Universitat Autònoma de Barcelona

# Development and application of computational tools for the coupled exploration of chemical and biological spaces

**José Emilio Sánchez Aparicio**

Universitat Autònoma de Barcelona
Facultat de Ciències
Departament de Química

# Development and application of computational tools for the coupled exploration of chemical and biological spaces

José Emilio Sánchez Aparicio

Tesi Doctoral

Programa de Doctorat en Bioinformàtica

2022

**Director i tutor acadèmic**

Jean-Didier Maréchal

*Memòria presentada per aspirar al Grau de Doctor per José Emilio Sánchez Aparicio*

Vist i plau,

Prof. Jean-Didier Maréchal

**Director**

Bellaterra, 20 de setembre de 2022

# Abstract

In most fields at the interface between chemistry and biology, a three-dimensional vision of the molecular systems is crucial, because it provides essential information about the structures and mechanisms of molecules. In the last decades, the rise of computer power, the improvement of the stability of the codes as well as the increase of savoir-faire of the modelers, have led to a massive expansion of the application of molecular modeling in chemistry and its interfaces. Despite this advance, a series of challenges have still to be solved that include how to deal with the prediction of substrate or cofactor binding routes, the identification of metal binding sites, and the search for chemical optimizations for a given receptor structure.

In this thesis, we aim at addressing these challenges by developing new tools to expand the limits of current software, and optimizing the application of pre-existing methods to biochemical systems that are difficult to handle nowadays.

The first part of the work reports the implementation, benchmark, and application of three novel pieces of software. Two of them are based on a multi-objective evolutionary algorithm: GPathFinder for the identification of ligand binding pathways in proteins, and GAlkemist for a hypothesis driven exploration of the chemical space. The third tool, called BioMetAll, aims at the prediction of metal-binding sites in proteins based on simple geometric descriptors of the protein backbone.

The second part of the work reports the application of GaudiMM, a multi-objective genetic algorithm optimized for molecular modeling tasks, to three real research cases on non-standard dockings. Also, the application of optimized computational workflows for the study of two bioinorganic systems (interactions between oxaliplatin and insulin, and mechanism of an artificial dirhodium cyclopropanase, is reported.

# Contents

# 1

# Biochemical structural models and computers

There are models everywhere. The model of the brand new stadium of your favorite football team, the LEGO set of the Star Wars Millennium Falcon, and the recipe for cooking a cake are examples of our daily life. Of course, the usefulness of models goes far beyond mere artistic or recreational use. In science, models are of central importance in many fields, such as meteorological forecast and climate change predictions, evolutionary models in biology, and the ideal gas model in thermodynamics.

In the area that refers to this thesis, models relate to the structural nature of biochemical systems. In the last decades, the rise of computer power, the im-

provement of the stability of the codes as well as the increase of savoir-faire of the modelers, have led to a massive expansion of the application of molecular modeling in chemistry and its interfaces. Despite this advance, a series of challenges have still to be solved that include how to deal with the prediction of substrate or cofactor binding routes, the identification of metal-binding sites, and the search for chemical optimizations for a given receptor structure. In this thesis, we aim at addressing these challenges by optimizing some pre-existing methods widely applied in molecular modeling but, more than anything else, developing new tools based on evolutionary algorithms.

The manuscript encopasses an introductive description of nowadays knowledge on biochemical models, the issues that arise in their computational treatment, and state-of-the-art application of computational tools in molecular modeling. Then, methodological grounds will be presented in more detail in chapter 2. The final part of the manuscript focuses on the developments and applications that have been performed in this doctorate.

## 1.1 Structural models of biochemical systems

In most fields at the interface between chemistry and biology, a three-dimensional vision of the molecular system is crucial. In this sense, molecular models provide essential information about the structures and mechanisms of molecules. A prototypical example is the three-dimensional double-helix atomic structure of the DNA molecule (Figure 1.1). The model was proposed in 1953 by Watson and Crick with the support of previous crucial X-ray crystallography work by Franklin and Wilkins.[1] This groundbreaking discovery has been fundamental in our understanding of evolution, has helped in the investigation and diagnosis of genetic diseases, and has opened up many areas of research, such as genome editing.

Because of the importance for the scientific community of such atomic molecular structures, the Brookhaven National Laboratory founded the Protein Data Bank (PDB)[4] in 1971. The PDB is a public database containing three-dimensional data of large biological molecules, which is freely accessible on the Internet and has

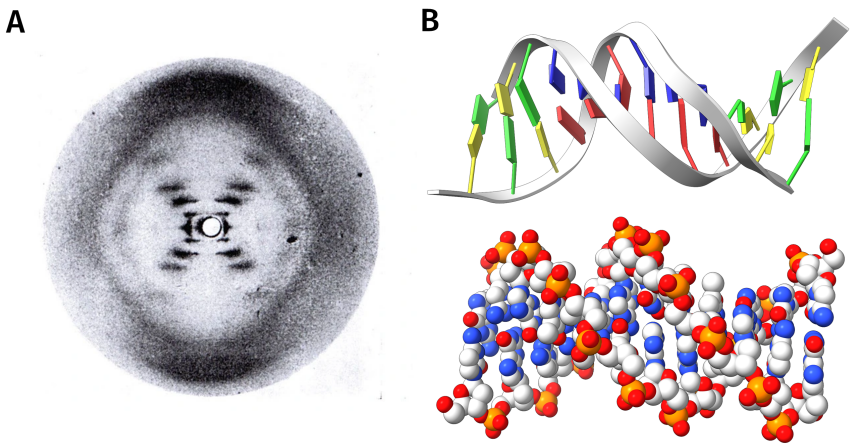



**Figure 1.1:** A prototypical example of a molecular model: the DNA molecule. **(A)** X-ray diffraction pattern of DNA published by R. Franklin in 1953,[2] key in deciphering the helical structure of the DNA molecule. **(B)** Example of atomic structure of a B-DNA dodecamer (PDB entry 1d65).[3] Nucleotide sequence is: CGCAAATTTGCG. Nucleotide and full atomic views are shown in the top and bottom positions, respectively.

become the standard repository of biomolecular structures. Atomic coordinates are obtained using different experimental methods, such as X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy; each entry of the database provides a digital file with the spatial arrangement of all the atoms in the system.

Public availability of atomic coordinates in digital format fostered the use of computers to study biomolecular structural models, but also limitations appeared very early. Living in the age of supercomputers and artificial intelligence, one might imagine that the computational capabilities would easily outstrip the simulation of a single molecule; that running calculations, for example, on an atomic model of the human body would be feasible. Reality hits hard when we compare the number of atoms of an adult human body[5] (~7 x $10^{27}$ atoms) with the capacity of today's most powerful computer, Fugaku, which executes 4.42 x $10^{17}$ floating point operations per second.[6] That is, performing a single mathematical operation for each atom of the human body would take 502 years using all of Fugaku's power. Perhaps the quantum computer revolution will bring us a paradigm shift in the long term, but in the meantime, we need to deal with computational limitations.

As a consequence of these limitations, the first thing we must take into account to deal computationally with any problem is an estimate of the number of mathematical operations necessary to solve it. For example, if we want to sample all the possible values that the polynomial function $f(x) = x^3 + 3x^2 + 4$ can take, the answer at first glance will be infinite operations, because $f$ is a continuous function with domain in all real numbers. We will only be able to deal with this problem computationally by limiting the precision (e.g. taking samples of $x$ with a step of 0.001) and the domain (e.g. only allowing values of $x$ between -5 and 5). With the constraints of the example, we will have a total of 10,000 values of $x$, which constitutes the space that we can now easily evaluate or explore with the computer. The study of biochemical structures also involves the intelligent definition of search spaces to allow their computational treatment, and the following paragraphs will be devoted to presenting them.

## 1.2 Search spaces needed for the study of biochemical systems

In this thesis, we are defining a *biochemical system* as a complex formed by one or more proteins and one or more (other) chemical entities. Other biological systems such as nucleic acids are not considered, although some of the approaches presented here could conveniently be adapted to them. Among the wide panorama of functions carried out by proteins, one of them will be of especial relevance in this doctorate: proteins acting as enzymes.

*Enzymes* are biological systems that act as catalysts by speeding up a specific chemical reaction. Sometimes, enzymes need the addition of a cofactor to become active, which can be another organic or an inorganic molecule. The mechanisms of natural enzymes have been studied for many decades[7] due to their undeniable importance in many metabolic pathways and biological processes. In fact, dysregulated enzyme activity can lead to disease states, making their study of crucial importance for the design of new drugs. Also, the understanding of enzymes' mechanisms of action has opened the avenue to the design of non-natural

or artificial enzymes, which are valuable for medicine, industrial chemistry, and energy production.[8] Several strategies have been successfully applied in the design of artificial enzymes, from pure chemical intuition to rational modification of the cofactor and/or the amino acids involved in catalysis, passing through the recent directed evolution methodology that awarded the Nobel Prize to F. Arnold in 2018.[9] In this regard, the use of computers to assist in the design and understanding of artificial enzymes has gained increasing attention.[10–12]

To study enzymes and the rest of biochemical systems, especially when done computationally, it is convenient to rely on some hierarchical classification of the atoms that compose the system. Sometimes we will also need additional descriptors besides atom coordinates, such as what are the covalent bonds or the atom types. This will allow a more complete description of the system and will facilitate the computational treatment of the information in a specific problem. The first classification we will state is the division between those atoms belonging to the *biological* part of the system –the protein(s)– and those atoms belonging to the *chemical* part –the rest of the atoms–. Along this chapter, we will consider as

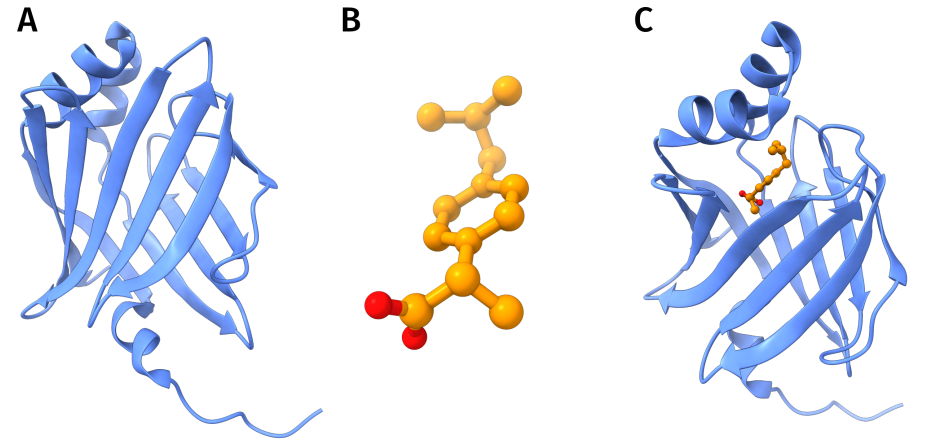
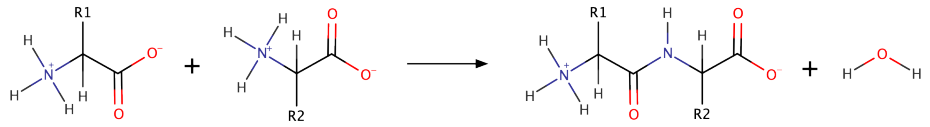



**Figure 1.2:** Example of a biochemical system (PDB entry 3p6h). **(A)** Biological part of the system: human adipocyte lipid-binding protein FABP4. **(B)** Chemical part of the system: (S)-ibuprofen. **(C)** Complete biochemical system: lipid-binding protein FABP4 (in blue cartoon representation) in complex with (S)-ibuprofen (in orange ball and sticks representation).

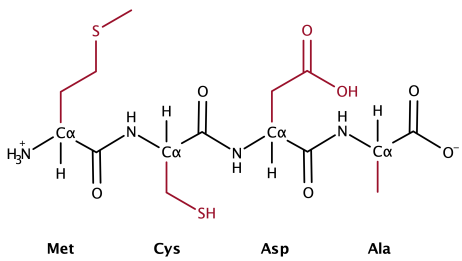
an example of a biochemical system the human adipocyte lipid-binding protein FABP4 in complex with (S)-ibuprofen (PDB entry 3p6h,[13] Figure 1.2). FABP4 belongs to the family of cytosolic fatty-acid binding proteins (FABPs), which are lipid chaperones that reversibly bind a wide variety of hydrophobic ligands.[14] As FABP4 has been associated to insulin resistance,[15] its selective inhibition is a promising research line for the treatment of type 2 diabetes. Also, FABPs have been used as a base to design artificial enzymes by chemical or genetic modification of their scaffolds.[16–19]

## The biological space

Proteins are large macromolecules formed by one or more chains of amino acid residues. There are 20 standard amino acids (Figure 1.3) encoded in the genetic code of an organism, each one with different chemical properties. In addition, selenocysteine and pyrrolysine may exist in some organisms. Two amino acids are linked together by a peptide bond, formed by the dehydration synthesis reaction between the carboxyl group of one amino acid and the amino group of the




**Scheme 1.1:** Dehydration condensation of two amino acids to form a peptide bond with expulsion of water.




**Scheme 1.2:** Example of polypeptide chain with sequence Met-Cys-Asp-Ala. Backbone atoms are shown in black, with alpha carbons labeled as $C_\alpha$. Side chain atoms are shown in dark red.

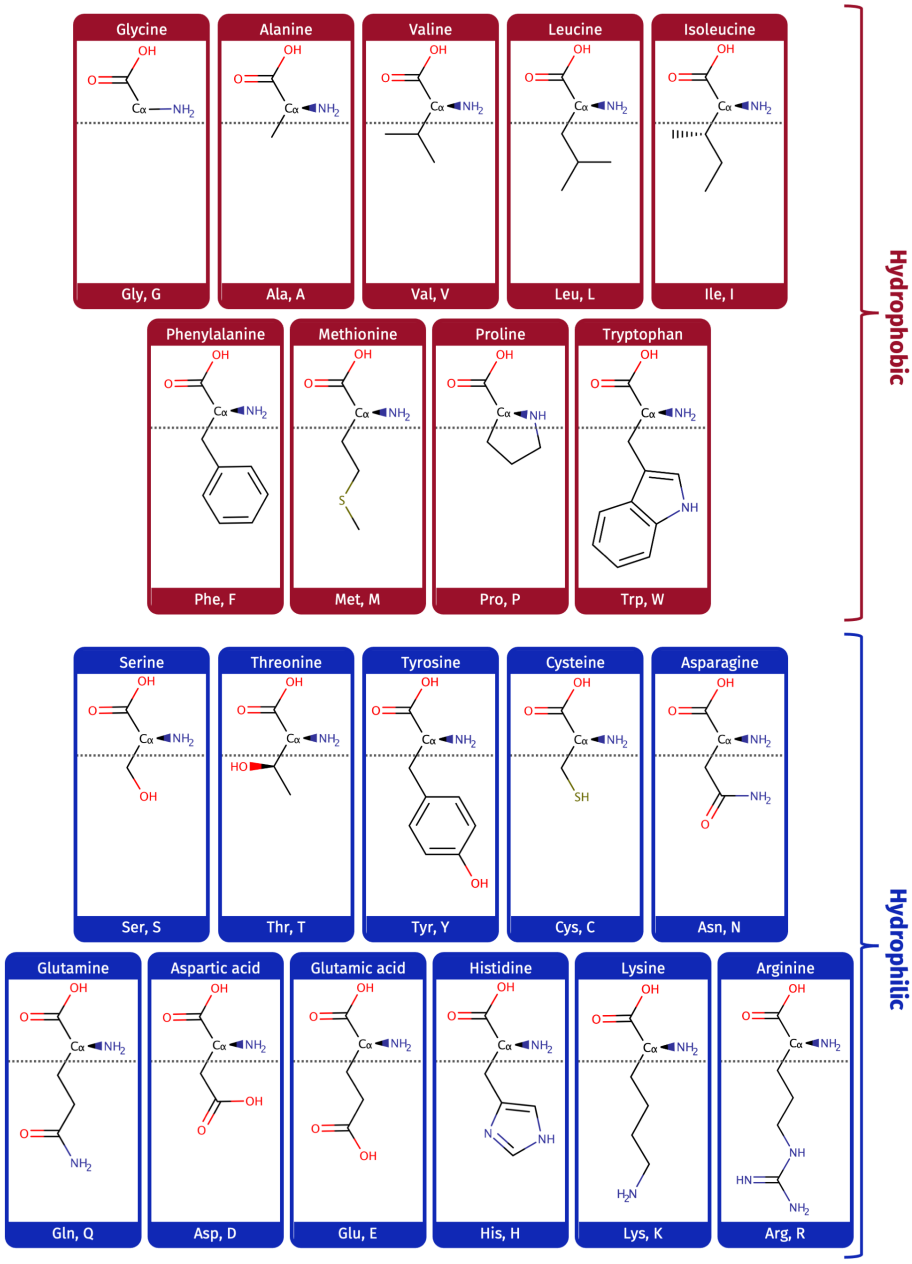



**Figure 1.3:** Structure of the 20 standard amino acids. Backbone and side chain atoms are respectively at the top and bottom positions of each structure, separated by a dashed line. The alpha carbon of each amino acid is labeled as $C_\alpha$. The residues are classified according to their hydrophobic or hydrophilic nature, following Crowe and Bradshaw criterium.[20] Other classifications, such as positively/negatively charged amino acids, are not considered in this scheme.

other amino acid (Scheme 1.1). The successive linking of all the amino acids is what ends building the polypeptide chain (Scheme 1.2), which starts with an N-terminus (containing an amino group) and ends with a C-terminus (containing a carboxylic acid group). The protein structure is usually divided into backbone and side chain atoms. The backbone atoms are common in all amino acids: a central carbon atom (the $\alpha$ carbon), bonded to an amino group ($NH_2$), a carboxyl group (COOH), and a hydrogen atom. The side chain atoms are particular for each amino acid (Figure 1.3) and are what define the amino acid chemical properties.

The ordered list of amino acids that compose a protein is called its *sequence*. It is important to note here that, as the structure (i.e. atoms and bonds) of every amino acid is clearly defined, the sequence of a protein unambiguously defines all its heavy atoms and how they are bonded. The protonation state of each amino acid will depend on the local microenvironment and pH. There exist several computational approaches to predict the hydrogen atoms of a biological molecule, such as the widely used H++ server[21] or the AddH option in UCSF Chimera.[22]

If we take into account that proteins generally have between 50 and 2,000 amino acids,[23] it gives us an idea of how large is the *biological space* that we have to cover to explore all the possible proteins that can exist. For example, suppose that we want to explore all the proteins that are 50 amino acids long. Each position in the sequence can be occupied by one of the 20 standard amino acids. As each of the positions in the sequence are independent, this gives a total of $20^{50}$ possible proteins of length 50. It becomes evident that, even for small proteins, it is not feasible to explore all the possibilities through brute force approaches. Therefore, we will usually refer to "exploring the biological space" limiting it to mutations on a few amino acids of the sequence: those that are supposed to be key in the function of the protein (e.g. the catalytic residues in an enzyme, or the amino acids that compose the binding site of a small molecule).

An illustrative example using the FABP4 protein is shown in Figure 1.4. The seven amino acids located within 4 Å of the ibuprofen molecule were selected

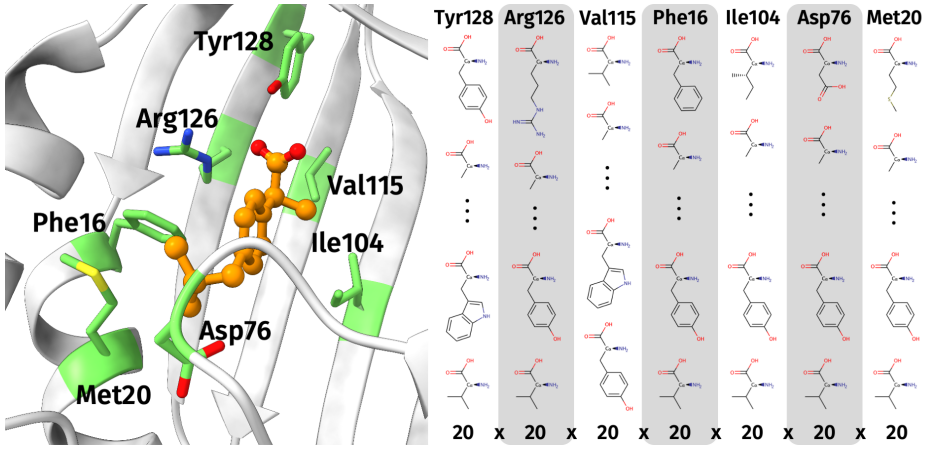



**Figure 1.4:** Exploration of the biological space for the residues at the binding site of FABP4. Residues composing the binding site are highlighted in green sticks, while the ibuprofen molecule is shown in orange balls and sticks. Each amino acid can remain the same or be mutated to any other standard amino acid, making a total of $20^7$ = 1.28 x $10^9$ combinations to explore.

as the "binding site" of the protein. A possible experiment would be to sample the biological space of the binding site, with the aim of improving the affinity of ibuprofen towards FABP4 and shedding light on the binding mechanisms of this protein. Even with this reduced biological space, a systematic exploration of all the possibilities would involve 1.28 x $10^9$ combinations, certainly beyond the reach of any experimental method, but possible with some computational approaches.

## The chemical space

Other chemical entities apart from a protein are present in a biochemical system, which we designated as the chemical part. One or more of the following entities could be important to the study of a system: drugs, organic and/or inorganic cofactors, glucose derivatives, lipids, and/or solvent molecules. Just for illustration, and continuing with the example of the human FABP4, more than 20 small molecules are reported in complex with that protein in the PDB. Indeed, this chemical promiscuity has been exploited for enzyme design using the FABP4 scaffold as starting point.[16]

In general, the number of atoms of the chemical part will be much less than that in the biological part. For example, the FABP4 protein (biological part) has 1185 heavy atoms, whereas the ibuprofen molecule (chemical part) has only 15. This difference in size between the biological and the chemical parts could give the impression that systematically sampling the *chemical space*, that is, the set of all possible molecules, could be computationally much cheaper than sampling the biological space. Nothing is further from reality.

The chemical structure of a standard amino acid is well known. Remember that when we talked about sampling the biological space we were focused on exploring all possible combinations of amino acids, not atoms or bonds. This is not the case in chemical space, where the chemical structure of the molecule is not known in advance if we are planning a blind sampling. Exploring all possibilities here will mean precisely trying all possible combinations of atoms and bonds between them to form all possible molecules. Analytically evaluating the number of different molecules that can exist is more difficult than in the case of biological space. 118 different elements are present in the latest version of the IUPAC periodic table,[24] which can be linked with a variety of covalent and/or ionic bond types. As an illustrative example, the chemical space of all drug-like molecules has been estimated to be $10^{60}$ different molecules[25] In any case, it now becomes apparent that the systematic sampling of biological or chemical spaces faces a similar problem: the rapid scaling of the number of solutions with the number of residues for the former and the number of atoms for the latter.

Chemical space exploration is a hot topic nowadays, and the in silico generation of molecules with desired properties has become a common step in drug discovery protocols,[26] among other applications such as new materials design.[27,28] There exist attempts to generate databases of molecules by systematically sampling a region of the chemical space. One of the most relevant works in this sense is "The chemical space project".[29] However, the big size of the databases generated, many of which containing more than a billion of molecules, affects their applicability. Methodologies to navigate more effectively the chemical space are generally more suitable, as they generate appropriate molecules to the problem at hand while considering much fewer so-

lutions. They can be implemented using several approaches, including deep learning[30–32] and genetic algorithms.[33–36] According to the coarseness of their molecular representation,[37] methods can be classified as atom-based,[35,38,39] fragment-based,[40–42] or reaction-based.[43–45]

## The conformational spaces

Both biological and chemical spaces are related to the chemical composition of the system, that is, what its atoms are and how they are bonded. As we have seen, this is not enough in structural biochemistry, where the study of a system is intimately related to the arrangement of atoms in three-dimensional space. The set of all atomic coordinates of a molecule constitutes a *conformation* of this molecule. Molecules, however, are rarely rigid entities and possess an inherent flexibility that allows them to function through intermolecular interactions. Therefore, considering a unique conformation for a biochemical system could not be sufficient for some purposes, making it necessary to sample at least a set of conformations energetically feasible. This is called exploring the conformational space. Of course, in biochemical systems, there will be a conformational space for the biological part and another for the chemical part. Both will be intrinsically related when it comes to their interactions.

In the case of proteins, two interrelated levels of flexibility are conventionally considered: global and local.[46] Global flexibility is usually associated with slow processes, and includes the folding of the entire protein and collective movements of large amplitude, involving, for instance, domain motions. To achieve these large movements, it is unavoidable that changes in the torsion angles phi and psi of the backbone bonds occur (Figure 1.5A). Local flexibility, on the contrary, is associated with faster and small amplitude motions, and usually affects only the rearrangement of a few amino acid side chains (Figure 1.5B).

Therefore, the computational sampling of these two types of movements can require very different methodologies. For global motions, traditional methods like classical molecular dynamics simulations (MD) can provide a good sampling sometimes, but usually fall short of providing an adequate description of

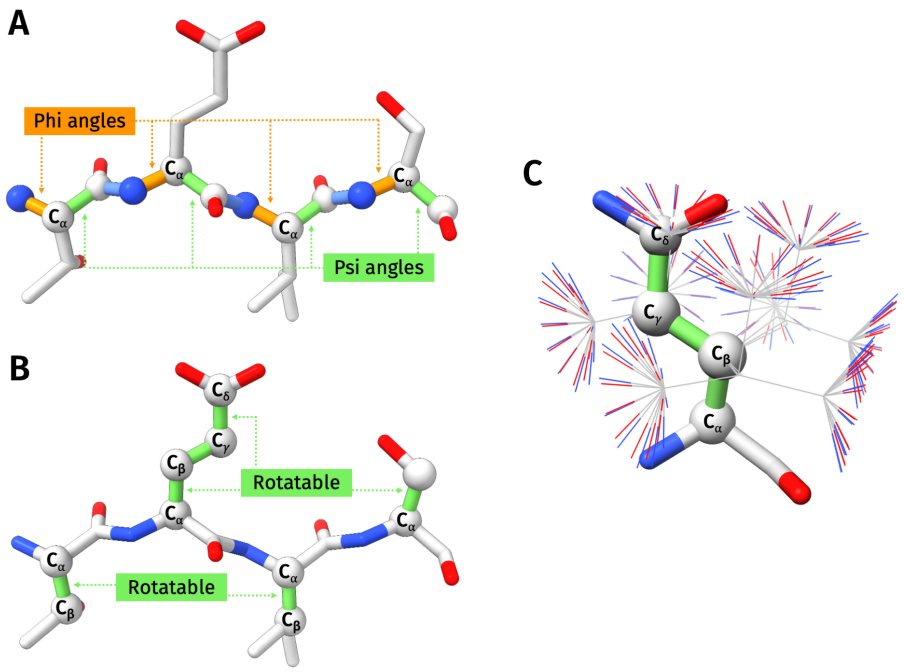



**Figure 1.5:** Conformational space of a protein. **(A)** The global flexibility of the protein depends on the backbone torsion angles phi (rotation of bonds N-$C_\alpha$, in orange) and psi (rotation of bonds $C_\alpha$-C in green). Rotations at the peptide bonds (in blue) are mostly fixed around 180º, due to their partial double-bond character. **(B)** The local flexibility of the protein depends on the rotatable bonds of the side chain, which are highlighted in green in the figure. **(C)** Examples of rotamers allowed for a Gln residue according to the Dunbrack library.[47] Rotatable bonds ($C_\alpha$-$C_\beta$, $C_\beta$-$C_\gamma$, and $C_\gamma$-$C_\delta$) are shown in green.

events at time scales beyond microseconds.[48] In this case, enhanced methods like gaussian accelerated molecular dynamics (GaMD[49]) could be an option. Also, analytical methods, exemplified by normal mode analysis (NMA) are widely used to provide physically plausible information on cooperative events,[50,51] although they may lack atomic details. For local motions, besides the MD-based approaches, one option would be to rely on the systematic sampling of the rotational bonds allowed to the different bonds of amino acid side chains. However, it has been demonstrated that side chains adopt only a subset of their in principle allowed conformations.[52] In this sense, several libraries of the so-called *rotamers* (i.e. conformations that the side chains could adopt, Figure 1.5C)

have been developed, which can be divided mainly into two groups: backbone-dependent[53,54,47] and backbone-independent[55–57] libraries.

For the chemical part of the system, a systematic sampling of all the torsion angles of the molecule can be feasible sometimes, but, even for relatively small molecules, it could produce a combinatorial explosion. By way of illustration, the arachidonic acid structure has 14 rotatable bonds (Figure 1.6). If we employ a precision of one degree for each of the torsion angles, a brute-force approach should sample $360^{14} = 6 \times 10^{35}$ different conformations. However, as in the case of the amino acid side chains, not all the theoretically available conformations are in reality adopted by the molecule. Only those energetically viable will be useful.

Therefore, the goal here is to obtain an ensemble that properly represents all possible low-energy conformations of the molecule. Again, MD-based methodologies are a safe bet when talking about conformational sampling. Due to the small size, *ab initio* MD could be a possibility if high accuracy is required.[58] Depending on the necessities of the problem, other less computationally demanding MD-based methods can be used: multiscale QM/MM-MD,[59–61] GaMD[49] and classical MD. Finally, conformer generators based on Molecular Mechanics (MM) and experimental evidence are a valid option when a good balance between speed

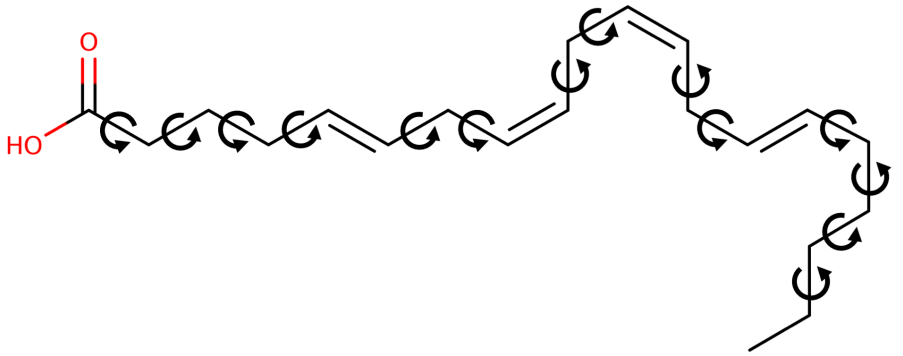



**Figure 1.6:** Arachidonic acid molecule. Its 14 rotatable bonds are indicated with curved arrows.

and accuracy is required.[62–64] Continuing with the example of arachidonic acid, a set of 951 low-energy conformations was found for this molecule with the RDKit ETKDG conformer generator.[62]

## The relative orientation of the different parts

The three spaces we have seen so far are related to a single part (biological or chemical) of the biochemical system, which is treated as an isolated element. It naturally follows that the last space we need to explore is the interaction between the different parts, that is, how each one is located in three-dimensional space in relation to the other(s). In fact, studying the physical interaction between a small molecule and a protein is one of the most scientifically interesting problems in molecular modeling, because it determines the biochemical properties of the system. The answer to questions such as "will this drug inhibit the function of the protein?", "how does this biological process work at the molecular level?", and "how could we improve the reaction catalyzed by this artificial enzyme?" will largely depend on our correct understanding of the interactions between the chemical and biological parts of the system.

Once again, a systematic sampling of all possibilities would imply computationally infeasible spaces. Assuming one conformation for the protein and another for the chemical part, we will need to explore all possible rotations of the chemical part by all possible translations of the chemical part with respect to the protein. Using a precision of one degree for angles and 0.01 Å for distances, it would give search spaces on the order of $10^{17}$ even for a small protein with a volume of $10^4$ Å$^3$. However, as in other cases explained above, only a very small proportion of all possible solutions will be feasible: those with lower energy, which indicates that the interaction between the chemical and biological parts is more favorable than in other spatial orientations. Therefore, computational methodologies intended for this task (e.g. docking programs) will need to implement efficient algorithms to find those biochemically relevant poses among the vast space of irrelevant possibilities (Figure 1.7).

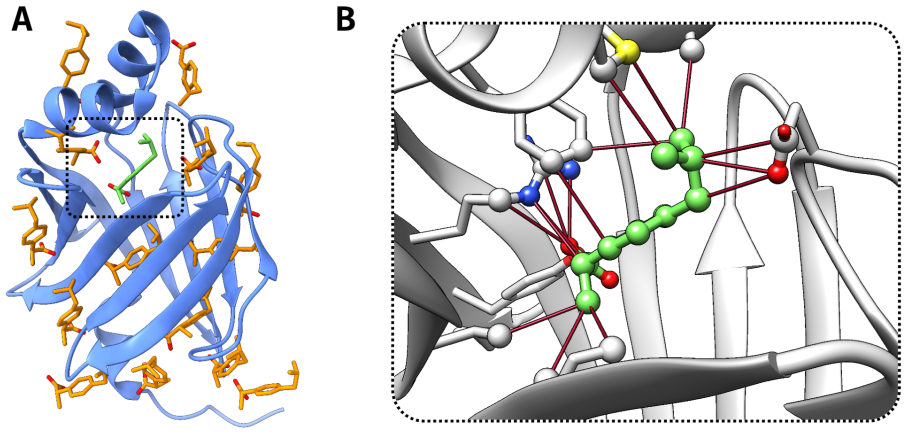



**Figure 1.7:** Exploring the relative orientation of the ibuprofen molecule with respect to the FABP4 protein structure. **(A)** Multiple options are possible. In orange there are depicted several locations and orientations of the ibuprofen molecule. In green it is highlighted the crystallographic pose (PDB entry 3p6h). **(B)** Detail of non-covalent ibuprofen-FABP4 interactions that stabilize the crystallographic pose. Interactions were obtained with the default options in UCSF Chimera software[22] (hydrogen bonds and contacts) and are highlighted in thin dark red sticks.

**Interrelationship between spaces** At this point, we have a description of all the spaces involved in the molecular modeling of a biochemical system. However, there is one issue that we have only touched on obliquely so far: the interrelationship between the different spaces. It is evident that a change in the atomic composition of the system will have consequences in its conformational space: even a small modification in the sequence of a protein can have a high impact on its conformation and therefore on its function, and similarly it happens with the chemical part of the system.[23] More subtle are the effects that the interactions between different molecules can have on their respective conformational spaces. In this sense, how the binding of a small molecule impacts on the protein conformations has been, and still is, the subject of a long and intense debate in the community. There are two main lines of thought: conformational selection and induced fit.[65,66] Conformational selection hypothesizes the existence of well-preorganized conformations of the unbound protein, among which the ligand "selects" the optimal one to allow binding (Figure 1.8A). In contrast, induced fit suggests that the geometries of the ligand and protein are displaced because of their interaction. In a way, there is no pre-equilibrium, and the pres-

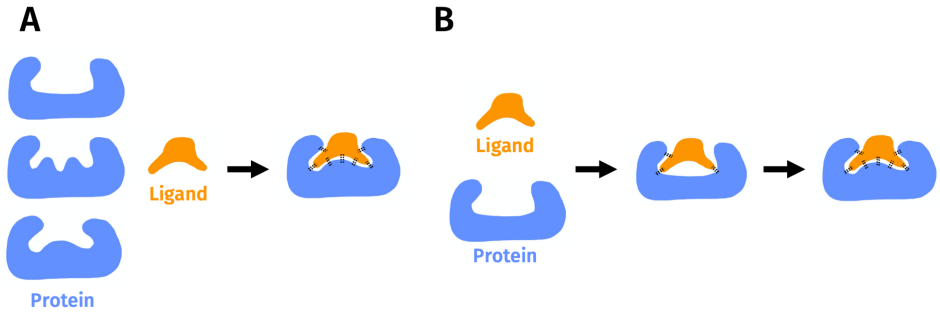



**Figure 1.8:** Schematic representation of a ligand binding to a protein. Proteins are depicted in blue, whereas ligands are shown in orange. Non-covalent protein-ligand interactions are represented by black dashed lines. **(A)** Conformational selection. One protein conformation among the unbound ensemble is optimal for ligand binding and thus it is the one "selected" by the ligand. **(B)** Induced fit. The ligand binds to a non-optimal unbound conformation of the protein. The binding causes a rearrangement in the protein to improve the accommodation of the ligand.

ence of the ligand modifies the protein conformation (Figure 1.8B). More recent studies suggest that conformational selection has a preponderant role, although induced fit can be sufficient to explain the binding process sometimes.[67,68] In any case, computational studies should take into account the additional complexity of the interrelations between spaces, either explicitly in the algorithm or by making assumptions for each particular system.

