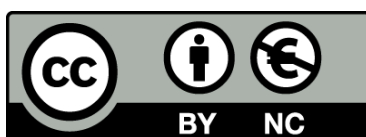




UNIVERSITAT<sub>DE</sub>  
BARCELONA

## Studying protein-ligand interactions using a Monte Carlo procedure

Daniel Lecina Casas



Aquesta tesi doctoral està subjecta a la llicència [Reconeixement- NoComercial 3.0. Espanya de Creative Commons](#).

Esta tesis doctoral está sujeta a la licencia [Reconocimiento - NoComercial 3.0. España de Creative Commons](#).

This doctoral thesis is licensed under the [Creative Commons Attribution-NonCommercial 3.0. Spain License](#).

# Studying protein-ligand interactions using a Monte Carlo procedure

Tesis Doctoral

Doctorat en Física  
Universitat de Barcelona  
2017

Autor: Daniel Lecina Casas  
Supervisor: Victor Guallar Tases  
Tutor: Giancarlo Franzese





# Table of Contents

|  |           |
|--|-----------|
| <b>i. Abstract .....</b>   | <b>5</b>  |
| <b>ii. List of Publications .....</b>  | <b>7</b>  |
| <b>iii. List of Abbreviations.....</b>   | <b>9</b>  |
| <b>iv. Motivation .....</b>  | <b>11</b> |
| <b>Introduction .....</b>  | <b>13</b> |
| <i>Proteins</i> .....  | 13        |
| Proteins as dynamic entities.....  | 14        |
| <i>Biomolecular Modeling</i> .....   | 17        |
| Quantum Modeling.....  | 17        |
| Classical Modeling.....  | 18        |
| <i>Molecular Simulations</i> .....   | 23        |
| Molecular Dynamics.....  | 23        |
| Monte Carlo.....   | 24        |
| PELE.....  | 27        |
| <i>Binding free energy</i> .....   | 33        |
| Absolute binding free energy.....  | 33        |
| Experimental estimations .....   | 34        |
| Computational estimations .....  | 35        |
| <i>Markov State Models</i> .....   | 38        |
| Construction, validation, and analysis .....   | 39        |
| <b>Objectives.....</b>   | <b>43</b> |
| <b>Results.....</b>  | <b>45</b> |
| <i>Publication 1 - Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques</i> .....                        | 45        |
| <i>Publication 2 - Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints</i> ..... | 57        |
| <i>Publication 3 - The unravelling of the complex pattern of tyrosinase inhibition</i> .....   | 69        |
| <i>Publication 4 - Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions</i> .....                | 81        |
| <i>Publication 5 - Adaptive simulations, towards interactive protein-ligand modeling</i> ....  | 93        |
| <i>Publication 6 – A Markov State Model benchmark using a Monte Carlo procedure as a propagator</i> .....                                      | 111       |

|  |            |
|--|------------|
| <b>Discussion</b> .....  | <b>121</b> |
| <i>Development of PELE</i> .....   | 121        |
| <i>Establishing a protocol to study protein-ligand binding</i> .....                                     | 123        |
| <i>Development of a procedure to overcome the sampling limitations associated to metastability</i> ..... | 128        |
| <b>Conclusions</b> .....   | <b>131</b> |
| <b>Appendices</b> .....  | <b>133</b> |
| <i>Resum de la tesi</i> .....  | 133        |
| <i>Resum individual de cada publicació</i> .....   | 137        |
| Resum de la primera publicació .....   | 137        |
| Resum de la segona publicació .....  | 137        |
| Resum de la tercera publicació .....   | 137        |
| Resum de la quarta publicació .....  | 137        |
| Resum de la cinquena publicació .....  | 138        |
| <i>Supporting information paper 2</i> .....  | 139        |
| <i>Supporting information paper 3</i> .....  | 160        |
| <i>Supporting information paper 4</i> .....  | 169        |
| <i>Supporting information paper 5</i> .....  | 180        |
| <b>Bibliography</b> .....  | <b>191</b> |

## **i. Abstract**

Biomolecular simulations have been widely used in the study of protein-ligand interactions. Comprehending the mechanisms involved in the prediction of binding affinities has a significant repercussion in the pharmaceutical industry. Notwithstanding the intrinsic difficulty of sampling the phase space, hardware and methodological developments make computer simulations a promising candidate in the resolution of biophysically relevant problems. In this context, the objective of the thesis is the development of a protocol that introduces a more efficient study of protein-ligand interactions, in view to disseminate PELE, a Monte Carlo sampling procedure, in drug discovery pipelines. Our main focus has been overcoming the sampling limitations caused by the ruggedness of the energy landscape, applying our protocol to perform atomistically detailed analyses in nuclear hormone receptors, G-protein coupled receptors, tyrosinases, and prolyl oligopeptidases, in collaboration with a pharmaceutical company and several experimental laboratories. Overall, we hope that the methodologies presented herein help streamline the drug design process.

**Keywords:** PELE, Monte Carlo, Markov State Models, protein-ligand binding, binding free energy, nuclear hormone receptors, G-protein coupled receptors, tyrosinases, prolyl oligopeptidases



## ii. List of Publications

### 1. Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques

**Authors:** Martin Kotev, Daniel Lecina, Teresa Tarragó, Ernest Giralt, Victor Guallar

**Journal:** Biophysical Journal, 108, 116–125 (2015)

### 2. Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints

**Authors:** Karl Edman, Ali Hosseini, Magnus K. Bjursell, Anna Aagaard, Lisa Wissler, Anders Gunnarsson, Tim Kaminski, Christian Köhler, Stefan Bäckström, Tina J. Jensen, Anders Cavallin, Ulla Karlsson, Ewa Nilsson, Daniel Lecina, Ryoji Takahashi, Christoph Grebner, Stefan Geschwindner, Matti Lepistö, Anders C. Hogner, Victor Guallar.

**Journal:** Structure, 23, 2280–2291 (2015)

### 3. The unravelling of the complex pattern of tyrosinase inhibition

**Authors:** Batel Deri, Margarita Kanteev, Mor Goldfeder, Daniel Lecina, Victor Guallar, Noam Adir, Ayelet Fishman

**Journal:** Scientific Reports 6, 34993 (2016)

### 4. Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions

**Authors:** Christoph Grebner\*, Daniel Lecina\*, Victor Gil, Johan Ulander, Pia Hansson, Anita Dellsen, Christian Tyrchan, Karl Edman, Anders Hogner, Victor Guallar. (*\* equally contributing authors*)

**Journal:** Biophysical Journal, 112, 1147-1156 (2017)

### 5. Adaptive simulations, towards interactive protein-ligand modeling

**Authors:** Daniel Lecina, Joan Francesc Gilabert, Victor Guallar

**Journal:** Scientific Reports (In Revision)

### 6. A Markov State Model benchmark using a Monte Carlo procedure as propagator

**Authors:** Daniel Lecina, Joan Francesc Gilabert, Christoph Grebner, Victor Guallar.

In Preparation

## Other Publications

### **7. The variability of the large genomic segment of Třahyná orthobunyavirus and an all-atom exploration of its anti-viral drug resistance**

**Authors:** Patrik Kilian, James J. Valdes, Daniel Lecina-Casas, Tomáš Chrudimský, Daniel Ružek

**Journal:** Infection, Genetics and Evolution, 20, 304–311 (2013)

### **8. Enhancing backbone sampling in Monte Carlo simulations using internal coordinates normal mode analysis**

**Authors:** Victor A. Gil, Daniel Lecina, Christoph Grebner, Victor Guallar

**Journal:** Bioorganic & Medicinal Chemistry, 24, 4855–4866, (2016)

### iii. List of Abbreviations

|                |  |
|----------------|--|
| <b>ANM</b>     | Anisotropic Network Model                            |
| <b>DFT</b>     | Density Functional Theory                            |
| <b>dibC</b>    | Desisobutyrylciclesonide                             |
| <b>EM</b>      | Electron Microscopy                                  |
| <b>FF</b>      | Force Field  |
| <b>FP</b>      | Fluorescent polarization                             |
| <b>GB</b>      | Generalized Born                                     |
| <b>GPCR</b>    | G-Protein Coupled Receptor                           |
| <b>GPU</b>     | Graphical Processing Unit                            |
| <b>GR</b>      | Glucocorticoid Receptor                              |
| <b>HPC</b>     | High-performance computing                           |
| <b>HQ</b>      | Hydroquinone   |
| <b>IC-NMA</b>  | Internal Coordinate Normal Mode Analysis             |
| <b>ITC</b>     | Isothermal Titration Calorimetry                     |
| <b>KA</b>      | Kojic Acid   |
| <b>LIE</b>     | Linear Interaction Energy                            |
| <b>MC</b>      | Monte Carlo  |
| <b>MCMC</b>    | Markov Chain Monte Carlo                             |
| <b>MD</b>      | Molecular Dynamics                                   |
| <b>MM</b>      | Molecular Mechanics                                  |
| <b>MM/PBSA</b> | Molecular Mechanics / Poisson Boltzmann Surface Area |
| <b>MM/GBSA</b> | Molecular Mechanics / Generalized Born Surface Area  |
| <b>MR</b>      | Mineralocorticoid Receptors                          |
| <b>MSM</b>     | Markov State Model                                   |
| <b>NCMC</b>    | Non-equilibrium Candidate Monte Carlo                |
| <b>NHR</b>     | Nuclear Hormone Receptor                             |
| <b>NMA</b>     | Normal Mode Analysis                                 |
| <b>NMR</b>     | Nuclear Magnetic Resonance                           |
| <b>PB</b>      | Poisson-Boltzmann                                    |
| <b>PCA</b>     | Principal Component Analysis                         |
| <b>PCCA</b>    | Perron Cluster Cluster Analysis                      |
| <b>PELE</b>    | Protein Energy Landscape Exploration                 |



|             |                                    |
|-------------|------------------------------------|
| <b>PLOP</b> | Protein Local Optimization Program |
| <b>PMF</b>  | Potential of Mean Force            |
| <b>POP</b>  | Prolyl Oligopeptidase              |
| <b>PR</b>   | Progesterone Receptor              |
| <b>QM</b>   | Quantum Mechanics                  |
| <b>RMSD</b> | Root-Mean-Square Deviation         |
| <b>RMSG</b> | Root-Mean-Square Gradient          |
| <b>SASA</b> | Solvent-Accessible Surface Area    |
| <b>SPR</b>  | Surface Plasmon Resonance          |
| <b>TN</b>   | Truncated Newton                   |

## iv. Motivation

When we suffer an illness, we are prescribed a treatment. If it is not successful, the physician changes the medication looking for the one that fits us the best. This is an impressively common problem; out of the ten highest-grossing drugs in United States<sup>1</sup>, none of them improved over the baseline condition to more than 25% of the population<sup>2</sup>. The situation can be dramatic depending on the illness. See for example the first one in the list: schizophrenia. Patients need to be treated more precisely, taking into account their individuality. In parallel, drugs need to respond to this new requirement and need to be more specific and efficient.

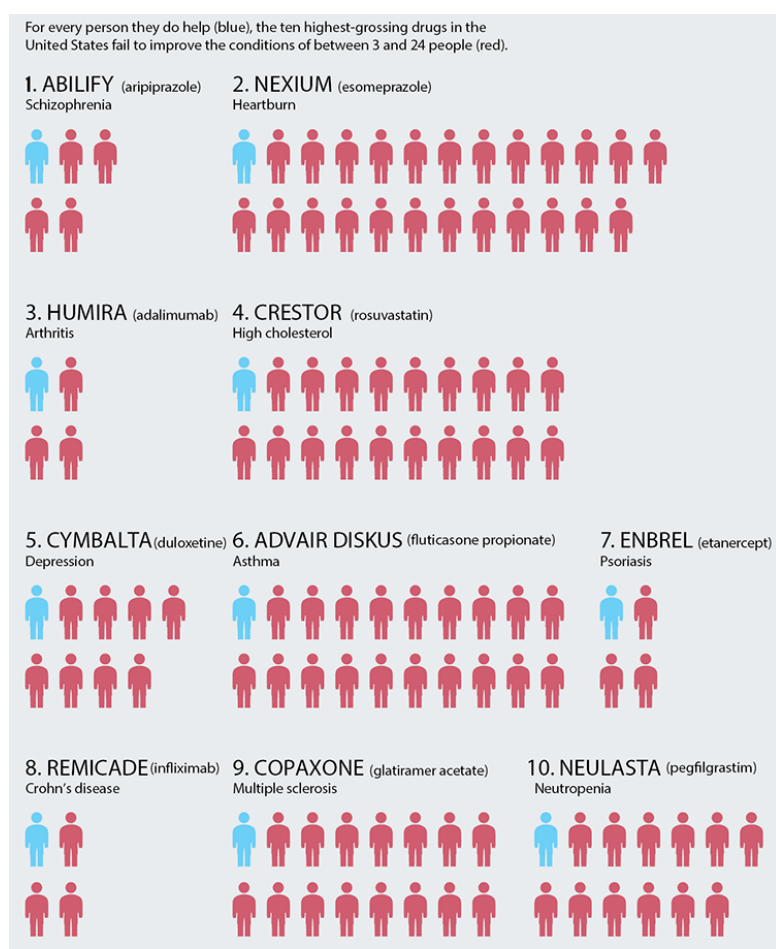


Figure 1. Efficacy of the ten highest-grossing drugs in the US in 2013. Source: Ref. 2

Computational tools have been gaining popularity lately in drug development due to their potential reducing costs and timelines. Also, they have been seen as a great complement for experimental assays, especially because they can give atomistic detail of the process under study. This has attracted the attention of pharmaceutical companies, motivating, for example, the collaboration of Sanofi and Schrödinger for up to \$120M<sup>3</sup> or the acquisition of Nimbus by Gilead Sciences for \$400M, with the possibility of adding up to \$1,200M depending on milestone results<sup>4</sup>.

Hardware developments are one of the causes of this new interest. One of the paradigms is D.E. Shaw's Anton supercomputer<sup>5</sup>, an especially designed supercomputer for running molecular simulations. It broke the millisecond simulation time, a timescale at which many *interesting* biological functions occur. We have also seen other examples, such as the MDGRAPE<sup>6</sup> or Blue Waters<sup>7</sup> supercomputers. On the other hand, these successes are sustained by major methodological improvements, which have been developed in conjunction. For example, force fields permit an increasingly more accurate description, or enhanced sampling techniques permit a better covering of the energy landscape. An excellent example of this conjunction is the Folding@home project<sup>8</sup>, developed in Pande's group. Instead of running few but very long simulations, they use a collaborative distributed protocol to run many shorter ones, which are combined with the Markov state model (MSM)<sup>9</sup> methodology to build a unique mathematical model. With this approach, they have also reached the millisecond aggregated time in protein folding simulations<sup>10</sup> (video: <https://goo.gl/Agthrw>). Following the same idea, the De Fabritiis' group has been able to reconstruct the ligand binding process<sup>11</sup> using GPU-enhanced simulations on GPUGRID<sup>12</sup>.

The aim of this dissertation is studying protein-ligand interactions. We aim to develop a protocol based on molecular simulations, in view to be used in an industrial drug design context. More specifically, our objective is improving lead optimization by accurately computing the free energy of association of protein receptors with small molecules, known as ligands. Among the current challenges in the field that are described below in this section, our main focus has been overcoming the sampling limitations caused by the ruggedness of the energy landscape.

With this in mind, the objectives are divided in three parts. First, we built a competitive sampling program, PELE (Protein Energy Landscape Exploration), a professional tool upon which we established the rest of procedures. PELE was written with reliability, efficiency, and maintainability in mind. This means that it permits the easy addition of leading algorithms while being consistent with the previous behavior. Then, we present our proposal to compute binding free energy differences. The procedure uses a combination of PELE, given its competitive sampling advantage, with MSMs, which serve to quantify the exploration. We show its applicability in current industrial problems, in collaboration with the pharmaceutical company AstraZeneca. Finally, we present an adaptive procedure that represents a significant improvement in sampling. With this technique we are able to map binding mechanisms in occluded binding sites in the order of minutes, taking into account protein and ligand flexibility. These developments open the door to a much faster, while accurate, screening of compounds.

Hopefully, in the following years we will see improvements in the drug design process that will be translated to more efficient and precise medicines, helping to erase the discouraging scenario drawn at the beginning. We firmly believe that computational tools will play a significant role.

## Introduction

In this section, we frame the thesis in the state of the art of molecular simulations in order contextualize the techniques presented in the results section. We first overview the dynamic nature of proteins, which evidences the need of using computational tools that can capture it. We then outline the main biomolecule mathematical models, and the main simulation methods: molecular dynamics and Monte Carlo. We pay particular emphasis to PELE, the simulation program of choice. Finally, we frame MSMs, the methodology that we use to quantify the energy landscape. The list of resulting publications is shown in section *ii*, and is referenced throughout the rest of the thesis.

## Proteins

Proteins are extraordinary molecular machines shaped by evolution and are responsible for most biological functions in our bodies. For example, hemoglobin **transports** oxygen to the cells, the G-Protein Coupled Receptors (GPCRs) are transmembrane proteins that emit **signals** upon stimuli, ATP synthase **synthesizes** ATP, or alcohol dehydrogenase is a **catalyzer** involved in the oxidation of alcohol. The list (sequence) of amino acids that forms a given protein is its primary structure and encodes its function. The protein is locally folded into a secondary structure (e.g. alpha helix or beta sheet being the most common ones), which is packed into a 3D structure known as the tertiary structure. Often, proteins are composed of non-covalently bound subunits, which may be different (heterodimer) or similar (homodimer), and is called the quaternary structure. In this thesis we will study their interactions with small molecules, namely ligands, having a particular interest in their binding process.

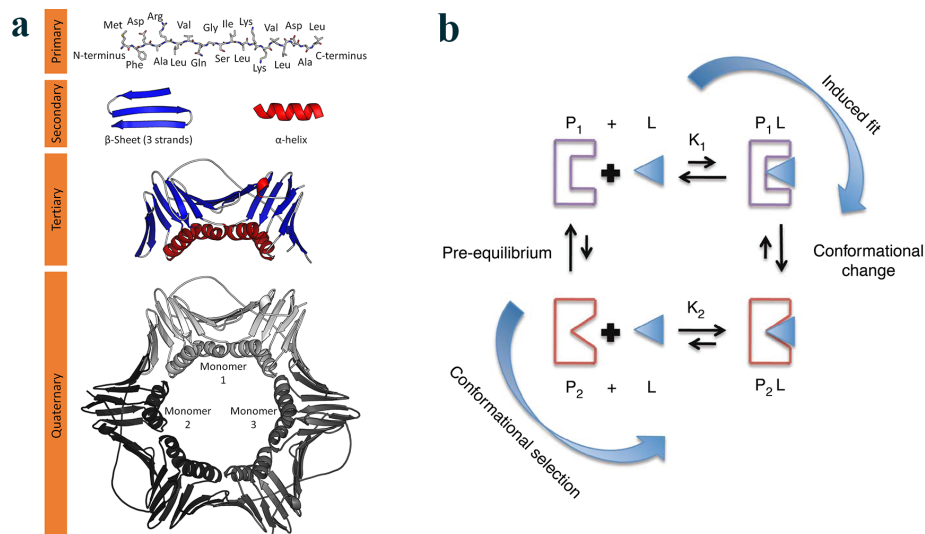


Figure 2. Protein structure and binding mechanisms. Panel (a): Primary, secondary, tertiary and quaternary structure of proliferating cell nuclear antigen. Source: <https://goo.gl/Jnterg>. Panel (b): Induced-fit and conformational selection binding mechanisms. Source: Ref. 13.

## Proteins as dynamic entities

Originally, proteins were considered as rigid objects, and their shape was linked to their action. This concept dates back to 1894 when Fisher<sup>14</sup> proposed the idea of the *lock-and-key* binding mechanism. In this rather simple model, the protein undergoes through minor conformational changes upon binding, and the ligand binds into a complementary protein pocket. According to a study carried out in Orozco's and Aloy's lab in 2011<sup>15</sup>, where they analyzed 2090 different complex transitions from unbound to bound, this mechanism could explain about two-thirds of bindings. In 1958, Koshland incorporated the idea of protein-ligand interactions mutually altering their shape upon binding, with an *induced-fit* mechanism<sup>16</sup>. It was not until 50 years later when a third model was proposed to play a role in ligand recognition, namely the *conformational-selection*<sup>13</sup>. It states that proteins coexist in different conformations, but there is a shift of population towards the ones that accommodate the ligand better.

Historically, it has been difficult to conceive proteins as dynamic entities. One of the main factors could be that we have traditionally obtained their structural information using *X-ray* crystallography. In 1958, Kendrew and coworkers determined the first protein crystal, myoglobin, and, today, *X-rays* account for more than hundred thousand solved structures, the 90% of the Protein Data Bank (source: <https://goo.gl/Cuhwo6>). It is a very powerful technique, able to determine atomic positions from the diffraction pattern with a high resolution for an extensive range of system sizes. However, its main limitation is that the complex under study must be in a crystal lattice, and some essential proteins are tough to crystallize, such as those in the cell membrane or intrinsically disordered proteins<sup>17</sup>. Also, crystals only give a static average picture. In a beautiful simile, Orellana compares in her Ph.D. thesis<sup>18</sup> the unfruitful attempts of painters to capture the galloping of a horse before the invention of cameras with scientists inferring protein dynamics from crystallographic structures, seen as resting horses.

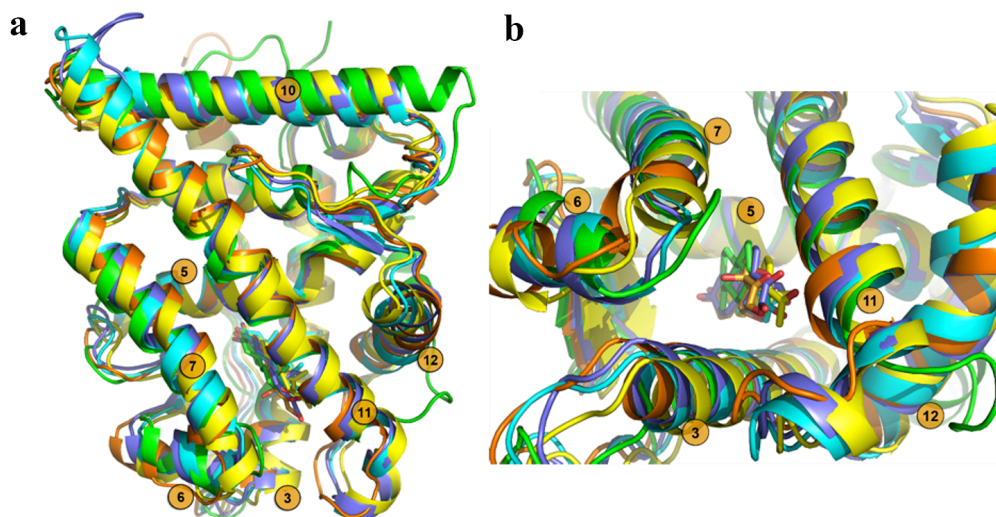


Figure 3. Crystal structures of the NHR family with the endogenous ligand: progesterone receptor (PR; yellow), estrogen receptor  $\alpha$  (green), androgen receptor (cyan), glucocorticoid receptor (orange) and mineralocorticoid receptor (ice blue). Panels (a) and (b) show two different perspectives. Adapted from publication 4.

Still, *X-ray* structures contain implicit information on plasticity. For example, while some parts of complexes may be solved with atomistic accuracy, other components, such as flexible loops, may not. This can be seen with the atomic B-factor descriptor, which accounts for (harmonic) dispersions around the average position. Another approach to study protein plasticity with *X-rays* is the use of the same (or similar) protein crystallized under different conditions, for example with different ligands or cofactors. In the current thesis, we present two articles where we combined this technique with principal component analysis (PCA): in the first, to study a conserved plasticity in the nuclear hormone receptor (NHR) family<sup>19</sup> (publication 2), and in the second, to enhance the sampling of protein flexibility in atomistic simulations<sup>20</sup> (publication 4, Fig. 3).

Since it solved the first protein structure in the mid 80's<sup>21</sup>, nuclear magnetic resonance (NMR) spectroscopy<sup>22</sup> became an alternative to *X-ray* diffraction, providing more evidence of protein flexibility. Its principal advantage over crystallography is that it allows the study protein dynamics in solution, in addition to atomic positions. It permits to obtain dynamics of fast events, such as side chain torsional rotations in picoseconds, and slow ones, such as larger collective motions in seconds (or even days). However, despite different attempts to study larger complexes<sup>23</sup>, its main drawback is the limitation on complex size (roughly 50-60kDa).

A third technique that is gaining popularity in the last decade is the cryo-electron microscopy (EM or cryo-EM)<sup>24</sup> (Fig. 4b). Proteins are placed in a thin layer in their native environment, and they are frozen using a cryogenic bath to be posteriorly studied with an electron microscope. In 2015, Rubinstein's lab published an amazing video of a eukaryotic vacuolar H<sup>+</sup>-ATPase pumping protons out of a cell<sup>25</sup>, which is used to control the pH in intracellular compartments (video: <https://goo.gl/WMgnxe>). The video is derived from three different snapshots with resolutions of 6.9Å, 7.6Å and 8.3Å, respectively. The relatively low resolutions that can be obtained with EM is caused by the alteration of molecules because of high-energy electrons<sup>26</sup>. However, high-resolution structures have already been solved, such as  $\beta$ -galactosidase<sup>27</sup>, suggesting that these limitations may be overcome in the future.

