PKR system in different biological conditions. The more comprehensive the collection of simulation conditions, the richer the comparative framework that allows elaborating on the knowledge of the structure and dynamics of PKR with respect to its biological function and its regulation mechanism. In turn, by providing a large dataset of MD trajectories, the project also becomes a suitable scenario to suggest an analytical approach like CEDA and test its performance and capabilities.

The present section provides a discussion about the setup procedures that have been implemented to achieve these objectives and obtain the trajectory dataset of PKR. This part of the project comprises the modeling of the initial structures and the parameterization strategy. The major decision points concerning these procedures are reviewed from a rational point of view, at both the scientific and technical levels, with special attention to how each decision has contributed to the goal and the quality of the experiments.

The modeling of the WT PKR enzyme, both as the apoprotein and as the holoprotein with different combinations of ligands, was possible thanks to the large volume of consistent crystallographic data that is available. The structure 2VGB from the PDB was selected as the main base structure. At the time when the project began, 6 different structures of PKR (or the equivalent region of the PKL isoenzyme) could be found at the PDB. From those, only the structure 2VGB corresponds to the WT sequence of the protein, whereas the rest include different point mutations. Furthermore, this model includes most of the natural ligands of the protein, only lacking MgADP.

<u>The absence of a fragment of the N-terminal domain</u>

The only downside of this model is that it lacks a significant portion of the N-terminal domain (the first 56 amino acids). This issue would not have been significantly remediated by using any of the other available models either. The newer structures of PKR/PKL that have been released lately have only achieved the crystallization of a few more residues, and the best models still lack 40 residues. Only now, with the release of the AlphaFold2 database [223], we would be able to work with a model of the full monomeric structure (Figure 5.1). However, the reported confidence score for the predicted fragment of the N-terminal domain is very low, and the model is not compatible with a tetrameric assembly *in silico* because it generates clashes with the adjacent subunits (not shown).

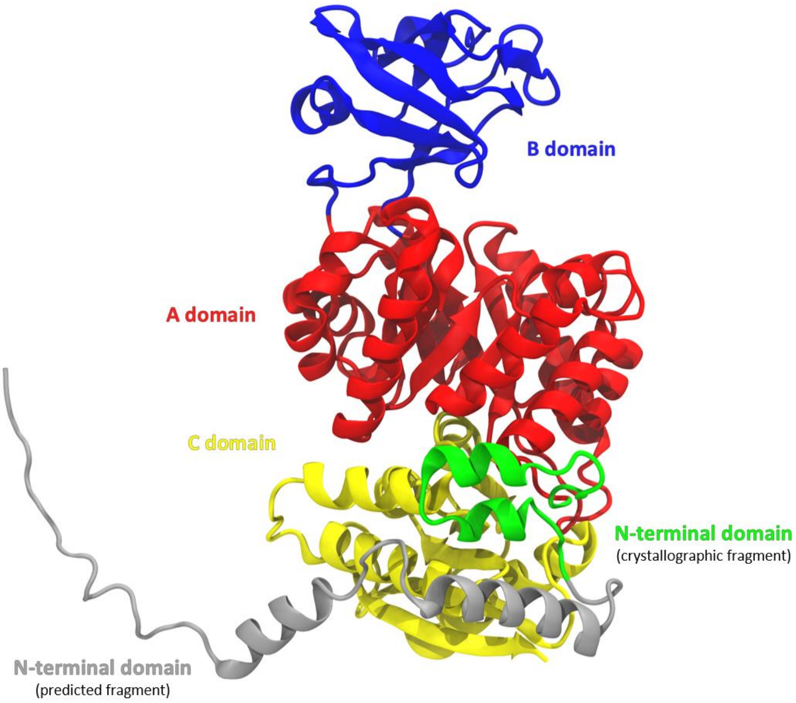


Figure 5.1. Predicted model of monomeric PKR by AlphaFold2. The N-terminal domain consists of a fragment equivalent to that of the PDB structure 2VGB (in green) and a predicted initial fragment that is absent in crystallographic models (in gray). NOTE. The image was generated with the software VMD.

The absence of this N-terminal region is a limitation with regard to the structural and dynamical study of this protein, since we may be missing its participation in the relevant motions of the protein. However, the evidence from crystallographic experiments and AlphaFold2 calculations suggest that this initial region of the domain is disordered. In fact, it is still not clear whether this fragment is directly involved in the modulation of enzymatic activity or fulfills any other functional roles [117]. In mature erythrocytes, the PKR isoenzyme is subject to some levels of post-translational proteolytic cleavage that removes the first 47 amino acids, generating both full-length and truncated proteins that have the capacity both for homo- and hetero-tetramerization [157], [521]. In addition, *in vitro* assays reported that the truncated PKR protein (lacking the first 49 amino acids) exhibits kinetic properties virtually identical to those of the WT enzyme [121]. It is nonetheless intriguing that this fragment appears to be exclusive to erythrocyte and liver isoenzymes, being especially lengthy in the former, based on sequence alignment data of the family of pyruvate kinases [117].

On the other hand, the serine residue at position 43 is a known phosphorylation site of these isoenzymes, which has been correlated with increased affinity for allosteric inhibitors ATP and alanine, and decreased affinity for PEP and the allosteric activator FBP [157]. A mechanism has been proposed whereby phosphorylation of Ser43 would interrupt a constitutive interaction between the N-terminal domain and the main body of the protein that is energetically coupled with stronger PEP binding, causing a decrease in apparent PEP affinity [155], [156]. In contrast, other studies reject this proposition showing that the phosphorylation event alone is not sufficient to alter enzyme kinetics or structure, and suggest that it must regulate activity by a different mechanism than directly altering enzyme kinetics [158].

Be that as it may, the region of the N-terminal that does crystallize folds in a small α-helical structure that interconnects the A and C domains of adjacent subunits of the homotetramer. Therefore, it may potentially be implicated in an orchestrated transmission of conformational changes between subunits, albeit not being essential as evidenced by its absence in PK isoenzymes of several species that display allosteric capabilities [117]. Recently, it has been found that the synthetic drug mitapivat enhances PKR activity by binding to a site buried between the N-terminal domain and the A and C domains of the adjacent subunit, presumably stabilizing the tetramer in its active state [53], [54]. Similarly, 17 newly synthesized compounds have been found to be allosteric modulators that bind to the same cryptic site [522]. The design of such exogenous allosteric modulators was inspired, in turn, by the earlier observations of similar compounds that bind to the homologous site in the PKM2 isoenzyme and that also induce allosteric response [214], [215]. Thus, this region of the N-terminal domain appears to have implications in the structure-dynamics-function relationship of this enzyme.

The chosen holo conditions of PKR

The analyzed holoenzyme states of PKR were rationally chosen to cover several conditions of the protein with partial and full sets of ligands. The goal was to provide a set of models with the potential to track the changes in the dynamical profile of the enzyme across the collection of trajectories, according to the functional capacities of each condition. The corresponding configurations were selected on the basis of the existing evidence of their occurrence in physiologic conditions. Moreover, the fact that PK isoenzymes are remarkably conserved both in architecture and sequence [119] enables the incorporation of structural data from a wider range of PK models, even if they correspond to isoenzymes other than PKR. Currently, the PDB contains more than 120 PDB models of PK, comprising diverse isoenzymes of various organisms and tissues, and co-crystallized with different combinations of ligands.

The configuration with a fully occupied active site is a self-evident choice for the study. Of course, no crystallographic structures come with both substrates as the chemical reaction would take place. Therefore, crystallographic experiments either employ unproductive substrate analogs or simply omit one of them. The structure 2VGB only lacks MgADP. The coordinates of these molecules were imported from the structure 4FXF, a model of human PKM2 with fully occupied active and allosteric sites [149]. This structure emulates the final state of the reaction featuring MgATP and oxalate as an analog of pyruvate. The structure 4FXF displays binding geometries for the rest of the ligands that are totally equivalent to those of 2VGB and many other models [139], [147], and was selected because it is the closest in sequence identity with PKR. The orientation of the γ-phosphate of ATP is compatible with the predicted displacement of this moiety after the phosphoryl-transfer reaction from PEP [140]. Therefore, removing the γ-phosphate from ATP and placing the resulting ADP molecule alongside PEP at the equivalent site in 2VGB rendered a good representation of the pre-reaction conditions. This procedure enabled representation of two holo conditions with fully occupied active sites, named PEP-ADP-holo and Full-holo, where the latter also includes FBP bound to the allosteric site.

The inclusion of the simplest holo condition, only with bound K^+, was an imperative given that PKR is dependent on K^+ for its enzymatic activity. This condition was named K-holo. K^+ is thought to induce the first rearrangements of the active site that subsequently allow either PEP or ADP to bind independently in a random sequential mechanism [140], [146], as this is the usual mechanism in type I K^+-activated enzymes [142]. Under physiologic conditions in the cytosol, K^+ is most likely constitutively bound to the apoprotein due to its abundance [142], [144], [145]. Several crystallographic experiments feature the spontaneous union of K^+ to its canonical binding site, again with a binding geometry consistently equivalent to that of 2VGB [128], [147], thus validating the modeling of this condition in this study.

The binding of the cofactor Mg^{2+} is also required for the stable and catalytically active configuration of the active site. Ligand-binding assays and crystallographic data prove that Mg^{2+} assists in the binding of PEP, although the process by which the bound state of both ligands is eventually reached is not clear. The group of Zhong *et al*. achieved the crystallization of the enzyme with both cofactors K^+ and Mg^{2+} in the absence of substrates or substrate analogs [137]. The model corresponds to a PK of *Trypanosoma brucei* and also features fructose 2,6-bisphosphate at the allosteric site, which stabilizes the enzyme in the R-state (active) conformation of the tetramer. Remarkably, Mg^{2+} was found at a binding site (which they called the Mg-3 site) that is shifted ~3 Å from the canonical (Mg-1) site but involves the same protein side chains. The Mg-3 site was hypothesized to retain Mg^{2+} after product release, working as a priming mechanism for the active site to accept the next PEP substrate molecule and have Mg^{2+} move back to the Mg-1 site. In turn, the binding of PEP would trigger the transition of the B domain from an open to a partially closed conformation. Interestingly, the only instance in which the Mg-3 had been previously spotted was in the crystallographic model of the constitutively active rabbit PKM1 co-crystallized with the substrate analog ʟ-phospholactate [138]. The model shows a range of open and partially closed B-domain conformations, with Mg^{2+} at the Mg-3 site only when the B domain is in the open conformation (and otherwise at the Mg-1 site).

Moreover, Zhong *et al*. also noted that Mg^{2+} (or equivalent divalent metal cations) had never been found in T-state PK structures despite being present at relatively high concentrations in the crystallization buffers [137]. The study of Yuan *et al*. [128], which comprises several crystallographic structures of human PKM2, likely corroborates this observation since Mg^{2+} can only be found in the single model that adopts the R state (with the allosteric activator serine). However, this model is in conflict with that of Zhong *et al*., since Mg^{2+} is located at the Mg-1 site despite the lack of substrate. Other models of the R-state tetramer in the absence of substrates and in the presence of allosteric activators do not detect K^+ at the active site but show Mg^{2+} at either the Mg-1 or the Mg-3 sites [153], [163], [523]. Finally, the crystallographic experiment by Wang *et al*. [127] provided the only instance of a PK in the T state (a model of human PKM2) that, despite failing to detect K^+, shows Mg^{2+} near the Mg-3 site.

With the current conflicting data, it seems that either Mg-1 or Mg-3 can potentially be putative binding sites of the cofactor Mg^{2+} in the absence of substrate. Both states might be in equilibrium. Further experiments should be carried out to ascertain this issue. In this study, the Mg-3 site was selected to model the conditions with cofactors K^+ and Mg^{2+} in the absence of substrate, following the structural evidence presented by Zhong *et al*. [137]. Accordingly, two holo conditions were modeled with this active-site configuration, named K-Mg-holo and FBP-holo, where the latter also includes FBP bound to the allosteric site. The simulation of the K-Mg-holo condition provides the chance to gain novel insight on the preferred conformations of the B domain when both cofactors are bound to the active site. On the other hand, the simulation of the FBP-holo condition allows for the study of the influence of the allosteric activator in the overall conformation of the enzyme.

Finally, the last relevant holo conditions are those with one substrate present and the other absent. Following the evidence for a random sequential mechanism induced by K^+ [140], [146], both models contain this metal. As argued earlier, since the binding of Mg^{2+} alone is yet to be further demonstrated, this cofactor was not included in the model that features MgADP as the only bound substrate (it is important to note that the ADP-bound Mg^{2+} ion is different from the cofactor Mg^{2+}). Furthermore, this particular configuration of ligands can be seen in a recent crystallographic structure [119]. This holo condition, named ADP-holo, allows for the study of the influence of MgADP in the closing of the B domain, as has been suggested by many studies [119], [124], [135], [137], [139], [147]. On the other hand, in the condition with PEP as the only substrate, both cofactors were included (with Mg^{2+} at the Mg-1 site), following the extensive evidence from crystallographic data. This holo condition, named PEP-holo, allows for the study of the cooperative effects exerted by PEP, which presumably manifest as a shift towards the R state [127], [147], [163].

The usage of the AMBER99SB-ILDN force field

In this project, the AMBER99SB-ILDN force field [456] was chosen to treat the system in general. This force field derives from the continuous work that started in the early 1990s, when the Kollman group [327] developed the so-called “second generation” AMBER94 force field with the aim of describing solvated systems like proteins and nucleic acids. Since then, subsequent refinement processes have allowed remedying its major inaccuracies, with an especial emphasis in upgrading the parameters for the protein torsion potentials [282], [284], [524]. Ultimately, the improvement of a few side-chain torsion potentials in 2010 accomplished a considerably better agreement with experimental data [456].

Even though in the subject of force fields it is very difficult to generate gold standards and assess their quality and comparability [524], AMBER99SB-ILDN has been a widely accepted force field, with a generalized use in the last decade [284], [356], [434], [525]. Not surprisingly, therefore, AMBER99SB-ILDN is the most recent force field of the AMBER family that is natively supported in the GROMACS MD suite.

Of course, in parallel, the development of new and better protein force fields has continued up to now. Currently, the AMBER documentation recommends AMBER14SB or the newest addition AMBER19SB (although the latter is stated to pair best with the OPC water model). In this project, the AMBER14SB force field would have been a suitable alternative, which is very similar in terms of compatibility, and the scenario would not have changed much with respect to the design of the MD protocol, except for having to implement a port of the force field for the GROMACS MD suite. In any case, AMBER99SB-ILDN was chosen because of its proven ability to satisfactorily describe the dynamical properties of a wide diversity of folded proteins, as seen in projects of massive simulation data like MoDEL [356], [434], [525].

Furthermore, the algorithm and the QM level of theory by which the parameters of the AMBER94/99 force fields were derived was explicitly disclosed, which allows researchers to make consistent extensions to include arbitrary molecules together with the protein system [459], [471], [524]. This advantage allowed for the inclusion of the parameters for the ligands PEP and FBP according to the GAFF force field and the RESP methodology (details in section 3.2.2).

A bonded/non-bonded hybrid model for the metal centers

The parameterization of the metal centers was the last and most challenging step to set up the holo models for subsequent MD simulation. The general force fields do not provide robust parameters that can describe the interactions of coordination complexes and reproduce the specific binding geometries of the metal-center environments. Thus, this stage of the project involved searching for the available specialized methodology and choosing the strategy that aligns with the needs of the study.

Here, it is important to note that the goal was not to delve deeply into the physicochemical properties of the active site or the catalytic mechanism. Simulations at the QM or QM/MM level would have been required for this purpose. Instead, the aim was to incorporate the known structural evidence about the arrangements of the occupied active site into MD simulations using the derived metal-center parameters. This procedure enabled inspection of the conformational space available to the protein with the local constraints imposed by ligand binding. On a practical level, this could manifest as shifts in conformational equilibria, the emergence or suppression of certain conformations, distinctive collective motions, etc.

In the field of the modeling of systems containing metal ions, a diverse array of strategies based on both quantum and classical regimes have been adopted over the past several decades. At present, no single methodology can be considered superior to all others. The optimal parameterization approach depends upon the specific scientific question under investigation. Some exhaustive reviews and comparative studies comprehensively cover this topic [482]–[485]. Broadly speaking, there are non-bonded and bonded models. While the former allow for dynamical change of coordination number, the latter offer good control over geometry of the first coordination sphere of the metals [483]. Following this primary distinction, the bonded model was the most appropriate choice for the present investigation. The bonded model is widely used with classical force fields and provides metal-center parameters that can readily simulate macromolecular systems with considerable accuracy [484]. In the last few years, several studies of the dynamical properties of metalloproteins and their internal motions via MD simulations have opted for the bonded model [383], [526]–[528].

The MCPB.py software [471] was selected to address this task, as it provides an optimized workflow developed by experts in the field to facilitate metal-center parameterization with the bonded model. MCPB.py supports various AMBER force fields and acts as a bridge between QM calculations and MD simulation software packages, enabling the direct incorporation of the generated parameters into the MD topology files. Parameter determination based on QM calculations offers considerable accuracy and is broadly applicable to a wide range of molecular configurations [471]. The implemented parameterization schemes within MCPB.py allow for the derivation of force constant parameters and atomic charges that fall reasonably within the ranges of values deemed correct in the literature.

In addition to the QM-derived parameters, the bonded model requires assigning appropriate van der Waals (VDW) parameters to the metal. The 12-6 Lennard-Jones (LJ) potential is the most widely employed expression for the repulsion and attraction terms of the VDW interaction. Li, Merz, and co-workers developed 12-6 LJ parameters for monovalent to tetravalent ions by trying to reproduce thermodynamic quantities, structural properties, and kinetic or dynamic properties. They derived different parameter sets depending on the targeted experimental value: hydration free energy (HFE), ion-oxygen distance (IOD), or coordination number (CN) of the first solvation sphere [469], [470]. In this study, the IOD parameter sets for K^+ and Mg^{2+} at the holo PKR system were incorporated, as they are recommended for MD simulations oriented to reproduce structural properties such as metal center geometries [455], [526].

The bonded model is unable to simulate ligand-exchange processes due to its inherent construction. A bonded/non-bonded hybrid model can be utilized to allow specific coordination sites to undergo ligand switching by modeling interactions at these sites solely with non-bonded parameters [484]. In this study, this approach was employed to facilitate the modeling of the interactions between the metal centers and the solvent. Accordingly, after the parameterization stage with QM calculations, the crystallographic water molecules of the model were not retained as permanent ligands in the MD simulations. This allowed for the exploration of the changes in coordination number at the metal centers according to the stability of each configuration in simulation.

The setup of the cluster model and the QM calculations

Some of the modeled configurations used in this study were significantly more complex than those shown in the MCPB.py usage examples. As recommended by the authors of the software [484], the basic protocol was further adapted and refined with the support of the literature and in accordance with best practices in the field. The process required multiple iterations that enabled the successive identification of weak points and subsequent refinement. As such, the resulting procedure in this thesis also served as a practical application of these methodologies and may provide guidance for future studies of metalloproteins or systems with comparable configurations.

In general, the cluster-model strategy represents a robust and versatile scheme that has been widely used to study the reactions and properties of enzymes, especially metalloenzymes, using QM methods. In this approach, only the metal binding site and a limited number of nearby atoms are explicitly included in the model, while the rest of the protein is ignored. Over the years, this strategy has solved many complex problems in computational enzymology, and models with 250–300 atoms are now routine thanks to the technical advances in computational infrastructure [144], [276], [277].

Chemical intuition plays an important role to address the practical issues that facilitate a correct design of the cluster model. The building of the model starts from the existing structural data of crystallographic structures. The first rational decision concerns the selection of the atoms to be included in the model in order to adequately describe the properties of the metal center in its structural and functional contexts. Besides the metal and its coordination ligands, additional residues or moieties of the second coordination sphere or even beyond may be included, as appropriate: i) those responsible for stabilization and binding of substrates, ii) those exerting substantial short-range and long-range electrostatic interactions, iii) those assumed to be involved in the catalytic mechanism, and iv) those involved in the optimal orientation of the side chains via nonspecific steric hindrance [276], [281], [526].

A systematic approach is to begin with the minimal set of atoms and to gradually include additional fragments, thus examining their effects and gaining insight into their chemical roles [277], [281]. Accordingly, the environment of each metal center in the holo conditions of PKR was meticulously inspected to identify the interactions with the immediate environment that are crucial for reproducing the corresponding geometries. Arg116 and Lys313 are catalytically relevant residues [139] at the second coordination spheres of K^+ and Mg^{2+} that interconnect both metal sites and mediate relevant electrostatic interactions. In turn, Ser286 constrains the side chain of Lys313 to its functional position and models the network of H-bonds between this residue and Thr157. The side chain of Glu161 incorporates the negative charge density that helps orient the hydroxyl group of Ser120 to interact with Glu161 and simultaneously coordinate to K^+. In the holo conditions where ADP is present at the active site, His121 (in its ε protonation configuration) helps accommodate the β-phosphate moiety by forming a H-bond. Furthermore, the side chains of Thr371 and Ser405 give structural context to the network of H-bonds between PEP and the water molecules coordinated to the ADP-bound Mg^{2+}. These two amino acids have been suggested to participate in the enzymatic reaction as general proton donors thanks to a proton-relay system with the H-bonding network of water molecules nearby [124], [139]. Finally, the included water molecules help describe the solvent-exposed portion of the metal centers. Some examples in the literature reported improvements in the accuracy of the cluster model after adding explicit water molecules [281]. In the PKR cluster models, they fill in the expected coordination sites and establish a robust network of H-bonds that signifies the explicit contribution of the solvent to the overall geometry. Without these, the orientation of several residues such as Asn118, Ser120 and PEP would be indeterminate at the boundaries of the cluster model.

As the cluster model expands in size, the computational cost becomes increasingly prohibitive. Therefore, it is important to determine the point at which a satisfactory balance between accuracy and computational efficiency is reached. Capping schemes are usually employed to reduce the size of the structure to only the essential fraction that exerts the relevant interactions. Amino-acid side chains can be truncated at the Cβ or Cα atoms, which are saturated with hydrogen atoms. When carbonyl groups or amide groups of the backbone are relevant, acetyl (ACE) and N-methyl amide (NME) groups represent a minimal version of the backbone [276], [471].

Frozen-atom schemes also let us represent structural constraints without the need of including further additional residues. Geometry optimizations can lead to large movements in regions that should be occupied by other residues, especially at the boundary of the cluster model. Peripheral atoms are usually fixed at their original crystallographic positions to avoid such artifacts and at the same time minimize the total number of atoms and avoid unnecessary calculations [276], [279]. Essentially, geometry optimizations should facilitate the proper relaxation of the metal-center components without compromising the conformational information imposed by the backbone of the protein. In general, fixing the following fragments has been shown to yield satisfactory results: i) truncation sites of side chains (α-carbon and the surrogate hydrogen atoms), ii) entire side chains further than the first coordination sphere that do not require structural optimization, iii) atoms involved in H-bonds or polar interactions with absent residues, and iv) secondary structure elements and backbone atoms in general [276], [277], [281]. These guidelines were taken into account to build the cluster model of each PKR holo condition. After careful monitorization, optimal configurations for the frozen-atom schemes were achieved (see the details in section 4.1.1.2).

The surrounding of the cluster model beyond the boundary is expected to exert a certain electrostatic influence that may produce a polarization effect. Metal-containing cluster models, especially when they possess non-zero net charge, are prone to be affected by unrealistic electron transfer. Implicit solvation models can compensate for this inaccuracy by simulating a homogeneous polarizable medium using a dielectric constant [276], [279], [281]. Although the value for this dielectric constant (ε) is arbitrary, values between 2 and 40 may properly describe solvation effects in agreement with experimental data [144], [276], [458], [526]. The value of a pure water medium corresponds to approximately 80. The value of 4 is the standard procedure to simulate the average effect of a protein matrix environment and the buried water molecules surrounding the cluster model [277], [529]. For more solvent-exposed sites a value of 20 may be used [479], [530]. In this study, several values of the dielectric constant were tested. Geometries obtained with a value of 20 showed best agreement with the experimental structures. It has been seen that the larger the size of the cluster model, the lower the impact of the continuum solvation model on accuracy [276], [277], [281], [480]. In the PKR cluster models, the dielectric constant strongly influenced the orientation of the charged residues at the boundary (*e.g.*, Asp156), whereas other residues less exposed to the solvent were not significantly affected.

In respect of the choice of an appropriate QM level of theory, Density Functional Theory (DFT) methods are the most widely used in theoretical mechanistic biochemistry, being able to model complex chemical systems like metalloproteins with a positive balance between computational efficiency and accuracy [276], [279]. The best predictions are usually achieved by combining DFT with a partial or full Hartree-Fock (HF) exchange contributions, resulting in hybrid functionals. Specifically, the most popular hybrid functional to study metalloproteins that bears compatibility with the AMBER force fields is the B3LYP hybrid functional [277], [526]. The incorporation of dispersion correction terms such as GD3BJ is essential because it improves the description of non-covalent interactions [277], [279]. For studies of metal-containing systems (especially the heavier transition metals) or when high accuracy is a priority, moving to a more modern method than B3LYP is also encouraged [280].

Similarly, established criteria guide the selection of the basis sets for QM calculations. As a general indication, given the limitations of computational resources, one should select the minimal basis set that guarantees the desired level of precision for each computational task. Accordingly, geometry optimizations can make use of a relatively small basis set (typically of double-zeta quality) to procure reliable geometries. Then, subsequent single-point energy (SPE) calculations may rely on considerably larger basis sets (typically of triple-zeta quality and including polarization and diffuse functions) to obtain accurate energies. Polarization functions should be included in the basis set to represent the electron distribution along chemical bonds, especially in H-bond interactions where there is substantial polarization. In addition, diffuse functions are an essential addition to accurately describe the electronic density of anions and coordination complexes, which spread out further from the atomic centers [276], [279], [280]. Multiple studies of systems that share a similar complexity with the PKR models follow these criteria [281], [383], [527], [531], [532]. Alternatively, Effective Core Potentials (ECP) or pseudopotentials may also be added to provide explicit representation of the core electrons of the metals [144], [484], [529], [531], [532].

After geometry optimization, a subsequent harmonic-frequency calculation facilitates the derivation of harmonic force constants and equilibrium values for the corresponding bond and angle parameters of the metal center. Ideally, this type of calculation should be done upon confirmation that the optimized structure indeed corresponds to a local minimum. To achieve this challenging objective, one can follow the recommendations that have been traditionally used to address the relevant methodological caveats [276], [533]. In this study, the geometries of each cluster model were optimized until reaching the convergence threshold values suggested in the MCPB.py protocol. In a few instances where linear angle values arose during the calculation, the associated indeterminacies were amended by applying a restart with a new estimate of the Hessian matrix [533]. Overall, the obtained geometries yielded satisfactory results, as evidenced by the performance of the resulting force constants in MD (see below). A possible refining stage could comprise further iterations near the final equilibrium geometry with tighter convergence values and the exploration of possible conflictive degrees of freedom.

Finally, the RESP algorithm is a standard practice for fitting the partial charges of metal-protein complexes where charge-transfer effects take place [484], [526], [531], [532]. A SPE calculation can be performed without the need of a full geometry optimization if the positions of hydrogen atoms are reasonable. As advised by the MCPB.py developers, and considering that the PKR models comprise a substantial part of the environment of the metal centers, the models were submitted to hydrogen-only geometry optimizations before fitting atomic charges [472].

Implementation of the QM-derived parameter sets

After the QM stage, bond and angle parameters can be generated with MCPB.py based on the Seminario method, which calculates force constants using a sub-matrix of the Cartesian Hessian matrix [471], [481]. The Seminario method is a recommended choice for the derivation of an harmonic approximation for the bonded terms of structural metal sites [485]. This method has the advantage of being simple and accurate enough for most applications. However, it also has a limitation that stems from its main characteristic. The parameterization implicitly incorporates the influence of the environment, which makes it adaptable to various systems but also incapable of representing isolated bonded terms. For studies that require the explicit decoupling of the bonded and non-bonded parameters, the so-called Automatic Parameterization Methods (APM) may be more suitable, as they involve scanning the parameter space and comparing the resulting properties with additional experimental or QM data [482], [484].

Conventionally, with the Seminario method, dihedral torsion parameters are disregarded because they are based on harmonic terms, whereas protein force fields employ a Fourier expansion to represent these parameters. Dihedral interactions incorporate the rotational barriers that arise from the physical repulsion between electron clouds of adjacent chemical bonds. A proper parameterization strategy of dihedral terms would entail a comprehensive refit by examining all the available degrees of freedom of the structure. However, this procedure is challenging, especially considering that dihedral terms are often coupled with the van der Waals and electrostatic terms that also affect the structure. In practice, it is assumed that all the dihedral angle values of a metal center are accessible at physiological temperature, and that the corresponding energy barriers are small and negligible. Therefore, they are usually set to zero as an acceptable trade-off between accuracy and effort [482]–[484].

After the acquisition of the QM-derived parameter sets, the performance of the models in MD simulation was evaluated. Some interactions at the metal centers benefited from a variation of the bonded model by which coordination bonds are modeled as individual harmonic restraints rather than being included as actual chemical bonds in the topology. In practice, the treatment of the bonded terms is virtually equivalent with either approach, although the decision affects how non-bonded interactions between neighboring atoms are considered.

In protein force fields, a scheme of exclusions is implemented with the bonded model to prevent double-counting of the VDW and electrostatic interactions between consecutively connected atoms, which may already be implicitly accounted for with the bonded parameters derived through QM [485], [524]. For pairs of atoms separated by one bond (1-2 pair) or an angle (1-3 pair), these interactions are neglected. For pairs of atoms separated by three bonds (1-4 pair; endpoints of a dihedral torsion angle), these interactions are scaled down by factors of 0.5 and 0.88 respectively (values correspond to the AMBER99SB-ILDN force field). However, under certain situations when modeling metal centers, reincorporating the full non-bonded set of terms may be advantageous in preventing structural distortion [472], [484], [485]. The implementation of this approach yielded stable models for the PKR metal centers that optimally reproduce the experimental geometries.