## The special case of metals

There is still one challenge more to mention in the computational study of biochemical systems: the presence of metallic species. Metals are of great importance for living organisms: they are present in approximately one third of the structures in the PDB and affect numerous biological processes, such as metabolism, muscle contraction, enzyme regulation, or oxygen transport.[69] However, the incorporation of metals in computational methodologies presents several technical difficulties we must overcome. First, the specificity and high diversity of the interactions that a metal atom can establish with its neighboring atoms, which can vary from almost pure electrostatic for the main group metals to suitable coordination bonds in the transition metals. As the force varies depending on the metal, the software will need to have a custom set of parameters

for each element to properly weigh and evaluate such interactions.[70] Second, even for a given metal, it could occur in several oxidation states, each with its own peculiarities. This is, for example, the case for iron, for which $Fe^{2+}$ and $Fe^{3+}$ are the most common oxidation states. It is not always necessary to consider this level of detail in the description of the model, but for some problems, such as the study of chemical reactions, it may be unavoidable. Finally, there is a compatibility issue between different computational methodologies. As we will see in the following section, it is common in molecular modeling to use computational workflows where different methods are applied sequentially to solve a problem. If adjusting one method to support the introduction of metals can be cumbersome, coordinating several increases the difficulty even more. Sometimes using specialized scripts such as MCPB.py[71] solves the problem, but other times more manual approaches are required.

One of the hallmarks of the InsiliChem group is the incorporation of metals in molecular modeling studies. Our group has worked intensively on the implementation, design, adaptation, and application of computational tools and protocols in this field. In this sense, this thesis is not an exception, and an important part of the efforts will be dedicated to this task. On the development side, we will evolve the group's previous efforts in understanding protein-metal interactions,[72–85] by conducting a comprehensive study of all the structures available in the MetalPDB,[86,87] ending with a new tool capable of predicting metal-binding sites in proteins (chapter 6). On the application side (chapter 8), we will use specialized computational workflows to the study of two systems: first, to decipher the interactions of a metallodrug –oxaliplatin– with a protein receptor –insulin–; and second, to shed light in some aspects of the mechanisms that lead to enantioselectivity in a cyclopropanation reaction catalyzed by a metalloenzyme.

## 1.3 Exploration of more than one space

The field of computational biochemistry is vast. But yet, only a fraction of the current stand-alone state-of-the-art methods are generally able to focus on more

than one given space at the time. Some methods are better to screen massive chemical spaces but this is only achievable with a limited conformational scope. Others are extremely valuable and accurate when it comes to studying massive conformational spaces, like molecular dynamics, but are only achievable on a limited number of systems. However, it is common for the understanding of a given biochemical system the need to account for a combination of biological, chemical, and/or conformational spaces. In those cases, the general way to move forward is by multiscale workflows, although other approaches based on the simultaneous exploration of such spaces, like some of the ones explored in this thesis, could open some new avenues.

## Multiscale workflows

The philosophy of molecular modeling based on computational workflows stands on designing a set of calculations that are executed sequentially. Each step along the process is provided by methods able to deal with different kinds of events, different computational cost, and leading to distinct energetic accuracies. The output of one step becomes the input of the next, and conclusions can be drawn from the analysis of each step. An advantage of this multi-step approach over simultaneous exploration is that it allows better control of the spaces explored. However, some interdependent effects between the different spaces could be lost, since there is no interrelationship between the workflow steps. The different steps usually involve calculations at different levels of theory, and that is why this approach is often also called "multiscale".

## Simultaneous exploration

In opposition to the concept of multiscale workflow, the other alternative is the simultaneous exploration of the different spaces that were introduced in the second section of this chapter: chemical, biological, conformational, and relative orientation between parts of the system. A clear advantage of this approach is to explicitly account for the interrelation among the different spaces, neutralizing the risk of losing good solutions in the way of a computational workflow. However, these methods entail a higher risk of combinatorial explosion in the

exploration, in which case the results of the simulation might not be representative of the system at all. To control this risk, **exploration must be restricted and guided**.

An example of a state-of-the-art methodology in this regard is AutoDock Vina 1.2.0,[88] which enables the simultaneous docking of multiple ligands, thus exploring their orientation relative to a protein, while at the same time exploring local protein flexibility (i.e. rotamers) and ligand(s) conformations through sampling of their rotatable bonds. To achieve good search efficiency, Vina guides the exploration with a Monte-Carlo[89] algorithm combined with the BFGS[90] gradient-based optimizer. To control the exploration and ensure a good result, the user is in charge of carefully selecting the constraints of the exploration to avoid combinatorial explosion. The basic parameters to configure in this case are: the exhaustiveness of the search, the size of the system through the search box, the degrees of flexibility allowed to the ligand(s), and which residues of the protein are allowed to rotate.

**GaudiMM** Several years ago, in our group InsiliChem, a tool was devised that could allow simultaneous exploration of as many spaces as needed for the modeling problem at hand. The philosophy behind the concept is to allow the researcher to make hypotheses regarding their system and let the software give the best solution(s) to the problem. This *hypothesis-driven* exploration would allow us to tackle non-standard modeling tasks not possible with other software. Some illustrative examples of research questions that could be addressed with the software are: "what would be the structure with best ligand-protein interaction if the distance between these two atoms is kept at $x$ Å?", "would be possible a folding of this peptide on a volume less than $y$ Å$^3$?", or "would be possible a rearrangement of the amino acid side-chains in the binding site of a protein to allow a metal binding?". The first version of the code was developed in the group mostly under the efforts of J. Rodríguez-Guerra. The methodological grounds of the tool, called GaudiMM,[91,92] will be presented in more detail in the following chapter.

## 1.4 Scope of the thesis

The work of this thesis is positioned within the context of current methodologies to the structural exploration of biochemical systems. Part of the effort is dedicated to expanding the limits of state-of-the-art simultaneous multi-space exploration, introducing within the framework of GaudiMM a module for exploring ligand binding routes (chapter 4) and another for sampling the chemical space (chapter 5). Also, the code of GaudiMM was updated and applied to several non-standard docking tasks (chapter 5 and chapter 7). The other part of the work focused on the introduction of metals in computational modeling protocols. In this sense, we developed a tool for the prediction and design of metal-binding sites in proteins (chapter 6) and we designed and applied new computational workflows for the study of metallodrug interactions and artificial metalloenzymes (chapter 8).

# 2

# Methodological grounds

One of the major takeaway messages of the introduction chapter is that the exploration of one or several spaces of complex (bio)chemical systems must be constrained and guided in today's molecular modeling experiments. This is mostly due to the relationship between the number of degrees of freedom considered and computational cost. There are two levels of complexity that the software must handle with the least amount of computation and highest level of quality: i) the generation of candidate structures that satisfy the experiment's constraints; and ii) the assessment of the goodness of the generated structures, which guides the exploration towards optimal solutions.

One or both of these components will appear in the discussion of all the methodologies presented in this chapter, which is divided into two sections. In section A, we will present the grounds of the state-of-the-art approaches that were used

in the application work of the thesis (chapter 7 and chapter 8). This part includes a description of the different levels of theory used in the energetic evaluation of systems, as well as the methods employed for the exploration of the conformational space. In section B, we will introduce the concept of optimization based on evolutionary algorithms. Specifically, we will describe the fundamentals of genetic algorithms and genetic programming, on which we built most of the development part of the thesis (chapter 4, chapter 5, and chapter 6).

# A. Well-established approaches in molecular modeling

## 2.1 Energetic evaluation of the system

Broadly speaking, the goal of any molecular modeling study is to provide a description of the energy of the system as accurately as possible, although other metrics may be important in some cases. If we assume that useful three-dimensional structures of biochemical systems must be in an accessible energy range, their energetic evaluation becomes of great importance to discern between "good" and "bad" structures. For this purpose, there are two main families of methods: those based on quantum chemistry, aimed at modeling electrons and atomic nuclei and providing maximum accuracy; and those based on classical mechanics that describe atoms or groups of atoms using Newton's laws of motion, discarding the electronics of the system for the sake of simpler and less computationally demanding calculations. Although an exhaustive description of the theory behind both methodologies is beyond the scope of this thesis, the following paragraphs will provide the key concepts to understand their application potential and limitations.

### Quantum mechanics

Methodologies based on quantum mechanics (QM) are those in which electrons are explicitly considered in the model and, therefore, allow to derive proper-

ties that depend on the electronic distribution. Specifically, QM methods allow the modeling of processes in which chemical reactions (i.e. bond breaking or bond formation) occur. In fact, the covalent bonds are not considered explicitly, but can be derived later from the nuclei positions and electronic distribution obtained in the calculation. Putting it in the context of the exploration spaces defined in the previous chapter, it means that QM methodologies allow sampling the conformational space (i.e. the spatial arrangement of atoms) at the same time as sampling part of the chemical space (i.e. bond breaking or creation). However, as we will see in the following paragraphs, the high accuracy of QM approaches comes with a high computational cost, which in practice makes them applicable nowadays only to small systems (hundreds of atoms).

The theory of quantum mechanics is mostly based on the Schrödinger equation, which postulates that the state of a given particle (e.g. an electron) in the system can be completely described by a *wavefunction* $\Psi(\mathbf{r}, t)$ that depends only on the spatial position of the particle ($\mathbf{r}$) and the time ($t$). If we consider a situation where the potential energy is only dependent on the particle's position and therefore independent of time, the Schrödinger equation can be written in its time-independent form (Equation 2.1), which allows us to describe the stationary states of the system.

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + V\right\}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{2.1}$$

$\hbar$ : Planck's constant
$m$ : mass of the particle
$V$ : potential
$\mathbf{r}$ : position vector

Where:

$$\nabla^2 = \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.2}$$

And can be abbreviated using the Hamiltonian operator:

$$\hat{H}\Psi = E\Psi \tag{2.3}$$

The Schrödinger equation can be solved analytically only for very simple systems: those containing a single electron. For more complex systems, we should rely on some approximations, being the most relevant ones: i) the Born-Oppenheimer approximation, that considers the fact of electrons being much lighter than nuclei and therefore simplifying the wavefunction to depend only on the position of the nuclei and not on their momenta. This allows us to solve the equation exactly for the simplest molecular species ($H_2^+$ and isotopically equivalent species).[93] ii) The perturbation theory, which is based on the idea that a complicated system can be solved starting from a simpler one and adding an additional Hamiltonian that represents a weak perturbation to the system. If the perturbation is small enough, the expressions obtained, although not exact, can be accurate. And iii), the variational method, which relies on substituting the exact wavefunction by an alternative trial orthonormal wavefunction $\Phi$. The method ensures that the energy calculated for the trial wavefunction $\Phi$ will always be equal or greater than the real energy ($E_\Phi \geqslant E_o$). Therefore, fine tuning the parameters of the trial wavefunction will allow us to accurately approximate the ground state energy.

**Hartree-Fock method** Approximations i) and iii) are at the core of the Hartree-Fock (HF) method, which assumes that each electron in the system can be approximated by one spin orbital.[94] A Slater determinant is constructed to describe the complete wavefunction as an anti-symmetric product of spin orbitals, accounting for electron coordinate changes in rows and spin orbital changes in columns (Equation 2.4). Assuming that the variational principle holds, the Slater determinant with lowest energy ($\delta E = 0$) will be the best approximation to the true energy. This condition of minimum energy with respect to changes in the orbitals is the base to obtain the HF equations. Usually, HF equations are not solved numerically; instead, a linear expansion of the orbitals is used over a standard basis function set, being the most commonly used Slater-type and Gaussian-type

orbitals.[94,95]

$$\Psi(\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_n) = \frac{1}{\sqrt{N!}}\begin{vmatrix} \varphi_1(\mathbf{r}_1) & \varphi_2(\mathbf{r}_1) & \cdots & \varphi_N(\mathbf{r}_1) \\ \varphi_1(\mathbf{r}_2) & \varphi_2(\mathbf{r}_2) & \cdots & \varphi_N(\mathbf{r}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{r}_N) & \varphi_2(\mathbf{r}_N) & \cdots & \varphi_N(\mathbf{r}_N) \end{vmatrix} \tag{2.4}$$

However, the fact of approximating each electron by a single spin orbital leads to one of the major drawbacks of the HF theory: it neglects the correlation between electrons. Instead of an explicit electron-electron repulsion, the HF method assumes an average electron charge cloud that interacts with every electron in the system. This introduces an error in the wavefuntion and the calculated energy (called *correlation energy*), which makes the method not accurate enough to make quantitative predictions.[93]

**Post-Hartree-Fock methods** To introduce electron correlation into the many-electron wavefunctions, several post-Hartree-Fock approaches were developed, which can be mainly classified into two groups: those which try to correct the single slater determinant approximation (e.g. configuration integration approaches), and those which introduce correlation energy through perturbation theory (e.g. Møller-Plesset perturbation theory).[96] Post-HF methods are characterized for a higher accuracy in the energy estimation with respect to HF methods, although at a (much) higher computational cost.

**Density functional theory** In contrast to HF and post-HF methods, where a direct approximation of the wavefunction is used, in density functional theory (DFT) the ground energy of the system is approximated using only the electronic density. The relation between the electronic density and the wavefunction is guaranteed by the Hohenberg-Kohn theorem.[97] The central problem in DFT methods is therefore to obtain adequate functionals that properly approximate the electronic density. Some common functionals are: local density approximation (LDA), generalized gradient approximation (GGA), and meta-generalized gradient approximation (MGGA). Also, there exist hybrid functionals, such as

B3LYP, B3PW91, mPW91, and PBE1PBE; they employ a weighted combination (calibrated against a reference dataset) of the expressions derived by LDA, GGA, MGGA, and HF methods.

DFT has attracted a wide interest in the last decades because it allows the inclusion of electron correlation. In general, DFT approaches can obtain an accuracy in between HF and post-HF methods, employing a slightly higher computational cost than HF. Both metrics –accuracy and cost– heavily depend on the funcional employed. In the application projects of this thesis, DFT calculations using hybrid functionals were employed to derive force field parameters for non-standard and metal atoms, for the QM/MM optimization of the oxaliplatin-insulin adducts, and for obtaining the reaction mechanisms in the cyclopropanation reaction catalyzed by a $Rh_2$ cofactor (chapter 8).

## Molecular mechanics

In opposition to QM approaches, in molecular mechanics (MM) methodologies the electronic configuration of the atom is not taken into account explicitly: the nucleus and the electrons around the nucleus are simplified to a perfect sphere, and covalent bonds are treated as springs (Figure 2.1). A direct and important consequence of this simplification is that covalent bonds between atoms should be defined *a priori*, which severely limits the application of MM methods in systems where chemical reactions (i.e. breaking and forming bonds) occur. However, the great advantage of MM models is that the energy evaluation is much faster than in QM methods, because it is based on the (simpler) classical laws of physics. This allows the application of the MM energy evaluation in bigger systems (thousands or even millions of atoms), such as biochemical complexes.

To estimate the potential energy ($E$) in a MM model, two contributions (covalent and noncovalent) are considered and evaluated for all the atoms of the system:

$$E = E_{covalent} + E_{noncovalent} \tag{2.5}$$

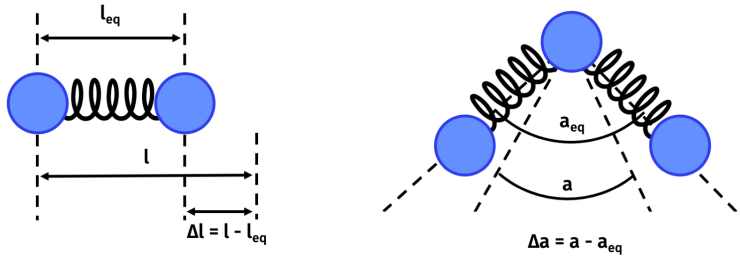



**Figure 2.1:** Ball-and-spring model for molecular representation. Atoms are treated as perfect spheres, while covalent bonds are treated as springs. Changes in bond lengths or angles result in changes in the energy of the molecule.

Only those covalently bonded atoms will have covalent contribution, which is composed by three main terms: bond stretching, angle bending and dihedral torsion (Equation 2.6). Although the exact functions to calculate each of these terms may differ from software to software, the stretching and bending terms are usually modeled as harmonic potentials centered around the equilibrium length and angle of the bond (Equation 2.7, Equation 2.8, and Figure 2.2A-B). In contrast, the functional form of the dihedral torsion term must take into account multiple energy minima and, therefore, needs more complex implementations, for example expressing it as a cosine series expansion (Equation 2.9, Figure 2.2C). Additionally, it may be necessary to include other terms, such as improper torsions (i.e. between atoms not bonded in the sequence), to model out-of-plane bending motions.[93] Finally, cross terms may be included in the energy estimation to account for interrelations between the other three terms (e.g. stretch-stretch, stretch-torsion, and bend-bend).

$$E_{covalent} = E_{str} + E_{bend} + E_{tors} + E_{cross} \tag{2.6}$$

$$E_{str}(l) = \frac{k}{2}(l - l_0)^2 \tag{2.7}$$

$l$ : displacement from $l_0$
$l_0$ : reference bond length
$k$ : stretching constant of the bond

$$E_{bend}(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \tag{2.8}$$

$\theta$ : angle displacement from $\theta_0$
$\theta_0$ : reference angle
$k$ : force constant

$$E_{tors}(\omega) = \sum_{n=0}^{N} \frac{V_n}{2}[1 + cos(n\omega - \gamma)] \tag{2.9}$$

$V_n$ : rotation barrier
$n$ : multiplicity
$\omega$ : torsion angle
$\gamma$ : phase factor

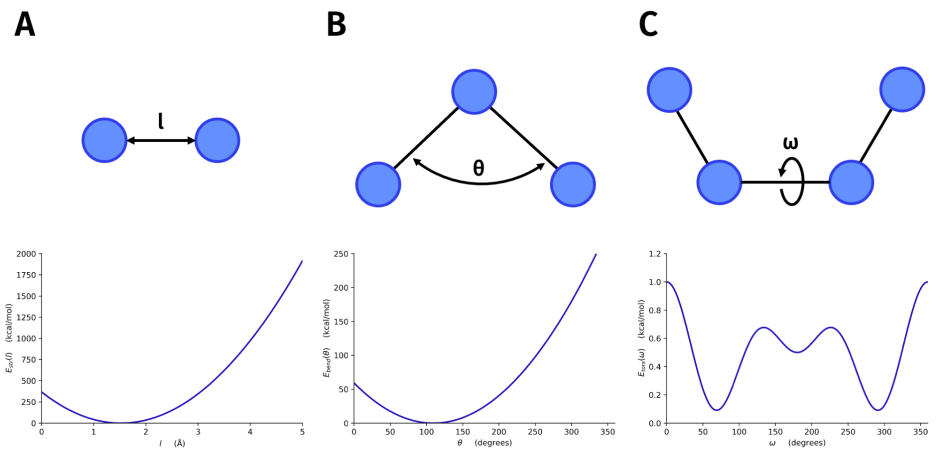



**Figure 2.2:** Covalent interactions. **(A)** Example of bond stretching energy for a $Csp^3$-$Csp^3$ bond, modeled as an harmonic potential. **(B)** Example of angle bending energy for a $Csp^3$-$Csp^3$-$Csp^3$ angle, modeled as an harmonic potential. **(C)** Example of angle torsion energy for a O-C-C-O angle, modeled as a cosine series expansion of two terms.

In turn, noncovalent interactions are estimated as the sum of van der Waals, hydrogen bonding, and electrostatic contributions (Equation 2.10). Note here that an exhaustive calculation of the noncovalent term would involve a high computational cost, since each atom interacts with every other atom in the system. Therefore, some simplifications are usually applied. In the case of van der Waals and hydrogen bonding contributions, a typical model to calculate the interaction between two atoms is the 12-6 Lennard-Jones potential (Equation 2.11, Figure 2.3A), which depends on two parameters: the internuclear distance at which the energy is minimum ($r_{\mathrm{m}}$) and the well depth ($\varepsilon$). In addition, to further speed up the calculation, a cutoff value ($r_0$) is used for the internuclear distance from which the interaction energy is considered zero. In the case of electrostatic interactions, they are usually estimated through the classical Coulomb potential, which, for two given atoms, depends on their point charges $q_{\mathrm{i}}$ and $q_{\mathrm{j}}$ (Equation 2.12, Figure 2.3B). Similarly to the case of the Lennard-Jones potential, a cutoff value for the internuclear distance can be introduced to speed up the calculation, although taking into account that here the interaction decays much more slowly.

$$E_{noncovalent} = E_{vdW} + E_{Hbond} + E_{elec} \tag{2.10}$$

$$E(r_{ij})^{LJ} = \varepsilon_{ij} \left\{ \left( \frac{r_m}{r_{ij}} \right)^{12} - 2 \left( \frac{r_m}{r_{ij}} \right)^{6} \right\} \tag{2.11}$$

$$E(r_{ij})^{Coul} = \frac{(q_i q_j)}{4\pi\varepsilon_0 r_{ij}} \tag{2.12}$$

In sum, we have seen how, in MM approaches, the estimation of potential energy is done by adding a set of more or less simplified functions. Each function depends on one or more parameters. The whole set of functions, together with the corresponding parameters for each atom/bond type, constitute what we call a *force field*. Parameters for a given function and atom/bond type can be

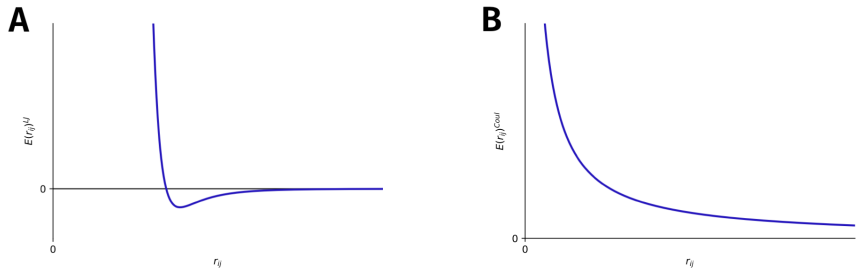



**Figure 2.3:** Noncovalent interactions. **(A)** Example of 12-6 LJ potential function between two atoms *i* and *j*. **(B)** Example of Coulomb potential function between two atoms *i* and *j*.

derived experimentally, from QM calculations, or from a combination of both approaches.

Over the years, many force fields have been designed with specific purposes in mind,[98–100] and more recently, even force fields derived from machine learning are emerging as an attractive alternative.[101] Within the framework of this thesis (chapter 8), we used the following force fields to carry out the simulations: the AMBER99SB[102] or AMBER14SB[103] for standard amino acids, the TIP3P model[104] for explicit water molecules, and the GAFF[105] for the rest of organic atoms. In the case of inorganic cofactors, a specific procedure was employed to generate the parameters of metal atoms. In our simulations, we opted for a bonded model, where an explicit coordination bond between the metal and its donors is present.[70] The metal-bonding force constants and equilibrium parameters were obtained for each particular case using the Seminario's method,[106] and integrated with the other parameters with the MCPB.py script.[107]

## QM/MM multiscale energy evaluation

As we have seen, the simulation of phenomena including chemical reactions will require energetic evaluation at QM level, being limited the MM level to those systems with a stable chemical structure. As the total cost of evaluating a structure rapidly scales with the number of atoms of the system, bigger systems will only be accessible when using MM approaches. To overcome this limitation, at

least in part, there is the possibility of combining methods with different accuracies to evaluate the energy of a biochemical system. This is called a *multiscale* model, and it earned Karplus, Levitt and Warshel the Nobel Prize in Chemistry in 2013.[108]

This hybrid approach makes it possible to simulate processes involving chemical reactions in larger biochemical systems than those accessible by pure QM methods. This is achieved by modeling the central part of the system (where the reaction occurs) at the QM level, while taking into account the conformational sampling of the surrounding atoms with a simpler MM-based method. Recovering, for the sake of illustration, the example of the FABP4-ibuprofen complex, we could simulate the ibuprofen molecule and its closest residues (119 atoms) at the QM level to observe if a chemical reaction is possible, while sampling the

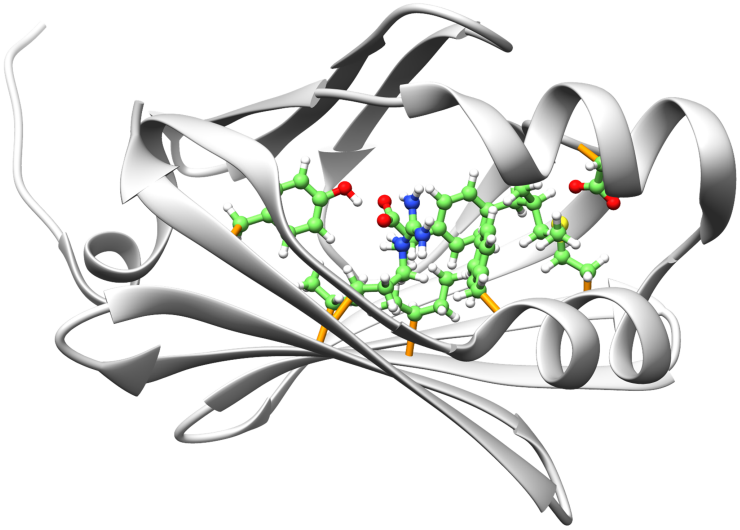


**Figure 2.4:** Division of the FABP4/Ibuprofen system into two parts for a multiscale QM/MM model. The QM part, shown in green ball and sticks, comprises 119 atoms: all the atoms of the ibuprofen molecule and side chain atoms of the binding site residues (Tyr128, Arg126, Val115, Phe16, Ile104, Asp76, and Met20). The MM part, shown in gray ribbon representation, comprises the remaining 2075 atoms. The linking points are the bonds between alpha and beta carbons of the amino acids involved in the QM part, shown in orange.

conformational space of the rest of the protein (2075 atoms) at the MM level (Figure 2.4). Note here that a boundary region exists between QM and MM parts. A proper choice and parametrization of these link points, together with a thoughtful choice of the QM region, are important aspects to achieve a successful result with QM/MM approaches.[109–112]

Concretizing the application of QM/MM in this thesis, we employed a multiscale strategy to optimize the structures in the project about the interaction of oxaliplatin with insulin (chapter 8). We used the Garleek[113] framework to configure and launch the calculations, which were carried out with Gaussian09[114] and Tinker[115,116] for the QM and MM regions, respectively. The MM part included only standard amino acids and was parameterized with the AMBER99SB[102] force field implemented in Tinker,[115,116] while the QM region was described at the DFT level.

## 2.2 Conformational search

Besides the method used to estimate the energy, every molecular modeling algorithm has a step aimed at generating suitable structures (i.e. conformations) that will then be energetically evaluated. Similarly to energy assessment, the accuracy and complexity of the exploration will impact the computational resources needed. The panorama of available methodologies is immense. In this section we describe those that are the most used in the application part of the thesis: i) energy optimization algorithms to find stable structures; ii) molecular dynamics to explore the evolution of the system over time; and iii) molecular docking as a semi-rigid exploration to find modes of interaction between the chemical and biological parts of the system. Again, an exhaustive description of the theory behind each methodology is beyond the scope of this thesis, but the key concepts to understand their application potential and limitations are provided.

## Energy optimization

An *energy optimization* process seeks a stable conformation of the biochemical system starting from an input structure. As stable structures are associated with physical significance, the goal is to obtain those conformations of the system that could be useful to understand its mechanisms and activity. Depending on the problem, the optimization will be directed to find the global minimum (i.e. most stable conformation), a local minimum (i.e. stable enough conformation), or a saddle point (i.e. transition state) of the potential energy function. In the two first cases, the optimization process is also known as *energy minimization*, because the algorithm seeks for a conformation with lower energy than that offered as input.

The potential energy function might be dependent on one or more coordinates (also known as decision variables or parameters of the function). If there exists only one coordinate, it is called a *potential energy curve* or energy profile (Figure 2.5A); whereas if the number of coordinates is higher than one, we talk of the *potential energy surface* (PES, Figure 2.5B). Note here that the PES is usually a *multimodal* function, meaning that it presents several local minima.

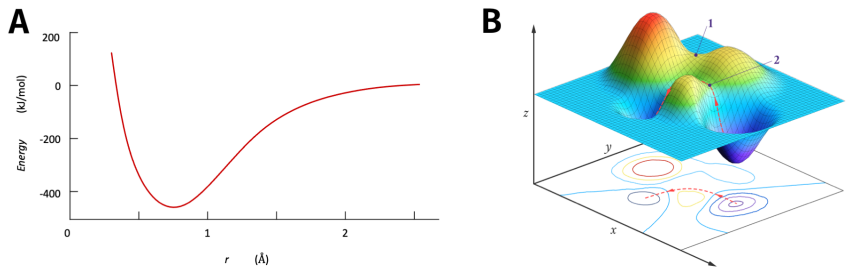



**Figure 2.5:** Potential energy function. **(A)** Example of potential energy curve for the covalent bond in a $H_2$ molecule, where *r* is the distance between the nuclei of the two H atoms. **(B)** Example of potential energy surface for a hypothetical endothermic reaction (from the global minimum to a local minimum of potential energy). Reaction coordinates are labeled as *x* and *y*. The *z* axis indicates the potential energy, which values are depicted in a blue-red range, being red the higher and blue the lower values. Saddle points of the PES are labeled as "1" and "2". Figure reproduced from Ümit Kaya via LibreTexts (CC BY-NC license).

Energy optimization processes were widely used in the application work of this thesis (chapter 8). In the case of the study of the POP-$Rh_2$ cyclopropanase, the reaction mechanism was obtained optimizing the energies at DFT level. In the study of interaction between oxaliplatin and insulin, an energy minimization at QM/MM level was carried out on the representative structures obtained in molecular dynamics simulations. In both cases, several steps of minimization at MM level were configured at the beginning of the MD protocol, in order to start the equilibration part of the simulation with a well-optimized structure.

## Molecular dynamics

*Molecular dynamics* (MD) simulates the evolution of the conformation of a biochemical system over time. The system, with a fixed number of particles $N$, and its thermodynamic properties can be treated on one of the following thermodynamic ensembles: i) *microcanonical* $(N, V, E)$, where the energy and volume are considered constant ; ii) *canonical* $(N, V, T)$, where the energy can change and the temperature is kept constant; iii) *isothermal-isobaric* $(N, P, T)$, in which the pressure and temperature are constant; and *grand canonical* $(\mu, V, T)$, where the chemistry potential $\mu$, as well as the volume and temperature, are constants.

The *trajectory* (positions and velocities of the particles along time) is obtained by solving the classical Newton's equations of motion (Equation 2.13) for the system in a step-by-step fashion. A time-step $\delta t$ is defined, usually in the range of a few femtoseconds. Then, given the atomic coordinates, velocities, and other dynamic information at time $t$, the respective quantities at time $t + \delta t$ are computed. The process is repeated by a numerical integration algorithm until the desired end simulation time is achieved.

$$m_i \frac{d\mathbf{v}_i(t)}{dt} = \mathbf{F}_i[\mathbf{r}(t)] \tag{2.13}$$

$\mathbf{v}_i$ : vector of velocities for particle i
$m_i$ : mass of particle i
$\mathbf{F}_i$ : total force acting on particle i

There exist a variety of integration methods.[117] One of the most widely used has been the Verlet algorithm, which obtains the positions at time $t$ from the Taylor expansions at times $t - \delta t$ and $t + \delta t$. However, this method presents two main limitations: the velocities are not explicitly calculated (which are needed to calculate the kinetic energy) and there exists a relevant loss of precision due to the truncation of the Taylor series. The leap-frog algorithm solves these problems by computing the velocities explicitly.[118] The positions and velocities are offset from each other by half a time step. In each step, they are updated as indicated in Equation 2.14 and Equation 2.15. It is important to note here that, as the positions are always half a time step later than the velocities, the calculation of the total energy of the system must be done carefully, using the same time step for the calculation of the potential and kinetic energies.

$$\mathbf{r}_i(t + \delta t) = \mathbf{r}_i(t) + \mathbf{v}_i\left(t + \frac{\delta}{2}t\right)\delta t \tag{2.14}$$

$$\mathbf{v}_i\left(t + \frac{\delta}{2}t\right) = \mathbf{v}_i\left(t - \frac{\delta}{2}t\right) + \frac{d\mathbf{v}_i(t)}{dt}\delta t \tag{2.15}$$

**Convergence of a MD simulation** One of the most difficult questions when running a MD simulation is the length that we need to ensure a good conformational sampling. As we have seen, biochemical processes can happen in a wide range of time scales. Even more, if we want to extract reliable statistics from the different conformations that our biochemical system can take, we need some measures that tell us that the error is small enough, that is, our simulation has *converged*. However, convergence in this case is almost a philosophical term. We will never be strictly sure that our simulation has converged; there will always be a chance to see something new in the next time step.[119]

In consequence, a more realistic goal is to assess the simulation convergence in the context of the specific objectives of the study at hand. In our case, we employ a pool of qualitative and visual analyses that can quickly suggest that the simulation has not explored sufficiently the required conformational space and, therefore, should be enlarged.[119–121] It concretizes in four analyses that share

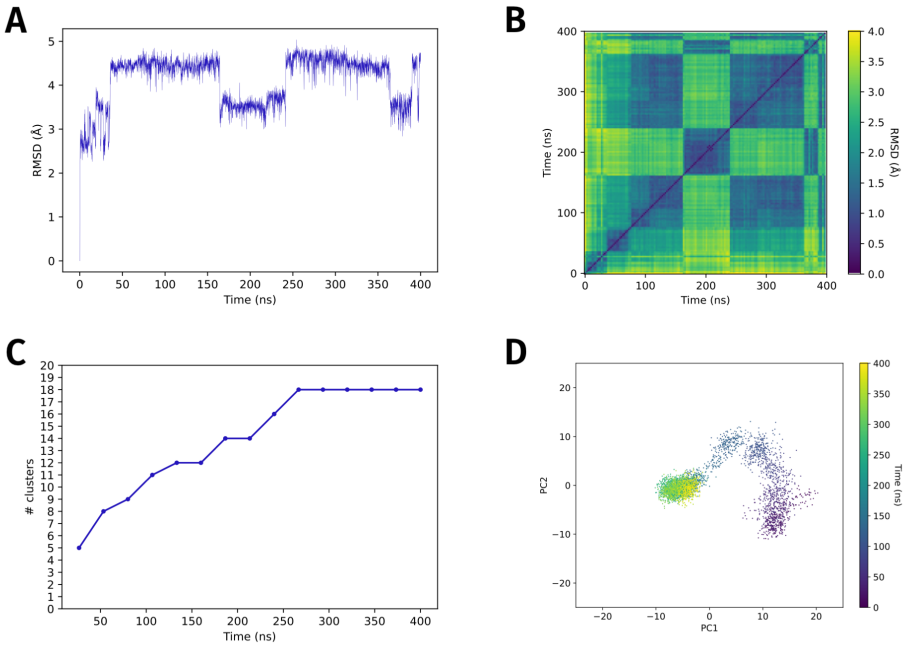



**Figure 2.6:** Example of qualitative convergence analysis. **(A)** RMSD from the first frame of the MD. **(B)** All-to-all frames RMSD. The RMSD value is indicated in a violet-yellow color range (being violet the lower and yellow the higher values). **(C)** Cluster counting analysis as described in Ref.[121] **(D)** PCA analysis. The first two principal components are plotted against each other. The MD time scale is indicated in a violet-yellow color range.

the same concept: there will be more probability that the conformational space is well sampled when we start to see the same conformations already explored and transitions between them. The four analyses (an illustrative example is shown in Figure 2.6) are: i) root mean square deviation (RMSD) from the first frame of the MD, ii) all-to-all frames RMSD, iii) principal component analysis (PCA), and iv) cluster counting. Note here that, traditionally, most molecular modeling studies analyze only the stabilization of the RMSD from the first frame of the trajectory (first analysis). However, this is often not enough, because the RMSD metric can hide different conformations under the same or similar RMSD value.

**Number of replicas of a MD simulation** Another topic that has been the object of intense debate in the community is the need to run multiple replicas of a MD

simulation starting from the same initial conformation. As the initial velocities of the system particles are assigned at random, MD approaches have an inherent stochastic component that could make it advisable to repeat several times a simulation to reduce the error and facilitate reproducibility. Although this reasoning would be undeniable if we had unlimited computational resources, obviously this is never the case. Therefore, the real question is: is it better to run one simulation of length $x$ or $N$ simulations of length $x/N$?

Again, to answer this question, we need to consider the specific needs of each research project. If we want to extract reliable statistics of the relative time that our system spends at each conformation, the best strategy will be to increment the number of replicas.[122] An example of this kind of study could be to analyze the percentage of simulated time that an amino acid side chain spends at a certain conformation. However, if we aim at seeing if our system can adopt a certain conformation, especially if it involves large-scale movements, the best strategy will be to enlarge one replica as much as possible. An example of this kind of study could be to check whether an alpha helix of the protein can unfold. If we split our available computational time into shorter replicas, we will have better statistics of quick processes but might lose slower movements like the alpha helix unfolding. In this thesis, the needs are closer to the second case. Therefore, we opted to enlarge a single replica controlling the quality of the exploration with the qualitative convergence analyses described before.