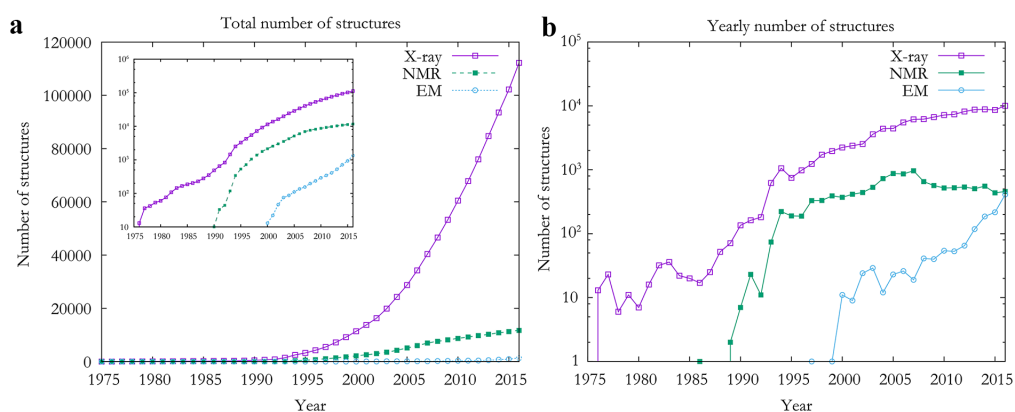


Figure 4. Evolution of available structures in the Protein Data Bank for the three most important experimental techniques in the 1975-2016 period. Panel (a): Total number of structures for each technique, shown in logarithmic scale in the inset box. There is approximately one order of magnitude difference between the three main experimental techniques. Panel (b): Yearly new available structures in logarithmic scale. Interestingly, EM has been gaining popularity in the last decade and solved a similar amount of structures than NMR in 2016. Data source: <https://goo.gl/Cuhwo6>.

The number of available structures in the Protein Data Bank has been growing almost exponentially over the last decades (Fig. 4a), which has allowed a much richer understanding of biological processes. One of the causes is the impressive development of experimental techniques, and future improvements will certainly answer many open questions and let reformulating new ones. However, experiments have important limitations, and computational modeling can be a great asset to complement their shortcomings, giving, for example, a detailed atomistic description of the dynamic processes under study at a variety of timescales<sup>28</sup>. At the time of writing this thesis, Pande and colleagues published an excellent example of this complementation<sup>29</sup>. They studied the induced changes upon binding for different ligands with protein kinase C, a membrane-associated protein. These ligands induce different activities on receptors, and experimental tools only provide partial knowledge due to the inherent problems determining its membrane activated structures. Using experimental structures as a starting point, computational tools have elucidated the full binding process and its consequences, observing different stable complex positionings in the membrane due to ligand-membrane and ligand-water interactions. The particular details may be valuable in the design of new drugs.

Overall, one of the main mindset changes in the last 50 years is the conception of proteins as dynamic entities. There is nowadays enough evidence that proteins are dynamic, and they cannot be adequately understood otherwise<sup>30-32</sup>. For example, in the binding process, the protein might undergo conformational changes, namely with the *induced-fit* and/or the *conformational selection* mechanisms. Thus, if we want to understand the binding mechanism, the *in silico* methods that we use must be precise, and able to handle flexibility accurately. In this thesis, we develop and use such a technique to study protein-ligand interactions.

## Biomolecular Modeling

In order to run computational simulations and study biophysical processes, we first need to build a theoretical model of biomolecules. Depending on the process under study, different approaches may be taken. When electrons are necessary in the description, such as in chemical reactions, quantum mechanics are used, but when they might not be explicitly accounted for, such as in ligand association or protein folding, classical physics are preferred. In this subsection, we briefly review some of the main representations.

### Quantum Modeling

Quantum Mechanics (QM) based methods are grounded on a quantum description at an electronic level. They give the most accurate system description but at a large computational cost. For this reason, QM is usually limited to study processes that cannot be adequately explained with classical physics, such as chemical reactions (*e.g.* involving bond formation/breaking) or interactions involving charge delocalization. As discussed below, these procedures have been applied in a drug design context, motivated by the accurate treatment of protein-ligand interactions<sup>33</sup>. Also, despite not being directly treated in this thesis, the study of electron transfer<sup>34</sup> is of particular interest in our research group and has been used in the (rational) design of enzymes, for example in peroxidases<sup>35</sup> or laccases<sup>36,37</sup>.

A first family of quantum methods is based on the resolution of the Schrödinger equation. Since it cannot be solved exactly for most practical systems, different simplifications and approximations are carried out. For example, the first simplification is only to consider the time-independence of the Hamiltonian, which yields the time-independent Schrödinger equation:

$$\hat{H}\Psi = E\Psi, \quad (1)$$

where  $\hat{H}$  is the Hamiltonian,  $\Psi$  is the wavefunction and  $E$  its associated energy. The Born-Oppenheimer approximation allows separating nucleic and electronic degrees of freedom by considering that the masses of the nuclei are much larger than the masses of the electrons. Also, relativistic corrections are usually not considered. In the Hartree-Fock method<sup>38</sup>, the N-electronic wavefunction is approximated with the product of single electron wavefunctions, and the electron repulsion is taken into account implicitly with a mean-field. Iteratively, an ansatz wavefunction is proposed, which modifies the mean-field, and a new solution that takes it into account can be proposed. The procedure is repeated until self-consistency is achieved, taking advantage of the variational principle in quantum mechanics, which states that the ground state wavefunction will give the lowest possible energy. Post-Hartree-Fock<sup>38</sup> methods include explicit electron correlation with higher order corrections in perturbation theory, which accounts, for example, for dispersion corrections at an increased computational cost.

The previous techniques are coined as *ab initio*, “from first principles” in Latin, given that no parameterized data is used. A second family is the semi-empirical methods, where calculations are speeded up with the use of parameters, coming from either *ab initio* calculations or experimental data. Whereas the computational time difference between quantum and classical methods calculations is of about



six orders of magnitude, with semi-empirical methods the difference is reduced to three, which broadens its applicability<sup>39</sup>, but limits its accuracy.

Density functional theory (DFT) encompasses a different family of techniques. Instead of solving Eq. (1) and finding the wavefunction that describes the system (a  $3N$ -dimensional function, being  $N$  the system size), these methods rely on the Hohenberg-Kohn theorems<sup>40</sup> to find the electron probability density (a 3-dimensional function). These theorems define an energy function that is minimized in the ground state, the latter depending solely on the electron density. The electron density is found using variational principles at a lower computational cost than Post-Hartree-Fock methods.

While these methods provide accurate quantum mechanical descriptions, they are impractical for calculations involving many atoms. Instead, hybrid methods such as Quantum Mechanics/Molecular Mechanics (QM/MM)<sup>41</sup> provide a better trade-off between accuracy and speed. In these, the system is divided into two parts: one using QM, typically the active site, and the rest with molecular mechanics. The treatment of the frontier between both regions is a challenge, and different proposals have been made. In this thesis, we used the B3LYP<sup>42</sup>/6-31 G\*\*+ functionals implemented in Jaguar<sup>43</sup> to perform quantum calculations on ligands to obtain their restrained electrostatic potential charges in the binding environment, combined with a classical description for the protein.

The previous families of quantum mechanical methods have been used in the framework of binding free energy calculations. Recently, Ryde and Söderhjelm wrote an excellent and detailed review on the topic<sup>33</sup>. The rationale behind its use is the rigorous treatment of protein-ligand interactions, especially in cases where charge delocalization is important, which is usually not considered using classical methods (see the following subsection). Nonetheless, accurate free energy estimations rely on sufficient sampling, and the elevated computational cost of quantum calculations is a severe limitation. In view to speeding them up, QM is often combined with MM; for example, using a quantum energetical description with the sampling provided by classical methods. Even so, free energy calculations are still most of the time restricted to end-point approaches (such as Molecular Mechanics/Poisson Boltzmann Surface Area, MM/PBSA, or Molecular Mechanics/Generalized Born Surface Area, MM/GBSA) or even single-structure calculations with only a few examples on more strict methodologies (such as umbrella sampling). By examining blind-test competitions, Ryde and Söderhjelm conclude that, despite major advances, nowadays quantum calculations do not suppose an advantage over classical ones regarding accuracy. They expose three different possible causes: the lack of sampling, the use of implicit solvents, or less error cancellation in quantum calculations with respect to classical ones.

### **Classical Modeling**

Classical molecular mechanics (also known as MM models) ignores electrons and uses a heuristic Hamiltonian to represent biomolecules as a set of interacting spheres that are governed by classical physics. These spheres may correspond to atom nuclei in high-resolution all-atom representations, or larger groups, such as residues or side-chains, in coarse-grained ones. The model is thus greatly simplified compared to those provided by QM, with a subsequent reduction of its computational cost that expands its applications. For example, most ligand association studies or rigorous free energy calculations have been performed

using MM methodologies. However, as we mentioned before, the simplification of the electronic degrees of freedom makes this approach unsuitable to study changes in the electron distribution, where QM or QM/MM techniques are preferred.

We owe to Shneior Lifson the basic formulation, known as Force Field (FF), at the birth of molecular simulations<sup>44</sup>. FFs consist of a parameterized energy function fitted to QM and/or experimental data, and aim to be general and **transferable** to describe different (but related) systems, which makes them appropriate for predictions<sup>45</sup>. This is achieved by assigning the same parameters to atoms with similar characteristics (*e.g.* atomic number, formal charge, bond order...), labeled with different atom types (complete list for std. AMBER: <https://goo.gl/HA2gJp>). However, transferability may not apply when the atoms to parameterize have not been in the training set, as may happen with ligands. For this reason, there have been efforts to generalize FF to small molecules, such as GAFF<sup>46</sup>, CGenFF<sup>47</sup>, GAAMP<sup>48</sup> or OPLS3<sup>49</sup>:

Popular FF choices for biomolecules include OPLS<sup>50</sup>, AMBER<sup>51</sup>, CHARMM<sup>52</sup> or GROMOS<sup>53</sup>. Their energy function is divided into two parts: the bonded and the non-bonded interactions. The bonded interactions account for covalent bonds and angular bends with harmonic potentials, and torsions with Fourier expansions. The non-bonded interactions consist of electrostatic and Lennard-Jones potential terms. We illustrate the Hamiltonian with the OPLS2005<sup>50</sup> FF:

$$E = \underbrace{\sum_{\text{bonds}} K_r(r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{torsions}, n} K_{\phi, n}[1 + \cos(n\phi_i - \delta_n)]}_{\text{Bonded}} + \underbrace{\sum_{i < j} k_e \frac{q_i q_j}{r_{ij}} f_{ij} + \sum_{i < j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] f_{ij}}_{\text{Non-bonded}} \quad (2)$$

where  $K_i$  are parameters representing the rigidity of bond length, angle bending and energy barriers in torsions, for  $i=r, \theta, \phi$ , respectively. Bond distances correspond to  $r$  and angles to  $\theta$ , with subscript “*eq*” for the equilibrium values (in isolation, when no other forces are present). The indices  $n$  and  $\delta_n$  correspond to the periodicity and phase of the torsion angle  $\phi$ , where the summation includes improper torsions to enforce planarity. In the non-bonded terms, the factor  $f_{ij}$  is 0 if the  $i$ -th and  $j$ -th atoms are separated by a distance of one or two bonds, 0.5 if they are separated by 3 bonds, and 1 otherwise. In the first non-bonded term, the Coulombic interaction, charges are denoted with  $q$ , the Coulombic factor with  $k_e$  and the distance between the  $i$ -th and  $j$ -th atoms with  $r_{ij}$ . The second term is the Lennard-Jones potential, and the parameters  $\epsilon_{ij}$  and  $\sigma_{ij}$  characterize its depth and average distance. It characterizes the repulsive forces of the Pauli exclusion principle and the London dispersion forces that result from mutually induced dipoles in atoms. The FF function is continuous and permits analytical second derivatives, so forces and Hessian can be derived to perform force calculations and minimizations, respectively.

Computationally speaking, the non-bonded are the most expensive interactions, because they involve the calculation of atom pairs. Especially problematic is the electrostatic interaction due to its slow distance decay: the potential decays with  $r$ , but the number of interacting particles grows with  $r^{d-1}$ , and simply assuming a cutoff may drive to artifacts. Fortunately, there are several workarounds to avoid the calculation of all atom pairs<sup>54</sup>. Explicit solvent molecules (see below) are often

simulated with periodic boundary conditions, and a common choice to compute long-range interactions is the particle mesh Ewald<sup>55</sup>, where the contributions within a cutoff are calculated in the direct space, and those beyond the cutoff in the reciprocal space, where they converge, using fast Fourier transform. Another option when periodic boundary conditions are not present is the Fast Multipole Method, which is based on the expansion of distant charges as a multipole series, that are converted to local field expansions with which distant particles interact<sup>56</sup>.

One of the main inaccuracies of classical FFs is the overlook of polarization, where partial charges do not vary according to the diverse charge environments and represent only a mean-field view. Actually, heterogeneous surroundings may induce different charge distribution on molecules, and accurate modeling ought to reflect them. There are different approaches to account for polarization, such as the Drude or the induced dipole model, which have been shown to yield accurate predictions of binding free energies<sup>57,58</sup>. The main drawback is that dealing with polarization efficiently, from a computational point of view, is still an open challenge<sup>59</sup>.

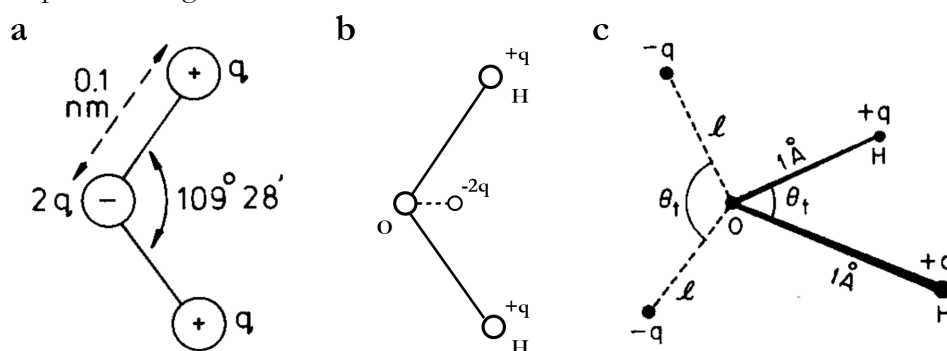


Figure 5. In panels (a,b,c) SPC, TIP4P and ST2 explicit water models, with 3, 4 and 5 interaction sites, respectively. Sources: panel (a) Ref. 60, and panel (c) Ref. 61.

The modeling of the solvent, protein's aqueous environment, deserves special attention. In MM, it may be represented either explicitly or implicitly. The explicit molecular representation permits studying its local properties, *i.e.* the effects of particular water molecules, such as in the bridging of interactions through hydrogen bonds or water clusters in cavities. Importantly, water is characterized by its ability as a donor and acceptor of hydrogen bonds, which can be (to some extent) explained with electrostatic attraction and electronic repulsion<sup>62</sup>, and is responsible for its extended networks<sup>63</sup>. Despite the apparent simplicity, the large number of available representations serves as a measure of the difficulty of its modeling<sup>62,64</sup>. Those include models with charges in the nuclei (*e.g.* SPC<sup>60</sup>, Fig. 5a), in fictitious interaction sites (*e.g.* TIP4P<sup>65</sup>, ST2<sup>61</sup>, Fig. 5b and 5c), polarizable (*e.g.* TTM3-F<sup>66</sup>, DPP2<sup>67</sup>, AMOEBA<sup>68</sup>), and molecules can be either rigid or flexible. These are fitted to reproduce experimental data, such as the density, radial distribution, diffusion coefficient, heat of vaporization or dielectric permittivity, just to name a few.

An alternative with a reduced computational cost is the implicit solvation, where the solvent is modeled with a continuous field that accounts for a thermal average interaction. It needs fewer calculations than the explicit representation, not only because of the smaller number of particles, but also because its behavior is already averaged out, and the relaxation time is instantaneous. In addition, the lack of steric clashes eases the design of stochastic algorithms and enhances the

search in algorithms based on equations of motion. Its major shortcoming, however, is that it cannot explain effects where local fluctuations are important, for example the formation of solvent-solute hydrogen bonds or in the distribution of waters within cavities.

Implicit models are often based on the assumption that the solvation free energy can be split into two terms:

$$\Delta G_{sol} = \Delta G_{pol} + \Delta G_{np} \quad (3)$$

The polar term is the reversible work to build the electrostatic interactions, and the non-polar is often split into a cavitation term that is reversible work to create the cavity in the solvent (to host the solute), and a van der Waals term, which accounts for dispersion and repulsion forces<sup>69</sup>.

The  $\Delta G_{pol}$  is computed obtaining the electrostatic potential with the Poisson-Boltzmann (PB) equation:

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \phi(\mathbf{r})] = -4\pi \rho(\mathbf{r}) \quad (4)$$

where  $\epsilon$  is the dielectric constant,  $\phi$  is the electrostatic potential, and  $\rho$  is the charge density. There are different approximations to the solution (see Ref. 69 for an exhaustive review), and a popular choice is the Generalized Born<sup>70</sup> (GB). The GB model uses the first term of the multipole expansion (Born approximation) and assumes that charges are distributed in the center of cavities (atom nuclei) with an internal dielectric  $\epsilon_{in}$  and an effective Born radius,  $\alpha$ . The latter depends on the particular molecular conformation: it corresponds to the van der Waals radius for atoms in isolation, and it becomes much larger for buried atoms. The  $\Delta G_{pol}$  is then written as the sum of Coulomb interactions in vacuum and the Born term (Eq. 3 to 5 in Ref. 70).

Following the GB approximation, in our software, PELE, we employ a variable dielectric model (VDGBNP)<sup>71</sup>. The motivation is the following. In non-polarizable FFs, such as OPLS2005, the parameterization of groups is carried out within neutral environments, which do not account for larger polarizations that may arise in the interaction with charged groups (*e.g.* in hydrogen bonds and or salt bridges). Thus, we empirically correct the dielectric values depending on the charge of the involved atoms (*e.g.* from  $\epsilon_{in}=1$  for non-charged side chains or backbone atoms, up to  $\epsilon_{in}=4$  for side chain atoms in lysine). This approach has been shown to improve the performance of side chain and loop predictions<sup>71</sup>.

All in all, the polar contribution can be written as:

$$\Delta G_{pol} = - \sum_{i \leq j} \left( \frac{1}{\epsilon_{in}(ij)} - \frac{e^{-\kappa f_{GB}}}{\epsilon_{solv}} \right) \frac{q_i q_j}{f_{GB}} \quad (5)$$

where  $\epsilon_{in}(ij)$  is the variable dielectric,  $\epsilon_{solv}$  is the solvent dielectric, and  $q$  is the charge. The exponential factor,  $e^{-\kappa f_{GB}}$ , improves the agreement with PB in the presence of ion screening, which extends the use of VDGBNP to charged polymers such as DNA chains<sup>72</sup>. Finally, the smoothing term,  $f_{GB}$ , is defined as:

$$f_{GB} = \sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp\left(-\frac{r_{ij}^2}{4\alpha_i \alpha_j}\right)} \quad (6)$$

When the distance between atoms,  $r_{ij}$ , is small,  $f_{GB} \approx \sqrt{\alpha_i \alpha_j}$ , and when it is large,  $f_{GB} \approx r_{ij}$ .

In GB surface area models (GB/SA), the  $\Delta G_{np}$  term is approximated to a linear function of the atomic solvent-accessible surface area (SASA):

$$\Delta G_{np} = \sum_i [\gamma_i \text{SASA}_i + S(\alpha_i)] \quad (7)$$

where the summation runs over all atoms,  $\gamma$  is a proportionality factor, and  $S$  is a switching function depending on the Born radii and is intended to represent the van der Waals interactions between the solvent and solute. VDGBNP uses Eq. 7 to account for the non-polar interaction, using the parameterization of Gallicchio and coworkers, which has been fit to experimental hydration free energy values of small organic molecules<sup>73</sup>.

The latest release of VDGBNP, VSGB 2.0, adds ideas from docking scoring functions. Several empirical corrections have been added to Eq. 5 to reproduce more accurately hydrogen bonds,  $\pi$ - $\pi$  packings, and self-interactions of Asn, Gln, Ser and Thr side chains with their own backbone, which, according to the authors, cannot be correctly reproduced with solely single point charge electrostatics and van der Waals effects<sup>74</sup>. The non-polar dependency on the SASA has been replaced by an all atom-pair sum that favors the packing, favoring dispersion interactions rather than penalizing hydrophobic solute contacts. Altogether provides a more accurate physical description with an improved energy estimation, which, in our experience, is reflected in the better correlation with experimental free energy values of MM/GBSA implemented in different versions of Maestro<sup>75</sup> (results not shown).

In the variable dielectric models, the Born radii are estimated with the contribution of each atom to the SASA, which involves surface element evaluations for each (neighboring) atom pair<sup>76</sup>. A faster alternative replaces the surface evaluation for a van der Waals volume integration using a parameterized set. Based on this idea, Hawkins and colleagues proposed the HTC model<sup>77</sup>. The volume integration is carried out with a pairwise-descreening approximation (assuming no van der Waals radii overlap), and a set of parameters was derived to reproduce the Born radii for small molecules. However, the assumption of no overlaps results in a poor estimation for deeply buried atoms. Later, the so-called OBC model corrected the functional form for the Born radii to show better agreement with PB for buried atoms, permitting its use with large macromolecules<sup>78</sup>. In our OBC implementation in PELE, following TINKER's<sup>79</sup> approach, we avoid the calculation of the SASA using the non-polar term of the ACE model (Eq. 2 in Ref. 80), which results in a considerable speedup with respect to VDGBNP.

The combination of Eq. 2 and Eq. 3 (the corresponding approximation of either VDGBNP or OBC) constitutes the Hamiltonian that we used to perform the present dissertation on protein-ligand interactions. The theoretical model is a tradeoff between accuracy and speed. It permits a *sufficiently* accurate description in a *reasonably* short computational time, from the point of view of lead optimization in drug design. The Hamiltonian provides the energy for each specific set of molecular coordinates. In the following section, we will overview the two main techniques to obtain the sampling.

## Molecular Simulations

Molecular simulations aim to mimic biomolecules computationally to elucidate biophysical/biochemical processes, such as protein-ligand association. They aim to generate atomistically detailed information of events occurring *in vitro* or *in vivo*. For example, appealing videos may be obtained, often used in “look and see” studies, where computers can be thought as computational microscopes. More importantly, molecular simulations can provide an ensemble of conformations that permits obtaining quantitative information, such as the binding free energy. In this section, we will overview the two most popular approaches to study ligand binding: classical Molecular Dynamics (MD) and Monte Carlo (MC) methods. We will devote particular emphasis to PELE, our in-house MC algorithm, which we used to perform the sampling.

### Molecular Dynamics

Classical MD is based on the iterative numerical resolution of Newton’s equations of motion and is the most common approach to simulate protein-ligand dynamics. MD is also often used to describe equilibrium properties such as free energy differences or to explore the energy landscape, usually in combination with other techniques such as simulated tempering<sup>81</sup> or replica exchange<sup>82</sup>.

The starting coordinates are usually obtained with either experimental techniques (see Proteins section) or homology modeling<sup>83</sup>. Then, the system is initialized to the simulation conditions with a progressive heating to the temperature and pressure (or volume) of interest. After the preparation, an iterative integration process is repeated until the desired simulation length is achieved, which is outlined in the following lines. In the beginning, the Hamiltonian is evaluated, which consists of a FF with either an implicit or an explicit solvation. Then, the forces that act on each atom nuclei are computed taking the Hamiltonian’s gradient. Finally, the equations of motion are integrated using a *small* time step. It is crucial to use a small time step in order to bound the integration error, which may otherwise result in the violation of conservation laws, such as the energy in the NVE ensemble. The limiting factor is atomic motions, being stretching motions involving hydrogens the fastest ones due to its small mass, and restricts the time step to 1 fs. Algorithms such as SHAKE<sup>84</sup> constrain hydrogen bonds and allow the use of time steps of 2 fs. Vibrations can be further reduced with the use of dummy atoms, where hydrogen mass is transferred to their respective bonded heavy atoms, which extends the time step to 4 fs<sup>85</sup>. Also, different ensembles can be simulated, such as NVE forbidding heat exchange with the environment, NVT with the addition of a thermostat, or NPT with the addition of a thermostat and a barostat. The outcome of the procedure is the evolution of the positions over time, the trajectory.

The ergodic hypothesis is assumed. Taking averages at different times in the trajectory becomes equivalent to taking ensemble averages, and trajectories thus represent the conformational ensemble<sup>86</sup>. In this sense, MD experiments are similar to those in the laboratory: after an initial preparation in the desired conditions, measures are obtained at different times, which will suffer from statistical fluctuations. The final measure is an average with a certain uncertainty that can be arbitrarily reduced by taking more measures (*i.e.* running longer simulations). Note that slight modifications of the initial conditions will result in completely uncorrelated trajectories after a certain simulation time (Lyapunov

instability)<sup>54</sup>. These trajectories are different from a numerical point of view, but similar from a statistical perspective, as a consequence of the ergodic hypothesis. This phenomenon makes software testing harder, as launching two simulations under the same initial conditions in two machines with different library implementations or different number of decimal floating points produces two numerically different trajectories, and checking for statistical equivalence is not straightforward (see Michael Shirts's proposal in the next paragraph). A common solution is the use of the same testing environment, for example with virtual machines.