In particular, this adjustment was instrumental in addressing the inaccuracies encountered in the simulation of the side chain of Ser120, a coordination ligand of K^+ (see section 4.1.1.2). The oxygen atom of this hydroxyl group is the only instance of a donor atom in the coordination complexes of PKR that corresponds to the penultimate atom of the chain of a chemical moiety. In its structural context, this functional group exhibits rotational freedom, as the hydrogen atom can reorient to transiently form H-bonds with either Glu161 or nearby water molecules without disrupting the O-K^+ coordination bond. The incompatibility of the standard bonded model with this region became apparent as soon as events of the switch of H-bond acceptors occurred in simulation. The lack of the conventional non-bonded terms caused structural instabilities due to unbalanced forces, potentially leading to simulation failure. For example, with the 1-4 interactions scaled down, exaggerated approaches were produced between the hydrogen atom and its 1-4 partners such as the δ oxygen of Asn118. This anomaly arises because the VDW interaction between such 1-4 pairs is zero while their electrostatic interaction is non-zero and attractive. This effect has been previously observed and reported in the literature (*e.g.*, for cases of two adjacent water molecules coordinated to the same metal) [484], [485].

This variation of the bonded model with harmonic restraints and full non-bonded terms was also applied to the rest of interactions of the K^+ metal center to harmonize the treatment of the whole coordination complex. In addition, the QM-derived equilibrium angle value between K^+ and the hydroxyl group of Ser120 was dismissed to further relax the moiety in MD simulation. Serine is not an uncommon coordination ligand in metalloproteins, especially in K^+ metal centers [144]. However, to the best of my knowledge, other studies that have parameterized metal centers that contain serine (or, equivalently, threonine or cysteine) either assumed its deprotonated state based on pK_a calculations [485], [534] or did not report the specific QM-derived parameter sets. Therefore, the strategy implemented here could not be compared with other studies. We might speculate that this adjustment is necessary for similar systems unless the rotational freedom of the coordinated hydroxyl group is restricted by steric hindrance or stable interactions.

<u>Evaluation of the metal centers in MD simulation</u>

The derived metal-center parameters successfully maintained the structural constraints of ligand binding at the active site along the MD trajectories, allowing for a comparative study of the dynamical behavior of PKR in apo and holo conditions. The geometrical and structural analyses of the metal centers show that the local conformation of the region was stable without significant distortion, whereas minor fluctuations were allowed to occur. The use of the RMSD and RMSF metrics provides a straightforward and informative approach to assess the flexibility of parameterized metal centers [527], [531], [535]. Accordingly, the protein region around the metal center, especially the coordinated amino acids, showed an overall higher conformational divergence in the apo condition. Local stiffness increased gradually as more ligands were bound in the active site.

In MD studies with parameterized metal centers, it is important to evaluate the interatomic distances between the metal and its coordination ligands over the course of the simulation, as well as to determine the predominant coordination geometries [109], [383], [531], [535]. Even though crystallographic metal-ligand distances and QM-derived equilibrium bond lengths in general did not bear equivalent values, the overall geometry of the coordination complexes was well reproduced. Metal-ligand distance values along the trajectories appear distributed around the corresponding equilibrium values and do not show abnormal deviations.

The K^+ metal center exhibited more mobility, as suggested by the wider distance value distributions and the higher number of upper outliers. This behavior could be attributed to the particular chemical and structural features of K^+ metal centers. Coordination complexes of this metal cation can adopt diverse geometries and are mainly governed by electrostatic interactions due to its low charge density. In addition, K^+-specific binding sites in enzymes such as PKR are often characterized by large cavities that can accommodate the particularly large ionic radius of K^+ (1.33 Å) [142], [145].

The average bond distances between K^+ and the coordinated amino acids and the substrate PEP range from 2.7 to 2.9 Å, which is consistent with the literature values. For example, Brás *et al*. performed a comprehensive analysis of the geometric properties of K^+ coordination spheres based on a representative set of structures from the PDB and validated by QM calculations [144]. They reported that the average oxygen-K^+ distances fall within the same range and vary inversely with the formal charge of the oxygen atom. This trend is partially observed in the simulations of this study, where the Asp156-K^+ and Ser120-K^+ distances show the lowest and highest values, respectively, in some instances of the holo conditions. An exception to this trend is the PEP-K^+ distance, which should be among the lowest distances considering that the oxygen belongs to a phosphate group. Interestingly, the crystallographic reference values of the K^+ metal center in structure 2VGB are higher, ranging from 2.95 to 3.44 Å [121], but they follow the pattern determined by the formal charge of the oxygen atom.

On the other hand, Mg^{2+} coordination spheres are more rigid and stable [142]. This feature has been reflected in narrower distributions of the distance values of both Mg^{2+} coordination complexes. Studies of the QM-calculated and experimental distance between Mg^{2+} and carboxylate ligands in monodentate fashion reveal characteristic values that are around 2.0-2.1 Å, with more distant instances not exceeding 2.6 Å [530], [536]. Coordination bonds of cofactor Mg^{2+} with Glu315 and Asp339 consistently reproduced this behavior both in the Mg-1 and Mg-3 sites. The tridentate configuration between Mg^{2+} and PEP results in slightly increased distance values. Finally, the interaction between PEP and the ADP-bound Mg^{2+} exhibits distance values higher than the corresponding parameterized equilibrium value. This suggests that the model is unable to accommodate the repulsive force exerted by the phosphate groups of both PEP and ADP.

The dynamical changes in the configuration of the metal centers were inspected with a straightforward strategy based on the use of distance cutoff values to detect the number of occurrences of certain atoms being near the metal. The particular cutoff values were empirically determined, starting from the tentative guess of 3.5 Å and gradually decreasing the value until finding an appropriate balance between the sensitivity and specificity of the detection experiment. After cutoff adjustment, the final values employed to detect coordinated water molecules were 3.1 Å and 2.8 Å for the K^+ and Mg^{2+} metal centers, respectively. Cutoff values of 3.59 Å and 3.0 Å have been used in previous studies [145], [536]. The selected values in this study should provide good estimates according to the experimental metal-water distances reported in the literature. Distances for K^+ in aqueous solution range from 2.65 to 2.97 Å, with 2.8 Å being the most common reported value for 6- and 7-coordinate complexes of this ion [145], [537]. In the case of the smaller Mg^{2+} ion, distances are typically lower, between 2.0 and 2.1 Å [530], [536]. All in all, this approach yielded satisfactory estimations of the relative abundance of coordination numbers and configurations of the PKR metal centers in MD simulation. However, alternative tools and methods exist for conducting a more refined analysis when necessary. For example, one could incorporate angle values along with distance values or employ a Radial Distribution Function (RDF) scan [527].

Remarkably, results suggest that the non-bonded portion of the models was able to simulate the spontaneous occupation of the vacant coordination sites. The coordination number of the complexes fluctuated along the trajectories mainly as a function of the hydration state. In general, the most abundant states for the K$^+$ metal center were the 5- and 6-coordinate states. Conversely, both Mg^{2+} metal centers were strongly stabilized in the 6-coordinate state, with invariable octahedral geometries.

Coordination numbers of metal ions in metalloproteins may range from 3 to 8, being 6 the most prevalent overall [142]. K$^+$-binding proteins are characteristic examples of this trend, preferentially adopting the large 6- or 7-coordinate states. Binding sites of K$^+$ are mostly composed of oxygen atoms with partial negative charges provided by either the side chain or the backbone of amino acids. However, the water molecule is the most common ligand of the first coordination sphere. Interestingly, the ligand combination Asn-Ser-Asp-Thr-Water-Water, characteristic of the K$^+$ metal center of PKR, is a prevalent configuration among K$^+$-binding proteins [142], [144]. Water-exchange events may be favored by the presence of the rest of protein ligands at K$^+$ metal centers, given the large ionic radius of the metal and the fast association and dissociation kinetics of K$^+$ complexes. When a phosphate group is a coordination ligand of K$^+$, the strongly negative charge of this moiety reduces the electrostatic potential of the metal cation, limiting its tendency to interact with other molecules [145]. This statement correlates with the fact that in presence of PEP, a higher proportion of states without water (*i.e.*, bonded-model–only states) were detected.

Mg^{2+} metal centers usually form stable octahedral geometries with six oxygen ligands. Such coordination complexes cannot accept further ligands due to the lack of vacant space in the vicinity of Mg^{2+} [142]. This fact may explain the consistent absence of higher coordination numbers in the trajectories of this study. Mg^{2+} binding sites in proteins contain at least one coordinated carboxylate group in monodentate fashion. The rest of the coordination sphere may be occupied by other moieties of the protein or external ligands, and completed by coordinated water molecules to form the octahedron. Furthermore, the rate of dissociation of water molecules in Mg^{2+} complexes is slower [142]. Accordingly, almost no variability in the configuration of the Mg^{2+} metal centers of PKR was observed along the simulations (states with coordination number less than 6 occurred with negligible frequencies).

5.2 The applicability of the CEDA approach

The CEDA methodology was developed in response to the objective of exploring analytical strategies that can leverage standard PCA to overcome its inherent limitations in comparative trajectory analysis. PCA is widely utilized in trajectory analysis due to its straightforward applicability and the fact that the captured dynamical features have the potential to afford biological interpretation, such as associating observed collective motions with their potential role in molecular function. However, when working with a set of trajectories, the usual approach is to conduct PCA on each trajectory as an independent experiment. Then, comparing PCs involves a visual examination and description of the observed similarities and differences among trajectories, resulting in comparisons that are qualitative rather than quantitative. An alternative approach, known as combined-PCA, is also frequently employed. This method involves applying PCA to a “multi-trajectory” that is formed by concatenating the various trajectories being studied, rendering a single set of PCs representative of the whole ensemble. The differences between combined-PCA and CEDA will be discussed in detail later in this section.

The CEDA protocol is divided into two parts. The first part is designed to detect the collective displacements that are most representative of a trajectory ensemble of a macromolecular system. The approach facilitates the dynamical study of the system, grounded in the principles of EDA, but allowing for the integration of the PCA output from independent trajectories within a unified framework of consensus PCs (CPCs). Thus, the available trajectories are considered as complementary views of a single conformational ensemble, enabling elucidation of the predominant dynamical behavior. The second part of the protocol introduces an analytical strategy to characterize the conformational distribution along the derived CPCs. This enables a comparative analysis between different trajectory ensembles by determining whether an alternative condition has explored a similar conformational space in terms of the reference CPCs.

Through the case study of this thesis, the practical application of CEDA has been effectively demonstrated. The protocol was employed to analyze three different regions of the structure of PKR, offering valuable insights into its potential and advantages for the dynamical study of biomacromolecules. The implementation proved successful, as it resulted in the identification of key characteristics pertaining to the dynamics of PKR. The detected dynamical events are consistent with the existing body of evidence regarding the conformational changes linked to the function of this protein, as will be discussed in the next section. Additionally, this exercise has facilitated the identification of possible caveats and upgrades to the strategy. With this perspective, the following discussion involves an examination of the benefits of the employed methodology, as well as several aspects of the process that could be further optimized to manage a broader range of macromolecular systems.

The influence of sampling heterogeneity between equivalent trajectories

First and foremost, the ensemble of equivalent trajectories of the reference condition, the WT apo condition, demonstrated a substantial degree of similarity of the explored conformational space. This scenario is determinant for the derivation of meaningful CPCs and, thus, also for the subsequent chance to perform comparisons with alternative conditions. Three relevant CPCs were produced per CEDA experiment, each accounting for a distinctive collective motion of the corresponding analyzed protein region. These CPCs were highly representative of the whole trajectory ensemble because they were derived from similar eigenvectors that came from all or most of the 20 equivalent trajectories of the reference condition (between 95% and 100% of coverage in the case of CPCs with indices #1 and #2, and between 65% and 95% in the case of CPCs with index #3).

The trajectories of the reference condition are called “equivalent” not because they are identical, but because they represent equivalent attempts to sample the conformational space around the employed initial structure in the given simulation time. Thus, none holds greater value than the rest. Due to the stochastic nature of MD, the more available trajectories, the greater the quality of the sampling [293], [445], [538], [539]. Therefore, any average property computed from a trajectory ensemble is statistically more significant and consistent between different experiments [109], [335], [365], [445]. For this reason, CPCs with high coverage and eigenvector similarity are good estimates of the true underlying collective motions of the system or, at least, more robust than regular PCs from a single trajectory. In turn, this implies that the applicability of CEDA is subject to the intrinsic dynamical nature of the system under study. For instance, extremely free and flexible large systems with a wide disparity of sampling spaces (*i.e.*, too many sampled energy minima between independent equivalent trajectories), are expected to produce very few or no meaningful CPCs.

From a more methodological point of view, the detection of similarity between eigenvectors depends on four factors: i) the consideration of sufficient eigenvectors in the clustering, ii) the employed measure of (dis)similarity, iii) the clustering algorithm, and iv) the criteria used to define consensus behavior.

Selecting the quantity of eigenvectors to incorporate into the clustering

Regarding the number of eigenvectors incorporated per trajectory in the clustering process, several considerations can be taken into account. Collective motions that co-occur in multiple trajectories may appear in eigenvectors of either the same or different index, by chance, according to their weight in the conformational variance of each simulation. The computation of matrices of inner products between pairwise trajectories is frequently performed to detect such shifts [441]. The application of a clustering algorithm, as in CEDA, offers a similar approach, but with the significant advantage of providing a comparison among all trajectories in a single operation. Furthermore, by visualizing the results of the clustering in a dendrogram, we obtain a comprehensive view of all eigenvectors agglomerated by similarity. Including an excessively large number of eigenvectors in the clustering implies adding noise to the dendrogram and may lead to a degradation of performance. Another potential constraint might be computational time, although it should not pose a significant obstacle.

The recommended approach may mirror that of conventional EDA, whereby the objective is to identify the "essential subspace" which comprises only the first eigenvectors that define the directions of the most important atomic fluctuations. This same principle could be applied in this context. A common criterion consists in constructing the so-called "scree plot" that shows the first eigenvalues (or their ratio to the sum of all eigenvalues) as a function of the eigenvector index. In globular proteins, this plot often displays an "elbow point", that is, an abrupt change of slope beyond which the subsequent PCs only provide marginal information and thus pertain to a more irrelevant subspace [379], [446]. The scree plots of this study (Figures 4.17, 4.38, and 4.52) exhibit clear elbow points at PC #3. However, a higher number of eigenvectors were incorporated, ranging from 6 to 20 per trajectory to include up to 95% of cumulative variance in each corresponding CEDA experiment. The objective was to test whether similarity among equivalent trajectories may be detected not only in terms of the essential subspace, but also of lower-variance eigenvectors.

The importance of PCs #1–3 was subsequently confirmed as they generated the most relevant clusters in all three CEDA experiments. In contrast, PCs with higher indices either formed minor clusters or remained strongly isolated with no significant similarity with others. Therefore, PCs #1–3 not only captured the essential subspace of the WT apo condition at the level of the individual trajectories, but also globally as a single conformational ensemble, forming CPCs of high coverage.

CPCs of lower coverage and higher PC indices were also found, although their retrieval depends on the criteria used to define consensus behavior. For instance, in the CEDA experiment involving the A and B domains of PKR, 7 of such CPCs (#4–10) were characterized in more detail by visually inspecting the captured collective motions and their estimated probability density distributions. The observed motions exhibited significantly lower amplitude compared to the major motions of CPCs #1–3, with unimodal density distributions mostly centered around zero (Figure 4.28). These characteristics suggest that these collective motions describe harmonic-like fluctuations around the conformation of the global average structure of the simulations [24], [379], [391]. These CPCs were not subsequently employed in the comparisons between alternative conditions as they describe minor consensus behavior (with coverage values around 20–35%). Moreover, in the additional CEDA experiment of the WT apo condition, designed to test replicability with a second trajectory ensemble, these collective motions did not robustly reemerge, as evidenced by cosine similarity calculations between both sets of CPCs (Figure 4.31). Still, this demonstrates that CEDA possesses the sensitivity to detect similarity between trajectories in terms of eigenvectors that would not be strictly classified within the essential subspace. This capability could be of interest for studies dedicated to the characterization of vibrational variations in single energy-minimum basins.

The distance measure between eigenvectors

Usually, clustering tasks progressively become more inaccurate as data dimensionality increases due to distance measures performing poorly (the so-called "curse of dimensionality") [540]. However, in certain contexts, specific properties of the dataset can allow for a meaningful usage of a particular measure, despite high dimensionality. In the case of eigenvectors, the topology of the dataset makes cosine distance a suitable measure. Importantly, trajectories must be subjected to equivalent structural superposition on a common reference structure prior to (Cartesian) PCA. Then, each trajectory adds a set of (orthogonal) eigenvectors to the pool for their clustering.

Eigenvector data points lie scattered on the surface of a $3N$-dimensional hypersphere with a unit radius, where N is the number of atoms of the analysis group. In such a scenario, similarity is optimally determined by the angle between eigenvectors, disregarding vector length. A given direction in this space corresponds to a particular collective variable constructed from the original variables (atomic Cartesian coordinates). Thus, homologous collective motions found in different trajectories will be represented as eigenvectors that adopt close directions. In fact, collinearity is the primary factor of similarity between eigenvectors, irrespective of direction; eigenvectors with opposite directions are equal, since they describe the same collective motion albeit reversed. This property enables expression of the cosine distance bounded in the interval [0, 1], where 0 indicates perfect collinearity and 1 indicates orthogonality.

Therefore, homologous eigenvectors among trajectories will tend to aggregate in close-to-collinear bundles that can potentially become a CPC when there is sufficient group similarity. Cosine distance manages to direct a meaningful clustering by detecting the occurrence of such data aggregations while marginalizing the otherwise scattered eigenvectors that do not correspond to consensus collective motions.

The clustering method

With respect to the choice of the clustering algorithm, agglomerative hierarchical clustering was selected because it facilitates an adequate control of the consensus behavior by first detecting the most similar individuals and then successively merging clusters. More specifically, the average-linkage algorithm was preferred to others because the resulting cophenetic distance preserved the scale of cosine distance, thus facilitating a direct interpretation of the dendrogram in the same range of values. Moreover, hierarchical clustering methods are agnostic to the employed distance measure and thus are compatible with cosine distance [503].

Other clustering methods besides agglomerative hierarchical clustering were tested, such as DBSCAN [541] and HDBSCAN [542]. However, they did not bring improvement in qualitative terms (data not shown), whereas they are less transparent due to the increase in the number of arbitrary hyperparameters. Nevertheless, the choice of the clustering method is open to discussion. Possible refinements could even involve combinations of methods, provided that the chosen strategy improves the quality of the obtained CPCs. Methods that require a predetermined number of clusters, such as k-means, are not recommended. This is because the pool of eigenvectors is potentially very heterogeneous, and not all eigenvectors need to be exhaustively classified as potential CPCs but only those that demonstrate a high degree of similarity and effectively represent common collective motions.

Determining criteria for consensus behavior

The extraction of CPCs is dependent on the selected cutoff values for cophenetic distance and coverage of the trajectories of the ensemble. The former parameter establishes the minimum group similarity that a cluster must hold for its eigenvectors to reflect the same collective motion. The latter parameter sets the minimum number of trajectories that must participate in the formation of a cluster for it to be considered representative of the whole ensemble.

The determination of consensus behavior based on these two parameters is subjective and relies on a balance between specificity and sensitivity. Given the unique characteristics of each system or structural region under investigation, it is impractical to suggest universal cutoff values for these criteria. Consequently, each experiment requires individual evaluation, with an initial focus on interpreting the information provided in the outcome of the clustering. With continued benchmarking across a variety of proteins or other macromolecular systems, it might be feasible to propose recommended ranges.

In this study, clusters with cophenetic distance values below 0.4–0.45 (depending on the experiment) and with a minimum coverage of 20% (*i.e.*, with members coming from at least 4 out of the 20 trajectories) were retained and processed as CPCs. Clusters that did not meet these criteria were discarded. As mentioned earlier, these criteria were selected to characterize various types of CPCs with both strong and mild consensus. Notably, the clusters with stronger consensus, which were later used in comparative analyses, started forming at cophenetic distance values below 0.1–0.2. They could have been procured with dendrogram cutoff points approximately at 0.3–0.35 without significant information loss. The adjustment of the cutoffs to the range of values 0.4–0.45 was implemented to maximize the size of these clusters without incorporating eigenvectors from repeated trajectories. Simultaneously, this range of values facilitated retrieval of other resulting clusters of lower coverage. These clusters are located in regions of the dendrograms that exhibit considerable variability and lack clear dendrogram cutoff points.

As demonstrated in this study, future CEDA experiments can adopt a similar rational approach to derive suitable CPCs. The first step involves selecting a tentative coverage value, which sets a preliminary standard for consensus quality based on the number of available trajectories. This criterion enables the evaluation of the feasibility of acquiring CPCs with such coverage, guided by the interpretation of dendrogram information. The complexity of this task depends on the observed degree of heterogeneity and the identification of clear dendrogram cutoff points. An iterative adjustment of the parameters may involve evaluation of the trade-off between enhancing sensitivity and specificity. Visualization of the potential collective motions and examination of the variance of the corresponding averaged vector components (*e.g.*, Figure 4.22) can further inform and validate this decision. In cases where there is no clear cophenetic distance cutoff or when outliers need to be discarded, a range of different values could be applied to specific areas of the dendrogram. Alternatively, if a section of the dendrogram exhibits too high variability, it may be disregarded. This flexible approach allows for the optimization of parameters to best adapt to the resolution of the dynamics of the system.

The interpretation of CPCs

The main source of dissimilarity between eigenvectors of a cluster is the presence of irregular local fluctuations that are captured together with the major homologous collective motions. Mobile regions such as termini tails or disordered loops undergo local fluctuations that may appear correlated in disparate orientations with a common and more essential collective motion. These have been identified as one of the main causes of the lack of convergence between equivalent trajectories [332], [447]. On the other hand, conformational transitions do not follow perfectly defined paths of motion. The same collective motion observed in different trajectories will exhibit variations in the orientation of the structure along the displacement. These variations also contribute to minor dissimilarities.

The CPC represents the average dynamical behavior of a cluster of eigenvectors. As such, CPCs accentuate the common qualities of the collective motions described by the members of the cluster. Consequently, the CPC provides the most representative (consensus) path of motion from the sampled variations of the same collective motion. In addition, the CPC achieves attenuation of the fraction of minor fluctuations scattered throughout the rest of the structure that may have appeared in correlation by chance. Thus, in essence, CPCs render denoised versions of the predominant collective motions that allow for sharper descriptions and comparisons of the relevant dynamics of a system while keeping the same interpretability as the regular PCs of EDA. The greater the number of homologous eigenvectors among trajectories, the more statistically robust and biologically meaningful is the resulting CPC.

Clear examples of the denoising effect of CPCs were demonstrated with the fluctuation of several flexible loops of the A domain. Due to their high dynamical activity, they appeared correlated with the found rigid-body motions of the B domain by chance. Upon derivation of the corresponding CPCs, it became clear that the A-domain loops had oscillated in divergent directions with lack of consensus when accompanying the larger displacements of the B domain and, therefore, their net fluctuations were almost none (Figures 4.22 and 4.23).

As similarity between eigenvectors gets moderately lower, it becomes less clear that they represent the same underlying collective motion. Mildly similar eigenvectors may still share a considerable fraction of collective atomic displacement in combination with other divergent components of motion. Whether these kinds of discrepancies should be interpreted as heterogeneous variations of a single collective motion or as entirely different motions with individual functional and mechanistic implications is dependent on the system of study and should be assessed on a case-by-case basis. For instance, CPCs #5 and #7 from the CEDA of the A and B domains of PKR were generated by different subsets of trajectories and exhibit qualitatively similar rigid-body motions of the B domain (Figure 4.21, subfigures e and g; Supplementary Videos S4.5 and S4.7). Their clusters would merge if the cophenetic distance cutoff was slightly increased from 0.4 to 0.42 (Figure 4.19).

An extreme instance of the aforementioned scenario arises when two eigenvectors from the same trajectory, which are orthogonal by definition, appear in close proximity within the dendrogram. This closeness could potentially result in their categorization as members of the same cluster, depending on the selected cutoffs. The interpretation of such instances is intricate, yet it is most plausible to infer that the generalized occurrence of these instances in the same experiment likely indicates that the selected cutoffs lead to an excessive increase in sensitivity, thereby compromising specificity. On the other hand, the presence of isolated instances might represent an accurate detection of a true positive.

The differences between combined-PCA and CEDA

Both combined-PCA and CEDA are strategies designed to extract consistent and reproducible dynamical properties from the average behavior of an ensemble of equivalent trajectories. These methods are based on the principle that independent simulations of the same system should be considered as a single conformational ensemble to yield statistically robust PCs. While combined-PCA relies on the application of a single PCA on the concatenated version of all trajectories, CEDA manages to integrate the output of all PCAs performed to each trajectory of the ensemble and derive a common set of consensus PCs. Although it has not been possible to include a comparative experiment between the two methods in this thesis, the differences in the nature of the PCs derived from each approach can be theoretically discussed. This discussion is informed by the interpretations of combined-PCA that are found in the literature [36], [109], [333], [365], [392], [443]–[446], which highlight certain biases inherent to this technique that CEDA potentially circumvents.

The limitations of combined-PCA primarily stem from two factors: i) the high sensitivity of PCA to outliers [379], [543], and ii) the emergence of “static modes” in the analysis of trajectories that explore different regions of the conformational space.

The first limitation concerns the fact that the eigenvectors of combined-PCA will be strongly influenced by the trajectories with the most deviated behavior with respect to the global average structure. Consequently, in scenarios where outliers are present (*e.g.*, a heterogeneous trajectory among a group of similar trajectories), the resulting PCs may be skewed towards these specific conformational outliers, thus obscuring the true predominant behavior of the ensemble. Due to the orthogonality of eigenvectors, any skewness in the first eigenvectors can potentially distort the subsequent eigenvectors [332], [395].

The second limitation manifests when the variance between average structures across trajectories exceeds the dispersion within individual trajectories. Or, in other words, when the conformational subspaces explored are markedly divergent. In such cases, the first PCs will represent the structural difference between the separately sampled subspaces rather than actual collective motions. The information contained in such PCs, referred to as “static modes”, is not much more informative than that obtained through simpler methods than PCA. Static modes cannot be reliably extrapolated to actual motion paths, as they have not been directly observed in simulations. The rest of the PCs may still afford actual dynamical information albeit with unclear precise meaning.

In practice, a given ensemble of equivalent trajectories can exhibit varying degrees of conformational heterogeneity within a broad spectrum of possibilities. Within the approach of combined-PCA, determining the occurrence of the aforementioned problems can be a complex task. Conversely, CEDA not only circumvents the emergence of static modes but also offers a more transparent approach that facilitates the precise interpretation of different potential scenarios. Subsets of trajectories exhibiting more heterogeneous or outlier behaviors will be more efficiently identified, due to either their absence in predominant clusters or the formation of distinct smaller clusters representing minority behaviors.

<u>**Comparing alternative conditions with CEDA**</u>

As reviewed in section 1.4.5 of the Introduction of this thesis, the comparative analysis of trajectory ensembles is a multifaceted challenge. It requires a combination of diverse analytical methods, metrics, and innovative data representations to express similarities and differences between the respective conformational ensembles. Currently, no single analytical method can encapsulate all the different perspectives from which this problem may be examined. Various strategies have been developed, each focusing on specific data properties, depending on the specific scientific inquiry being addressed.

Within the context of the CEDA framework, we attempt to answer the following questions: 1) Given a collective motion that is characteristic of a system under a specific reference condition, does this motion also manifest in an alternative condition? 2) If so, are the conformational distributions for this collective motion consistent across both conditions? Thus, the application of CEDA primarily depends on the extraction of meaningful CPCs from a trajectory ensemble of the reference condition. This enables a subsequent comparative analysis between conditions by examining the trajectory ensemble of the alternative condition in terms of the information provided by the reference CPCs.

The outcomes of such comparisons may be interpreted using different data analysis techniques and graphical representations. In this thesis, the emphasis is on comparing the resulting conformational distributions, based on the estimation of the underlying probability density distribution of each ensemble and the subsequent application of similarity metrics between statistical distributions. This approach is consistent with other methodologies in this scientific field aimed at characterizing biologically significant conformational changes associated with phenomena such as ligand binding, enzyme catalysis, and allosteric responses [249], [432].

Upon projecting the trajectory data onto the relevant CPCs and generating the corresponding density distributions, the comparative assessment can be conducted both in qualitative and quantitative terms. From a qualitative perspective, one may describe the main features that are visually discernible from the plotted data to infer differential dynamical behavior. For instance, if the distributions share a common span of values along a given CPC, this indicates that both conditions sampled a comparable region of the conformational space. Consequently, the corresponding collective motion is characteristic of the system under both conditions. Conversely, if the distributions span different ranges of values, this suggests a potential difference in the types of conformations favored under each condition. When the sampled conformational spaces coincide, assessing the consistency of proportions across the spectrum of CPC values yields valuable information. These observations, together with the characterization of the collective motions in the structural context, facilitate the study of the potential functional implications of the identified conformational diversity.

On the other hand, in this study, three simple metrics have been proposed to provide quantitative scores for such assessments. The interpretation of each metric is straightforward and can be directly correlated with the insights gained through visual inspection of the distributions to achieve a comprehensive analysis. Alternatively, when the number of comparative analyses is substantial, these metrics facilitate extracting the main information from each assessment and compiling an overview of the results in the form of rankings, as demonstrated in the comparative study between WT and mutant variants of PKR (section 4.2.3). However, the experience from this study suggests that summarizing the array of potential differences that the distributions may exhibit into three general scores is not a trivial task. Consequently, there exists an opportunity to develop more specialized metrics.