## Application to molecular dockings

The philosophy behind energy optimization processes and classical MD simulations is to guide the exploration with the energy evaluation function. Generally, no other constraints are imposed on the flexibility allowed to the system. In these cases, the initial structure is therefore of great importance in determining the conformational space that can be explored, which will be relatively close to it. When we want to do bigger trips in the exploration, other approaches are needed, such as *semi-rigid* ones.

One example is the prediction of the binding site and interaction modes of a

small molecule into a protein. This problem can be tackled with protein-ligand *docking* programs. The flexibility of the biological molecule is limited (usually to a few side-chains) and the exploration focuses on large movements, rotations, and conformational sampling of the small molecule (i.e. ligand). To achieve the high speed that characterizes these methodologies, they combine this restricted exploration with the use of a simplified MM forcefield to estimate the binding energy (*scoring function*). The result is a series of protein-ligand relative positions (*poses*), which are ordered according to the binding energy estimate provided by the scoring function. Precisely, the lack of accuracy of the scoring functions is one of the aspects that is usually blamed on docking. However, it should be remembered that the primary goal of the method is not to determine binding affinity with great accuracy, but rather to obtain reliable poses.

Among the wide panorama of docking software available,[123] several of them (AutoDock4,[124] AutoDock Vina,[125] and GOLD[126–128]) were used in the application chapters of this thesis. Also, the GaudiMM modeling platform[91] was configured to perform some challenging docking tasks. Although the specific methodological details of each case are reported in the respective works, in the following paragraphs we provide a general description of the exploratory and evaluative capabilities of each software.

**AutoDock4** The primary exploratory algorithm employed in AutoDock4[124] is a genetic algorithm with Lamarckian local optimization.[129] Previously to run the algorithm, the user is in charge of defining the zone of the protein where the ligand is able to bind. This is done by defining a rectangular-shaped box where the whole ligand should be contained. The method allows for the conformational sampling of the ligand and user-selected amino acid side chains, by allowing rotation around their torsional degrees of freedom. The binding poses are evaluated through a semiempirical free energy force field that includes six pairwise evaluations ($V$) and an estimate of the conformational entropy lost upon binding ($\Delta S_{conf}$):[130]

$$\Delta G = \left(V_{bound}^{L-L} - V_{unbound}^{L-L}\right) + \left(V_{bound}^{P-P} - V_{unbound}^{P-P}\right) + \left(V_{bound}^{P-L} - V_{unbound}^{P-L} + \Delta S_{conf}\right) \tag{2.16}$$

**AutoDock Vina** AutoDock Vina[125] (or simply, Vina) is an evolution of the AutoDock4 suite developed by the same group with the primary goals of accelerating calculations while providing a higher accuracy in the binding mode predictions. The constraints in the exploration are the same as in AutoDock4: a restricted conformational sampling and a user-defined search box. In fact, this latter aspect is of crucial importance to the performance of the program, being the optimum box dimension of 2.9 times the radius of gyration of the ligand.[131]

The exploratory and evaluative modules in Vina were completely redesigned. For the conformational and relative orientation sampling, they employ an iterated local search global optimizer,[132,133] together with a Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm for the local optimization.[90] The binding energy estimation, in its turn, is done with a much computationally-cheaper scoring function, inspired in X-Score.[134] For every pair of atoms that can move relative to each other, the program computes three terms: i) steric interactions based on the van der Waals radii; ii) hydrophobic interactions when the atoms belong to hydrophobic types; and iii) hydrogen bonding when the atoms have suitable types to be a donor-acceptor pair. The weighted sum of these three terms for all the atom-pairs constitutes the conformation-dependent part of the binding energy estimation. A conformation-independent term depending on the number of rotatable bonds of the ligand is then added to obtain the final score of a pose.

**GOLD** The Cambridge crystallographic data center (CCDC) developed the GOLD software[126–128] with the aim of providing highly accurate and flexible docking. Unlike AutoDock4 and Vina, which are offered under open-source licenses, GOLD has a proprietary license. The exploratory phase of the method is based on a genetic algorithm (see more about GA in section B of this chapter) to sample the conformational spaces of the chemical and biological components.

The allowed conformational sampling includes the torsional angles of the ligand, besides user-selected side-chains, which can be constrained to well-known rotamer libraries. The relative protein-ligand orientation, however, is not sampled directly by the genetic algorithm. Instead, GOLD uses a preprocessing method of the binding cavity to constraint the search, which is based on fitting points that takes into account hydrogen-bond donors and acceptors, besides hydrophobic fitting points. Regarding scoring functions, GOLD has been incorporating several ones. Among them, GoldScore[126] and Chemscore[128] are maybe the most utilized and accurate, although ChemPLP[135] compares well with them and is faster to calculate[136] (4x faster than GoldScore), which makes it suitable for virtual screening applications.

In this thesis, we took advantage of the GOLD capacity to correctly identify hydrogen bonds to use it for the docking of inorganic compounds (chapter 8). For this, we used a concept previously developed and benchmarked in our group,[84] which consists of mimicking the metal-donor interaction (i.e. coordination bond) by a hydrogen-donor interaction (Figure 2.7). A "fake" hydrogen atom is attached to the metal atom at a distance of 0.75 Å and following the coordination geometry characteristic of the metal. Also, the H_BOND_LEN parameter of the GoldScore function is modified to 2.0, to properly represent the average coordination bond length.

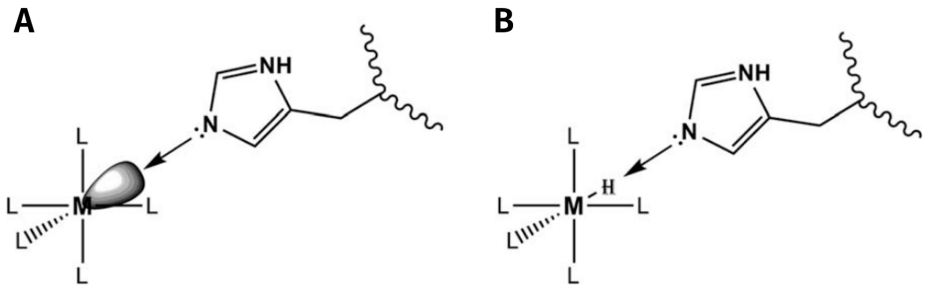



**Figure 2.7:** Metal-donor coordination bond is mimicked by a hydrogen bond interaction. **(A)** Metal-donor coordination bond. **(B)** A dummy hydrogen atom is placed at a distance of 0.75 Å from the center of the metal following its coordination geometry. GOLD and its GoldScore function will treat the interaction as a hydrogen bond. Figure reproduced from Ref.[84] with permission from John Wiley and Sons.

# B. Evolutionary algorithms in molecular modeling

As we have seen, all the problems we encounter in the structural modeling of biochemical systems share two characteristics: i) a (wide) set of solutions have to be sampled (geometrical and/or chemical spaces); ii) one or several solutions have to be selected as the most suitable for our problem. In mathematics and computer science, these are treated as optimization problems. In this section, we will first introduce the single-objective optimization problem definition, and then extend it to multi-objective optimization. After that, we will introduce a particular class of optimization algorithms (evolutionary algorithms), which are the basis of most of the developments contained in this thesis.

## 2.3 Single-objective optimization

We can define an *optimization problem* as the task of finding the best solution from the whole set of solutions. Therefore, the goal is to minimize/maximize a single *objective function*, that depends on the solution vector $\mathbf{x}$:

$$\begin{aligned} &\text{Maximize/minimize } f(\mathbf{x}); \\ &\qquad \text{Subject to } g_i(\mathbf{x}) \geqslant 0, && \text{j=1, 2, ..., J;} \\ &\qquad\qquad h_k(\mathbf{x}) = 0, && \text{k=1, 2, ..., K;} \\ &\qquad\qquad x_i^{(L)} \leqslant x_i \leqslant x_i^{(U)}, && \text{i=1, 2, ..., N.} \end{aligned} \tag{2.17}$$

The solution vector is composed of $n$ decision variables: $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$. The optimization can be done either minimizing or maximizing a single objective function $f(\mathbf{x})$, although both problems can be considered equivalent (multiplying the objective function by -1). The functions $g_j(\mathbf{x})$ and $h_k(\mathbf{x})$ define the so-called inequality and equality constraints of the problem, respectively. The last set of constraints are called the variable bounds and define the *decision space*,

restricting each decision variable $x_i$ to take a value within a lower $x_i^L$ and upper $x_i^U$ bound. Therefore, these multiple constraints serve to define the *search space* of the problem, which will be the set of all feasible solutions: those that satisfy all of the $(J + K)$ constraints and all of the $2N$ variable bounds stated above. If none of the previously defined constraints exist, the problem is an unconstrained optimization problem and the search space is infinite.

It is worth noting here that, depending on the form of the objective function and the constraints, several categories of optimization problems emerge. If all constraints and the objective function are *linear*, the optimization problem can be classified as linear. However, most real-world problems (including molecular modeling) are nonlinear, which makes the problem harder because solving techniques often do not have mathematical convergence proofs, that is, it cannot be ensured that the solution is a global optimum with a given precision.[137] Moreover, optimization problems can be classified as convex when both the objective function and the search space are *convex* or *nonconvex* otherwise. Again, nonconvexity of the objective function or the search space makes the optimization problem harder to solve.[137]

## 2.4 Multi-objective optimization

Until now, we have understood an optimization problem as the minimization (or maximization) of **one** objective function. However, in most practical decision-making problems, it becomes evident that multiple objectives or multiple criteria are needed. The formal definition of multi-objective optimization follows naturally from the definition of single-objective optimization by generalizing the single objective function to a list of objective functions:

$$\begin{aligned} \text{Maximize/minimize } & f_m(\mathbf{x}), && \text{m=1, 2, ..., M;} \\ \text{Subject to } & g_i(\mathbf{x}) \geqslant 0, && \text{j=1, 2, ..., J;} \\ & h_k(\mathbf{x}) = 0, && \text{k=1, 2, ..., K;} \\ & x_i^{(L)} \leqslant x_i \leqslant x_i^{(U)}, && \text{i=1, 2, ..., N.} \end{aligned} \tag{2.18}$$

Trying to avoid its inherent complexities, multi-objective optimization is commonly handled by grouping all objectives into one function and then treating it as a single objective problem. An example in molecular modeling are docking scores, which generally build the binding energy estimate by weighting several factors such as steric interactions, hydrophobic interactions, and hydrogen bonding.[125,128,138] The main problem with this approach arises precisely from the need to weigh, *a priori*, the different (sub)objectives that generate the global objective function.

Let's take an example from everyday life to illustrate the problem: suppose we want to buy an apartment among those available for sale in our city with an area between 90 and 110 $m^2$. All available apartments would then be our search space. To make our decision we want to take into account the following three factors: price, energy efficiency and the general state of conservation of the apartment. At first glance, we realize that each objective is in conflict with the others: it is unlikely that the cheapest apartment is at the same time the most energy efficient and the best maintained. Therefore, the optimal solution to this type of problem can no longer be a single solution, but rather a set of solutions. Optimal solutions would certainly include the cheapest, the most energy efficient and the best maintained apartments, but also those apartments that have a good compromise between the three objectives. This set of optimal solutions is called the *Pareto frontier*. Deciding on one of the Pareto frontier solutions is a matter of prioritizing which objective/s are most relevant. In this specific case, after seeing which apartments are considered optimal by the algorithm, we can decide if we absolutely prioritize one objective (e.g. price) over the others, or we opt for one of the apartments that balances the three objectives.

Going back to the docking score example, if we group the different (sub)objectives into one, we are deciding **in advance** (by assigning their weights) the priorities of the different (sub)objectives. This is the reason why a lot of research has been devoted to designing good docking scores, that is, ascertaining a set of (sub)objectives and their corresponding weights that make the binding energy estimate suitable for a general case.[139–142] However, a clear drawback arises with this approach: an imbalance in the objective weights can lead to a wrong

prediction of the best solution. Although docking programs usually provide a set of best-scored solutions (not just the best-scored one), the risk of missing the "perfect pose" is high if the system under study has particularities outside the scope of the score design.

An example in the framework of this thesis could be the docking of an inorganic compound. In addition to the interactions between the organic part of the ligand and the protein, here we look for poses with a coordination bond between the metal and a donor amino acid. In a single-objective approach, the score representing the coordination bond will be integrated with the other interactions. As a consequence, if the results do not provide any structure with a coordination bond, it will be impossible to discern whether it is geometrically impossible or the score weighting is prioritizing other interactions. However, in a multi-objective approach, you could configure one of the objectives to assess only the coordination bond. This approach will proportionate a more granular information that can be analyzed *a posteriori*, for example, to decide if the poses with better coordination score deserves prioritization over the ones with better score in the other interactions.

In the end, deciding between a single- or a multi-objective approximation algorithm is similar to the chicken-and-egg dilemma: if you save the post-analysis effort inherent to multi-objective algorithms, you should put that effort into ensuring a good set of weights to build your only objective, and *vice versa*. In this manuscript, we will often refer to this type of decision when developing new methodologies and applying them to biochemical systems.

## 2.5 Evolutionary algorithms

There exist multiple examples of single- and multi-objective optimization algorithms in the literature. Many of them have been successfully applied in molecular modeling tasks,[143] although in this case the most common approach is to stick to single-objective optimization. For example, first-order numerical methods like steepest descent and conjugate gradient are used in energy minimization and conformational analysis of molecules. Other numerical methods based on

the computation of Hessian matrices, such as the quasi-Newton algorithm BFGS, have also been employed for similar tasks.[125] However, when the problem involves multimodal objective functions (i.e., with several local minima), numerical methods become harder to apply, and other approaches such as heuristics and meta-heuristics are more suitable. Some families of heuristic and meta-heuristic methodologies are Monte Carlo and evolutionary algorithms. This last family is of special relevance for this thesis, since most of the developments are based on its foundations.

Evolutionary algorithms (EAs) are a family of heuristic search methods inspired by nature, more specifically by the main concept of Darwinian evolution: the fittest will survive. The most general workflow of an EA (Figure 2.8) comprises three steps: i) initialization of a random population of individuals; ii) generation of a children population through the application of evolutionary operators such as crossover and mutation; and iii) selection of individuals that will survive the next generation. Steps ii) and iii) are repeated cyclically until a previously established number of generations is reached or a termination criterion is met. Note here that the exploration power comes from step ii) of the algorithm and therefore evolutionary operators will play a crucial role. Note also that step iii) involves deciding somehow which individuals are the *fittest* and therefore deserve to survive; this is done by evaluating the objective function in all individuals and then applying a selection algorithm. Finally, note that the random generation of the first population, together with the inherent random nature of the evolutionary operators, confere the EAs a stochastic component that must be taken into account in their application.

Due to the EAs rapid convergence to a set of feasible solutions, they are usually applied to optimization tasks that cannot be easily solved in polynomial time, such as NP-Hard problems or problems with search spaces that would take a long time to process exhaustively. One example of such applications is molecular modeling, where different types of EAs have been successfully developed and applied.[144] Nowadays, EAs are classified into five main subfamilies: genetic algorithms (GAs), genetic programming (GP), differential evolution, evolution strategy, and evolutionary programming.[145] In the following paragraphs we will

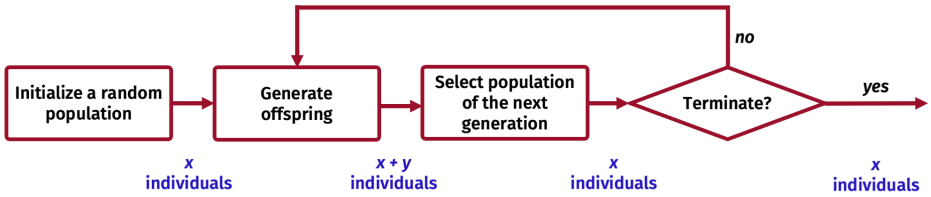



**Figure 2.8:** General workflow of an evolutionary algorithm. The steps of the algorithm are shown inside dark red boxes. The population size at the end of each step is indicated in blue, assuming that the selection algorithm keeps it constant.

introduce the first two families, which are the methodological basis for most of the developments contained in this thesis (chapter 4 and chapter 5). It is not our aim to give an exhaustive description of the methods, which can be found in excellent textbooks such as K. Deb for multi-objective GAs[137] and J. Koza for GP.[146,147] Instead, we will focus on a comprehensive overview of their key concepts and the issues that affect their application to molecular modeling tasks.

## Genetic algorithms

The main peculiarity of genetic algorithms (GAs) with respect to the general definition of EA is that individuals are encoded in a linear sequence of elements of a fixed length, called the *chromosome* of the individual. A standard codification used for chromosomes is the binary string, which consists in an array of bits (0s and 1s) that uniquely represents the individual. Another widely used encoding is a float array, which offers more usability when the definition of the individual needs a large amount of real numbers.

One direct consequence of this chromosomal encoding is that the objective function for evaluating solutions should "understand" it. Therefore, either the objective function is adapted to deal directly with the chromosomes or a conversion of the chromosome to the domain of the objective function is needed. The latter is called *expressing* the individual. For example, in the case of a small molecule, a possible encoding for its conformation could be an array of float numbers with as many elements as rotatable bonds the molecule has. The expression of the chromosome will consist of applying the torsion angles contained in the float

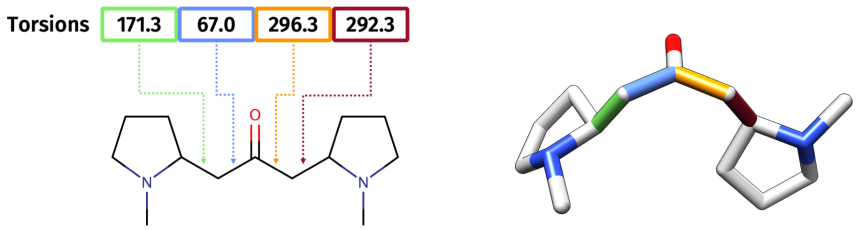



**Figure 2.9:** Encoding and expression of a small molecule conformation. The encoding (left) is done with a list of floats, each one representing the torsion angle of a rotatable bond of the molecule. The expression (right) is done by applying the torsion angles to the three dimensional structure of the molecule.

array to an initial conformation of the molecule (Figure 2.9). The result of the expression will be the 3D position of all the atoms in the molecule, which is then passed to the objective function to evaluate its fitness (e.g. energy).

Another implication of the particular encoding in GAs is the need for specific evolutionary operators to handle the exploration of the search space. It is of paramount importance that these operators allow us to generate solutions in the search space that are both diverse and, at the same time, better (i.e. with higher fitness) on average than randomly generated ones. The former will allow us to explore zones of the search space that have not yet been explored, avoiding being trapped in a local minimum when the objective function is multimodal. The latter will ensure faster convergence towards a set of optimal solutions. Two evolutionary operators are usually defined in GAs: crossover and mutation.

The philosophy of the crossover operator in GAs is analogous to mating during sexual reproduction in biology: combining the genetic information of two parents to generate a child. The child will share some characteristics with one parent and the rest with the other parent. Since the selection pressure of the GA will favor those characteristics that lead to better fitness, the crossover operator will increase the probabilities of convergence towards optimal solutions once the "winning" genetic characteristics are present in the population. Examples of traditional crossover operators are point and uniform crossovers (Figure 2.10). In the case of chromosomes formed by a float array, the simulated binary crossover

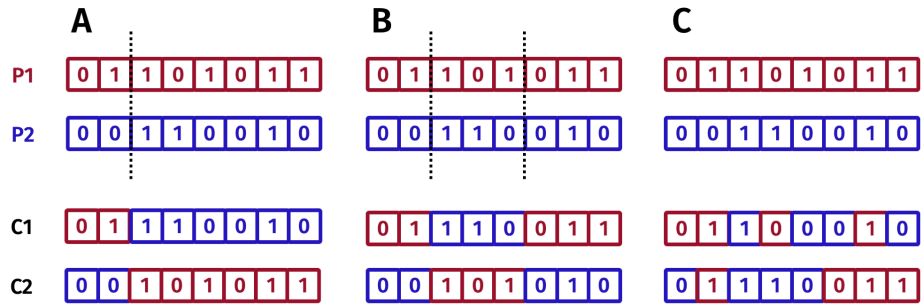



**Figure 2.10:** Examples of crossover operators. First parent chromosomes (P1) are shown in dark red, whereas second parent chromosomes (P2) are shown in blue. Children are denominated C1 and C2. **(A)** One-point crossover. A random crossover point is picked (in dashed line). Bits to the right of the crossover point are swapped between P1 and P2, generating two children. **(B)** Two-point crossover. Two crossover points are randomly chosen (in dashed lines). The bits in between the two crossover points are swapped, generating two children. **(C)** The two children are generated by choosing each bit from either parent with equal probability.

has demonstrated good performance in several optimization problems.[148,149]

On the other hand, the main goal of the mutation operator, analogous to mutations in biology, is to maintain diversity in the pool of chromosomes. Replacing one position of the chromosome with a randomly generated value (bit, float, or the corresponding type) is the traditional approach, although more specific mutation operators are used in some cases, such as polynomial mutation in the case of float-array chromosomes.[150] Typically, the mutation operator is applied with a certain probability on the offspring individuals generated by crossover. What is the advisable probability of a mutation operator in GAs has been the subject of extensive debate in the community and remains an open question.[151] The great diversity of optimization problems, together with the different evolutionary operators available, make GAs difficult to parameterize and sometimes it must be done in a personalized way for the problem at hand.

Finally, to complete the picture of the GA workflow, it is essential to introduce the basic concepts of selection, which is in charge of applying evolutionary pressure to guide the population towards an optimal set of solutions. Selection algorithms are introduced in two points of the GA workflow: before the crossover, to select the individuals that will participate in the offspring generation; and

after the evaluation of the objective function, to select the individuals that will survive to the next generation.

In the first case, several strategies can be applied, from the purely random selection of parents to the selection of a portion of the individuals with the highest fitness score as the only ones allowed to have offspring, which is called selection by truncation. However, a compromise is usually sought between prioritizing the fittest individuals (with supposedly better characteristics) while maintaining the possibility that the least fit individuals have at least a small chance of having descendance. In this sense, two approaches can be used: tournament selection, repeatedly selecting the best individual from a randomly chosen subset, and roulette-wheel selection, where individuals are randomly chosen with a probability proportional to their fitness score.

Note here that an ordering criterion is needed to prioritize the individuals in those approaches that are not purely random. Whereas in single-objective optimization problems it becomes natural to directly use the fitness score (at most normalizing it for the case of a roulette-wheel selection), for the case of multi-objective optimization things are a bit more complicated. An approach in the latter case could be to choose one of the objectives and prioritize those individuals with the best scores on that objective. However, it would be contrary to the notion that all objectives should be of equal importance. This is when the concepts of *dominance*[a] and *rank* arise (Figure 2.11). If one individual dominates another, it is considered better. Depending on the dominance between individuals, a rank can be assigned to each of them: the individuals that are non-dominated will constitute the first rank, the individuals that are dominated only by those of the first rank will constitute the second rank, and so on. The rank of an individual will therefore be the metric for the selection operators. In the event of a tie in rank, other factors such as diversity may be taken into account.

[a] A solution $\mathbf{x}^{(1)}$ is said to dominate another solution $\mathbf{x}^{(2)}$, if both conditions 1 and 2 are true:
1. The solution $\mathbf{x}^{(1)}$ is no worse than $\mathbf{x}^{(2)}$ in all objectives: $f_j(\mathbf{x}^{(1)}) \trianglerighteq f_j(\mathbf{x}^{(2)})$ for all $j = 1, 2, \ldots, M$.
2. The solution $\mathbf{x}^{(1)}$ is strictly better than $\mathbf{x}^{(2)}$ in at least one objective: $f_{\hat{j}}(\mathbf{x}^{(1)}) \triangleleft f_{\hat{j}}(\mathbf{x}^{(2)})$ for at least one $\hat{j} \in \{1, 2, \ldots, M\}$.

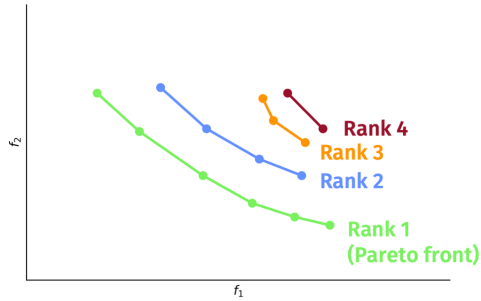



**Figure 2.11:** Pool of solutions ordered by rank in a GA with two objective functions, $f_1$ and $f_2$. Both functions are supposed to be minimized. Therefore, the best pool of solutions are the ones shown in green, which are denominated the "Pareto front" or first rank. The rest of the solutions are ordered by rank, following the rules of dominance.

In the case of selecting surviving individuals at the end of a generation, there are also several strategies. One option would be to consider only those individuals belonging to the offspring as candidates to survival. It would imply that a part of the parent population (at least the fittest individuals) should be cloned in the offspring. Otherwise, there is a risk of losing good individuals along the way. Another option to handle this issue is to select the survivors from the full set of parents and offspring, which is the approach followed in the developments of this thesis. Therefore, the selection when there is only one objective function will be as easy as ordering the individuals by fitness, and then selecting as many as the size of the population (which remains constant throughout the optimization). In the case of multi-objective optimization, more specialized algorithms are needed. That is the case of the non-dominated sorting genetic algorithm (NSGA), whose versions II and III are good options to maintain a balance between well-scored and diverse individuals.[152–154]

On the whole, in this subsection we have seen how the peculiar chromosomal encoding used in GAs has profound effects on the different operators used throughout the optimization. As we will see along the different chapters of this manuscript, one of the difficulties in applying this workflow in the structural modeling of biochemical systems is the proper choice and parametrization of all

these steps.

**GaudiMM** As previously mentioned in the introductory chapter, the InsiliChem group developed GaudiMM (Genetic Algorithms with Unrestricted Descriptors for Intuitive Molecular Modeling) as a general framework that could be used in molecular modeling tasks.[91,92] The platform was designed with three goals in mind: i) a code architecture as modular as possible; ii) clear distinction between the different stages of the optimization process (i.e. exploration, evaluation and selection); and iii) a native support for multi-objective evaluation. The philosophy behind GaudiMM is that the user can design a molecular modeling experiment by selecting the suitable explorative and evaluative modules for their problem, which is called the "recipe". This approach is particularly powerful when the research project involves hypothesis-driven modeling, that is, we can establish some constraints *a priori* and guiding obtained from experiments or from informed knowledge. As the code is written in Python and is open-source licensed, there is also the possibility to tweak existing modules or even create new ones.

The workflow of a GaudiMM calculation (Figure 2.12) follows the general postulates of a multi-objective GA optimization process. However, some relevant changes were introduced to adapt the algorithm to molecular modeling tasks and the modular philosophy of the platform. First, GaudiMM implements a multi-gene approach: each explorative module has its own gene definition and therefore a specific type of chromosome (Table 2.1). For example, the "Torsion" gene, in charge of varying the rotatable bonds of a molecule, has a chromosome composed by a list of float numbers representing the different torsion angles, while the "Search" gene, in charge of translating and rotating a molecule, has the transformation matrix as a chromosome. As a direct consequence of the multi-gene schema, each gene accounts with its specific crossover and mutation operators, which are sequentially executed in the offspring generation stage of the GA. Finally, an important effort was dedicated to construct various geometric objectives and adapt well-known evaluative functions (Table 2.2) to ensure high versatility in the optimization problems that GaudiMM can handle.

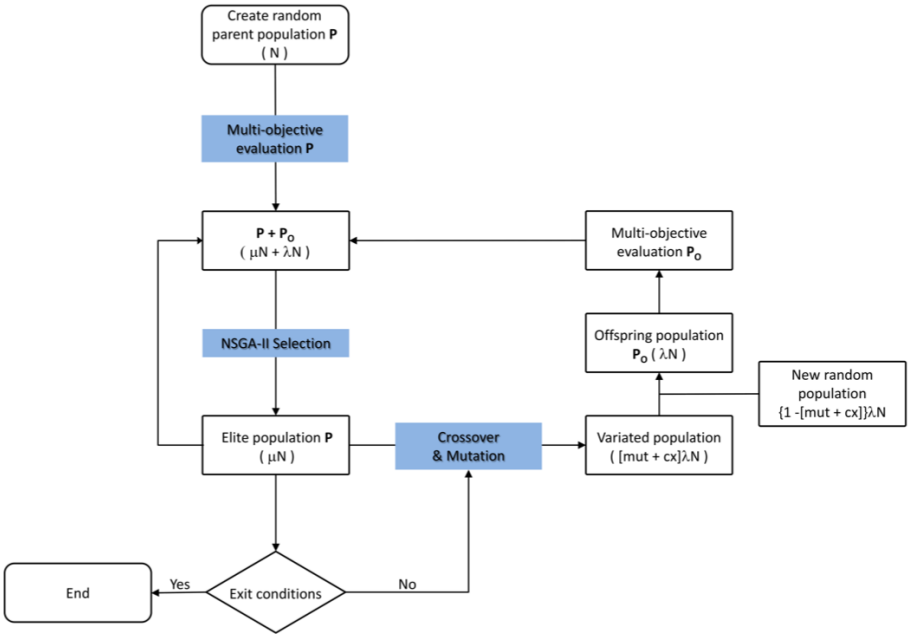



**Figure 2.12:** Workflow of a GaudiMM calculation. $N$ is the number of individuals in the initial population **P**. $\mu$ is related to the number of individuals selected for the following generation. $\lambda$ is related to the number of individuals produced at each generation as offspring (population $\mathbf{P}_0$). The parameters *mut* and *cx* are the probabilities associated with mutation and crossover operators, respectively. Figure adapted from Ref.[91] with permission from John Wiley and Sons.

| Name | Description | Chromosome type |
|---|---|---|
| Molecule | Load and build structures | Integer |
| Rotamers | Explore side chain flexibility | List of floats |
| Mutamers | Explore mutation of residues | Tuple (string, float) |
| NormalModes | Explore collective motions | Integer |
| Search | Translation and rotation of Molecules | Transformation matrix |
| Torsion | Dihedral rotation of bonds | List of floats |

**Table 2.1:** List of genes available in the original version of GaudiMM. Table adapted from Ref.[91] with permission from John Wiley and Sons.

| Name | Description |
|---|---|
| Angle | Optimize angle of three atoms, or dihedral of four atoms |
| Contacts | Minimize steric clashes, maximize hydrophobic interactions |
| Coordination | Optimize coordination geometry of metal center |
| Distance | Optimize distance between two or more atoms |
| DSX | Docking scoring function |
| Energy | Minimize molecular mechanics potential energy |
| HBonds | Detect hydrogen bonds |
| Inertia | Align axes of inertia of two or more molecules |
| LigScore | Docking scoring function |
| Solvation | Measure solvent accessible solvent area |
| Volume | Measure volume ocupied by molecule |

**Table 2.2:** List of objectives available in the original version of GaudiMM. Table adapted from Ref.[91] with permission from John Wiley and Sons.

The original version of GaudiMM has been successfully applied in several molecular modeling projects, which cannot easily be addressed by mainstream modeling approaches, in particular, nonstandard dockings.[155–158] Its code has also been extended and adapted to the problem of a metallic moiety docking into a protein,[85] which has also been applied in several works, involving the study of metallodrugs and metalloenzymes.[159–162] In this latter case, the approach is focused on the coordination geometries characteristic of each metal. A fingerprint of the ideal geometry of the metal is used as a descriptor to assess the "coordination" score of a specific pose. Therefore, the exploration will be guided to those poses with good coordination geometries. Note here that the interaction of the rest of the inorganic compound with the protein is not taken into account in this approach. In GaudiMM's philosophy, this would mean that we'll need an additional objective to assess these interactions. Also, it is worth noting that due to the less constrained exploration of GaudiMM (compared to the GOLD approach explained in section A), sometimes the resulting geometries might not be good enough. In particular, when we introduce a large set of side chains that are allowed to rotate, it will be worthwhile to refine the resulting poses with an additional set of more constrained calculations, which we generally do using the

GOLD method.

Despite GaudiMM's success, some challenges remained. First, to expand the exploratory capabilities of the software. While conformational sampling of biological and chemical molecules was sufficiently covered in the first version (via the "Torsion", "NormalModes" and "Rotamers" genes), other interesting and cutting-edge modeling problems, such as chemical space exploration and discovery of ligand binding routes, were impossible to tackle. Secondly, there is still room to push the limits of GaudiMM application in research, especially in those cases that involve non-standard dockings, which are hard or forbidden territory for other state-of-the-art software. Finally, the first version of GaudiMM was thought of and developed as a proof of concept of the multi-objective GA framework. As a consequence, it was not designed with performance in mind and several technical aspects deserve a revision to improve its execution and allow new extensions.

## Genetic programming

Another interesting variant of evolutionary algorithms is genetic programming (GP). The concept of GP was invented by N. Cramer[163] in 1985 and further developed by J. R. Koza in 1992,[146] with the idea of automatically designing computer programs through evolution. Its fundamental methodological novelty is the introduction of a parse tree structure to encode the programs, in contrast to the fixed-length linear chromosome of traditional GAs. The tree nodes are operator functions, while each terminal node is an operand. This flexible structure soon proved its usefulness for representing mathematical expressions and computer programs in LISP (Figure 2.13). Of course, the introduction of the nonlinear tree representation made it necessary to use specific genetic operators. Crossover and mutation operators are commonly used to manipulate trees, but different implementations exist depending on the optimization problem.[164] Other genetic operators such as permutation, transposition or inversion are also employed in some applications.

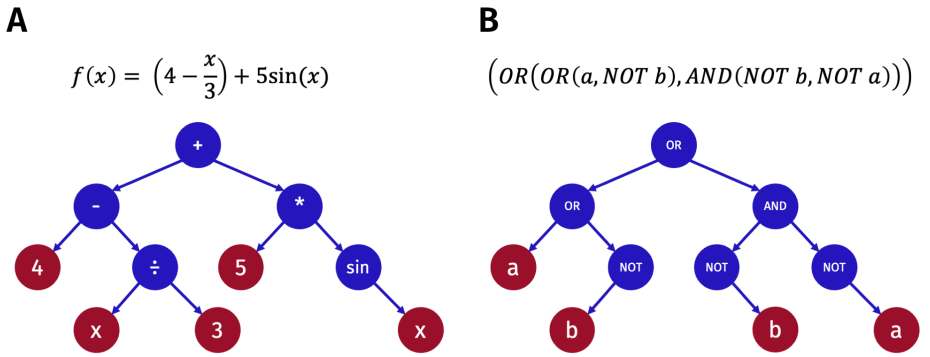



**Figure 2.13:** Examples of tree encoding used in genetic programming. **(A)** Tree representation of a mathematical function. **(B)** Tree representation of a LISP program.

GP has been successfully applied to many optimization problems, including engineering fields as diverse as the design of analog electrical circuits[165] and the fiber-to-yarn process in the textile industry.[166] In the end, the idea of parse tree encoding can be easily extrapolated to any problem that can be expressed as a group of operators and operands, especially in those cases where there is no ideal solution but a set of "sufficiently good" ones. However, over the years and as usual, the experience revealed some drawbacks with the use of GPs. The most relevant would be: i) its tendency to generate increasingly larger trees throughout the evolutionary process; ii) the difficulty of finding genetic operators that generate a high proportion of valid offspring; and iii) the evolution is very constrained to the shapes of the trees generated in the initial random population. In order to address some of these issues, C. Ferreira evolved the concept to what she named gene expression programming (GEP),[167] recovering a linear representation of chromosomes with fixed length (but not only numerical as in traditional GAs), and an expression operator to convert the chromosomes to parse trees.

Although the vast majority of reported GP and GEP applications implement them in single-objective optimization problems, nothing prevents the use of such tree-inspired non-numerical representation with multiple objective functions. This idea, together with a set of operators and operands capable of build-

ing molecules, is the origin of the new module to explore the chemical space within the GaudiMM framework, developed in chapter 5 of this manuscript.

## 2.6 Development framework

To end this methodological chapter, in this section we will introduce the more technical aspects regarding the development framework over which the different pieces of software of this thesis were built.