Because of maybe erroneous approximations (*e.g.* too much integration error, too short cutoffs, inaccurate thermostats...) or potential bugs in the software, algorithms should be validated. For example, the total energy should be conserved in the NVE ensemble, the instantaneous pressure and temperature in the NPT, velocities should sample the Maxwell-Boltzmann distribution in simulations with thermostat... Michael Shirts proposed a more sophisticated validation test that can be performed with the program "checkensemble" (<https://goo.gl/0ngROX>). It measures the likelihood that a particular simulation protocol follows the Boltzmann distribution, and that deviations from it are caused by statistical fluctuations<sup>87</sup>. This is achieved by comparing the ratio of Boltzmann weights for the same value of an extensive property (*e.g.* potential energy) and two different values of a parameter (*e.g.* temperature) such that its dependence on an intrinsic system characteristic (*e.g.* density of states) cancels out and the ratio has a linear dependence on the user-selected parameter. For example, by analyzing 20 ns simulations of 900 TIP3P<sup>65</sup> water molecules at two different temperatures, Shirts found out that the Berendsen thermostat<sup>88</sup> fails considerably to sample the Boltzmann distribution in the NVT ensemble.

## Monte Carlo

The origin of MC dates back to the post-World War II period in Los Alamos National Laboratory. While playing Canfield Solitaire, Stanislaw Ulam was thinking on the probability of drawing a successful card combination. After an initial struggle with the combinatorial problem, he envisioned the calculation of probabilities with a much simpler approach. He would use a counting scheme instead, repeating the game one hundred times and obtaining statistical sampling. This perfectly captures the spirit of MC, named after Ulam's uncle by Nicholas Metropolis, who would not hesitate to borrow money to go to the casino in Monte Carlo. Ulam and John von Neumann rapidly saw its potential and took advantage of the advent of the first electronic computer, ENIAC, to predict neutron chain reactions in nuclear fission. Shortly after, it would be applied in a broader range of problems that could not be solved with theoretical approaches. A detailed historical review of this exciting period is presented in a special issue of Los Alamos Science devoted to Ulam<sup>89</sup> (<https://goo.gl/7LhfgM>). Nowadays MC is a general term that comprises a broad spectrum of methodologies that use statistical sampling to perform numerical calculations in computer simulations. From the point of view of biomolecular modeling, it often regards those that make use of Markov chains to perform Boltzmann sampling, known as Markov Chain Monte Carlo (MCMC), but can be used to perform low energy conformation searches<sup>90,91</sup>.

In protein-ligand binding simulations, the high dimensionality of the phase space does not allow the direct calculation of the partition function, so the absolute

probability distribution is unknown. Nonetheless, we do know the relative probabilities between any pair of states  $i$  (initial) and  $n$  (new):

$$\frac{\pi_n(\beta)}{\pi_i(\beta)} = e^{-\beta(E_n - E_i)} \quad (8)$$

where  $\pi$  is the probability,  $\beta$  is the inverse temperature, and  $E$  is the internal energy. Then, our strategy will be to perform a random walk respecting their relative probabilities, which will concentrate the sampling in regions that have a significant statistical weight<sup>92</sup>.

In order to derive the transition probabilities,  $p_{in}$ , we use the fact that once the system is in the stationary state, the Boltzmann distribution, transition probabilities should not destroy it. In other words, the average flux leaving a state  $i$  must be equal to the flux coming from any other state to it, which is known as the (global) balance condition:

$$\pi_i \sum_j p_{ij} = \sum_j \pi_j p_{ji} \quad (9)$$

Imposing the more strict pairwise cancelation of terms yields the sufficient but not necessary condition of equilibrium known as detailed balance:

$$\pi_i p_{in} = \pi_n p_{ni} \quad (10)$$

One can easily find cases where detailed balance is not fulfilled despite the stationary distribution is reached. The system in Fig. 6a is one of them; it is in equilibrium, as it does not have any sink or sources of flux, but pairwise fluxes are not canceled out, and therefore detailed balance is not satisfied. However, detailed balance is still encouraged because it eases the design of algorithms and it ensures the elimination of any possible systematic error that would not be removed with more sampling<sup>54</sup>.

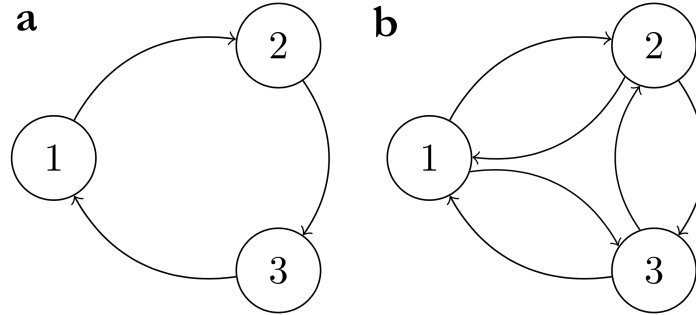


Figure 6. Three-state systems and their stationary transition fluxes represented as arrows, which are assumed of equal magnitude. Despite being in equilibrium, the system in panel (a) satisfies global balance, but not detailed balance. The system in panel (b) is in equilibrium and fulfills both global and detailed balance.

The transition probability,  $p_{in}$ , is given by the probability to propose that transition,  $\alpha_{in}$ , times the probability to accept it,  $\text{acceptance}_{in}$ :

$$p_{in} = \alpha_{in} \cdot \text{acceptance}_{in} \quad (11)$$

Combining Eq. 8, 10 and 11, we obtain:

$$\frac{\text{acceptance}_{in}}{\text{acceptance}_{ni}} = \frac{\alpha_{ni}}{\alpha_{in}} e^{-\beta(E_n - E_i)} \quad (12)$$



A choice that satisfies Eq. 12 is an extension of the Metropolis criterion<sup>93</sup>:

$$\text{acceptance}_{in} = \min \left( 1, \frac{\alpha_{ni} e^{-\beta(E_n - E_i)}}{\alpha_{in}} \right) \quad (13)$$

Starting in an initial configuration,  $i$ , we propose a new state,  $n$ . Then, we generate a random number between 0 and 1. The transition is accepted if the number is smaller than the acceptance criterion (Eq. 13); otherwise, it is rejected and the procedure starts back from state  $i$ . Note that if there is no bias in the proposal, such as in purely random proposals,  $\alpha$  will be symmetrical, and  $\alpha_{ni}/\alpha_{in}=1$ .

A major difference to MD is that transitions do not need to have a physical meaning. This results, at least theoretically, in a more efficient exploration of the phase space, as energies can be merely crossed rather than surmounted by thermal fluctuations. For example, let us imagine the exploration of an energy landscape with two states at  $r=A$  and  $r=B$  separated by an energy barrier, where  $r$  is a generic coordinate (Fig. 7a). Using MD, we would need to wait for thermal fluctuations to drive the crossing from  $r=A$  towards  $r=B$ , which takes an exponential time of the barrier height. On the contrary, in MC we could simply propose a displacement  $\Delta r=B-A$  and reach it in a single step. In publication 5, we show a real case application in a cross-docking study. We used Glide<sup>94</sup> to dock the ligand from PDB ID:3KBA into the PR protein with PDB ID:1A28. Glide succeeded to find the deeply buried binding site but failed to reproduce the binding pose, where the ligand was rotated 180° with respect to the native (Fig. 7b). Using an MC that allowed proposals with large rotations, we were able to recover the experimental pose in a single MC step, despite the large energy barrier due to the limited binding pocket volume.

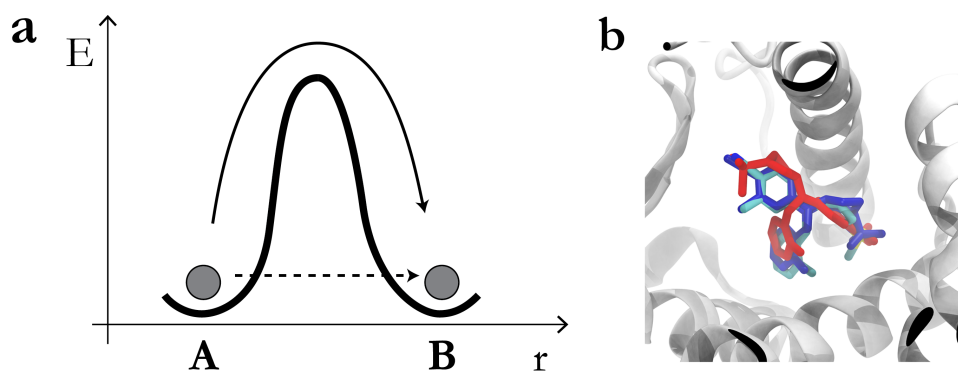


Figure 7. Energy barrier crossing. Panel (a): schematic representation of energy barrier crossing in MD and MC. In the y-axis we represent the energy and in the x-axis a generic coordinate  $r$  with two minima in  $r=A$  and  $r=B$ . MD surmounts energy barriers with thermal fluctuations (continuous line), whereas MC can simply move from  $r=A$  to  $r=B$  (dashed line). Panel (b): MC barrier traversing in cross docking. We show the native structure (atom-type colored ligand and white cartoon protein), the wrongly docked ligand structure (red, flipped 180°), and the resulting after an MC exploration (dark blue). Source: publication 5.

A second key difference with MD is the absence of an absolute time step between transitions. This can be inferred from the previous energy barrier model, where the crossing time is one step regardless of the barrier height. However, there is an underlying concept of the dynamics, as suggested by Rey and Skolnick<sup>95</sup>. They discuss that one could arguably construct transition proposals that capture the main traits of a process under study, and therefore, their fundamental dynamics. Consequently, the total number of MC steps would be

related to the real time. They prove their argument showing identical folding pathways of a  $\alpha$ -helical hairpin using an MC protocol and Brownian dynamics. Shakhnovich and coworkers later used a similar analogy with master-equation formalism to state that MC provides a kinetic coarse-grained picture of the dynamics, which they used to predict folding kinetics<sup>96,97</sup>. Another example is kinetic MC<sup>98</sup>, which uses a similar rationale as Markov State Models (MSM, see Markov State Model section). The energy landscape is seen as a collection of energy basins separated by barriers. When the residence time is long enough, the system eventually forgets its history and can be modeled as a Markov process with transition rates characterizing the jumps between adjacent basins. If these can be fully determined, one can build an MC procedure indistinguishable from MD. In our toy model, if we want to study the energy barrier crossing, we can propose displacements  $\Delta r \ll B-A$  and get an insight of its height, and consequently, its dynamics. Similarly, in a mechanistic study of ligand binding, we can assign small ligand displacements (translations and rotations) avoiding unphysical jumps over protein backbone or side chain, being 1 or 2 Å a sound choice for translations ( $\sim$ van der Waals radii). With these examples we have shown evidence of the connection with dynamics. However, we should emphasize that the analogy is heuristic, and, in general, there is not a correspondence with an absolute time; for example, different choices of  $\Delta r$  or ligand perturbations will result in different kinetics.

The biggest challenge of MC in biophysics is the inherent difficulty of generating uncorrelated protein-ligand poses with a significant statistical weight because only a reduced fraction of the possible perturbations are energetically favorable. The inherent flexibility of both protein and ligand is one of the causes. Internal coordinates take into account the coordinated movement of atoms with torsions and suppose a significant advantage over Cartersians<sup>99-101</sup>. However, random moves easily lead to clashes (*e.g.* see a single backbone dihedral rotation), and proposals are typically reduced to local variations of the initial structure<sup>102-104</sup>. Still, we should remark that their extensive exploration has allowed precise free energy estimations<sup>105,106</sup>. In order to improve the success rate, the cooperative movement of a larger number of degrees of freedom, such as whole protein domains<sup>107</sup>, should be taken into account to enhance the exploration. Following this idea, PELE utilizes protein structure prediction algorithms and can introduce larger conformational changes.

## PELE

The major contribution of PELE<sup>108</sup> is the addition of protein structure prediction algorithms in the MC proposals, which allows an efficient traversal of energy barriers. Importantly, this enables exhaustive protein-ligand sampling, reproducing the conformational selection and induced-fit binding mechanisms. Some of its applications comprise studies of ligand migration<sup>109-111</sup>, induced-fit ligand docking<sup>36,37,112-114</sup> and free energy calculations<sup>19,20,115,116</sup>. PELE has been extensively used throughout the thesis to perform protein-ligand sampling.

In MC algorithms, an energy increase of a few kcal/mol can dramatically reduce the acceptance rate. For example, an increment of 10 kcal/mol yields an average acceptance of only  $4 \cdot 10^{-6}$  % at physiological temperature (Eq. 13). Such a difference can be obtained with only a few degrees of freedom; the bonded coefficients of Eq. 2 have the following orders of magnitude:  $K_r \sim 10^2$

kcal/(mol·Å),  $K_\theta \sim 10\text{-}10^2$  kcal/(mol·degree) and  $K_\phi \sim 1\text{-}10$  kcal/mol. They give an idea of: 1) the importance of coordinated movements, since neglecting a few terms easily results in a rejection, and 2) the severe ruggedness the energy landscape. With this scenario, (at least) two strategies could be followed. In the first one, we could try to overcome energy barriers with a random approach of small variations of the initial structure in order to run a large number of proposals in a given wall-clock time. In a second strategy, to which PELE belongs, we spend computational time to carefully plan MC proposals in order to enhance the acceptance whilst making more uncorrelated proposals. PELE iterations are composed of two main blocks: perturbation and relaxation. In the perturbation, all degrees of freedom are modified in a coordinated fashion, and in the relaxation, the overall structure is locally relaxed to avoid drastic energy increases and to enhance the exploration (see *basin-hopping* methods below). Overall, each PELE step roughly takes around one minute in a single Mare Nostrum III computing core (SandyBridge-EP 2.6GHz).

Originally, PELE only supported OPLS2005 and the implicit solvents SGBNP<sup>76</sup> and VDGBNP<sup>71</sup> (see Biomolecular modeling section). OPLS2005 is derived from the original OPLS-AA<sup>117</sup> and includes ligand support and torsional corrections from Refs. 50 and 118. PELE was posteriorly rewritten into C++ and integrated: AMBER99<sup>119</sup>; AMBER99SB, which includes supplementary backbone torsional dihedral corrections<sup>120</sup>; AMBER99SBBS0, which also contains support for nucleic acids<sup>121</sup>; and the OBC implicit solvent. Besides, the code was parallelized using CUDA and OpenMP, reducing further the wall-clock time.

### Algorithm

At the beginning of the simulation, the complex is initialized. Then, we repeat the MC procedure either until convergence is reached (in sampling simulations) or the desired conformation is found (in search simulations). The MC step is divided into two main stages, perturbation and relaxation. The first is composed of a ligand and a normal mode protein perturbation and the second of a side chain prediction and a minimization (Fig. 8). Then, the proposal is accepted or rejected according to the Metropolis criterion. Optionally in search simulations, trajectories are allowed to communicate between steps, after the Metropolis criterion, to drive the simulation towards a goal.

*Initialization.* This phase is devoted to building the complex and involves constructing the topology and assigning the FF parameters and rotamer libraries. These are already pre-calculated for proteins, RNA and DNA, but should be computed for heteroatoms. In an automatic preparation procedure<sup>112</sup>, we identify the atom type with *hetgrp\_ffgen* from the Schrödinger suite and flexible bonds and rings with MacroModel<sup>122</sup>. We can use the default ones, or use QSite<sup>123</sup> or Jaguar<sup>43</sup> in the Schrödinger suite to compute quantum charges, for example in the binding site. Then, a rigid core, which acts as a protein backbone, is assigned to minimize the longest flexible functional group. Throughout the code, flexible groups are treated in the same way as protein side chains and are also optimized in the side chain prediction. During the execution, each rotatable bond is linked to a pseudo-rotamer library that comprehends all possible dihedrals for a given resolution (e.g. a library with a resolution of 10° contains: 0°, 10°, 20°, ... , 350°). The list of all non-clashing combinations constitutes the rotamer library and is built with an efficient tree algorithm based on dead-end elimination. This rotamer

library solely aims to represent ligand flexibility and remove steric overlaps rather than representing a set of common conformations, as opposed to protein rotamer libraries<sup>124</sup>. Also, note that it may become prohibitive for long flexible groups. For example, a group with five rotatable bonds described with a resolution of  $10^\circ$  results in over 60 million possible rotamers, whereas the library for lysine, with the same number of rotatable bonds, contains around 24 thousand. A workaround to reduce the size is lowering either the resolution or the number of rotatable bonds to be optimized in the side chain prediction (these would still be considered as flexible in the rest of the code).

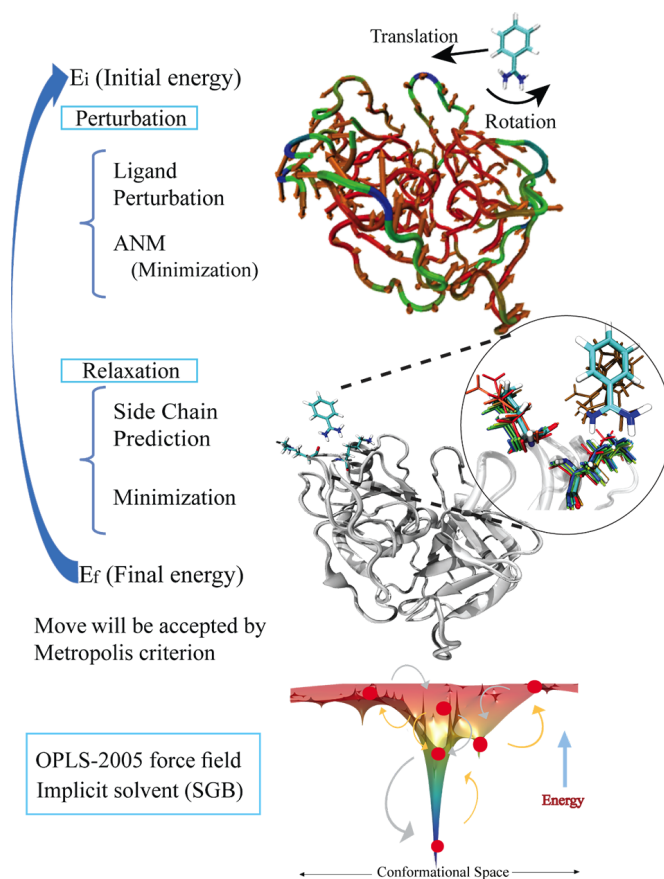


Figure 8. Schematic representation of a PELE step. (Image author: Ryoji Takahashi)

*Ligand perturbation.* The ligand is randomly translated and rotated. If there are clashes, we identify the type of steric overlap, and if it involves a flexible group, we try to relieve it with random combinations of rotamers. On the contrary, if it only implicates core (rigid) groups the movement is discarded and we try a new random translation and rotation. The result of the procedure is a non-clashing random conformation.

When studying binding mechanisms, we try to capture the associated conformational changes (*e.g.* domain movements, side chain bond rotations, hydrogen bond formation...) using small ligand translations ( $\sim 1-2\text{\AA}$ ) and rotations ( $\sim 20^\circ-60^\circ$ ). In search studies, where we are only interested in the final pose, we use arbitrarily large values instead, and also, the whole ligand perturbation is repeated several times with the same initial conformation ( $\sim 5-10$ ), selecting the lowest energy pose for the next phase.

*Normal mode perturbation.* The protein is perturbed following a normal mode analysis (NMA)<sup>125</sup> procedure, the anisotropic network model (ANM)<sup>126</sup>, where the low-frequency modes have been shown to describe global motions accurately<sup>127</sup> and provide an “extremely correct picture” of protein dynamics<sup>128</sup>. We lately included an internal coordinate NMA (IC-NMA)<sup>129</sup> protocol, but the integration in PELE is still in progress, and currently it only supports proteins. Both methods are thoroughly discussed in Victor Gil’s Ph.D. thesis<sup>130</sup>.

We model protein motions with the fluctuations around an equilibrium position of a coarse-grained elastic network. In the ANM, the elastic network connects neighboring  $C_\alpha$ ’s with harmonic potentials, which are centered at an equilibrium distance (*e.g.* experimental result) and have a constant force constant,  $k$ . In our implementation<sup>131</sup>, these  $k$  are chosen depending on the  $C_\alpha$ - $C_\alpha$  distance,  $r$ , as it improves the correlation with experimental B-factors<sup>132</sup>. As derived in literature<sup>125</sup>, displacements are described by normal modes, and their frequencies are computed diagonalizing the Hessian matrix.

$C_\alpha$  displacement proposals are computed as a random linear combination of the lowest frequency normal modes ( $\sim 6$  modes). The sense of the movement is randomly chosen, and the magnitude of the displacement is  $\sim 1$  Å. Transferring the coarse-grained proposal to the complete all-atom model is not direct. In our workaround we include harmonic restraints in each  $C_\alpha$ ’s towards the proposed positions, followed by the application of a minimization procedure. As a result, the protein is perturbed in the chosen direction while keeping the covalent structure of the molecule.

The IC-NMA model describes protein motions with backbone dihedrals rather than with  $C_\alpha$  Cartesian coordinates<sup>129</sup>. Each residue is divided into two subunits, and all those neighboring subunits are connected with an elastic network of harmonic potentials. In an analogous procedure to the Cartesian coordinate ANM (from now on, simply ANM), rotational displacements are found diagonalizing the Hessian matrix.

The sense of the rotation is chosen randomly, and the maximum angle is typically  $\sim 5^\circ$ . As opposed to the ANM, applying the coarse-grained proposal is straightforward. However, little backbone dihedral rotations may easily result in steric clashes, and special care must be taken with packed side chains. Therefore, torsional rotations are applied in an iterative procedure of small increments, and side chains are able to readapt in-between. Importantly, it does not require of backbone minimizations, which has significant implications in the sampling, as discussed below. Also, IC-NMA takes a better account of the collective behavior of proteins, which results in lower energy increments in the perturbation compared to ANM. Once it is fully integrated into PELE, it will allow removing minimizations, a faster protein-ligand sampling and a better coverage of the protein conformational space, as we show in Ref. 129.

*Side Chain Prediction*<sup>71,133,134</sup>. The resulting structure after the perturbation needs to be refined. For example, hydrogen bonds, salt bridges, and  $\pi$ - $\pi$  stackings<sup>a</sup> may have been broken or may be potentially formed, in other cases, soft steric clashes need to be released, and flexible groups need to accommodate to the new

---

<sup>a</sup> These are poorly captured by classical FF unless correction terms are added, such as in VSGB 2.0 (see Biomolecule Modeling section).

scenario. Also, transient openings may be crucial for the binding, as we discuss in publication 5 in the binding of tiotropium to an M3 muscarinic acetylcholine receptor. The side chain prediction is responsible for such task.

The underlying idea of the side chain prediction algorithm is that performance is only slightly degraded by breaking down the hard combinatorial problem of all possible conformations of all flexible groups into a much easier linear one, where the prediction is made with discretized proposals, a list of pre-selected rotamers, applied group by group whilst keeping the rest fixed<sup>135,136</sup>.

We predict the conformation of all those flexible groups that either lay in the vicinity of the ligand ( $\sim 6\text{\AA}$ ) or have undergone a large energy increase in the previous perturbation step. In the first place, if the starting dihedral conformation is not in the rotamer library, it is added. Then, initial conformations are optionally randomized, and, in an iterative procedure that continues until convergence (or until an upper limit of iterations is reached), single side chain predictions are sequentially performed maintaining the rest of groups fixed.

We now outline the single side chain prediction. The energy is calculated for all non-clashing rotamers. Then, starting from the lowest energy rotamer, known as the representative, we build clusters in an incremental procedure, adding iteratively all those that are close to any cluster element. Once a cluster is finished, if there are unassigned rotamers, the process is repeated starting a new cluster with the remaining lowest energy rotamer. Typically 1-3 clusters are generated (Fig. 9). Finally, the Boltzmann contribution is calculated for each cluster and the predicted rotamer is the representative of the cluster with the largest Boltzmann weight. The final pose is discretized according to the resolution of our library and needs to be refined to describe the continuous energy landscape accurately. Smoothly varying libraries would already account for this fine-tuning<sup>137,138</sup>.

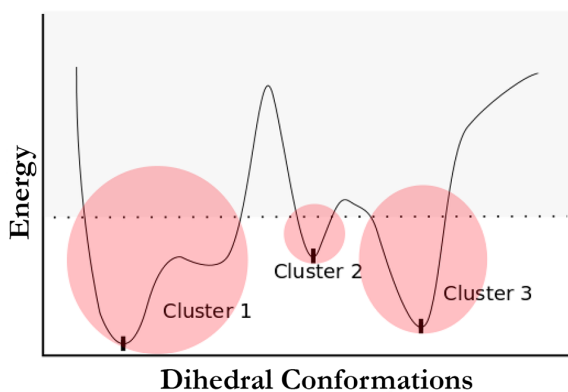


Figure 9. Rotamers represent the possible dihedral conformations (x-axis), and the energy (y-axis) is evaluated for all those that do not have steric overlaps (white background). Rotamers are clustered according to dihedral similarity (red circles), and the prediction corresponds to the representative of the cluster with largest Boltzmann weight.