First, the metric termed *overlap* serves as a preliminary reference to assess whether the alternative condition operates inside or outside the same conformational space as the reference condition. A low overlap value suggests that the alternative condition exhibits distinct conformations. Secondly, the metric termed *coverage* provides insights into the extent of reference conformational space that the alternative condition covered. For instance, a scenario where the alternative condition maintains high overlap yet displays reduced coverage suggests increased structural rigidity relative to the reference condition. Lastly, the Bhattacharyya coefficient (BC) quantifies the consistency of the proportions between both distributions, within the span of the reference conformational space. As applied in this study, the BC is dependent on the quality of the coverage. High coverage implies that variations in the BC are mainly due to alterations in relative proportions or shifts in local maxima between the distributions. Mild coverage automatically correlates with reduced BC due to the partial lack of correspondence in spatial regions. The combination of both incomplete coverage and distribution shifts results in a further decrease in the BC.

In prospective applications of CEDA, additional complementary techniques could be integrated to uncover specific properties or characteristics that are not discernible through density distribution comparison. For instance, network analysis metrics could be employed to identify specific amino-acid interactions [439]. Alternatively, other time-dependent properties could be measured, such as the rate of conformational changes or their probability of co-occurrence across different regions of the macromolecular structure.

5.3 Structure, dynamics, and function in PKR

The study of PKR through the analysis of MD simulations has provided new support for several of the proposed conformational changes that are associated with the transition between the inactive and active states of the enzyme. Despite the multiple consistent descriptions of the conformations adopted by PKs with different ligands, derived from numerous comparative studies of its crystal structures, the characterization of the mechanism of the enzyme remains incomplete without the validation of the dynamical events leading to the observed differences between static structures. This study identified and captured conformational transitions of PKR that are pivotal in the function of the enzyme. The implementation of studies that utilize dynamics-derived data, such as this one and the one of Naithani *et al.* [164], is enabling a more comprehensive characterization of the functional dynamical behavior of the structure.

The Results chapter of this thesis includes annotations regarding insights obtained from each experiment conducted. The purpose of this section is to provide a structured summary of the contributions made by this study to our understanding of this protein, viewed through the lens of the structure-dynamics-function paradigm, and in alignment with observations reported in previous studies. In addition to this summary, future analyses are proposed to further investigate the findings of this study and provide a more detailed understanding of the involved structural rearrangements.

The most relevant observations were obtained primarily through the application of CEDA, which facilitated the identification of distinct collective motions and enabled the detection of conformational diversity among the various states of the protein-ligand complex. The stability (RMSD) and flexibility (RMSF) analyses of the trajectories provided complementary information.

Given the modular structure of PKR, each CEDA experiment was designed to focus on a specific combination of functional domains. The analyzed regions were selected to characterize the structural reorganizations that have an impact on the active site of the enzyme, both its local environment and the regions interconnecting its instances across the tetramer.

Thus, firstly, the analysis of the A and B domains facilitated the exploration of the dynamical behavior of the active site and the influence of cofactor and substrate binding on its local conformation. Secondly, the analysis of the A-A' pair of domains provided insights into the relative motions between subunits across the A-A' interface, which constitutes the most direct communication path between active sites. Finally, the analysis of the AC-C'A' pair of cores facilitated the inspection of conformational changes of larger scale in the enzyme, which are also directly coupled with the functional role of the allosteric site and the local reorganizations of the C-C' interface.

In all these analyses, the A domain served as a fitting group. Given the high rigidity characteristic of the TIM barrel core [134], the A domain provided an optimal reference region, enabling the alignment of all structures with equivalent spatial orientations. Furthermore, this approach facilitated the systematic description of the array of relative motions of the other domains with respect to the A domain of one subunit.

<u>The dynamical behavior of the B domain</u>

The B domain demonstrated marked mobility, being the major source of structural diversity along the trajectories of the apo condition (Figure 4.4). This domain fluctuates mainly as a rigid body along diverse directions with respect to the A domain, thanks to the flexible capabilities of the linker fragment between these two domains that serves as a hinge mechanism. This type of hinge regions are common in multidomain proteins and characteristically enable sampling of a large ensemble of heterogeneous conformations with low transition barriers between the states [35]. It has been suggested that hinge-bending motions of domains can take place at relatively fast timescales, on the nanosecond-microsecond time scale [261], consistent with the behavior of the B domains in the simulations of this study.

The conformational variability of the B domain in the apo condition manifested with very diverse amplitudes of motion, not only between trajectory replicates but also between the subunits of the same tetramer. The asymmetrical dynamical behavior of the B domains was consistently shown by comparisons of the corresponding RMSD and RMSF profiles (Figures 4.4 and 4.6), as well as by the apparent diversity in the range of conformational spectra generated along the most relevant CPCs of the corresponding CEDA experiment (Figures 4.24, 4.26, 4.27). These observations reinforce the suggestions made in previous studies [121], [132], [138], [164].

Recently, a study of type I PK of *Escherichia coli* revealed that the removal of the B domain produces a protein that achieves regular folding of the rest of the structure, retains a low level of catalytic activity with a reduced binding affinity for PEP, and retains allosteric activation [544]. The study concluded with the hypothesis that the insertion of the B domain was favored by natural selection to optimize the enzymatic capabilities of a PK ancestor devoid of B domain. Consistent with this idea, the absence of symmetrical behavior in the motions of the B domain among the subunits of the tetramer suggests that this region is not dynamically coupled with other regions of the structure. Rather, its role seems to be entirely related to the optimization of the local configuration of the active site, assisting in cofactor/substrate binding and enhancing catalytic activity.

The conformational change of the B domain that is more relevant from this point of view is the transition between open and closed forms, with the B domain acting as a lid to cover or uncover the active site at the top of the A domain. The studies of diverse PK isoenzymes show that this conformational change is correlated with the absence or presence of ligands at the active site. The more open forms tend to be found in ligand-unbound structures, whereas partially and fully closed conformations are favored upon the binding of PEP and ADP, respectively [121], [124], [135], [137], [139], [140], [147], [160].

The collective motion consistent with this transition, called the opening/closing motion, was captured in CPC #1 (Figure 4.21a, Supplementary Video S4.1). This CPC was derived from the trajectory ensemble in the apo condition, but also exhibited total correspondence with the first CPC of the trajectory ensembles of each holo condition (Figure 4.37). Therefore, results suggest that the interchange between open and closed forms of the B domain is a dynamical phenomenon inherent in the structure of the protein, both in the absence and the presence of ligands at the active site. The conformational profiles along CPC #1 show two distinctive regions of the spectrum across apo and holo conditions that correspond to the open and closed conformations (Figure 4.34, top). This scenario is consistent with the existence of a conformational equilibrium whereby these states represent local minima in the FEL of the structure. Ligand binding follows the population shift model: in the apo condition the open forms are energetically more favorable, whereas in holo conditions the closed forms are stabilized.

Transitions in the apo condition predominantly occurred from a closed form (characteristic of the initial crystallographic structure 2VGB) to the open forms. However, importantly, the open-to-closed transition was also detected in a few trajectories, confirming that the apo protein is able to oscillate between both forms within the time span of these simulations. Extensions of this study should be aimed at further validating this statement by providing new trajectory replicates in the apo condition starting from diverse initial conformations.

In contrast, all holo conditions primarily retained the closed conformation of the B domain. The fact that simulations with bound K^+ and Mg^{2+} stabilized the closed conformation suggests that the cofactors provide the proper conformation to receive the substrates and catalyze the reaction more efficiently. Studies have shown that multi-substrate enzymes that feature domain closure motions may utilize multiple fast cycles of domain opening and closing to enable reorganization of ligand binding and reach optimal configuration for the reaction [261]. Sampling of the more open conformations of the B domain was no longer allowed in the simulations with bound PEP. Such behavior was also demonstrated via RMSD and RMSF measurements, which indicate a progressive decrease of mobility of the B domain as more ligands were bound to the active site, with PEP exerting the strongest effect (Figures 4.5 and 4.9). Finally, the local environment of the active site increased in rigidity in holo conditions, as expected from the restraints imposed by ligand binding (Figures 4.11 and 4.12).

The concerted collective motions of the tetrameric core

The main body of the PKR tetramer, composed of the four AC cores, is a large and compact structure that exhibits markedly lower mobility than the B domains. Each domain, A or C, interacts with two other adjacent domains through two interfaces: one intramolecularly with a domain of the other kind (the A-C interface within the AC core itself), and one intermolecularly with a domain of the same type (the A-A′ and C-C′ interfaces between subunits). Given this interconnected architecture, structural rearrangements between any pair of domains propagate to the rest of the structure.

CEDA experiments were conducted independently on the A-A′ pair of domains and the AC-C′A′ pair of cores to facilitate the decomposition of the dynamical information of the assembly and detect the underlying collective motions between the domains in an isolated manner. This approach enabled identification of dynamical events that best represent the differences across apo and holo conditions. The corresponding collective motions were captured in the second CPC of each CEDA experiment. Relative to the A domain, the A′ domain undergoes a seesaw-like swinging motion, pivoting about the contact point of the A-A′ interface (Figure 4.42b, Supplementary Video S4.12). On the other hand, the C-C′A′ block undergoes a swinging motion along the plane of the tetramer, facilitated by the joint-like capabilities of the A-C interface (Figure 4.56b, Supplementary Video S4.20).

A dynamical coupling between these two collective motions was identified on the basis of two characteristic features exhibited by both conformational profiles along their respective CPCs (Figures 4.49 and 4.62, central panels). Firstly, they displayed the same differential patterns across apo and holo conditions, in qualitative terms. Secondly, the main factor contributing to the observed differences in the sampling was the presence or absence of PEP at the active site. PEP-bound simulations were predominantly constrained within a conformational space near the initial crystallographic structure 2VGB (active state), whereas PEP-unbound simulations generally left the initial conformation and adopted a range of conformations around the center of the spectra. The FBP-holo condition (a PEP-unbound condition) also displayed distinctive behavior that will be discussed later in this section.

The observed correspondence between the conformational profiles suggests that these two collective motions potentially occur in a concerted manner, being complementary perspectives of a single transition. Or, at least, the probability of observing the conformational change at one site correlates with the probability of observing it at the other site. As described in section 4.1.3.2.3, the overlay of both collective motions (Supplementary Video S4.30) revealed that the joint structural rearrangements are in alignment with the main descriptions of the conformational changes between the T (inactive) and R (active) states in PKs [124], [127], [128], [132], [147]–[149], [159], [162]–[165], thus establishing consistency between crystallographic and dynamical data.

Given that the initial structure of the simulations is in the active state (with bound PEP), the observed transition in PEP-unbound conditions can be interpreted mainly as a relaxation of the structure towards the inactive conformation. However, PEP-bound simulations did explore the conformational space characteristic of the inactive conformations in minor proportions, which suggests that both states are in equilibrium and that the binding of PEP is consistent with a population shift model. Taking into account the observations from both this study and the available literature, the R-to-T transition can be described with the following sequence of events.

The release of PEP results in the loosening of interactions within the active site, thus allowing the cavity to widen. The Aα6′ helix, which contains the PEP-binding residues Gly338 and Asp339, now is able to slide away from the inner part of the active-site cleft. This local rearrangement disrupts an R-state–stabilizing intersubunit interaction between these two residues and Arg385 of the adjacent subunit, which is located at the top of Aα7. The loss of this interaction makes this α-helix recede from Aα6′ and favor a T-state–stabilizing interaction between Arg385 and Asp390 between opposing subunits. These rearrangements at the A-A′ interface occur symmetrically between both instances of Aα6′ and Aα7. Consequently, the active sites between the adjacent A domains become less entangled with each other, which results in their mutual separation and a subsequent rotation of the A domains relative to each other [128], [132], [147], [149], [159], [162]–[164]. The dynamical events associated to this process were captured in CPC #2 of the A-A′ pair of domains: the relative rotation between adjacent A domains (Figure 4.66 and Supplementary Video S4.29) and the correlated fluctuation of the width of the active-site cleft involving Aα6′ and Aα7 and their neighboring components (Figure 4.43b and Supplementary Video S4.17).

Importantly, these structural rearrangements are currently considered to be the basis of the homotropic cooperativity mechanism in PK. The binding event of PEP to one active site triggers the reverse process whereby A-domain rotations induce the narrowing of the adjacent (unoccupied) active site which becomes more affine to PEP binding.

The rotational motion of the A domain in the R-to-T transition is transmitted to the C domain of the same subunit. This event, in turn, reaches the C-C′ interface and induces local rearrangements of its elements, with corresponding changes in the intersubunit interactions. In the R state, the region close to the center of the tetramer establishes tight interactions across the C-C′ interface. Specifically, Lys465 at Cα2 extends towards the Cα1+Cα2+Cα3 region of the adjacent subunit and interacts with Pro446, Glu453, and Tyr487. These interactions break or weaken in the transition to the T state due to the widening of the gap at this section of the C-C′ interface. Conversely, the gap at the opposite section of the interface narrows and gains T-state–stabilizing intersubunit interactions. Provided that FBP is not bound to the allosteric site, Trp558 and Arg559 located at the L-Cβ4-Cβ5 loop (FBP binding loop) interact with Asp530 (Cα5) and/or Arg569 (Cβ5) of the adjacent subunit. Interactions between the adjacent Cβ5 strands are preserved in the transition. Finally, the rearrangements of the C-C′ interface potentially induce the symmetrical rotational motions in the adjacent subunit, completing the mechanism of transmission of dynamical information across the tetramer [124], [128], [132], [148], [149], [159], [162], [165]. The dynamical events associated to this process were captured in CPC #2 of the AC-C′A′ pair of cores (Figure 4.65 and Supplementary Video S4.28).

The observed dynamical events in this study align with the interpretation of the “domain-rotation” model [132] on the conformational change between the T and R states (Figure 1.12). This is because two separate pivot points have been detected: one at Cα4 that allows the A and C domains of the same subunit to rotate relative to each other, and another at the C-C′ interface, approximately where the adjacent Cα2 helices interact, that allows the subunits of the C-C′ dimer to rotate relative to each other.

Insights on the allosteric activation by FBP

Importantly, the described structural rearrangements not only imply that dynamical information is transmitted among the active sites of the tetramer, but also that the process is dynamically coupled with the local conformation of the allosteric site. Notably, in the collective motion of the AC-C'A' pair of cores associated with the R-to-T transition in the apo simulations, the L-Cβ4-Cβ5 loops of both subunits exhibit an unfolding motion correlated with the major rearrangements (Figure 4.65 and Supplementary Video S4.28). This suggests that upon removal of the FBP molecule from the initial structure, the loop relaxed in simulation. This behavior is consistent with other studies that show that, in the absence of FBP, this loop tends to be more stabilized in the open position [165].

Conversely, in the FBP-holo condition, simulations were performed with FBP bound to the allosteric site, therefore preventing the unfolding of these loops in the structure. Moreover, these simulations exhibited characteristic sampling of intermediate conformations between those of the PEP-bound (active) and PEP-unbound (inactive) conditions, along the CPCs that captured the R-to-T transition (Figures 4.49 and 4.62, central panels). While such effects are indicative of a differential dynamical behavior induced by the presence of FBP in the allosteric site, it is noteworthy that the active conformation was not fully retained.

Considering this, do these observations account for sufficient demonstration of the allosteric effect? The classical MWC model of allostery [74] would have certainly required retention of the "canonical" active conformation. However, the current view of allostery, through the lens of the ensemble allosteric model (EAM), understands the phenomenon from the perspective of the properties of the full native conformational ensemble of the system [38]. Accordingly, determining the conformational heterogeneity of the system and detecting a redistribution of the probabilities of the entire ensemble of states upon the allosteric perturbation is more meaningful than identifying a single active conformation [44]. The allosteric perturbation (*i.e.*, the binding of FBP) reshaped the energy landscape, stabilizing an intermediate state to the detriment of the inactive state. This phenomenon may consequently suffice to lower the energy barrier associated with the transition to the active state, thus increasing the statistical probability of finding instances of the active site with higher affinity to PEP binding.

In terms of the amino-acid networks described in the literature [128], [148], [149], [165], with the L-Cβ4-Cβ5 loop in a closed conformation locking FBP at the allosteric site, the characteristic T-state–stabilizing interactions between the residues of this loop and Cα5 and Cβ5 of the adjacent subunit cannot be properly established. Therefore, the "canonical" T state is not stabilized either, potentially leaving the structure in an intermediate state whereby the R-state–stabilizing interactions may be achieved with lower energy barriers.

Possible expansions of the dynamical study of PKR

On the basis of the insight provided by this study, subsequent analyses may be proposed to achieve a more comprehensive characterization of the enzymatic mechanism.

Firstly, this study did not include a scan of the most characteristic amino-acid interactions for each specific state of the conformational spectrum. Such research is crucial to validate the descriptions of interactions reported by other studies. A potential strategy could involve extracting samples of frames from multiple trajectories that explore the same points of the conformational spectrum. This would enable the verification of the consistency of interaction networks across trajectories. By conducting a thorough examination of various points of the conformational spectrum, it may be possible to obtain a description of the dynamical changes in the interaction network throughout each collective motion.

Alternatively, the application of methods derived from network theory or Markov state models could be employed to achieve similar objectives.

Secondly, it would be beneficial to expand the collection of trajectories by incorporating conditions with asymmetrical ligand binding between the subunits of the tetramer. For instance, it would be interesting to investigate whether a tetramer with PEP bound only to the active site of one subunit maintains the active tetrameric conformation with the same consistency as the PEP-bound conditions examined in this study. A similar experimental design could be implemented with FBP. Additionally, the dynamical behavior of PKR structures with other known allosteric modulators, such as alanine (which should favor the T state), or allosteric drugs like mitapivat (which should favor the R state), could be explored.

Lastly, future analyses could aim to examine and compare the transition events of collective motions in a time-resolved manner. That is, ascertaining whether the conformational changes captured by a single or multiple CPCs occur concurrently or not in a given trajectory or across trajectories. While this type of analysis was not needed to identify the asymmetrical behavior of the B domains within the tetramer, it could further validate the nature of the concerted collective motions between the A-A′ and C-C′ interfaces, or even detect other underlying orchestrations between motions.

5.4 Dynamical alterations of the analyzed PKR variants

Despite the extensive array of known sequence variants of the *PKLR* gene, only a small subset holds a fully validated pathogenicity status. This validation is typically achieved through experimental assays that are subsequently uploaded in repositories such as ClinVar, or when robust evidence is accumulated from multiple clinical cases. For most variants identified in patients with pyruvate kinase deficiency (PKD), only sequence information is available. Only in cases where distinct symptomatology is manifested, variants can be potentially categorized with a particular degree of severity.

However, since performing subsequent functional assays in a systematic way is impractical due to cost and time restrictions, these variants often remain unanalyzed at the molecular level. Some specialized studies have provided insights into the implications of missense variants in the context of the structure of PKR. For instance, Valentini *et al*. and Zanella *et al*. [118], [121] observed that several of the known pathogenic variants cluster in specific regions of the structure that are critical to the function and stability of the enzyme. These regions include: i) the hydrophobic core of the A domain, ii) the α-helices of the A-A′ interface, iii) the A-C interface, and iv) the FBP binding site. Similarly, Pendergrass *et al*. [213] highlighted that the initial fragment of the C domain (Cα1 to mid Cα2) is particularly devoid of identified PKD mutations, aligning with the now recognized functional significance of this region in establishing R-state–stabilizing interactions at the central region of the tetramer.

Modeling and simulating variants of PKR

In general, the knowledge that can be derived solely from the information of sequence or static structures is insufficient to assist the clinical community in the colossal task of validating the interpretation of genomic data. The implementation of computational methods is a critical necessity for this endeavor. The standard use of dynamical data to assess the functional impact of protein variants represents one of the next frontiers in the field of pathogenicity prediction. As outlined in section 1.3.6 of the Introduction chapter, when expression and kinetic assays cannot be implemented, or when insights derived from structural information fail to provide clear functional implications, dynamics-based methods such as MD may offer a more robust approach to uncover anomalies in molecular behavior.

Simulating missense variants commonly involves modeling the corresponding amino-acid substitution from a WT structure. This strategy facilitates the use of mutant models without depending on experimental methods. However, it is important to note that the resulting model is an approximation. In the absence of experimental validation, it is impossible to verify that the new sequence achieves a folding equivalent to the WT protein *in vivo*.

If there are indications suggesting that a specific mutation could significantly impact folding or the machinery of gene expression/transcription, it is not recommended to proceed with the modeling of such a mutation. For example, in this study, the Arg479His mutation was excluded as a simulation candidate, despite its known severity and particularly high allelic frequency, especially within the Amish community [212]. This mutation affects a position at a splicing site of *PKLR*, resulting in significantly reduced transcript levels. Although the corresponding mutated protein is still expressed, the study of the molecular perturbation of the amino-acid substitution is not of interest, as the insufficient protein product is the primary cause of PKD [121], [545], [546].

In contrast, the determination of crystallographic structures of other variants, such as Thr384Met or Arg486Trp [121], has confirmed that the corresponding proteins achieve a folding equivalent to the WT protein, thus validating the use of *in silico* models for these mutations.

These models of missense variants can serve as the initial structure for MD simulations to compare their dynamical behavior with that of the WT protein. This strategy has been referred to as “relaxation MD” [175], since it describes the relaxation of the native structure of the protein upon introduction of a mutation. While the pertinence of this approach may be debatable, numerous studies have demonstrated good correlations between their predictions and experimental data, and have been able to extract meaningful insights into the mechanisms through which mutations impair protein function [175].

An additional factor to consider in the simulation of PKR variants aimed at characterizing pathogenicity is ascertaining whether the variant under study was found in homozygous, heterozygous, or compound heterozygous genotypes. For instance, two missense variants in compound heterozygosity produce two types of monomer in the cellular environment that can potentially heterotetramerize. The properties exhibited between the resulting heterotetramers of different combinations can vary significantly. Therefore, homotetrameric models of a specific mutation only account for one of the extreme scenarios of this phenomenon, and the derived insight will only be applicable to cases of homozygosity for that variant.

General assessment of the detected dynamical alterations

The construction of a catalog of the known missense variants of PKR and the subsequent selection of a representative subset have facilitated the generation of an extensive collection of system simulations. This task has in turn provided the chance to test the analytical capabilities of the CEDA framework with a comprehensive set of comparisons between WT and mutant variants.

The identification of dynamical alterations in the trajectories of PKR variants has been conducted within the context of the insights targeted by the CEDA framework that were discussed in section 5.2. Specifically, the comparative assessments focused on determining whether the target condition (mutant) explores the conformational space in a manner similar to the reference condition (WT) in terms of the paths of collective motions of the latter.

Overall, the simulated variants exhibited modest signs of dynamical alteration, with differences less pronounced than those observed in the comparative assessments between the WT apo *vs*. holo conditions. Nevertheless, the assessment enabled identification of variants that exhibit distinctively altered behaviors. The protein regions with more instances of dynamical alteration were predominantly domain interfaces, either intrasubunit or intersubunit.

In some instances, anomalies have been detected in the conformational profile of variants of known pathological phenotype. These results serve as a support to interpret the pathogenicity of these variants in terms of the disruption of dynamical behavior. Some examples are Gln505Glu [546], Ser120Phe [547], [548], Ala115Pro [547], Arg163Cys [521], Gly263Trp [549], or Arg510Gln, the most commonly reported variant in United States and Northern/Central Europe and known for its severe clinical phenotype [212], [546], [550]. In contrast, the variant Arg486Trp, despite being a notorious pathogenic variant [546], [550], did not show clear alterations.

In some instances of variants affecting the same positions, such as Gly263Trp *vs*. Gly263Ala and Gln505Glu *vs*. Gln505Arg, the pathogenic variants exhibited altered behavior whereas the potentially neutral variants did not. However, in general, there was no clear distinction between altered and unaltered variants that coincides with the classification between damaging and neutral variants according to the available annotation resources. The datasets of neutral variants are generally retrieved from data portals of the large-scale sequencing projects, typically disregarding entries with less than a minimum number of observations in different studies, and/or with allele frequencies lower than 1% [173], [551]. In this project, such filters could not be considered since the gathered variants were mostly found in only a few individuals. Therefore, they should be regarded as “putatively neutral” [204].

Such a scenario is nevertheless interesting, since the application of this analysis facilitated the characterization of variants that have been observed in the human population and for which no details are known other than sequence differences. For instance, variant Pro521Ser was retrieved from the ExAC and gnomAD databases and yet exhibited significant alteration in terms of the analyzed dynamical properties. Interestingly, results suggest that, in the absence of FBP, it displays a conformational profile comparable to that of the WT enzyme in the FBP-holo condition (Figures 4.63 and 4.81e). Therefore, this variant might not be truly neutral, even though it did not sample abnormal conformations.

The case of Val506Ile is similar, being one of the variants with the highest allele frequency in databases such as ExAC/gnomAD. It has been classified as a “polymorphism” or “likely benign” in repositories like Humsavar. Despite this classification, there exist several instances of reported pathogenicity in the scientific literature [213], [552]. The analyses conducted in this study have identified alterations in the conformational profile that are similar to those observed for the Pro521Ser variant (Figure 4.78).

The contribution of this study to the field of pathogenicity prediction

The report on the detected dynamical alterations of the examined PKR variants offers a classification that can be used in a complementary manner with other sources of pathogenicity assessment. The results were presented in the form of rankings, with the variant exhibiting the most significant differential behavior at the top, and the one with the least at the bottom.

For the comparative assessments between wild-type (WT) and mutant systems in the apo condition, the quantitative indicator that best differentiated the behavior of the maximum number of variants was selected. This strategy provided the most meaningful comparisons with respect to the reference values of the intrinsic sampling variability of the WT apo condition, in each of the examined regions of the enzyme (Figures 4.73, 4.76, and 4.79). For the comparative assessments in holo conditions, the sum of the three quantitative scores was considered. These approaches leveraged the trajectory similarity metrics proposed in this implementation of CEDA, emphasizing variants with potentially significant dynamical alterations.

However, as discussed in section 5.2, there is potential for identifying more specialized metrics. The diverse range of possible dynamical alterations may not be captured solely by the systematic application of a single score. For instance, the markedly low overlap score of the variant Ala115Pro was obscured by its high coverage score in the comparative analysis of the A-A′ pair of domains in the apo condition (Figure 4.77). Finding an appropriate balance in the specificity of the metrics is challenging. It is crucial to develop metrics that are sufficiently generic to analyze large sets of mutations of diverse nature (*e.g.*, location in the structure, functional role, degree of evolutionary conservation) and yet specific enough to extract the most meaningful information from each variant.

A logical progression for this study involves the development of a predictive tool specifically designed for PKR. This tool would incorporate the scores derived from this study (with the possible implementation of new iterations to refine the methodology). Unlike general-purpose predictive tools, this class of tools is designed to evaluate the pathogenicity of variants in the context of a specific gene, protein, or protein family. The precision of these tools is progressively improved by integrating the available experimental data from the specific system under study. This process facilitates the training of a predictive model that becomes increasingly reliable over time [216], [233].

In a recent development, the research group led by Fenton and associates has introduced a comprehensive database, known as the PYK-SubstitutionOME [523]. This database encompasses the biochemical characterizations of more than 1000 variants in PK isoenzymes. The database collates data from a diverse range of studies and aims to set rigorous benchmarks for testing protein predictions and improving the continuous advancement of prediction algorithms. The research conducted in this thesis could potentially form highly advantageous synergies with such initiatives. The ultimate objective is to persist in the integration of information extracted from various experiments, thereby amplifying the potential contributions to the field of pathogenicity prediction.

Chapter 6 Conclusions

1. A comprehensive set of molecular dynamics (MD) simulations of human erythrocyte pyruvate kinase (PKR) has been procured. The collection of trajectories comprises multiple conditions of the enzymatic complex with its natural ligands, as well as a large array of human genomic missense variants of the protein.

2. The insights from the available crystallographic data of PKR have been successfully integrated in the MD simulations of the holoenzyme complex by means of a parametrization strategy, revalidating the approach of the cluster model as a direct and versatile method to parameterize coordination complexes in metalloproteins for subsequent MD simulations. Accordingly, the practical application of the methodology in this thesis may serve as a reference for setting up other macromolecules with similar complexity and configuration.

3. The Consensus Essential Dynamics Analysis (CEDA) strategy was developed. CEDA manages to integrate the output from the Principal Components Analysis (PCA) of independent MD simulations of a macromolecular system in a unified framework. The derived set of collective variables, called the Consensus Principal Components (CPCs), accentuate the common qualities of the collective displacements displayed by a trajectory ensemble and attenuate the representation of sporadic minor variations.

4. In the framework of CEDA, the underlying similarities and differences between trajectory ensembles can be evaluated by comparison of the conformational profiles in terms of a single reference set of CPCs. Three simple metrics, namely, the overlap, the coverage and the Bhattacharyya coefficient, have been proposed to provide quantitative scores for such comparative assessments.

5. Distinctive consensus motions of the protein domains and functional sites of PKR were identified, providing deeper insights into the structure-dynamics-function relationship of the enzyme. The findings of this study are coherent and provide further mechanistic insight for several of the conformational transitions that had been previously proposed via crystallographic studies of pyruvate kinase isoenzymes.

6. The identified concerted collective motions of the tetrameric core of PKR are consistent with the classical descriptions of the transition between the inactive and active states of this protein and provide a mechanistic interpretation of the allosteric phenomenon that is compatible with the current general descriptions of allostery in terms of the ensemble allosteric model (EAM).

7. The dynamical activity of the B domain of PKR seems to be entirely related to the optimization of the local configuration of the active site, not being dynamically coupled with other regions of the structure.

8. A comprehensive database of missense mutations of PKR was constructed by integrating data from the available literature as well as from a wide range of public databases of human genetic variants. The analysis with CEDA has enabled detection of altered dynamical behavior in

variants either with a previously validated pathogenic status or for which no functional details were known in this regard.

9. The report on the detected PKR variants that exhibit dynamical alterations offers a classification that can complement other sources of pathogenicity assessment. The assessment of the functional consequences of protein variants with dynamics-based approaches like CEDA affords valuable insight that is complementary to that obtainable by other methods and has the potential to improve specific predictive models.

References

[1] R. H. Garrett and C. M. Grisham, *Biochemistry*. 2017.

[2] C. Tanford and J. Reynolds, *Nature's robots: a history of proteins*. OUP Oxford, 2003.