One of the goals of the InsiliChem group is to provide the community with programs as accessible as possible. This involves the licensing under permissive open-source licenses (Apache v.2[168] or BSD-3-Clause[169]), but also a readable code architecture to allow other researchers to improve or even adapt the code to their particular needs. To achieve this, we make use of the Python[170] programming language, together with a variety of well-known packages. In addition, the whole code of the different programs is uploaded to the group's GitHub (https://github.com/insilichem), and each program accounts with its corresponding documentation. Finally, the programs developed in this thesis have been deployed in widely used package repositories, such as PyPI[171] and/or conda.[172]

### Python: the gold-standard programming language in science

Python is an open-source, object-oriented, and interpreted programming language, which is licensed under the permissive Python Software Foundation License. The first version of Python appeared in 1991, designed by Guido van Rossum as a successor to the ABC programming language. Two major code revisions have been released: Python 2, released in 2000 and discontinued in 2020, introduced new features such as list comprehensions and Unicode support. Python 3, released in 2008, is the current active version of the language and it is not fully backward compatible. This last aspect is of major importance for this thesis, since the original GaudiMM code was developed in Python 2.7

due to the restrictions imposed by some code dependencies (in particular, the UCSF Chimera code[22,173]). The inevitable obsolescence of the code and several incompatibilities that arose with other dependencies that switched to Python 3, forced us to update the entire GaudiMM framework, as will be explained in chapter 5.

Python's high-level, general-purpose design, coupled with its focus on code readability, made it so popular that it became one of the most widely used programming languages in the world.[174] In the field of scientific software, Python has occupied the top position among the most popular languages in recent years.[175,176] It is used in a wide range of development projects: from relatively simple tasks like data analysis or file parsing and conversion, to the creation of complete software packages.

Besides the standard libraries contained in Python, there exist thousands of external libraries and packages developed and maintained by the community and licensed under different conditions. In this thesis, the most relevant packages that have been integrated as dependencies in the developments are: PyChimera[173] to integrate the chemical functions of UCSF Chimera[22] in GPathFinder,[177] the DEAP package[178] to implement the multi-objective genetic algorithm in GPathFinder,[177] RDKit to integrate all its chemical functions in the new version of GaudiMM and GAlkemist, and an extensive use of the high-performance matrix operations provided by NumPy[179] in all the developments.[177,180] A comprehensive list of all the dependencies can be found at the GitHub page of each program.

# 3

# Objectives

Molecular modeling is a fast-growing field. In the last decades, increasing computational capabilities accessible at reasonable cost, coupled with a greater number of qualified scientists, have made it possible to incorporate molecular modeling software into cutting-edge research.

However, a systematic sampling of all the spaces necessary for the exploration of biologically significant systems is far from being a state-of-the-art capability. In consequence, we need to rely on constrained and clever exploration algorithms. Among the main challenges to address, the two following are behind the motivation of this thesis: first, the expansion of the explorative limits of current software and protocols; and second, the incorporation of metallic species in MM-based methodologies, since they are often underrepresented if we compare it with their biological importance.

These two general axes concretize in the following objectives:

- To evolve GaudiMM's capabilities for the coupled exploration of biochemical spaces, developing new modules for the identification of ligand binding pathways and for chemical space exploration.
- To develop a novel computational method to predict metal-binding sites in proteins.
- To apply GaudiMM in the framework of research projects requiring non-standard docking approaches.
- To optimize a multi-scale modeling workflow for the understanding of interactions between a metallodrug and its target, and apply it to the specific case of the interactions between oxaliplatin and insulin.
- To optimize a multi-scale modeling workflow for the understanding of the mechanisms of a metalloenzyme, and apply it to the specific case of a dirhodium cyclopropanase.

# 4

# Identification of ligand binding pathways with GPathFinder

The traditional approach to study the recognition mechanism between proteins and small molecules is to look at the interactions in the binding site of the protein. Sometimes, the binding occurs at the surface of the protein, in relatively solvent-exposed zones. In these cases, the study of the local interactions at the binding site could be enough to ascertain the protein-ligand recognition mechanism. However, in a general case, we can expect that the binding happens at a much buried place of the protein, favoring hydrophobic interactions between the ligand and the protein. In these cases, the study of the complete binding route or pathway becomes unavoidable to have a complete picture of the key determinants of binding affinity and selectivity.

This concept can be concretized in the "tunnel and gate" model. We define a *tunnel* or *pathway* as a transport route from one point of the protein to another. The pathway might have a functional role controlling the specificity of the protein. For example, the width of the tunnel at its narrowest point will determine the maximum size of a ligand able to bind the protein, or the chemical properties of the amino acids along the tunnel will improve/worsen the affinity towards ligands containing certain chemical groups. In the case of *gates*, their role is more dynamic, because they involve a mechanism of open-close in some part of the protein (typically inside a tunnel) that can control the access to the binding site or other parts of the protein. Altogether, the tunnel and gate model highlights the importance of studying the complete binding process when the binding site is buried in the core of the protein. In particular, it can be key in our understanding of how a reaction is produced in an enzyme and, in drug discovery protocols, to see if a drug can have chances to achieve the binding site and therefore inhibit/activate the function of a protein.[181]

Ligand binding pathways can be sometimes identified by experimental methods, such as the case of the human indoleamine 2,3-dioxygenase 1 (hIDO1) in complex with a suicide inhibitor, reported in PDB entries 6dpq, 6dpr, and 6mq6. In this work,[182] they reported three crystallographic structures showing the inhibitor in different places of the protein, allowing to identify the binding pathway (Figure 4.1). However, the most (by far) usual case is to find in the PDB a single structure of the biochemical complex with the ligand at the binding site, because it is, in principle, the most stable configuration. Therefore, computational methods are of great importance to study ligand binding pathways.

One option to tackle the problem is to use enhanced sampling techniques in MD simulations. Methodologies like steered MD, metadynamics, Gaussian accelerated MD, and the recent ligand Gaussian accelerated MD could be very valuable.[49,183,184] However, here we are interested in faster approaches that can work on a single workstation, while maintaining enough accuracy. In this sense, state-of-the-art methods can be classified into three families: i) tunnel searching in the protein structure without taking into account the ligand explicitly;[185–191] ii) studying the protein-ligand interactions along a previously identified tun-

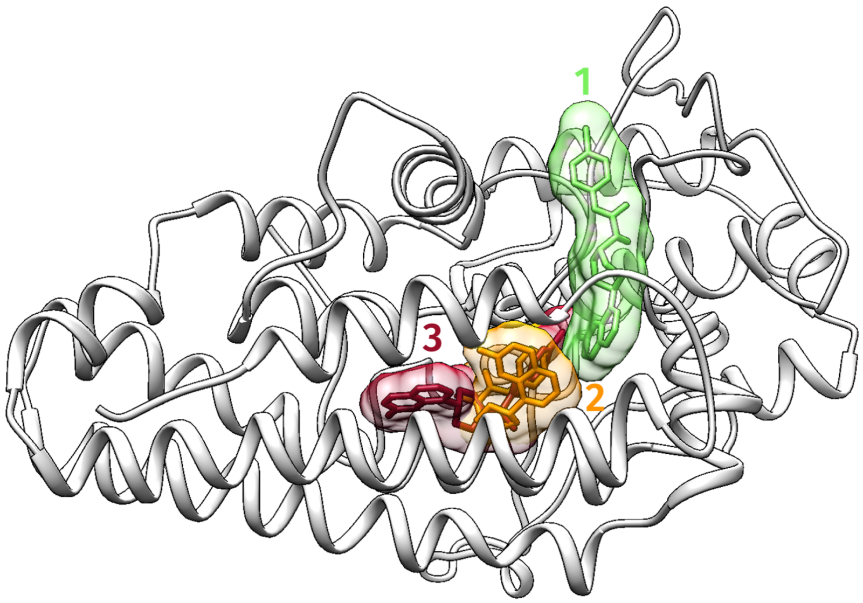



**Figure 4.1:** Schematic representation of the binding process of a suicide inhibitor to 2,3-dioxygenase 1. Three consecutive positions of the inhibitor are shown in green (step 1), orange (step 2) and dark red (step 3). For clarity, only one conformation of the protein is depicted. Structures were obtained from the PDB (entries 6dpq, 6dpr, and 6mq6).

nel, by taking into account the flexibility of the protein with different degrees depending on the method;[192–194] and iii) identifying all possible tunnels while taking explicitly into account the ligand and the flexibility of the protein.[195–198] Whatever the method, their accuracy depends on the quality of the sampling and the energetic evaluation, which can range from very simple geometric functions like steric clashes to standard MM force fields.

In our case, we propose a novel tool ubicated in the third family. The original idea was born in the group and developed in the final master thesis of P. Orenes, where the protein-ligand affinity was evaluated by placing the ligand in a grid of points covering the whole volume of the protein and then tracing the best pathway (that with best affinity) from the solvent to the binding site. We quickly realized that the optimization of these pathways was a suitable task for

the multi-objective genetic algorithm of GaudiMM.[91] In fact, the implementation under the GaudiMM framework offered us a much wider set of possibilities both in the sampling and the evaluation of the protein-ligand poses along the pathway. The development was concretized in a new "path" gene integrated in a customized GaudiMM environment that we deployed separately under the name of GPathFinder[177] (https://github.com/insilichem/gpathfinder). Regarding the sampling, GPathFinder combines the exploration of the ligand conformational flexibility with local (side chains around each ligand position) and global (backbone movements) degrees of flexibility for the protein. To evaluate the binding pathways, we took advantage of the multi-objective capabilities of GaudiMM, which allow to combine geometric criteria like steric clashes and smoothness of the pathway with a fast energetic evaluation through a docking scoring function (Autodock Vina[125] or smina[199]).

One of the most time-consuming steps in the development of the project was to properly configure and benchmark the algorithm. First, several parameters of the genetic algorithm, such as number of generations and population size, were fine-tuned to achieve a good balance between accuracy and speed (a typical GPathFinder calculation can be run on a desktop computer in a few hours). Second, we parameterized and balanced the proportion of the custom crossover and mutation operators that we built for the path gene. Finally, we successfully benchmarked GPathFinder on 20 biochemical systems whose binding pathways had already been reported in the literature, representing the broadest benchmark done so far in pathway determination software.

We also demonstrated the usefulness of GPathFinder in three illustrative cases where we did a more detailed analysis and compared the results with those already available in the literature: i) transport of glycerol across three different families of aquaporins, showing that the precision of the method was sufficient to differentiate those aquaporins that can transport glycerol. ii) Binding process of a suicide inhibitor to hIDO1, where we found the same route as in the avaliable crystallographic structures (mentioned above). This case study allowed us to identify a limitation of the method: a big conformational change in a loop of the binding site entrance prevented the algorithm from finding the exact mecha-

nism observed in the crystallographic structures. And iii) identification of binding routes used by the 0XV ligand[a] to access the binding site of the human cytochrome P450 2C19, starting from the crystallographic structure with entry 4gqs.[200]

## 4.1 Application of GPathFinder

In addition to the three show-cases of the GPathFinder article, we also applied a beta-version of the program to the discovery of an exo-hydrolase processive catalytic mechanism.[201] Glycoside hydrolases, such as the *Hordeum* exo-hydrolase HvExoI studied here, are enzymes which catalyze the hydrolysis of oligo- and polysaccharides, a process that is essential to understanding the global carbon cycle. Previous structural work on HvExoI showed that the glucose (Glc) product released from the hydrolysis reaction remains trapped in the active site of the enzyme until an incoming substrate arrives. However, an open question in those studies was to determine how Glc is displaced from the active site to allow the new catalytic cycle.

In this work, several experimental and computational techniques were combined to examine the whole catalytic cycle of HvExoI, ultimately proposing a "substrate-product assisted processive catalysis" (Figure 4.2). First, the refinement of the native HvExoI structure (1,65 Å resolution) revealed that Glc is bound at 0.5 occupancy at each -1 and +1 subsites, suggesting that it may be mobile between both subsites (Figure 4.2, step 2). The strength and the conformational states of the HvExoI-Glc complex were assessed by surface plasmon resonance (SPR) analysis, NMR spectroscopy and QM/MM metadynamics simulations, and compared with a recombinant HvExoI produced in *Pichia pastoris*. The next step was to investigate the molecular basis of the Glc displacement after the incoming of a new substrate (Figure 4.2, steps 3 and 4). Six substrates with potential to displace Glc were experimentally tested: two deoxy-Glc derivatives, two alkyl-glucoside derivatives, a hydrophilic polymer polyethylene

[a] 0XV: (4-hydroxy-3,5-dimethylphenyl)(2-methyl-1-benzofuran-3-yl)methanone

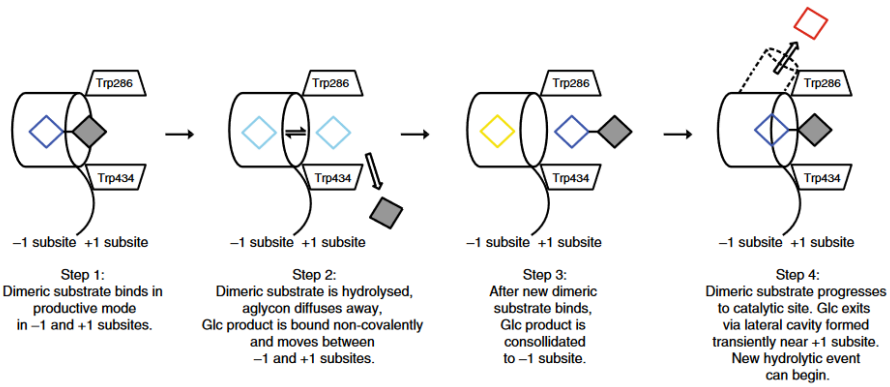



**Figure 4.2:** Scheme of the processive catalytic mechanism. The dimeric substrate is represented in empty blue and filled gray squares. The Glc product is represented by an empty cyan, yellow or red square for step 2, step 3 and step 4, respectively. Figure reproduced from Ref.[201]

glycol (PEG), and the thioanalogue methyl 2-thio-β-sophoroside (G2SG-OMe). Whereas deoxy-Glc derivatives could not remove the Glc from the binding site, the alkyl-glucosides and PEG were able to do so. Interestingly, G2SG-OMe was observed to bind to HvExoI in a markedly different position, which was considered an intermediate structure (PDB entry 6md6) and therefore we selected it as one of the starting structures in the molecular modeling experiment that we carried out to finally assess the unbinding route of Glc.

It was at this stage of the study when we employed a beta-version of GPathFinder to simulate the unbinding route of the Glc. We started the simulation with the Glc at the -1 subsite, while the incoming substrate (G2OG[b] or G3OG[c]) was ubicated at the entrance of the binding site (Figure 4.2, step 3). We tested four different scenarios using structures previously obtained with a combination of docking, MD simulations, and normal mode analysis (NMA): i) ternary HvExoI:Glc:G2OG complex generated by docking on the crystallographic structure (PDB entry 6md6); ii) ternary HvExoI:Glc:G2OG complex generated by docking on PDB entry 6md6 and applying NMA to consider the global protein

[b]G2OG: β-D-glucopyranosyl-(1,2)-D-glucose
[c]G3OG: β-D-glucopyranosyl-(1,3)-D-glucose

flexibility; iii) ternary HvExoI:Glc:G3OG complex derived from a MD snapshot; and iv) ternary HvExoI:Glc:G3OG complex derived from a MD snapshot and applying NMA to consider the global protein flexibility.

Only steric clashes evaluation was available at that development version of GPathFinder. The three cases accounting with a pre-sampling of the global flexibility of the protein (through MD and/or NMA) resulted in feasible escaping routes for the Glc. However, in case i), where only the crystallographic structure was used, the high steric clashes prevented the Glc exit. In all three successful cases, the unbinding pathway was roughly the same, through an adjacent lateral cavity (Figure 4.2, step 4) formed by a cork-like motion of domains 1 and 2 of the protein. This pathway was further assessed in higher detail with the PELE software[196] and by the design of a HvExoI variant to experimentally observe the role of two key residues.

Altogether, this experience with the use of GPathFinder in a real research scenario allowed us to realize the importance of incorporating protein motions in these calculations. In the final version of the program, we addressed this challenge with an automated sampling of NMA-based structures and allowing the user to incorporate a pool of snapshots from a MD simulation that are automatically sampled in the pathway generation.

## 4.2 Chapter conclusions and future work

In this chapter we have seen how we developed and successfully applied a novel tool for the simulation of ligand (un)binding processes in proteins, which provides multi-objective evaluation of the generated pathways through a combination of geometric criteria and a docking-based score. Both works (development of GPathFinder and application on the exo-hydrolase case) were published in the form of scientific articles.[177,201] Future expansions of this work involve dealing with limitations regarding structure rearrangements of the proteins (such as the one observed in the case study ii), the design of a user-friendly interface to configure and analyze calculations, and including the evaluation of metallic species.

# 5

# Hypothesis driven exploration of the chemical space with GAlkemist

From ancient alchemists to modern chemical industries, humans have always been interested in processes to transform matter. Although scientists have put a lot of effort into discovering and designing new molecules, the chemical space is far from being fully explored. For example, the ChEMBL database[202] accounts today (August 2022) with 2.3 million biologically active compounds, while the drug-likeness molecular space is estimated to be in the order of $10^{60}$ molecules.[25] Exploring those spaces through brute-force approaches is infeasi-

ble, even with the most powerful supercomputers. As a consequence, the efficient exploration of the chemical space by computational means has become a vivid field of research.[37]

Many tools exist for the generation and optimization of molecules, which often make intensive use of evolutionary algorithms[203–207,39,208–211] or deep learning[212–219,44,220–225] to guide the exploration towards the desired fraction of the chemical space. However, the ratio of success drops when applying these programs in real research scenarios, for example, in drug discovery pipelines. Several reasons could explain the fact that only a small number of drugs have been designed with the aid of a molecule generator. For sure, the novelty of the methods (17 out of the 25 programs cited before have less than three years of life) plays a crucial role, as drug research projects usually involve longer time scales. But also other factors have been identified. In an excellent review on the subject, J. Meyers and coworkers raised several challenges: the difficulty of designing the objective function to guide the search, the need for better benchmarks nearer to the reality of *in vitro* validation, the lack of the 3D environment in most of the current methods, and the difficulty to interpret some generative models that are based on black-box technologies like deep learning. Altogether, the main challenge can be summarized in their phrase: "generative methods should be flexible in their usage such that they can complement routine design strategies in medicinal chemistry".

If we briefly analyze the drug discovery process, we realize that, in fact, the design of a new molecule does not start from scratch. There is usually a motivation or a purpose in mind to start the project: a disease to be cured. Other information, such as the molecular targets associated with the disease, previous drugs that have been partially successful, and other diseases with similar mechanisms, can complete the framework of the project and focus the search on a specific molecule profile. That was the case of the recent discovery of nirmatrelvir, one of the active compounds composing the Paxlovid medicine for the treatment of COVID-19. In this project,[226] the researchers started from one compound that was identified as a potent inhibitor of the SARS-CoV-1 main protease ($M^{pro}$). They submitted the structure of the original compound to several rounds of re-

finement with two goals in mind: to improve the pharmacokinetic properties of the drug and to better fit the SARS-CoV-2 $M^{pro}$ binding site. After testing several hypotheses of chemical groups that could produce the desired effects, they finally obtained the nirmatrelvir drug, which was the compound with the best compromise between all the metrics analyzed.

Translating the essence of the discovery process to the language of optimization algorithms, we identify several requirements. First, a multi-objective approach capable of accounting with a diversity of descriptors mainly related to the structure of the molecule and its interactions with the biological target. Note here that these descriptors will often provide contradictory results, because an improvement in one objective can produce a worsening in another. Second, if we envision a tool that could assist along a human-in-the-loop pipeline, the model should be understandable, allowing the researcher to intervene and adjust the parameters when needed. And finally, the integration of 3D conformational exploration can be key, as was in the case of nirmatrelvir mentioned before.

We felt that these requirements for a *hypothesis driven* generation of molecules were close to what the GaudiMM platform[91] offered in its original version in 2017: a multi-objective evaluation capable of tackling the simultaneous optimization of several descriptors, and simple 3D descriptors to allow the researcher to generate and test conformational hypotheses. We therefore decided to complete and adapt the tool to the problem of molecule generation. The main piece, which we report in this chapter, is GAlkemist, a module ("gene", in the GaudiMM nomenclature) to allow the exploration of the chemical space, that is, generating molecules. Also, we incorporated several descriptors ("objectives") related to the structural properties of small molecules.

## 5.1 Computational methodology

Truth to be said, what seemed like a fairly straightforward implementation of a new module within the GaudiMM environment quickly became complicated. A cascade of technical issues arose that ultimately lead to the impossibility of GAlkemist integration. As a context to understand the issues, we need to know

that the original version of GaudiMM heavily depends on UCSF Chimera[22,173] code to implement the majority of its molecular functionalities. As a 3D molecular visualization program, UCSF Chimera code is mainly focused on functionalities around the conformation of molecules, but has more limited chemical capabilities. As we needed a more specialized library to implement the chemical sampling and descriptors, we decided to use the well-established RDKit library.[227] Although difficult to manage, the interrelationship between the two frameworks was possible with the use of interconversion functions.

However, with the discontinuation of Python 2.7 in 2020, the developers of both libraries made different decisions: whereas RDKit switched all its code to Python 3, UCSF Chimera decided to focus on the new version of their visualizer (ChimeraX[228]), and left Chimera code in Python 2.7 without further maintenance. If we wanted to continue with the GAlkemist development, these decisions basically left us with two options: adapt the code of GaudiMM either to the new ChimeraX or the RDKit library. We opted for the last option, which involved the entire rewriting of the code and therefore supposed a significant delay in the GAlkemist development. Currently GaudiMM v.2 is in the last stage of development, in which almost all of the members of the InsiliChem group are involved.

## GaudiMM v.2

We planned the development of GaudiMM v.2 with three technical goals in mind: i) reduce code dependencies to avoid conflicts; ii) rely on only one molecular library (RDKit) for the molecular objects; and iii) improve the performance of the code. Also, we took the opportunity to include other scientific improvements. First, we have incorporated the conformational sampling of small molecules through knowledge-based conformer generators[63,64] (Figure 5.1). With the experience of using GaudiMM v.1 (see chapter 7) we found it useful to limit the search space, especially in those cases involving highly flexible molecules. We also expect that conformer generators will help in ring-containing molecules like glyco-lipids, which is a line of research of the group. Second, we have included a new version of the selection algorithm, NSGA-III,[153,154] capable

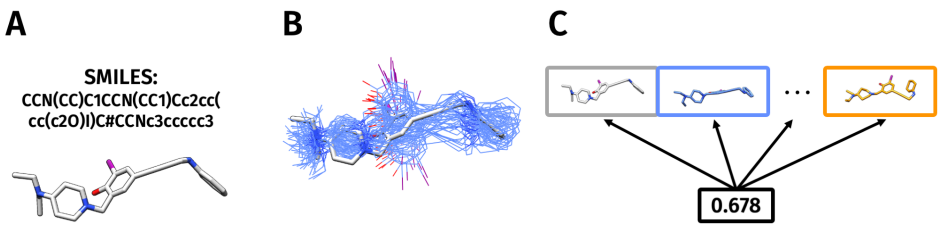



**Figure 5.1:** Gene for conformer generation in GaudiMM v.2. **(A)** The input is the code SMILES of the molecule or a base 3D conformation. **(B)** The conformer gene generates a pool of conformers for the desired molecule. **(C)** The conformers are stored in a list, ordered by similarity to the most-stable conformer. The encoding chromosome consists on a float number that identifies the position in the list of the selected conformer.

of handling the evaluation of many-objectives. Although not usual in the case of conformational sampling, we expect that handling more than 3-4 objectives will be useful in the use of chemical sampling, because several physico-chemical descriptors are often included in the optimization. Finally, all genes and objectives of the original GaudiMM version have been rethought to increment their potential. For example, we found that traditional rotamer libraries overly constrain the conformational space in the case of metal binding. To address this issue, the "rotamers" gene now includes the possibility of free rotation of the side chains in addition to the already implemented sampling through rotamer libraries.

## GAlkemist exploration: genetic programming

Switching to the GAlkemist implementation, we designed its exploration capabilities with flexibility in mind. Typically, three approaches exist to computationally construct molecules: atom-based, making small changes in the molecule at each step (like adding an atom or a bond); fragment-based, which allow bigger changes in the molecule by the incorporation of libraries of fragments; and reaction-based, which tries to build molecules based on the knowledge of existent reaction mechanisms. We opted for a hybrid between the first and second approaches, allowing the user to configure operations in GAlkemist ranging from the mutation of an atom to the merging of two large fragments (Figure 5.2).

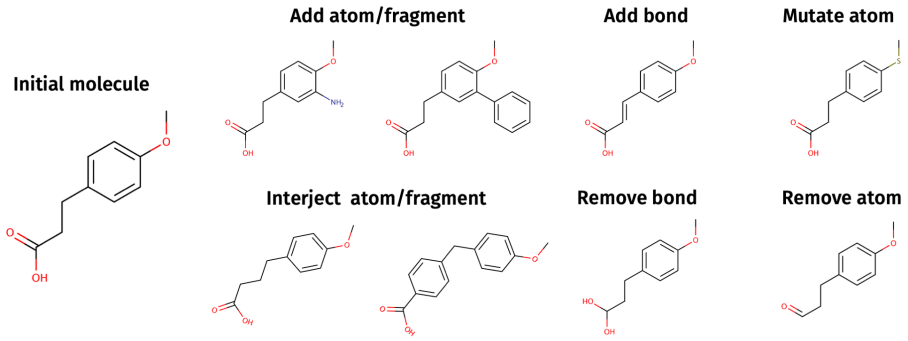



**Figure 5.2:** GAlkemist operations. Starting from an initial molecule (that can be as small as a single atom), the gene will perform a series of operations among the ones indicated. The operations allowed can be configured by the user. The chemical space allowed to explore is constrained by the valence rules, the library of fragments provided by the user, and the list of atom types also configured by the user.

For the implementation within the genetic algorithm of GaudiMM, we adapted the genetic programming methodology to tackle the problem of molecular design. In this framework, we define a set of valid operations among the ones introduced above (Figure 5.2). The "program" to evolve would be a sequence of such operations that, starting from a base molecule or scratch, gives a molecule as a result. The chromosome encoding the "program" is composed of three lists: the first contains the operations, another contains molecular fragments (i.e. operands), and the last contains float constants (Figure 5.3). This encoding facilitates the integration within the GaudiMM environment, because standard crossover and mutation operators can be applied to these lists. It is important to note here that molecular graphs or 3D conformations can be used to represent the molecules making use of other GaudiMM modules, such as the conformer generator mentioned above. This flexibility allows the application of the methodology not only to pure chemical design but also in structural modeling experiments involving coupled chemical and conformational sampling.

## GAlkemist evaluation: chemical descriptors

For the guiding of the exploration, we took advantage of the multi-objective capabilities of GaudiMM. As the descriptors for 3D structure evaluation were suf-

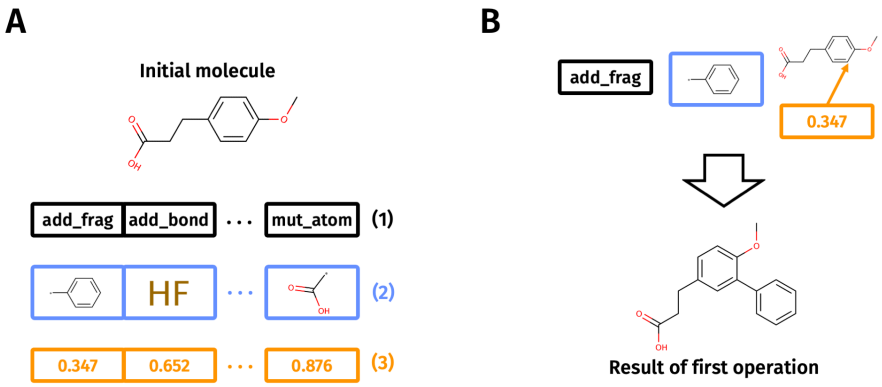



**Figure 5.3:** GAlkemist encoding uses a genetic programming approach. **(A)** The chromosome is composed of three lists. List (1) contains operations. List (2) contains molecular fragments or atoms. List (3) contains float numbers. These lists are submitted to the traditional crossover and mutation operators of the genetic algorithm to evolve different molecules. **(B)** The final molecule encoded in the chromosome is obtained through its expression or "execution". The expression starts from the initial molecule and picks the first operation contained in list (1). Depending on the operation type, one or none molecular fragment from list (2) and one or two floats from list (3) will be required and used in order. For example, the "add fragment" operation requires one molecular fragment (the one to add) and one float (to identify the atom object of the addition). The result from an operation will be the initial molecule of the following operation, until the end of list (1) is achieved.

| Name | Description |
|---|---|
| Structure | Allows to define a structural profile of the molecule, by defining ranges for the number of atoms, presence of chemical groups, number of rings, etc. |
| Similarity | Directs the search towards molecules that are similar to a target (similarity degree measured through RDKit fingerprints) |
| Drug-likeness | Optimizes the molecular properties associated with the drug-likeness of the molecule, including molecular weight, logP, HB acceptors/donors, etc. as implemented in the QED module of RDKit[229] |
| Synthesizability | Optimizes the synthetic accessibility score as described in Ref.[230] |

**Table 5.1:** Chemical descriptors incorporated in GAlkemist.

ficiently covered in the original version, here we focused on adapting several well-known physico-chemical descriptors whose functionality was provided by the RDKit library (Table 5.1).

## 5.2 Benchmark

With the aim of validating the whole implementation of GAlkemist, we first performed benchmarks on the new genetic algorithm (GA) of GaudiMM v.2, that was coded from scratch without relying on any external dependence. Then, we validated the chemical exploration module. As mentioned in the introduction of this chapter, the lack of proper benchmarks to measure the performance of molecule generators is still an issue. Usually, the tools are validated in *ad-hoc* benchmarks that are biased towards their functionalities, in standard benchmarks like GuacaMol[231] that are nearer to theoretical examples than to real research cases, or in very specific situations because the program was designed with an applicative scenario in mind. In our case we opted to assess the exploratory capabilities of GAlkemist with the goal-directed benchmark of GuacaMol[231] and report an illustrative case to show GAlkemist's integrative capabilities with the conformational exploration of GaudiMM. We intentionally left out of this work the application of GAlkemist in real research cases, which will be the subject of future work in the group.

### Benchmark of GaudiMM v.2 genetic algorithm

The implementation of the GA core of GaudiMM v.2 is based on well-known algorithms, which have been the object of extensive use and assessment. However, the code has been written from scratch, and our interest was to test possible errors and ensure correct functionality. Besides the usual unit tests to check every piece of the code, we performed a benchmark on a usual multi-objective optimization problem. Concretely, we picked the ZDT1 two-objective test function, which is a classical multi-objective problem easy to visualize.[232] Mathematically, ZDT1 is expressed as following:

$$
\begin{aligned}
f_1(x_1) &= x_1 \\
f_2(\mathbf{x}) &= g \cdot h \\
g(x_2,...,x_D) &= 1 + 9 \cdot \sum_{d=2}^{D} \frac{x_d}{(D-1)} \\
h(f_1, g) &= 1 - \sqrt{\frac{f_1}{g}}
\end{aligned}
\tag{5.1}
$$

Where $\mathbf{x}$ is a solution vector of $D$ decision variables, and all decision variables fall between 0 and 1.

The problem therefore consists of optimizing $f_1$ and $f_2$ at the same time, and to obtain a Pareto frontier as close as possible to the optimal set, which can be calculated with the expression $f_2 = 1 - \sqrt{f_1}$ (Figure 5.4A). We executed the optimization problem in GaudiMM v.2 during 200 generations, with a population of 200 individuals, crossover proportion of 0.8, mutation proportion of 0.2, and 10 decision variables. The results (Figure 5.4B) show that the GA engine is working properly, because the obtained Pareto frontier is practically overlapping the optimal one (RMSD in the objective space $< 0.01$).

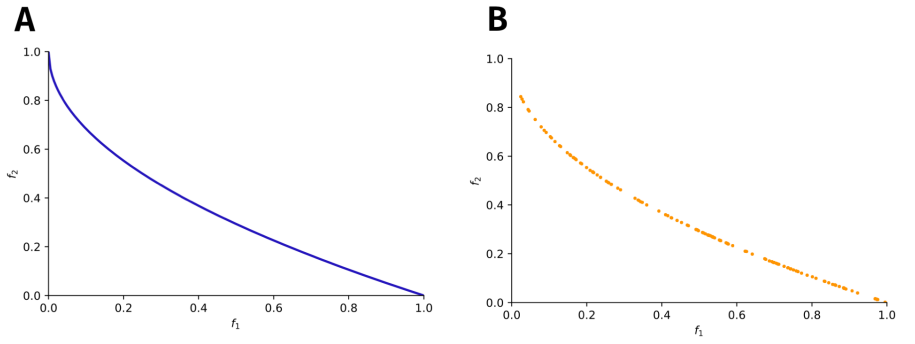



**Figure 5.4:** ZDT1 optimization problem. **(A)** Optimal Pareto front representation. **(B)** Obtained pareto front after GaudiMM v.2 optimization process, starting from 200 randomly generated individuals.

## Benchmark of GAlkemist exploration capabilities

With the aim of validating the exploration capabilities of the GAlkemist module, we employed the GuacaMol[231] goal-directed benchmark, which is a state-of-the-art set of molecular optimization problems that are not trivial to solve. Concretely, the benchmark involves the following types of problems: rediscover a given molecule, find molecules similar to a given target, generate isomers of a given molecular formula, generate median molecules, and the optimization of drug-likeness physico-chemical properties of a target. The GuacaMol article reports the results of four "baseline" methods: i) a single-objective genetic algorithm that uses as a chromosome the SMILES string representing the molecule (SMILES GA), ii) a graph-based Monte Carlo tree search algorithm (GA MCTS), iii) a graph-based single-objective genetic algorithm (Graph GA), and iv) a long-short-term memory neural network to predict SMILES strings.

We performed the benchmark twice. First, we configured GAlkemist in an atom-based fashion, without allowing operations based on fragments. Second, we allowed the fragment operations and proportioned a library of fragments based on the molecule library used by the "Graph GA" baseline approach, the one which obtained the best results on the benchmark and also the nearest in algorithmic implementation to our approach. In both cases the initial molecule was set to a single carbon atom, which is one of the smaller possible starting points and rarely used in other programs. The size of the population was set to 100 individuals (except for the isomer tasks, where it was set to the maximum number of isomers). We evolved the random initial population through 100 generations, applying a proportion of crossover/mutation of 0.5. We run each test five times and report the best result (Table 5.2).

In the first GAlkemist configuration, the results should be classified as modest. In all three rediscovery tasks, GAlkemist performance was below all baseline methods except "Graph MCTS". The same happened with the generation of median molecules, which also was worse in GAlkemist. However, in the rest of tasks, GAlkemist presents comparable results as state-of-the-art methods. Analyzing the molecules generated, we found that the main difficulty in GAlkemist

| Task | SMILES GA | Graph MCTS | Graph GA | SMILES LSTM | GAlkemist (1) | GAlkemist (2) |
|---|---|---|---|---|---|---|
| Celecoxib rediscovery | 0.732 | 0.355 | 1.000 | 1.000 | 0.643 | 1.000 |
| Troglitazone rediscovery | 0.515 | 0.311 | 1.000 | 1.000 | 0.486 | 1.000 |
| Thiothixene rediscovery | 0.598 | 0.311 | 1.000 | 1.000 | 0.512 | 1.000 |
| Aripiprazole similarity | 0.834 | 0.380 | 1.000 | 1.000 | 1.000 | 1.000 |
| Albuterol similarity | 0.907 | 0.749 | 1.000 | 1.000 | 1.000 | 1.000 |
| Mestranol similarity | 0.790 | 0.402 | 1.000 | 1.000 | 1.000 | 1.000 |
| $C_{11}H_{24}$ isomers | 0.829 | 0.410 | 0.971 | 0.993 | 0.968 | 1.000 |
| $C_9H_{10}N_2O_2PF_2Cl$ | 0.889 | 0.631 | 0.982 | 0.879 | 0.972 | 1.000 |
| Median molecules 1 | 0.334 | 0.225 | 0.406 | 0.438 | 0.283 | 0.416 |
| Median molecules 2 | 0.380 | 0.170 | 0.432 | 0.422 | 0.237 | 0.458 |
| Osimertinib MPO | 0.886 | 0.784 | 0.953 | 0.907 | 0.892 | 0.954 |
| Fexofenadine MPO | 0.931 | 0.695 | 0.998 | 0.959 | 0.924 | 0.984 |
| Ranolazine MPO | 0.881 | 0.616 | 0.920 | 0.855 | 0.898 | 0.823 |
| Perindopril MPO | 0.661 | 0.385 | 0.792 | 0.808 | 0.745 | 0.806 |
| Amlodipine MPO | 0.722 | 0.533 | 0.894 | 0.894 | 0.800 | 0.914 |
| Sitagliptin MPO | 0.689 | 0.458 | 0.891 | 0.545 | 0.763 | 0.878 |
| Zaleplon MPO | 0.413 | 0.488 | 0.754 | 0.669 | 0.689 | 0.802 |
| **Average** | **0,705** | **0.465** | **0.882** | **0.845** | **0.754** | **0.884** |

**Table 5.2:** Results of the GuacaMol benchmark.

arose with the generation of aromatic rings and other cyclic structures which, in fact, is a known issue in atom-based approaches. As the three target molecules of the rediscovery benchmark contain rings, it is not surprising the bad result of GAlkemist. Similarly it happened with the task about generating median molecules, where all four target molecules also contained cyclic motifs.