*Minimization.* In this phase, the resulting structure is minimized with a multiscale truncated Newton (TN)<sup>139</sup>, which is based on the TNPACK implementation<sup>140</sup> and responds to the necessity of quick energy minimizations. Specifically, it is two times faster than the original TNPACK and about one order of magnitude faster than standard procedures such as the conjugate gradient<sup>141</sup>. This speedup is achieved dividing the interactions into short and long-range. The long-range

interactions are assumed to change more slowly than the short ones and are thus updated less frequently. For example, we only compute the long-range energy gradient about 1 to 3 times per minimization whereas the short-range is computed up to 65 times. The corresponding long-range energy contribution is estimated with a first order Taylor expansion around the last updated position, assuming its Hessian (second derivatives) to be 0. Born radii are recomputed periodically in a similar fashion and presumed constant throughout the rest of the minimization.

The multiscale minimization continues either until a maximum number of iterations is reached, or the root-mean-square gradient (RMSG) or the energy difference between iterations fall below a threshold value. In this thesis we solely use the RMSG to tune the convergence criterion.

*Communication (after the Metropolis criterion).* In PELE, rather than characterizing the energy landscape with a unique long simulation, we use an ensemble of tens or hundreds of trajectories to reduce the wall-clock time. In the particular case of search simulations, we can characterize the search goal with a reaction coordinate and exploit the communication between trajectories to concentrate the sampling in the region of interest. In this optional phase, the trajectories that are far from the goal according to a user-defined criterion are spawned to the pose with the best reaction coordinate, which generally speeds up the search but has two important flaws. First, it may impose a severe bias towards the best pose and the exploration may easily become trapped in metastable minima. Secondly, it is not a valid sampling procedure and can only be used in search simulations, as thermodynamic information, such as free energies, cannot be extracted. In this thesis, we introduce an iterative sampling procedure of adaptive simulations that successfully addresses these two issues (publication 5).

## Binding free energy

Having summarized the sampling protocols, we proceed to overview an important thermodynamic concept that helps us characterize the protein-ligand binding process: the absolute binding free energy. Zhou and Gilson wrote an excellent and didactic summary on non-covalent binding thermodynamics<sup>142</sup>, and the reader may refer to it for a deeper analysis. Posteriorly in this section, we present some experimental and computational techniques to carry out its calculation.

### Absolute binding free energy

Let us consider a 3-species solute with concentrations  $C_i$ , with  $i = P, L, PL$ , standing for protein, ligand, and complex, respectively. It can be shown that the free energy change upon binding at constant pressure can be expressed as:

$$\Delta G_b = -k_b T \ln \left( \frac{C_P C_L K_a}{C_{PL}} \right) \quad (14)$$

where  $k_b$  is the Boltzmann constant,  $T$  the temperature, and  $K_a$  the binding constant.

$K_a$  is a relevant thermodynamic magnitude that corresponds to the ratio of species concentrations at equilibrium,  $\Delta G_b = 0$ .  $K_a$  is the inverse of the dissociation constant. The latter is commonly used in experiments, and is equal to the (non-complexed) ligand concentration when the protein and complex concentration are the same. Therefore, the lower the dissociation constant, the higher is the affinity between the protein and the ligand.

The binding free energy can be also written as:

$$\Delta G_b = \Delta H - T\Delta S \quad (15)$$

where  $\Delta H$  is the enthalpy change and  $\Delta S$  the entropy change. Consequently, minimizing the enthalpy and maximizing the entropy upon binding can optimize the protein-ligand binding affinity. The enthalpy change is typically associated with electrostatic and van der Waals interactions and usually favors the binding, for example with hydrogen bond or salt bridge formations. This term favors ligands with stronger interactions in the binding pose compared to the bulk solvent. On the other side, the entropy change is related to the loss (*e.g.* ligand translational or rotational entropy) or gain (*e.g.* water release<sup>143</sup> or backbone low-frequency vibrations<sup>144</sup>) of degrees of freedom. For example, because they suffer a minor entropy loss, rigid ligands are often found to bind more tightly than flexible ones<sup>145</sup>. Docking programs (see below) sometimes overlook entropy in the characterization of poses, but it may be a strong component, such as in HIV protease inhibitors<sup>146</sup> (Fig. 10).

In view to compare affinity measures from different sources<sup>147</sup> (*e.g.* computational or experimental), we use a concentration-independent measure. Specifically, we work with the absolute binding free energy, which is the free energy change in standard conditions when the three species have a standard concentration  $C^\circ$  (= 1M):

$$\Delta G_b^\circ = -k_b T \ln (C^\circ K_a) \quad (16)$$



In an abuse of the language, the absolute binding free energy is often referred in the literature as the binding free energy,  $\Delta G_b$ , or  $\Delta G$ , and in this work, we follow the convention unless otherwise stated.

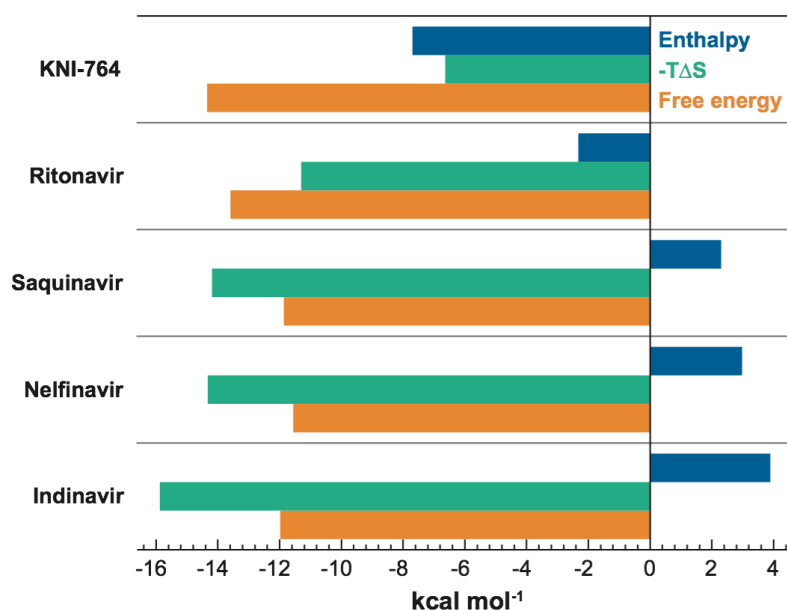


Figure 10. Free energies, enthalpies and  $-T\Delta S$  for five HIV protease inhibitors. The entropic term may be much stronger than the enthalpic one, and it should not be overlooked. Source: Ref. 148.

### Experimental estimations

There are several different experimental techniques to measure thermodynamic properties of protein and ligand association. These measures serve as a reference to test our computational estimations. In this subsection, we review some of them.

Isothermal titration calorimetry<sup>149,150</sup> (ITC) is a gold standard because it allows measuring  $K_d$ ,  $\Delta G$ ,  $\Delta H$  and  $\Delta S$  at the same time. In this technique, the ligand is titrated into a protein solution, generating or absorbing heat. This heat variation is measured by the calorimeter while maintaining a constant temperature. Then, these data are fitted to obtain  $K_d$  and  $\Delta H$ . Finally,  $\Delta G$  and  $\Delta S$  are found by means of Eq. 14 and 15. ITC has many advantages, such as its precision or its capability to study the association in their native aqueous environment. Regarding its precision, in a blind experiment involving 17 groups<sup>151</sup>, it was found that the errors in  $K_d$  and  $\Delta H$  were in the order of  $\sim 24\%$  when comparing among different groups. Note that this translates in much smaller errors in  $\Delta G$  due to the logarithm. This value is much larger than the experimental error provided by the individual laboratories and is most likely due to an underestimation of the concentration error. ITC also has some inconveniences. For instance, it is hard to distinguish the heat exchange from the binding from that coming from other interactions<sup>152</sup> or it may be difficult to assess binding free energies for  $\Delta G > -5$  kcal/mol, due to the significant signal to noise ratio<sup>153</sup>.

Surface plasmon resonance<sup>154,155</sup> (SPR) is an optical method that is based on measuring the refraction index change near a metal surface, *e.g.* gold, and can measure binding kinetics in addition to affinities. One binding partner, which acts as bait, is kept fixed on a sensor, whereas the other is micro-flowed over the surface and binding events are detected as changes in the reflected angle. As the

bait starts to be exposed to the flow, the refraction index steadily changes, and the association rate index can be measured. Once equilibrium is reached we can obtain the binding affinity, and when the surface ceases to be exposed to the flow, the dissociation rate can be measured.

A third method is fluorescent polarization<sup>156,157</sup> (FP) and is based on the idea that when a fluorescent ligand is excited by polarized light, its emission becomes more rapidly unpolarized in unbound ligands than in bound ones. FP has many advantages, for example the reasonable costs compared to ITC or SPR, the relatively easy automatization makes it suitable for high-throughput screening (HTS), it does not require of large samples to obtain measures, or it provides non-destructive measurements. Notwithstanding FP can also provide enthalpy and entropy measures, they are less reliable than those of ITC. This technique requires fluorescent ligands, so it may be necessary to obtain a conjugate fluorophore, which is not guaranteed to interact equivalently with the receptor.

We would like to emphasize that we should be cautious when comparing computational and experimental results. In the first place, both experimental and computational measures should be converted to absolute binding free energies. Also, both measures may not have been taken under the same conditions, which is a possible source of error according to David Mobley<sup>158</sup>. Computer experiments take place in idealized conditions that may not hold in experiments. For diverse reasons, such as experimental costs, experimental values are often given in terms of IC<sub>50</sub>, which is converted to affinity values in a process involving the Cheng-Prusoff equation, and this may be more or less accurate depending on the conditions<sup>159</sup>. Altogether, although they are our gold standard, comparison with experimental values is not straightforward, and care must be taken.

### **Computational estimations**

We already saw that computational tools are a great asset since they can complement experiments. Moreover, they would have a greater impact if they could be used to make predictions<sup>160</sup>. For example, the goal of structure-based drug design is to use structural information to obtain high-affinity ligands that bind to a specific site in a given receptor, and obtaining accurate binding free energies is crucial for that. However, estimating binding free energies with computational methods is an open challenge. In the SAMPL5 blind contest<sup>161</sup>, the best prediction obtained a root-mean-square error of 2 kcal/mol in a host-guest system. Host-guest systems provide a simplified picture of the protein-ligand complexes and are easier to sample. For this reason, notwithstanding some very precise calculations have been reported<sup>162</sup>, the error is (in general) not expected to be lower in the real case prospective applications. As overviewed above, the ruggedness of the high-dimensional energy landscape is one cause behind this difficulty. As pointed out by Gilson and Zhou in a different and also outstanding review<sup>163</sup>, an additional reason is the small value that results from the subtraction of large numbers: the complexed and uncomplexed protein and ligand energies, which are involved in the binding free energy estimation.

Depending on the tradeoff between speed and accuracy, there is a broad spectrum of computational techniques that goes from fast procedures that are capable of screening a large number of compounds with a reduced precision, to those that are able to reach a greater accuracy at a more expensive cost.

In one side of the spectrum we find docking methods<sup>164</sup>, which are able to screen large libraries of compounds in drug design<sup>165</sup>. Some common docking programs are AutoDock Vina<sup>166</sup>, DOCK<sup>167</sup>, Glide<sup>94</sup> and rDock<sup>168</sup>. These aim to find the best binding pose and generally provide a reduced collection of promising poses ranked with a scoring function that represents the free energy<sup>169</sup>. The scoring function can be based on different principles, and may be for example force-field-based, empirical or knowledge-based (or a consensus between them); the accuracy and speed will motivate its choice. Aside from the accuracy, a significant drawback is the typically limited conception of flexibility for the sake of speed. Two common alternatives to overcome this limitation are ensemble docking<sup>170</sup>, where the ligand is docked to an ensemble of previously generated receptors rather than to a single static protein, or the use of normal modes to sample protein flexibility<sup>171,172</sup>. PELE has also been widely used as a docking program<sup>173</sup> and utilizes the latter approach, which has allowed gaining a distinctive recognition in the CSAR14 contest<sup>174</sup>. However, we should emphasize that considering protein and ligand flexibility imposes a major penalty in execution time, and thus PELE is not suitable for the virtual screening of large libraries.

A different compromise is that of end-point free energy methods, such as the Linear Interaction Energy (LIE) method<sup>175</sup>, MM/PBSA<sup>176</sup> or MM/GBSA<sup>177</sup>. Rather than using a reduced number of docked poses, these methods rely on sampling to provide free energy estimations. More specifically, they sample the end points, namely the bound complex and optionally the unbound protein and ligand. Focusing on the more popular MM/PBSA and MM/GBSA, their free energy estimation is given by:

$$G = E_{bond} + E_{vdW} + E_{elec} + G_{sol} - TS \quad (17)$$

where,  $E_{bond}$  accounts for the bonded interactions (bond, angle, torsions),  $E_{vdW}$  for van der Waals terms,  $E_{elec}$  for electrostatics,  $G_{sol}$  for solvation free energy, and  $-TS$  for the entropy contribution.

Different implementations are available, and a common choice involves the use of a single snapshot to obtain the complex, protein and ligand contributions, as it has been shown to perform better due to the lack of exhaustive sampling<sup>178</sup>. In this case,  $\Delta G$  can be written as:

$$\Delta G = \langle G_{PL} - G_P - G_L \rangle_{PL} \quad (18)$$

In PELE we perform single-point (unaveraged) evaluations of Eq. 18 neglecting entropic terms in order to score poses on the fly, and is often found in our publications under the name of “binding energy”. We use it to discriminate the best binder at an inexpensive cost but for more accurate results MSM is preferred (see below).

Entropy is often overlooked in Eq. 18, which explains the common large overestimations in the free energy<sup>179</sup>. The underlying reason is that it is a computationally expensive term, difficult to converge<sup>180</sup>, and it is not clear whether it improves results<sup>177</sup>. In publication 1 we use Eq. 18 including approximate entropic terms to study the binding of a substrate and inhibitor in a prolyl oligopeptidase (POP). Aside from translational, rotational and quantum mechanical vibrational terms<sup>181</sup>, we account for protein and ligand flexibility loss upon binding employing rotamers.

The other end of the spectrum encompasses rigorous free energy methods, also known as pathway methods. The latter may be physical, such as in steered MD<sup>182</sup>,

or non-physical, such as in alchemical transformations<sup>183</sup>. They require the sampling of the whole pathway and therefore have a more expensive associated cost, especially because molecular simulations tend to get trapped in long-lived metastable minima<sup>184,185</sup>. For this reason, methodologies that make use of physical pathways are usually found in combination with biased sampling techniques and these may require the definition of reaction coordinates. Some examples are umbrella sampling<sup>186–188</sup>, metadynamics<sup>189–191</sup>, steered MD<sup>182,192</sup> combined with Jarzynski’s non-equilibrium equality<sup>193</sup> or adaptive force bias<sup>194</sup>.

Alchemical transformations are often used to compute the relative binding energy between ligands (*i.e.*  $\Delta\Delta G$ ), and are based on the progressive transmutation of a ligand into another one using overlapping windows. To our knowledge, it is currently the most precise technique and a remarkable example is the FEP+ software of Schrödinger, which has shown typical deviations of less than 1 kcal/mol in 330 transformations in a retrospective study, and 1 kcal/mol in a prospective study<sup>162</sup>. Its main caveat is the sampling problem associated for example with large energy barriers separating bound conformations or large chemical transmutations between ligands<sup>195,196</sup>.

Finally, the sampling difficulties mentioned earlier motivated alternatives to thermodynamics-based sampling methods such as mining minima<sup>197</sup> of Gilson and colleagues or the conformational factorization<sup>198</sup> of Wales and coworkers, which are based on the evaluation of local minima contributions.

In this thesis, we present a methodology that represents a compromise between rigorous free energy calculations and end-point methods. Specifically, we used a combination of PELE and MSM to describe the energy landscape and estimate binding free energies. As we saw, PELE provides a considerable speedup compared to standard thermodynamics-based methods that reproduce protein and ligand flexibility (*e.g.* MD or rigorous MC). However, there is no free lunch, and the price that we have to pay is the unfulfillment of detailed balance, which may have an impact in our free energy estimations. In publication 6 we explain the technique that we used to assess the effects of minimizations. Also, we developed a protocol of adaptive simulations based on reinforcement learning<sup>199</sup> in order to overcome these sampling limitations and is presented in publication 5.

## Markov State Models

MSM<sup>9</sup> are mathematical models constructed from molecular simulation data that approximate long-timescale dynamics to Markov chains. MSMs require the process under study to be divided into long-lived states so that the Markovian assumption is valid, *i.e.* the transition probabilities are memoryless and solely depend on the current state, which is a reasonable assumption in protein-ligand association. This technique projects complex behaviors onto a simplified model and is, therefore, a suitable method to gain insight into the large amounts of data required to study biophysical processes. As opposed to other techniques, this projection is done *a posteriori* and does not influence the dynamics. Note that this projection emerges from the system, and does not need of human intervention<sup>200</sup>. Common applications of MSMs are the study of protein folding<sup>184,201–203</sup> or protein-ligand binding<sup>11,115,204–206</sup>. In this thesis, we use MSMs to characterize binding mechanisms and thermodynamic properties of protein-ligand binding, such as binding free energies, as it is one of the main objectives. Notwithstanding they have been mostly used with MD, we use MSMs to describe PELE explorations, bearing in mind the limitations of not sampling with a real time (see MC section for more details).

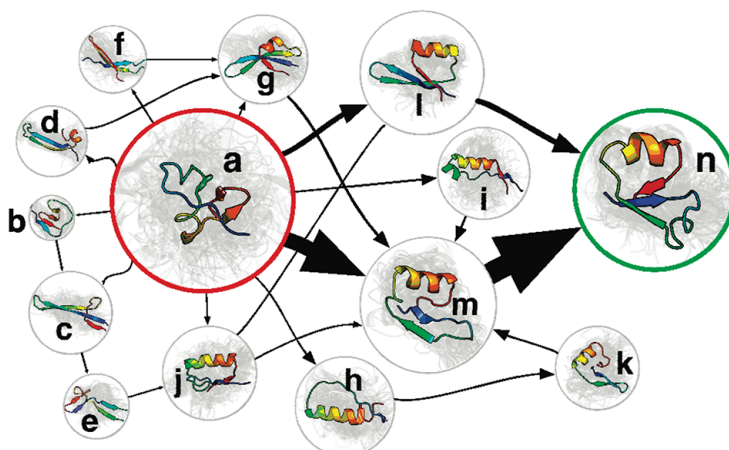


Figure 11. Folding pathway of NTL9(1-39). The arrow thickness is proportional to the flux, whereas the circle size to its free energy. Source: Ref. 10.

Molecular simulations are often involved in qualitative (“look and see”) studies, such as in the observation of rare events. Markov modeling is a suitable approach to quantify them and assess for example whether certain features have statistical significance or not. Importantly, the exploration does not need to proceed from a unique and long trajectory and individual simulations only need to characterize locally the Markov chain. Hence, events that take place in long timescales can be described using much shorter simulations<sup>207</sup>. A major consequence is that it provides a game-changing protocol to plan simulations, as can be seen in collaborative projects such as Folding@home<sup>8</sup> or GPUGRID<sup>12</sup>, or in the use of the cloud platform Google Exacycle<sup>208</sup>. Remarkably, this allowed reaching the microsecond milestone<sup>10</sup>. Also, it opened the door to adaptive sampling, where simulation starting points are adaptively chosen according to certain criteria, such as minimizing the model uncertainty<sup>209,210</sup> or according to the cluster residence time<sup>211</sup>, and provide substantial speedups compared to non-adaptive simulations.

Sampling is still a caveat for Markov models; despite only needing local equilibrium, obtaining sufficient statistics to estimate local transition probabilities may be a challenge in long-timescale processes, as it may be protein-ligand dissociation. For this reason, it is convenient to discretize the slowest process, for example with the aid of time-independent component analysis (tICA)<sup>212</sup>, or discretizing the binding pathway, as we propose in publication 4.

### Construction, validation, and analysis

In this subsection we outline the process of MSM building and validation (Fig. 12), for a more in-depth analysis, the reader may refer elsewhere<sup>9</sup>. There are currently two main software projects to ease the building, validation, and analysis of MSMs: MSMBuilder<sup>213</sup> and pyEMMA<sup>214</sup>. In publications 2, 3, and 4, we used EMMA 1.3<sup>215</sup>, while in publication 6 we upgraded to pyEMMA.

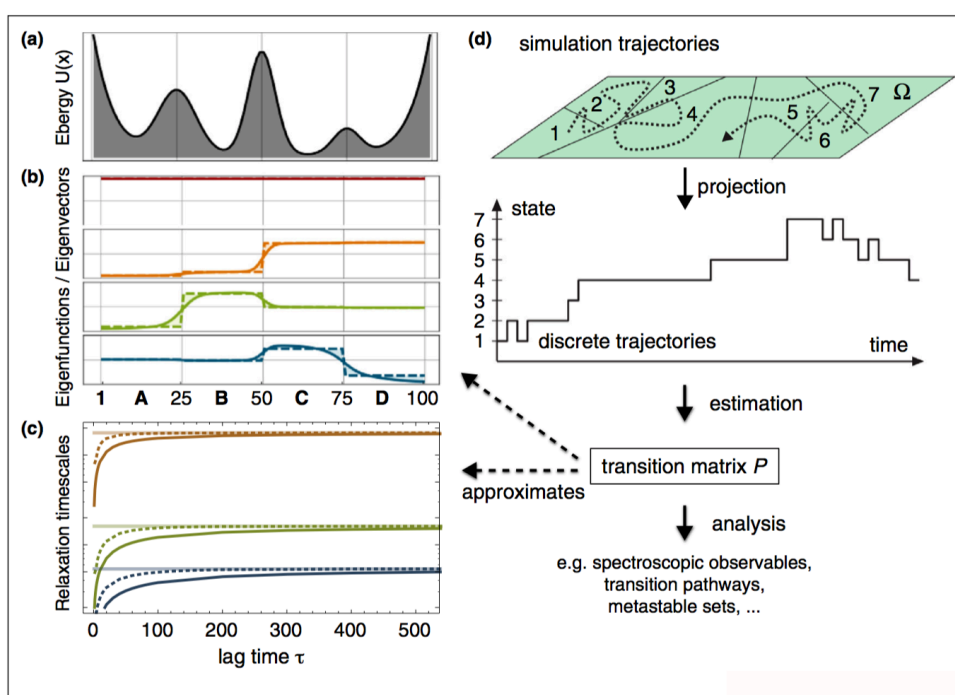


Figure 12. An illustrative example of MSMs. Panel (a): Four-well energy potential. Panel (b): The Markov model's eigenvectors approximate to the real dynamics' eigenfunctions. Panel (c): The MSM timescales converge to the true relaxation timescales, and the convergence rate depends for example on the discretization. Panel (d): Outline of the MSM construction: Trajectories are projected onto a discretized space, from which the transition matrix can be estimated. Its largest eigenvalues and eigenvectors approximate to the real dynamics. Adapted from Ref. 216.

The first step involves **running** the **simulations** until sufficient statistics are gathered and convergence is reached. With PELE, we typically use hundreds of processors during 24-48h. We assess convergence studying the variation of the free energy estimation<sup>115</sup> or with a metric based on the relative entropy (see SI in publication 4).

Then, the conformational space is **partitioned** into a set of non-overlapping states, defined according to a metric that must be defined *a priori*, and the dynamics are projected onto it. At this point, tICA may be used to reduce the dimensionality and provide the slowest collective coordinates automatically. In

our case, we employ the ligand center of mass or a ligand atom’s coordinates and use a geometrical clustering. In particular, we use the *k-means* clustering method and choose the number of clusters according to a convergence criterion that ensures a correct discretization (publication 6), typically in the range of hundreds. Finally, the space is partitioned into microstates using a Voronoi tessellation, and the real trajectory is projected onto it. The discretized trajectory describes the ligand center of mass evolution throughout the simulation in the reduced cluster space.

Generally, the probability transition between clusters does not satisfy the Markov assumption, and, as shown by Prinz and colleagues<sup>200</sup>, the discretization error can be reduced with either a better partition of the slow process or a sufficiently long lag time (Fig. 13). Tentative transition matrices are estimated for different lag times, and we **choose a lag time** for which the implied timescales have converged. The lag time determines the sequential discretization (temporal/steps in the case of MD/MC). An MSM that is estimated at a certain lag time is not suitable to study processes that decay faster than that, and thus the lag time must be much smaller than the implied timescales of interest. The lag time may be reduced improving the discretization to reduce intra-cluster barriers<sup>217</sup> (Fig. 13).