[3] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 19, pp. 6679–6685, 2005, doi: 10.1073/pnas.0408930102.

[4] K. Henzler-Wildman and D. Kern, "Dynamic personalities of proteins," *Nature*, vol. 450, no. 7172, pp. 964–972, 2007, doi: 10.1038/nature06522.

[5] M. Orozco, "A theoretical view of protein dynamics," *Chem. Soc. Rev.*, vol. 43, no. 14, pp. 5051–5066, 2014, doi: 10.1039/c3cs60474h.

[6] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry (8th edition)*. 2021.

[7] E. Medina, D. R. Latham, and H. Sanabria, "Unraveling protein's structural dynamics: from configurational dynamics to ensemble switching guides functional mesoscale assemblies," *Curr. Opin. Struct. Biol.*, vol. 66, pp. 129–138, 2021, doi: 10.1016/j.sbi.2020.10.016.

[8] R. G. Smock and L. M. Gierasch, "Sending signals dynamically," *Science (80-.).*, vol. 324, no. 5924, pp. 198–203, 2009, doi: 10.1126/science.1169377.

[9] C. Micheletti, "Comparing proteins by their internal dynamics: Exploring structure-function relationships beyond static structural alignments," *Phys. Life Rev.*, vol. 10, no. 1, pp. 1–26, 2013, doi: 10.1016/j.plrev.2012.10.009.

[10] P. E. Bourne and J. Gu, *Structural Bioinformatics*, 2nd ed. 2011.

[11] X. Liu *et al.*, "Structural Insights into the Process of GPCR-G Protein Complex Formation," *Cell*, vol. 177, no. 5, pp. 1243–1251, 2019, doi: 10.1016/j.cell.2019.04.021.

[12] W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," *J. Mol. Graph.*, vol. 14, no. 1, pp. 33–38, 1996, doi: 10.1016/0263-7855(96)00018-5.

[13] S. Zhang, H. Li, J. M. Krieger, I. Bahar, and B. Ozkan, "Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity," *Mol. Biol. Evol.*, vol. 36, no. 9, pp. 2053–2068, 2019, doi: 10.1093/molbev/msz102.

[14] N. Dawson, I. Sillitoe, R. L. Marsden, and C. A. Orengo, "The Classification of Protein Domains," in *Bioinformatics. Methods in Molecular Biology*, vol. 1525, 2017, pp. 137–164.

[15] C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma, "Structure of hen egg-white lysozyme: A three-dimensional Fourier synthesis at 2 Å resolution," *Nature*, vol. 206, no. 4986, pp. 757–761, 1965, doi: 10.1038/206757a0.

[16] L. N. Johnson and D. C. Phillips, "Structure of some crystalline lysozyme-inhibitor complexes determined by X-ray analysis at 6 Å resolution," *Nature*, vol. 206, no. 4986, pp. 761–763, 1965, doi: 10.1038/206761a0.

[17] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, "Molecular dynamics simulations: Advances and applications," *Adv. Appl. Bioinforma. Chem.*, vol. 8, no. 1, 2015, doi: 10.2147/AABC.S70333.

[18] A. B. Kolomeisky and M. E. Fisher, "Molecular motors: A theorist's perspective," *Annu. Rev. Phys. Chem.*, vol. 58, no. 1, pp. 675–695, 2007, doi: 10.1146/annurev.physchem.58.032806.104532.

[19] I. Bahar, C. Chennubhotla, and D. Tobi, "Intrinsic dynamics of enzymes in the unbound state and relation to allosteric regulation," *Curr. Opin. Struct. Biol.*, vol. 17, no. 6, pp. 633–640, 2007, doi: 10.1016/j.sbi.2007.09.011.

[20] M. A. Maria-Solano, E. Serrano-Hervás, A. Romero-Rivera, J. Iglesias-Fernández, and S. Osuna, "Role of conformational dynamics in the evolution of novel enzyme function," *Chem. Commun.*, vol. 54, no. 50, pp. 6622–6634, 2018, doi: 10.1039/c8cc02426j.

[21] A. Stank, D. B. Kokh, J. C. Fuller, and R. C. Wade, "Protein Binding Pocket Dynamics," *Acc. Chem. Res.*, vol. 49, no. 5, pp. 809–815, 2016, doi: 10.1021/acs.accounts.5b00516.

[22] T. Shinoda *et al.*, "Distinct conformation-mediated functions of an active site loop in the catalytic reactions of NAD-

dependent D-lactate dehydrogenase and formate dehydrogenase," *J. Biol. Chem.*, vol. 280, no. 17, pp. 17068–17075, 2005, doi: 10.1074/jbc.M500970200.

[23] M. Karplus, "Role of conformation transitions in adenylate kinase," in *Proceedings of the National Academy of Sciences of the United States of America*, 2010, vol. 107, no. 17, p. E71, doi: 10.1073/pnas.1002180107.

[24] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio, and R. Nussinov, "The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery," *Chem. Rev.*, vol. 116, no. 11, pp. 6391–6423, 2016, doi: 10.1021/acs.chemrev.5b00623.

[25] Y. Wang, K. Bugge, B. B. Kragelund, and K. Lindorff-Larsen, "Role of protein dynamics in transmembrane receptor signalling," *Curr. Opin. Struct. Biol.*, vol. 48, pp. 74–82, 2018, doi: 10.1016/j.sbi.2017.10.017.

[26] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, "The energy landscapes and motions of proteins," *Science (80-.).*, vol. 254, no. 5038, pp. 1598–1603, 1991, doi: 10.1126/science.1749933.

[27] M. D. Miller and G. N. Phillips, "Moving beyond static snapshots: Protein dynamics and the Protein Data Bank," *J. Biol. Chem.*, vol. 296, no. 100749, 2021, doi: 10.1016/j.jbc.2021.100749.

[28] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nat. Struct. Biol.*, vol. 4, no. 1, pp. 10–19, 1997, doi: 10.1038/nsb0197-10.

[29] P. W. Fenimore *et al.*, "Concepts and problems in protein dynamics," *Chem. Phys.*, vol. 424, pp. 2–6, 2013, doi: 10.1016/j.chemphys.2013.06.023.

[30] H. Frauenfelder *et al.*, "A unified model of protein dynamics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 13, pp. 5129–5134, 2009, doi: 10.1073/pnas.0900336106.

[31] C. Levinthal, "How to fold graciously," in *Mössbauer Spectroscopy in Biological Systems (Proceedings of a Meeting Held at Allerton House, Monticello, Illinois)*, 1969, pp. 22–24.

[32] P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: A kinetic approach to the sequence-structure relationship," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 89, no. 18, pp. 8721–8725, 1992, doi: 10.1073/pnas.89.18.8721.

[33] C.-J. Tsai, S. Kumar, B. Ma, and R. Nussinov, "Folding funnels, binding funnels, and protein function," *Protein Sci.*, vol. 8, no. 6, pp. 1181–1190, 1999, doi: 10.1110/ps.8.6.1181.

[34] P. Mishra and S. K. Jha, "The native state conformational heterogeneity in the energy landscape of protein folding," *Biophys. Chem.*, vol. 283, no. 106761, 2022, doi: 10.1016/j.bpc.2022.106761.

[35] G. Wei, W. Xi, R. Nussinov, and B. Ma, "Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? the Diverse Functional Roles of Conformational Ensembles in the Cell," *Chem. Rev.*, vol. 116, no. 11, pp. 6516–6551, 2016, doi: 10.1021/acs.chemrev.5b00562.

[36] L. Q. Yang *et al.*, "Protein dynamics and motions in relation to their functions: Several case studies and the underlying mechanisms," *J. Biomol. Struct. Dyn.*, vol. 32, no. 3, pp. 372–393, 2014, doi: 10.1080/07391102.2013.770372.

[37] S. Raman, N. Taylor, N. Genuth, S. Fields, and G. M. Church, "Engineering allostery," *Trends Genet.*, vol. 30, no. 12, pp. 521–528, 2014, doi: 10.1016/j.tig.2014.09.004.

[38] S. J. Wodak *et al.*, "Allostery in Its Many Disguises: From Theory to Applications," *Structure*, vol. 27, no. 4, pp. 566–578, 2019, doi: 10.1016/j.str.2019.01.003.

[39] R. A. Laskowski, F. Gerick, and J. M. Thornton, "The structural basis of allosteric regulation in proteins," *FEBS Lett.*, vol. 583, no. 11, pp. 1692–1698, 2009, doi: 10.1016/j.febslet.2009.03.019.

[40] R. Nussinov and C. J. Tsai, "Allostery in disease and in drug discovery," *Cell*, vol. 153, no. 2, pp. 293–305, 2013, doi: 10.1016/j.cell.2013.03.034.

[41] J. Liu and R. Nussinov, "Allostery: An Overview of Its History, Concepts, Methods, and Applications," *PLoS Comput. Biol.*, vol. 12, no. 6, p. e1004966, 2016, doi: 10.1371/journal.pcbi.1004966.

[42] J. Monod, *Chance and Necessity: Essay on the Natural Philosophy of Modern Biology*. New York: Vintage Books, 1977.

[43] A. W. Fenton, "Allostery: an illustrated definition for the 'second secret of life,'" *Trends Biochem. Sci.*, vol. 33, no. 9, pp. 420–425, 2008, doi: 10.1016/j.tibs.2008.05.009.

[44] H. N. Motlagh, J. O. Wrabl, J. Li, and V. J. Hilser, "The ensemble nature of allostery," *Nature*, vol. 508, no. 7496, pp.

331–339, 2014, doi: 10.1038/nature13001.

[45] O. Jardetzky, “Protein dynamics and conformational transitions in allosteric proteins,” *Prog. Biophys. Mol. Biol.*, vol. 65, no. 3, pp. 171–219, 1996, doi: 10.1016/S0079-6107(96)00010-7.

[46] N. M. Goodey and S. J. Benkovic, “Allosteric regulation and catalysis emerge via a common route,” *Nat. Chem. Biol.*, vol. 4, no. 8, pp. 474–482, 2008, doi: 10.1038/nchembio.98.

[47] C. J. Tsai, A. Del Sol, and R. Nussinov, “Protein allostery, signal transmission and dynamics: A classification scheme of allosteric mechanisms,” *Mol. Biosyst.*, vol. 5, no. 3, pp. 207–216, 2009, doi: 10.1039/b819720b.

[48] A. J. Faure, J. Domingo, J. M. Schmiedel, C. Hidalgo-Carcedo, G. Diss, and B. Lehner, “Mapping the energetic and allosteric landscapes of protein binding domains,” *Nature*, vol. 604, no. 7904, pp. 175–183, 2022, doi: 10.1038/s41586-022-04586-4.

[49] J. P. Changeux, “The concept of allosteric modulation: an overview,” *Drug Discov. Today Technol.*, vol. 10, no. 2, pp. e223–e228, 2013, doi: 10.1016/j.ddtec.2012.07.007.

[50] K. Gunasekaran, B. Ma, and R. Nussinov, “Is allostery an intrinsic property of all dynamic proteins?,” *Proteins Struct. Funct. Genet.*, vol. 57, no. 3, pp. 433–443, 2004, doi: 10.1002/prot.20232.

[51] S. R. Tzeng and C. G. Kalodimos, “Protein dynamics and allostery: An NMR view,” *Curr. Opin. Struct. Biol.*, vol. 21, no. 1, pp. 62–67, 2011, doi: 10.1016/j.sbi.2010.10.007.

[52] J. A. Hardy and J. A. Wells, “Searching for new allosteric sites in enzymes,” *Curr. Opin. Struct. Biol.*, vol. 14, no. 6, pp. 706–715, 2004, doi: 10.1016/j.sbi.2004.10.009.

[53] C. Kung *et al.*, “AG-348 enhances pyruvate kinase activity in red blood cells from patients with pyruvate kinase deficiency,” *Blood*, vol. 130, no. 11, pp. 1347–1356, 2017, doi: 10.1182/blood-2016-11-753525.

[54] M. A. E. Rab *et al.*, “AG-348 (Mitapivat), an allosteric activator of red blood cell pyruvate kinase, increases enzymatic activity, protein stability, and ATP levels over a broad range of PKLR genotypes,” *Haematologica*, vol. 106, no. 1, pp. 238–249, 2021, doi: 10.3324/haematol.2019.238865.

[55] M. J. van Dijk *et al.*, “Activation of pyruvate kinase as therapeutic option for rare hemolytic anemias: Shedding new light on an old enzyme,” *Blood Rev.*, vol. 61, p. 101103, 2023, doi: 10.1016/j.blre.2023.101103.

[56] S. Lu, X. He, D. Ni, and J. Zhang, “Allosteric Modulator Discovery: From Serendipity to Structure-Based Design,” *J. Med. Chem.*, vol. 62, no. 14, pp. 6405–6421, 2019, doi: 10.1021/acs.jmedchem.8b01749.

[57] J. E. Lindsley and J. Rutter, “Whence cometh the allosterome?,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 28, pp. 10533–10535, 2006, doi: 10.1073/pnas.0604452103.

[58] N. V. Dokholyan, “Controlling Allosteric Networks in Proteins,” *Chem. Rev.*, vol. 116, no. 11, pp. 6463–6487, 2016, doi: 10.1021/acs.chemrev.5b00544.

[59] J. Monod and F. Jacob, “General conclusions: teleonomic mechanisms in cellular metabolism, growth, and differentiation,” *Cold Spring Harb. Symp. Quant. Biol.*, vol. 26, pp. 389–401, 1961, doi: 10.1101/sqb.1961.026.01.048.

[60] J. P. Changeux, “The feedback control mechanisms of biosynthetic L-threonine deaminase by L-isoleucine.,” *Cold Spring Harb. Symp. Quant. Biol.*, vol. 26, pp. 313–318, 1961, doi: 10.1101/SQB.1961.026.01.037.

[61] J. Monod, J. P. Changeux, and F. Jacob, “Allosteric proteins and cellular control systems,” *J. Mol. Biol.*, vol. 6, no. 4, pp. 306–329, 1963, doi: 10.1016/S0022-2836(63)80091-1.

[62] C. Bohr, K. Hasselbalch, and A. Krogh, “Über einen in biologischer Beziehung wichtigen Einfluss, den die Kohlensäurespannung des Blutes auf dessen Sauerstoffbindung übt,” *Skand. Arch. Physiol.*, vol. 16, no. 2, pp. 402–412, 1904.

[63] G. S. Adair, A. V. Bock, and H. Field Jr, “The Hemoglobin System: VI. The Oxygen Dissociation Curve of Hemoglobin,” *J. Biol. Chem.*, vol. 63, no. 2, pp. 529–545, 1925.

[64] L. Pauling, “The Oxygen Equilibrium of Hemoglobin and Its Structural Interpretation,” *Proc. Natl. Acad. Sci.*, vol. 21, no. 4, pp. 181–191, 1935, doi: 10.1073/pnas.21.4.186.

[65] G. T. Cori, S. P. Colowick, and C. F. Cori, “The action of nucleotides in the disruptive phosphorylation of glycogen,” *J. Biol. Chem.*, vol. 123, no. 2, pp. 381–389, 1938, doi: 10.1016/s0021-9258(18)74126-4.

[66] A. Novick and L. Szilard, "Experiments with the chemostat on the rates of amino acid synthesis in bacteria," *Dynamics of Growth Processes*. pp. 21–32, 1954.

[67] H. E. Umbarger, "Evidence for a negative-feedback mechanism in the biosynthesis of isoleucine," *Science (80-.).*, vol. 123, no. 3202, p. 848, 1956, doi: 10.1126/science.123.3202.848.

[68] A. B. Pardee and R. A. Yates, "Control of pyrimidine biosynthesis in Escherichia coli by a feedback mechanism," *J. Biol. Chem.*, vol. 221, no. 2, pp. 757–770, 1956, doi: 10.1016/s0021-9258(18)65188-9.

[69] Q. Cui and M. Karplus, "Allostery and cooperativity revisited," *Protein Sci.*, vol. 17, no. 8, pp. 1295–1307, 2008, doi: 10.1110/ps.03259908.

[70] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, vol. 181, no. 4610, pp. 662–666, 1958, doi: 10.1038/181662a0.

[71] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North, "Structure of hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis," *Nature*, vol. 185, no. 4711, pp. 416–422, 1960.

[72] D. E. Koshland, "Application of a Theory of Enzyme Specificity to Protein Synthesis," *Proc. Natl. Acad. Sci.*, vol. 44, no. 2, pp. 98–104, 1958, doi: 10.1073/pnas.44.2.98.

[73] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme," *Berichte der Dtsch. Chem. Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, 1894, doi: 10.1002/cber.18940270364.

[74] J. Monod, J. Wyman, and J. P. Changeux, "On the nature of allosteric transitions: A plausible model," *J. Mol. Biol.*, vol. 12, no. 1, pp. 88–118, 1965, doi: 10.1016/S0022-2836(65)80285-6.

[75] D. E. Koshland, J. G. Nemethy, and D. Filmer, "Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits," *Biochemistry*, vol. 5, no. 1, pp. 365–385, 1966, doi: 10.1021/bi00865a047.

[76] D. E. Koshland and K. Hamadani, "Proteomics and models for enzyme cooperativity," *J. Biol. Chem.*, vol. 277, no. 49, pp. 46841–46844, 2002, doi: 10.1074/jbc.R200014200.

[77] M. Eigen, "New looks and outlooks on physical enzymology," *Q. Rev. Biophys.*, vol. 1, no. 1, pp. 3–33, 1968, doi: 10.1017/S0033583500000445.

[78] G. Weber, "Ligand Binding and Internal Equilibria in Proteins," *Biochemistry*, vol. 11, no. 5, pp. 864–878, 1972, doi: 10.1021/bi00755a028.

[79] P. Ascenzi, A. Bocedi, A. Bolli, M. Fasano, S. Notari, and F. Polticelli, "Allosteric modulation of monomeric proteins," *Biochem. Mol. Biol. Educ.*, vol. 33, no. 3, pp. 169–176, 2005, doi: 10.1002/bmb.2005.494033032470.

[80] V. J. Hilser and E. B. Thompson, "Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 20, pp. 8311–8315, 2007, doi: 10.1073/pnas.0700329104.

[81] N. Popovych, S. Sun, R. H. Ebright, and C. G. Kalodimos, "Dynamically driven protein allostery," *Nat. Struct. Mol. Biol.*, vol. 13, no. 9, pp. 831–838, 2006, doi: 10.1038/nsmb1132.

[82] C. J. Tsai, A. del Sol, and R. Nussinov, "Allostery: Absence of a Change in Shape Does Not Imply that Allostery Is Not at Play," *J. Mol. Biol.*, vol. 378, no. 1, pp. 1–11, 2008, doi: 10.1016/j.jmb.2008.02.034.

[83] A. Peselis, A. Gao, and A. Serganov, "Cooperativity, allostery and synergism in ligand binding to riboswitches," *Biochimie*, vol. 117, pp. 100–109, 2015, doi: 10.1016/j.biochi.2015.06.028.

[84] S. A. Harris, E. Gavathiotis, M. S. Searle, M. Orozco, and C. A. Laughton, "Cooperativity in drug-DNA recognition: A molecular dynamics study," *J. Am. Chem. Soc.*, vol. 123, no. 50, pp. 12658–12663, 2001, doi: 10.1021/ja016233n.

[85] D. Kern and E. R. P. Zuiderweg, "The role of dynamics in allosteric regulation," *Curr. Opin. Struct. Biol.*, vol. 13, no. 6, pp. 748–757, 2003, doi: 10.1016/j.sbi.2003.10.008.

[86] O. Schueler-Furman and S. J. Wodak, "Computational approaches to investigating allostery," *Curr. Opin. Struct. Biol.*, vol. 41, pp. 159–171, 2016, doi: 10.1016/j.sbi.2016.06.017.

[87] V. J. Hilser, J. O. Wrabl, and H. N. Motlagh, "Structural and energetic basis of allostery," *Annu. Rev. Biophys.*, vol. 41, no. 1, pp. 585–609, 2012, doi: 10.1146/annurev-biophys-050511-102319.

[88] M. Civera, E. Moroni, L. Sorrentino, F. Vasile, and S. Sattin, "Chemical and Biophysical Approaches to Allosteric

Modulation," *European J. Org. Chem.*, vol. 2021, no. 30, pp. 4245–4259, 2021, doi: 10.1002/ejoc.202100506.

[89] R. Van Der Lee *et al.*, "Classification of intrinsically disordered regions and proteins," *Chem. Rev.*, vol. 114, no. 13, pp. 6589–6631, 2014, doi: 10.1021/cr400525m.

[90] T. Flock, R. J. Weatheritt, N. S. Latysheva, and M. M. Babu, "Controlling entropy to tune the functions of intrinsically disordered regions," *Curr. Opin. Struct. Biol.*, vol. 26, no. 1, pp. 62–72, 2014, doi: 10.1016/j.sbi.2014.05.007.

[91] A. Cooper and D. T. F. Dryden, "Allostery without conformational change - A plausible model," *Eur. Biophys. J.*, vol. 11, no. 2, pp. 103–109, 1984, doi: 10.1007/BF00276625.

[92] A. del Sol, C. J. Tsai, B. Ma, and R. Nussinov, "The Origin of Allosteric Functional Modulation: Multiple Pre-existing Pathways," *Structure*, vol. 17, no. 8, pp. 1042–1050, 2009, doi: 10.1016/j.str.2009.06.008.

[93] R. Nussinov and C. J. Tsai, "Allostery without a conformational change? Revisiting the paradigm," *Curr. Opin. Struct. Biol.*, vol. 30, pp. 17–24, 2015, doi: 10.1016/j.sbi.2014.11.005.

[94] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan, "Evolutionarily conserved networks of residues mediate allosteric communication in proteins," *Nat. Struct. Biol.*, vol. 10, no. 1, pp. 59–69, 2003, doi: 10.1038/nsb881.

[95] C. J. Tsai and R. Nussinov, "A Unified View of 'How Allostery Works,'" *PLoS Comput. Biol.*, vol. 10, no. 2, p. e1003394, 2014, doi: 10.1371/journal.pcbi.1003394.

[96] J. O. Wrabl, J. Gu, T. Liu, T. P. Schrank, S. T. Whitten, and V. J. Hilser, "The role of protein conformational fluctuations in allostery, function, and evolution," *Biophys. Chem.*, vol. 159, no. 1, pp. 129–141, 2011, doi: 10.1016/j.bpc.2011.05.020.

[97] E. K. Jaffe, "Morpheeins - A new structural paradigm for allosteric regulation," *Trends Biochem. Sci.*, vol. 30, no. 9, pp. 490–497, 2005, doi: 10.1016/j.tibs.2005.07.003.

[98] A. C. Storer and A. Cornish-Bowden, "Kinetic evidence for a 'Mnemonical' mechanism for rat liver glucokinase," *Biochem. J.*, vol. 165, no. 1, pp. 61–69, 1977, doi: 10.1042/bj1650061.

[99] K. Kamata, M. Mitsuya, T. Nishimura, J. I. Eiki, and Y. Nagata, "Structural basis for allosteric regulation of the monomeric allosteric enzyme human glucokinase," *Structure*, vol. 12, no. 3, pp. 429–438, 2004, doi: 10.1016/j.str.2004.02.005.

[100] Z. Huang *et al.*, "ASD: A comprehensive database of allosteric proteins and modulators," *Nucleic Acids Res.*, vol. 39, no. suppl_1, pp. D663–D669, 2011, doi: 10.1093/nar/gkq1022.

[101] M. V. LeVine and H. Weinstein, "AIM for allostery: Using the ising model to understand information processing and transmission in allosteric biomolecular systems," *Entropy*, vol. 17, no. 5, pp. 2895–2918, 2015, doi: 10.3390/e17052895.

[102] Q. Huang, L. Lai, and Z. Liu, "Quantitative Analysis of Dynamic Allostery," *J. Chem. Inf. Model.*, vol. 62, no. 10, pp. 2538–2549, 2022, doi: 10.1021/acs.jcim.2c00138.

[103] J. Guo and H. X. Zhou, "Protein Allostery and Conformational Dynamics," *Chem. Rev.*, vol. 116, no. 11, pp. 6503–6515, 2016, doi: 10.1021/acs.chemrev.5b00590.

[104] J. S. Fraser *et al.*, "Accessing protein conformational ensembles using room-temperature X-ray crystallography," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 39, pp. 16247–16252, 2011, doi: 10.1073/pnas.1111325108.

[105] D. A. Keedy, "Journey to the center of the protein: Allostery from multitemperature multiconformer X-ray crystallography," *Acta Crystallogr. Sect. D Struct. Biol.*, vol. 75, no. 2, pp. 123–137, 2019, doi: 10.1107/S2059798318017941.

[106] J. E. Knapp, R. Pahl, V. Šrajer, and W. E. Royer, "Allosteric action in real time: Time-resolved crystallographic studies of a cooperative dimeric hemoglobin," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 20, pp. 7649–7654, 2006, doi: 10.1073/pnas.0509411103.

[107] S. Grutsch, S. Brüschweiler, and M. Tollinger, "NMR Methods to Study Dynamic Allostery," *PLoS Comput. Biol.*, vol. 12, no. 3, p. e1004620, 2016, doi: 10.1371/journal.pcbi.1004620.

[108] S. Hertig, N. R. Latorraca, and R. O. Dror, "Revealing Atomic-Level Mechanisms of Protein Allostery with Molecular Dynamics Simulations," *PLoS Comput. Biol.*, vol. 12, no. 6, p. e1004746, 2016, doi: 10.1371/journal.pcbi.1004746.

[109] E. Papaleo, "Investigating Conformational Dynamics and Allostery in the p53 DNA-Binding Domain Using Molecular Simulations," in *Allostery. Methods in Molecular Biology*, vol. 2253, 2021, pp. 221–244.

[110] Y. Tsuchiya, K. Taneishi, and Y. Yonezawa, “Autoencoder-Based Detection of Dynamic Allostery Triggered by Ligand Binding Based on Molecular Dynamics,” *J. Chem. Inf. Model.*, vol. 59, no. 9, pp. 4043–4051, 2019, doi: 10.1021/acs.jcim.9b00426.

[111] L. Meireles, M. Gur, A. Bakan, and I. Bahar, “Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins,” *Protein Sci.*, vol. 20, no. 10, pp. 1645–1658, 2011, doi: 10.1002/pro.711.

[112] A. Panjkovich and X. Daura, “Exploiting protein flexibility to predict the location of allosteric sites,” *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–12, 2012, doi: 10.1186/1471-2105-13-273.

[113] A. Goncearenco, S. Mitternacht, T. Yong, B. Eisenhaber, F. Eisenhaber, and I. N. Berezovsky, “SPACER: Server for predicting allosteric communication and effects of regulation.,” *Nucleic Acids Res.*, vol. 41, no. w1, pp. W266–W272, 2013, doi: 10.1093/nar/gkt460.

[114] C. Kaya, A. Armutlulu, S. Ekesan, and T. Haliloglu, “MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues.,” *Nucleic Acids Res.*, vol. 41, no. W1, pp. W249–W255, 2013, doi: 10.1093/nar/gkt284.

[115] W. Huang *et al.*, “Allosite: A method for predicting allosteric sites,” *Bioinformatics*, vol. 29, no. 18, pp. 2357–2359, 2013, doi: 10.1093/bioinformatics/btt399.

[116] S. W. Lockless and R. Ranganathan, “Evolutionarily conserved pathways of energetic connectivity in protein families,” *Science (80-.).*, vol. 286, no. 5438, pp. 295–299, 1999, doi: 10.1126/science.286.5438.295.

[117] M. E. Muñoz and E. Ponce, “Pyruvate kinase: current status of regulatory and functional properties,” *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.*, vol. 135, no. 2, pp. 197–218, Jun. 2003, doi: 10.1016/S1096-4959(03)00081-2.

[118] A. Zanella, E. Fermo, P. Bianchi, and G. Valentini, “Red cell pyruvate kinase deficiency: molecular and clinical aspects,” *Br. J. Haematol.*, vol. 130, no. 1, pp. 11–25, 2005, doi: 10.1111/j.1365-2141.2005.05527.x.

[119] N. Schormann, K. L. Hayden, P. Lee, S. Banerjee, and D. Chattopadhyay, “An Overview of Structure, Function and Regulation of Pyruvate Kinases,” *Protein Sci.*, 2019, doi: 10.1002/pro.3691.

[120] H. Muirhead *et al.*, “The structure of cat muscle pyruvate kinase,” *EMBO J.*, vol. 5, no. 3, pp. 475–481, 1986, [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1166788&tool=pmcentrez&rendertype=abstract.

[121] G. Valentini *et al.*, “Structure and function of human erythrocyte pyruvate kinase: Molecular basis of nonspherocytic hemolytic anemia,” *J. Biol. Chem.*, vol. 277, no. 26, pp. 23807–23814, 2002, doi: 10.1074/jbc.M202107200.

[122] H. P. Morgan, W. Zhong, I. W. McNae, P. A. M. Michels, L. A. Fothergill-Gilmore, and M. D. Walkinshaw, “Structures of pyruvate kinases display evolutionarily divergent allosteric strategies.,” *R. Soc. open Sci.*, vol. 1, no. 1, p. 140120, 2014, doi: 10.1098/rsos.140120.

[123] A. W. Fenton and A. Y. Alontaga, “The impact of ions on allosteric functions in human liver pyruvate kinase,” *Methods Enzymol.*, vol. 466, pp. 83–107, 2009, doi: 10.1016/s0076-6879(09)66005-5.

[124] J. D. Dombrauckas, B. D. Santarsiero, and A. D. Mesecar, “Structural basis for tumor pyruvate kinase M2 allosteric regulation and catalysis,” *Biochemistry*, vol. 44, no. 27, pp. 9417–9429, 2005, doi: 10.1021/bi0474923.

[125] W. J. Israelsen and M. G. Vander Heiden, “Pyruvate kinase: Function, regulation and role in cancer,” *Semin. Cell Dev. Biol.*, vol. 43, pp. 43–51, 2015, doi: 10.1016/j.semcdb.2015.08.004.