When we incorporated the fragment library in the second experiment, the results of GAlkemist became completely comparable to state-of-the-art methodologies, confirming the exploration capabilities of the approach. Although the results can cause the impression that the atom-based mode is worthless, we should remember that one of the goals of the approach is to allow a hypothesis driven generation of molecules, where a higher customization of the search could be

needed. However, it becomes clear in the light of the benchmark results, that this high flexibility comes with more difficulty in the method configuration, which should be done carefully.

## 5.3 Application in structural molecular modeling

Finally, with the aim of showing GAlkemist capabilities to sample the chemical space while simultaneously exploring the interactions with a biological target, we performed a proof-of-concept experiment with the FABP4 protein. As mentioned in chapter 1, this lipid-binding protein has been reported to reversibly bind a wide variety of hydrophobic ligands.[14] A common moiety observed in these ligands is the presence of an aromatic ring at its center, like the ones present in PDB entries 3p6d, 3p6e, 3p6f, and 3p6h (Figure 5.5A).

Therefore, we thought that an interesting experiment could be to start the optimization around a benzene molecule located at the binding site of the protein. GAlkemist was instructed to preserve the benzene moiety, while allowed

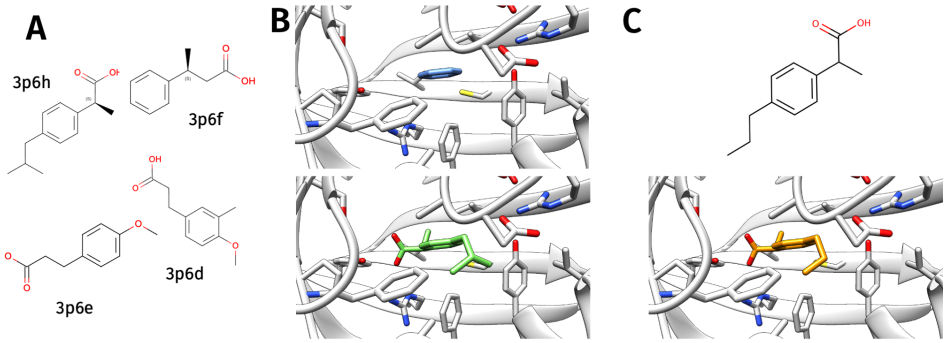



**Figure 5.5:** Proof of concept of GAlkemist chemical space exploration coupled with sampling the 3D interactions at FABP4 binding site. **(A)** Crystallographic ligands (PDB entry for each ligand is indicated). **(B)** Initial configuration of the calculation was a benzene molecule (in blue) located at the binding site (upper panel). Crystallographic binding site of PDB entry 3p6h is shown at bottom (ligand in green). **(C)** Best-scored chemical structure and binding pose (ligand in orange).

to make any atom-based operation using oxygen, nitrogen, carbon, and sulfur atoms. 3D structures of the molecules were obtained by applying the RDKit ETKDG conformer generator[63] at each expression of the GAlkemist gene. Rotation and translation of the ligand was allowed inside a sphere of radius 8 Å centered at the protein binding site. A random population of 20 individuals was evolved through 100 generations, using the standard crossover and mutation operators in a ratio of 0.5. The exploration was guided with the Vina scoring function[125] and the experiment was repeated five times.

If we consider the five best-scored poses of each run, the results showed a recurring presence of molecules with chemical structures similar to the four mentioned crystallographic ligands (Tanimoto similarity $\geqslant 0.75$ when considering the ECFC4 fingerprint), although no exact match was obtained. The carboxylic moiety was present in all the poses ranked in the top-3 of every run. Interestingly, the best-scored pose (Vina score = -7.4 kcal/mol) showed a very good superposition with the PDB entry 3p6h (Figure 5.4C). Altogether, we valorate this proof of concept as a promising result. We do not consider worrying the absence of exact chemical matches with the experimental ligands, as the protein is characterized by a high promiscuity and the exploration was guided with a single objective based on simplified MM energy.

## 5.4 Chapter conclusions and future work

In this chapter we have presented the development and benchmark of GAlkemist, a module to explore the chemical space that can be considered a hybrid approach in between atom- and fragment-based molecule generators. We have also presented the GAlkemist integration within the GaudiMM environment, which allows it to incorporate multi-objective capabilities in the guiding of the exploration, as well as the simultaneous exploration of chemical and 3D conformational spaces. Although a few state-of-the-art methods incorporate one of these two features, GAlkemist is, to the best of our knowledge, the first tool to incorporate both. We believe that GAlkmetist could be a valuable tool to assist in molecule discovery pipelines, as it was shown in the proof-of-concept exper-

iment. Short-term future goals include the deployment of a production version of the tool under an open source license, the development of tutorials to instruct how to configure the program, and publication of the scientific article. A longer term goal of the group will be to apply GAlkemist to real research problems.

# 6

# Prediction of metal-binding sites in proteins with BioMetAll

Metal ions and metallic compounds are associated with a large list of biological processes, from oxygen transport to photosynthesis, which make them essential for life.[233] In particular, one half of all existing enzymes is estimated to contain metal ions.[234] Altogether, it gives a hint about the high importance of studying biochemical mechanisms where a metal ion is involved, which could open the avenue to the rational design of metallo-drugs and artificial metalloenzymes. The first step in our understanding of such bioinorganic processes is to identify where the metal ion is located in the protein, that is, the *metal-binding site*. Generally, it will be defined by the list of amino acids involved in the first coordination sphere of the metal.

As well as in the case of organic compounds, metallic species in proteins can be identified by experimental means, such as X-ray diffraction, electron microscopy, and NMR. In fact, around one third of the structures in the PDB[235] contain a metal ion, and there even exists a database, called MetalPDB,[86,87] dedicated to provide information and statistics about all those metal-containing structures. However, these experimentally-obtained structures often do not account for the complete picture of metal binding. A particularly illustrative example is the case of $Zn^{2+}$ binding in human serum albumin (case study 1 of the base-article of this chapter[180]). Up to six structures for this system are reported in the PDB, each one obtained with a specific concentration of $Zn^{2+}$.[236] Whereas only one primary zinc-binding site is observed at lower $Zn^{2+}$ concentrations (Figure 6.1A), several secondary and tertiary binding sites appear when the concentration is increased (Figure 6.1B-C). These non-primary sites could eventually be important in some biological processes and are usually absent in crystallographic structures. Therefore, the use of computational means could be a valuable complement or even the only option to unravel the metal-binding sites.

Despite its importance, there are not many computational tools available for metal-binding site predictions. The options encompass two families: those based on the sequence of the protein and those structure-based.[237–239] Focusing only

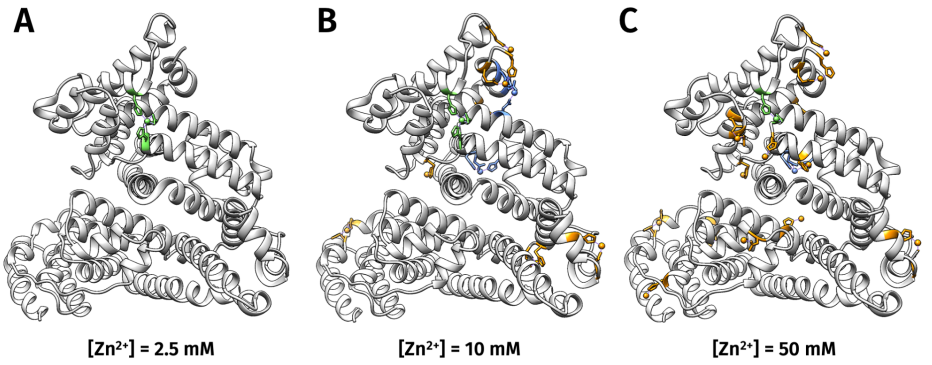



**Figure 6.1:** Binding sites of $Zn^{2+}$ in human serum albumin at different concentrations of $Zn^{2+}$. Primary, secondary, and tertiary binding sites are shown in green, blue, and orange, respectively. **(A)** $Zn^{2+}$ concentration 2.5 mM (PDB entry 5iih). **(B)** $Zn^{2+}$ concentration 10 mM (PDB entry 5iiu). **(C)** $Zn^{2+}$ concentration 50 mM (PDB entry 5ij5).

in the sequence of the protein has the limitation of neglecting the paper of conformational changes in the formation of a metal-binding site. In turn, existing structure-based predictors[240–247] had, in our view, some areas for improvement: i) all are based on perfect matches with templates of the first-coordination sphere of reported crystallographic sites, which orients them towards finding only primary metal-binding sites; ii) some of them are metal-specific, which limits their application; iii) none of them allow to explore the biological space of the protein, that is, proposing amino-acid mutations that could be useful to create or improve a metal-binding site in a metalloenzyme design scenario; iv) the deployment through the use of web-servers, together with the high time of response, difficult their application in an exhaustive screening of structures; and v) the majority of them are released under a free-for-academic-use license, and none provide the source code.

To address these challenges we decided to embark on the development of a new tool for structure-based prediction of metal-binding sites, called BioMetAll[180] (https://github.com/insilichem/biometall); with the idea of incorporating it in the future as an evaluation method within the GaudiMM framework. We licensed the software under the permissive BSD-3 clause license,[169] allowing for a free use and modification even for commercial purposes. BioMetAll is a command-line application available to install from the PiPy repository, as well as through precompiled executables that can be run on Windows, macOS, and Linux.

Previous work carried out in our group offered us a valuable clue as to how we might handle the evaluation of metal-binding sites. In a 2011 study,[248] they reported a first statistical analysis on 400 iron-containing structures, where it was observed that potential coordinating amino acids had their α-carbon located in a sphere up to 7-9 Å from the metal position. In a more recent work,[85] a filter was used to identify protein areas for metal-protein docking. Specifically, they selected those areas where there was a potential coordinating amino acid (which could be Asp, His, Glu, or Cys) with its β-carbon in a range from 2.5 to 5.0 Å. Taken together, both works suggested that a few geometric descriptors related to the protein backbone could offer useful information to identify potential metal-

binding sites. This allowed us to formulate the backbone-preorganization hypothesis, meaning that the geometry of the backbone constrains the space where a metal can bind and thus would allow a prediction of metal-binding sites.

Assuming the backbone preorganization hypothesis, we carried out a statistical analysis on all the available structures in the MetalPDB[86,87] (c.a. 170,000 metal-binding sites). We identified three geometric descriptors (distances and angles) involving the metal, α-, and β-carbons when the donor atom belongs to the side chain of the amino acid (Figure 6.2A). For the less-abundant case of a backbone oxygen donor, we identified two descriptors (Figure 6.2B). Based on these data, we developed BioMetAll, which evaluates the descriptors over a grid of probes covering the whole protein volume. If a probe satisfies all the geometric constraints, it is considered suitable for metal coordination. Those probes that share the same coordinating amino acids form a metal-binding site. With this approach, the software outputs not only the list of potential coordinating amino acids for each site, but also a 3D representation of the area suitable for coordination. The metal-binding sites are ordered based on the number of probes, which means that those sites with a greater number of valid probes are considered better.

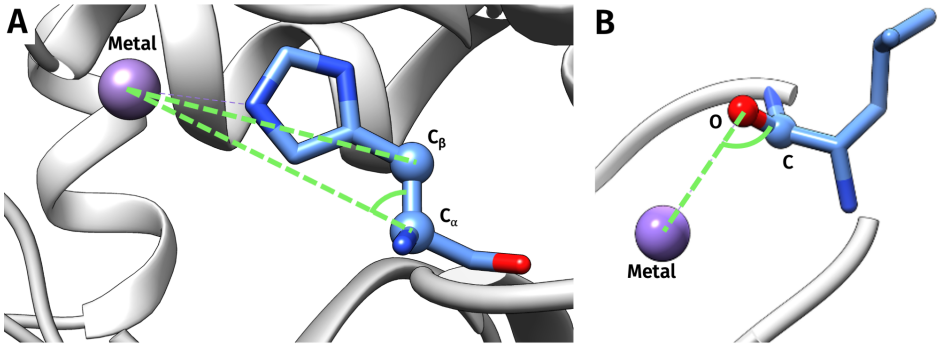



**Figure 6.2:** Geometric descriptors employed in BioMetAll binding-site evaluation. **(A)** For side-chain donors, three descriptors are considered: i) distance metal-$C_\alpha$, ii) distance metal-$C_\beta$, and iii) angle metal-$C_\alpha$-$C_\beta$. **(B)** For backbone oxygen donors, two descriptors are considered: i) distance metal-backbone oxygen and ii) angle metal-backbone oxygen-backbone carbon.

We benchmarked BioMetAll on a set of 93 already characterized structures, instructing the program to find the experimental coordination motif. All experimental motifs were found among the solutions proposed by BioMetAll, outperforming the state-of-the-art IonCom[247] and MIB[242] predictors. In addition to the primary experimental sites, BioMetAll also identified additional sites that are not necessarily false positives but could be transient environments for metal binding. In this sense, ordering the sites by number of probes allowed us to identify those most-stable sites –the crystallographic ones–, but only to a limited extent (in 75% of the cases the best-scored solution was located in the experimental binding site). Therefore, a limitation to address in future work is to build a more accurate scoring method to better classify the binding sites.

We also applied BioMetAll in three illustrative cases. They allowed us to confirm that the method was able to detect non-primary sites (we used the case mentioned above on human serum albumin) and to identify possible channels for metal binding in hemocyanins. Finally, we showed how BioMetAll is able to find existing metal-binding sites in the protein, but also to explore the biological space and propose mutations that could create a new site or improve an existing one. We believe that this feature of the program will be of great interest for the rational design of metalloenzymes.

## 6.1 Example of BioMetAll application

From its first release in December 2020, BioMetAll has been applied in several research projects, some of them already published in the form of scientific articles.[249–253] For example, in a recent work by L. Roldán and coworkers,[251] the program was used within a multiscale computational pipeline to hypothesize possible metal-binding modes in β-amyloid fibrils.

Plaque deposits of the β-amyloid peptide outside the cells are a hallmark in the diagnosis of Alzheimer's disease. However, the factors influencing the aggregation of the β-amyloid peptides to form fibers are still under debate, and one of the hypotheses is the influence of metals in the process. In this work they proposed to computationally study the metal-ion binding at two steps of the plaque

formation process: when the β-amyloid is still in monomeric form, to evaluate whether the metal binding could induce preorganized structures that have more tendency to aggregate; and in the oligomeric form, to assess if metal binding could have a paper in the stabilization of the β-amyloid fibrils. It was in the last case where they used BioMetAll within a two-step protocol. First, a screening with BioMetAll was performed in the structure of the β-amyloid fibril (PDB entry 2mxu[254]), which allowed them to find the most probable metal-binding sites. Then, with a docking experiment employing GOLD,[84,128] they found two coordination modes (one for $Cu^{2+}$ and the other for $Al^{3+}$) at the sites previously identified (Figure 6.3). The obtained structures confirmed the possibility of metal binding in the oligomeric state and opened the avenue to study the impact of metals in the stabilization of the fibril.

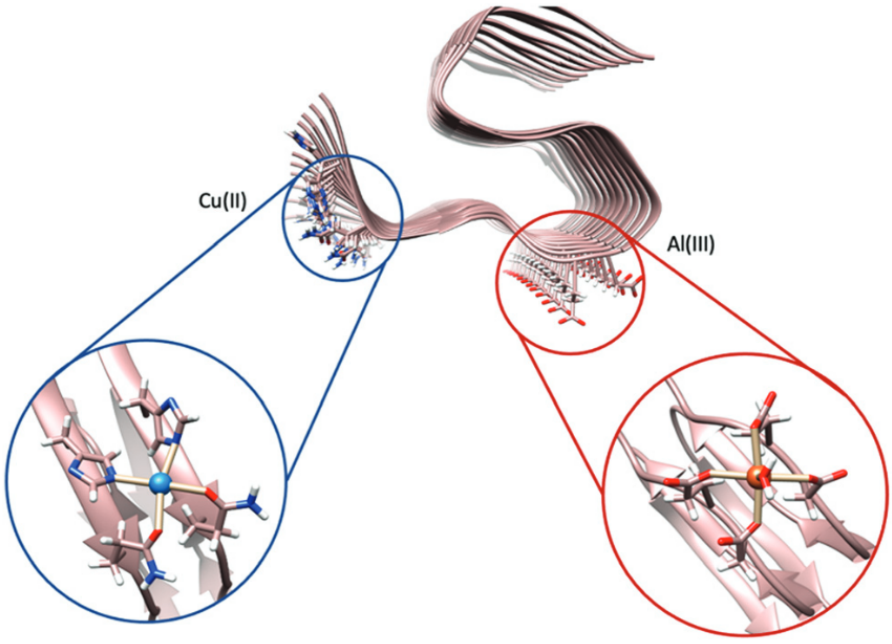



**Figure 6.3:** Main predicted metal-binding sites for the β-amyloid fibril. Two docking examples of $Cu^{2+}$ and $Al^{3+}$ are highlighted in blue and red circles, respectively. Figure reproduced from Ref.[251] with permission from the PCCP Owner Societies.

## 6.2 Chapter conclusions and future work

In this chapter we have seen how we successfully developed and benchmarked BioMetAll, a novel tool for the prediction of metal-binding sites in proteins, which provides an evaluation of the candidate binding sites based on simple geometric descriptors of the protein backbone. We published the method in the form of a scientific article.[180] Currently, we are developing a new scoring method that goes beyond just counting valid probes. This new score will presumably allow better discrimination of those more stable sites, and we plan to integrate it within the GaudiMM framework. One of the projects in this sense is to build an extension of GPathFinder capable of identifying binding pathways of metallic species.

# 7

# Application of GaudiMM to non-standard dockings

The primary goal of a protein-ligand docking experiment is to predict the predominant binding mode of a ligand at the protein binding site. Conventional docking programs are generally highly efficient from a computational point of view. Due to their specialized search algorithms and quick evaluation functions, you can obtain in a matter of seconds or minutes the solution/s to your system even employing a desktop computer. Their fast execution time has also allowed the incorporation of docking methods to workflows where a high number of executions are needed, for example, in virtual screening campaigns for drug discovery.

However, when the requirements of the system under study fall outside the scope of the docking program design, things become harder or even impossible to manage. Some limitations of standard docking approaches are: i) their single-objective scoring function, which restricts their application when other than protein-ligand interactions must be considered; ii) the constrained exploration space they use to avoid falling into a combinatorial explosion, limiting both the flexibility allowed to the protein and the size of the binding site; and iii) the type of ligands that can be considered, which are highly conditioned by the ligand sets employed to design and validate the program.

A more generic and modular framework like the GaudiMM platform[91] might be of great utility in the case of such *non-standard dockings*. The multi-objective nature of GaudiMM evaluation, coupled with the exploration power of its genetic algorithm, can help to address the limitations mentioned above. In this chapter, we aim at continuing the line of the InsiliChem group in pushing the limits of the application of GaudiMM in docking.[157–160,78,85] Concretely, we report three research cases that are associated with their own particularities about exploration and evaluation:

i) The docking of a polyfluoroalkyl $sp^2$-glycolipid, which involved the sampling of more than 20 rotatable bonds for the ligand, in the limit of conventional docking programs' capabilities. In this case, we opted for a consensus approach using four different pieces of software: AutoDock4[124] AutoDock Vina,[125] GOLD,[128] and GaudiMM.[91] For the standard docking programs, their parameters were fine-tuned to account for the big search space. In GaudiMM, we implemented the exploration through a pre-sampling of the most-stable ligand conformers.

ii) The docking of non-covalent complexes formed by the union of two disaccharides, which involved the presence of inter-ligand interactions besides the usual protein-ligand scenario. In this case, we opted for a two-step protocol, first ascertaining the inter-ligand interactions with a DFT approach, and then docking the resulting complexes into the enzymes using GaudiMM.

iii) The docking of vitamin $B_{12}$ into sugarcane-derived activated carbon. Here, the main challenge was the big size of the receptors (with a volume of c.a. 1,000

$nm^3$ and a number of atoms between 32,000 and 52,000), which needed to be sampled in a blind-fashion. We took advantage of the exploration capabilities of GaudiMM, guiding the results towards geometrically and energetically feasible solutions.

In all three cases, the docking with GaudiMM was a contribution to a broader work that was carried out in collaboration with different national and international groups. While the first two studies have already been published in the form of scientific articles,[255,256] the third is still in the process of being published. For the former, we provide a brief description to understand the context avoiding unnecessary duplication with already published work. For the latter, a more detailed description of the computational methods and results is also provided. Finally, it is worth noting that, in addition to the three cases reported here, you will find another interesting application of GaudiMM on covalent docking of a metallic cofactor in chapter 8.

## 7.1 Binding of a polyfluoroalkyl $sp^2$-iminosugar glycolipid in the p38 mitogen activated protein kinase

Immunomodulatory glycolipids, such as α-galactosylceramide (KRN7000), have demonstrated high therapeutic potential as anti-tumorals and in microbial infections.[257] However, they often provoke a cytokine storm that adversely impacts the immune response. Therefore, the synthesis of analogues with less cytokine secretion induction is a promising line of research. In this sense, the substitution of the monosaccharide glycone (Figure 7.1A) by a $sp^2$-iminosugar glycomimetic moiety (Figure 7.1B) facilitates the α-stereocontrol of the synthesis.

In this work,[255] the synthesis of six polyfluoroalkyl $sp^2$-iminosugar glycolipids is reported, as well as their evaluation for anti-proliferative, anti-leishmanial, and anti-inflammatory activities in cells. All six compounds share the same overall

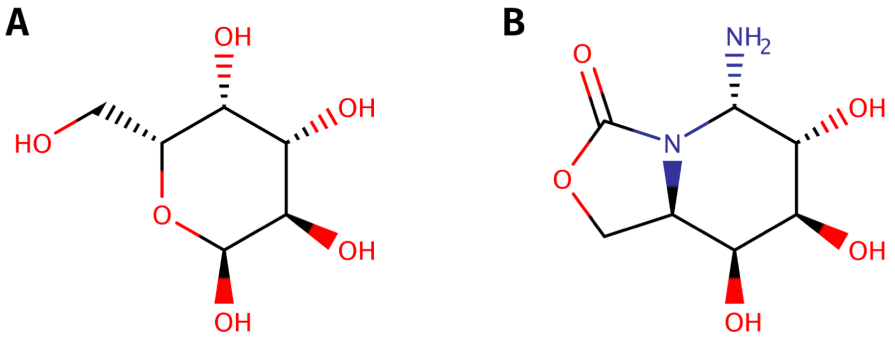



**Figure 7.1:** Comparative between the monosaccharide glycone motif present in the KRN7000 molecule and the $sp^2$-iminosugar glycomimetic moiety used in this study. **(A)** Glycone motif. **(B)** $Sp^2$-iminosugar glycomimetic motif.

structure: a $sp^2$-iminosugar glycone linked to a polyfluorocarbon motif by a non-amethylene portion. They differ in the length of the fluoro tail (with three, five, or seven $CF_2$ groups) and the α-anomeric configuration of the glycone (which can be α-D-*gluco*-like or α-D-*galacto*-like). An example of compound is given in Figure 7.2, with an α-D-*galacto*-like configuration of the glycone head and five $CF_2$ groups in the tail.

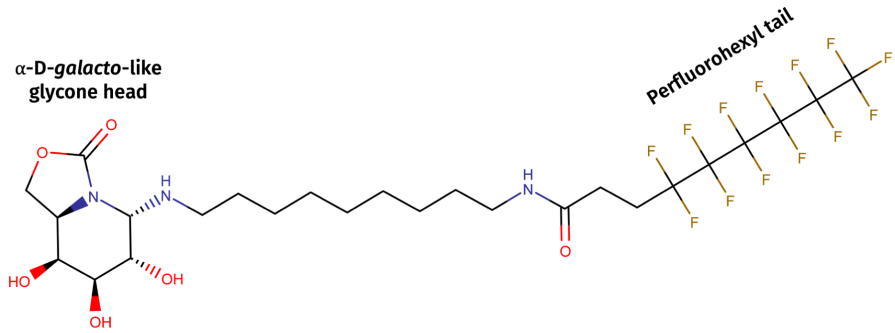



**Figure 7.2:** Structure of the OGJ (D-galacto) perfluorohexyl sp2-glycolipid (compound 26 in the base-article of this section).[255]

The experimental evaluation of the six compounds showed that those with perfluorohexyl tails (five $CF_2$ groups plus one $CF_3$ group) present better overall results. Strikingly, for the case of anti-inflammatory activity, the compound with D-*galacto* head (shown in Figure 7.2) had about 10-fold higher efficiency than the compound with D-*gluco* head. Therefore, our molecular modeling experiment was focused on these two perfluorohexyl $sp^2$-glycolipids.

Previous studies on other lipids suggested a mechanism of action through the binding into the allosteric lipid binding site of the p38 mitogen activated protein kinase (p38 MAPK).[258,259] However, the introduction of the polyfluorocarbon moiety implies an increase in the chain rigidity that could affect protein binding. Hence, the primary question we wanted to address with a molecular docking approach was: is it possible the binding of the perfluorohexyl $sp^2$-glycolipid in the allosteric site of the p38 MAPK? In an affirmative case, a subsequent question was if some differences could be observed between the D-*gluco* and D-*galacto* dockings that suggest an explanation for their different anti-inflammatory efficiency.

The challenge in this case was related to the large size of the compound (47 heavy atoms) and its high flexibility (22 rotatable bonds). Conventional docking programs are benchmarked on test sets where the ligands are often outside of these values. For example, the full CCDC/Astex data set, on which GOLD scoring functions were validated, has only 10 of 305 ligands with a number of heavy atoms $\geqslant 50$, and 23 of 305 ligands with a number of rotatable bonds $\geqslant 22$.[260] Although we do not have a detailed list of the success/failure cases of the benchmark, the overall success ratio (i.e. the top-ranked GOLD solution is within 2.0 Å of the exeperimental binding mode) of Goldscore is $68.4 \pm 1.2\%$.[128] It clearly gives us a hint that we are playing with the limits of the methods.

For that reason, we opted for a consensus approach, carrying out the dockings with four different programs: AutoDock4,[124] AutoDock Vina,[125] GOLD,[128] and GaudiMM.[91] The parameters of the conventional docking programs were configured to take into account the particularities of the ligand. For the GaudiMM approach, we decided to pre-calculate the ligand conformational space with the

ETKDG conformer generator[63] implemented in the RDKit library,[227] and focus the evaluation on LigScore,[138] a different scoring function than those used by the other three methods. It is worth noting here that the conformational sampling via knowledge-based conformer generators turned out to be an interesting approach, and we decided to implement it as a gene in the new GaudiMM version.

Docking results confirmed the geometric and energetic feasibility of the perfluorohexyl binding into the lipid binding site of p38 MAPK, therefore supporting the hypothesis of this mechanism of action. In addition, a difference in the binding modes between the D-*galacto* and D-*gluco* was observed. Whereas the C-4 axial hydroxyl of the D-*galacto* compound was found to form a hydrogen bond with amino acid Asn196, this situation was not observed in any docking pose for the D-*gluco* case, which we hypothesize might be a factor to explain the different anti-inflammatory profiles of both compounds.

## 7.2 Binding of disaccharide complexes into YKL-39 and hHyal-1 enzymes

Capsular systems formed by the non-covalent union of oppositely-charged polyelectrolytes are useful for several biomedical applications. For example, they can help to overcome the traditional problems with the delivery of high molecular weight drugs, as well as being used as tissue adhesives and scaffolds for tissue engineering.[261,262] One aspect that has not yet been addressed regarding these complexes is the molecular mechanisms of their enzymatic decomposition, which are of scientific relevance given the intended medical use.

In this work,[256] our aim was to study the interaction of various polyelectrolyte complexes with the key enzymes responsible for their metabolic decomposition (Table 7.1). For that, we used disaccharide complexes as a model for the polysaccharides. An example is given in Figure 7.3, where the CHI/FUR complex is formed by the union of chitosan (CHI) and furcellaran (FUR). We employed a two-step protocol. First, single disaccharide ligands and the four disaccharide complexes were modeled through DFT calculations by P. Paneth's group. The

result of this step was the three-dimensional models of the disaccharide complexes, such as the one shown in Figure 7.3C. Then, the ligands and disaccharide complexes were docked into the enzymes using GaudiMM. It was in this second step of the protocol where we collaborated intensively to properly configure the GaudiMM recipe and analyze the results.

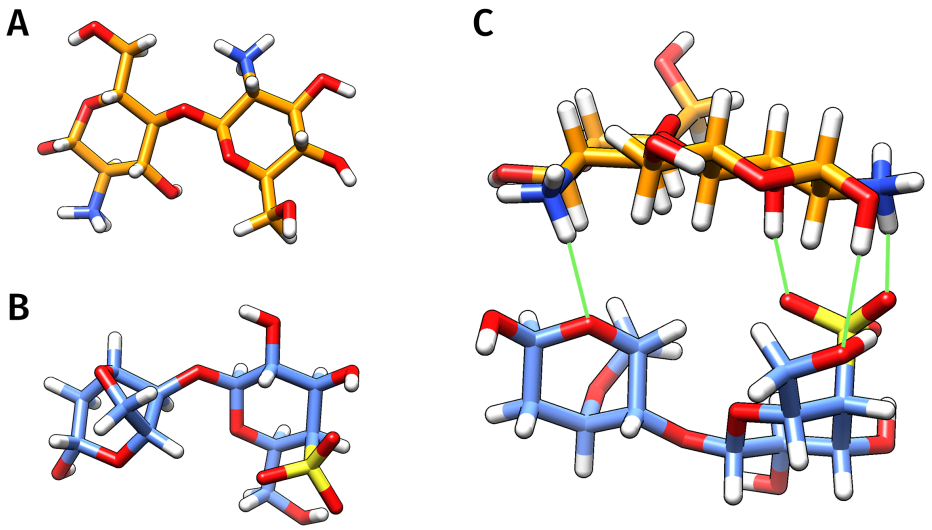



**Figure 7.3:** Formation of the chitosan/furcellaran complex. **(A)** Chitosan disaccharide. **(B)** Furcellaran disaccharide. **(C)** Chitosan/furcellaran complex, formed by the non-covalent union of the two disaccharides. Hydrogen bonds stabilizing the complex are shown in thin green sticks.

| Enzyme | Function | Disaccharides docked | Complexes docked |
|---|---|---|---|
| Human cartilage chitinase 3-like Protein 2 (YKL-39) | Hydrolyzes the glycosidic bond in chitin | CHI, FUR | CHI/FUR |
| Lysosomal human hyaluronidase-1 (hHyal-1) | Hydrolyzes the hyaluronic acid | CHI, HA, CB, HP | HA/CHI, HA/CB, HP/CB |

CHI: chitosan; FUR: furcelleran; HA: hyaluronic acid; CB: cellobiose; HP: heparin

**Table 7.1:** Enzymes and disaccharide complexes analyzed in this work.

GaudiMM calculations included four objectives to guide the search, accounting with geometric and energetic evaluation, as well as giving a relevant paper to the hydrogen bond network that plays a crucial role in these systems. Concretely, the objectives to optimize were: i) hydrogen bond network between the disaccharide complex and the enzyme; ii) MM energy using the Amber99SB forcefield for the standard residues and GAFF-based parameterization[105] for the rest of the atoms; iii) steric clashes of the whole structure; and iv) Vina docking score[125] between the disaccharide complex and the enzyme.

This multi-objective evaluation allowed us to select those final poses presenting a best consensus between all the descriptors we thought important to describe the system. It is worth noting here that, besides the multi-objective evaluation, the mere docking of two non-covalently bound ligands is something not common in standard docking software. Although there exist some programs that can handle multi-ligand docking, such as the latest version of Autodock Vina,[88] the ligands are always treated as separate entities without taking into account the previously calculated network of non-covalent interactions between the two disaccharides forming the complex.

In fact, at the beginning of the project, we contemplated a one-step protocol where we docked directly the two disaccharides inside the enzyme, without a previous calculation of the disaccharide complexes. However, we realized that the inter-ligand interactions were not taken into account in docking software like Autodock Vina 1.2,[88] which made the experiment impossible with standard approaches. Carrying out the experiment with GaudiMM was in theory possible, but additional objectives were necessary to account with the inter-ligand interactions (at least, the hydrogen bond network and energy evaluation between disaccharides). The total of objectives increased to at least six, which made it not advisable to follow this approach for two main reasons. First, the NSGA-II selection algorithm incorporated in GaudiMM is not designed to handle more than 3-4 objectives.[153] And second, the results would be more difficult to analyze, because the relative importance of the different objectives was not clear *a priori*. In particular, deciding whether to prioritize those solutions with better protein-ligand or inter-ligand interactions was an impossible decision to make without

further information. As the DFT calculations showed a strong interaction between disaccharides, we decided to divide the protocol in the two mentioned steps, assuming that the disaccharide complexes were acting as a single entity.

As a general conclusion of the docking part of this work, we observed that the binding of the disaccharide complex was possible in all cases in the native binding site of the enzymes. The complexes always show higher affinities than the individual disaccharides in both enzymes, while the CHI/FUR complex has also higher affinity than the native YKL-39 ligand (for the hHyal-1 case there was not a good *holo* structure available to compare). On the methodological side, we realized that the incorporation of a selection algorithm capable of managing more than four objectives, such as the NSGA-III, was an interesting option to consider in the future version of GaudiMM.

## 7.3 Interactions between sugarcane-derived activated carbon and vitamin $B_{12}$

The intensive use of chlorinated pesticides such as chlordecone (CLD) has caused a serious environmental problem of water and soil contamination in the French West Indies.[263] The extension of CLD contamination to the food chain constitutes also a public health issue, because CLD interferes with reproduction, is suspected of being an endocrine disruptor, and is classified as possibly carcinogenic to humans by the International Agency for Research on Cancer.[264]

A promising approach to allow an environmentally friendly degradation of CLD is the use of nanohybrid materials formed by the non-covalent union of porphyrins and carbon materials such as carbon nanotubes, single- and multi-walled nanotubes, and activated carbon (AC).[265] Non-covalent binding has an advantage as a functionalization process over the covalent technique because it produces stable materials that do not disrupt electronic, optical, or catalytic properties of both components of the nanohybrid.[266,267]

The ultimate goal of this work was to obtain a hybrid material that could be used for the degradation of CLD. Concretely, a non-covalent material formed by a sugarcane AC and vitamin $B_{12}$ (VB12) was the main object of study. VB12 is an organocobalt porphyrin (Figure 7.4) which is known for its ability to reduce chlorinated pollutants.[268–271] Even more, a recently published work showed that VB12 is able to achieve the degradation of CLD leading to the opening of the CLD cage structure to produce pentachloroindene.[265,272] Therefore, the development of this hybrid AC-VB12 material definitely constitutes a promising approach to address the problem of water and soil contamination by chlorinated pesticides.