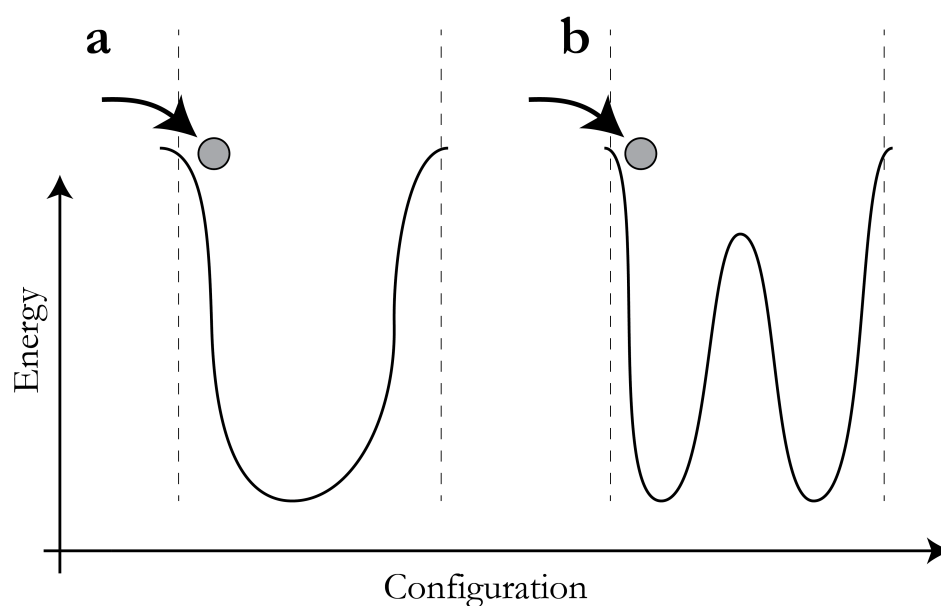


Figure 13. Illustration of the lag time. An energy potential surface (solid line) and the state boundaries (dashed line) are sketched. Panel **(a)**: In a state composed of a single energy well, transition probabilities will depend on history in short timescales, for example, it will be more likely to jump to the left state rather than to the right state when coming from the left one. After a certain time, the system relaxes in the energy well and satisfies the Markov assumption (becomes memoryless). Panel **(b)**: The lag time will increase with internal energy barriers, as the system has to surmount them in order to become memoryless.

At this point, the transition matrix is **estimated** from the counting matrix, which counts for all transitions between any pair of states at a given lag time. Before proceeding with the analysis, the model is **validated** with the Chapman-Kolmogorov test as described in Ref. 200, which assesses the consistency of the MSM with the original data. More specifically, it tests whether the estimated transition probabilities between groups of microstates, typically metastable states, agree with the observed values at different times within the statistical uncertainty.

If the model does not pass the test, *i.e.* it is not Markovian, it is often due to a poor discretization or a too short lag time.

Different **analysis** can be conducted on the model. A fundamental value in this thesis is the stationary distribution,  $\pi$ , obtained with the normalized first left eigenvector of the transition matrix. In our model, this value corresponds to the probability of finding the ligand's center of mass (or, alternatively, a ligand's atom coordinates) in a given microstate, averaging out the rest of coordinates in the phase space (that is, all possible protein and ligand, and eventual ion or water conformations). The potential of mean force<sup>186</sup> (PMF),  $\mathcal{W}$ , which we use to compute free energy differences, is defined in our reaction coordinate,  $\mathbf{r}$ , as:

$$\mathcal{W}(\mathbf{r}) = -k_b T \ln(p(\mathbf{r})) \quad (19)$$

where  $k_b$  is the Boltzmann constant,  $T$  is the temperature and  $p(\mathbf{r})$  is the average probability distribution along  $\mathbf{r}$ , the ligand's center of mass coordinates. As discussed earlier, the computational free energy estimation needs to be converted to the standard binding free energy in order to be compared with experiments, and we followed the procedure described in Refs. 11 and 115, which is an extension of the one-dimensional PMF described in Ref. 188.

The number of microstates is often too large and makes the binding mechanism analysis difficult. For this reason, kinetically similar microstates are lumped together into macrostates (Fig. 11). The number of macrostates is a user-defined value that depends on the characteristics of the system under study and the desired level of detail<sup>217</sup>. In this thesis, we used the Perron cluster cluster analysis<sup>218</sup> (PCCA) implemented in EMMA, which uses the right eigenvectors to lump together microstates, assuming that kinetically close microstates have similar eigenvector values. Some analyses that we performed were computing transition probabilities between macrostates, using transition path theory<sup>207</sup> to study fluxes the bulk and the binding pose (publication 4), and computing macrostate probabilities (publication 3).





## Objectives

The previous section exposed some advanced computational techniques to study protein-ligand binding. These methodologies reach the compromise of accurately describing atomistic interactions while still being able to reproduce protein plasticity. However, as we showed, they often face restrictions in their application to drug design. The main objective of this thesis has been the development of computational methodologies to overcome some of these limitations and the subsequent application to real-case situations. This goal has been developed in collaboration with a pharmaceutical company and experimental research laboratories.

The specific objectives are:

### 1. Development of PELE

Competitive state-of-the-art computational tools must be able to take advantage of cutting-edge algorithms. The first version of PELE lacked of tests, and the different software components were not entirely independent, which hampered the reliability and maintainability. For this reason, the first objective was rewriting PELE and obtaining a competitive program. The main purpose of the recoding was extending its support to include new algorithms, a graphical interface, and machine-learning libraries or to take advantage of the multi-core paradigm, just to name a few. This is an indispensable objective to accomplish objectives 2 and 3, and led to Publications 2, 3, 4, 5, 7 and 8.

### 2. Establishing a protocol to study protein-ligand binding

The study of protein-ligand binding is one of the biggest open challenges in biomolecular modeling and has significant implications in drug design. Sampling a representative ensemble of the energy landscape is essential for reliable binding affinity estimations, and the major drawback of computational tools is the inherent complexity of traversing the energy landscape efficiently. Also, accurately describing binding mechanisms with unbiased and atomistic simulations remains an open challenge for receptors with occluded binding sites due to the long associated timescales.

The objective is to establish a protocol using PELE in combination with MSM to build a potential of mean force describing protein-ligand interactions and study binding mechanisms. Results are shown in publications 7, 1, 2 and 3.

### **3. Development of a procedure to overcome the sampling limitations associated to metastability**

As described in the introduction, the use of PELE enhances the conformational exploration compared to other methods such as MD, but still suffers from metastability-related problems. This objective involves improving the sampling, in order to provide a more efficient management of computational resources. In publication 5, we address this issue with a procedure based on an adaptive reinforcement learning protocol combined with PELE, and in publication 6 we apply it to speed up free energy estimations.

## Results

In this section we include the results derived from this thesis. The supporting information is found in the appendices.

### **Publication 1 - Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques**

**Authors:** Martin Kotev, [Daniel Lecina](#), Teresa Tarragó, Ernest Giralt, Victor Guallar

**Journal:** Biophysical Journal, 108, 116–125 (2015)

#### **Summary:**

In this publication we study the ligand migration of the inhibitor Z-pro-prolinal in prolyl oligopeptidase (POP). POP is a large protease that presents a deeply buried active site; for this reason it is a challenging system for standard sampling techniques and PELE was used instead. Upon the observation of multiple binding events, we found that ligand entrance is produced through the pore in the  $\beta$ -propeller domain. Furthermore, we modeled the binding of an undecapeptide substrate and the release of a dipeptide product by means of a biased protocol. The dissociation occurs through a flexible 18-amino acid residues loop, a different path to the one followed by the substrate.

#### **Author contribution:**

My tasks involved helping in the overall ligand migration simulations, the entropy change estimation upon binding and writing the corresponding section of the manuscript.



## Article

# Unveiling Prolyl Oligopeptidase Ligand Migration by Comprehensive Computational Techniques

Martin Kotev,<sup>1</sup> Daniel Lecina,<sup>1</sup> Teresa Tarragó,<sup>2</sup> Ernest Giralt,<sup>2,3,\*</sup> and Víctor Guallar<sup>1,4,\*</sup>

<sup>1</sup>Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain; <sup>2</sup>Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain; <sup>3</sup>Department of Organic Chemistry, University of Barcelona (UB), Barcelona, Spain; and <sup>4</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

**ABSTRACT** Prolyl oligopeptidase (POP) is a large 80 kDa protease, which cleaves oligopeptides at the C-terminal side of proline residues and constitutes an important pharmaceutical target. Despite the existence of several crystallographic structures, there is an open debate about migration (entrance and exit) pathways for ligands, and their coupling with protein dynamics. Recent studies have shown the capabilities of molecular dynamics and classical force fields in describing spontaneous binding events and nonbiased ligand migration pathways. Due to POP's size and to the buried nature of its active site, an exhaustive sampling by means of conventional long enough molecular dynamics trajectories is still a nearly impossible task. Such a level of sampling, however, is possible with the breakthrough protein energy landscape exploration technique. Here, we present an exhaustive sampling of POP with a known inhibitor, Z-pro-prolinal. In >3000 trajectories Z-pro-prolinal explores all the accessible surface area, showing multiple entrance events into the large internal cavity through the pore in the  $\beta$ -propeller domain. Moreover, we modeled a natural substrate binding and product release by predicting the entrance of an undecapeptide substrate, followed by manual active site cleavage and nonbiased exit of one of the products (a dipeptide). The product exit shows preference from a flexible 18-amino acid residues loop, pointing to an overall mechanism where entrance and exit occur in different sites.

## INTRODUCTION

Prolyl oligopeptidase (POP; EC 3.4.21.26) (also known as prolyl endopeptidase, PREP, or postproline cleaving enzyme) is a serine protease that cleaves postproline bonds in short peptides (1). POP inhibitors might be valuable compounds in a variety of clinical conditions of the brain, such as the cognitive disturbances present in schizophrenia and bipolar affective disorder, as indicated by their neuroprotective and cognition-enhancing effects in experiments with animals (2). For these reasons, a plethora of POP inhibitors have been developed during the last 10 years for treatment of several central nervous system disorders (3,4). Two basic groups of inhibitors have been proposed: forming a covalent bond with the catalytic serine and noncovalent ones. Both of them dock at the same specific proline pocket, the main difference being the presence or lack of chemical groups capable of covalently binding to Ser-554. The development of POP inhibitors, however, has been based almost exclusively on modification of the canonical peptidomimetic compound Z-prolyl-prolinal (ZPP) that fits into the POP active site. This strategy does not take into account other possible POP binding surfaces such as surfaces involved in the entry of substrates and/or exit of products, which may trigger the discovery of innovative peptide scaffolds

with biological activity. In addition to its enzymatic role, POP interacts with several proteins,  $\alpha$ -synuclein being one of the most relevant. POP accelerates aggregation of  $\alpha$ -synuclein in vitro, a process that can be reversed by specific inhibitors (5,6). Moreover, nuclear magnetic resonance spectroscopy studies have revealed that POP is a highly dynamic protein and that active site inhibition shifts this conformational equilibrium toward a less dynamic form (7). POP structural fluctuations and its importance for substrate/inhibitor delivery, however, is centering particular attention (8,9).

The crystal structures of POP indicate two domains, a catalytic one bearing the Ser-His-Asp triad and the so-called  $\beta$ -propeller domain, which covers a huge cavity around the catalytic center (1). Ligand access to this catalytic center, however, is under debate. Two main entry/exit areas have been investigated since release of the first crystal structures of porcine POP 16 years ago. The first one is a pore in the  $\beta$ -propeller domain, whereas the other is a ~18-residue flexible loop (some authors call it loop A (10)) standing close to the active site. The diameter of the pore (distance between two approximately opposite  $\alpha$ -carbons) is around 11–13 Å (Fig. 1). This means that appropriate conformational orientations of some side chains in the area could open a passage for some inhibitors or small peptides. Two lysine side chains and two glutamic acid ones form salt bridges, which, together with a few hydrogen bonds reduce the propeller

Submitted June 23, 2014, and accepted for publication November 17, 2014.

\*Correspondence: victor.guallar@bsc.es or ernest.giralt@irbbarcelona.org

Editor: Bert de Groot.

© 2015 by the Biophysical Society  
0006-3495/15/01/0116/10 \$2.00

<http://dx.doi.org/10.1016/j.bpj.2014.11.3453>



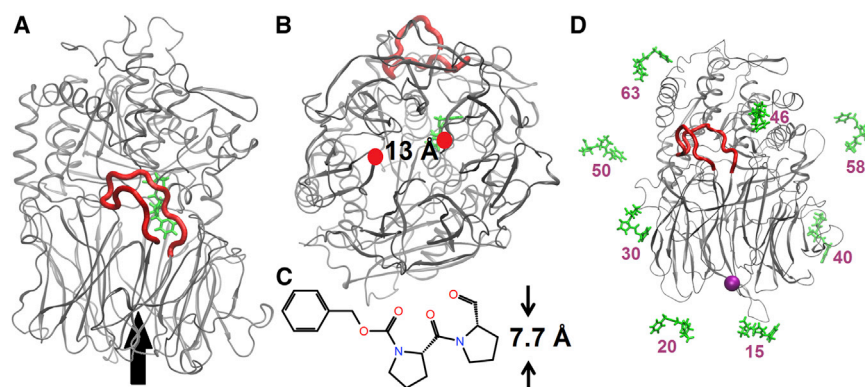


FIGURE 1 (A) Side view of the flexible loop (underlined in red) with a bound ZPP inhibitor (green). The arrow points to the  $\beta$ -propeller pore. (B) Bottom view showing also the pore distance between two  $\alpha$ -carbons (red circles). (C) Structure and maximum width of ZPP. (D) All eight starting positions of ZPP (in green) for the entrance simulations and their approximate distance (purple) in Å to the center of the pore (purple bead). To see this figure in color, go online.

pore (Protein Data Base (PDB) entry 1QFS). Some experimental data shows forming of a disulfide bridge to block this pore but in this case the bridge is created aside and does not cover the central part of the pore (11). The mobility of the flexible loop, on the other side, has been suggested by trypsin cleavage assays (10). Recent experimental studies, however, question its involvement in ligand delivery. In the work of Szeltner and co-workers (10) a heptadecapeptide is better cleaved from a mutated POP containing a loop covalently locked by a disulfide bridge to the catalytic domain.

Of importance, there is an additional crystal structure from the bacterium *Aeromonas punctata* (12) where the two domains present a large opening, pointing to a clear entrance into the active site. In fact the domains are almost separated and only held by two covalent bonds—the hinge between the domains. However, there is no mammalian crystal structure showing such conformation, neither is there clear experimental proof of this opening (10). Furthermore, a porcine POP crystal structure, 99% similar to the human one, shows clear differences to the bacterial one in the non-covalent forces keeping together the two domains (12). Other studies also suggest that local conformational changes related to some flexible loops but not the entire domains could be responsible for the access to the active site of POP (13).

Computational studies have also addressed POP's dynamics and its possible ligand migration pathways. Molecular dynamics (MD) simulations showed significant loop opening and exposure to the bulk solvent (13). Some authors in their previous study have used steered MD and umbrella sampling simulations to force the inhibitor ZPP exiting from the active site (8). In this work, the inhibitor was pulled in two possible directions: the loop one and to the  $\beta$ -propeller pore. Results show that the exit of ZPP is energetically more favorable through the loop region (8). Docking results and subsequent MD simulations from a docked pose of an inhibitor in the  $\beta$ -propeller pore have shown that the ligand can reduce some distance traveling toward the active site, which depicts potentiality of  $\beta$ -propeller in ingesting a ligand (9). None of the published simulations have shown indications

for the interdomain opening. All of them reveal stable closed POP structures during the simulations except for some loop motions (13).

Using special purpose machines or graphical processors units, a nonbiased search accessing microsecond timescale simulations has recently been performed on small or medium systems (14,15). These computational approaches represent a significant computational cost, being still prohibitive when dealing with complex systems (buried active sites) such as POP. To address this issue we have used protein energy landscape exploration (PELE), a novel computational technique capable of exploring the nonbiased ligand diffusion and proteins dynamics (16). PELE combines a Monte Carlo stochastic approach with protein structure prediction algorithms, and it is capable of accurately reproducing long-timescale processes in a 1–2 order of magnitude faster manner than MD (17–20). Such a technological development, together with the use of the supercomputer Mare Nostrum, has allowed us to run 3000 trajectories, for an extensive exploration of ZPP interaction with both mammalian and bacterial POP. Our results indicate that entrance happens mainly through the bottom pore, with only smaller molecules being able to enter through the bacterial opened loop. Furthermore, we simulated the catalytic process of entering an 11-amino acid residue peptide as a substrate and the exiting of one of the two products. This full catalytic event indicates entrance through the  $\beta$ -propeller pore and exit of the cleaved small peptide through the loop area.

## MATERIALS AND METHODS

### System preparation

Initial coordinates for the closed POP structures were taken from the PDB entries 1QFS (mammalian porcine) (1) and 3IVM (bacterial *A. punctata*) (12). Semiopen bacterial POP coordinates were obtained from the PDB structure 3IUQ (12). Hydrogen atoms and titratable side chains were optimized with the Protein Preparation Wizard tool from Schrödinger (21) at physiological pH. The covalent bond with the ligand was broken (with the corresponding hydrogen additions) to assure the free exploration. The second ligand in PDB 3IVM was removed. Two missing flexible fragments (residues 194–201 and 654–660) of the 3IUQ PDB entry were recovered and filled with the Prime software (21).

## PELE

The PELE algorithm is based on a consecutive iteration of three main steps: a ligand and protein (backbone) perturbation, a side-chain sampling, and a minimization (22,23). Thus, the procedure begins by a ligand perturbation involving a random translation and rotation of the ligand. In the case of the protein, the perturbation is based on the  $\alpha$ -carbons anisotropic network model (ANM) (24); all atoms are displaced by a minimization where the  $\alpha$ -carbons are forced to follow a randomly picked low eigenvector (within the lowest six modes) obtained in the ANM approach. In particular, three consecutive perturbations of 1.5 Å in the same mode (and direction) were used before randomly picking a new mode. The ANM network model used identical springs connecting all  $\alpha$ -carbons within a 15 Å cutoff (additional details on the ANM setup can be found in (20)). The algorithm defines the most excited side chains with the largest changes in energy after the ANM move and these are included in the next step, the side-chain prediction. Here, PELE proceeds by optimizing all side chains local to the ligand in a defined distance (6 Å) together with the hot side chains determined in the ANM step (22,23). The last procedure involves the minimization of the entire system, keeping the  $\alpha$ -carbon with a weak constraint after the ANM move. These steps compose a move that is accepted (a new local minimum) or rejected based on a Metropolis criterion, forming a stochastic trajectory. PELE runs were carried out at a temperature of 1000° K. As emphasized in our original work (16,20), this high Metropolis temperature does not correspond to a real thermal bath, the effective temperature being significantly lower. PELE uses an OPLS (optimized potentials for liquid simulations) all-atom force field (OPLS-AA) (25) with an implicit surface generalized Born (SGB) continuum solvent model (26).

PELE's combination of random perturbations and protein structure prediction algorithms results in an effective exploration of the protein energy landscape, capable of reproducing large conformational changes associated with ligand migration (16–20). The method provides MD quality results (20) at a significantly faster rate. When compared to docking techniques, it provides a good induced fit description, allowing the docking in difficult cases (apo, cross-docking, etc.) (19). Moreover, when combined with Markov state models, PELE provides absolute binding free energies in a similar fashion to extensive (and more expensive) MD techniques (17).

## PELE entrance/exit protocols

For ZPP, rotations and translations alternate between two different values: small ones using 30° rotation and 0.75 Å translation, and big ones with 60° and 1.50 Å, which were independently and randomly switched (with 50% overall probability). Two different ANM options were used for sampling the protein backbone. First type includes a random switch among the first (lowest) six calculated modes. The second type, aiming to bias the protein opening, used a dominant ANM mode describing the movement of opening and closing of the two domains.

For the entrance simulations, the ligand was placed at eight different random positions in the protein surface (see Fig. 1 D for the exact initial ligand positions). When studying the exit pathways, the ligand was chosen always to start from the active site in an equivalent position to the crystallographic structures.

## Entropy corrections

Entropy loss estimates for the bound complex (respect to the solution value) were divided in the following contributions: translational, rotational, conformational, and vibrational. Translational, rotational, and vibrational entropies were obtained using the standard ideal gas approximation (for example as described in the Gaussian thermochemistry site, [http://www.gaussian.com/g\\_whitepap/thermo.htm](http://www.gaussian.com/g_whitepap/thermo.htm)). Conformational entropy was obtained by screening all available dihedral conformations for the ligand and the neighboring protein side chains (in direct contact,

$<2$  Å):  $\Delta S_{\text{conf}} = k_B \ln (\Omega_C / \Omega_P \Omega_L)$ , where  $\Omega_C$  is the available dihedrals for the complex,  $\Omega_L$  for the isolated ligand and  $\Omega_P$  for the isolated protein.

## Loop prediction

Loop prediction calculations were executed with the Prime package from Schrödinger software (21). The protocol includes a default sampling algorithm and ultra extended loop refinement method, specifically designed to overcome sampling problems with long loops (>10 residues). Side-chain refinement was limited to residues with a side-chain heavy atom within 7.5 Å of any  $\beta$ -carbon from the loop. Energy cutoff for the final minimization refinement was varied to 20 kcal/mol (default is 10 kcal/mol). The loop prediction included residues 190–208 for the porcine POP and residues 190–205 for the bacterial one.

## MD

MD simulations were performed using the Desmond MD program (27,28). The bound ZPP structure (PDB entry 1QFS) was solvated in an orthorhombic box of 19 012 water molecules, and 66 sodium and 49 chloride ions were added to neutralize and create a 0.14 M solution of NaCl. We used the OPLS-AA force field and the simple point charge water model. The default relaxation protocol in Desmond was used, followed by a 70 ns production run in the NPT ensemble using the Nose-Hoover thermostat and the Martyna-Tobias-Klein barostat (29,30). The smooth particle mesh Ewald method was used for the long-range interactions.

## RESULTS

### Loop prediction calculations and MD simulations

Loop prediction calculations obtained with the Prime software indicate preference for the closed state. For both porcine and bacterial, the first loop pose corresponds to a structure in close agreement with the crystallographic closed one with 0.5 Å and 0.9 Å  $\alpha$ -carbon root mean-square deviation (RMSD), respectively. Higher energy poses, however, introduce some degree of opening. In porcine, we find the most open structure as the fifth pose, with an energy increase of 9.1 kcal/mol, shown in Fig. 2 A in black. Easier opening is observed in the bacterial POP simulations. In this case, the second result by score, shown in Fig. 2 A in red, represents the most opened loop geometry, with an energy increase of only 2.3 kcal/mol. In both porcine and bacterial POP the loop is involved in interactions with another small flexible part of  $\beta$ -propeller domain. This is the loop constructed by residues 215–222 for the porcine and the same one for the bacterial analog (residues 212–219). The two most open structures were chosen as our initial models for the open state simulations in PELE (called porcine open and bacterial open, see Table 1).

The analysis of the MD simulations for the porcine structure also indicates some degree of loop opening. The loop starts the opening at ~10 ns of the simulations and after passing through a semiopen conformation tends to partially close again. Together with loop motion toward opening, ZPP starts moving in a direction showing partial exiting through the loop with the phenyl ring as a leading residue.



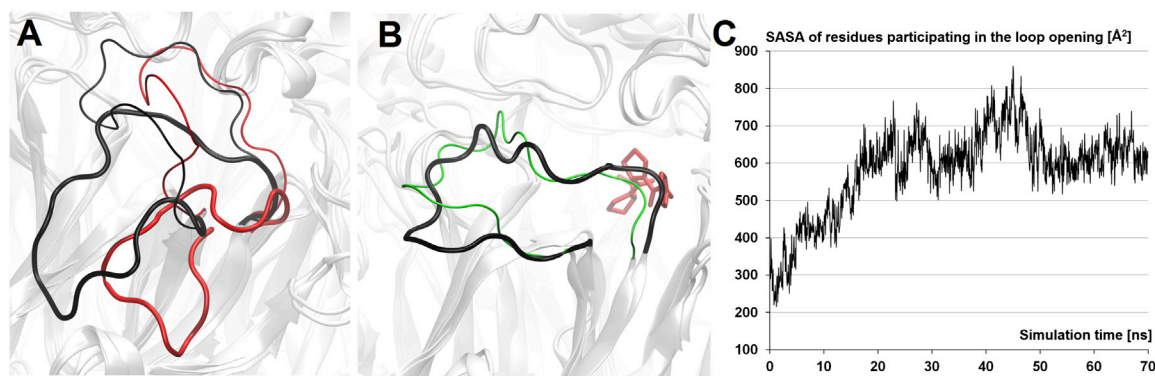


FIGURE 2 (A) Comparison of loop shapes from porcine (*thin black tube*) and bacterial (*thin red tube*) POP crystal structures, with PDB entries 1QFS and 3IVM, against most open structures from loop prediction calculations in the same colors but thicker tubes. (B) The 45 ns MD snapshot (*green*) is compared to the loop predicted structure (*black*). ZPP's position in the active site is underlined in *red licorice*. (C) Time evolution of SASA for residues 200–207, 590–594, and 641–644. Data were updated every 50 ps. To see this figure in color, go online.

The maximum open loop snapshot along the MD simulation, occurring at the 45 ns, is shown in Fig. 2 B in green. This opening is also clear when inspecting the evolution of the solvent accessible surface area for the main residues involved in the loop structure, as shown in Fig. 2 C. Interestingly, many of the structures from Prime's loop predictions have RMSD differences from MD snapshots lower than 1 Å (Fig. 2 B).

### PELE explorations

The summary of PELE simulations exploring the entrance and exit pathways for the ZPP ligand in both porcine and bacterial POP is shown in Table 1. As a reminder, to model the open state, we used the most opened loop structures shown in Fig. 2 A (obtained with loop prediction techniques).