[126] M. S. Jurica, A. Mesecar, P. J. Heath, W. Shi, T. Nowak, and B. L. Stoddard, “The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate,” *Structure*, vol. 6, pp. 195–210, 1998, doi: 10.1016/S0969-2126(98)00021-5.

[127] P. Wang, C. Sun, T. Zhu, and Y. Xu, “Structural insight into mechanisms for dynamic regulation of PKM2,” *Protein Cell*, vol. 6, no. 4, pp. 275–287, 2015, doi: 10.1007/s13238-015-0132-x.

[128] M. Yuan *et al.*, “An allostatic mechanism for M2 pyruvate kinase as an amino-acid sensor,” *Biochem. J.*, vol. 475, no. 10, pp. 1821–1837, 2018, doi: 10.1042/BCJ20180171.

[129] T. Noguchi, K. Yamada, H. Inoue, T. Matsuda, and T. Tanaka, “The L- and R-type isozymes of rat pyruvate kinase are produced from a single gene by use of different promoters.,” *J. Biol. Chem.*, vol. 262, no. 29, pp. 14366–14371, 1987.

[130] H. Kanno, H. Fujii, and S. Miwa, “Structural analysis of human pyruvate kinase L-gene and identification of the

promoter activity in erythroid cells," *Biochem. Biophys. Res. Commun.*, vol. 188, no. 2, pp. 516–523, 1992, doi: 10.1016/0006-291X(92)91086-6.

[131] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, p. 980, 2003, doi: 10.1038/nsb1203-980.

[132] A. Mattevi, G. Valentini, M. Rizzi, M. L. Speranza, M. Bolognesi, and A. Coda, "Crystal structure of Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition.," *Structure*, vol. 3, no. 7, pp. 729–741, 1995, doi: 10.1016/S0969-2126(01)00207-6.

[133] T. M. Larsen, L. T. Laughlin, H. M. Holden, I. Rayment, and G. H. Reed, "Structure of Rabbit Muscle Pyruvate Kinase Complexed with Mn2+, K+, and Pyruvate," *Biochemistry*, vol. 33, no. 20, pp. 6301–6309, 1994, doi: 10.1021/bi00186a033.

[134] R. K. Wierenga, "The TIM-barrel fold: A versatile framework for efficient enzymes," *FEBS Lett.*, vol. 492, no. 3, pp. 193–198, 2001, doi: 10.1016/S0014-5793(01)02236-0.

[135] L. B. Tulloch, H. P. Morgan, V. Hannaert, P. A. M. Michels, L. A. Fothergill-Gilmore, and M. D. Walkinshaw, "Sulphate Removal Induces a Major Conformational Change in Leishmania mexicana Pyruvate Kinase in the Crystalline State," *J. Mol. Biol.*, vol. 383, no. 3, pp. 615–626, 2008, doi: 10.1016/j.jmb.2008.08.037.

[136] B. O. Community, "Blender - a 3D modelling and rendering package," *Blender Foundation*. 2023, [Online]. Available: http://www.blender.org.

[137] W. Zhong, H. P. Morgan, I. W. McNae, P. A. M. Michels, L. A. Fothergill-Gilmore, and M. D. Walkinshaw, "'In crystallo' substrate binding triggers major domain movements and reveals magnesium as a co-activator of Trypanosoma brucei pyruvate kinase," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 69, no. 9, pp. 1768–1779, 2013, doi: 10.1107/S0907444913013875.

[138] T. M. Larsen, M. M. Benning, G. E. Wesenberg, I. Rayment, and G. H. Reed, "Ligand-induced domain movement in pyruvate kinase: Structure of the enzyme from rabbit muscle with Mg2+, K+, and L-phospholactate at 2.7 Å resolution," *Arch. Biochem. Biophys.*, vol. 345, no. 2, pp. 199–206, 1997, doi: 10.1006/abbi.1997.0257.

[139] T. M. Larsen, M. M. Benning, I. Rayment, and G. H. Reed, "Structure of the Bis(Mg2+)-ATP-oxalate complex of the rabbit muscle pyruvate kinase at 2.1 Å resolution: ATP binding over a barrel," *Biochemistry*, vol. 37, no. 18, pp. 6247–6255, 1998, doi: 10.1021/bi980243s.

[140] J. Oria-Hernández, N. Cabrera, R. Pérez-Montfort, and L. Ramírez-Silva, "Pyruvate kinase revisited: The activating effect of K+," *J. Biol. Chem.*, vol. 280, no. 45, pp. 37924–37929, 2005, doi: 10.1074/jbc.M508490200.

[141] L. Ramírez-Silva *et al.*, "The K+-Dependent and-Independent Pyruvate Kinases Acquire the Active Conformation by Different Mechanisms," *Int. J. Mol. Sci.*, vol. 23, no. 3, p. 1347, 2022, doi: 10.3390/ijms23031347.

[142] E. A. Permyakov, "Metal Binding Proteins," *Encyclopedia*, vol. 1, no. 1, pp. 261–292, 2021, doi: 10.3390/encyclopedia1010024.

[143] L. T. Laughlin and G. H. Reed, "The monovalent cation requirement of rabbit muscle pyruvate kinase is eliminated by substitution of lysine for glutamate 117," *Arch. Biochem. Biophys.*, vol. 348, no. 2, pp. 262–267, 1997, doi: 10.1006/abbi.1997.0448.

[144] N. F. Brás *et al.*, "Analyses of cobalt-ligand and potassium-ligand bond lengths in metalloproteins: Trends and patterns," *J. Mol. Model.*, vol. 20, no. 6, pp. 1–14, 2014, doi: 10.1007/s00894-014-2271-z.

[145] M. Vašák and J. Schnabl, "Sodium and Potassium Ions in Proteins and Enzyme Catalysis," in *The Alkali Metal Ions: Their Role for Life. Metal Ions in Life Sciences*, vol. 16, 2016, pp. 259–290.

[146] Y. Ou, W. Tao, Y. Zhang, G. Wu, and S. Yu, "The conformational change of rabbit muscle pyruvate kinase induced by activating cations and its substrates," *Int. J. Biol. Macromol.*, vol. 47, no. 2, pp. 228–232, 2010, doi: 10.1016/j.ijbiomac.2010.04.017.

[147] H. P. Morgan *et al.*, "Allosteric mechanism of pyruvate kinase from Leishmania mexicana uses a rock and lock model," *J. Biol. Chem.*, vol. 285, no. 17, pp. 12892–12898, 2010, doi: 10.1074/jbc.M109.079905.

[148] A. Ishwar, Q. Tang, and A. W. Fenton, "Distinguishing the interactions in the fructose 1,6-bisphosphate binding site of human liver pyruvate kinase that contribute to allostery," *Biochemistry*, vol. 54, no. 7, pp. 1516–1524, 2015, doi: 10.1021/bi501426w.

[149] H. P. Morgan *et al.*, "M2 pyruvate kinase provides a mechanism for nutrient sensing and regulation of cell

proliferation," *Proc. Natl. Acad. Sci.*, vol. 110, no. 15, pp. 5881–5886, 2013, doi: 10.1073/pnas.1217157110.

[150] A. F. M. Gavriilidou, F. P. Holding, D. Mayer, J. E. Coyle, D. B. Veprintsev, and R. Zenobi, "Native Mass Spectrometry Gives Insight into the Allosteric Binding Mechanism of M2 Pyruvate Kinase to Fructose-1,6-Bisphosphate," *Biochemistry*, vol. 57, no. 11, pp. 1685–1689, 2018, doi: 10.1021/acs.biochem.7b01270.

[151] A. Y. Alontaga and A. W. Fenton, "Effector analogues detect varied allosteric roles for conserved protein-effector interactions in pyruvate kinase isozymes," *Biochemistry*, vol. 50, no. 11, pp. 1934–1939, 2011, doi: 10.1021/bi200052e.

[152] Q. Tang and A. W. Fenton, "Whole-protein alanine-scanning mutagenesis of allostery: A large percentage of a protein can contribute to mechanism," *Hum. Mutat.*, vol. 38, no. 9, pp. 1132–1143, 2017, doi: 10.1002/humu.23231.

[153] B. Chaneton *et al.*, "Serine is a natural ligand and allosteric activator of pyruvate kinase M2.," *Nature*, vol. 491, no. 7424, pp. 458–62, 2012, doi: 10.1038/nature11540.

[154] S. Nandi and M. Dey, "Biochemical and structural insights into how amino acids regulate pyruvate kinase muscle isoform 2," *J. Biol. Chem.*, vol. 295, no. 16, pp. 5390–5403, 2020, doi: 10.1074/jbc.RA120.013030.

[155] A. W. Fenton and Q. Tang, "An activating interaction between the unphosphorylated N-terminus of human liver pyruvate kinase and the main body of the protein is interrupted by phosphorylation," *Biochemistry*, vol. 48, no. 18, pp. 3816–3818, 2009, doi: 10.1021/bi900421f.

[156] C. B. Prasannan, Q. Tang, and A. W. Fenton, "Allosteric regulation of human liver pyruvate kinase by peptides that mimic the phosphorylated/dephosphorylated N-terminus," in *Allostery. Methods in Molecular Biology*, vol. 796, 2012, pp. 335–349.

[157] R. Van Bruggen *et al.*, "Modulation of malaria phenotypes by pyruvate kinase (PKLR) variants in a Thai population," *PLoS One*, vol. 10, no. 12, pp. 1–18, 2015, doi: 10.1371/journal.pone.0144555.

[158] B. M. Gassaway *et al.*, "Distinct Hepatic PKA and CDK Signaling Pathways Control Activity-Independent Pyruvate Kinase Phosphorylation and Hepatic Glucose Production," *Cell Rep.*, vol. 29, no. 11, pp. 3394–3404, 2019, doi: 10.1016/j.celrep.2019.11.009.

[159] J. C. Lee, "Modulation of allostery of pyruvate kinase by shifting of an ensemble of microstates," *Acta Biochim. Biophys. Sin. (Shanghai).*, vol. 40, no. 7, pp. 663–669, 2008, doi: 10.1111/j.1745-7270.2008.00445.x.

[160] K. A. Donovan *et al.*, "Conformational dynamics and allostery in pyruvate kinase," *J. Biol. Chem.*, vol. 291, no. 17, pp. 9244–9256, 2016, doi: 10.1074/jbc.M115.676270.

[161] A. Mattevi, M. Bolognesi, and G. Valentini, "The allosteric regulation of pyruvate kinase," *Fed. Eur. Biochem. Soc. Lett.*, vol. 389, no. 1, pp. 15–19, 1996.

[162] R. H. E. Friesen, R. J. Castellani, J. C. Lee, and W. Braun, "Allostery in rabbit pyruvate kinase: Development of a strategy to elucidate the mechanism," *Biochemistry*, vol. 37, no. 44, pp. 15266–15276, 1998, doi: 10.1021/bi981273y.

[163] J. E. Pinto Torres *et al.*, "Structural and kinetic characterization of Trypanosoma congolense pyruvate kinase," *Mol. Biochem. Parasitol.*, vol. 236, p. 111263, 2020, doi: 10.1016/j.molbiopara.2020.111263.

[164] A. Naithani, P. Taylor, B. Erman, and M. D. Walkinshaw, "A Molecular Dynamics Study of Allosteric Transitions in Leishmania mexicana Pyruvate Kinase," *Biophys. J.*, vol. 109, no. 6, pp. 1149–1156, 2015, doi: 10.1016/j.bpj.2015.05.040.

[165] J. S. McFarlane, T. A. Ronnebaum, K. M. Meneely, A. Chilton, A. W. Fenton, and A. L. Lamb, "Changes in the allosteric site of human liver pyruvate kinase upon activator binding include the breakage of an intersubunit cation-π bond," *Acta Crystallogr. Sect. F Struct. Biol. Commun.*, vol. 75, no. 6, pp. 461–469, 2019, doi: 10.1107/S2053230X19007209.

[166] P. Kalaiarasan, N. Subbarao, and R. N. K. Bamezai, "Molecular simulation of Tyr105 phosphorylated pyruvate kinase M2 to understand its structure and dynamics," *J. Mol. Model.*, vol. 20, no. 2447, pp. 1–12, 2014, doi: 10.1007/s00894-014-2447-6.

[167] P. Kalaiarasan, B. Kumar, R. Chopra, V. Gupta, N. Subbarao, and R. N. K. Bamezai, "In silico screening, genotyping, molecular dynamics simulation and activity studies of SNPs in Pyruvate Kinase M2," *PLoS One*, vol. 10, no. 3, p. e0120469, 2015, doi: 10.1371/journal.pone.0120469.

[168] P. Gollapalli and M. Hanumanthappa, "Conformational Flexibility and Dynamic Properties in Allosteric Regulation of Mycobacterium Tuberculosis Pyruvate Kinase," *MOJ Proteomics Bioinforma.*, vol. 4, no. 4, p. 00128, 2016, doi:

10.15406/mojpb.2016.04.00128.

[169] J. Yang, H. Liu, X. Liu, C. Gu, R. Luo, and H. F. Chen, "Synergistic Allosteric Mechanism of Fructose-1,6-bisphosphate and Serine for Pyruvate Kinase M2 via Dynamics Fluctuation Network Analysis," *J. Chem. Inf. Model.*, vol. 56, no. 6, pp. 1184–1192, 2016, doi: 10.1021/acs.jcim.6b00115.

[170] V. Li, C. Lee, Y. Lee, and H. Kim, "Molecular dynamics simulation of pyruvate kinase to investigate improved thermostability of artificially selected strain in Enterococcus faecium," *Genes and Genomics*, vol. 45, no. 6, pp. 741–747, 2023, doi: 10.1007/s13258-023-01373-x.

[171] H. Ellegren and N. Galtier, "Determinants of genetic diversity," *Nat. Rev. Genet.*, vol. 17, no. 7, pp. 422–433, 2016, doi: 10.1038/nrg.2016.58.

[172] K. Shameer, L. P. Tripathi, K. R. Kalari, J. T. Dudley, and R. Sowdhamini, "Interpreting functional effects of coding variants: Challenges in proteome-scale prediction, annotation and assessment," *Brief. Bioinform.*, vol. 17, no. 5, pp. 841–862, 2016, doi: 10.1093/bib/bbv084.

[173] J. Thusberg, A. Olatubosun, and M. Vihinen, "Performance of mutation pathogenicity prediction methods on missense variants," *Hum. Mutat.*, vol. 32, no. 4, pp. 358–368, 2011, doi: 10.1002/humu.21445.

[174] M. Lin, S. Whitmire, J. Chen, A. Farrel, X. Shi, and J. T. Guo, "Effects of short indels on protein structure and function in human genomes," *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, 2017, doi: 10.1038/s41598-017-09287-x.

[175] J. J. Galano-Frutos, H. García-Cebollada, and J. Sancho, "Molecular dynamics simulations for genetic interpretation in protein coding regions: where we are, where to go and when," *Brief. Bioinform.*, vol. 22, no. 1, pp. 3–19, 2021, doi: 10.1093/bib/bbz146.

[176] K. Eilbeck, A. Quinlan, and M. Yandell, "Settling the score: Variant prioritization and Mendelian disease," *Nat. Rev. Genet.*, vol. 18, no. 10, pp. 559–612, 2017, doi: 10.1038/nrg.2017.52.

[177] A. Anna and G. Monika, "Splicing mutations in human genetic disorders: examples, detection, and confirmation," *J. Appl. Genet.*, vol. 59, no. 3, pp. 253–268, 2018, doi: 10.1007/s13353-018-0444-7.

[178] L. A. Miosge *et al.*, "Comparison of predicted and actual consequences of missense mutations," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 37, pp. E5189–E5198, 2015, doi: 10.1073/pnas.1511585112.

[179] C. Feinauer and M. Weigt, "Context-Aware Prediction of Pathogenicity of Missense Mutations Involved in Human Disease," *bioRxiv*, 2017, doi: http://dx.doi.org/10.1101/103051.

[180] P. B. McGarvey *et al.*, "UniProt genomic mapping for deciphering functional effects of missense variants," *Hum. Mutat.*, vol. 40, no. 6, pp. 694–705, 2019, doi: 10.1002/humu.23738.

[181] D. Navío, M. Rosell, J. Aguirre, X. de la Cruz, and J. Fernández-Recio, "Structural and computational characterization of disease-related mutations involved in protein-protein interfaces," *Int. J. Mol. Sci.*, vol. 20, no. 7, p. 1583, 2019, doi: 10.3390/ijms20071583.

[182] J. F. Sayılgan, T. Haliloğlu, and M. Gönen, "Protein dynamics analysis reveals that missense mutations in cancer-related genes appear frequently on hinge-neighboring residues," *Proteins Struct. Funct. Bioinforma.*, vol. 87, no. 6, pp. 512–519, 2019, doi: 10.1002/prot.25673.

[183] Y. Liu, W. S. B. Yeung, P. C. N. Chiu, and D. Cao, "Computational approaches for predicting variant impact: An overview from resources, principles to applications," *Front. Genet.*, vol. 13, no. 981005, 2022, doi: 10.3389/fgene.2022.981005.

[184] P. Sneha and C. G. Priya Doss, "Molecular Dynamics: New Frontier in Personalized Medicine," in *Advances in Protein Chemistry and Structural Biology*, vol. 102, 2016, pp. 181–224.

[185] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, 2015, doi: 10.1038/nature15393.

[186] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A vision for the future of genomics research," *Nature*, vol. 422, no. 6934, pp. 835–847, 2003, doi: 10.1038/nature01626.

[187] W. Fu *et al.*, "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants," *Nature*, vol. 493, no. 7431, pp. 216–220, 2013, doi: 10.1038/nature11690.

[188] M. Lek *et al.*, "Analysis of protein-coding genetic variation in 60,706 humans," *Nature*, vol. 536, no. 7616, pp. 285–291, 2016, doi: 10.1038/nature19057.

[189] K. J. Karczewski *et al.*, “The mutational constraint spectrum quantified from variation in 141,456 humans,” *Nature*, vol. 581, no. 7809, pp. 434–443, 2020, doi: 10.1038/s41586-020-2308-7.

[190] R. Chen and M. Snyder, “Promise of personalized omics to precision medicine,” *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 5, no. 1, pp. 73–82, 2013, doi: 10.1002/wsbm.1198.

[191] J. van der Greef, T. Hankemeier, and R. N. McBurney, “Metabolomics-based systems biology and personalized medicine: Moving towards n = 1 clinical trials?,” *Pharmacogenomics*, vol. 7, no. 7, pp. 1087–1094, 2006, doi: 10.2217/14622416.7.7.1087.

[192] R. Karki, D. Pandya, R. C. Elston, and C. Ferlini, “Defining ‘mutation’ and ‘polymorphism’ in the era of personal genomics,” *BMC Med. Genomics*, vol. 8, no. 1, pp. 1–7, 2015, doi: 10.1186/s12920-015-0115-z.

[193] S. T. Sherry *et al.*, “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Res.*, vol. 29, no. 1, pp. 308–311, 2001, doi: 10.1093/nar/29.1.308.

[194] S. E. Hunt *et al.*, “Ensembl variation resources,” *Database (Oxford).*, vol. 2018, 2018, doi: 10.1093/database/bay119.

[195] F. Cunningham *et al.*, “Ensembl 2022,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D988–D995, 2022, doi: 10.1093/nar/gkab1049.

[196] J. McClellan and M. C. King, “Genetic heterogeneity in human disease,” *Cell*, vol. 141, no. 2, pp. 210–217, 2010, doi: 10.1016/j.cell.2010.03.032.

[197] The Cancer Genome Atlas Research Network, “Comprehensive genomic characterization defines human glioblastoma genes and core pathways,” *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008, doi: 10.1038/nature07385.

[198] The International Cancer Genome Consortium, “International network of cancer genome projects,” *Nature*, vol. 464, no. 7291, 2010, doi: 10.1038/nature08987.

[199] P. Bianchi *et al.*, “Addressing the diagnostic gaps in pyruvate kinase deficiency: Consensus recommendations on the diagnosis of pyruvate kinase deficiency,” *Am. J. Hematol.*, vol. 94, no. 1, pp. 149–161, 2019, doi: 10.1002/ajh.25325.

[200] M. S. Hassan, A. A. Shaalan, M. I. Dessouky, A. E. Abdelnaiem, and M. ElHefnawi, “Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity,” *Genomics*, vol. 111, no. 4, pp. 869–882, 2019, doi: 10.1016/j.ygeno.2018.05.013.

[201] L. Ponzoni and I. Bahar, “Structural dynamics is a determinant of the functional significance of missense variants,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 115, no. 16, pp. 4164–4169, 2018, doi: 10.1073/pnas.1715896115.

[202] F. Gutierrez-Rodrigues and R. T. Calado, “The interpretation of rare or novel variants: damaging vs. disease-causing,” *Hematol. Transfus. Cell Ther.*, vol. 40, no. 1, pp. 3–4, 2018, doi: 10.1016/j.bjhh.2017.10.003.

[203] P. D. Stenson *et al.*, “The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting,” *Hum. Genet.*, vol. 139, pp. 1197–1207, 2020, doi: 10.1007/s00439-020-02199-3.

[204] L. Gerasimavicius, X. Liu, and J. A. Marsh, “Identification of pathogenic missense mutations using protein stability predictors,” *Sci. Rep.*, vol. 10, no. 1, p. 15387, 2020, doi: 10.1038/s41598-020-72404-w.

[205] S. Richards *et al.*, “Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,” *Genet. Med.*, vol. 17, no. 5, pp. 405–424, 2015, doi: 10.1038/gim.2015.30.Standards.

[206] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, “OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D789–D798, 2015, doi: 10.1093/nar/gku1205.

[207] A. Mottaz, F. P. A. David, A. L. Veuthey, and Y. L. Yip, “Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar,” *Bioinformatics*, vol. 26, no. 6, pp. 851–852, 2010, doi: 10.1093/bioinformatics/btq028.

[208] A. Bateman *et al.*, “UniProt: the universal protein knowledgebase in 2021,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, 2021, doi: 10.1093/nar/gkaa1100.

[209] M. J. Landrum *et al.*, “ClinVar: Improving access to variant interpretations and supporting evidence,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1062–D1067, 2018, doi: 10.1093/nar/gkx1153.

[210] I. F. A. C. Fokkema, P. E. M. Taschner, G. C. P. Schaafsma, J. Celli, J. F. J. Laros, and J. T. den Dunnen, “LOVD v.2.0:

The next generation in gene variant databases," *Hum. Mutat.*, vol. 32, no. 5, pp. 557–563, 2011, doi: 10.1002/humu.21438.

[211] R. F. Grace and W. Barcellini, "Management of pyruvate kinase deficiency in children and adults," *Blood*, vol. 136, no. 11, pp. 1241–1249, 2020, doi: 10.1182/blood.2019000945.

[212] N. Luke, K. Hillier, H. Al-Samkari, and R. F. Grace, "Updates and advances in pyruvate kinase deficiency," *Trends Mol. Med.*, vol. 29, no. 5, pp. 406–418, 2023, doi: 10.1016/j.molmed.2023.02.005.

[213] D. C. Pendergrass, R. Williams, J. B. Blair, and A. W. Fenton, "Mining for allosteric information: Natural mutations and positional sequence conservation in pyruvate kinase," *IUBMB Life*, vol. 58, no. 1, pp. 31–38, 2006, doi: 10.1080/15216540500531705.

[214] D. Anastasiou *et al.*, "Pyruvate kinase M2 activators promote tetramer formation and suppress tumorigenesis," *Nat. Chem. Biol.*, vol. 8, no. 10, pp. 839–847, 2012, doi: 10.1038/nchembio.1060.

[215] Y. Matsui *et al.*, "Discovery and structure-guided fragment-linking of 4-(2,3-dichlorobenzoyl)-1-methyl-pyrrole-2-carboxamide as a pyruvate kinase M2 activator," *Bioorg. Med. Chem.*, vol. 25, no. 13, pp. 3540–3546, 2017, doi: 10.1016/j.bmc.2017.05.004.

[216] V. López-Ferrando, A. Gazzo, X. De La Cruz, M. Orozco, and J. L. Gelpí, "PMut: A web-based tool for the annotation of pathological variants on proteins, 2017 update," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W222–W228, 2017, doi: 10.1093/nar/gkx313.

[217] J. Aguirre, N. Padilla, S. Özkan, C. Riera, L. Feliubadaló, and X. De La Cruz, "Choosing variant interpretation tools for clinical applications: context matters," *bioRxiv*, 2022, doi: 10.1101/2022.02.17.480823.

[218] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, 2003, doi: 10.1093/nar/gkg509.

[219] P. D. Thomas, D. Ebert, A. Muruganujan, T. Mushayahama, L. P. Albou, and H. Mi, "PANTHER: Making genome-scale phylogenetics accessible to all," *Protein Sci.*, vol. 31, no. 1, pp. 8–22, 2022, doi: 10.1002/pro.4218.

[220] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: Application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, p. e118, 2011, doi: 10.1093/nar/gkr407.

[221] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the Functional Effect of Amino Acid Substitutions and Indels," *PLoS One*, vol. 7, no. 10, p. e46688, 2012, doi: 10.1371/journal.pone.0046688.

[222] H. A. Shihab *et al.*, "Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models," *Hum. Mutat.*, vol. 34, no. 1, pp. 57–65, 2013, doi: 10.1002/humu.22225.

[223] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.

[224] M. Varadi *et al.*, "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D439–D444, 2022, doi: 10.1093/nar/gkab1061.

[225] V. Pejaver *et al.*, "Inferring the molecular and phenotypic impact of amino acid variants with MutPred2," *Nat. Commun.*, vol. 11, no. 1, pp. 1–13, 2020, doi: 10.1038/s41467-020-19669-x.

[226] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, "Functional annotations improve the predictive score of human disease-related mutations in proteins," *Hum. Mutat.*, vol. 30, no. 8, pp. 1237–1244, 2009, doi: 10.1002/humu.21047.

[227] I. A. Adzhubei *et al.*, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, no. 4, pp. 248–249, 2010, doi: 10.1038/nmeth0410-248.

[228] R. Steinhaus, S. Proft, M. Schuelke, D. N. Cooper, J. M. Schwarz, and D. Seelow, "MutationTaster2021," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W446–W451, 2021, doi: 10.1093/nar/gkab266.

[229] A. González-Pérez and N. López-Bigas, "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel," *Am. J. Hum. Genet.*, vol. 88, no. 4, pp. 440–449, 2011, doi: 10.1016/j.ajhg.2011.03.004.

[230] M. Kircher, D. M. Witten, P. Jain, B. J. O'roak, G. M. Cooper, and J. Shendure, "A general framework for estimating the relative pathogenicity of human genetic variants," *Nat. Genet.*, vol. 46, no. 3, pp. 310–315, 2014, doi:

10.1038/ng.2892.

[231] N. M. Ioannidis *et al.*, “REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants,” *Am. J. Hum. Genet.*, vol. 99, no. 4, pp. 877–885, 2016, doi: 10.1016/j.ajhg.2016.08.016.

[232] N. Alirezaie, K. D. Kernohan, T. Hartley, J. Majewski, and T. D. Hocking, “ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants,” *Am. J. Hum. Genet.*, vol. 103, no. 4, pp. 474–483, 2018, doi: 10.1016/j.ajhg.2018.08.005.

[233] C. Riera, N. Padilla, and X. de la Cruz, “The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions,” *Hum. Mutat.*, vol. 37, no. 10, pp. 1013–1024, 2016, doi: 10.1002/humu.23048.

[234] N. Fleming, B. Kinsella, and C. Ing, “Predicting Protein Thermostability Upon Mutation Using Molecular Dynamics Timeseries Data,” *bioRxiv*, 2016, doi: 10.1101/078246.

[235] C. G. Priya Doss, C. Chakraborty, L. Chen, and H. Zhu, “Integrating in silico prediction methods, molecular docking, and molecular dynamics simulation to predict the impact of ALK missense mutations in structural perspective,” *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/895831.

[236] A. Kumar and R. Purohit, “Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs,” *PLoS Comput. Biol.*, vol. 10, no. 4, p. e1003318, 2014, doi: 10.1371/journal.pcbi.1003318.

[237] V. E. Angarica, M. Orozco, and J. Sancho, “Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: Linking SNPs with disease phenotypes in familial hypercholesterolemia,” *Hum. Mol. Genet.*, vol. 25, no. 6, pp. 1233–1246, 2016, doi: 10.1093/hmg/ddw004.

[238] L. M. Espinoza-Fonseca, “Pathogenic mutation R959W alters recognition dynamics of dysferlin inner DysF domain,” *Mol. Biosyst.*, vol. 12, no. 3, pp. 973–981, 2016, doi: 10.1039/c5mb00772k.

[239] B. Pandey *et al.*, “Alanine mutation of the catalytic sites of Pantothenate Synthetase causes distinct conformational changes in the ATP binding region,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–13, 2018, doi: 10.1038/s41598-017-19075-2.

[240] V. Frappier, M. Chartier, and R. J. Najmanovich, “ENCoM server: Exploring protein conformational space and the effect of mutations on protein function and stability,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W395–W400, 2015, doi: 10.1093/nar/gkv343.

[241] C. H. M. Rodrigues, D. E. V. Pires, and D. B. Ascher, “DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability,” *Nucleic Acids Res.*, vol. 46, no. W1, pp. W350–W355, 2018, doi: 10.1093/nar/gky300.

[242] A. K. Padhi, S. V. Vasaikar, B. Jayaram, and J. Gomes, “ANGDelMut - a web-based tool for predicting and analyzing functional loss mechanisms of amyotrophic lateral sclerosis-associated angiogenin mutations,” *F1000Research*, vol. 2, p. 227, 2013, doi: 10.12688/f1000research.2-227.v3.

[243] P. Schadzek *et al.*, “Data of the molecular dynamics simulations of mutations in the human connexin46 docking interface,” *Data Br.*, vol. 7, pp. 93–99, 2016, doi: 10.1016/j.dib.2016.01.067.