The study was focused on ascertaining the absorption and non-covalent interactions of VB12 on the sugarcane-derived AC. Our participation in this collabora-

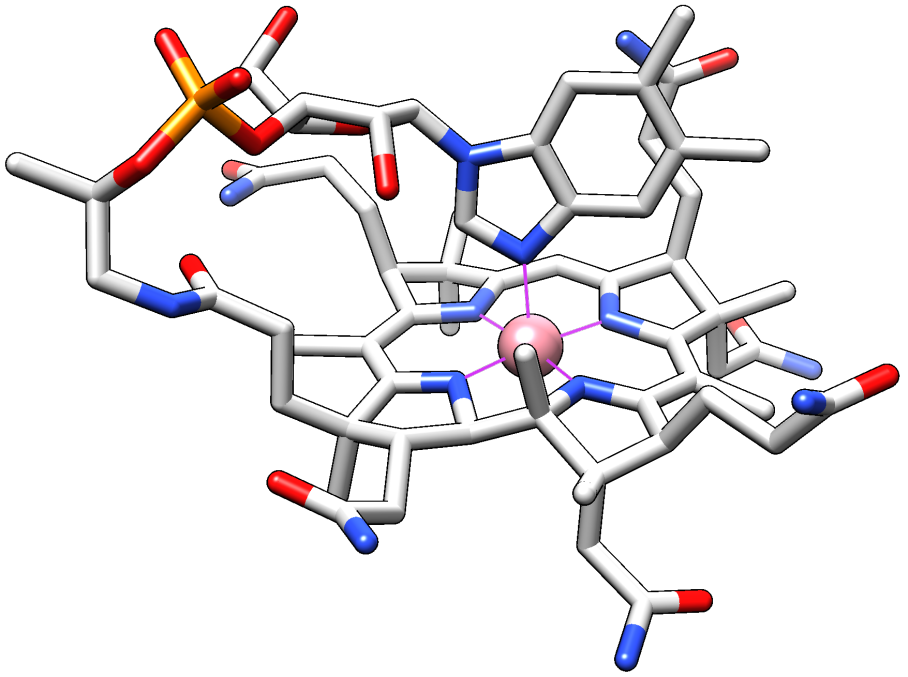


**Figure 7.4:** Three dimensional structure of vitamin $B_{12}$, obtained from PDB entry 1ddy.[273]

tive work was centered on performing a molecular modeling protocol based on a docking approach to get insight on the non-covalent interactions between VB12 and AC, and obtain some 3D models.

## Computational details

First, the atomic structure of VB12 was obtained from the PDB (entry 1ddy[273]). Subsequently, this structure was docked into the activated carbon atomistic models proposed by F.S. Cannon and coworkers:[274] three AC structures of size 200 x 200 x 200 Å (here denominated cubes AC1, AC2 and AC3) which contain a different percentage of sheets with curvature (70%, 50% and 30%, respectively) and fit the experimental characterizations of elemental composition and porosity of AC. Each one of these models are composed of eight smaller cuboids of 100 x 100 x 100 Å. In the dockings performed here, all non-redundant "small" cuboids (eight for AC1, five for AC2, and five for AC3) were sampled.

Two different approaches were envisaged at the beginning of the study: i) screening of the different cuboids for cavities with volumes superior to the dimension of VB12 (1200 $Å^3$) followed by docking in the best sites, and ii) blind docking of the ligand directly in the entire cuboid. For volume calculations, the SURFNET algorithm[275] as implemented in UCSF Chimera[22] was used with a cut-off distance of 6 Å and a prove size of 18 $Å^3$. Regarding dockings, they were performed with the GaudiMM platform.[91] Calculations were carried out allowing full flexibility to the VB12 torsional angles and using two objectives to evaluate its binding within the AC structure: clashes in order to minimize bad contacts of the VB12 molecule with the skeleton of AC and Vina score to obtain an approximated energetic value.[125] The .yaml input file used in the GaudiMM calculations is provided at the end of this chapter.

Test calculations of both procedures were performed on cuboid AC1.1 (selected randomly). Similar results were obtained with both approaches. Therefore, the one consisting of blind dockings was selected for the rest of the study, since it globally required fewer computational resources and the results were faster to analyze. Note here that the space to cover is extremely large (each cuboid

has a volume of c.a. 1,000 $nm^3$ and a number of atoms between 32,000 and 52,000). In this scenario, the genetic algorithm of GaudiMM revealed a very good capacity for exploration. To ensure a good covering, four independent replicas were performed for each of the 18 cuboids.

## Results and discussion

GaudiMM calculations were performed on the 18 non-redundant cuboids of the AC models from Cannon and coworkers.[274] From the four replicas run for each cuboid, all best solutions present no bad contacts (i.e. steric clashes) between the VB12 molecule and the AC structure, hence sustaining that the VB12 could find hosting cavities with excellent matching (Table 7.2). When looking at the energetic prediction by the GaudiMM descriptor associated with the Vina scor-

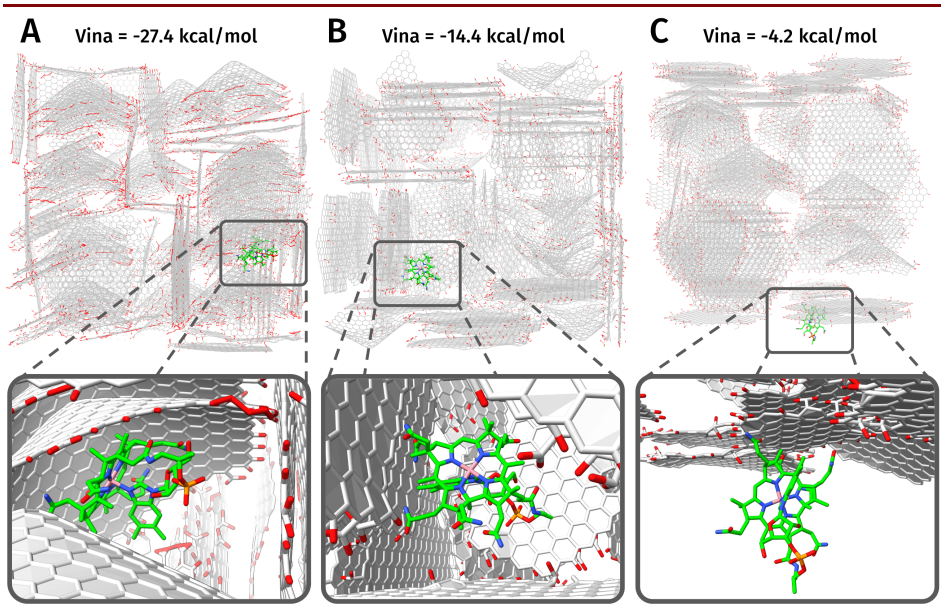



**Figure 7.5:** Docking solutions for the VB12 molecule bound to AC structure. VB12 is depicted in green sticks, while AC is shown in gray. **(A)** Vina score = -27.4 kcal/mol. VB12 is bound near a conic sheet of the AC structure. **(B)** Vina score = -14.4 kcal/mol. VB12 is bound near several plain sheets of the AC structure. **(C)** Vina score = -4.2 kcal/mol. VB12 is bound at the exterior surface of the AC structure.

ing function, calculations show a variation in binding energy, ranging from -1 to -27.4 kcal/mol (Table 7.2). The most stable interactions are observed in model AC1, belonging the best-scored structure to a binding in cuboid AC1.2 (-27.4 kcal/mol).

The origin of this stability was further investigated by visual inspection of all the complexes. Strikingly, it is observed that the vitamin $B_{12}$ is systematically incorporated into a cavity with a conic form in complexes with a good Vina score (lower than -16.0 kcal/mol, Figure 7.5A). In docking poses with intermediate Vina score (between -12.0 and -16.0 kcal/mol, Figure 7.5B) the VB12 generally interacts with one or more plain sheet/s of the AC, while in the worst-scored complexes (higher than -12.0 kcal/mol, Figure 7.5C) the VB12 is usually found at the exterior surface of the cuboid. Calculations therefore suggest that those

| | Replica 1 | | Replica 2 | | Replica 3 | | Replica 4 | |
|---|---|---|---|---|---|---|---|---|
| | Clashes | Vina | Clashes | Vina | Clashes | Vina | Clashes | Vina |
| **Cube AC1** | | | | | | | | |
| AC1.1 | 0 | -18.869 | 0 | -12.627 | 0 | -21.856 | 0 | -11.773 |
| AC1.2 | 0 | -11.925 | 0 | -18.394 | 0 | -9.264 | 0 | -27.395 |
| AC1.3 | 0 | -21.889 | 0 | -12.380 | 0 | -10.253 | 0 | -10.175 |
| AC1.4 | 0 | -8.811 | 0 | -10.261 | 0 | -10.967 | 0 | -12.834 |
| AC1.5 | 0 | -7.958 | 0 | -8.568 | 0 | -9.725 | 0 | -0.997 |
| AC1.6 | 0 | -12.325 | 0 | -12.507 | 0 | -10.786 | 0 | -13.311 |
| AC1.7 | 0 | -14.519 | 0 | -10.569 | 0 | -12.877 | 0 | -8.443 |
| AC1.8 | 0 | -12.557 | 0 | -13.934 | 0 | -12.922 | 0 | -10.059 |
| **Cube AC2** | | | | | | | | |
| AC2.1 | 1.661 | -11.776 | 0 | -8.624 | 0 | -23.528 | 0 | -14.500 |
| AC2.2 | 0 | -9.162 | 0 | -15.800 | 0 | -15.330 | 0 | -9.861 |
| AC2.3 | 0 | -12.602 | 0 | -4.211 | 0 | -13.029 | 0 | -10.548 |
| AC2.4 | 0 | -16.726 | 0 | -11.247 | 0 | -10.704 | 0 | -16.304 |
| AC2.5 | 0 | -9.659 | 0 | -14.604 | 0 | -14.403 | 0 | -14.770 |
| **Cube AC3** | | | | | | | | |
| AC3.1 | 0 | -10.438 | 0 | -19.253 | 0 | -16.519 | 0 | -8.455 |
| AC3.2 | 0 | -11.276 | 0 | -17.098 | 0 | -5.156 | 0 | -15.194 |
| AC3.3 | 0 | -15.103 | 0 | -15.553 | 0 | -19.563 | 0 | -15.175 |
| AC3.4 | 0 | -11.071 | 0 | -10.043 | 0 | -9.269 | 0 | -9.113 |
| AC3.5 | 0 | -10.731 | 0 | -8.743 | 0 | -14.941 | 0 | -14.440 |

**Table 7.2:** Clashes ($Å^3$) and vina (kcal/mol) scores for the best structure obtained in each calculation.

sites near curved sheets are preferred for the binding of VB12. Such sites are comprehensively the most stable for VB12 because they provide a very high hydrophobic interaction with a large part of VB12. Sometimes, the binding pose allows one of the main faces of the porphyrin to be exposed to the void and, subsequently, able to act as a catalyst with an incoming substrate.

### Conclusions

Altogether, the molecular modeling study performed here suggests that VB12 is definitely able to bind into AC through non-covalent interactions, because good geometric complementarity and good binding-energy scores were observed. Specifically, the best results were obtained in the vicinity of the most-curved shapes of the AC, which also offers a suitable space for the subsequent catalytic step in the CLD degradation process.

## 7.4 Chapter conclusions and future work

In this chapter we have seen how a generic molecular modeling platform, such as GaudiMM, can help to determine the ligand binding poses in systems that conventional docking approaches struggle with, or are even unable to handle. These non-standard docking cases often imply a higher amount of sampling, which is out of the scope of conventional programs' design. Although being, of course, a slower approach (minutes/hours vs. seconds/minutes), GaudiMM's multi-objective genetic algorithm proved to be a good choice when this large amount of sampling is needed. From the lessons learned in these three research cases, we obtained some ideas to implement in the future version of GaudiMM, such as the conformational exploration through knowledge-base conformer generators, and the incorporation of a selection algorithm capable of efficiently managing more than four objectives.

# Annex: GaudiMM .yaml file used for the docking of VB12 in AC

```
output:
    path: ./cubeAC1_1_00  # Name of the cuboid and number of replica
    name: cubeAC1_1_00
    verbose: True
    pareto: False
    check_every: 0

ga:
    cx_eta: 5
    cx_pb: 0.5
    generations: 200
    mu: 1
    mut_eta: 5
    mut_indpb: 1.0
    mut_pb: 0.5
    population: 100

similarity:
    module: gaudi.similarity.rmsd
    args: [[Ligand], 0.5]
    kwargs: {}

genes:
    -   name: Ligand
        module: gaudi.genes.molecule
        path: ./mol_files/B12_with_H.pdb   #File of VB12 structure
    -   name: Receptor
        module: gaudi.genes.molecule
        path: ./mol_files/AC1_1.pdb    #File of AC cuboid structure
```

```
    -   name: Search
        module: gaudi.genes.search
        radius: 65
        precision: 5
        rotate: True
        target: Ligand
        center: [0, 100, 0]
        interpolation: 0.5
    -   name: Torsion
        module: gaudi.genes.torsion
        target: Ligand

objectives:
    -   name: Clashes
        module: gaudi.objectives.contacts
        which: clashes
        weight: -1.0
        probes: [Ligand]
        radius: 5.0
    -   name: Vina
        module: gaudi.objectives.vina
        weight: -1.0
        receptor: Receptor
        ligand: Ligand
```

# 8

# Multiscale workflows applied to bioinorganic systems

Despite all the efforts made (including those of this thesis) in the development of novel computational tools, each time with greater exploration capacity and evaluation accuracy, the truth is that the complexity of biochemical systems sometimes make it impossible to rely on a single method to achieve a comprehensive understanding of their molecular mechanisms. This is specially the case when events occurring on different time scales are crucial to the function of the system, for example, the unfolding of a protein helix followed by the binding of a ligand, followed in turn by a reaction at the protein binding site. In these cases, the combination in a sequential workflow of several tools, each one for the specific study of a part of the mechanism, can be of great help.

In this chapter, we present the optimization and application in real research cases of two multiscale workflows, one for the study of the interactions between a metallodrug (oxaliplatin) and a protein (insulin),[276] and the other for the study of the enantioselective profile of a reaction (cyclopropanation) catalyzed by a metalloenzyme formed by the covalent anchoring of a dirhodium homogeneous catalyst into a prolyl oligopeptidase.[277] Our aim here will be to focus the explanation on the computational workflows employed and what were the main challenges to address, while providing the context of the project and the main results obtained.

## 8.1 Interaction of oxaliplatin with insulin

Although the use of metallic compounds for medical treatments has been present for more than a century,[278] the eclosion of metallodrugs has its turning point in the approval by the FDA in 1978 of cisplatin as an anticancer agent. Since then, dozens of inorganic compounds have entered the FDA/EU market, and the field of medicinal inorganic chemistry has experienced a tremendous increase.[279] However, if we compare the numbers with the available organic drugs, we realize that metallodrugs still represent a small fraction. In this framework, understanding the unique mechanisms of action of metallodrugs is of special interest for the development of these treatments.

The successful effect of cisplatin as an anticancer agent fostered the emergence of second and third generations of platinum drugs aiming to address some of cisplatin drawbacks, such as its high toxicity and side effects, low bioavailability, and resistance profiles.[280–284] One of these new-generation drugs is oxaliplatin (Figure 8.1A), especially effective against metastatic colorectal cancer. Despite its improvements with respect to cisplatin regarding toxicity levels, some issues remain about resistance profiles in certain patients. In particular, high levels of insulin (Figure 8.1B) has been related to resistance in colon cancer cell lines via activation of PI3K/Akt pathway,[285] and direct interactions of cisplatin and oxaliplatin with insulin have been demonstrated.[286–290] Although the molecular details of the interaction mechanism between oxaliplatin and insulin have poten-

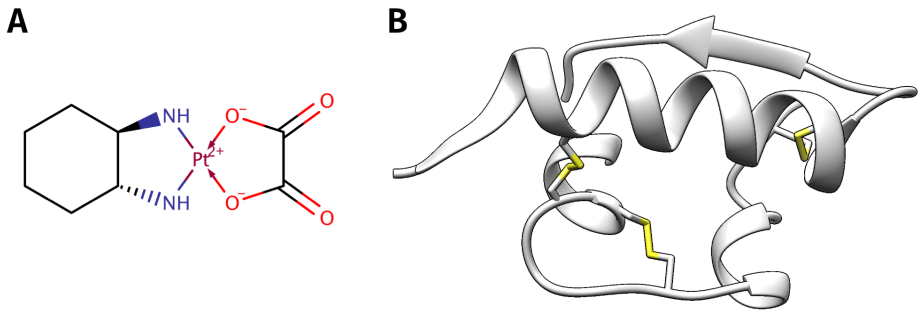



**Figure 8.1:** Biochemical system object of study. **(A)** Oxaliplatin molecule, which is composed of a square planar platinum(II) center, complexed with 1R,2R-diaminocyclohexane and a labile oxalate ligand. One or both hydroxide ions of the oxalate ligand can be replaced by amino acids of the protein. **(B)** Three-dimensional structure of insulin molecule, obtained from PDB entry 1zni.[291] Insulin is a small protein composed of two peptides, linked by two interchain and one intrachain disulphide bridges, which are shown in sticks.

tial to help in the development of more effective treatments, three-dimensional structures of the insulin-oxaliplatin adducts had not yet been obtained. In this work,[276] we aimed at applying a computational workflow to understand such interaction mechanisms and obtain accurate 3D structures of the oxaliplatin-insulin complex.

## Computational workflow employed

The core of the multiscale workflow that we designed (Figure 8.2) is based in a common approach used in drug discovery pipelines: a first step employing molecular docking to obtain the most relevant binding poses of the drug within the protein, followed by classical MD simulations to assess the stability of the drug-target adducts and the possible effects on the protein conformation. The main challenge was to incorporate the explicit treatment of the metal into these two methodologies.

In the case of the docking, we opted for the GOLD protocol previously developed in our group,[84] because the protein structure was already preorganized for metal binding, and its small size (51 amino acids) allowed for a direct screening of all possible metal-binding sites. If a wider screening was required to assess the

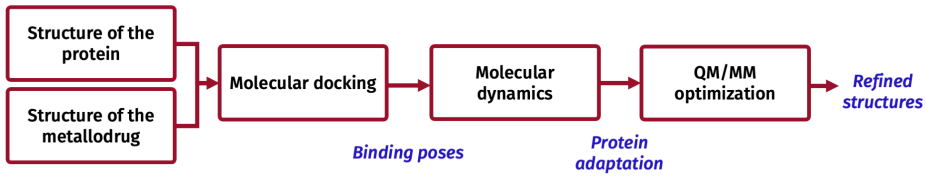



**Figure 8.2:** Computational workflow employed. Starting from three-dimensional structures of the insulin and oxaliplatin molecules, the workflow consists of three sequential steps: molecular docking to obtain the main binding poses, classical molecular dynamics to assess the protein adaptation, and a final refinement of the most representative structures by QM/MM optimization.

location of the metal-binding sites or the preorganization of the protein, methods like BioMetAll[180] or GaudiMM[91,85] could be a good option to incorporate as a previous step in the workflow. In the case of the MD simulations, we opted for a bonded model for the Pt-donor coordination bond, whose forcefield parameters were calculated at the DFT level. Note here that this model assumes the hypothesis that the coordination bonds obtained in the previous docking step effectively exist and are stable.

Other considerations regarding the workflow are its initial and final steps. The input of this workflow requires three-dimensional structures of both the protein and the metallodrug. In the case of this study, we used the structure of insulin contained in the PDB entry 1zni,[291] and extracted the oxaliplatin structure from the PDB entry 4s18.[292] However, sometimes there is no availability of experimental structures that can be used to start the workflow. In these cases, an option could be to include a previous modeling step to obtain such structures. For the protein, we could employ tools like AlphaFold2[293] or RoseTTA Fold,[294] which have demonstrated high accuracy in the prediction of a protein structure from its sequence. For the metallodrug, the best option would be to parameterize and optimize the structure at the QM level. Finally, note that we included a last step of QM/MM optimization after the MD simulations. This step served two objectives: first, to refine the most representative structures of the oxaliplatin-insulin complex obtained from the MD simulations, ensuring as much as possible their accuracy; and second, to confirm the stability and geometry of the Pt(II) first coordination sphere, which was included in the QM part of the model.

## Main results obtained

For the docking step we assumed that, in aqueous solution and at physiological pH, the oxaliplatin loses the weak oxalate ligand, therefore forming $[Pt^{II}(dach)(OH)_2]$[a], where one or both $OH^-$ ions can be replaced by an amino acid from insulin. We tested both hypothesis (one or two coordinating amino acids), which resulted in two main binding modes (namely α and β): one where the Pt ion coordinates with His5B and Cys7B (mode α), and the other where the Pt ion only coordinates with His10B and one $OH^-$ of oxaliplatin remains (mode β). An additional mode γ was also considered: the Pt ion binds to His5B and Cys7B as in mode α, but the binding favors the reduction of the interchain disulphide bridge between Cys7A and Cys7B. Altogether, the docking results were in coherence with the available experimental data by Møller and coworkers.[295]

The analysis of the MD simulations revealed a highly stable folding of the insulin molecule for the binding mode α, probably due to the three disulphide bridges that connect the helices. For the case of mode β, the simulation switches between two sub-states, with the $OH^-$ moiety of the oxaliplatin forming and breaking a hydrogen bond with Glu13. The overall scaffold of the protein, although more flexible than in the case of mode α, is conserved. Interestingly, the simulation of mode γ showed big conformational changes in the insulin scaffold: the unfolding of the chain B containing the $[Pt^{II}(dach)]^{2+}$ moiety, and a reorientation of one of the helices in chain A (Gly1–Thr8).

Finally, we performed a cluster analysis on the three obtained trajectories. The most representative structure of each binding mode was submitted to a QM/MM optimization, obtaining as a result geometries of the first coordination sphere of Pt that appear coherent with available crystallized structures of oxaliplatin in complex with other proteins.

[a] Where "dach" is 1R,2R-diaminocyclohexane

## 8.2 Enantioselectivity in a cyclopropanation reaction catalyzed by an artificial metalloenzyme

Artificial metalloenzymes are a family of biocatalysts obtained by the insertion of homogeneous catalysts into proteic hosts,[296,297] which have the advantages over other de novo enzyme design of bringing a stable first coordination sphere of the metal and facilitating the engineering of the protein with the same techniques as other enzymes (e.g. mutagenesis and directed evolution).

A wide variety of protein scaffolds have been used for metalloenzyme design.[298] In this work,[277] we chose to study the cyclopropanase that J. Lewis and coworkers developed recently,[299,300] due to its scientific and methodological interest. The enzymes object of study here catalyze the enantioselective cyclopropanation of styrene and are built by a dirhodium homogeneous catalyst (Figure 8.3C) covalently anchored into *Pyrococcus furiosus* prolyl oligopeptidase (POP, Figure 8.3A). Three variants were constructed with different enantioselective profiles by playing with cofactor anchoring and amino acids mutations (Figure 8.3B): GSH, with 92% ee for the S,R enantiomer, HFF, with 92% ee for the S,R enantiomer, and RFY, with 80% ee for the R,S enantiomer. In their work, Lewis and coworkers suggest that the global motion of the protein scaffold could drive the changes in the enantioselective profile between the different mutants by altering the dynamics of the opening of the interdomain region.[301–303] Therefore, this system offered us an excellent framework to apply our multiscale philosophy, also integrating new approaches that can deal with global motions of the protein and substrate diffusion.

### Computational workflow employed

In the study of biochemical systems where a reaction occurs, it is common (and often indispensable) to start the workflow (Figure 8.4) by ascertaining the energetic pathway of the reaction, in this case a cyclopropanation catalyzed by the

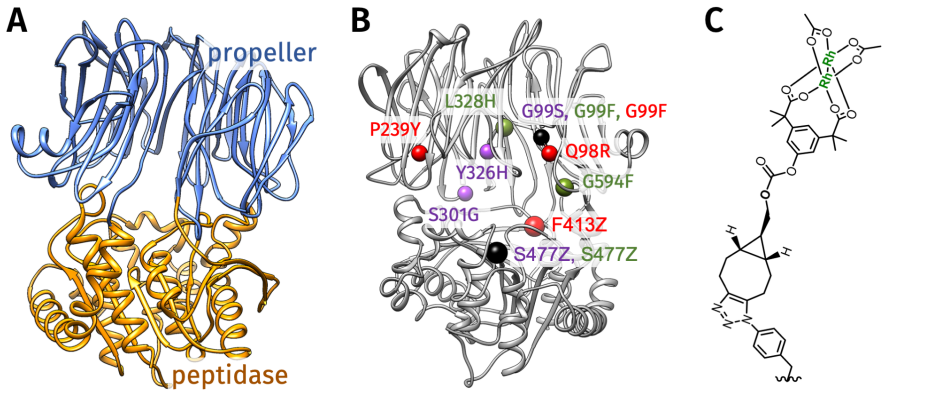



**Figure 8.3:** Biochemical system object of study. **(A)** Overall scaffold of the prolyl oligopeptidase enzyme. POP is a protein with a two-domain architecture: a peptidase domain with an α/β-hydrolase fold capped by a seven-bladed β-propeller domain. **(B)** Amino acids which were mutated in the different variants studied: GSH (in purple), HFF (in green), and RFY (in red). **(C)** Amino acids which were mutated in the different variants studied: GSH (in purple), HFF (in green), and RFY (in red). Structure of the covalent dirhodium cofactor that catalyzes the reaction (named Z in the mutation scheme). Figure reproduced from Ref.[277]

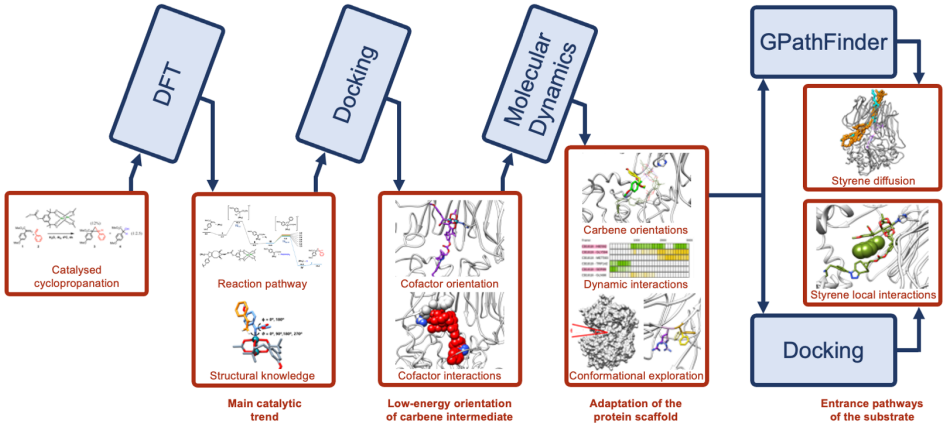



**Figure 8.4:** Computational workflow employed.

dirhodium cofactor. Our option in this case was to make a DFT model in water of the metallic cofactor and the two substrates of the reaction, without including any contribution of the protein at this stage. Another option could have been to include the entire biochemical system in a multiscale QM/MM model. In any case, the main goal of this first step is to obtain reliable information about the intermediate structures that could be of interest in the reaction pathway, and we valorated that the contribution of the protein was not necessary here. As a result , we obtained the Gibbs energy profile of the reaction, which revealed two key intermediates (Figure 8.5), named "intermediate II" and intermediate "III" in the article. The formation of the rhodium carbenoid (intermediate II) is the rate determining step of the reaction, while the formation of the rhodium carbenoid-styrene adduct (intermediate III) is the enantio determining step.

Once obtained the main catalytic trend of the reaction and the key intermediate structures, the second part of the workflow is aimed at the study of their interaction with the protein. There are multiple options to follow at this stage, and the concrete approach could depend on the experimental information available, the results obtained in the previous step of the workflow, and the intuition of the modeler, among other factors. In our case, we followed the logic of the reaction

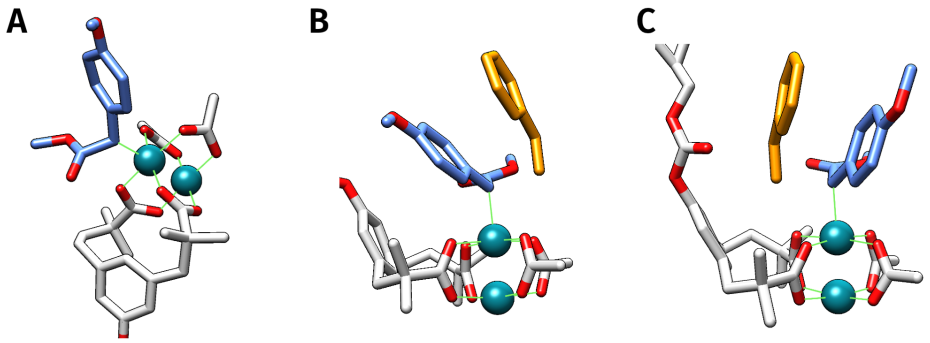



**Figure 8.5:** Key intermediates of the reaction. **(A)** The formation of the rhodium carbenoid (intermediate II) is the rate determining step of the reaction. Carbene is shown in blue, while the head of the cofactor is shown in gray sticks. **(B)** The formation of the carbenoid-styrene adduct is the enantio determining step. In this case, the intermediate III leading to the S,R enantiomer of the product is represented. Carbene is shown in blue, while styrene is shown in orange. **(C)** Intermediate III leading to the R,S enantiomer of the product.

pathway, first modeling the rhodium carbenoid into the protein, and then proceeding with the incorporation of the styrene to study the enantio determining step.

For the intermediate II, we employed a docking + MD approach, to obtain first a three-dimensional model and then a dynamic conformational sampling of the best-scored docking poses. Note that these simulations, in particular the dockings, faced several challenges related to the inorganic and covalent nature of the cofactor, and the large space needed to sample. In fact, we needed to split the docking stage into two separate calculations: first, with the help of the GaudiMM extension for metal docking,[85] we allowed full flexibility to the dirhodium cofactor, and sampled several global conformations of the POP scaffold while scanning the local flexibility of all the amino acids that could interact (i.e. form a coordination bond) with the rhodium atoms. It allowed us to obtain the best candidate structures assuming that a coordination bond between the cofactor and the protein would favor the reaction. These candidate structures were submitted to further refinement with the GOLD approach for metal docking that was previously developed in our group.[84] In this case, a much more restricted space was explored, because the objective was to obtain the more accurate structures that were further submitted to the MD procedure.

Finally, for the study of the enantio determining step, we aimed at evaluating the accessibility for the styrene to the two faces of the carbene (Figure 8.5B,C). Often, the modeling of an enzyme focuses only on the active site, but interactions with amino acids along the binding pathway of a substrate can be also of great importance. Therefore, we decided to explore the whole binding pathway of the styrene molecule with GPathFinder[177] in all the variants of the enzyme. For the variant where the binding pathways did not show a clear tendency (HFF), we opted to complement the modeling with a more traditional approach, ascertaining which of the two poses was more favored with docking.

### Main results obtained

Regarding the first part of the workflow, we obtained the reaction profile and optimized structures of the key intermediates of the reaction, which were used as starting points for the subsequent steps of the workflow. The energy barriers for the carbene and cyclopropanoid formation were estimated at about 19 and 10 kcal/mol, respectively.

The study of the carbenoid formation inside the enzyme showed that some amino acids mutated in the variants have a direct interaction with the cofactor and/or the substrate either in form of a coordination bond (His326 and His328 for GSH and HFF, respectively) or a strong hydrophobic patch between Phe99 and Arg98 and the carbene in RFY. Differences between variants also highlight global rearrangements of the protein scaffold, in particular regarding its inter-domain region.

Finally, based on the hypothesis that these motions could affect substrate (styrene) binding, we performed GPathFinder calculations in the three variants and analyzed their propensity to lead to S,R and R,S pre-reactive geometries. This hypothesis appears valid for GSH and RFY, where the enantioselective trends agree with those observed experimentally. For HFF, though, styrene binding pathways are not explicative and additional dockings were necessary to show that the experimental ee may relate to the organization of the second coordination sphere.

## 8.3 Chapter conclusions and future work

In this chapter we have seen how the combination of various modeling approaches can help to provide a comprehensive understanding of complex processes, such as drug-target interactions and enzyme-catalyzed reactions. The explicit treatment of the metal atoms throughout the workflows was an additional level of complexity that we overcame with the use and customization of methodologies previously developed in the group.

Some parts of the workflows can be easily extrapolated to other systems. For example, the use of docking+MD to ascertain the interactions between the chemical and biological parts and the adaptation of the protein. Other parts are more tailored to the particularities of the systems studied here, but some trends can be drawn, such as the importance of taking into account interactions outside the active site of the protein, or the need to sample the global motions of the protein.

Future perspectives for these types of protocols involve dealing with the interrelationships between the different steps. As having a unique energetic framework for the entire system is not possible, mechanisms to control the risk of losing information that could dramatically affect other levels of the workflow is something that deserves further attention.

# 9

# General conclusions

This thesis aimed at expanding the limits of current state-of-the-art methodologies for the exploration of biochemical spaces, while optimizing existing approaches and integrating them into multiscale workflows for the study of bioinorganic systems. Such objectives have been reached throughout a series of developments, show-cases and applications on real research cases. Altogether, it gives proof that methods with relatively simple algorithmic grounds can provide useful information about the molecular mechanisms of very complex systems. As a consequence of working in this scenario, where we rely on non-exhaustive sampling, a great part of the efforts has been devoted to benchmark and optimize the parametrization of these tools.

On the side of the developments, the achievements can be summarized in the following four points:

1. GPathFinder has been presented as a well-balanced tool for the exploration of ligand binding pathways in proteins. A typical calculation can be run in a few hours on a desktop computer, while the accuracy was validated in the widest benchmark performed so far in this kind of methodology. GPathFinder integration with the multi-objective capacities of the GaudiMM environment gives versatility to guide the exploration, from simple geometric descriptors to simplified energetic evaluation at the molecular mechanics level. In this sense, the modular architecture of the implementation opens the avenue for the incorporation of other descriptors, with the ultimate goal of simulating the binding of inorganic compounds.

2. As a previous step for including the exploration of the chemical space, the GaudiMM core has been actualized to incorporate a selection algorithm capable of managing many-objective optimization, and to integrate the RDKit chemistry library. Also, a module for the knowledge-based generation of ligand conformers have been included, and the whole code was updated to the new version of Python and unnecessary dependencies were removed.

3. GAlkemist has been presented as a novel approach for the exploration of the chemical space of small molecules, coupled with the conformational exploration of the system. Together with the multi-objective capabilities of the GaudiMM environment, GAlkemist provides useful support for molecule discovery processes based on the generation of informed hypotheses, as it has been shown in an illustrative case. A future horizon of this program will be its application in real research cases.

4. BioMetAll has been presented as an approach for the prediction of metal-binding sites in proteins. Based on simple geometric descriptors of the protein backbone, the methodology outperforms all state-of-the-art methods, as was demonstrated in a wide benchmark and three show-cases on cutting-edge applications. Due to its speed, BioMetAll allows for the screening of large conformational ensembles. A future direction for this project is to implement an evolved scoring function explicitly accounting with the metal type and integrate it in the GaudiMM environment.

On the side of the application in real case scenarios, the achievements can be summarized in the following three points:

1. A beta version of GPathFinder was applied to the discovery of the exit route of the glucose product of a hydrolysis reaction in a *Hordeum* exo-hydrolase. The GPathFinder calculation ultimately contributed to the proposal of a "substrate-product assisted processive" catalytic mechanism for this enzyme, which was further experimentally validated by other co authors of this work.

2. In the framework of several collaborations with national and international groups, the capabilities of GaudiMM were applied for the task of non-standard dockings. These cases were difficult or impossible to tackle with more traditional approaches and ultimately lead to results that shed light on the molecular mechanisms of the three involved systems: polyfluoroalkyl $sp^2$-iminosugar glycolipid into the p38 mitogen activated protein kinase, disaccharide complexes into YKL-39 and hHyal-1 enzymes, and vitamin $B_{12}$ into sugarcane-derived activated carbon.

3. Multiscale workflows were optimized and applied for the study of two types of bioinorganic systems: metallodrug interactions with a protein host, and molecular mechanisms of an artificial metalloenzyme. In those cases where the coordination sphere of the metal could eventually adapt during the binding process, these multiscale workflows, combining QM or QM/MM optimization with methods based on molecular mechanics evaluation, such as protein-ligand docking, GPathFinder, and classical molecular dynamics, demonstrated to be an efficient framework to better understand the binding mechanism.

# References


[1] Leslie A. Pray. "Discovery of DNA structure and function: Watson and Crick". In: *Nature Education* 1.1 (2008), p. 100.

[2] Rosalind E. Franklin and R. G. Gosling. "Molecular configuration in sodium thymonucleate". In: *Nature* 171.4356 (Apr. 1953), pp. 740–741. DOI: 10.1038/171740a0.

[3] Karen J. Edwards et al. "Molecular structure of the B-DNA dodecamer d(CGCAAATTTGCG)2 An examination of propeller twist and minor-groove water structure at 2.2Å resolution". In: *Journal of Molecular Biology* 226.4 (Aug. 1992), pp. 1161–1173. DOI: 10.1016/0022-2836(92)91059-X.

[4] "Crystallography: Protein Data Bank". In: *Nature New Biology* 233.42 (Oct. 1971), pp. 223–223. DOI: 10.1038/newbio233223b0.

[5] A. M. Helmenstine. *How many atoms there are in the human body*. Aug. 2021. URL: thoughtco.com/how-many-atoms-are-in-human-body-603872 (visited on 07/25/2022).

[6] *Top 500, the list*. URL: https://www.top500.org/lists/top500/2021/11/ (visited on 04/21/2022).