#### ZPP entrance pathway

As seen in Table 1, out of the 400 simulations for each system we obtain approximately the same number of entrances by the  $\beta$ -propeller pore in all of them (referred to as bottom pathway in Table 1). As indicated in the Materials and Methods, ZPP initial positions were randomly placed in the protein surface (Fig. 1 D). The remaining nonentering trajectories present structures where the ligand is associated

with the surface (with some minor excursions into the bulk solvent). Furthermore, within each simulation the ligand explores a large fraction of the protein surface (see, for example, Fig. 3 or Movie S1 in the Supporting Material). In Fig. 3 we show a cross-section image where we display the ZPP ligand with blue beads (protein not shown) for the 50,000 snapshots along the porcine closed simulation. Clearly, we observe how the ligand covers the protein surface getting inside through the bottom entrance. In  $\sim 60\%$  of these entrance events, ZPP enters the  $\beta$ -propeller pore by the hydrophobic phenyl moiety (Fig. 4 A). From the protein site, the most displaced blade of the  $\beta$ -propeller, after overlapping with the crystal structure, is the one bearing His-180 (Fig. 4 B). Two other blades, ones bearing Glu-134 and Lys-82, also show significant displacement. In the remainder 40% entrance events ZPP enters by the proline moiety, showing similar displacement of the blades. These three blade changes (*marked with stars* on Fig. 4 B), however, do not enlarge the pore significantly, and they seem to be induced by internal protein adjustments rather than by interaction with ZPP; analogous changes are seen with and without inhibitor.

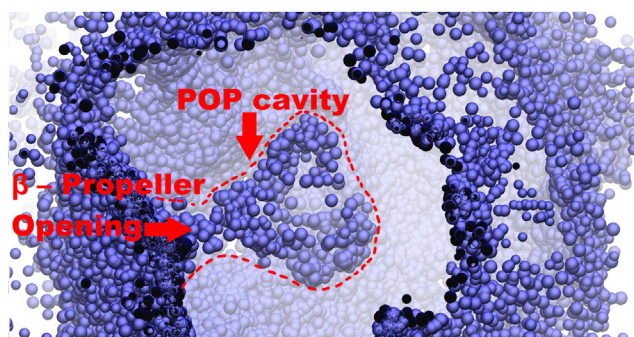


FIGURE 3 Cross section of POP from the side depicting the large cavity (*enclosed in red dashed curve*) and the ZPP surface exploration and bottom entrance. ZPP is presented with blue beads and the protein is omitted. To see this figure in color, go online.

TABLE 1 Entrance and exit pathways simulations

| PELE experiment   | ENTRANCE simulations                            | EXIT simulations          |
|-------------------|---|---------------------------|
| Protein structure | Pathway/successful/total number of trajectories |                           |
| Porcine closed    | bottom <sup>a</sup> /12/400                     | bottom/5/400              |
| Porcine open      | bottom/13/400                                   | bottom/5/400              |
| Bacterial closed  | bottom/13/400                                   | bottom/1/400              |
| Bacterial open    | bottom/14/400 <sup>b</sup>                      | loop/10/400 <sup>c</sup>  |
|                   | loop/7/400 <sup>b</sup>                         | bottom/1/400 <sup>c</sup> |

<sup>a</sup>Bottom refers to the  $\beta$ -propeller pore.

<sup>b,c</sup>Same set of trajectories.

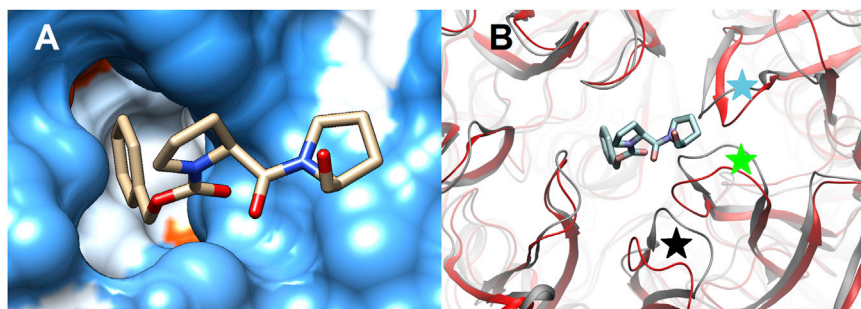


FIGURE 4 (A) Snapshot of entering ZPP through the  $\beta$ -propeller pore of porcine POP (surface presentation). (B) Bottom view of the entrance, gray cartoon, with an overlapped crystal structure (PDB entry 1QFS) shown in red cartoon. Black, green, and blue stars indicate bigger changes in the blades bearing His-180, Glu-134, and Lys-82. To see this figure in color, go online.

Our simulations indicate that the number of bottom entrances in both species is independent of the nature of the loop. Moreover, the mammalian POP does not show any ligand entrance by the loop even when starting by the open state. The bacterial one, however, shows seven entrances by the loop pathway when starting the simulation with the open state, the only instance where we observe entrance through the loop pathway.

#### ZPP exit pathway

Statistics on exit simulations, where the ligand starts in its active site position, show significantly different results from the entrance ones. In all cases the bottom exit was less probable than the entrance. Furthermore, contrary to the entrance, the exit through the bottom shows different results between porcine and bacterial, five events for porcine and only one for bacterial. Nevertheless, the exit through the bottom is still independent of the loop state.

Fig. 5 shows an entrance (*green*) and an exit (*red*) bottom trajectory for the closed porcine state. Fig. 5 A displays the ligand RMSD (*to the bound x-ray crystal*) along the PELE trajectory and the protein-ligand interaction energy. The entrance trajectory has initial high RMSD and interaction energies, decreasing accordingly along the entrance pathway. As expected, the opposite behavior is seen for the exit trajectory: an increase in RMSD and interaction energy. We should notice here that ZPP is a covalent inhibitor and that RMSD values are obtained in comparison to the bound crystal. The best binding ligand poses adopts an analogous crystal orientation but missing the last  $\sim 1\text{--}2$  Å translation of the covalent bound formation, giving an overall RMSD  $\sim 5$  Å. The ligand exits the bottom at approximately the 650 step, where we see an important barrier ( $\sim 6$  kcal/mol) in interaction energy. We want to note once more that the entrance trajectory starts significantly apart from the bottom pore.

Entropic contributions for ZPP at the bound state indicate a 31.2 kcal/mol correction to the binding free energy ( $-\Delta S$  term), obtained from 11.1, 15.0, 1.9, and 4.2 contributions from the translational, rotational, vibrational, and configurational entropy terms, respectively. This number, together with the PELE interaction energy, indicates an overall favorable binding event for ZPP in the noncovalent initial stage of

the binding process. Moreover, we want to point to the nice correlation between the interaction energy and the RMSD in the last approach to the active site (*lower right green and left red corners in the bottom panel of Fig. 5 A*), indicating biological relevance for these pathways.

The most interesting aspect of the exit simulations, however, is the presence of 10 exit events through the loop pathway in the bacterial open state, see Fig. 6 A. Fig. 6 B shows a representative orientation of ZPP when crossing the loop, with the phenyl leading the pathway. In most of the cases exiting was observed with the participation of five residues, Trp-579, Phe-174, Tyr-233, Arg-232, and Tyr-190. Interestingly, this orientation adopted by the ligand is similar to the one observed in our MD simulation of porcine POP as a response of loop semiopening (Fig. 6 B).

#### Porcine POP interdomain opening

PELE can simulate protein motion according to the displacement of  $\alpha$ -carbon-based ANM (24), an elastic model capable of describing large conformational changes. Inspection of the lowest six ANM modes showed that either the first or the second mode (depending on the initial structure) is associated with the interdomain opening direction. Thus, we forced PELE to sample this opening mode as the main ANM mode in porcine POP. Although open structures (similar to the open bacterial crystal) were produced when following (forcing) this mode, all of our attempts, including extreme temperatures, were unsuccessful in stabilizing them; the open structures spontaneously revert back to the closed one when not forcing the opening sampling mode.

#### Undcapeptide (substrate) entering and dipeptide (product) exiting simulations

Taking into account the results on ZPP, we modeled the entrance of a 11-residue peptide through the  $\beta$ -propeller pore in porcine POP. We also wanted to simulate the peptide cleavage in two products and the exit of the smaller (and more mobile) one in the presence of the other product in POP's cavity. For this purpose the Phe-Gly-Cys-Gly-Ala-Ser-Ala-Gly-Pro-Ala-Gly peptide, with two residues after the Pro cleavage point, was built. To facilitate the experiment the smaller product peptide (exiting part of

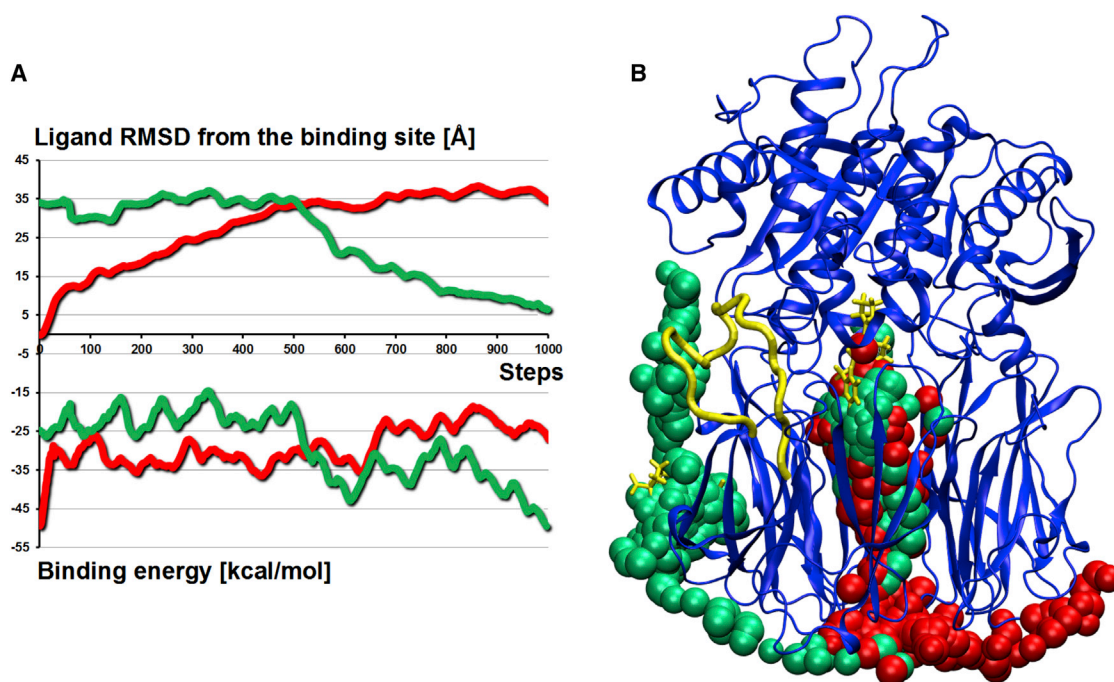


FIGURE 5 (A) Ligand interaction energy profile for a representative ZPP's entrance (*green*) and exit (*red*) trajectories. (B) Superposition of same trajectories (with same colors scheme) on one POP structure. ZPP positions along the trajectories are shown with beads, the flexible loop in *yellow tube*, and the initial inhibitor position in *yellow licorice*. To see this figure in color, go online.

the simulations) was chosen to be a dipeptide (Ala-Gly). The undecapeptide was placed around the bottom pore and guided to the  $\alpha$ -carbon of the catalytic Ser-554 using the spawning algorithm in PELE. This algorithm aims to reduce the distance between two atoms (the  $\alpha$ -carbons of Ser-554 and the substrate Pro) by random perturbation of the ligand and by using a tolerance distance window, 3 Å in our simulation. Every time the trajectory has a distance value larger (by the tolerance value) than the best registered distance, it will abandon the search and start with the best coordinates. Obviously, the best registered distance is updated when a shorter distance is found. Applying such protocol, the substrate cannot move further away, and explores freely, in a reduced window, possible structures that will reduce the desired distance. In this way, we can

model difficult cases like the entrance of a large substrate by the bottom pore.

Fig. 7 shows the interaction energy profile and the Ser-554-Pro distance along the guided entrance process. As mentioned previously, we should keep in mind that this value reflects only internal energies which, due to the peptide size (forming numerous hydrogen bonds), are significantly larger. Rotational, vibrational, and translational entropic corrections amount for  $\sim 34$  kcal/mol. Conformational entropy is out of our reach due to the presence of 30 rotatable bonds. Nevertheless, it was recently estimated to be on the order of  $\sim 60$  kcal/mol for a nine-residue flexible peptide, giving rise to a total corrections on the order of  $\sim 94$  kcal/mol (31). As seen in Fig. 7, no significant energy barrier is observed along the initial entrance process, in

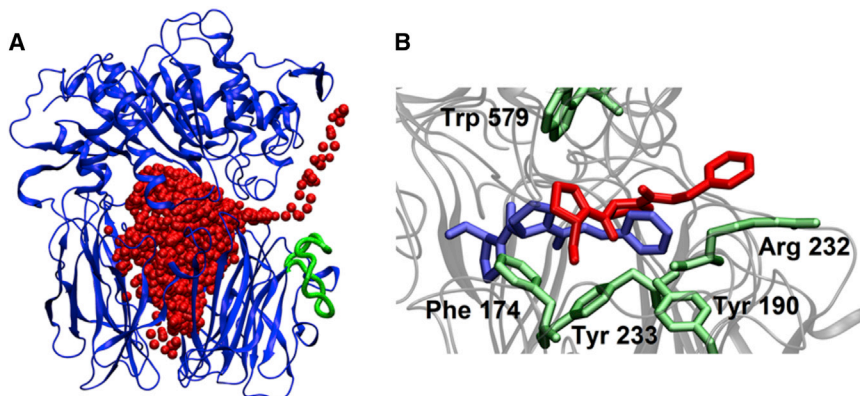


FIGURE 6 (A) A representative ZPP exit simulation for the open-loop (shown in *green*) bacterial POP. ZPP is shown in *red beads*. (B) A detailed view of the exit through the open loop and participating residues (in *green*). ZPP in blue corresponds to a superimposed structure from a MD snapshot in porcine POP. To see this figure in color, go online.



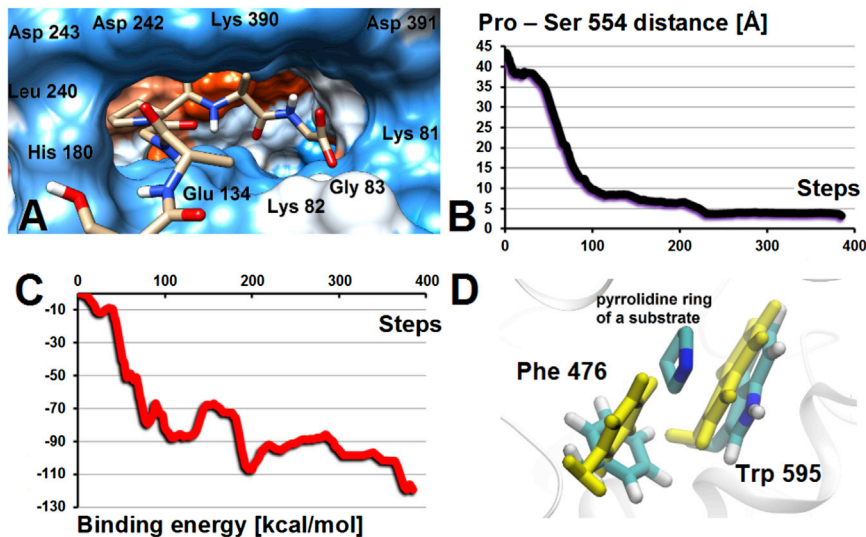


FIGURE 7 (A) Entering snapshot of the undecapeptide (C-terminus shown) and the residues forming the entrance. (B) Average distance (*angstroms*) between the carbonyl carbon of the undecapeptide proline and the oxygen atom of POP's Ser-554 for all entering trajectories. (C) Average binding energy profile for the entering trajectories (in kcal/mol). (D) Details of Phe-476 rearrangement to accommodate the pyrrolidine proline ring. To see this figure in color, go online.

agreement with a smooth reduction in the guiding distance. Thus, it seems like the big internal cavity can accommodate and easily allow the passage of large peptides. Around step 150 (Fig. 7) and after reaching a low Ser-554-Pro distance, we observe a significant side-chain rearrangement, giving rise to a better fitting (lower interaction energies) in the active site pocket. In particular, it involves mainly residues Phe-476 and Trp-595, two main actors in the active site proline pocket (as seen in the crystals 1QFS and 1E8N), where we observe changes from a closed state (Fig. 7 D, yellow) to an open one (Fig. 7 D, atom type color) to better accommodate the pyrrolidine proline ring.

When the proline  $\alpha$ -carbon reached  $\sim 4$  Å from the hydroxyl oxygen of Ser-554 (part of the catalytic triad) we cleaved the substrate into two peptide fragments. At this point, we repeated the nonbiased exit simulations for the small dipeptide product as performed with ZPP. We used the open loop state and we ensured that the remaining nine-residue product peptide was not blocking the bottom pore, facilitating the possible exit of the two-residue product fragment along both pathways. Of importance, and contrary to the results with ZPP, from a total of 400 trajectories, we observe now 38 exits along the loop pathway, with only seven events through the bottom. [Movie S2](#) from the full process is deposited in the [Supporting Material](#). We should emphasize that in this simulation only the two-residue fragment is perturbed (asked to leave) in PELE's simulation, and that it does it in the presence of a bulkier nine-residue fragment, which remains the entire time in the POP's cavity. An additional (and last) simulation was performed after removing (by deleting it) the non-peptide from the POP's cavity. Thus, here the small product was let free to explore all internal volume before leaving the protein. In this case, we observed 19 exits through the loop opening and 29 through the bottom from a total number of 400 trajectories.

The dipeptide exit through the loop happens mostly from two different areas associated with a larger opening around residues Leu-206 and Thr-204 and a smaller one around Thr-202 (Fig. 8 B). Both combined, could result into an opening similar to the one predicted for bacterial POP (Fig. 2, red). For this small dipeptide product, five exit trajectories cross another loop (some authors call it loop B (13)), involving catalytic domain residues 578–604; three of them involve close interactions to Tyr-589. In only one trajectory, ZPP exits around residue Pro-74 between the hinge keeping together the catalytic and  $\beta$ -propeller domains (exits 578–604 and through the hinge are not shown on Fig. 8). Exiting around loop 578–604 shows another potential flexible part of POP. Finally, analogous to ZPP, entropic contributions for the dipeptide product gave 25.5 kcal/mol, which added to an  $\sim 10$  kcal/mol interaction energy in the solvent gave a total exothermic energy profile for product release.

## DISCUSSION

The loop prediction results, both using Prime and MD, indicate that the loop in the mammalian POP has significant mobility, confirming previous experimental and computational results (10,13). The excellent agreement of the predicted structures with the experimental ones, 0.5 Å and 0.9 Å  $\alpha$ -carbon RMSD with porcine and bacterial crystals, respectively, indicates the quality of Prime's algorithm in sampling long loops and gives credit to the open structure predictions. Moreover, we observe a large degree of overlapping between the loop prediction techniques and the MD results for the semiopen loops. Thus, one would expect that being able to run a longer MD we would observe a larger opening of the loop, similar to the one predicted by Prime. Moreover, a similar argument could be expected for domain opening. Our mammalian simulations do not

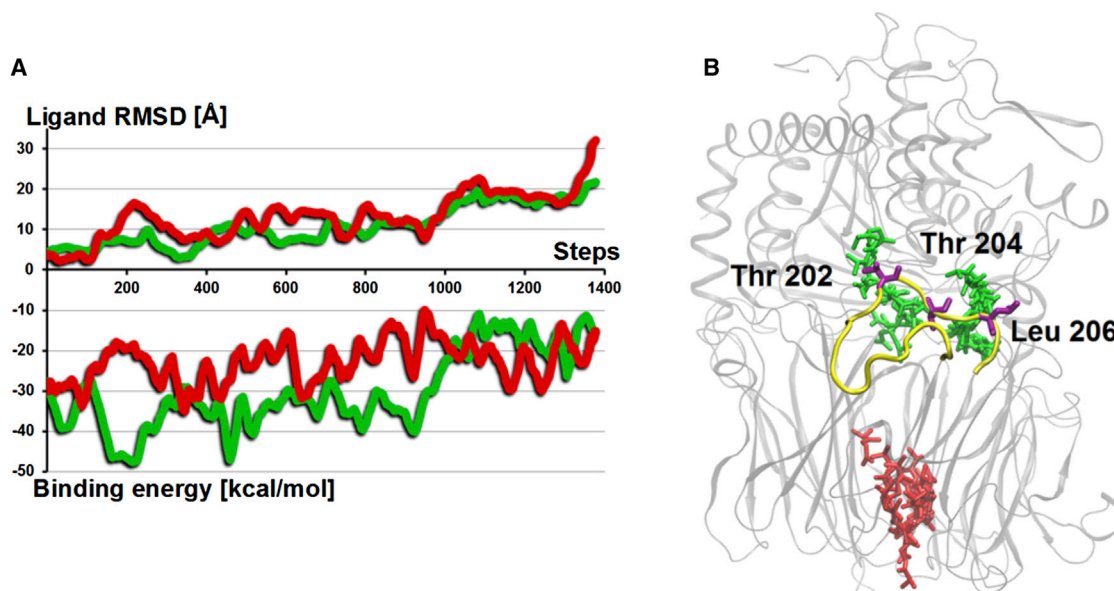


FIGURE 8 (A) Ligand interaction energy and RMSD profiles for the bottom (*red*, exiting at  $\sim 30$  Å of ligand RMSD) and loop (*green*, exiting at  $\sim 15$  Å of ligand RMSD) exit pathways for the Ala-Gly product. (B) Representative exit snapshots along the exit pathways (same color scheme). In *purple licorice*, we underline some residues from the loop (in *yellow*) where the dipeptide exits POP's cavity. To see this figure in color, go online.

present significant interdomain conformational change as the one present in the bacterial crystals. Inspection of the interdomain contacts seems to confirm larger difficulty in opening mammalian POP. Normal mode analysis, however, still indicates that the lowest modes describe domain-domain movement. Thus, one could expect that considerably larger MD simulations could introduce partially opened structures.

In previous experimental work, it showed not only the mobility of this loop but also the important contributions to the substrate enter/exit mechanisms, i.e., lower activity of a porcine POP mutant lacking this loop (10). The largest predicted opening in porcine POP is, however, not as significant as the one observed in bacterial POP. Nevertheless, the exit of the two-residue product during the simulation of the undecapeptide substrate seems to indicate that the loop opening is enough for some small product release (see below).

Our simulations indicate a clear preference for the bottom entrance. Only for the bacterial opened state do we observe partial entrance by the loop pathway, yet the statistics for this state show higher occurrence for the bottom entering (Table 1). We should keep in mind that simulations were performed with a relatively small molecule size (compared to average POP size substrates) like ZPP. Thus, for larger peptides one would expect even a larger contribution of bottom entrances.

The entrance by the bottom pore is in agreement with recent umbrella sampling simulations (8) where the authors monitor the energy profile when forcing exit pathways. In another study, using MD from a docked inhibitor in the bottom of the large internal cavity, the authors show spon-

taneous migration of the ligand toward the active site region (9). Along the different entrances through the bottom pore, we find nine residues (Glu-134, His-180, Leu-240, Ser-241, Asp-242, Asp-243, Gln-388, Lys-389, and Lys-390) having contacts closer than 4 Å with ZPP atoms. One could expect that mutations that introduce bulkier side chains in some of these positions would result in a weaker inhibitory activity (with possibly a large alteration of binding kinetics) of ZPP. Within our statistical limitations, our results indicate no orientation preference along the  $\beta$ -propeller pore entrance for small inhibitors. Furthermore, such size molecules do easily rotate and translate in the POP's internal huge cavity. Bigger peptide substrates, however, might need some guiding to pass preferably with its C-terminus. The amount of Lys side chains around the bottom pore (Lys-81, 82, 84, 157, 162, 183, 389, 390) could be this guiding tool (some of them shown in Fig. 7 A).

Although the bottom entrance in bacterial POP has similar probability to the mammalian one, exiting by the same pathway is severely more restricted (Table 1). A close look at both structures reveals clear differences in the pore residues. Mammalian Lys-81 and Lys-389, are replaced by the bulkier Arg-83 and His-377. Similarly, Asp-242 in porcine is replaced by a longer Glu-240. More importantly, Arg-135 replaces Glu-134. We observe, in  $\sim 50\%$  of cases where ZPP is in close proximity to exit by the pore, how Arg-135 blocks ZPP passage by interacting with Glu-240 and Asp-237 (Fig. 9). In the reverse (entrance) cases, however, the inhibitor molecule has more mobility and interacts closely with the pore, leading to bigger changes (including Arg-135) and to more successful entry trials—total number

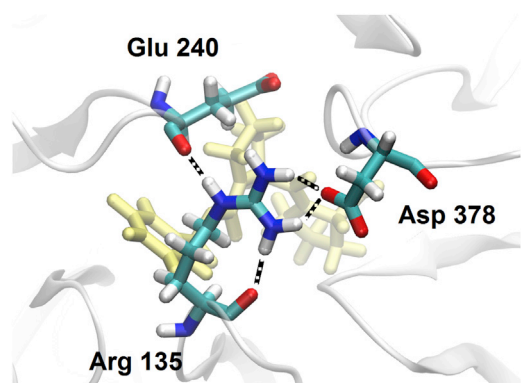


FIGURE 9 Bottom view of the  $\beta$ -propeller (white ribbons) of bacterial POP showing Arg-135 blocking the center of the pore by interacting with Glu-240 and Asp-378. ZPP is shown with yellow licorice. To see this figure in color, go online.

of 27 pore entrances to only two exits for the bottom pathway of bacterial POP (Table 1).