[244] J. A. Bauer, Ľ. Borko, J. Pavlović, E. Kutejová, and V. Bauerová-Hlinková, “Disease-associated mutations alter the dynamic motion of the N-terminal domain of the human cardiac ryanodine receptor,” *J. Biomol. Struct. Dyn.*, vol. 38, no. 4, pp. 1054–1070, 2019, doi: 10.1080/07391102.2019.1600027.

[245] H. Nagarajan, S. Narayanaswamy, and U. Vetrivel, “Mutational landscape screening of methylene tetrahydrofolate reductase to predict homocystinuria associated variants: An integrative computational approach,” *Mutat. Res. - Fundam. Mol. Mech. Mutagen.*, vol. 819–820, p. 111687, 2020, doi: 10.1016/j.mrfmmm.2020.111687.

[246] A. Shafie *et al.*, “Investigating single amino acid substitutions in PIM1 kinase: A structural genomics approach,” *PLoS One*, vol. 16, no. 10, p. e0258929, 2021, doi: 10.1371/journal.pone.0258929.

[247] D. E. Elmore and D. A. Dougherty, “Molecular dynamics simulations of wild-type and mutant forms of the Mycobacterium tuberculosis MscL channel,” *Biophys. J.*, vol. 81, no. 3, pp. 1345–1359, 2001, doi: 10.1016/S0006-3495(01)75791-8.

[248] M. W. van der Kamp *et al.*, “Dynameomics: A Comprehensive Database of Protein Dynamics,” *Structure*, vol. 18, no. 4, pp. 423–435, 2010, doi: 10.1016/j.str.2010.01.012.

[249] K. Lindorff-Larsen and J. Ferkinghoff-Borg, “Similarity measures for protein ensembles,” *PLoS One*, vol. 4, no. 1, p. e4203, 2009, doi: 10.1371/journal.pone.0004203.

[250] R. O. Dror, M. Ø. Jensen, D. W. Borhani, and D. E. Shaw, “Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations,” *J. Gen. Physiol.*, vol. 135, no. 6, pp. 555–562, 2010, doi: 10.1085/jgp.200910373.

[251] H. W. Wang and J. W. Wang, “How cryo-electron microscopy and X-ray crystallography complement each other,” *Protein Sci.*, vol. 26, no. 1, pp. 32–39, 2017, doi: 10.1002/pro.3022.

[252] A. Srivastava, T. Nagai, A. Srivastava, O. Miyashita, and F. Tama, “Role of computational methods in going beyond x-ray crystallography to explore protein structure and dynamics,” *Int. J. Mol. Sci.*, vol. 19, no. 11, p. 3401, 2018, doi: 10.3390/ijms19113401.

[253] A. A. Kermani, “A guide to membrane protein X-ray crystallography,” *FEBS J.*, vol. 288, no. 20, pp. 5788–5804, 2021, doi: 10.1111/febs.15676.

[254] Y. Tsuchiya, H. Nakamura, and K. Kinoshita, “Discrimination between biological interfaces and crystal-packing contacts,” *Adv. Appl. Bioinforma. Chem.*, vol. 1, pp. 99–113, 2012.

[255] L. C. Johansson, B. Stauch, A. Ishchenko, and V. Cherezov, “A Bright Future for Serial Femtosecond Crystallography with XFELs,” *Trends Biochem. Sci.*, vol. 42, no. 9, pp. 749–762, 2017, doi: 10.1016/j.tibs.2017.06.007.

[256] M. Carroni and H. R. Saibil, “Cryo electron microscopy to determine the structure of macromolecular complexes,” *Methods*, vol. 95, pp. 78–85, 2016, doi: 10.1016/j.ymeth.2015.11.023.

[257] D. Ban, T. M. Sabo, C. Griesinger, and D. Lee, “Measuring dynamic and kinetic information in the previously inaccessible supra-τc window of nanoseconds to microseconds by solution NMR spectroscopy,” *Molecules*, vol. 18, no. 10, pp. 11904–11937, 2013, doi: 10.3390/molecules181011904.

[258] E. B. Gibbs, E. C. Cook, and S. A. Showalter, “Application of NMR to studies of intrinsically disordered proteins,” *Arch. Biochem. Biophys.*, vol. 628, pp. 57–70, 2017, doi: 10.1016/j.abb.2017.05.008.

[259] L. Konermann, J. Pan, and Y. H. Liu, “Hydrogen exchange mass spectrometry for studying protein structure and dynamics,” *Chem. Soc. Rev.*, vol. 40, no. 3, pp. 1224–1234, 2011, doi: 10.1039/c0cs00113a.

[260] A. Leitner, M. Faini, F. Stengel, and R. Aebersold, “Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines,” *Trends Biochem. Sci.*, vol. 41, no. 1, pp. 20–32, 2016, doi: 10.1016/j.tibs.2015.10.008.

[261] G. Haran and H. Mazal, “How fast are the motions of tertiary-structure elements in proteins?,” *J. Chem. Phys.*, vol. 153, no. 13, p. 130902, 2020, doi: 10.1063/5.0024972.

[262] L. Makowski, “Characterization of proteins with wide-angle X-ray solution scattering (WAXS),” *J. Struct. Funct. Genomics*, vol. 11, no. 1, pp. 9–19, 2010, doi: 10.1007/s10969-009-9075-x.

[263] A. Barth, “Infrared spectroscopy of proteins,” *Biochim. Biophys. Acta - Bioenerg.*, vol. 1767, no. 9, pp. 1073–1101, 2007, doi: 10.1016/j.bbabio.2007.06.004.

[264] M. Gaczynska and P. A. Osmulski, “AFM of biological complexes: What can we learn?,” *Curr. Opin. Colloid Interface Sci.*, vol. 13, no. 5, pp. 351–367, 2008, doi: 10.1016/j.cocis.2008.01.004.

[265] M. T. Muhammed and E. Aki-Yalcin, “Homology modeling in drug discovery: Overview, current applications, and future perspectives,” *Chem. Biol. Drug Des.*, vol. 93, no. 1, pp. 12–20, 2019, doi: 10.1111/cbdd.13388.

[266] C. N. Cavasotto and S. S. Phatak, “Homology modeling in drug discovery: current trends and applications,” *Drug Discov. Today*, vol. 14, no. 13–14, pp. 676–683, 2009, doi: 10.1016/j.drudis.2009.04.006.

[267] S. Genheden, A. Reymer, P. Saenz-Méndez, and L. A. Eriksson, “Chapter 1. Computational Chemistry and Molecular Modelling Basics,” in *Computational Tools for Chemical Biology*, 2017, pp. 1–38.

[268] A. Šali and T. L. Blundell, “Comparative protein modelling by satisfaction of spatial restraints,” *J. Mol. Biol.*, vol. 234, no. 3, pp. 779–815, 1993, doi: 10.1006/jmbi.1993.1626.

[269] Y. Zhang, “I-TASSER server for protein 3D structure prediction,” *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–8, 2008, doi: 10.1186/1471-2105-9-40.

[270] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, “SWISS-MODEL: An automated protein homology-modeling server,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3381–3385, 2003, doi: 10.1093/nar/gkg520.

[271] B. Adhikari and J. Cheng, “Protein residue contacts and prediction methods,” in *Data Mining Techniques for the Life Sciences. Methods in Molecular Biology*, vol. 1415, 2016, pp. 463–476.

[272] M. Torrisi, G. Pollastri, and Q. Le, “Deep learning methods in protein structure prediction,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1301–1310, 2020, doi: 10.1016/j.csbj.2019.12.011.

[273] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack, “Improved prediction of protein side-chain conformations with SCWRL4,” *Proteins Struct. Funct. Bioinforma.*, vol. 77, no. 4, pp. 778–795, 2009, doi: 10.1002/prot.22488.

[274] P. Andrio *et al.*, “BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows,” *Sci. data*, vol. 6, no. 1, 2019, doi: 10.1038/s41597-019-0177-4.

[275] P. Saura, M. Röpke, A. P. Gamiz-Hernandez, and V. R. I. Kaila, “Quantum Chemical and QM/MM Models in Biochemistry,” in *Biomolecular Simulations. Methods in Molecular Biology*, vol. 2022, 2019, pp. 75–104.

[276] L. Bertini *et al.*, “Quantum mechanical methods for the investigation of metalloproteins and related bioinorganic compounds,” in *Metalloproteins. Methods in Molecular Biology*, vol. 1122, Humana Press Inc., 2014, pp. 207–268.

[277] F. Himo, “Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions,” *J. Am. Chem. Soc.*, vol. 139, no. 20, pp. 6780–6786, 2017, doi: 10.1021/jacs.7b02671.

[278] W. Kohn, A. D. Becke, and R. G. Parr, “Density functional theory of electronic structure,” *J. Phys. Chem.*, vol. 100, no. 31, pp. 12974–12980, 1996, doi: 10.1021/jp960669l.

[279] M. T. Stiebritz and Y. Hu, “Computational methods for modeling metalloproteins,” in *Metalloproteins. Methods in Molecular Biology*, vol. 1876, 2019, pp. 245–266.

[280] A. J. Thakkar, *Quantum Chemistry: A concise introduction for students of physics, chemistry, biochemistry and materials science*, Second Edi. 2017.

[281] S. Ahmadi, L. Barrios Herrera, M. Chehelamirani, J. Hostaš, S. Jalife, and D. R. Salahub, “Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review,” *Int. J. Quantum Chem.*, vol. 118, no. 9, pp. 1–34, 2018, doi: 10.1002/qua.25558.

[282] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, “Systematic validation of protein force fields against experimental data,” *PLoS One*, vol. 7, no. 2, p. e32131, 2012, doi: 10.1371/journal.pone.0032131.

[283] E. Papaleo, “Integrating atomistic molecular dynamics simulations, experiments, and network analysis to study protein dynamics: Strength in unity,” *Front. Mol. Biosci.*, vol. 2, p. 28, 2015, doi: 10.3389/fmolb.2015.00028.

[284] P. S. Nerenberg and T. Head-Gordon, “New developments in force fields for biomolecular simulations,” *Curr. Opin. Struct. Biol.*, vol. 49, pp. 129–138, 2018, doi: 10.1016/j.sbi.2018.02.002.

[285] R. O. Dror, R. M. Dirks, J. P. Grossman, H. Xu, and D. E. Shaw, “Biomolecular simulation: A computational microscope for molecular biology,” *Annual Review of Biophysics*, vol. 41, no. 1. pp. 429–452, 2012, doi: 10.1146/annurev-biophys-042910-155245.

[286] M. W. Van Der Kamp and A. J. Mulholland, “Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology,” *Biochemistry*, vol. 52, no. 16, pp. 2708–2728, 2013, doi: 10.1021/bi400215w.

[287] S. Takada, “Coarse-grained molecular simulations of large biomolecules,” *Curr. Opin. Struct. Biol.*, vol. 22, no. 2, pp. 130–137, 2012, doi: 10.1016/j.sbi.2012.01.010.

[288] M. M. Tirion, “Large amplitude elastic motions in proteins from a single-parameter, atomic analysis,” *Phys. Rev. Lett.*, vol. 77, no. 9, p. 1905, 1996, doi: 10.1103/PhysRevLett.77.1905.

[289] C. Clementi, H. Nymeyer, and J. N. Onuchic, “Topological and energetic factors: What determines the structural details of the transition state ensemble and ‘en-route’ intermediates for protein folding? An investigation for small globular proteins,” *J. Mol. Biol.*, vol. 298, no. 5, pp. 937–953, 2000, doi: 10.1006/jmbi.2000.3693.

[290] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S. J. Marrink, “The MARTINI coarse-grained force field: Extension to proteins,” *J. Chem. Theory Comput.*, vol. 4, no. 5, pp. 819–834, 2008, doi: 10.1021/ct700324x.

[291] L. Skjaerven, S. M. Hollup, and N. Reuter, “Normal mode analysis for proteins,” *J. Mol. Struct. THEOCHEM*, vol. 898, no. 1–3, pp. 42–48, 2009, doi: 10.1016/j.theochem.2008.09.024.

[292] S. Hayward and B. L. De Groot, “Normal modes and essential dynamics,” in *Molecular Modeling of Proteins. Methods in Molecular Biology*, vol. 443, 2008, pp. 89–106.

[293] L. Skjaerven, A. Martinez, and N. Reuter, “Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit,” *Proteins Struct. Funct. Bioinforma.*, vol. 79, no. 1, pp. 232–243, 2011, doi: 10.1002/prot.22875.

[294] D. J. Earl and M. W. Deem, “Monte Carlo simulations,” in *Molecular Modeling of Proteins. Methods in Molecular Biology*, vol. 443, 2008, pp. 25–36.

[295] K. Roy, S. Kar, and R. N. Das, “Chapter 5 – Computational Chemistry,” in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*, 2015, pp. 151–189.

[296] B. Gautam, “Energy Minimization,” in *Homology Molecular Modeling - Perspectives and Applications*, 2020.

[297] M. Orozco and F. J. Luque, “Theoretical methods for the description of the solvent effect in biomolecular systems,” *Chem. Rev.*, vol. 100, no. 11, pp. 4187–4226, 2000, doi: 10.1021/cr990052a.

[298] P. Dauber-Osguthorpe and A. T. Hagler, “Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there?,” *J. Comput. Aided. Mol. Des.*, vol. 33, no. 2, pp. 133–203, 2019, doi: 10.1007/s10822-018-0111-4.

[299] R. Anandakrishnan, A. Drozdetski, R. C. Walker, and A. V. Onufriev, “Speed of conformational change: Comparing explicit and implicit solvent molecular dynamics simulations,” *Biophys. J.*, vol. 108, no. 5, pp. 1153–1164, 2015, doi: 10.1016/j.bpj.2014.12.047.

[300] E. H. Lee, J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten, “Discovery Through the Computational Microscope,” *Structure*, vol. 17, no. 10, pp. 1295–1306, 2009, doi: 10.1016/j.str.2009.09.001.

[301] J. A. McCammon, B. R. Gelin, and M. Karplus, “Dynamics of folded proteins,” *Nature*, vol. 267, no. 5612, 1977, doi: 10.1038/267585a0.

[302] M. Karplus and J. A. McCammon, “Molecular dynamics simulations of biomolecules,” *Nat. Struct. Biol.*, vol. 9, no. 9, pp. 646–652, 2002, doi: 10.1038/nsb0902-646.

[303] M. Gecht, M. Siggel, M. Linke, G. Hummer, and J. Köfinger, “MDBenchmark: A toolkit to optimize the performance of molecular dynamics simulations,” *J. Chem. Phys.*, vol. 153, no. 144105, 2020, doi: 10.1063/5.0019045.

[304] A. Hospital, F. Battistini, R. Soliva, J. L. Gelpí, and M. Orozco, “Surviving the deluge of biosimulation data,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 10, no. 3, 2020, doi: 10.1002/wcms.1449.

[305] S. A. Hollingsworth and R. O. Dror, “Molecular Dynamics Simulation for All,” *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018, doi: 10.1016/j.neuron.2018.08.011.

[306] C. L. Brooks III, D. A. Case, S. Plimpton, B. Roux, D. Van Der Spoel, and E. Tajkhorshid, “Classical molecular dynamics,” *Journal of Chemical Physics*, vol. 154, no. 10. American Institute of Physics Inc., p. 100401, Mar. 14, 2021, doi: 10.1063/5.0045455.

[307] P. L. Freddolino and K. Schulten, “Common structural transitions in explicit-solvent simulations of villin headpiece folding,” *Biophys. J.*, vol. 97, no. 8, pp. 2338–2347, 2009, doi: 10.1016/j.bpj.2009.08.012.

[308] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, “Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39),” *J. Am. Chem. Soc.*, vol. 132, no. 5, pp. 1526–1528, 2010, doi: 10.1021/ja9090353.

[309] M. A. Young, S. Gonfloni, G. Superti-Furga, B. Roux, and J. Kuriyan, “Dynamic coupling between the SH2 and SH3 domains of c-Src and Hck underlies their inactivation by C-Terminal tyrosine phosphorylation,” *Cell*, vol. 105, no. 1, pp. 115–126, 2001, doi: 10.1016/S0092-8674(01)00301-4.

[310] J. Ma, P. B. Sigler, Z. Xu, and M. Karplus, “A dynamic model for the allosteric mechanism of GroEL,” *J. Mol. Biol.*, vol. 302, no. 2, pp. 303–313, 2000, doi: 10.1006/jmbi.2000.4014.

[311] S. Piana and D. E. Shaw, “Atomic-Level Description of Protein Folding inside the GroEL Cavity,” *J. Phys. Chem. B*, vol. 122, no. 49, pp. 11440–11449, 2018, doi: 10.1021/acs.jpcb.8b07366.

[312] J. R. Perilla *et al.*, “Molecular dynamics simulations of large macromolecular complexes,” *Curr. Opin. Struct. Biol.*, vol. 31, pp. 64–74, 2015, doi: 10.1016/j.sbi.2015.03.007.

[313] L. V. Bock *et al.*, “Energy barriers and driving forces in tRNA translocation through the ribosome,” *Nat. Struct. Mol. Biol.*, vol. 20, no. 12, pp. 1390–1396, 2013, doi: 10.1038/nsmb.2690.

[314] D. E. Chandler, J. Strümpfer, M. Sener, S. Scheuring, and K. Schulten, “Light harvesting by lamellar chromatophores in Rhodospirillum photometricum,” *Biophys. J.*, vol. 106, no. 11, pp. 2503–2510, 2014, doi:

10.1016/j.bpj.2014.04.030.

[315] C. Boiteux, S. Kraszewski, C. Ramseyer, and C. Girardet, “Ion conductance vs. pore gating and selectivity in KcsA channel: Modeling achievements and perspectives,” *J. Mol. Model.*, vol. 13, no. 6–7, pp. 699–713, 2007, doi: 10.1007/s00894-007-0202-y.

[316] K. Tai, T. Shen, U. Börjesson, M. Philippopoulos, and J. A. McCammon, “Analysis of a 10-ns molecular dynamics simulation of mouse acetylcholinesterase,” *Biophys. J.*, vol. 81, no. 2, pp. 715–724, 2001, doi: 10.1016/S0006-3495(01)75736-0.

[317] R. O. Dror *et al.*, “Structural basis for nucleotide exchange in heterotrimeric G proteins,” *Science (80-.).*, vol. 348, no. 6241, pp. 1361–1365, 2015, doi: 10.1126/science.aaa5264.

[318] N. Špačková, I. Berger, and J. Šponer, “Nanosecond molecular dynamics simulations of parallel and antiparallel guanine quadruplex DNA molecules,” *J. Am. Chem. Soc.*, vol. 121, no. 23, pp. 5519–5534, 1999, doi: 10.1021/ja984449s.

[319] R. Štefl *et al.*, “Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substates,” *Biophys. J.*, vol. 85, no. 3, pp. 1787–1804, 2003, doi: 10.1016/S0006-3495(03)74608-6.

[320] B. Ma and A. J. Levine, “Probing potential binding modes of the p53 tetramer to DNA based on the symmetries encoded in p53 response elements,” *Nucleic Acids Res.*, vol. 35, no. 22, pp. 7733–7747, 2007, doi: 10.1093/nar/gkm890.

[321] I. Buch, T. Giorgino, and G. De Fabritiis, “Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 25, pp. 10184–10189, 2011, doi: 10.1073/pnas.1103547108.

[322] N. X. Wang and J. J. Zheng, “Computational studies of H5N1 influenza virus resistance to oseltamivir,” *Protein Sci.*, vol. 18, no. 4, pp. 707–715, 2009, doi: 10.1002/pro.77.

[323] M. Wieczór *et al.*, “Pre-exascale HPC approaches for molecular dynamics simulations. Covid-19 research: A use case,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 13, no. 1, p. e1622, 2023, doi: 10.1002/wcms.1622.

[324] R. E. Amaro and A. J. Mulholland, “A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19,” *J. Chem. Inf. Model.*, vol. 60, no. 6, pp. 2653–2656, 2020, doi: 10.1021/acs.jcim.0c00319.

[325] M. Torrens-Fontanals, A. Peralta-García, C. Talarico, R. Guixà-González, T. Giorgino, and J. Selent, “SCoV2-MD: A database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions,” *Nucleic Acids Res.*, vol. 50, no. D1, pp. D858–D866, 2022, doi: 10.1093/nar/gkab977.

[326] J. Huang and A. D. MacKerell, “Force field development and simulations of intrinsically disordered proteins,” *Curr. Opin. Struct. Biol.*, vol. 48, pp. 40–48, 2018, doi: 10.1016/j.sbi.2017.10.008.

[327] W. D. Cornell *et al.*, “A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules,” *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, 1995, doi: 10.1021/ja00124a002.

[328] A. D. MacKerell *et al.*, “All-atom empirical potential for molecular modeling and dynamics studies of proteins,” *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, 1998, doi: 10.1021/jp973084f.

[329] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, “Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids,” *J. Am. Chem. Soc.*, vol. 118, no. 45, pp. 11225–11236, 1996, doi: 10.1021/ja9621760.

[330] S. Piana, J. L. Klepeis, and D. E. Shaw, “Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations,” *Curr. Opin. Struct. Biol.*, vol. 24, pp. 98–105, 2014, doi: 10.1016/j.sbi.2013.12.006.

[331] J. Behler, “Perspective: Machine learning potentials for atomistic simulations,” *J. Chem. Phys.*, vol. 145, no. 17, 2016, doi: 10.1063/1.4966192.

[332] A. Grossfield, S. E. Feller, and M. C. Pitman, “Convergence of molecular dynamics simulations of membrane proteins,” *Proteins Struct. Funct. Genet.*, vol. 67, no. 1, pp. 31–40, 2007, doi: 10.1002/prot.21308.

[333] E. Papaleo, P. Mereghetti, P. Fantucci, R. Grandori, and L. De Gioia, “Free-energy landscape, principal component analysis, and structural clustering to identify representative conformations from molecular dynamics simulations: The myoglobin case,” *J. Mol. Graph. Model.*, vol. 27, no. 8, pp. 889–899, 2009, doi: 10.1016/j.jmgm.2009.01.006.

[334] G. Paris, C. Ramseyer, and M. Enescu, "A principal component analysis of the dynamics of subdomains and binding sites in human serum albumin," *Biopolymers*, vol. 101, no. 5, pp. 561–572, 2014, doi: 10.1002/bip.22418.

[335] B. Knapp, L. Ospina, and C. M. Deane, "Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas," *J. Chem. Theory Comput.*, vol. 14, no. 12, pp. 6127–6138, 2018, doi: 10.1021/acs.jctc.8b00391.

[336] G. Zhao *et al.*, "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics," *Nature*, vol. 497, no. 7451, pp. 643–646, 2013, doi: 10.1038/nature12162.

[337] A. Singharoy *et al.*, "Atoms to Phenotypes: Molecular Design Principles of Cellular Energy Metabolism," *Cell*, vol. 179, no. 5, pp. 1098–1111, 2019, doi: 10.1016/j.cell.2019.10.021.

[338] D. E. Shaw *et al.*, "Anton, a special-purpose machine for molecular dynamics simulation," *Commun. ACM*, vol. 51, no. 7, p. 91, 2008, doi: 10.1145/1364782.1364802.

[339] I. Ohmura, G. Morimoto, Y. Ohno, A. Hasegawa, and M. Taiji, "MDGRAPE-4: A special-purpose computer system formolecular dynamics simulations," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 372, no. 2021, p. 20130387, 2014, doi: 10.1098/rsta.2013.0387.

[340] D. E. Shaw, "Millisecond-Long Molecular Dynamics Simulations of Proteins on a Special-Purpose Machine," *Biophys. J.*, vol. 104, no. 2, p. 45A, 2013, doi: 10.1016/j.bpj.2012.11.289.

[341] D. E. Shaw *et al.*, "Anton 3: Twenty Microseconds of Molecular Dynamics Simulation before Lunch," 2021, doi: 10.1145/3458817.3487397.

[342] S. Singh and V. K. Singh, "Molecular Dynamics Simulation: Methods and Application," in *Frontiers in Protein Structure, Function, and Dynamics*, 2020, pp. 213–238.

[343] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," *J. Comput. Chem.*, vol. 29, no. 11, pp. 1859–1865, 2008, doi: 10.1002/jcc.20945.

[344] A. Hospital, P. Andrio, C. Fenollosa, D. Cicin-Sain, M. Orozco, and J. L. Gelpí, "MDWeb and MDMoby: An integrated web-based platform for molecular dynamics simulations," *Bioinformatics*, vol. 28, no. 9, pp. 1278–1279, 2012, doi: 10.1093/bioinformatics/bts139.

[345] J. V. Ribeiro *et al.*, "QwikMD - Integrative Molecular Dynamics Toolkit for Novices and Experts," *Sci. Rep.*, vol. 6, p. 26536, 2016, doi: 10.1038/srep26536.

[346] S. Doerr, M. J. Harvey, F. Noé, and G. De Fabritiis, "HTMD: High-Throughput Molecular Dynamics for Molecular Discovery," *J. Chem. Theory Comput.*, vol. 12, no. 4, pp. 1845–1852, 2016, doi: 10.1021/acs.jctc.6b00049.

[347] G. Bayarri, P. Andrio, A. Hospital, M. Orozco, and J. L. Gelpí, "BioExcel Building Blocks Workflows (BioBB-Wfs), an integrated web-based platform for biomolecular simulations," *Nucleic Acids Res.*, vol. gkac380, 2022, doi: 10.1093/nar/gkac380.

[348] J. Lemkul, "From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]," *Living J. Comput. Mol. Sci.*, vol. 1, no. 1, p. 5068, 2019, doi: 10.33011/livecoms.1.1.5068.

[349] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, and D. M. Zuckerman, "Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]," *Living J. Comput. Mol. Sci.*, vol. 1, no. 1, p. 5067, 2018, doi: 10.33011/livecoms.1.1.5067.

[350] E. Braun *et al.*, "Best Practices for Foundations in Molecular Simulations [Article v1.0]," *Living J. Comput. Mol. Sci.*, vol. 1, pp. 1–28, 2019, doi: 10.33011/livecoms.1.1.5957.

[351] D. R. Roe and B. R. Brooks, "A protocol for preparing explicitly solvated systems for stable molecular dynamics simulations," *J. Chem. Phys.*, vol. 153, no. 5, 2020, doi: 10.1063/5.0013849.

[352] M. J. Abraham *et al.*, "Sharing Data from Molecular Simulations," *J. Chem. Inf. Model.*, vol. 59, no. 10, pp. 4093–4099, 2019, doi: 10.1021/acs.jcim.9b00665.

[353] P. W. Hildebrand, A. S. Rose, and J. K. S. Tiemann, "Bringing Molecular Dynamics Simulation Data into View," *Trends in Biochemical Sciences*, vol. 44, no. 11. Elsevier Ltd, pp. 902–913, 2019, doi: 10.1016/j.tibs.2019.06.004.

[354] A. Elofsson, B. Hess, E. Lindahl, A. Onufriev, D. van der Spoel, and A. Wallqvist, "Ten simple rules on how to create open access and reproducible molecular simulations of biological systems," *PLoS Comput. Biol.*, vol. 15, no. 1, p. e1006649, 2019, doi: 10.1371/journal.pcbi.1006649.

[355] K. Tai *et al.*, “BioSimGrid: Towards a worldwide repository for biomolecular simulations,” *Org. Biomol. Chem.*, vol. 2, no. 22, pp. 3219–3221, 2004, doi: 10.1039/b411352g.

[356] T. Meyer *et al.*, “MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories,” *Structure*, vol. 18, no. 11, pp. 1399–1409, 2010, doi: 10.1016/j.str.2010.07.013.

[357] A. Hospital *et al.*, “BIGNASim: A NoSQL database structure and analysis portal for nucleic acids simulation data,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D272–D278, 2016, doi: 10.1093/nar/gkv1301.

[358] R. Sun, Z. Li, and T. C. Bishop, “TMB-iBIOMES: An iBIOMES-Lite database of Nucleosome Trajectories and meta-analysis,” *ChemRxiv*, 2019.

[359] I. Rodríguez-Espigares *et al.*, “GPCRmd uncovers the dynamics of the 3D-GPCRome,” *Nat. Methods*, vol. 17, no. 8, pp. 777–787, 2020, doi: 10.1038/s41592-020-0884-y.

[360] M. D. Wilkinson *et al.*, “Comment: The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, no. 1, pp. 1–9, 2016, doi: 10.1038/sdata.2016.18.

[361] M. Wieczór, A. Hospital, G. Bayarri, J. Czub, and M. Orozco, “Molywood: Streamlining the design and rendering of molecular movies,” *Bioinformatics*, vol. 36, no. 17, pp. 4660–4661, 2020, doi: 10.1093/bioinformatics/btaa584.

[362] G. Bayarri, A. Hospital, and M. Orozco, “3dRS, a Web-Based Tool to Share Interactive Representations of 3D Biomolecular Structures and Molecular Dynamics Trajectories,” *Front. Mol. Biosci.*, vol. 8, p. 726232, 2021, doi: 10.3389/fmolb.2021.726232.

[363] T. D. Romo and A. Grossfield, “Block covariance overlap method and convergence in molecular dynamics simulation,” *J. Chem. Theory Comput.*, vol. 7, no. 8, pp. 2464–2472, 2011, doi: 10.1021/ct2002754.

[364] L. Sawle and K. Ghosh, “Convergence of Molecular Dynamics Simulation of Protein Native States: Feasibility vs Self-Consistency Dilemma,” *J. Chem. Theory Comput.*, vol. 12, no. 2, pp. 861–869, 2016, doi: 10.1021/acs.jctc.5b00999.

[365] P. V. Coveney and S. Wan, “On the calculation of equilibrium thermodynamic properties from molecular dynamics,” *Phys. Chem. Chem. Phys.*, vol. 18, no. 44, pp. 30236–30240, 2016, doi: 10.1039/c6cp02349e.

[366] Y. Sugita and Y. Okamoto, “Replica-exchange molecular dynamics method for protein folding,” *Chem. Phys. Lett.*, vol. 314, no. 1–2, pp. 141–151, 1999, doi: 10.1016/S0009-2614(99)01123-9.