[7] Christian M. Heckmann and Francesca Paradisi. "Looking back: a short history of the discovery of enzymes and how they became powerful chemical tools". In: *ChemCatChem* 12.24 (Dec. 2020), pp. 6082–6102. DOI: 10.1002/cctc.202001107.

[8] Vikas Nanda and Ronald L. Koder. “Designing artificial enzymes by intuition and computation”. In: *Nature Chemistry* 2.1 (Jan. 2010), pp. 15–24. DOI: 10.1038/nchem.473.

[9] Frances H. Arnold. “Directed evolution: bringing new chemistry to life”. In: *Angewandte Chemie International Edition* 57.16 (Apr. 2018), pp. 4143–4148. DOI: 10.1002/anie.201708408.

[10] Hein J. Wijma and Dick B. Janssen. “Computational design gains momentum in enzyme catalysis engineering”. In: *FEBS Journal* 280.13 (July 2013), pp. 2948–2960. DOI: 10.1111/febs.12324.

[11] Maria P Frushicheva et al. “Computer aided enzyme design and catalytic concepts”. In: *Current Opinion in Chemical Biology* 21 (Aug. 2014), pp. 56–62. DOI: 10.1016/j.cbpa.2014.03.022.

[12] Emanuele Monza et al. “Molecular modeling in enzyme design, toward in silico guided directed evolution”. In: *Directed Enzyme Evolution: Advances and Applications*. Ed. by Miguel Alcalde. Cham: Springer International Publishing, 2017, pp. 257–284.

[13] Javier M. González and S. Zoë Fisher. “Structural analysis of ibuprofen binding to human adipocyte fatty-acid binding protein (FABP4)”. In: *Acta Crystallographica Section F Structural Biology Communications* 71.2 (Feb. 2015), pp. 163–170. DOI: 10.1107/S2053230X14027897.

[14] Masato Furuhashi and Gökhan S. Hotamisligil. “Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets”. In: *Nature Reviews Drug Discovery* 7.6 (June 2008), pp. 489–503. DOI: 10.1038/nrd2589.

[15] Marcin Trojnar et al. “Associations between fatty acid-binding protein 4–A proinflammatory adipokine and insulin resistance, gestational and type 2 diabetes mellitus”. In: *Cells* 8.3 (Mar. 2019), p. 227. DOI: 10.3390/cells8030227.

[16] Hao Kuang et al. “Enantioselective reductive amination of -keto acids to -amino acids by a pyridoxamine cofactor in a protein cavity”. In: *Journal of the American Chemical Society* 118.44 (Jan. 1996), pp. 10702–10706. DOI: 10.1021/ja954271z.

[17] Ronald R Davies et al. “Artificial metalloenzymes based on protein cavities: exploring the effect of altering the metal ligand attachment position by site directed mutagenesis”. In: *Bioorganic & Medicinal Chemistry Letters* 9.1 (1999), pp. 79–84. DOI: 10.1016/s0960-894x(98)00684-2.

[18] Dietmar Häring et al. “Converting a fatty acid binding protein to an artificial transaminase: novel catalysts by chemical and genetic modification of a protein cavity”. In: *Journal of Molecular Catalysis B: Enzymatic* 11.4-6 (Jan. 2001), pp. 967–970. DOI: 10.1016/S1381-1177(00)00051-5.

[19] Dietmar Häring and Mark D. Distefano. “Enzymes by design: chemogenetic assembly of transamination active sites containing lysine residues for covalent catalysis”. In: *Bioconjugate Chemistry* 12.3 (May 2001), pp. 385–390. DOI: 10.1021/bc000117c.

[20] Jonathan Crowe and Tony Bradshaw. *Chemistry for the biosciences: the essential concepts*. 2nd ed. New York: Oxford University Press, 2010.

[21] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev. “H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations”. In: *Nucleic Acids Research* 40.W1 (July 2012), W537–W541. DOI: 10.1093/nar/gks375.

[22] Eric F. Pettersen et al. “UCSF Chimera - A visualization system for exploratory research and analysis”. In: *Journal of Computational Chemistry* 25.13 (2004), pp. 1605–1612. DOI: 10.1002/jcc.20084.

[23] Bruce Alberts. *Molecular biology of the cell*. Seventh edition. New York: W. W. Norton & Company, 2022.

[24] Thomas Prohaska et al. “Standard atomic weights of the elements 2021 (IUPAC Technical Report)”. In: *Pure and Applied Chemistry* (May 2022). DOI: 10.1515/pac-2019-0603.

[25] Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. “The art and practice of structure-based drug design: a molecular modeling perspective”. In: *Medicinal Research Reviews* 16.1 (Jan. 1996), pp. 3–50. DOI: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.

[26] Gisbert Schneider. “Automating drug discovery”. In: *Nature Reviews Drug Discovery* 17.2 (Feb. 2018), pp. 97–113. DOI: 10.1038/nrd.2017.232.

[27] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. "Inverse molecular design using machine learning: generative models for matter engineering". In: *Science* 361.6400 (July 2018), pp. 360–365. DOI: 10.1126/science.aat2663.

[28] Addis S. Fuhr and Bobby G. Sumpter. "Deep generative models for materials discovery and machine learning-accelerated innovation". In: *Frontiers in Materials* 9 (Mar. 2022), p. 865270. DOI: 10.3389/fmats.2022.865270.

[29] Jean-Louis Reymond. "The chemical space project". In: *Accounts of Chemical Research* 48.3 (Mar. 2015), pp. 722–730. DOI: 10.1021/ar500432k.

[30] Youjun Xu et al. "Deep learning for molecular generation". In: *Future Medicinal Chemistry* 11.6 (Mar. 2019), pp. 567–597. DOI: 10.4155/fmc-2018-0358.

[31] Daniel C. Elton et al. "Deep learning for molecular design—a review of the state of the art". In: *Molecular Systems Design & Engineering* 4.4 (2019), pp. 828–849. DOI: 10.1039/C9ME00039A.

[32] Yu Cheng et al. "Molecular design in drug discovery: a comprehensive review of deep generative models". In: *Briefings in Bioinformatics* 22.6 (Nov. 2021), bbab344. DOI: 10.1093/bib/bbab344.

[33] R. C. Glen and A. W. R. Payne. "A genetic algorithm for the automated generation of molecules within constraints". In: *Journal of Computer-Aided Molecular Design* 9.2 (Apr. 1995), pp. 181–202. DOI: 10.1007/BF00124408.

[34] Nathan Brown et al. "A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules". In: *Journal of Chemical Information and Computer Sciences* 44.3 (May 2004), pp. 1079–1087. DOI: 10.1021/ci034290p.

[35] Jan H. Jensen. "A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space". In: *Chemical Science* 10.12 (2019), pp. 3567–3572. DOI: 10.1039/C8SC05372C.

[36] Emilie S. Henault, Maria H. Rasmussen, and Jan H. Jensen. "Chemical space exploration: how genetic algorithms find the needle in the

haystack". In: *PeerJ Physical Chemistry* 2 (July 2020), e11. DOI: 10.7717/peerj-pchem.11.

[37] Joshua Meyers, Benedek Fabian, and Nathan Brown. "De novo molecular design and generative models". In: *Drug Discovery Today* 26.11 (Nov. 2021), pp. 2707–2715. DOI: 10.1016/j.drudis.2021.05.019.

[38] Naruki Yoshikawa et al. "Population-based de novo molecule generation, using grammatical evolution". In: *Chemistry Letters* 47.11 (Nov. 2018), pp. 1431–1434. DOI: 10.1246/cl.180665.

[39] Robin Winter et al. "Efficient multi-objective molecular optimization in a continuous latent space". In: *Chemical Science* 10.34 (2019), pp. 8016–8024. DOI: 10.1039/C9SC01928F.

[40] Xiao Qing Lewell et al. "RECAP Retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry". In: *Journal of Chemical Information and Computer Sciences* 38.3 (May 1998), pp. 511–522. DOI: 10.1021/ci970429i.

[41] Nicholas C. Firth et al. "MOARF, an integrated workflow for multiobjective optimization: implementation, synthesis, and biological evaluation". In: *Journal of Chemical Information and Modeling* 55.6 (June 2015), pp. 1169–1180. DOI: 10.1021/acs.jcim.5b00073.

[42] Niclas Ståhl et al. "Deep reinforcement learning for multiparameter optimization in *de novo* drug design". In: *Journal of Chemical Information and Modeling* 59.7 (July 2019), pp. 3166–3176. DOI: 10.1021/acs.jcim.9b00325.

[43] H. Maarten Vinkers et al. "SYNOPSIS: SYNthesize and OPtimize System in Silico". In: *Journal of Medicinal Chemistry* 46.13 (June 2003), pp. 2765–2773. DOI: 10.1021/jm030809x.

[44] Alexander Button et al. "Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis". In: *Nature Machine Intelligence* 1.7 (July 2019), pp. 307–315. DOI: 10.1038/s42256-019-0067-7.

[45] Ksenia Korovina et al. "ChemBO: bayesian optimization of small organic molecules with synthesizable recommendations". In: *Proceedings of the*

*Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by Silvia Chiappa and Roberto Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, Aug. 2020, pp. 3393–3403.

[46] Melchor Sánchez Martínez. "Protein Flexibility: from local to global motions. A computational study". PhD thesis. Universitat de Barcelona, 2014. URL: http://hdl.handle.net/2445/64883.

[47] Maxim V. Shapovalov and Roland L. Dunbrack. "A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions". In: *Structure* 19.6 (June 2011), pp. 844–858. DOI: 10.1016/j.str.2011.03.019.

[48] Burak T. Kaynak et al. "Sampling of protein conformational space using hybrid simulations: a critical assessment of recent methods". In: *Frontiers in Molecular Biosciences* 9 (Feb. 2022), p. 832847. DOI: 10.3389/fmolb.2022.832847.

[49] Yinglong Miao, Victoria A. Feher, and J. Andrew McCammon. "Gaussian Accelerated Molecular Dynamics: unconstrained enhanced sampling and free energy calculation". In: *Journal of Chemical Theory and Computation* 11.8 (Aug. 2015), pp. 3584–3595. DOI: 10.1021/acs.jctc.5b00436.

[50] Ivet Bahar et al. "Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins". In: *Chemical Reviews* 110.3 (Mar. 2010), pp. 1463–1497. DOI: 10.1021/cr900095e.

[51] Osamu Miyashita, Florence Tama, and Pablo Chacón. "Normal mode analysis techniques in structural biology". en. In: 1st ed. Wiley, Oct. 2014.

[52] Roland L Dunbrack. "Rotamer libraries in the 21st century". In: *Current Opinion in Structural Biology* 12.4 (Aug. 2002), pp. 431–440. DOI: 10.1016/S0959-440X(02)00344-5.

[53] Roland L. Dunbrack and Martin Karplus. "Backbone-dependent rotamer library for proteins application to side-chain prediction". In: *Journal of Molecular Biology* 230.2 (Mar. 1993), pp. 543–574. DOI: 10.1006/jmbi.1993.1170.

[54] Roland L. Dunbrack and Fred E. Cohen. "Bayesian statistical analysis of protein side-chain rotamer preferences". In: *Protein Science* 6.8 (Aug. 1997), pp. 1661–1681. DOI: 10.1002/pro.5560060807.

[55] Jay W. Ponder and Frederic M. Richards. "Tertiary templates for proteins". In: *Journal of Molecular Biology* 193.4 (Feb. 1987), pp. 775–791. DOI: 10.1016/0022-2836(87)90358-5.

[56] Simon C. Lovell et al. "The penultimate rotamer library". In: *Proteins: Structure, Function, and Genetics* 40.3 (Aug. 2000), pp. 389–408. DOI: 10.1002/1097-0134(20000815)40:3<389::AID-PROT50>3.0.CO;2-2.

[57] Zhexin Xiang and Barry Honig. "Extending the accuracy limits of prediction for side-chain conformations". In: *Journal of Molecular Biology* 311.2 (Aug. 2001), pp. 421–430. DOI: 10.1006/jmbi.2001.4865.

[58] Radu Iftimie, Peter Minary, and Mark E. Tuckerman. "*Ab initio* molecular dynamics: concepts, recent developments, and future trends". In: *Proceedings of the National Academy of Sciences* 102.19 (May 2005), pp. 6654–6659. DOI: 10.1073/pnas.0500193102.

[59] A. Warshel and M. Levitt. "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme". In: *Journal of Molecular Biology* 103.2 (May 1976), pp. 227–249. DOI: 10.1016/0022-2836(76)90311-9.

[60] Martin J. Field, Paul A. Bash, and Martin Karplus. "A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations". In: *Journal of Computational Chemistry* 11.6 (July 1990), pp. 700–733. DOI: 10.1002/jcc.540110605.

[61] Donghong Min et al. "Enhancing QM/MM molecular dynamics sampling in explicit environments via an orthogonal-space-random-walk-based strategy". In: *The Journal of Physical Chemistry B* 115.14 (Apr. 2011), pp. 3924–3935. DOI: 10.1021/jp109454q.

[62] Sereina Riniker and Gregory A. Landrum. "Better informed distance geometry: using what we know to improve conformation generation". In: *Journal of Chemical Information and Modeling* 55.12 (Dec. 2015), pp. 2562–2574. DOI: 10.1021/acs.jcim.5b00654.

[63] Jason C. Cole et al. "Knowledge-based conformer generation using the cambridge structural database". In: *Journal of Chemical Information and Modeling* 58.3 (Mar. 2018), pp. 615–629. DOI: 10.1021/acs.jcim.7b00697.

[64] Shuzhe Wang et al. “Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences”. In: *Journal of Chemical Information and Modeling* 60.4 (Apr. 2020), pp. 2044–2058. DOI: 10.1021/acs.jcim.0c00025.

[65] D. E. Koshland. “Application of a theory of enzyme specificity to protein synthesis”. In: *Proceedings of the National Academy of Sciences* 44.2 (Feb. 1958), pp. 98–104. DOI: 10.1073/pnas.44.2.98.

[66] Buyong Ma et al. “Folding funnels and binding mechanisms”. In: *Protein Engineering, Design and Selection* 12.9 (Sept. 1999), pp. 713–720. DOI: 10.1093/protein/12.9.713.

[67] Austin D. Vogt and Enrico Di Cera. “Conformational selection is a dominant mechanism of ligand binding”. In: *Biochemistry* 52.34 (Aug. 2013), pp. 5723–5729. DOI: 10.1021/bi400929b.

[68] Austin D. Vogt et al. “Essential role of conformational selection in ligand binding”. In: *Biophysical Chemistry* 186 (Feb. 2014), pp. 13–21. DOI: 10.1016/j.bpc.2013.09.003.

[69] Dieter Rehder. *Bioinorganic chemistry*. Oxford (GB): Oxford university press, 2014.

[70] Pengfei Li and Kenneth M. Merz. “Metal ion modeling using classical mechanics”. In: *Chemical Reviews* 117.3 (Feb. 2017), pp. 1564–1686. DOI: 10.1021/acs.chemrev.6b00440.

[71] Pengfei Li and Kenneth M. Merz. “MCPB.py: a Python based metal center parameter builder”. In: *Journal of Chemical Information and Modeling* 56.4 (2016), pp. 599–604. DOI: 10.1021/acs.jcim.5b00674.

[72] Mathieu Allard et al. “Incorporation of manganese complexes into xylanase: new artificial metalloenzymes for enantioselective epoxidation”. In: *ChemBioChem* 13.2 (Jan. 2012), pp. 240–251. DOI: 10.1002/cbic.201100659.

[73] Jon I. Mujika et al. “Elucidating the 3D structures of Al(III)–Aβ complexes: a template free strategy based on the pre-organization hypothesis”. In: *Chemical Science* 8.7 (2017), pp. 5041–5049. DOI: 10.1039/C7SC01296A.

[74] Victor Muñoz Robles et al. “What can molecular modelling bring to the design of artificial inorganic cofactors?” In: *Faraday Discuss.* 148 (2011), pp. 137–159. DOI: 10.1039/C004578K.

[75] Lur Alonso-Cotchico et al. “Molecular modeling for artificial metalloenzyme design and optimization”. In: *Accounts of Chemical Research* 53.4 (Apr. 2020), pp. 896–905. DOI: 10.1021/acs.accounts.0c00031.

[76] Giuseppe Sciortino et al. “Elucidation of binding site and chiral specificity of oxidovanadium drugs with lysozyme through theoretical calculations”. In: *Inorganic Chemistry* 56.21 (Nov. 2017), pp. 12938–12951. DOI: 10.1021/acs.inorgchem.7b01732.

[77] Giuseppe Sciortino et al. “Decoding surface interaction of $V^{IV}O$ metallodrug candidates with lysozyme”. In: *Inorganic Chemistry* 57.8 (Apr. 2018), pp. 4456–4469. DOI: 10.1021/acs.inorgchem.8b00134.

[78] Giuseppe Sciortino et al. “Integrated ESI-MS/EPR/computational characterization of the binding of metal species to proteins: vanadium drug–myoglobin application”. In: *Inorganic Chemistry Frontiers* 6.6 (2019), pp. 1561–1578. DOI: 10.1039/C9QI00179D.

[79] Giuseppe Sciortino et al. “Effect of secondary interactions, steric hindrance and electric charge on the interaction of V $^{IV}$O species with proteins”. In: *New Journal of Chemistry* 43.45 (2019), pp. 17647–17660. DOI: 10.1039/C9NJ01956A.

[80] Valeria Ugone et al. “Interaction of vanadium(IV) species with ubiquitin: a combined instrumental and computational approach”. In: *Inorganic Chemistry* 58.12 (June 2019), pp. 8064–8078. DOI: 10.1021/acs.inorgchem.9b00807.

[81] Daniele Sanna et al. “$V^{IV}O$ complexes with antibacterial quinolone ligands and their interaction with serum proteins”. In: *Dalton Transactions* 47.7 (2018), pp. 2164–2182. DOI: 10.1039/C7DT04216G.

[82] Giuseppe Sciortino et al. “Biospeciation of potential vanadium drugs of acetylacetonate in the presence of proteins”. In: *Frontiers in Chemistry* 8 (May 2020), p. 345. DOI: 10.3389/fchem.2020.00345.

[83] Giuseppe Sciortino, Eugenio Garribba, and Jean-Didier Maréchal. “Validation and Applications of Protein–Ligand Docking Approaches Im-

proved for Metalloligands with Multiple Vacant Sites". en. In: *Inorg. Chem.* 58.1 (Jan. 2019), pp. 294–306. DOI: 10.1021/acs.inorgchem.8b02374. URL: https://pubs.acs.org/doi/10.1021/acs.inorgchem.8b02374 (visited on 08/03/2022).

[84] Giuseppe Sciortino et al. "Prediction of the interaction of metallic moieties with proteins: an update for protein-ligand docking techniques". In: *Journal of Computational Chemistry* 39.1 (Jan. 2018), pp. 42–51. DOI: 10.1002/jcc.25080.

[85] Giuseppe Sciortino et al. "Simple coordination geometry descriptors allow to accurately predict metal-binding sites in proteins". In: *ACS Omega* 4.2 (Feb. 2019), pp. 3726–3731. DOI: 10.1021/acsomega.8b03457.

[86] Claudia Andreini et al. "MetalPDB: a database of metal sites in biological macromolecular structures". In: *Nucleic Acids Research* 41.D1 (Jan. 2013), pp. D312–D319. DOI: 10.1093/nar/gks1063.

[87] Valeria Putignano et al. "MetalPDB in 2018: a database of metal sites in biological macromolecular structures". In: *Nucleic Acids Research* 46.D1 (Jan. 2018), pp. D459–D464. DOI: 10.1093/nar/gkx989.

[88] Jerome Eberhardt et al. "AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings". In: *Journal of Chemical Information and Modeling* 61.8 (Aug. 2021), pp. 3891–3898. DOI: 10.1021/acs.jcim.1c00203.

[89] Nicholas Metropolis et al. "Equation of state calculations by fast computing machines". In: *The Journal of Chemical Physics* 21.6 (June 1953), pp. 1087–1092. DOI: 10.1063/1.1699114.

[90] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. 2nd ed. Springer series in operations research. OCLC: ocm68629100. New York: Springer, 2006.

[91] Jaime Rodríguez-Guerra Pedregal et al. "GaudiMM: a modular multiobjective platform for molecular modeling". In: *Journal of Computational Chemistry* 38.24 (Sept. 2017), pp. 2118–2126. DOI: 10.1002/jcc.24847.

[92] Jaime Rodríguez-Guerra Pedregal. "Development and application of a computational platform for complex molecular design". PhD thesis. Uni-

versitat Autònoma de Barcelona, 2018. URL: https://ddd.uab.cat/record/201498.

[93] Andrew R. Leach. *Molecular modelling: principles and applications*. 2nd ed. Harlow, England ; New York: Prentice Hall, 2001.

[94] Donald A. McQuarrie. *Quantum chemistry*. 2nd ed. Sausalito, Calif: University Science Books, 2008.

[95] P. W. Atkins and Ronald Friedman. *Molecular quantum mechanics*. 5th ed. Oxford ; New York: Oxford University Press, 2011.

[96] Matija Zlatar and Maja Gruden. "Introduction to ligand field theory and computational chemistry". In: *Practical Approaches to Biological Inorganic Chemistry*. Elsevier, 2020, pp. 17–67.

[97] W. Kohn and L. J. Sham. "Self-consistent equations including exchange and correlation effects". In: *Physical Review* 140.4A (Nov. 1965), A1133–A1138. DOI: 10.1103/PhysRev.140.A1133.

[98] Zhifeng Jing et al. "Polarizable force fields for biomolecular simulations: recent advances and applications". In: *Annual Review of Biophysics* 48.1 (May 2019), pp. 371–394. DOI: 10.1146/annurev-biophys-070317-033349.

[99] Judith A. Harrison et al. "Review of force fields and intermolecular potentials used in atomistic computational materials research". In: *Applied Physics Reviews* 5.3 (Sept. 2018), p. 031104. DOI: 10.1063/1.5020808.

[100] Paul Robustelli, Stefano Piana, and David E. Shaw. "Developing a molecular dynamics force field for both folded and disordered protein states". In: *Proceedings of the National Academy of Sciences* 115.21 (May 2018). DOI: 10.1073/pnas.1800690115.

[101] Oliver T. Unke et al. "Machine learning force fields". In: *Chemical Reviews* 121.16 (Aug. 2021), pp. 10142–10186. DOI: 10.1021/acs.chemrev.0c01111. (Visited on 08/09/2022).

[102] Viktor Hornak et al. "Comparison of multiple Amber force fields and development of improved protein backbone parameters". In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (Nov. 2006), pp. 712–725. DOI: 10.1002/prot.21123.

[103] James A. Maier et al. “ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB”. In: *Journal of Chemical Theory and Computation* 11.8 (Aug. 2015), pp. 3696–3713. DOI: 10.1021/acs.jctc.5b00255.

[104] Pekka Mark and Lennart Nilsson. “Structure and dynamics of the TIP3P, SPC, and SPC/E Water models at 298 K”. In: *The Journal of Physical Chemistry A* 105.43 (Nov. 2001), pp. 9954–9960. DOI: 10.1021/jp003020w.

[105] Junmei Wang et al. “Development and testing of a general amber force field”. In: *Journal of Computational Chemistry* 25.9 (July 2004), pp. 1157–1174. DOI: 10.1002/jcc.20035.

[106] Jorge M. Seminario. “Calculation of intramolecular force fields from second-derivative tensors”. In: *International Journal of Quantum Chemistry* 60.7 (1996), pp. 1271–1277. DOI: 10.1002/(SICI)1097-461X(1996)60:7<1271::AID-QUA8>3.0.CO;2-W.

[107] Pengfei Li and Kenneth M. Merz. “MCPB.py: a Python based metal center parameter builder”. In: *Journal of Chemical Information and Modeling* 56.4 (Apr. 2016), pp. 599–604. DOI: 10.1021/acs.jcim.5b00674.

[108] Ruth Nussinov. “The significance of the 2013 Nobel prize in chemistry and the challenges ahead”. In: *PLoS Computational Biology* 10.1 (Jan. 2014), e1003423. DOI: 10.1371/journal.pcbi.1003423.

[109] Johannes Grotendorst, Norbert Attig, and Institute for Advanced Simulation, eds. *Multiscale simulation methods in molecular sciences*. NIC series / John von Neumann Institute for Computing 42. Jülich: Supercomputing Centre, Forschungszentrum Jülich, 2009.

[110] Rimsha Mehmood and Heather J. Kulik. “Both configuration and QM region size matter: zinc stability in QM/MM models of DNA methyltransferase”. In: *Journal of Chemical Theory and Computation* 16.5 (May 2020), pp. 3121–3134. DOI: 10.1021/acs.jctc.0c00153.

[111] Heather J. Kulik et al. “How large should the QM region be in QM/MM calculations? The case of catechol O-Methyltransferase”. In: *The Journal of Physical Chemistry B* 120.44 (Nov. 2016), pp. 11381–11394. DOI: 10.1021/acs.jpcb.6b07814.

[112] Haitao Kang and Mingna Zheng. "Influence of the quantum mechanical region size in QM/MM modelling: a case study of fluoroacetate dehalogenase catalyzed C F bond cleavage". In: *Computational and Theoretical Chemistry* 1204 (Oct. 2021), p. 113399. DOI: 10.1016/j.comptc.2021.113399.

[113] Jaime Rodríguez-Guerra Pedregal et al. "GARLEEK: Adding an extra flavor to ONIOM". In: *Journal of Computational Chemistry* 40.2 (Jan. 2019), pp. 381–386. DOI: 10.1002/jcc.25612.

[114] M. J. Frisch et al. "Gaussian 09, Revision D.01". In: *Gaussian Inc.* (2009).

[115] Rohit V. Pappu, Reece K. Hart, and Jay W. Ponder. "Analysis and application of potential energy smoothing and search methods for global optimization". In: *The Journal of Physical Chemistry B* 102.48 (Nov. 1998), pp. 9725–9742. DOI: 10.1021/jp982255t.

[116] Joshua A. Rackers et al. "Tinker 8: software tools for molecular design". In: *Journal of Chemical Theory and Computation* 14.10 (Oct. 2018), pp. 5273–5289. DOI: 10.1021/acs.jctc.8b00529.

[117] Nawaf Bou-Rabee. "Time integrators for molecular dynamics". In: *Entropy* 16.1 (Dec. 2013), pp. 138–162. DOI: 10.3390/e16010138.

[118] Etienne Forest and Ronald D. Ruth. "Fourth-order symplectic integration". In: *Physica D: Nonlinear Phenomena* 43.1 (May 1990), pp. 105–117. DOI: 10.1016/0167-2789(90)90019-L.

[119] Alan Grossfield and Daniel M. Zuckerman. "Chapter 2. Quantifying uncertainty and sampling quality in biomolecular simulations". In: *Annual Reports in Computational Chemistry* 5 (2009), pp. 23–48. DOI: 10.1016/S1574-1400(09)00502-7.

[120] Alan Grossfield et al. "Best practices for quantification of uncertainty and sampling quality in molecular simulations [Article v1.0]". In: *Living Journal of Computational Molecular Science* 1.1 (2019). DOI: 10.33011/livecoms.1.1.5067.

[121] Lorna J. Smith, Xavier Daura, and Wilfred F. van Gunsteren. "Assessing equilibration and convergence in biomolecular simulations". In: *Proteins: Structure, Function, and Genetics* 48.3 (Aug. 2002), pp. 487–496. DOI: 10.1002/prot.10144.

[122] Bernhard Knapp, Luis Ospina, and Charlotte M. Deane. "Avoiding false positive conclusions in molecular simulation: the importance of replicas". In: *Journal of Chemical Theory and Computation* 14.12 (Dec. 2018), pp. 6127–6138. DOI: 10.1021/acs.jctc.8b00391.

[123] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. "Software for molecular docking: a review". In: *Biophysical Reviews* 9.2 (Apr. 2017), pp. 91–102. DOI: 10.1007/s12551-016-0247-1.

[124] Garrett M. Morris et al. "AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility". In: *Journal of Computational Chemistry* 30.16 (Dec. 2009), pp. 2785–2791. DOI: 10.1002/jcc.21256.

[125] Oleg Trott and Arthur J. Olson. "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading". In: *Journal of Computational Chemistry* 31.2 (2010), pp. 455–461. DOI: 10.1002/jcc.21334.

[126] Gareth Jones et al. "Development and validation of a genetic algorithm for flexible docking". In: *Journal of Molecular Biology* 267.3 (Apr. 1997), pp. 727–748. DOI: 10.1006/jmbi.1996.0897.

[127] Gareth Jones, Peter Willett, and Robert C. Glen. "Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation". In: *Journal of Molecular Biology* 245.1 (Jan. 1995), pp. 43–53. DOI: 10.1016/S0022-2836(95)80037-9.

[128] Marcel L Verdonk et al. "Improved protein–ligand docking using GOLD". In: *Proteins: Structure, Function, and Bioinformatics* 52.4 (Sept. 2003), pp. 609–623. DOI: 10.1002/prot.10465.

[129] Garrett M. Morris et al. "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function". In: *Journal of Computational Chemistry* 19.14 (Nov. 1998), pp. 1639–1662. DOI: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.

[130] Ruth Huey et al. "A semiempirical free energy force field with charge-based desolvation". In: *Journal of Computational Chemistry* 28.6 (Apr. 2007), pp. 1145–1152. DOI: 10.1002/jcc.20634.

[131] Wei P. Feinstein and Michal Brylinski. “Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets”. In: *Journal of Cheminformatics* 7.1 (Dec. 2015), p. 18. DOI: 10.1186/s13321-015-0067-5.

[132] John Baxter. “Local optima avoidance in depot location”. In: *The Journal of the Operational Research Society* 32.9 (Sept. 1981), p. 815. DOI: 10.2307/2581397.

[133] Christian Blum, ed. *Hybrid metaheuristics: an emerging approach to optimization*. Studies in computational intelligence v. 114. Berlin: Springer, 2008.

[134] Renxiao Wang, Luhua Lai, and Shaomeng Wang. “Further development and validation of empirical scoring functions for structure-based binding affinity prediction”. In: *Journal of Computer-Aided Molecular Design* 16.1 (2002), pp. 11–26. DOI: 10.1023/A:1016357811882.

[135] Oliver Korb, Thomas Stützle, and Thomas E. Exner. “Empirical scoring functions for advanced proteinligand docking with PLANTS”. In: *Journal of Chemical Information and Modeling* 49.1 (Jan. 2009), pp. 84–96. DOI: 10.1021/ci800298z.

[136] John W. Liebeschuetz, Jason C. Cole, and Oliver Korb. “Pose prediction and virtual screening performance of GOLD scoring functions in a standardized test”. In: *Journal of Computer-Aided Molecular Design* 26.6 (June 2012), pp. 737–748. DOI: 10.1007/s10822-012-9551-4.

[137] Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*. Reprint. Wiley paperback series. Chichester Weinheim: Wiley, 2009.

[138] André Krammer et al. “LigScore: a novel scoring function for predicting binding affinities”. In: *Journal of Molecular Graphics and Modelling* 23.5 (2005), pp. 395–407. DOI: 10.1016/j.jmgm.2004.11.007.

[139] Douglas B. Kitchen et al. “Docking and scoring in virtual screening for drug discovery: methods and applications”. In: *Nature Reviews Drug Discovery* 3.11 (Nov. 2004), pp. 935–949. DOI: 10.1038/nrd1549.

[140] B. Coupez and R. A. Lewis. "Docking and scoring - theoretically easy, practically impossible?" In: *Current Medicinal Chemistry* 13.25 (Oct. 2006), pp. 2995–3003. DOI: 10.2174/092986706778521797.

[141] Zhe Wang et al. "Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power". In: *Physical Chemistry Chemical Physics* 18.18 (2016), pp. 12964–12975. DOI: 10.1039/C6CP01555G.

[142] Jin Li, Ailing Fu, and Le Zhang. "An overview of scoring functions used for protein–ligand interactions in molecular docking". In: *Interdisciplinary Sciences: Computational Life Sciences* 11.2 (June 2019), pp. 320–328. DOI: 10.1007/s12539-019-00327-w.

[143] Christodoulos A Floudas and P. M Pardalos. *Optimization in computational chemistry and molecular biology: local and global approaches*. 1st ed. NY: Springer New York, 2000.

[144] David Clark and David E. Clark, eds. *Evolutionary algorithms in molecular design*. Methods and principles in medicinal chemistry 8. Weinheim: Wiley-VCH, 2000.

[145] Adam Slowik and Halina Kwasnicka. "Evolutionary algorithms and their applications to engineering problems". In: *Neural Computing and Applications* 32.16 (Aug. 2020), pp. 12363–12379. DOI: 10.1007/s00521-020-04832-8.

[146] John R. Koza. *Genetic programming: on the programming of computers by means of natural selection*. Complex adaptive systems. Cambridge, Mass: MIT Press, 1992.

[147] John R. Koza, ed. *Genetic programming. 4: routine human-competitive machine intelligence*. 1. paperback print. Genetic programming series 5. Cambridge, Mass: MIT Press, 2005.

[148] Kalyanmoy Deb and Hans-Georg Beyer. "Self-adaptive genetic algorithms with simulated binary crossover". In: *Evolutionary Computation* 9.2 (June 2001), pp. 197–221. DOI: 10.1162/106365601750190406.

[149] Linqiang Pan et al. "Adaptive simulated binary crossover for rotated multi-objective optimization". In: *Swarm and Evolutionary Computation* 60 (Feb. 2021), p. 100759. DOI: 10.1016/j.swevo.2020.100759.

[150] Kalyanmoy Deb and Debayan Deb. "Analysing mutation schemes for real-parameter genetic algorithms". In: *International Journal of Artificial Intelligence and Soft Computing* 4.1 (2014), pp. 1–28. DOI: 10.1504/IJAISC.2014.059280.

[151] Ahmad Hassanat et al. "Choosing mutation and crossover ratios for genetic algorithms—a review with a new dynamic approach". In: *Information* 10.12 (Dec. 2019), p. 390. DOI: 10.3390/info10120390.

[152] K. Deb et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2 (Apr. 2002), pp. 182–197. DOI: 10.1109/4235.996017.

[153] Kalyanmoy Deb and Himanshu Jain. "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints". In: *IEEE Transactions on Evolutionary Computation* 18.4 (Aug. 2014), pp. 577–601. DOI: 10.1109/TEVC.2013.2281535.

[154] Himanshu Jain and Kalyanmoy Deb. "An evolutionary many-objective optimization algorithm using reference-point based nondominated sorting approach, part II: handling constraints and extending to an adaptive approach". In: *IEEE Transactions on Evolutionary Computation* 18.4 (Aug. 2014), pp. 602–622. DOI: 10.1109/TEVC.2013.2281534.

[155] Agnieszka Krzemińska et al. "Theoretical studies of cyanophycin dipeptides as inhibitors of tyrosinases". In: *International Journal of Molecular Sciences* 23.6 (Mar. 2022), p. 3335. DOI: 10.3390/ijms23063335.

[156] Manik Das et al. "Effect of ancillary ligand on DNA and protein interaction of the two Zn (II) and Co (III) complexes: experimental and theoretical study". In: *Journal of Biomolecular Structure and Dynamics* (Nov. 2021), pp. 1–16. DOI: 10.1080/07391102.2021.2001377.

[157] Leire Dublang et al. "Inhibition of the human Hsc70 system by small ligands as a potential anticancer approach". In: *Cancers* 13.12 (June 2021), p. 2936. DOI: 10.3390/cancers13122936.

[158] Jacobo Gómez-González et al. "Selective recognition of A/T-rich DNA 3-way junctions with a three-fold symmetric tripeptide". In: *Chemical Communications* 58.56 (2022), pp. 7769–7772. DOI: 10.1039/D2CC02874C.

[159] Fadri Christoffel et al. “Design and evolution of chimeric streptavidin for protein-enabled dual gold catalysis”. In: *Nature Catalysis* 4.8 (Aug. 2021), pp. 643–653. DOI: 10.1038/s41929-021-00651-9.

[160] Soraya Learte-Aymamí et al. “Controlling oncogenic KRAS signaling pathways with a Palladium-responsive peptide”. In: *Communications Chemistry* 5.1 (Dec. 2022), p. 75. DOI: 10.1038/s42004-022-00691-7.

[161] Giuseppe Sciortino et al. “Integrated ESI-MS/EPR/computational characterization of the binding of metal species to proteins: vanadium drug–myoglobin application”. In: *Inorganic Chemistry Frontiers* 6.6 (2019), pp. 1561–1578. DOI: 10.1039/C9QI00179D.

[162] Vadim R. Viviani, Gabriel F. Pelentir, and Vanessa R. Bevilaqua. “Bioluminescence color-tuning firefly luciferases: engineering and prospects for real-time intracellular pH imaging and heavy metal biosensing”. In: *Biosensors* 12.6 (June 2022), p. 400. DOI: 10.3390/bios12060400.