Our results correlate the larger degree of loop opening in bacterial POP with the appearance of exit events through it. Along the exit pathway, the hydrophobic interactions of ZPP's proline with Phe-174 and Tyr-233 closely resemble the interactions of the inhibitor in the active site (for example with residues Phe-476 and Trp-595 in porcine POP). In addition, the orientation of ZPP when exiting, Fig. 6 B, agrees with the one adopted when entering: the proline moiety finding first this hydrophobic cavity. Orientations of the side chains of Phe-174 and Tyr-233 change dynamically through both processes, adopting conformations that consecutively interact with the phenyl or proline rings in ZPP. Additional residues showing large motion in some exiting trajectories are Tyr-190 (also located in porcine POP) and Arg-232.

In an attempt to model the entire process, we diffused an 11-residue peptide from the bottom pore to the porcine active site. The substrate reached the active site with a smooth energy profile, agreeing with the preference observed in ZPP for the bottom entrance. Interestingly, after the cleavage and forming the two-residue product we observe a preference for the dipeptide exiting through the loop pathway. This difference could come not only because of the partial shielding of the bottom pore from the remainder nonapeptide, but also because of the size of the leaving ligand. We studied this hypothesis with an additional 400 simulations where we removed the nonapeptide. The 19/29 ratio for loop/bottom exits shows higher occurrence through the bottom pore, although the loop one still participates in 40% of the successful cases. This means that the option for exiting products through the bottom pore will be further impeded with the presence of a large molecule (i.e., our modeled nonapeptide product) in the POP cavity or even totally obstructed. Of significant importance in this case is the conformation of the N-terminus of the substrate and which part of the

POP  $\beta$ -propeller domain has been occupied. Our 11-residue peptide simulations show that the tail of the peptide prefers extended conformations covering the area close to the bottom. Extensive preliminary active site search and docking calculation by us, showed good interactions in proximity of the  $\beta$ -propeller pore. These observations seem to agree with the only crystal structure with a bigger substrate—a bulky octapeptide in porcine POP, PDB entry 1E8N (32). In this crystal, the N-terminus is pointing toward the bottom, whereas the C-terminus is not well resolved. Thus, all together this indicates that longer peptide will obstruct the bottom passage and drive the products release through the loop opening.

Our binding energy and entropy estimates indicate large compensation effects. Although this topic has been under debate, numerous recent calorimetric studies seem to support its importance (33,34). We have studied three peptide-like substrates with large flexibility and entropy loss upon binding, in agreement with recent observations (31). Interestingly, the polar large cavity in POP seems to have evolved to compensate for this reduction in mobility by increasing the number of protein-ligand interactions. This is clear when inspecting the binding energy plots where we observe a sudden large increase (in absolute value) once the ligand enters the cavity (see, for example, Fig. 7 C at steps ~50). Due to the large errors in entropy calculations, however, any attempt to obtain accurate binding free energies should use more sophisticated methods; our binding energies are only of qualitative nature.

In addition to the large polar cavity, our PELE and MD simulations show a hydrophobic pocket (similar to the one for proline in the active site of POP) buried and uncovered by the loop motions, which could be a trigger mechanism for peptide release. Thus, product exit seems to follow three steps: initial ligand binding to the predocking site followed by a larger opening of the loop, which pulls out the products in the same direction and then exit through it. This mechanism would agree with the experimental results showing that mutants lacking a flexible loop convert POP to an inefficient enzyme (10).

In summary, our extensive computational analysis reveals a clear preference for ligand entrance through the  $\beta$ -propeller pore. Exit conditions, however, seem to be more specific of the species, degree of loop opening, and nature of the substrate. Overall, cleavage of a small peptide at the active site seems to be correlated with its exit along the loop. This loop is shown to be very flexible in our simulations; modeling domain-domain opening in mammalian POP (if present) will require considerably longer simulations.

## SUPPORTING MATERIAL

Two movies are available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(14\)04666-9](http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)04666-9).

## ACKNOWLEDGMENTS

This study was supported by The European Research Council (2009-Adg25027-PELE) to V.G., MICIN-FEDER (BIO2013-40716R) to E.G., the Generalitat de Catalunya (XRB and 2014SGR-521) to E.G., and SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government to D.L.

## REFERENCES

- Fülöp, V., Z. Böcskei, and L. Polgár. 1998. Prolyl oligopeptidase: an unusual  $\beta$ -propeller domain regulates proteolysis. *Cell*. 94:161–170.
- Brandt, I., S. Scharpé, and A.-M. Lambeir. 2007. Suggested functions for prolyl oligopeptidase: a puzzling paradox. *Clin. Chim. Acta*. 377:50–61.
- Lawandi, J., S. Gerber-Lemaire, ..., N. Moitessier. 2010. Inhibitors of prolyl oligopeptidases for the therapy of human diseases: defining diseases and inhibitors. *J. Med. Chem.* 53:3423–3438.
- López, A., T. Tarragó, and E. Giralt. 2011. Low molecular weight inhibitors of Prolyl Oligopeptidase: a review of compounds patented from 2003 to 2010. *Expert Opin. Ther. Pat.* 21:1023–1044.
- Brandt, I., M. Gérard, ..., A.-M. Lambeir. 2008. Prolyl oligopeptidase stimulates the aggregation of  $\alpha$ -synuclein. *Peptides*. 29:1472–1478.
- Myöhänen, T. T., M. J. Hannula, ..., A.-M. Lambeir. 2012. A prolyl oligopeptidase inhibitor, KYP-2047, reduces  $\alpha$ -synuclein protein levels and aggregates in cellular and animal models of Parkinson's disease. *Br. J. Pharmacol.* 166:1097–1113.
- Kichik, N., T. Tarragó, ..., E. Giralt. 2011.  $^{15}\text{N}$  relaxation NMR studies of prolyl oligopeptidase, an 80 kDa enzyme, reveal a pre-existing equilibrium between different conformational states. *ChemBioChem*. 12:2737–2739.
- St-Pierre, J.-F., M. Karttunen, ..., A. Bunker. 2011. Use of umbrella sampling to calculate the entrance/exit pathway for Z-Pro-proline inhibitor in prolyl oligopeptidase. *J. Chem. Theory Comput.* 7:1583–1594.
- Kaushik, S., and R. Sowdhamini. 2011. Structural analysis of prolyl oligopeptidases using molecular docking and dynamics: insights into conformational changes and ligand binding. *PLoS One*. 6:e26251.
- Szeltner, Z., T. Juhász, ..., L. Polgár. 2013. The loops facing the active site of prolyl oligopeptidase are crucial components in substrate gating and specificity. *Biochim. Biophys. Acta*. 1834:98–111.
- Szeltner, Z., D. Rea, ..., L. Polgár. 2004. Concerted structural changes in the peptidase and the propeller domains of prolyl oligopeptidase are required for substrate binding. *J. Mol. Biol.* 340:627–637.
- Li, M., C. Chen, ..., T. K. Chiu. 2010. Induced-fit mechanism for prolyl endopeptidase. *J. Biol. Chem.* 285:21487–21495.
- Kaszuba, K., T. Róg, ..., A. Bunker. 2012. Molecular dynamics, crystallography and mutagenesis studies on the substrate gating mechanism of prolyl oligopeptidase. *Biochimie*. 94:1398–1411.
- Buch, I., T. Giorgino, and G. De Fabritiis. 2011. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA*. 108:10184–10189.
- Shan, Y., E. T. Kim, ..., D. E. Shaw. 2011. How does a drug molecule find its target binding site? *J. Am. Chem. Soc.* 133:9181–9183.
- Borrelli, K. W., A. Vitalis, ..., V. Guallar. 2005. PELE: protein energy landscape exploration. A novel Monte Carlo based technique. *J. Chem. Theory Comput.* 1:1304–1311.
- Takahashi, R., V. A. Gil, and V. Guallar. 2014. Monte Carlo free ligand diffusion with Markov state model analysis and absolute binding free energy calculations. *J. Chem. Theory Comput.* 10:282–288.
- Madadkar-Sobhani, A., and V. Guallar. 2013. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Res.* 41 (Web Server issue):W322–W328.
- Borrelli, K. W., B. Cossins, and V. Guallar. 2010. Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J. Comput. Chem.* 31:1224–1235.
- Cossins, B. P., A. Hosseini, and V. Guallar. 2012. Exploration of protein conformational change with PELE and meta-dynamics. *J. Chem. Theory Comput.* 8:959–965.
- Schrödinger Suite. 2013. Protein Preparation Wizard; Epik version 2.6, Schrödinger, LLC, New York, NY; Impact version 6.1, Schrödinger, LLC; Prime version 3.4, Schrödinger, LLC.
- Jacobson, M. P., R. A. Friesner, ..., B. Honig. 2002. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* 320:597–608.
- Jacobson, M. P., G. A. Kaminski, ..., C. S. Rapp. 2002. Force field validation using protein side chain prediction. *J. Phys. Chem. B*. 106:11673–11680.
- Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2:173–181.
- Kaminski, G. A., R. A. Friesner, ..., W. L. Jorgensen. 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*. 105:6474–6487.
- Ghosh, A., C. S. Rapp, and R. A. Friesner. 1998. Generalized Born model based on a surface integral formulation. *J. Phys. Chem. B*. 102:10983–10990.
- Molecular Dynamics System, 2012, version 3.1, D. E. Shaw Research, New York, NY. Maestro-Desmond Interoperability Tools, version 3.1, Schrödinger LLC, New York, NY.
- Bowers, K. J., E. Chow, ..., D. E. Shaw. 2006. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *Proc. ACM/IEEE Conf. Supercomputing (SC06)*. Tampa, Florida, November 11–17, 2006.
- Hoover, W. G. 1986. Constant-pressure equations of motion. *Phys. Rev. A*. 34:2499–2500.
- Martyna, G. J., D. J. Tobias, and M. L. Klein. 1994. Constant pressure molecular-dynamics algorithms. *J. Chem. Phys.* 101:4177–4189.
- Killian, B. J., J. Y. Kravitz, ..., M. K. Gilson. 2009. Configurational entropy in protein-peptide binding: computational study of Tsg101 ubiquitin E2 variant domain with an HIV-derived PTAP nonapeptide. *J. Mol. Biol.* 389:315–335.
- Fülöp, V., Z. Szeltner, ..., L. Polgár. 2001. Structures of prolyl oligopeptidase substrate/inhibitor complexes. Use of inhibitor binding for titration of the catalytic histidine residue. *J. Biol. Chem.* 276:1262–1266.
- Chodera, J. D., and D. L. Mobley. 2013. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* 42:121–142.
- Portman, K. L., J. Long, ..., D. J. Scott. 2014. Enthalpy/entropy compensation effects from cavity desolvation underpin broad ligand binding selectivity for rat odorant binding protein 3. *Biochemistry*. 53:2371–2379.

## Publication 2 - Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints

**Authors:** Karl Edman, Ali Hosseini, Magnus K. Bjursell, Anna Aagaard, Lisa Wissler, Anders Gunnarsson, Tim Kaminski, Christian Köhler, Stefan Bäckström, Tina J. Jensen, Anders Cavallin, Ulla Karlsson, Ewa Nilsson, Daniel Lecina, Ryoji Takahashi, Christoph Grebner, Stefan Geschwindner, Matti Lepistö, Anders C. Hogner, Victor Guallar.

**Journal:** Structure, 23, 2280–2291 (2015)

### Summary:

In this work, we studied the flexibility of NHRs. In particular, we solved *X-ray* structures of glucocorticoid (GR) and mineralocorticoid receptors (MR) to identify a conserved plasticity at the helix 6-7 region. To support the idea that it constitutes an integral part of the binding event, we launched entrance, exit, and refinement simulations. NHRs present deeply buried binding sites, and the study of ligand migration is currently out of reach for standard all-atom unbiased procedures. For this reason, PELE was used for the sampling in this study. We developed a procedure of confined sampling around the shared entrance point (~10-15Å) and we were able to compute the binding free energy difference between dexamethasone and desisobutyrylciclesonide (dibC), both complexed with MR. Residence time measures correlate with the magnitude of structural rearrangements in both structures. All in all, we show that nature has conserved the capacity to open up this region, which impose different evolutionary constraints across the steroid receptors.

### Author contribution:

My task in this publication was the analysis with MSM of 400 unbiased MC trajectories for the MR with dexamethasone and dibC. The analysis included computing the binding free energies in addition to the potential of mean force.

**Comment:** An error in the units was corrected in figure 7.



# Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints

Karl Edman,<sup>1,\*</sup> Ali Hosseini,<sup>2</sup> Magnus K. Bjursell,<sup>3</sup> Anna Aagaard,<sup>1</sup> Lisa Wissler,<sup>1</sup> Anders Gunnarsson,<sup>1</sup> Tim Kaminski,<sup>1</sup> Christian Köhler,<sup>4</sup> Stefan Bäckström,<sup>1</sup> Tina J. Jensen,<sup>4</sup> Anders Cavallin,<sup>4</sup> Ulla Karlsson,<sup>1</sup> Ewa Nilsson,<sup>1</sup> Daniel Lecina,<sup>2</sup> Ryoji Takahashi,<sup>2</sup> Christoph Grebner,<sup>5</sup> Stefan Geschwindner,<sup>1</sup> Matti Lepistö,<sup>4</sup> Anders C. Hogner,<sup>5,\*</sup> and Víctor Guallar<sup>2,6,\*</sup>

<sup>1</sup>Discovery Sciences, AstraZeneca, Mölndal, Pepparedsleden 1, 43183 Mölndal, Sweden

<sup>2</sup>Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Jordi Girona 29, 08034 Barcelona, Spain

<sup>3</sup>R&D Information, AstraZeneca, Pepparedsleden 1, 43183 Mölndal, Sweden

<sup>4</sup>RIA, AstraZeneca, Pepparedsleden 1, 43183 Mölndal, Sweden

<sup>5</sup>CVMD, AstraZeneca, Pepparedsleden 1, 43183 Mölndal, Sweden

<sup>6</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain

\*Correspondence: [anders.hogner@astrazeneca.com](mailto:anders.hogner@astrazeneca.com) (A.C.H.), [victor.guallar@bsc.es](mailto:victor.guallar@bsc.es) (V.G.), [karl.edman@astrazeneca.com](mailto:karl.edman@astrazeneca.com) (K.E.)  
<http://dx.doi.org/10.1016/j.str.2015.09.012>

## SUMMARY

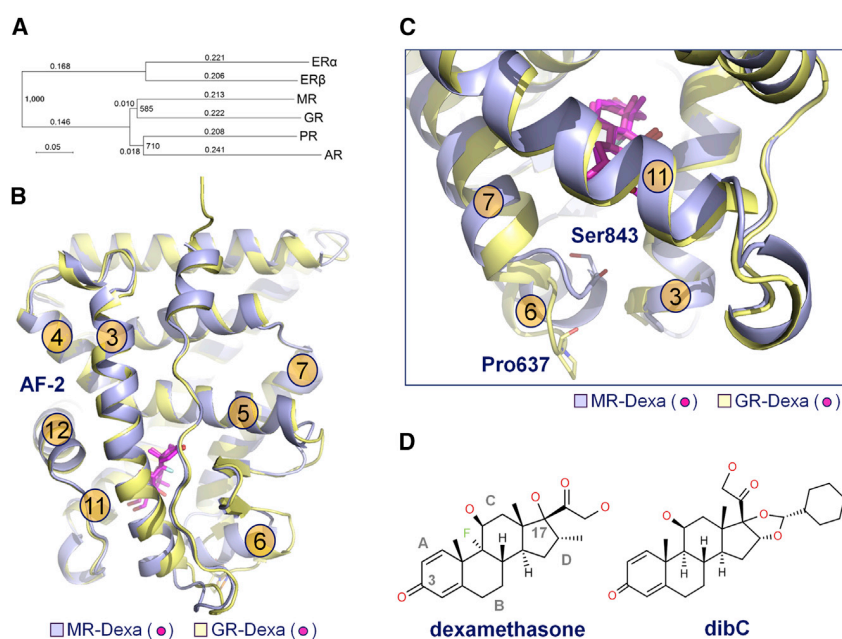
Steroid receptor drugs have been available for more than half a century, but details of the ligand binding mechanism have remained elusive. We solved X-ray structures of the glucocorticoid and mineralocorticoid receptors to identify a conserved plasticity at the helix 6–7 region that extends the ligand binding pocket toward the receptor surface. Since none of the endogenous ligands exploit this region, we hypothesized that it constitutes an integral part of the binding event. Extensive all-atom unbiased ligand exit and entrance simulations corroborate a ligand binding pathway that gives the observed structural plasticity a key functional role. Kinetic measurements reveal that the receptor residence time correlates with structural rearrangements observed in both structures and simulations. Ultimately, our findings reveal why nature has conserved the capacity to open up this region, and highlight how differences in the details of the ligand entry process result in differential evolutionary constraints across the steroid receptors.

## INTRODUCTION

Biological functions originate from, and are maintained by, a combination of genomic drift and selection. The traditional method to derive evolutionary relationships is to compare primary sequences, tertiary structures, and protein function. However, while changes in the amino acid sequence and placement of key residues provide useful insights into lineage, this only provides the basic framework for mechanistic detail. A more complete functional understanding requires protein plasticity to be considered. Moreover, comparing protein flexibility of related systems adds an important dimension when exploring evolutionary trajectories (Bhabha et al., 2013).

The steroid receptor family consists of five closely related receptors: the mineralocorticoid receptor (MR), the glucocorticoid receptor (GR), the androgen receptor (AR), the progesterone receptor (PR), and the estrogen receptors (ER $\alpha$  and ER $\beta$ ) (Figure 1A). All these receptors bind cholesterol derivatives and play a critical role in fundamental biological processes, ranging from pregnancy to early development, the stress response, and electrolyte homeostasis (Evans, 1988; Mangelsdorf et al., 1995). Continual pharmaceutical efforts have resulted in several efficacious drugs across the family (Cole, 2006; Gravez et al., 2013; Shelley et al., 2008; Sitruk-Ware and Nath, 2010; Alexander et al., 2013). However, target class-related side effects limit the prescription of these drugs for many indications, and the scope for further improvement is considered to be high (Bertocchio et al., 2011). The receptors share a common architecture with three separate domains: the N-terminal domain (NTD), the DNA binding domain, and the ligand binding domain (LBD). Besides recognizing the ligand pharmacophore, the LBD also contains the activation function 2 (AF-2), which is important for transmitting ligand binding information and partially driving the co-regulator interaction fingerprint (Grone-meyer et al., 2004). In the resting state, the receptors are associated with chaperone proteins in the cytoplasm. Ligand activation leads to a partial release of chaperone proteins, followed almost always by nuclear translocation. In the nucleus, the receptors dimerize and form ligand and context-specific protein complexes, resulting in activation and/or repression of gene transcription.

All steroid receptor LBD structures exhibit the typical three-layered  $\alpha$ -helical fold that fully encloses the various compounds in the ligand binding pocket (Bledsoe et al., 2002; Williams and Sigler, 1998; Fagart et al., 2005; Matias et al., 2000) (Figure 1B). When overlaying the steroid receptors, the largest structural difference in proximity to the ligand is located in the region where helices 3, 7, and 11 meet (Li et al., 2005). Figure 1C shows a detailed comparison of GR with its paralog MR. An outward tilt of the helix 6–7 (H6-H7) interface in GR results in an expanded ligand binding pocket, and the most potent GR ligands contain large substituents extending in this direction



**Figure 1. Evolutionary Relationship of the Steroid Receptors with Structural Comparison of GR- and MR-LBD**

(A) Evolutionary relationship of the steroid hormone receptors (ER $\alpha$ , ER $\beta$ , MR, GR, PR, and AR). Decimal numbers = distance; integers = bootstrap value.

(B) GR (yellow) in complex with dexamethasone (magenta) overlaid on MR (light blue) in complex with dexamethasone (magenta). The AF-2 surface is located where helices 3, 4, and 12 meet.

(C) Details near the region where helices 3, 7, and 11 meet.

(D) The chemical structures of dexamethasone and dibC. The steroidal A, B, C, and D rings and positions 3 and 17 are marked on the dexamethasone structure.

(17 $\alpha$ ). Despite the smaller pocket in MR, several ligands with bulky 17 $\alpha$  substituents on the steroidal D-ring, such as desisobutyrylciclesonide (dibC, the active metabolite of the prodrug ciclesonide), are more potent in the MR binding assay than the endogenous agonist aldosterone.

Plasticity in the H6-H7 region has been reported for ER $\alpha$ , AR, and PR (Andrieu et al., 2015; Nettles et al., 2007; Kohn et al., 2012), and appears to be a conserved feature across the nuclear receptor superfamily (Soisson et al., 2008; Hughes et al., 2012). To build a detailed understanding for how the differences in receptor design influence the H6-H7 rearrangements, we determined the X-ray structures of both MR and GR in complex with dexamethasone and dibC (Figure 1D). The structures revealed that when binding a ligand with a large 17 $\alpha$  substituent, MR is fully capable of adopting an open structural conformation, and that the nature of these rearrangements is clearly distinct from analogous changes in GR. Why has nature preserved the capacity to open up this region across the steroid receptor family, even though it is not exploited by the endogenous ligands? Our hypothesis is that the observed plasticity is an integral part of the ligand entry mechanism.

To test this hypothesis, we performed comprehensive all-atom unbiased simulations. In these studies, we linked the observed plasticity in the H6-H7 region to the ligand binding mechanism. While the simulations clearly identified a common binding trajectory for the two receptors, they also highlighted detailed differences in the entry and exit processes. By employing surface plasmon resonance (SPR) and single-molecule microscopy (SMM), we showed that these differences correlate with distinct ligand-receptor residence times. Finally, we performed a bioinformatics analysis whereby we confirmed that GR has relaxed evolutionary constraints on the H6-H7 amino acid sequence relative all other steroid receptors. The link to the ligand binding utility provides a functional understanding for these observations.

## RESULTS

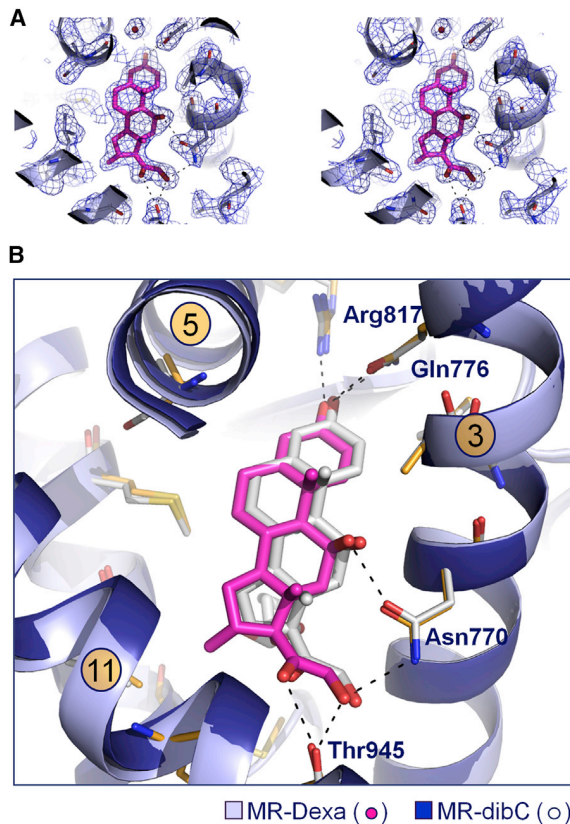
### A Conserved Plasticity

Dexamethasone was originally developed as a GR-specific agonist (Alexander et al., 2013) and was used to determine the first GR-LBD structure (Bledsoe et al.,

2002). However, dexamethasone was later shown to also be a potent MR ligand in a functional reporter gene assay (Rupprecht et al., 1993). The X-ray structure of MR in complex with dexamethasone (MR:Dexa, Figure 2A) is similar to the corresponding GR:Dexa structure (normalized root-mean-square deviation [RMSD] of 0.37 Å for 100 C $\alpha$  atoms) (Table 1). However, examining the region where helices 3, 7, and 11 meet confirms that the 17 $\alpha$  subpocket is considerably smaller in the MR structure than in the GR structure (Figure 1C). This is reflected in the total volume of the MR:Dexa ligand binding pocket, which is approximately 543 Å<sup>3</sup> compared with 572 Å<sup>3</sup> in the GR:Dexa structure (Figure S1).

It has been proposed that structural differences in the loop between helices 6 and 7 are primarily due to replacement of Ser843<sup>MR</sup> by Pro637<sup>GR</sup>, which alters the geometrical constraints of this region and allows GR to adopt a more open conformation (Li et al., 2005). However, despite the limited size of the MR subpocket, dibC has higher affinity than aldosterone in the scintillation proximity assay using tritiated aldosterone and MR-LBD fusion protein ( $K_i$  for dibC is 0.18 nM compared with 1.0 nM for aldosterone, Figure S2). To study the structural flexibility associated with large 17 $\alpha$  substituents, we determined the complex structures of MR:dibC and GR:dibC (Table 1).