[367] G. A. Huber and S. Kim, “Weighted-ensemble Brownian dynamics simulations for protein association reactions,” *Biophys. J.*, vol. 70, no. 1, pp. 97–110, 1996, doi: 10.1016/S0006-3495(96)79552-8.

[368] C. Micheletti, A. Laio, and M. Parrinello, “Reconstructing the density of states by history-dependent metadynamics,” *Phys. Rev. Lett.*, vol. 92, no. 17, p. 170601, 2004, doi: 10.1103/PhysRevLett.92.170601.

[369] J. Schlitter, M. Engels, P. KrüGer, E. Jacoby, and A. Wollmer, “Targeted molecular dynamics simulation of conformational change - application to the T R transition in insulin,” *Mol. Simul.*, vol. 10, no. 2–6, pp. 291–308, 1993, doi: 10.1080/08927029308022170.

[370] B. Isralewitz, S. Izrailev, and K. Schulten, “Binding pathway of retinal to bacterio-opsin: A prediction by molecular dynamics simulations,” *Biophys. J.*, vol. 73, no. 6, pp. 2972–2979, 1997, doi: 10.1016/S0006-3495(97)78326-7.

[371] L. Maragliano and E. Vanden-Eijnden, “A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations,” *Chem. Phys. Lett.*, vol. 426, no. 1–3, pp. 168–175, 2006, doi: 10.1016/j.cplett.2006.05.062.

[372] D. Hamelberg, J. Mongan, and J. A. McCammon, “Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules,” *J. Chem. Phys.*, vol. 120, no. 24, pp. 11919–11929, 2004, doi: 10.1063/1.1755656.

[373] W. L. DeLano, “The PyMOL Molecular Graphics System, Version 2.3,” *Schrödinger LLC*. 2020.

[374] S. L. Kazmirski, A. Li, and V. Daggett, “Analysis methods for comparison of multiple molecular dynamics trajectories: Applications to protein unfolding pathways and denatured ensembles,” *J. Mol. Biol.*, vol. 290, no. 1, pp. 283–304, 1999, doi: 10.1006/jmbi.1999.2843.

[375] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, “Relation between free energy landscapes of proteins and dynamics,” *J. Chem. Theory Comput.*, vol. 6, no. 2, pp. 583–595, 2010, doi: 10.1021/ct9005745.

[376] F. Sittel and G. Stock, “Perspective: Identification of collective variables and metastable states of protein dynamics,” *J. Chem. Phys.*, vol. 149, no. 15, p. 150901, 2018, doi: 10.1063/1.5049637.

[377] M. Bernetti, M. Bertazzo, and M. Masetti, “Data-driven molecular dynamics: A multifaceted challenge,” *Pharmaceuticals*, vol. 13, no. 9, 2020, doi: 10.3390/ph13090253.

[378] I. Daidone and A. Amadei, “Essential dynamics: Foundation and applications,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 2, no. 5, pp. 762–770, 2012, doi: 10.1002/wcms.1099.

[379] C. C. David and D. J. Jacobs, “Principal component analysis: A method for determining the essential dynamics of proteins,” in *Protein Dynamics. Methods in Molecular Biology*, vol. 1084, Humana Press Inc., 2014, pp. 193–226.

[380] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, “Unsupervised Learning Methods for Molecular Simulation Data,” *Chem. Rev.*, vol. 121, no. 16, pp. 9722–9758, 2021, doi: 10.1021/acs.chemrev.0c01195.

[381] C. Böde, I. A. Kovács, M. S. Szalay, R. Palotai, T. Korcsmáros, and P. Csermely, “Network analysis of protein dynamics,” *FEBS Lett.*, vol. 581, no. 15, pp. 2776–2782, 2007, doi: 10.1016/j.febslet.2007.05.021.

[382] Z. Liang, G. M. Verkhivker, and G. Hu, “Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications,” *Brief. Bioinform.*, vol. 21, no. 3, pp. 815–835, 2020, doi: 10.1093/bib/bbz029.

[383] L. Chebon-Bore, T. A. Sanyanga, C. V. Manyumwa, A. Khairallah, and Ö. T. Bishop, “Decoding the molecular effects of atovaquone linked resistant mutations on plasmodium falciparum cytb-isp complex in the phospholipid bilayer membrane,” *Int. J. Mol. Sci.*, vol. 22, no. 4, p. 2138, 2021, doi: 10.3390/ijms22042138.

[384] J. D. Chodera and F. Noé, “Markov state models of biomolecular conformational dynamics,” *Curr. Opin. Struct. Biol.*, vol. 25, pp. 135–144, 2014, doi: 10.1016/j.sbi.2014.04.002.

[385] L. Busto-Moner, C. J. Feng, A. Antoszewski, A. Tokmakoff, and A. R. Dinner, “Structural Ensemble of the Insulin Monomer,” *Biochemistry*, vol. 60, no. 42, pp. 3125–3136, 2021, doi: 10.1021/acs.biochem.1c00583.

[386] A. Kitao, “Principal Component Analysis and Related Methods for Investigating the Dynamics of Biological Macromolecules,” *J*, vol. 5, no. 2, pp. 298–317, 2022, doi: 10.3390/j5020021.

[387] R. M. Levy, A. R. Srinivasan, W. K. Olson, and J. A. McCammon, “Quasi-harmonic method for studying very low frequency modes in proteins,” *Biopolymers*, vol. 23, no. 6, pp. 1099–1112, 1984, doi: 10.1002/bip.360230610.

[388] T. Ichiye and M. Karplus, “Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations,” *Proteins Struct. Funct. Bioinforma.*, vol. 11, no. 3, pp. 205–217, 1991, doi: 10.1002/prot.340110305.

[389] A. Kitao, F. Hirata, and N. Go, “The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum,” *Chem. Phys.*, vol. 158, no. 2–3, pp. 447–472, 1991, doi: 10.1016/0301-0104(91)87082-7.

[390] A. E. García, “Large-amplitude nonlinear motions in proteins,” *Phys. Rev. Lett.*, vol. 68, no. 17, p. 2696, 1992, doi: 10.1103/PhysRevLett.68.2696.

[391] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, “Essential dynamics of proteins,” *Proteins Struct. Funct. Bioinforma.*, vol. 17, no. 4, pp. 412–425, 1993, doi: 10.1002/prot.340170408.

[392] D. M. F. Van Aalten, A. Amadei, A. B. M. Linssen, V. G. H. Eijsink, G. Vriend, and H. J. C. Berendsen, “The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water,” *Proteins Struct. Funct. Bioinforma.*, vol. 22, no. 1, pp. 45–54, 1995, doi: 10.1002/prot.340220107.

[393] A. E. García and J. G. Harman, “Simulations of CRP:(cAMP)2 in noncrystalline environments show a subunit transition from the open to the closed conformation,” *Protein Sci.*, vol. 5, no. 1, pp. 62–71, 1996, doi: 10.1002/pro.5560050108.

[394] Z. Zhang and W. Wriggers, “Local feature analysis: A statistical theory for reproducible essential dynamics of large macromolecules,” *Proteins Struct. Funct. Genet.*, vol. 64, no. 2, pp. 391–403, 2006, doi: 10.1002/prot.20983.

[395] M. A. Balsera, W. Wriggers, Y. Oono, and K. Schulten, “Principal component analysis and long time protein dynamics,” *J. Phys. Chem.*, vol. 100, no. 7, pp. 2567–2572, 1996, doi: 10.1021/jp9536920.

[396] A. Kitao and N. Go, “Investigating protein dynamics in collective coordinate space,” *Curr. Opin. Struct. Biol.*, vol. 9, no. 2, 1999, doi: 10.1016/S0959-440X(99)80023-2.

[397] B. Hess, “Similarities between principal components of protein dynamics and random diffusion,” *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 62, no. 6, pp. 8438–8448, Dec. 2000, doi:

10.1103/PhysRevE.62.8438.

[398] M. Lambrughi *et al.*, "Analyzing Biomolecular Ensembles," in *Biomolecular Simulations. Methods in Molecular Biology*, vol. 2022, 2019.

[399] M. B. Kubitzki and B. L. De Groot, "Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange," *Biophys. J.*, vol. 92, no. 12, pp. 4262–4270, 2007, doi: 10.1529/biophysj.106.103101.

[400] T. Meyer *et al.*, "Essential dynamics: A tool for efficient trajectory compression and management," *J. Chem. Theory Comput.*, vol. 2, no. 2, pp. 251–258, 2006, doi: 10.1021/ct050285b.

[401] R. Chaudhuri, O. Carrillo, C. A. Laughton, and M. Orozco, "Application of drug-perturbed essential dynamics/molecular dynamics (ED/MD) to virtual screening and rational drug design," *J. Chem. Theory Comput.*, vol. 8, no. 7, pp. 2204–2214, 2012, doi: 10.1021/ct300223c.

[402] L. Zanetti-Polzi, S. Corni, I. Daidone, and A. Amadei, "Extending the essential dynamics analysis to investigate molecular properties: Application to the redox potential of proteins," *Phys. Chem. Chem. Phys.*, vol. 18, no. 27, pp. 18450–18459, 2016, doi: 10.1039/c6cp03394f.

[403] F. Sittel, A. Jain, and G. Stock, "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates," *J. Chem. Phys.*, vol. 141, p. 014111, 2014, doi: 10.1063/1.4885338.

[404] C. C. David, E. R. A. Singam, and D. J. Jacobs, "JED: A Java Essential Dynamics Program for comparative analysis of protein trajectories," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–9, 2017, doi: 10.1186/s12859-017-1676-y.

[405] M. Ernst, F. Sittel, and G. Stock, "Contact- and distance-based principal component analysis of protein dynamics," *J. Chem. Phys.*, vol. 143, p. 244114, 2015, doi: 10.1063/1.4938249.

[406] Y. Mu, P. H. Nguyen, and G. Stock, "Energy landscape of a small peptide revealed by dihedral angle principal component analysis," *Proteins Struct. Funct. Genet.*, vol. 58, no. 1, pp. 45–52, 2005, doi: 10.1002/prot.20310.

[407] K. Sargsyan, J. Wright, and C. Lim, "GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics," *Nucleic Acids Res.*, vol. 40, no. 3, p. e25, 2012, doi: 10.1093/nar/gkr1069.

[408] B. Eltzner, S. Huckemann, and K. V. Mardia, "Torus principal component analysis with applications to RNA structure," *Ann. Appl. Stat.*, vol. 12, no. 2, pp. 1332–1359, 2018, doi: 10.1214/17-AOAS1115.

[409] H. Zhou, F. Wang, and P. Tao, "T-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations," *J. Chem. Theory Comput.*, vol. 14, no. 11, pp. 5499–5510, 2018, doi: 10.1021/acs.jctc.8b00652.

[410] F. Trozzi, X. Wang, and P. Tao, "UMAP as a Dimensionality Reduction Tool for Molecular Dynamics Simulations of Biomacromolecules: A Comparison Study," *J. Phys. Chem. B*, vol. 125, no. 19, pp. 5022–5034, 2021, doi: 10.1021/acs.jpcb.1c02081.

[411] P. H. Nguyen, "Complexity of free energy landscapes of peptides revealed by nonlinear principal component analysis," *Proteins Struct. Funct. Genet.*, vol. 65, no. 4, pp. 898–913, 2006, doi: 10.1002/prot.21185.

[412] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998, doi: 10.1162/089976698300017467.

[413] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952, doi: 10.1007/BF02288916.

[414] J. M. Troyer and F. E. Cohen, "Protein conformational landscapes: Energy minimization and clustering of a long molecular dynamics trajectory," *Proteins Struct. Funct. Bioinforma.*, vol. 23, no. 1, pp. 97–110, 1995, doi: 10.1002/prot.340230111.

[415] Y. Naritomi and S. Fuchigami, "Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: The case of domain motions," *J. Chem. Phys.*, vol. 134, no. 6, p. 065101, 2011, doi: 10.1063/1.3554380.

[416] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," *J. Chem. Phys.*, vol. 139, no. 1, p. 015102, 2013, doi: 10.1063/1.4811489.

[417] V. C. de Souza, L. Goliatt, and P. V. Z. Capriles, "Insight About Nonlinear Dimensionality Reduction Methods Applied to Protein Molecular Dynamics," in *Bioinformatics and Biomedical Engineering. International Work-Conference on*

Bioinformatics and Biomedical Engineering (IWBBIO). Lecture Notes in Computer Science, 2019, vol. 11466, pp. 219–239, doi: 10.1007/978-3-030-17935-9_21.

[418] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science (80-.).*, vol. 290, no. 5500, pp. 2319–2323, 2000, doi: 10.1126/science.290.5500.2319.

[419] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 26, pp. 9885–9890, 2006, doi: 10.1073/pnas.0603553103.

[420] D. K. Agrafiotis and H. Xu, "A self-organizing principle for learning nonlinear manifolds," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 25, pp. 15869–15872, 2002, doi: 10.1073/pnas.242424399.

[421] R. R. Coifman *et al.*, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 21, pp. 7426–7431, 2005, doi: 10.1073/pnas.0500334102.

[422] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, "Systematic determination of order parameters for chain dynamics using diffusion maps," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 31, pp. 13597–13602, 2010, doi: 10.1073/pnas.1003293107.

[423] O. F. Lange and H. Grubmüller, "Full correlation analysis of conformational protein dynamics," *Proteins Struct. Funct. Genet.*, vol. 70, no. 4, pp. 1294–1312, 2008, doi: 10.1002/prot.21618.

[424] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Simplifying the representation of complex free-energy landscapes using sketch-map," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 32, pp. 13023–13028, 2011, doi: 10.1073/pnas.1108486108.

[425] A. I. Albu and G. Czibula, "Analysing protein dynamics using machine learning based generative models," in *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 2020, pp. 000135–000140, doi: 10.1109/SACI49304.2020.9118834.

[426] W. Chen and A. L. Ferguson, "Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration," *J. Comput. Chem.*, vol. 39, no. 25, pp. 2079–2102, 2018, doi: 10.1002/jcc.25520.

[427] M. T. Degiacomi, "Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space," *Structure*, vol. 27, no. 6, pp. 1034–1040, 2019, doi: 10.1016/j.str.2019.03.018.

[428] O. Fleetwood, M. A. Kasimova, A. M. Westerlund, and L. Delemotte, "Molecular Insights from Conformational Ensembles via Machine Learning," *Biophys. J.*, vol. 118, no. 3, pp. 765–780, 2020, doi: 10.1016/j.bpj.2019.12.016.

[429] D. Bhowmik, S. Gao, M. T. Young, and A. Ramanathan, "Deep clustering of protein folding simulations," *BMC Bioinformatics*, vol. 19, no. 18, pp. 47–58, 2018, doi: 10.1186/s12859-018-2507-5.

[430] C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, "Variational encoding of complex dynamics," *Phys. Rev. E*, vol. 97, no. 6, p. 062412, 2018, doi: 10.1103/PhysRevE.97.062412.

[431] H. Tian, X. Jiang, F. Trozzi, S. Xiao, E. C. Larson, and P. Tao, "Explore Protein Conformational Space With Variational Autoencoder," *Front. Mol. Biosci.*, vol. 8, no. 781635, 2021, doi: 10.3389/fmolb.2021.781635.

[432] M. Tiberti, E. Papaleo, T. Bengtsen, W. Boomsma, and K. Lindorff-Larsen, "ENCORE: Software for Quantitative Ensemble Comparison," *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004415, 2015, doi: 10.1371/journal.pcbi.1004415.

[433] R. Brüschweiler, "Efficient RMSD measures for the comparison of two molecular ensembles," *Proteins Struct. Funct. Genet.*, vol. 50, no. 1, pp. 26–34, 2003, doi: 10.1002/prot.10250.

[434] M. Rueda *et al.*, "A consensus view of protein dynamics," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 3, pp. 796–801, 2007, doi: 10.1073/pnas.0605534104.

[435] K. C. Wolfe and G. S. Chirikjian, "Quantitative comparison of conformational ensembles," *Entropy*, vol. 14, no. 2, pp. 213–232, 2012, doi: 10.3390/e14020213.

[436] S. Yang, L. Salmon, and H. M. Al-Hashimi, "Measuring similarity between dynamic ensembles of biomolecules," *Nat. Methods*, vol. 11, no. 5, pp. 552–554, 2014, doi: 10.1038/nmeth.2921.

[437] J. Farmer, F. Kanwal, N. Nikulsin, M. C. B. Tsilimigras, and D. J. Jacobs, "Statistical measures to quantify similarity between molecular dynamics simulation trajectories," *Entropy*, vol. 19, no. 12, p. 646, 2017, doi: 10.3390/e19120646.

[438] B. Hess, “Convergence of sampling in protein simulations,” *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, vol. 65, no. 3, p. 031910, 2002, doi: 10.1103/PhysRevE.65.031910.

[439] X. Q. Yao and D. Hamelberg, “Detecting Functional Dynamics in Proteins with Comparative Perturbed-Ensembles Analysis,” *Acc. Chem. Res.*, vol. 52, no. 12, pp. 3455–3464, 2019, doi: 10.1021/acs.accounts.9b00485.

[440] M. Grosso, A. Kalstein, G. Parisi, A. E. Roitberg, and S. Fernandez-Alberti, “On the analysis and comparison of conformer-specific essential dynamics upon ligand binding to a protein,” *J. Chem. Phys.*, vol. 142, no. 24, p. 245101, 2015, doi: 10.1063/1.4922925.

[441] M. Martínez-Archundia, J. Correa-Basurto, S. Montaño, and J. L. Rosas-Trigueros, “Studying the collective motions of the adenosine A2A receptor as a result of ligand binding using principal component analysis,” *J. Biomol. Struct. Dyn.*, vol. 37, no. 18, pp. 4685–4700, 2019, doi: 10.1080/07391102.2018.1564700.

[442] Y. Li *et al.*, “Molecular dynamics simulations reveal distinct differences in conformational dynamics and thermodynamics between the unliganded and CD4-bound states of HIV-1 gp120,” *Phys. Chem. Chem. Phys.*, vol. 22, no. 10, pp. 5548–5560, 2020, doi: 10.1039/c9cp06706j.

[443] G. Pierdominici-Sottile and J. Palma, “New insights into the meaning and usefulness of principal component analysis of concatenated trajectories,” *J. Comput. Chem.*, vol. 36, no. 7, pp. 424–432, 2015, doi: 10.1002/jcc.23811.

[444] B. L. De Groot, S. Hayward, D. M. F. Van Aalten, A. Amadei, and H. J. C. Berendsen, “Domain motions in bacteriophage T4 lysozyme: A comparison between molecular dynamics and crystallographic data,” *Proteins Struct. Funct. Genet.*, vol. 31, no. 2, pp. 116–127, 1998, doi: 10.1002/(SICI)1097-0134(19980501)31:2<116::AID-PROT2>3.0.CO;2-K.

[445] R. Cossio-Pérez, J. Palma, and G. Pierdominici-Sottile, “Consistent Principal Component Modes from Molecular Dynamics Simulations of Proteins,” *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 826–834, 2017, doi: 10.1021/acs.jcim.6b00646.

[446] J. Palma and G. Pierdominici-Sottile, “On the Uses of PCA to Characterise Molecular Dynamics Simulations of Biological Macromolecules: Basics and Tips for an Effective Use,” *ChemPhysChem*, vol. 24, no. 2, p. e202200491, 2023, doi: 10.1002/cphc.202200491.

[447] H. W. Ng, C. A. Laughton, and S. W. Doughty, “Molecular dynamics simulations of the adenosine A2a receptor: Structural stability, sampling, and convergence,” *J. Chem. Inf. Model.*, vol. 53, no. 5, 2013, doi: 10.1021/ci300610w.

[448] M. K. E. Braza, J. D. N. Gazmen, E. T. Yu, and R. B. Nellas, “Ligand-Induced Conformational Dynamics of A Tyramine Receptor from Sitophilus oryzae,” *Sci. Rep.*, vol. 9, no. 1, p. 16275, 2019, doi: 10.1038/s41598-019-52478-x.

[449] M. Ahmad, V. Helms, O. V. Kalinina, and T. Lengauer, “Relative Principal Components Analysis: Application to Analyzing Biomolecular Conformational Changes,” *J. Chem. Theory Comput.*, vol. 15, no. 4, pp. 2166–2178, 2019, doi: 10.1021/acs.jctc.8b01074.

[450] L. Jordà, “Supplementary Material to ‘Analysis of consensus motions in proteins through molecular dynamics simulations’ [Thesis],” *Zenodo*, doi: 10.5281/zenodo.10017455, 2023.

[451] C. R. Søndergaard, M. H. M. Olsson, M. Rostkowski, and J. H. Jensen, “Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values,” *J. Chem. Theory Comput.*, vol. 7, no. 7, pp. 2284–2295, 2011, doi: 10.1021/ct200133y.

[452] T. J. Dolinsky *et al.*, “PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations,” *Nucleic Acids Res.*, vol. 35, no. suppl_2, pp. W522–W525, 2007, doi: 10.1093/nar/gkm276.

[453] M. J. Abraham *et al.*, “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,” *SoftwareX*, vol. 1–2, pp. 19–25, 2015, doi: 10.1016/j.softx.2015.06.001.

[454] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, “Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation,” *J. Mol. Biol.*, vol. 285, no. 4, pp. 1735–1747, Jan. 1999, doi: 10.1006/jmbi.1998.2401.

[455] D. A. Case *et al.*, “AMBER 2018,” *University of California, San Francisco*. 2018, [Online]. Available: http://ambermd.org/doc12/Amber18.pdf.

[456] K. Lindorff-Larsen *et al.*, “Improved side-chain torsion potentials for the Amber ff99SB protein force field,” *Proteins Struct. Funct. Bioinforma.*, vol. 78, no. 8, pp. 1950–1958, 2010, doi: 10.1002/prot.22711.

[457] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, “Development and testing of a general Amber force field,” *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004, doi: 10.1002/jcc.20035.

[458] A. W. Sousa Da Silva and W. F. Vranken, “ACPYPE - AnteChamber PYthon Parser interfacE,” *BMC Res. Notes*, vol. 5, pp. 1–8, 2012, doi: 10.1186/1756-0500-5-367.

[459] B. A. Leland, D. A. Paul, B. P. Krueger, and R. C. Walker, “AMBER Advanced Tutorials A1: Setting up an advanced system (including charge derivation),” *Dept. of Chemistry, Hope College; San Diego Supercomputer Center, University of California, San Diego*, 2008. http://ambermd.org/tutorials/advanced/tutorial1/index.htm (accessed Feb. 28, 2020).

[460] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman, “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model,” *J. Phys. Chem.*, vol. 97, no. 40, pp. 10269–10280, 1993, doi: 10.1021/j100142a004.

[461] C. Møller and M. S. Plesset, “Note on an approximation treatment for many-electron systems,” *Phys. Rev.*, vol. 46, no. 7, p. 618, 1934, doi: 10.1103/PhysRev.46.618.

[462] R. Ditchfield, W. J. Hehre, and J. A. Pople, “Self-consistent molecular-orbital methods. IX. An extended gaussian-type basis for molecular-orbital studies of organic molecules,” *J. Chem. Phys.*, vol. 54, no. 2, pp. 724–728, 1971, doi: 10.1063/1.1674902.

[463] P. C. Hariharan and J. A. Pople, “The influence of polarization functions on molecular orbital hydrogenation energies,” *Theor. Chim. Acta*, vol. 28, no. 3, pp. 213–222, 1973, doi: 10.1007/BF00533485.

[464] U. C. Singh and P. A. Kollman, “An approach to computing electrostatic charges for molecules,” *J. Comput. Chem.*, vol. 5, no. 2, pp. 129–145, 1984, doi: 10.1002/jcc.540050204.

[465] B. H. Besler, K. M. Merz Jr, and P. A. Kollman, “Atomic charges derived from semiempirical methods,” *J. Comput. Chem.*, vol. 11, no. 4, pp. 431–439, 1990, doi: 10.1002/jcc.540110404.

[466] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollman, “Application of RESP Charges To Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation,” *J. Am. Chem. Soc.*, vol. 115, no. 21, pp. 9620–9631, 1993, doi: 10.1021/ja00074a030.

[467] M. J. Frisch *et al.*, “Gaussian 16, Rev. B.01,” *Gaussian, Inc., Wallingford, CT*. 2016.

[468] K. L. Meagher, L. T. Redman, and H. A. Carlson, “Development of polyphosphate parameters for use with the AMBER force field,” *J. Comput. Chem.*, vol. 24, no. 9, pp. 1016–1025, 2003, doi: 10.1002/jcc.10262.

[469] P. Li, L. F. Song, and K. M. Merz Jr, “Systematic parameterization of monovalent ions employing the nonbonded model,” *J. Chem. Theory Comput.*, vol. 11, no. 4, pp. 1645–1657, 2015, doi: 10.1021/ct500918t.

[470] P. Li, B. P. Roberts, D. K. Chakravorty, and K. M. Merz Jr, “Rational design of particle mesh ewald compatible lennard-jones parameters for +2 metal cations in explicit solvent,” *J. Chem. Theory Comput.*, vol. 9, no. 6, pp. 2733–2748, 2013, doi: 10.1021/ct400146w.

[471] P. Li and K. M. Merz Jr, “MCPB.py: A Python Based Metal Center Parameter Builder,” *J. Chem. Inf. Model.*, vol. 56, no. 4, pp. 599–604, 2016, doi: 10.1021/acs.jcim.5b00674.

[472] P. Li and K. M. Merz Jr, “Developing Nonstandard Parameters: Metal Ion Modeling Tutorial,” *AMBER Tutorials Website (not peer reviewed)*. http://ambermd.org/tutorials/advanced/tutorial20/index.php (accessed Mar. 01, 2023).

[473] A. D. Becke, “Density-functional exchange-energy approximation with correct asymptotic behavior,” *Phys. Rev. A*, vol. 38, no. 6, pp. 3098–3100, 1988, doi: 10.1103/PhysRevA.38.3098.

[474] C. Lee, W. Yang, and R. G. Parr, “Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density,” *Phys. Rev. B*, vol. 37, no. 2, pp. 785–789, 1988, doi: 10.1103/PhysRevB.37.785.

[475] A. D. Becke, “Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing,” *J. Chem. Phys.*, vol. 104, no. 3, pp. 1040–1046, 1996, doi: 10.1063/1.470829.

[476] S. Grimme, S. Ehrlich, and L. Goerigk, “Effect of the damping function in dispersion corrected density functional theory,” *J. Comput. Chem.*, vol. 32, no. 7, pp. 1456–1465, 2011, doi: 10.1002/jcc.21759.

[477] T. Clark, J. Chandrasekhar, G. W. Spitznagel, and P. V. R. Schleyer, “Efficient diffuse function-augmented basis sets for anion calculations. III. The 3-21+G basis set for first-row elements, Li–F,” *J. Comput. Chem.*, vol. 4, no. 3, pp. 294–301, 1983, doi: 10.1002/jcc.540040303.

[478] M. Cossi, N. Rega, G. Scalmani, and V. Barone, “Energies, structures, and electronic properties of molecules in

solution with the C-PCM solvation model," *J. Comput. Chem.*, vol. 24, no. 6, pp. 669–681, 2003, doi: 10.1002/jcc.10189.

[479] E. Rezabal, J. M. Mercero, X. Lopez, and J. M. Ugalde, "A theoretical study of the principles regulating the specificity for Al(III) against Mg(II) in protein cavities," *J. Inorg. Biochem.*, vol. 101, no. 9 SPEC. ISS., pp. 1192–1200, 2007, doi: 10.1016/j.jinorgbio.2007.06.010.

[480] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, "Computational enzymatic catalysis - Clarifying enzymatic mechanisms with the help of computers," *Phys. Chem. Chem. Phys.*, vol. 14, no. 36, pp. 12431–12441, 2012, doi: 10.1039/c2cp41180f.

[481] J. M. Seminario, "Calculation of intramolecular force fields from second-derivative tensors," *Int. J. Quantum Chem.*, vol. 60, no. 7, pp. 1271–1277, 1996, doi: 10.1002/(SICI)1097-461X(1996)60:7<1271::AID-QUA8>3.0.CO;2-W.

[482] D. Sala, F. Musiani, and A. Rosato, "Application of Molecular Dynamics to the Investigation of Metalloproteins Involved in Metal Homeostasis," *Eur. J. Inorg. Chem.*, vol. 2018, no. 43, pp. 4661–4677, 2018, doi: 10.1002/ejic.201800602.

[483] J. Andrys, J. Heider, and T. Borowski, "Comparison of different approaches to derive classical bonded force-field parameters for a transition metal cofactor: a case study for non-heme iron site of ectoine synthase," *Theor. Chem. Acc.*, vol. 140, no. 8, p. 115, 2021, doi: 10.1007/s00214-021-02796-z.

[484] P. Li and K. M. Merz Jr, "Metal Ion Modeling Using Classical Mechanics," *Chem. Rev.*, vol. 117, no. 3, pp. 1564–1686, 2017, doi: 10.1021/acs.chemrev.6b00440.

[485] L. Hu and U. Ryde, "Comparison of methods to obtain force-field parameters for metal sites," *J. Chem. Theory Comput.*, vol. 7, no. 8, pp. 2452–2463, 2011, doi: 10.1021/ct100725a.

[486] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, 1983, doi: 10.1063/1.445869.

[487] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, no. 1, 2007, doi: 10.1063/1.2408420.

[488] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, 1984, doi: 10.1063/1.448118.

[489] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, 1981, doi: 10.1063/1.328693.

[490] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N · log(N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993, doi: 10.1063/1.464397.

[491] W. F. van Gunsteren and H. J. C. Berendsen, "A Leap-Frog Algorithm for Stochastic Dynamics," *Mol. Simul.*, vol. 1, no. 3, pp. 173–185, 1988, doi: 10.1080/08927028808080941.

[492] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A Linear Constraint Solver for molecular simulations," *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997, doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[493] O. Carugo, "Statistical validation of the root-mean-square-distance, a measure of protein structural proximity," *Protein Eng. Des. Sel.*, vol. 20, no. 1, pp. 33–37, 2007, doi: 10.1093/protein/gzl051.