[163] Nichael Lynn Cramer. “A representation for the adaptive generation of simple sequential programs”. In: *Proceedings of an International Conference on Genetic Algorithms and the Applications* 24-26 (July 1985), pp. 183–187.

[164] Riccardo Poli et al. *A field guide to genetic programming*. OCLC: 837998350. [Morrisville, NC: Lulu Press], 2008.

[165] John R Koza et al. “The design of analog circuits by means of genetic programming”. In: *Evolutionary design by computers*. Morgan Kaufmann, 1999, p. 365385.

[166] S. Sette and L. Boullart. “Genetic programming: principles and applications”. In: *Engineering Applications of Artificial Intelligence* 14.6 (Dec. 2001), pp. 727–736. DOI: 10.1016/S0952-1976(02)00013-1.

[167] Cândida Ferreira. *Gene expression programming*. 2nd ed. Studies in Computational Intelligence. Springer Berlin, Heidelberg, 2002.

[168] *Apache license, version 2.0*. URL: https://www.apache.org/licenses/LICENSE-2.0 (visited on 08/04/2022).

[169] *The 3-clause BSD license*. URL: https://opensource.org/licenses/BSD-3-Clause (visited on 08/04/2022).

[170] *Python*. URL: https://www.python.org (visited on 08/04/2022).

[171] *The Python package index*. URL: https://pypi.org (visited on 08/04/2022).

[172] *Conda*. URL: https://docs.conda.io/projects/conda/en/latest/ (visited on 08/04/2022).

[173] Jaime Rodríguez-Guerra Pedregal and Jean-Didier Maréchal. "PyChimera: use UCSF Chimera modules in any Python 2.7 project". In: *Bioinformatics* 34.10 (May 2018), pp. 1784–1785. DOI: 10.1093/bioinformatics/bty021.

[174] Peter Wang. *Why Python: the factors leading to the language's ascendancy*. URL: https://www.anaconda.com/blog/why-python (visited on 08/04/2022).

[175] *TIOBE programming community index*. URL: https://www.tiobe.com/tiobe-index (visited on 08/04/2022).

[176] *Top programming languages for data scientists in 2022*. URL: https://www.datacamp.com/blog/top-programming-languages-for-data-scientists-in-2022 (visited on 08/04/2022).

[177] José-Emilio Sánchez-Aparicio et al. "GPathFinder: identification of ligand-binding pathways by a multi-objective genetic algorithm". In: *International Journal of Molecular Sciences* 20.13 (Jan. 2019), p. 3155. DOI: 10.3390/ijms20133155.

[178] Felix-Antoine Fortin. "DEAP: evolutionary algorithms made easy". In: *Journal of Machine Learning Research* 13.1 (2012), pp. 2171–2175.

[179] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[180] José-Emilio Sánchez-Aparicio et al. "BioMetAll: identifying metal-binding sites in proteins from backbone preorganization". In: *Journal of Chemical Information and Modeling* 61.1 (Jan. 2021), pp. 311–323. DOI: 10.1021/acs.jcim.0c00827.

[181] Sergio M. Marques et al. "Enzyme tunnels and gates as relevant targets in drug design". In: *Medicinal Research Reviews* 37.5 (Sept. 2017), pp. 1095–1139. DOI: 10.1002/med.21430.

[182] Khoa N. Pham and Syun-Ru Yeh. "Mapping the binding trajectory of a suicide inhibitor in human indoleamine 2,3-dioxygenase 1". In: *Journal*

*of the American Chemical Society* 140.44 (Nov. 2018), pp. 14538–14541. DOI: 10.1021/jacs.8b07994.

[183] Phuc-Chau Do, Eric H. Lee, and Ly Le. "Steered molecular dynamics simulation in rational drug design". In: *Journal of Chemical Information and Modeling* 58.8 (Aug. 2018), pp. 1473–1482. DOI: 10.1021/acs.jcim.8b00261.

[184] Yinglong Miao, Apurba Bhattarai, and Jinan Wang. "Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics". In: *Journal of Chemical Theory and Computation* 16.9 (Sept. 2020), pp. 5526–5547. DOI: 10.1021/acs.jctc.0c00395.

[185] Eva Chovancova et al. "CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures". In: *PLoS Computational Biology* 8.10 (Oct. 2012). Ed. by Andreas Prlic, e1002708. DOI: 10.1371/journal.pcbi.1002708.

[186] Talha Bin Masood et al. "CHEXVIS: a tool for molecular channel extraction and visualization". In: *BMC Bioinformatics* 16.1 (Dec. 2015), p. 119. DOI: 10.1186/s12859-015-0545-9.

[187] David G. Levitt and Leonard J. Banaszak. "POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids". In: *Journal of Molecular Graphics* 10.4 (Dec. 1992), pp. 229–234. DOI: 10.1016/0263-7855(92)80074-N.

[188] Ashutosh Tripathi and Glen E. Kellogg. "A novel and efficient tool for locating and characterizing protein cavities and binding sites". In: *Proteins: Structure, Function, and Bioinformatics* 78.4 (Mar. 2010), pp. 825–842. DOI: 10.1002/prot.22608.

[189] J. Dundas et al. "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues". In: *Nucleic Acids Research* 34.Web Server (July 2006), W116–W118. DOI: 10.1093/nar/gkl282.

[190] David Sehnal et al. "MOLE 2.0: advanced approach for analysis of biomacromolecular channels". In: *Journal of Cheminformatics* 5.1 (Dec. 2013), p. 39. DOI: 10.1186/1758-2946-5-39.

[191] Eitan Yaffe et al. “MolAxis: efficient and accurate identification of channels in macromolecules”. In: *Proteins: Structure, Function, and Bioinformatics* 73.1 (Apr. 2008), pp. 72–86. DOI: 10.1002/prot.22052.

[192] Pratyush Tiwary et al. “Kinetics of protein–ligand unbinding: predicting pathways, rates, and rate-limiting steps”. In: *Proceedings of the National Academy of Sciences* 112.5 (Feb. 2015). DOI: 10.1073/pnas.1424461112.

[193] Ondrej Vavra et al. “CaverDock: a molecular docking-based tool to analyse ligand transport through protein tunnels and channels”. In: *Bioinformatics* 35.23 (Dec. 2019). Ed. by Yann Ponty, pp. 4986–4993. DOI: 10.1093/bioinformatics/btz386.

[194] P.-H. Lee et al. “SLITHER: a web server for generating contiguous conformations of substrate molecules entering into deep active sites of proteins or migrating through channels in membrane transporters”. In: *Nucleic Acids Research* 37.Web Server (July 2009), W559–W564. DOI: 10.1093/nar/gkp359.

[195] Didier Devaurs et al. “MoMA-LigPath: a web server to simulate protein–ligand unbinding”. In: *Nucleic Acids Research* 41.W1 (July 2013), W297–W302. DOI: 10.1093/nar/gkt380.

[196] Kenneth W. Borrelli et al. “PELE: protein energy landscape exploration. A novel Monte Carlo based technique”. In: *Journal of Chemical Theory and Computation* 1.6 (Nov. 2005), pp. 1304–1311. DOI: 10.1021/ct0501811.

[197] Oliver Carrillo and Modesto Orozco. “GRID-MD—A tool for massive simulation of protein channels”. In: *Proteins: Structure, Function, and Bioinformatics* 70.3 (Feb. 2008), pp. 892–899. DOI: 10.1002/prot.21592.

[198] Minh Khoa Nguyen, Léonard Jaillet, and Stéphane Redon. “ART-RRT: As-Rigid-As-Possible exploration of ligand unbinding pathways”. In: *Journal of Computational Chemistry* 39.11 (Apr. 2018), pp. 665–678. DOI: 10.1002/jcc.25132.

[199] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. “Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise”. In: *Journal of Chemical Information and Modeling* 53.8 (Aug. 2013), pp. 1893–1904. DOI: 10.1021/ci300604z.

[200] R. Leila Reynald et al. “Structural characterization of human cytochrome P450 2C19”. In: *Journal of Biological Chemistry* 287.53 (Dec. 2012), pp. 44581–44591. DOI: 10.1074/jbc.M112.424895.

[201] Victor A. Streltsov et al. “Discovery of processive catalysis by an exo-hydrolase with a pocket-shaped active site”. In: *Nature Communications* 10.1 (Dec. 2019), p. 2222. DOI: 10.1038/s41467-019-09691-z.

[202] Anna Gaulton et al. “The ChEMBL database in 2017”. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D945–D954. DOI: 10.1093/nar/gkw1074.

[203] Nicholas C. Firth et al. “MOARF, an integrated workflow for multiobjective optimization: implementation, synthesis, and biological evaluation”. In: *Journal of Chemical Information and Modeling* 55.6 (June 2015), pp. 1169–1180. DOI: 10.1021/acs.jcim.5b00073.

[204] Naruki Yoshikawa et al. “Population-based de novo molecule generation, using grammatical evolution”. In: *Chemistry Letters* 47.11 (Nov. 2018), pp. 1431–1434. DOI: 10.1246/cl.180665.

[205] Yongbeom Kwon and Juyong Lee. “MolFinder: an evolutionary algorithm for the global optimization of molecular properties and the extensive exploration of chemical space using SMILES”. In: *Journal of Cheminformatics* 13.1 (Dec. 2021), p. 24. DOI: 10.1186/s13321-021-00501-7.

[206] Nathan Brown et al. “A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules”. In: *Journal of Chemical Information and Computer Sciences* 44.3 (May 2004), pp. 1079–1087. DOI: 10.1021/ci034290p.

[207] Jan H. Jensen. “A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space”. In: *Chemical Science* 10.12 (2019), pp. 3567–3572. DOI: 10.1039/C8SC05372C.

[208] H. Maarten Vinkers et al. “SYNOPSIS: SYNthesize and OPtimize System in Silico”. In: *Journal of Medicinal Chemistry* 46.13 (June 2003), pp. 2765–2773. DOI: 10.1021/jm030809x.

[209] Jacob O. Spiegel and Jacob D. Durrant. “AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization”. In: *Jour-*

*nal of Cheminformatics* 12.1 (Dec. 2020), p. 25. DOI: 10.1186/s13321-020-00429-4.

[210] Eric-Wubbo Lameijer et al. "The Molecule Evoluator. An interactive evolutionary algorithm for the design of drug-like molecules". In: *Journal of Chemical Information and Modeling* 46.2 (Mar. 2006), pp. 545–552. DOI: 10.1021/ci050369d.

[211] Pavel Polishchuk. "CReM: chemically reasonable mutations framework for structure generation". In: *Journal of Cheminformatics* 12.1 (Dec. 2020), p. 28. DOI: 10.1186/s13321-020-00431-w.

[212] Thomas Blaschke et al. "REINVENT 2.0: an AI tool for de novo drug design". In: *Journal of Chemical Information and Modeling* 60.12 (Dec. 2020), pp. 5918–5922. DOI: 10.1021/acs.jcim.0c00915.

[213] Marwin H. S. Segler et al. "Generating focused molecule libraries for drug discovery with recurrent neural networks". In: *ACS Central Science* 4.1 (Jan. 2018), pp. 120–131. DOI: 10.1021/acscentsci.7b00512.

[214] Qi Liu et al. "Constrained graph variational autoencoders for molecule design". In: *Advances in Neural Information Processing Systems*. Curran Associates Inc., 2018, pp. 7806–7815.

[215] Zhenpeng Zhou et al. "Optimization of molecules via deep reinforcement learning". In: *Scientific Reports* 9.1 (Dec. 2019), p. 10752. DOI: 10.1038/s41598-019-47148-x.

[216] Rocío Mercado et al. "Graph networks for molecular design". In: *Machine Learning: Science and Technology* 2.2 (June 2021), p. 025023. DOI: 10.1088/2632-2153/abcf91. (Visited on 09/13/2022).

[217] Peter Pogány et al. "De novo molecule design by translating from reduced graphs to SMILES". In: *Journal of Chemical Information and Modeling* 59.3 (Mar. 2019), pp. 1136–1146. DOI: 10.1021/acs.jcim.8b00626.

[218] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation". In: Stockholm, 2018.

[219] Niclas Ståhl et al. "Deep reinforcement learning for multiparameter optimization in *de novo* drug design". In: *Journal of Chemical Information and Modeling* 59.7 (July 2019), pp. 3166–3176. DOI: 10.1021/acs.jcim.9b00325.

[220] John Bradshaw et al. “A model to search for synthesizable molecules”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 713. Curran Associates, Inc., 2019, pp. 7937–7949.

[221] Julien Horwood and Emmanuel Noutahi. “Molecular design in synthetically accessible chemical space via deep reinforcement learning”. In: *ACS Omega* 5.51 (Dec. 2020), pp. 32984–32994. DOI: 10.1021/acsomega.0c04153.

[222] Sai Krishna Gottipati et al. “Learning to navigate the synthetically accessible chemical space using reinforcement learning”. In: 2020.

[223] Jacques Boitreaud et al. “OptiMol : optimization of binding affinities in chemical space for drug discovery”. In: *Journal of Chemical Information and Modeling* 60.12 (Dec. 2020), pp. 5658–5666. DOI: 10.1021/acs.jcim.0c00833.

[224] Mingyuan Xu, Ting Ran, and Hongming Chen. “De novo molecule design through the molecular generative model conditioned by 3D information of protein binding sites”. In: *Journal of Chemical Information and Modeling* 61.7 (July 2021), pp. 3240–3254. DOI: 10.1021/acs.jcim.0c01494.

[225] Martin Simonovsky and Joshua Meyers. “DeeplyTough: learning structural comparison of protein binding bites”. In: *Journal of Chemical Information and Modeling* 60.4 (Apr. 2020), pp. 2356–2366. DOI: 10.1021/acs.jcim.9b00554.

[226] Dafydd R. Owen et al. “An oral SARS-CoV-2 $M^{pro}$ inhibitor clinical candidate for the treatment of COVID-19”. In: *Science* 374.6575 (Dec. 2021), pp. 1586–1593. DOI: 10.1126/science.abl4784.

[227] *RDKit: open-source cheminformatics*. URL: rdkit.org (visited on 08/30/2022).

[228] Eric F. Pettersen et al. “UCSF ChimeraX : structure visualization for researchers, educators, and developers”. In: *Protein Science* 30.1 (Jan. 2021), pp. 70–82. DOI: 10.1002/pro.3943.

[229] G. Richard Bickerton et al. “Quantifying the chemical beauty of drugs”. In: *Nature Chemistry* 4.2 (Feb. 2012), pp. 90–98. DOI: 10.1038/nchem.1243.

[230] Peter Ertl and Ansgar Schuffenhauer. "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". In: *Journal of Cheminformatics* 1.1 (Dec. 2009), p. 8. DOI: 10.1186/1758-2946-1-8.

[231] Nathan Brown et al. "GuacaMol: benchmarking models for de novo molecular design". In: *Journal of Chemical Information and Modeling* 59.3 (Mar. 2019), pp. 1096–1108. DOI: 10.1021/acs.jcim.8b00839.

[232] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. "Comparison of multiobjective evolutionary algorithms: empirical results". In: *Evolutionary Computation* 8.2 (June 2000), pp. 173–195. DOI: 10.1162/106365600568202.

[233] Michael Moustakas. "The role of metal ions in biology, biochemistry and medicine". In: *Materials* 14.3 (Jan. 2021), p. 549. DOI: 10.3390/ma14030549.

[234] Dieter Rehder. *Bioinorganic chemistry*. Oxford: Oxford University Press, 2014.

[235] Helen M. Berman et al. "The Protein Data Bank". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 235–242. DOI: 10.1093/nar/28.1.235.

[236] Katarzyna B. Handing et al. "Circulatory zinc transport is controlled by distinct interdomain sites on mammalian albumins". In: *Chemical Science* 7.11 (2016), pp. 6635–6648. DOI: 10.1039/C6SC02267G.

[237] Laura Riccardi, Vito Genna, and Marco De Vivo. "Metal–ligand interactions in drug design". In: *Nature Reviews Chemistry* 2.7 (July 2018), pp. 100–112. DOI: 10.1038/s41570-018-0018-6.

[238] Medhavi Mallick, Ambarish Sharan Vidyarthi, and Shankaracharya. "Tools for predicting metal binding sites in protein: a review". In: *Current Bioinformatics* 6.4 (Dec. 2011), pp. 444–449. DOI: 10.2174/157489311798072990.

[239] Gunseli Bayram Akcapinar and Osman Ugur Sezerman. "Computational approaches for *de novo* design and redesign of metal-binding sites on proteins". In: *Bioscience Reports* 37.2 (Apr. 2017), BSR20160179. DOI: 10.1042/BSR20160179.

[240] J. C. Ebert and R. B. Altman. "Robust recognition of zinc binding sites in proteins". In: *Protein Science* 17.1 (Nov. 2007), pp. 54–65. DOI: 10.1110/ps.073138508.

[241] Wei Zhao et al. "Structure-based de novo prediction of zinc-binding sites in proteins of unknown function". In: *Bioinformatics* 27.9 (May 2011), pp. 1262–1268. DOI: 10.1093/bioinformatics/btr133.

[242] Yu-Feng Lin et al. "MIB: metal ion-binding site prediction and docking server". In: *Journal of Chemical Information and Modeling* 56.12 (Dec. 2016), pp. 2287–2291. DOI: 10.1021/acs.jcim.6b00407.

[243] Michal Brylinski and Jeffrey Skolnick. "FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level". In: *Proteins* 79.3 (Mar. 2011), pp. 735–751. DOI: 10.1002/prot.22913.

[244] Jaspreet Singh Sodhi et al. "Predicting metal-binding site residues in low-resolution structural models". In: *Journal of Molecular Biology* 342.1 (Sept. 2004), pp. 307–320. DOI: 10.1016/j.jmb.2004.07.019.

[245] Mark N. Wass, Lawrence A. Kelley, and Michael J. E. Sternberg. "3DLigandSite: predicting ligand-binding sites using similar structures". In: *Nucleic Acids Research* 38.suppl_2 (July 2010), W469–W473. DOI: 10.1093/nar/gkq406.

[246] Chih-Hao Lu et al. "Prediction of metal ion–binding sites in proteins using the fragment transformation method". In: *PLoS ONE* 7.6 (June 2012). Ed. by Beata G. Vertessy, e39252. DOI: 10.1371/journal.pone.0039252.

[247] Xiuzhen Hu et al. "Recognizing metal and acid radical ion-binding sites by integrating *ab initio* modeling with template-based transferals". In: *Bioinformatics* 32.21 (Nov. 2016), pp. 3260–3269. DOI: 10.1093/bioinformatics/btw396.

[248] Victor Muñoz Robles et al. "What can molecular modelling bring to the design of artificial inorganic cofactors?" In: *Faraday Discuss.* 148 (2011), pp. 137–159. DOI: 10.1039/C004578K.

[249] Giuseppe Sciortino, Manuel Aureliano, and Eugenio Garribba. "Rationalizing the decavanadate(V) and oxidovanadium(IV) binding to G-actin and the competition with decaniobate(V) and ATP". In: *Inorganic Chem-*

*istry* 60.1 (Jan. 2021), pp. 334–344. DOI: 10.1021/acs.inorgchem.0c02971.

[250] Giuseppe Sciortino, Jean-Didier Maréchal, and Eugenio Garribba. "Integrated experimental/computational approaches to characterize the systems formed by vanadium with proteins and enzymes". In: *Inorganic Chemistry Frontiers* 8.8 (2021), pp. 1951–1974. DOI: 10.1039/D0QI01507E.

[251] Lorena Roldán-Martín et al. "Impact of Cu(II) and Al(III) on the conformational landscape of amyloid$\beta_{1\text{-}42}$". In: *Physical Chemistry Chemical Physics* 23.23 (2021), pp. 13023–13032. DOI: 10.1039/D1CP01561C.

[252] Vanesa Ramirez-Bello, Javier Martinez-Seoane, and Arline Fern. "Zinc and copper ions induce aggregation of human β-crystallins". In: *Molecules* 27.9 (2022), p. 2970. DOI: 10.3390/molecules27092970.

[253] Douglas M M Soares et al. "Reannotation of fly Amanita l-DOPA dioxygenase gene enables its cloning and heterologous expression". In: *ACS Omega* 7.18 (2022), pp. 16070–16079. DOI: 10.1021/acsomega.2c01365.

[254] Yiling Xiao et al. "Aβ(1–42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease". In: *Nature Structural & Molecular Biology* 22.6 (June 2015), pp. 499–505. DOI: 10.1038/nsmb.2991.

[255] Elena M. Sánchez-Fernández et al. "Synthesis of polyfluoroalkyl $sp^2$-iminosugar glycolipids and evaluation of their immunomodulatory properties towards anti-tumor, anti-leishmanial and anti-inflammatory therapies". In: *European Journal of Medicinal Chemistry* 182 (Nov. 2019), p. 111604. DOI: 10.1016/j.ejmech.2019.111604.

[256] Agnieszka Krzemińska et al. "Influence of association on binding of disaccharides to YKL-39 and hHyal-1 enzymes". In: *International Journal of Molecular Sciences* 23.14 (2022), p. 7705. DOI: 10.3390/ijms23147705.

[257] E. Kobayashi et al. "KRN7000, a novel immunomodulator, and its antitumor activities". In: *Oncology Research* 7.10-11 (1995), pp. 529–534.

[258] Netanel Tzarum et al. "Lipid molecules induce p38α activation via a novel molecular switch". In: *Journal of Molecular Biology* 424.5 (Dec. 2012), pp. 339–353. DOI: 10.1016/j.jmb.2012.10.007.

[259] Elena Alcalde-Estévez et al. “The sp 2 -iminosugar glycolipid 1-dodecylsulfonyl-5 N ,6 O -oxomethylidenenojirimycin (DSO 2 -ONJ) as selective anti-inflammatory agent by modulation of hemeoxygenase-1 in Bv.2 microglial cells and retinal explants”. In: *Food and Chemical Toxicology* 111 (Jan. 2018), pp. 454–466. DOI: 10.1016/j.fct.2017.11.050.

[260] J. Willem M. Nissink et al. “A new test set for validating predictions of protein-ligand interaction”. In: *Proteins: Structure, Function, and Bioinformatics* 49.4 (Dec. 2002), pp. 457–471. DOI: 10.1002/prot.10232.

[261] Masayuki Ishihara et al. “Polyelectrolyte complexes of natural polymers and their biomedical applications”. In: *Polymers* 11.4 (Apr. 2019), p. 672. DOI: 10.3390/polym11040672.

[262] Vedran Milosavljevic et al. “Encapsulation of doxorubicin in furcellaran/chitosan nanocapsules by layer-by-layer technique for selectively controlled drug delivery”. In: *Biomacromolecules* 21.2 (Feb. 2020), pp. 418–434. DOI: 10.1021/acs.biomac.9b01175.

[263] Luc Multigner et al. “Chlordecone exposure and adverse effects in French West Indies populations”. In: *Environmental Science and Pollution Research* 23.1 (Jan. 2016), pp. 3–8. DOI: 10.1007/s11356-015-4621-5.

[264] Carine Dubuisson et al. “Impact of subsistence production on the management options to reduce the food exposure of the Martinican population to Chlordecone”. In: *Regulatory Toxicology and Pharmacology* 49.1 (Oct. 2007), pp. 5–16. DOI: 10.1016/j.yrtph.2007.04.008.

[265] Ronald Ranguin et al. “Development and characterisation of a nanostructured hybrid material with vitamin $B_{12}$ and bagasse-derived activated carbon for anaerobic chlordecone (Kepone) removal”. In: *Environmental Science and Pollution Research* 27.33 (Nov. 2020), pp. 41122–41131. DOI: 10.1007/s11356-020-08201-9.

[266] Van-Duong Dao et al. “AuNP/graphene nanohybrid prepared by dry plasma reduction as a low-cost counter electrode material for dye-sensitized solar cells”. In: *Electrochimica Acta* 156 (Feb. 2015), pp. 138–146. DOI: 10.1016/j.electacta.2014.12.109.

[267] Yuanyuan Yin et al. “Atmospheric pressure synthesis of nitrogen doped graphene quantum dots for fabrication of BiOBr nanohybrids with en-

hanced visible-light photoactivity and photostability". In: *Carbon* 96 (Jan. 2016), pp. 1157–1165. DOI: 10.1016/j.carbon.2015.10.068.

[268] Guy Glod et al. "Cobalamin-mediated reduction of *cis*- and *trans*-dichloroethene, 1,1-dichloroethene, and vinyl chloride in homogeneous aqueous solution: reaction kinetics and mechanistic considerations". In: *Environmental Science & Technology* 31.11 (Nov. 1997), pp. 3154–3160. DOI: 10.1021/es9701220.

[269] Y.-H. Kim and E. R. Carraway. "Reductive dechlorination of PCE and TCE by vitamin $B_{12}$ and ZVMs". In: *Environmental Technology* 23.10 (Oct. 2002), pp. 1135–1145. DOI: 10.1080/09593332308618332.

[270] Mark H. Smith and Sandra L. Woods. "Regiospecificity of chlorophenol reductive dechlorination by vitamin $B_{12s}$". In: *Applied and Environmental Microbiology* 60.11 (Nov. 1994), pp. 4111–4115. DOI: 10.1128/aem.60.11.4111-4115.1994.

[271] Pei-Chun Chiu and Martin Reinhard. "Transformation of Carbon Tetrachloride by Reduced Vitamin $B_{12}$ in Aqueous Cysteine Solution". In: *Environmental Science & Technology* 30.6 (May 1996), pp. 1882–1889. DOI: 10.1021/es950477o.

[272] Ronald Ranguin et al. "Study of chlordecone desorption from activated carbons and subsequent dechlorination by reduced cobalamin". In: *Environmental Science and Pollution Research* 24.33 (Nov. 2017), pp. 25488–25499. DOI: 10.1007/s11356-017-9542-z.

[273] Django Sussman, Charles Wilson, and Jay C. Nix. "The structural basis for molecular recognition by the vitamin $B_{12}$ RNA aptamer". In: *Nature Structural Biology* 7.1 (Jan. 2000), pp. 53–57. DOI: 10.1038/71253.

[274] Yang Huang et al. "Activated carbon efficient atomistic model construction that depicts experimentally-determined characteristics". In: *Carbon* 83 (Mar. 2015), pp. 1–14. DOI: 10.1016/j.carbon.2014.11.012.

[275] Roman A. Laskowski. "SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions". In: *Journal of Molecular Graphics* 13.5 (Oct. 1995), pp. 323–330. DOI: 10.1016/0263-7855(95)00073-9.

[276] Giuseppe Sciortino et al. “Computational insight into the interaction of oxaliplatin with insulin”. In: *Metallomics* 11.4 (Apr. 2019), pp. 765–773. DOI: 10.1039/c8mt00341f.

[277] José-Emilio Sánchez-Aparicio et al. “Successes and challenges in multiscale modelling of artificial metalloenzymes: the case study of POP-$Rh_2$ cyclopropanase”. In: *Faraday Discussions* 234 (2022), pp. 349–366. DOI: 10.1039/D1FD00069A.

[278] Katja Dralle Mjos and Chris Orvig. “Metallodrugs in medicinal inorganic chemistry”. In: *Chemical Reviews* 114.8 (Apr. 2014), pp. 4540–4563. DOI: 10.1021/cr400460s.

[279] Elizabeth J. Anthony et al. “Metallodrugs are unique: opportunities and challenges of discovery and development”. In: *Chemical Science* 11.48 (2020), pp. 12888–12917. DOI: 10.1039/D0SC04082G.

[280] M. Gielen and Edward R. T. Tiekink, eds. *Metallotherapeutic drugs and metal-based diagnostic agents: the use of metals in medicine*. Hoboken, N.J: Wiley, 2005.

[281] Zijian Guo and Peter J. Sadler. “Metals in medicine”. In: *Angewandte Chemie International Edition* 38.11 (June 1999), pp. 1512–1531. DOI: 10.1002/(SICI)1521-3773(19990601)38:11<1512::AID-ANIE1512>3.0.CO;2-Y.

[282] Serenella Medici et al. “Noble metals in medicine: latest advances”. In: *Coordination Chemistry Reviews* 284 (Feb. 2015), pp. 329–350. DOI: 10.1016/j.ccr.2014.08.002.

[283] Chris J. Jones and John Thornback. *Medicinal applications of coordination chemistry*. Cambridge: Royal Society of Chemistry, 2007.

[284] James C. Dabrowiak. *Metals in medicine*. Second edition. Inorganic chemistry: a textbook series. Hoboken, NJ: John Wiley & Sons, Inc, 2017.

[285] Jiezhong Chen et al. “Insulin caused drug resistance to oxaliplatin in colon cancer cell line HT29”. In: *Journal of Gastrointestinal Oncology* 2.1 (2011), pp. 27–33. DOI: 10.3978/j.issn.2078-6891.2010.028.

[286] Angela Casini et al. “ESI mass spectrometry and X-ray diffraction studies of adducts between anticancer platinum drugs and hen egg white

lysozyme". In: *Chemical Communications* 2 (2007), pp. 156–158. DOI: 10.1039/B611122J.

[287] Jing Li et al. "Mass spectrometric studies on the interaction of cisplatin and insulin". In: *Amino Acids* 48.4 (Apr. 2016), pp. 1033–1043. DOI: 10.1007/s00726-015-2159-y.

[288] Estefanía Moreno-Gordaliza et al. "Novel insights into the bottom-up mass spectrometry proteomics approach for the characterization of Pt-binding proteins: the insulin-cisplatin case study". In: *The Analyst* 135.6 (2010), p. 1288. DOI: 10.1039/b927110d.

[289] Estefanía Moreno-Gordaliza et al. "Top-down mass spectrometric approach for the full characterization of insulincisplatin adducts". In: *Analytical Chemistry* 81.9 (May 2009), pp. 3507–3516. DOI: 10.1021/ac900046v.

[290] Rupasri Mandal, Michael B. Sawyer, and Xing-Fang Li. "Mass spectrometry study of hemoglobin-oxaliplatin complexes in colorectal cancer patients and potential association with chemotherapeutic responses". In: *Rapid Communications in Mass Spectrometry* 20.17 (Sept. 2006), pp. 2533–2538. DOI: 10.1002/rcm.2622.

[291] Graham Bentley et al. "Structure of insulin in 4-zinc insulin". In: *Nature* 261.5556 (May 1976), pp. 166–168. DOI: 10.1038/261166a0.

[292] Luigi Messori, Tiziano Marzo, and Antonello Merlino. "Interactions of carboplatin and oxaliplatin with proteins: insights from X-ray structures and mass spectrometry studies of their ribonuclease A adducts". In: *Journal of Inorganic Biochemistry* 153 (Dec. 2015), pp. 136–142. DOI: 10.1016/j.jinorgbio.2015.07.011.

[293] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596 (July 2021). DOI: 10.1038/s41586-021-03819-2.

[294] Minkyung Baek et al. "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557 (Aug. 2021), pp. 871–876. DOI: 10.1126/science.abj8754.

[295] Charlotte Møller et al. "Determination of the binding sites for oxaliplatin on insulin using mass spectrometry-based approaches". In: *Ana-*

*lytical and Bioanalytical Chemistry* 401.5 (Sept. 2011), pp. 1619–1629. DOI: 10.1007/s00216-011-5239-1.

[296] Ivana Drienovská et al. "Design of an enantioselective artificial metallohydratase enzyme containing an unnatural metal-binding amino acid". In: *Chemical Science* 8.10 (2017), pp. 7228–7235. DOI: 10.1039/C7SC03477F.

[297] Victor Muñoz Robles et al. "Computational Insights on an Artificial Imine Reductase Based on the Biotin–Streptavidin Technology". In: *ACS Catalysis* 4.3 (Mar. 2014), pp. 833–842. DOI: 10.1021/cs400921n.

[298] Fabian Schwizer et al. "Artificial Metalloenzymes: Reaction Scope and Optimization Strategies". In: *Chemical Reviews* 118.1 (Jan. 2018), pp. 142–231. DOI: 10.1021/acs.chemrev.7b00014.

[299] Poonam Srivastava et al. "Engineering a dirhodium artificial metalloenzyme for selective olefin cyclopropanation". In: *Nature Communications* 6.1 (Nov. 2015), p. 7789. DOI: 10.1038/ncomms8789.

[300] Hao Yang et al. "Evolving artificial metalloenzymes via random mutagenesis". In: *Nature Chemistry* 10.3 (Mar. 2018), pp. 318–324. DOI: 10.1038/nchem.2927.

[301] Ken Ellis-Guardiola et al. "Crystal Structure and Conformational Dynamics of Pyrococcus furiosus Prolyl Oligopeptidase". In: *Biochemistry* 58.12 (Mar. 2019), pp. 1616–1626. DOI: 10.1021/acs.biochem.9b00031.

[302] Jared C. Lewis. "Beyond the Second Coordination Sphere: Engineering Dirhodium Artificial Metalloenzymes To Enable Protein Control of Transition Metal Catalysis". In: *Accounts of Chemical Research* 52.3 (Mar. 2019), pp. 576–584. DOI: 10.1021/acs.accounts.8b00625.

[303] David M. Upp et al. "Engineering Dirhodium Artificial Metalloenzymes for Diazo Coupling Cascade Reactions**". In: *Angewandte Chemie International Edition* 60.44 (Aug. 2021), pp. 23672–23677. DOI: 10.1002/anie.202107982.

# Appendix. Publications from this thesis

(1) Sánchez-Aparicio, J.-E.; Sciortino, G.; Herrmannsdoerfer, D. V.; Chueca, P. O.; Pedregal, J. R.-G.; Maréchal, J.-D. GPathFinder: Identification of Ligand-Binding Pathways by a Multi-Objective Genetic Algorithm. *Int. J. Mol. Sci.* **2019**, *20* (13), 3155. https://doi.org/10.3390/ijms20133155.

(2) Streltsov, V. A.; Luang, S.; Peisley, A.; Varghese, J. N.; Ketudat Cairns, J. R.; Fort, S.; Hijnen, M.; Tvaroška, I.; Ardá, A.; Jiménez-Barbero, J.; Alfonso-Prieto, M.; Rovira, C.; Mendoza, F.; Tiessler-Sala, L.; Sánchez-Aparicio, J.-E.; Rodríguez-Guerra, J.; Lluch, J. M.; Maréchal, J.-D.; Masgrau, L.; Hrmova, M. Discovery of Processive Catalysis by an Exo-Hydrolase with a Pocket-Shaped Active Site. *Nat. Commun.* **2019**, *10* (1), 2222. https://doi.org/10.1038/s41467-019-09691-z.

(3) Sánchez-Aparicio, J.-E.; Tiessler-Sala, L.; Velasco-Carneros, L.; Roldán-Martín, L.; Sciortino, G.; Maréchal, J.-D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *J. Chem. Inf. Model.* **2021**, *61* (1), 311–323. https://doi.org/10.1021/acs.jcim.0c00827.

(4) Sánchez-Fernández, E. M.; García-Moreno, M. I.; Arroba, A. I.; Aguilar-Diosdado, M.; Padrón, J. M.; García-Hernández, R.; Gamarro, F.; Fustero, S.; Sánchez-Aparicio, J.-E.; Masgrau, L.; García Fernández, J. M.; Ortiz Mellet, C. Synthesis of Polyfluoroalkyl Sp2-Iminosugar Glycolipids and Evaluation of Their Immunomodulatory Properties towards Anti-Tumor, Anti-Leishmanial and Anti-Inflammatory Therapies. *Eur. J. Med. Chem.* **2019**, *182*, 111604. https://doi.org/10.1016/j.ejmech.2019.111604.

(5) Krzemińska, A.; Sánchez-Aparicio, J.-E.; Maréchal, J.-D.; Paneth, A.; Paneth, P. Influence of Association on Binding of Disaccharides to YKL-39 and HHyal-1 Enzymes. *Int. J. Mol. Sci.* **2022**, *23* (14), 7705. https://doi.org/10.3390/ijms23147705.

(6) Sciortino, G.; Sánchez-Aparicio, J.-E.; Rodríguez-Guerra Pedregal, J.; Garribba, E.; Maréchal, J.-D. Computational Insight into the Interaction of Oxaliplatin with Insulin. *Metallomics* **2019**, *11* (4), 765–773. https://doi.org/10.1039/c8mt00341f.

(7) Sánchez-Aparicio, J.-E.; Sciortino, G.; Mates-Torres, E.; Lledós, A.; Maréchal, J.-D. Successes and Challenges in Multiscale Modelling of Artificial Metalloenzymes: The Case Study of POP-Rh2 Cyclopropanase. *Faraday Discuss.* **2022**, *234*, 349–366. https://doi.org/10.1039/D1FD00069A.