[494] J. W. Pitera, "Expected distributions of root-mean-square positional deviations in proteins," *J. Phys. Chem. B*, vol. 118, no. 24, pp. 6526–6530, 2014, doi: 10.1021/jp412776d.

[495] W. F. Van Gunsteren, X. Daura, and A. E. Mark, "Computation of free energy," *Helv. Chim. Acta*, vol. 85, no. 10, pp. 3113–3129, 2002, doi: 10.1002/1522-2675(200210)85:10<3113::AID-HLCA3113>3.0.CO;2-0.

[496] J. Wereszczynski and J. A. McCammon, "Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition," *Q. Rev. Biophys.*, vol. 45, no. 1, pp. 1–25, 2012, doi: 10.1017/S0033583511000096.

[497] V. Gapsys, S. Michielssens, J. H. Peters, B. L. de Groot, and H. Leonov, "Calculation of Binding Free Energies," in *Molecular Modeling of Proteins. Methods in Molecular Biology*, vol. 1215, 2015, pp. 173–209.

[498] H. J. Berendsen and S. Hayward, "Collective protein dynamics in relation to function," *Curr. Opin. Struct. Biol.*, vol. 10, no. 2, 2000, doi: 10.1016/S0959-440X(00)00061-0.

[499] L. L. Palese, “Random Matrix Theory in molecular dynamics analysis,” *Biophys. Chem.*, vol. 196, pp. 1–9, 2015, doi: 10.1016/j.bpc.2014.08.007.

[500] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, “Principal Component Analysis for Protein Folding Dynamics,” *J. Mol. Biol.*, 2009, doi: 10.1016/j.jmb.2008.10.018.

[501] M. M. Deza and E. Deza, *Encyclopedia of distances*. 2009.

[502] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. 2012.

[503] F. Murtagh and P. Contreras, “Algorithms for hierarchical clustering: An overview,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 86–97, 2012, doi: 10.1002/widm.53.

[504] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. 2008.

[505] Y. Loewenstein, E. Portugaly, M. Fromer, and M. Linial, “Efficient algorithms for accurate hierarchical clustering of huge datasets: Tackling the entire protein space,” *Bioinformatics*, vol. 24, no. 13, pp. i41–i49, 2008, doi: 10.1093/bioinformatics/btn174.

[506] R. R. Sokal and C. D. Michener, “A statistical method for evaluating systematic relationships,” *Univ. Kansas Sci. Bull.*, vol. 38, pp. 1409–1438, 1958.

[507] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[508] M. Rosenblatt, “Remarks on Some Nonparametric Estimates of a Density Function,” *Ann. Math. Stat.*, vol. 27, no. 3, pp. 832–837, 1956, doi: 10.1214/aoms/1177728190.

[509] E. Parzen, “On Estimation of a Probability Density Function and Mode,” *Ann. Math. Stat.*, vol. 33, no. 3, pp. 1065–1076, 1962, doi: 10.1214/aoms/1177704472.

[510] Y. C. Chen, “A tutorial on kernel density estimation and recent advances,” *Biostat. Epidemiol.*, vol. 1, no. 1, pp. 161–187, 2017, doi: 10.1080/24709360.2017.1396742.

[511] A. Łysiak and M. Szmajda, “Empirical Comparison of the Feature Evaluation Methods Based on Statistical Measures,” *IEEE Access*, vol. 9, pp. 27868–27883, 2021, doi: 10.1109/ACCESS.2021.3058428.

[512] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[513] A. K. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bull. Calcuta Math. Soc.*, vol. 35, no. 99–109, 1943.

[514] F. J. Aherne, N. A. Thacker, and P. I. Rockett, “The Bhattacharyya metric as an absolute similarity measure for frequency coded data,” *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998.

[515] K. E. Sundell and J. E. Saylor, “Two-Dimensional Quantitative Comparison of Density Distributions in Detrital Geochronology and Geochemistry,” *Geochemistry, Geophys. Geosystems*, vol. 22, no. 4, p. e2020GC009559, 2021, doi: 10.1029/2020GC009559.

[516] J. G. Tate *et al.*, “COSMIC: The Catalogue Of Somatic Mutations In Cancer,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D941–D947, 2019, doi: 10.1093/nar/gky1015.

[517] H. M. Dingerdissen, J. Torcivia-Rodriguez, Y. Hu, T. C. Chang, R. Mazumder, and R. Kahsay, “BioMuta and BioXpress: Mutation and expression knowledgebases for cancer biomarker discovery,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1128–D1136, 2018, doi: 10.1093/nar/gkx907.

[518] R. A. Hoskins, S. Repo, D. Barsky, G. Andreoletti, J. Moult, and S. E. Brenner, “Reports from CAGI: The Critical Assessment of Genome Interpretation,” *Hum. Mutat.*, vol. 38, no. 9, pp. 1039–1041, 2017, doi: 10.1002/humu.23290.

[519] Q. Xu *et al.*, “Benchmarking predictions of allostery in liver pyruvate kinase in CAGI4,” *Hum. Mutat.*, vol. 38, no. 9, pp. 1123–1131, 2017, doi: 10.1002/humu.23222.

[520] R. A. Laskowski and M. B. Swindells, “LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery,” *J. Chem. Inf. Model.*, vol. 51, no. 10, pp. 2778–2786, 2011, doi: 10.1021/ci200227u.

[521] R. Van Wijk, E. G. Huizinga, A. C. W. Van Wesel, B. A. Van Oirschot, M. A. Hadders, and W. W. Van Solinge, “Fifteen novel mutations in PKLR associated with pyruvate kinase (PK) deficiency: Structural implications of amino acid substitutions in PK,” *Hum. Mutat.*, vol. 30, no. 3, pp. 446–453, 2009, doi: 10.1002/humu.20915.

[522] A. Nain-Perez *et al.*, “Tuning liver pyruvate kinase activity up or down with a new class of allosteric modulators,” *Eur. J. Med. Chem.*, vol. 250, p. 115177, 2023, doi: 10.1016/j.ejmech.2023.115177.

[523] L. Swint-Kruse *et al.*, “PYK-SubstitutionOME: an integrated database containing allosteric coupling, ligand affinity and mutational, structural, pathological, bioinformatic and computational information about pyruvate kinase isozymes,” *Database*, vol. 2023, p. baad030, 2023, doi: 10.1093/database/baad030.

[524] J. W. Ponder and D. A. Case, “Force fields for protein simulations,” *Adv. Protein Chem.*, vol. 66, pp. 27–85, 2003, doi: 10.1016/S0065-3233(03)66002-X.

[525] A. Stein, M. Rueda, A. Panjkovich, M. Orozco, and P. Aloy, “A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks,” *Structure*, vol. 19, no. 6, pp. 881–889, 2011, doi: 10.1016/j.str.2011.03.009.

[526] N. Díaz and D. Suárez, “Molecular Dynamics Studies of Matrix Metalloproteases,” in *Matrix Metalloproteases. Methods in Molecular Biology*, vol. 1579, 2017, pp. 111–134.

[527] J. Torras, M. Maccarrone, and E. Dainese, “Molecular dynamics study on the Apo- and Holo-forms of 5-lipoxygenase,” *Biotechnol. Appl. Biochem.*, vol. 65, no. 1, pp. 54–61, 2018, doi: 10.1002/bab.1583.

[528] L. Banci, “Molecular dynamics simulations of metalloproteins,” *Curr. Opin. Chem. Biol.*, vol. 7, no. 1, pp. 143–149, 2003, doi: https://dx.doi.org/10.1016/S1367-5931(02)00014-5.

[529] O. Amata, T. Marino, N. Russo, and M. Toscano, “A proposal for mitochondrial processing peptidase catalytic mechanism,” *J. Am. Chem. Soc.*, vol. 133, no. 44, pp. 17824–17831, 2011, doi: 10.1021/ja207065v.

[530] T. Dudev, L. Y. Chang, and C. Lim, “Factors governing the substitution of La3+ for Ca2+ and Mg2+ in metalloproteins: A DFT/CDM study,” *J. Am. Chem. Soc.*, vol. 127, no. 11, pp. 4091–4103, 2005, doi: 10.1021/ja044404t.

[531] R. P. P. Neves, S. F. Sousa, P. A. Fernandes, and M. J. Ramos, “Parameters for molecular dynamics simulations of manganese-containing metalloproteins,” *J. Chem. Theory Comput.*, vol. 9, no. 6, pp. 2718–2732, 2013, doi: 10.1021/ct400055v.

[532] P. Ferreira, N. M. F. S. A. Cerqueira, N. F. Brás, P. A. Fernandes, and M. J. Ramos, “Parametrization of Molybdenum Cofactors for the AMBER Force Field,” *J. Chem. Theory Comput.*, vol. 14, no. 5, pp. 2538–2548, 2018, doi: 10.1021/acs.jctc.8b00137.

[533] H. B. Schlegel, “Some Practical Suggestions for Optimizing Geometries and Locating Transition States,” in *New Theoretical Concepts for Understanding Organic Reactions. NATO ASI Series*, vol. 267, 1989, pp. 33–53.

[534] B. K. Mai, N. M. Neris, Y. Yang, and P. Liu, “C-N Bond Forming Radical Rebound Is the Enantioselectivity-Determining Step in P411-Catalyzed Enantioselective C(sp3)-H Amination: A Combined Computational and Experimental Investigation,” *J. Am. Chem. Soc.*, vol. 144, no. 25, pp. 11215–11225, 2022, doi: 10.1021/jacs.2c02283.

[535] P. Li and K. M. Merz, “Parameterization of a Dioxygen Binding Metal Site Using the MCPB.py Program,” in *Structural Genomics. Methods in Molecular Biology*, vol. 2199, 2021, pp. 257–275.

[536] T. Y. Yang, T. Dudev, and C. Lim, “Mononuclear versus binuclear metal-binding sites: Metal-binding affinity and selectivity from PDB survey and DFT/CDM calculations,” *J. Am. Chem. Soc.*, vol. 130, no. 12, pp. 3844–3852, 2008, doi: 10.1021/ja076277h.

[537] J. Mähler and I. Persson, “A study of the hydration of the alkali metal ions in aqueous solution,” *Inorg. Chem.*, vol. 51, no. 1, pp. 425–438, 2012, doi: 10.1021/ic2018693.

[538] J. J. Perez, M. S. Tomas, and J. Rubio-Martinez, “Assessment of the Sampling Performance of Multiple-Copy Dynamics versus a Unique Trajectory,” *J. Chem. Inf. Model.*, vol. 56, no. 10, 2016, doi: 10.1021/acs.jcim.6b00347.

[539] M. Nemec and D. Hoffmann, “Quantitative Assessment of Molecular Dynamics Sampling for Flexible Systems,” *J. Chem. Theory Comput.*, vol. 13, no. 2, 2017, doi: 10.1021/acs.jctc.6b00823.

[540] I. Assent, “Clustering high dimensional data,” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 4, pp. 340–350, 2012, doi: 10.1002/widm.1062.

[541] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[542] R. J. G. B. Campello, D. Moulavi, and J. Sander, “Density-based clustering based on hierarchical density estimates,”

in *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*, 2013, vol. 7819, doi: 10.1007/978-3-642-37456-2_14.

[543] S. Neumayer, M. Nimmer, S. Setzer, and G. Steidl, "On the Robust PCA and Weiszfeld's Algorithm," *Appl. Math. Optim.*, vol. 82, no. 3, pp. 1017–1048, 2020, doi: 10.1007/s00245-019-09566-1.

[544] E. Sugrue, D. Coombes, D. Wood, T. Zhu, K. A. Donovan, and R. C. J. Dobson, "The lid domain is important, but not essential, for catalysis of Escherichia coli pyruvate kinase," *Eur. Biophys. J.*, vol. 49, no. 8, pp. 761–772, 2020, doi: 10.1007/s00249-020-01466-5.

[545] R. Van Wijk, A. C. W. Van Wesel, A. A. M. Thomas, G. Rijksen, and W. W. Van Solinge, "Ex vivo analysis of aberrant splicing induced by two donor site mutations in PKLR of a patient with severe pyruvate kinase deficiency," *Br. J. Haematol.*, vol. 125, no. 2, pp. 253–263, 2004, doi: 10.1111/j.1365-2141.2004.04895.x.

[546] A. Zanella, E. Fermo, P. Bianchi, L. R. Chiarelli, and G. Valentini, "Pyruvate kinase deficiency: The genotype-phenotype association," *Blood Rev.*, vol. 21, no. 4, pp. 217–231, 2007, doi: 10.1016/j.blre.2007.01.001.

[547] H. Rouger, C. Valentin, C. T. Craescu, F. Galactéros, and M. Cohen-Solal, "Five unknown mutations in the LR pyruvate kinase gene associated with severe hereditary nonspherocytic haemolytic anaemia in France," *Br. J. Haematol.*, vol. 92, no. 4, pp. 825–830, 1996, doi: 10.1046/j.1365-2141.1996.405941.x.

[548] W. Kugler *et al.*, "Eight novel mutations and consequences on mRNA and protein level in pyruvate kinase-deficient patients with nonspherocytic hemolytic anemia," *Hum. Mutat.*, vol. 15, no. 3, pp. 261–272, 2000, doi: 10.1002/(SICI)1098-1004(200003)15:3<261::AID-HUMU7>3.0.CO;2-T.

[549] A. Zanella *et al.*, "Molecular Characterization of PK-LR Gene in Pyruvate Kinase–Deficient Italian Patients," *Blood*, vol. 89, no. 10, pp. 3847–3852, 1997, doi: 10.1046/j.1365-2141.1998.01013.x.

[550] M. C. C. M. Svidnicki *et al.*, "Novel mutations associated with pyruvate kinase deficiency in Brazil," *Hematol. Transfus. Cell Ther.*, vol. 40, no. 1, pp. 5–11, 2018, doi: 10.1016/j.bjhh.2017.08.007.

[551] C. Dong *et al.*, "Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies," *Hum. Mol. Genet.*, vol. 24, no. 8, pp. 2125–2137, 2015, doi: 10.1093/hmg/ddu733.

[552] R. Zarza *et al.*, "Molecular characterization of the PK-LR gene in pyruvate kinase deficient spanish patients," *Br. J. Haematol.*, vol. 103, no. 2, pp. 377–382, 1998, doi: 10.1046/j.1365-2141.1998.01013.x.

Appendix A

Contents and access to the Supplementary Material

This section provides a detail of the contents uploaded to the online repository Zenodo that serve as the supplementary material of this thesis.

- Jordà, L. (2023). Supplementary Material to "Analysis of consensus motions in proteins through molecular dynamics simulations" [Thesis]. Zenodo. https://doi.org/10.5281/zenodo.10017455

The purposes of this material are: i) to facilitate reproducibility and further exploration of technical aspects of the methodology, and ii) to enable visual inspection of the collective motions from the study of human erythrocyte pyruvate kinase (PKR).

The contents are structured in three main directories (compressed as *ZIP* folders):

1. Directory *QM_related_files*
    - Contains the relevant files to trace and reproduce the quantum mechanics (QM) calculations of the project. These comprise geometry optimizations, force-constant calculations, and single-point energy calculations with the Gaussian16 software.
        - Input files (ASCII files with the *.com* extension). They contain the input structures and the set of parameters and decisions submitted in every type of calculation, including the frozen-atom schemes of the cluster models for the parameterization of metal centers.
        - Log files (ASCII files with the *.log* extension). They contain detailed information of every step and stage in the calculations (parameters, processes, decisions, values…).
        - Printed information from the checkpoint files (ASCII files with the *.gjf* extension), namely, the molecular structure (charge, multiplicity, coordinates), the basis set, the Molecular Orbitals, and the forces (if force constants were requested to be calculated).

2. Directory *MD_related_files*
    - Subdirectory *force_field_custom_entries*
        - Contains text files (ASCII files with the *.txt* extension) with descriptions of the steps followed to include new custom residue types, atom types and their respective non-bonded terms in the files of the AMBER99SB-ILDN force field included in the native installation of the GROMACS software. The procedure was adapted from the GROMACS online "Short How-To guides", and involves modifying the *aminoacids.rtp*, *atomtypes.atp*, and *ffnonbonded.itp* files of the force field, and the *residuetypes.dat* file.
    - Subdirectory *MDP_files*
        - Contains the GROMACS MDP files Molecular Dynamics Parameters; ASCII files with the *.mdp* extension) that correspond to several of the stages of the MD workflow of this project (Figure 3.2 from the Methods chapter), namely,

energy minimizations, incorporation of dissolved ions, and NVT/NPT simulations. The MDP format contains a list of keywords that are used in the framework of GROMACS to set up the different types of simulation that require use of the *grompp* tool.

  - Subdirectory *simulation_systems*
    - Contains the structure files (ASCII files in PDB format with the *.pdb* extension) and the GROMACS topology files (ASCII files with the *.top* and *.itp* extensions) of all the systems subjected to MD in this project.
    - In the case of the systems in the apo condition, the files correspond to the configuration after the placement of hydrogen atoms, right before the stage of energy minimization in vacuum.
    - In the case of the systems in holo conditions, the files correspond to the configuration after the incorporation of the QM-derived metal-center parameters, right before the repetition of the energy minimization in vacuum.
    - Importantly, this is the location where all the QM-derived atomic charges and bonded parameters of each type of system can be found.

3. Directory *supplementary_videos*
   - Contains the 30 Supplementary Videos (S4.1 to S4.30) in MP4 and GIF formats (*.mp4* and *.gif* extensions) that enable visualization of the collective motions reported in the Results chapter. Each video is paired with a text file (ASCII file with the *.txt* extension) that includes a brief description of the type of visualization and references to figures of the Results chapter.

On the other hand, the complete set of MD simulations, as well as standard trajectory analyses can be found at the online database PKLR from the Molecular Dynamics Data Bank (https://pklr.mddbr.eu), a project funded by the European Union's Horizon Europe programme under grant agreement 101094651.

Appendix B

Bibliographic references for the selected missense variants of PKR

Table 1

Bibliographic references for the selected missense variants of PKR

Variant	Subset [a]	Relevant references [b]
Leu73Pro	P	[1]
Ser80Pro	P	[2], [3]
Glu81Lys	N	-
Ala115Pro	P	[4]
Ser120Phe	P	[2], [4]
His124Gln	P	[5], [6]
Glu125Ala	N	-
Glu129Lys	N	-
Ser130Tyr	P	[7]
Gly143Ser	P	[8]
Leu155Pro	P	[9]
Thr157Pro	N	-
Arg163Cys	P	[1]
Glu172Gln	P	[10]
Glu172Gly	N	-
Ala257Thr	N	-
Gly263Ala	N	-
Gly263Trp	P	[11]
Ala295Thr	N	-
Ala295Val	P	[12]
Pro303Leu	N	-
Gly307Ser	N	-
Ile310Asn	P	[1]
Glu315Lys	P	[12]
Leu327Val	N	-
Gly332Ser	P	[2], [13], [14]
Arg337Gln	P	[6], [10]
Asp339His	P	[10]
Arg359Cys	P	[15]
Thr371Ile	P	[16]
Thr384Met	P	[17]
Arg385Lys	P	[1], [5], [16]
Asp390Asn	P	[17], [18]
Ala394Asp	P	[17]
Ala394Val	P	[17]
Ile402Val	N	-
Met403Ile	P	[19], [20]
Met403Thr	N	-
Thr408Ile	P	[10]
Gly411Ser	P	[21]
Ala430Thr	P	[5]

Table 1 (Continued)

Gly458Ala	N	-
Gly458Asp	P	[1], [22]
Arg486Gln	N	-
Arg486Trp	P	[13], [16], [17], [23], [24]
Ile494Thr	P	[25]
Arg504Leu	P	[12], [13], [17]
Gln505Arg	N	-
Gln505Glu	P	[17]
Val506Ile	P	[10]
Arg510Gln	P	[2], [4], [6], [13], [17], [23], [26]–[28]
Arg518His	N	-
Pro521Ser	N	-
Arg531Cys	P	[29], [30]
Arg531His	N	-
Arg532Trp	P	[17], [27]
Val552Ala	N	-
Val552Met	P	[31]
Gly557Ala	P	[5], [24]
Arg559Gln	N	-
Arg559Gly	P	[31], [32]

[a] The variants from the pathogenic subset (labeled as “P”) have been suggested to be implicated in pyruvate kinase deficiency, as reported in clinical studies or experimental assays (either from public repositories or the literature). The variants from the potentially neutral subset (labeled as “N”) were retrieved from the data portals of large-scale genomic sequencing projects and have not been associated with pyruvate kinase deficiency.

[b] This table only includes some representative bibliographic references of each variant. Only variants from the pathogenic subset are reported in the literature.

[1] R. Van Wijk, E. G. Huizinga, A. C. W. Van Wesel, B. A. Van Oirschot, M. A. Hadders, and W. W. Van Solinge, “Fifteen novel mutations in PKLR associated with pyruvate kinase (PK) deficiency: Structural implications of amino acid substitutions in PK,” Hum. Mutat., vol. 30, no. 3, pp. 446–453, 2009, doi: 10.1002/humu.20915.

[2] W. Kugler et al., “Eight novel mutations and consequences on mRNA and protein level in pyruvate kinase-deficient patients with nonspherocytic hemolytic anemia,” Hum. Mutat., vol. 15, no. 3, pp. 261–272, 2000, doi: 10.1002/(SICI)1098-1004(200003)15:3<261::AID-HUMU7>3.0.CO;2-T.

[3] R. Uenaka et al., “Compound heterozygosis mutations affecting both hepatic and erythrocyte isozymes of pyruvate kinase,” Biochem. Biophys. Res. Commun., vol. 208, no. 3, pp. 991–998, 1995.

[4] H. Rouger, C. Valentin, C. T. Craescu, F. Galactéros, and M. Cohen-Solal, “Five unknown mutations in the LR pyruvate kinase gene associated with severe hereditary nonspherocytic haemolytic anaemia in France,” Br. J. Haematol., vol. 92, no. 4, pp. 825–830, 1996, doi: 10.1046/j.1365-2141.1996.405941.x.

[5] L. Montllor, M. del M. Mañú-Pereira, E. Llaudet-Planas, P. Gómez Ramírez, J. Sevilla Navarro, and J. L. Vives-Corrons, “Red cell pyruvate kinase deficiency in Spain: A study of 15 cases,” Med. Clínica (English Ed., vol. 148, no. 1, pp. 23–27, 2017, doi: 10.1016/j.medcle.2016.10.037.

[6] J. M. Nieto et al., “Next Generation Sequencing for diagnosis of inherited hemolytic anemias,” European Hematology Association. 2018.

[7] M. Cohen-Solal et al., “A new sickle cell disease phenotype associating Hb S trait, severe pyruvate kinase deficiency (PK Conakry), and an α2 globin gene variant (Hb Conakry),” Br. J. Haematol., vol. 103, no. 4, pp. 950–956, 1998, doi: 10.1046/j.1365-2141.1998.01094.x.

[8] P. Kedar et al., “Spectrum of novel mutations in the human PKLR gene in pyruvate kinase-deficient Indian patients with heterogeneous clinical phenotypes,” Clin. Genet., vol. 75, no. 2, pp. 157–162, 2009, doi: 10.1111/j.1399-0004.2008.01079.x.

[9] L. Baronciani and E. Beutler, “Analysis of pyruvate kinase-deficiency mutations that produce nonspherocytic hemolytic anemia.,” Proc. Natl. Acad. Sci. U. S. A., vol. 90, no. 9, pp. 4324–7, 1993, doi: 10.1073/pnas.90.9.4324.

[10] R. Zarza et al., “Molecular characterization of the PK-LR gene in pyruvate kinase deficient spanish patients,” Br. J. Haematol., vol. 103, no. 2, pp. 377–382, 1998, doi: 10.1046/j.1365-2141.1998.01013.x.

[11] A. Zanella et al., “Molecular Characterization of PK-LR Gene in Pyruvate Kinase–Deficient Italian Patients,” Blood, vol. 89, no. 10, pp. 3847–3852, 1997, doi: 10.1046/j.1365-2141.1998.01013.x.

[12] A. Demina, K. I. Varughese, J. Barbot, L. Forman, and E. Beutler, “Six previously undescribed pyruvate kinase mutations causing enzyme deficiency,” Blood, vol. 92, no. 2, pp. 647–652, 1998, [Online]. Available: http://research.bmn.com/medline/search/record?uid=98322173.

[13] M. C. C. M. Svidnicki et al., “Novel mutations associated with pyruvate kinase deficiency in Brazil,” Hematol. Transfus. Cell Ther., vol. 40, no. 1, pp. 5–11, 2018, doi: 10.1016/j.bjhh.2017.08.007.

[14] W. Kugler, P. Laspe, M. Stahl, W. Schröter, and M. Lakomek, “Identification of a novel promoter mutation in the human pyruvate kinase (PK) LR gene of a patient with severe haemolytic anaemia,” Br. J. Haematol., vol. 105, no. 3, pp. 596–598, 1999, doi: 10.1046/j.1365-2141.1999.01386.x.

[15] H. Kanno et al., “Frame shift mutation, exon skipping, and a two-codon deletion caused by splice site mutations account for pyruvate kinase deficiency,” Blood, vol. 89, no. 11, pp. 4213–4218, 1997, doi: 10.1182/blood.v89.11.4213.

[16] S. Pissard et al., “Pyruvate kinase deficiency in France: A 3-year study reveals 27 new mutations,” Br. J. Haematol., vol. 133, no. 6, pp. 683–689, 2006, doi: 10.1111/j.1365-2141.2006.06076.x.

[17] A. Zanella, E. Fermo, P. Bianchi, L. R. Chiarelli, and G. Valentini, “Pyruvate kinase deficiency: The genotype-phenotype association,” Blood Rev., vol. 21, no. 4, pp. 217–231, 2007, doi: 10.1016/j.blre.2007.01.001.

[18] N. Karadsheh, T. Gelbart, and R. Naffa, “Hemolytic anemia associated with a novel heterozygote mutation 1183A in the PK-LR gene (PK- Jordan),” Int. J. Lab. Hematol., vol. 36, pp. e66–e68, 2014, doi: 10.1080/13518040701205365.

[19] E. Fermo et al., “Red cell pyruvate kinase deficiency: 17 New mutations of the PK-LR gene,” Br. J. Haematol., vol. 129, no. 6, pp. 839–846, 2005, doi: 10.1111/j.1365-2141.2005.05520.x.

[20] A. Minucci et al., “Worsening of the clinical-hematological picture in a patient with a rare PK-LR compound heterozygosis after mitral replacement,” Clin. Biochem., vol. 44, no. 14–15, pp. 1261–1263, 2011, doi: 10.1016/j.clinbiochem.2011.07.007.

[21] J. O. Park-Hah, H. Kanno, D. K. Won, and H. Fujii, “A novel homozygous mutation of PKLR gene in a pyruvate-kinase-deficient Korean family,” Acta Haematol., vol. 113, no. 3, pp. 208–211, 2005, doi: 10.1159/000084453.

[22] P. Bianchi et al., “Molecular Characterization of 140 Patients in the Pyruvate Kinase Deficiency (PKD) Natural History Study (NHS): Report of 20 New Variants,” Blood, vol. 126, no. 23, p. 3337, 2015, doi: 10.1182/blood.v126.23.3337.3337.

[23] L. Pastore et al., “Novel mutations and structural implications in R-type pyruvate kinase- deficient patients from southern Italy,” Hum. Mutat., vol. 11, no. 2, 1998, doi: 10.1002/(SICI)1098-1004(1998)11:2<127::AID-HUMU5>3.0.CO;2-G.

[24] L. Manco, M. L. Ribeiro, H. Almeida, O. Freitas, A. Abade, and G. Tamagnini, “PK-LR gene mutations in pyruvate kinase deficient Portuguese patients,” Br. J. Haematol., vol. 105, no. 3, pp. 591–595, 1999, doi: 10.1046/j.1365-2141.1999.01387.x.

[25] S. van Straaten et al., “Worldwide study of hematopoietic allogeneic stem cell transplantation in pyruvate kinase,” Haematologica, vol. 103, no. 2, pp. e82–e86, 2018.

[26] W. W. Van Solinge et al., “Molecular modelling of human red blood cell pyruvate kinase: Structural implications of a novel G1091 to A mutation causing severe nonspherocytic hemolytic anemia,” Blood, vol. 90, no. 12, pp. 4987–4995, 1997, doi: 10.1182/blood.v90.12.4987.4987_4987_4995.

[27] M. Lakomek, P. Huppke, B. Neubauer, A. Pekrun, H. Winkler, and W. Schröter, “Mutations in the R-type pyruvate kinase gene and altered enzyme kinetic properties in patients with hemolytic anemia due to pyruvate kinase deficiency,” Ann. Hematol., vol. 69, pp. 253–260, 1994, doi: 10.1007/BF01700280.

[28] H. Wang, W. Chu, S. K. Das, Q. Ren, S. J. Hasstedt, and S. C. Elbein, “Liver pyruvate kinase polymorphisms are associated with type 2 diabetes in Northern European Caucasians,” Diabetes, vol. 51, no. 9, pp. 2861–2865, 2002, doi: 10.2337/diabetes.51.9.2861.

[29] L. Baronciani, P. Bianchi, and A. Zanella, “Hematologically important mutations: Red cell pyruvate kinase (2nd update),” Blood Cells, Mol. Dis., vol. 24, no. 3, pp. 273–279, 1998, doi: 10.1006/bcmd.1998.0193.

[30] T. Utsugisawa et al., “Pyruvate kinase deficiency in Japan: a summary of clinical feature, laboratory data and enzymatic diagnosis,” European Hematology Association. 2018.

[31] L. Baronciani et al., “Study of the molecular defects in pyruvate kinase deficient patients affected by nonspherocytic hemolytic anemia,” Blood Cells, Mol. Dis., vol. 21, no. 1, pp. 49–55, 1995, doi: 10.1006/bcmd.1995.0008.

[32] S. Unal and F. Gumruk, “Molecular Analyses of Pyruvate Kinase Deficient Turkish Patients from a Single Center,” Pediatr. Hematol. Oncol., vol. 32, no. 5, pp. 354–361, 2015, doi: 10.3109/08880018.2015.1010671.