The structure of MR:dibC superimposes well on the MR:Dexa structure (normalized RMSD 0.28 Å for 100 C $\alpha$  atoms). dibC is placed in a nearly identical position as dexamethasone in the binding pocket, with all polar interactions conserved (Figure 2B). In addition, the AF-2 surface remains virtually unchanged, with key interactions to the NCOA1 peptide intact. However, while these two receptor conformations are closely related, dibC induces a large rearrangement of the H6-H7 loop region, essentially extending the ligand binding pocket toward the receptor surface (Figure 3A). Specifically, side chains of Ser843<sup>MR</sup>, Met845<sup>MR</sup>, and Cys849<sup>MR</sup> in the MR:Dexa complex occupy the same volume as the cyclohexyl motif of dibC, forcing the



**Figure 2. Comparison of the Complex Structures of MR:Dexa and MR:dibC**

(A) Stereo view of the  $2mF_o-dF_c$  density map of the MR:Dexa ligand binding pocket.

(B) The structure of MR (light blue) in complex with dexamethasone (magenta) superimposed on MR (dark blue) in complex with dibC (white). The steroid template overlays nearly perfectly (RMSD 0.28 Å) with all hydrophilic interactions conserved.

receptor to adopt a new conformation (Figure 3B). This leads to a repositioning of helix 6 and an extension of helix 7. While Ser843<sup>MR</sup> was previously buried within the protein and engaged in a hydrogen bond to the backbone nitrogen of Met845<sup>MR</sup>, it is now exposed to the solvent, forming the new start of helix 7 (Figure 3A). Recent data suggest that phosphorylation of this residue affects both ligand binding and receptor translocation into the nucleus (Shibata et al., 2013). The structural changes observed here explain how the receptor may use the local plasticity to make Ser843<sup>MR</sup> available for modification.

The size of the  $17\alpha$  pocket in the MR:dibC complex increases significantly (total ligand binding pocket volume 714 Å<sup>3</sup>, Figure S1), and the superposition on the GR:Dexa structure shows that this region now adopts a more closely related structural state (Figure 3C). Finally, while GR in complex with dibC (Figure 3D) expands the  $17\alpha$  pocket (total ligand binding pocket volume 661 Å<sup>3</sup>, Figure S1) relative to the GR:Dexa structure, it does not alter any of the secondary structural elements. Instead, the H6-H7 region appears to be shifted in a rigid way in response to cyclohexyl of dibC. While plasticity in the H6-H7 region seems

to be conserved across these two receptors, the details of the ligand-driven rearrangements are different.

To quantify the flexibility in the H6-H7 region across the steroid receptor family, we performed principal component analysis for all X-ray structures from the PDB for each receptor. This allows visualization of the variance between structures as a set of normal modes. While the description of this variance will be highly dependent on what regions of the binding pocket are exploited by the various ligands, the mode describing H6-H7 motion is one of the strong features (Figure S3). However, for MR the H6-H7 motion is only prominent if we include the MR:dibC structure from this work, emphasizing that the MR:dibC structure describes a novel structural conformation.

### Modeling Nonbiased Entry and Exit Pathways

Spontaneous ligand binding events have been investigated using molecular dynamics in both exposed (Buch et al., 2011) and partially exposed binding sites (Dror et al., 2011). However, nuclear receptors have fully occluded binding pockets that likely require significant rearrangements for ligand entry. Therefore, we decided to use protein energy landscape exploration (PELE) (Borrelli et al., 2005), an alternative approach that uses Monte Carlo algorithms with structural prediction for efficient sampling of the protein-ligand energy landscape. For ligand escape simulations, the MR and GR X-ray complex structures were used as the starting position. For ligand binding studies, the ligand was randomly placed in the bulk solvent and allowed to freely migrate. All simulations were completed in the presence and absence of a co-factor peptide at the AF-2 site (NCOA1 residues 1,430–1,441 for MR and NCOA2 residues 741–753 for GR). In addition, both the wild-type protein sequences and the specific mutants present in the X-ray structures were used.

### Ligand Dissociation

For all permutations of both MR and GR, we performed three separate exit simulations, observing only one exit trajectory perforating the surface where helices 3, 7, and 11 meet. Figure 4A illustrates the MR:Dexa exit pathway simulation with the array of dexamethasone positions superimposed on the initial MR structure. Notably, ligand motion is coupled with significant rearrangement of the protein backbone along the migration pathway. In particular, the loop connecting helices 6 and 7 is shifted outward to accommodate ligand release (Figure 4B). Interestingly, the simulated protein movements mimic the differences between the MR:Dexa and MR:dibC structures shown in light and dark blue, respectively. Root-mean-square fluctuations (RMSF) along the exit trajectory (Figure 4C) clearly show that the movements of the H6-H7 region are considerably larger than for the rest of the protein.

Figure 5 shows the corresponding simulation for GR:Dexa (equivalent simulations for MR:dibC and GR:dibC resulted in the same exit trajectory). Based on the complete set of ligand dissociation simulations it is apparent that both MR and GR have the same ligand unbinding pathway. In addition, while ligand exit is associated with similar protein motions, the fluctuations in the H6-H7 region are significantly larger for MR than for GR (Figure 5C). This is in agreement with the idea that GR would require smaller rearrangements, because the receptor is more open to begin with.

**Table 1. Data Collection and Refinement Statistics**

|  | MR:Dexa               | MR:dibC                | GR:Dexa                | GR:dibC                |
|--|-----------------------|------------------------|------------------------|------------------------|
| Data Collection <sup>a</sup>           |                       |                        |                        |                        |
| PDB ID                                 | 4UDA                  | 4UDB                   | 4UDC                   | 4UDD                   |
| Space group                            | P212121               | P41212                 | P3221                  | P3221                  |
| a, b, c (Å)                            | 73.00, 81.40, 45.23   | 75.92, 75.92, 117.00   | 84.66, 84.66, 105.91   | 87.20, 87.20, 102.89   |
| α, β, γ (°)                            | 90.00, 90.00, 90.00   | 90.00, 90.00, 90.00    | 90.00, 90.00, 120.00   | 90.00, 90.00, 120.00   |
| Resolution (Å)                         | 40.7–2.03 (2.17–2.03) | 48.79–2.36 (2.55–2.36) | 31.81–2.50 (2.67–2.50) | 40.14–1.80 (1.85–1.80) |
| R <sub>sym</sub> (R <sub>merge</sub> ) | 0.06 (0.50)           | 0.13 (1.30)            | 0.08 (0.55)            | 0.08 (1.05)            |
| I/σI                                   | 13.10 (2.30)          | 15.10 (1.90)           | 8.80 (1.60)            | 7.40 (0.70)            |
| Completeness (%)                       | 83.9 (83.7)           | 100.0 (100.0)          | 99.6 (99.5)            | 99.9 (100.0)           |
| Redundancy                             | 3.3 (2.5)             | 12.6 (11.7)            | 4.1 (4.2)              | 3.5 (3.6)              |
| Refinement                             |                       |                        |                        |                        |
| Resolution (Å)                         | 2.03                  | 2.36                   | 2.50                   | 1.80                   |
| No. of reflections                     | 15,085                | 14,672                 | 15,559                 | 42,339                 |
| R <sub>work</sub> /R <sub>free</sub>   | 0.185/0.240           | 0.182/0.218            | 0.210/0.253            | 0.213/0.224            |
| No. of atoms                           |                       |                        |                        |                        |
| Protein                                | 2,080                 | 2,118                  | 2,133                  | 2,184                  |
| Ligand/ion                             | 34                    | 49                     | 64                     | 146                    |
| Water                                  | 101                   | 60                     | 83                     | 250                    |
| B factors                              |                       |                        |                        |                        |
| Protein                                | 30.14                 | 53.25                  | 49.72                  | 33.25                  |
| Ligand/ion                             | 22.12                 | 44.16                  | 34.51                  | 23.55                  |
| Water                                  | 36.03                 | 56.86                  | 46.23                  | 46.95                  |
| RMSD                                   |                       |                        |                        |                        |
| Bond lengths (Å)                       | 0.010                 | 0.010                  | 0.010                  | 0.010                  |
| Bond angles (°)                        | 1.01                  | 1.04                   | 1.12                   | 1.06                   |
| MolProbity score                       |                       |                        |                        |                        |
| Clashscore                             | 2                     | 1                      | 1                      | 1                      |
| Ramachandran outliers (%)              | 0                     | 0                      | 0.4                    | 0                      |
| Side-chain outliers (%)                | 1.7                   | 1.7                    | 2.5                    | 0.8                    |

<sup>a</sup>Values in parentheses represent highest-resolution shell.

### Ligand Association

To investigate ligand entry, we randomly placed dexamethasone in the bulk solvent and released it to freely probe the protein surface. For each receptor we performed five runs with 64 independent trajectories over 48 hr. Each run yielded one to two trajectories whereby the ligand entered the binding pocket. In all runs the ligand is free to move without any predefined search direction.

Figure 6A shows the evolution of the ligand heavy atom RMSD to the crystallographic complex for one of the MR:Dexa runs. It is clear that most of the trajectories explore the receptor surface with some excursions into the bulk solvent. However, the blue and red trajectories enter the ligand binding pocket at steps ~50 and ~210, respectively. While the entry along the blue trajectory is relatively fast, the red demonstrates the unbiased nature of the simulation, probing a large portion of the receptor surface before finding the entrance pathway. Figure 6B shows representative ligand centers of mass along these trajectories superimposed on the initial protein structure, with the entry to the binding pocket denoted by a surface representation. The corresponding ligand entry simulation for GR is shown in Figure S4. In keeping with the ligand escape simulations for all

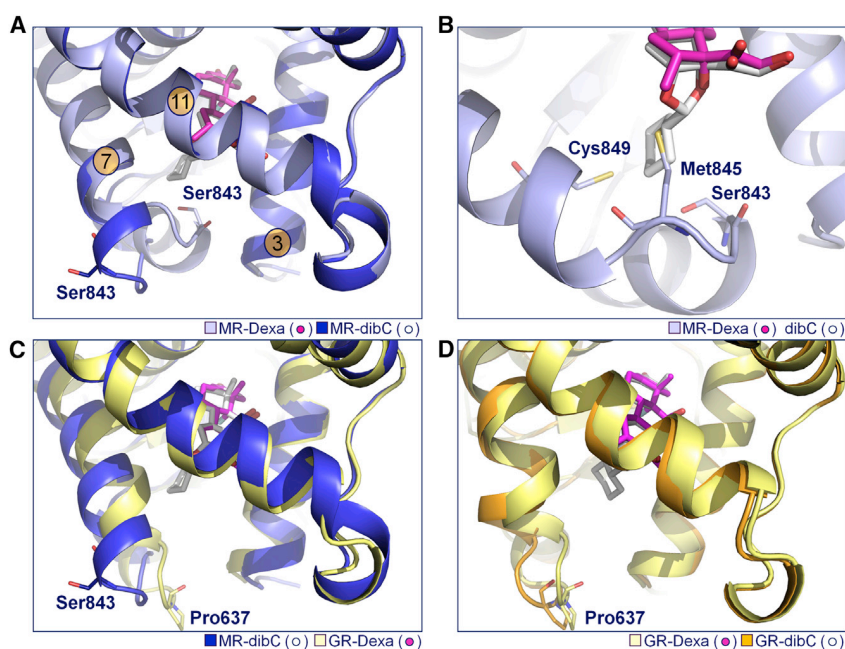
runs in both systems, trajectories entering the ligand binding pocket pierce the protein surface at the H3-H7-H11 junction. The MR:Dexa binding event is demonstrated in greater detail in the Movie S1.

While the mutants used in the X-ray structures did not influence the simulations significantly, removal of co-factor peptide at the AF-2 resulted in larger fluctuations in both helix 12 and the H3-H7-H11 junction along the exit and entrance trajectories. However, the ligand entry pathway remained unchanged. The presence of co-regulator peptide has been shown to affect the ligand binding kinetics (Pfaff and Fletterick, 2010).

### Active-Site Ligand Refinement and Binding Free Energy

Once the entrance path to the MR binding pocket had been located, we refined the free search with local enhanced sampling to obtain a precise pose for the best binder. This procedure does not add any bias in the ligand search direction, but limits the sampling to the region around the entrance point (typically 10–15 Å). Figure 7A shows the interaction energy profile plotted against the ligand heavy atom RMSD to the crystallographic complex for the MR:Dexa refining process (400 trajectories).





**Figure 3. Comparison of the Complex Structures of MR:Dexa, MR:dibC, GR:Dexa, and GR:dibC**

(A) MR (light blue) in complex with dexamethasone (magenta) overlaid on MR (dark blue) in complex with dibC (white).

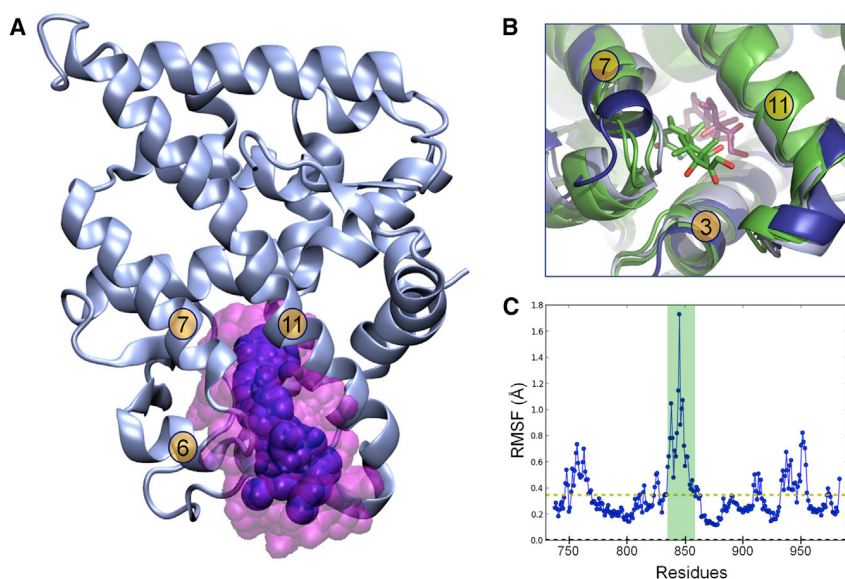
(B) The cyclohexyl motif of dibC comes into direct conflict with residues from H7 (MR:Dexa), enforcing a new structural state.

(C) MR (dark blue) in complex with dibC (white) superimposed on GR (yellow) in complex with dexamethasone (magenta).

(D) GR (yellow) in complex with dexamethasone (magenta) overlaid on GR (orange) in complex with dibC (white).

The lowest binding energies are derived from poses located within 0.75 Å RMSD of the X-ray ligand conformation. The sampling places dexamethasone in the accurate orientation with the A-ring 3-keto moiety pointing toward the Arg817<sup>MR</sup>-Gln776<sup>MR</sup> pair from helices 5 and 3, and the D-ring hydroxyacetyl approaching the Asn770<sup>MR</sup> on the N-terminal half of helix 3 (Figure 7B). Studying the protein-ligand interaction energy plot in more detail (Figure 7A), it is interesting that the surface exploration exhibits local minima near RMSD of 12 Å. In the crystal structure of GR:Dexa and GR:dibC, this site is occupied by a steroid-like CHAPS molecule that is part of the protein formulation (Figure S5). In addition, for MR a nonsteroidal antagonist has been observed at this position (Hasui et al., 2011). As such, the

region may correspond to a peripheral binding site at the H3-H7-H11 junction, and the energy barrier located at the 11- to 12-Å segment in Figure 7A reflects the energy cost associated with the surface-crossing event through the entry channel. The fast performance of PELE, together with the local restriction in the refinement exploration, facilitates running hundreds of trajectories. Based upon Markov state model (MSM) analysis (Takahashi et al., 2014), we used these data to calculate the binding free energies for MR:Dexa and MR:dibC. While absolute values might be slightly shifted due to the absence of an exhaustive surface/bulk exploration, relative values should be in reasonable agreement, because both ligands share entry point and binding site. Figure 7C shows a 2D projection of the potential mean field (PMF) obtained for MR:Dexa along the 400 refinement trajectories. The red area corresponds to the bulk exploration, whereas the global minimum, shown in blue, corresponds to ligand positions matching the experimental structure (Figures 7A and 7B). Integration of the PMF volume at the active site,

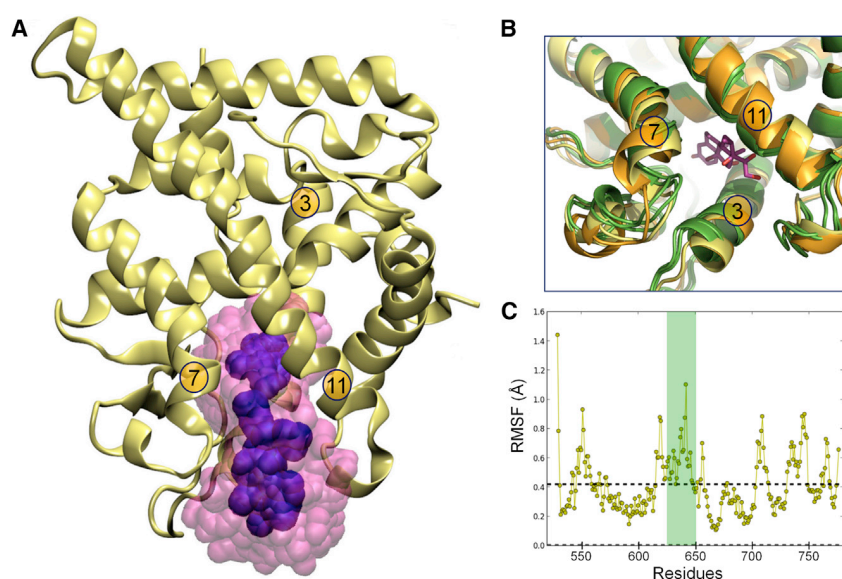


**Figure 4. Ligand Exit Pathway for the MR:Dexa Complex**

(A) The ligand center of mass is highlighted as blue beads. The ligand atoms are shown as transparent space fill.

(B) Detail of the backbone rearrangement along the exit pathway. The MR:Dexa and MR:dibC X-ray structures are shown in light and dark blue, respectively, with dexamethasone in the binding pocket in magenta. Three protein cartoon snapshots and one pose of dexamethasone as it passes through the receptor surface from the exit simulations are shown in green.

(C) C<sub>α</sub> RMSF relative the average structure along the MR:Dexa exit pathway plotted for each residue. The dotted line denotes the average RMSF across the LBD. Helices 6 and 7 are marked with green shading.



**Figure 5. Ligand Exit Pathway for the GR:Dexa Complex**

(A) The ligand center of mass is highlighted as blue beads. The ligand atoms are shown as transparent space fill.

(B) Detail of the backbone rearrangement along the exit pathway. The GR:Dexa and GR:dibC X-ray structures are shown in light yellow and orange, respectively. Three snapshots from the exit simulations are shown in green, and dexamethasone in the binding pocket is shown for reference in magenta.

(C)  $C_{\alpha}$  RMSF relative the average structure along the GR:Dexa exit pathway where helices 6 and 7 are marked with green shading.

where we observe a smooth function (as opposed to the bulk solvent or entrance pathway), converges to a binding free energy of  $-7.5$  kcal/mol for dexamethasone and  $-9.3$  kcal/mol for dibC. The difference in binding free energy of  $1.8$  kcal/mol is in quantitative agreement with the experimental difference of  $2.09$  kcal/mol (derived from the  $K_i$  values of  $6.3$  nM for dexamethasone and  $0.18$  nM for dibC).

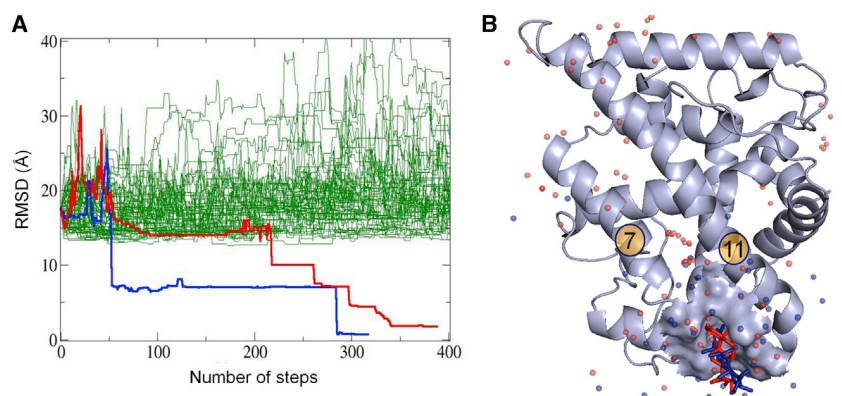
### Residence Time Measurements

The ligand entry and exit mechanism establishes a functional role for helices 6 and 7 as a gatekeeper. In addition, the simulations revealed that the structural rearrangements required for ligand entry and exit are significantly different for GR and MR. As a consequence, the ligand binding kinetics should differ for the two receptors. Using both SPR and SMM (Gunnarsson et al., 2015), we measured the residence time of both dexamethasone and dibC by monitoring the time-resolved change in receptor binding to a surface-immobilized co-regulator peptide upon addition of  $>10$ -fold concentration excess of a reference compound (Figure S6). The data from all experiments are summarized in Table 2. In all instances,  $k_{\text{off}}$  is larger for GR than for MR, hence the residence time is longer in MR. This is in agree-

ment with the observations that MR requires a larger rearrangement of the H6-H7 region compared with GR (Figures 4 and 5). In addition, dexamethasone has a larger  $k_{\text{off}}$  than dibC, reflecting the fact that dibC is a bulkier ligand. Finally, while the different measurement methods result in the same pattern for both GR and MR and dexamethasone and dibC, providing confidence to the analysis, the systematically larger off rates using SMM likely reflect the temperature difference at which the experiments were conducted ( $20^{\circ}\text{C}$  for SMM and  $10^{\circ}\text{C}$  for SPR).

### Differential Selection Pressure

Studies on the evolution of GR from the ancestral corticoid receptor revealed that GR has accumulated a number of mutations on and in the proximity of helix 7 that prevents reversal of evolution (Bridgham et al., 2009). As our findings suggest that there is an intimate link to the ligand binding function, we decided to investigate the evolutionary consequences across the whole steroid receptor family. To explore this, sequence clusters for each receptor were downloaded from OrthoDB (Waterhouse et al., 2013). The sequences for each receptor were aligned using ClustalX version 2.0 (Larkin et al., 2007) and the pairwise species overlap with GR was selected for each receptor. Each residue position was then assigned a variability score based on the number of different amino acids at that position across the various species. All receptor sequences were overlaid on the GR

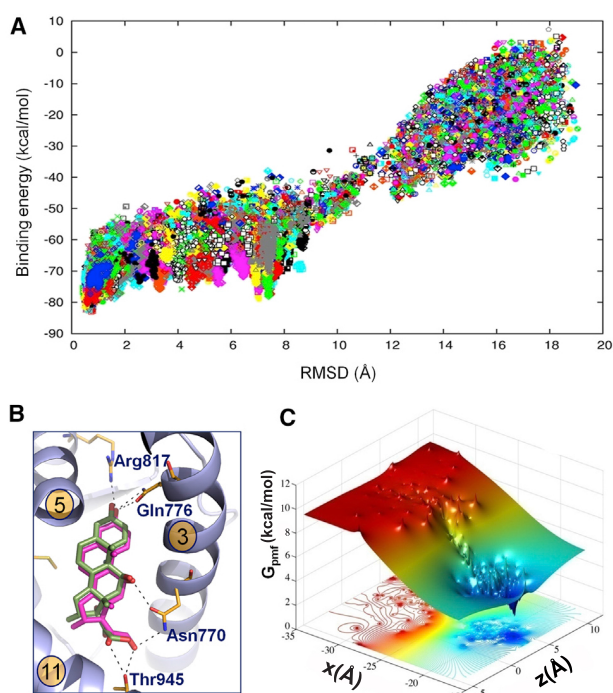


**Figure 6. Unbiased Simulation of Dexamethasone Entering the MR Binding Pocket**

(A) Each line represents the ligand heavy atom RMSD to the ligand from the crystallographic structure for a single trajectory. Two of the trajectories represented by blue and red lines enter the ligand binding pocket at steps 52 and 214, respectively.

(B) The ligand center of mass for the two trajectories that enter the binding pocket are shown as blue and red spheres. The region where the ligands enter the binding pocket is emphasized as a surface with two ligands from the simulations shown in full stick representation.





**Figure 7. Refined Ligand Binding Simulations and Estimated Binding Free Energy**

(A) The protein-ligand interaction energy plotted against the ligand heavy atom RMSD to the crystallographic structure along the 400 refinement trajectories in MR:Dexa.

(B) MR (blue) in complex with dexamethasone (magenta) overlaid on the lowest interaction energy structure after the refined exploration (green).

(C) X-Z 2D projection of the PMF obtained in the MSM analysis for the same process.

sequence using X-ray structures to define the equivalent positions. Finally, we plotted the variability score against the amino acid sequence for all receptor pairs (Figure 8). The data confirm that important structural elements of the receptors are relatively conserved. For example, the variability score for the AF-2 surface (the N-terminal end of H12, H4, and the C-terminal end of H3), which is directly involved in the protein-protein interaction transmitting the ligand activation signal, is consistently low for all receptors. However, H6-H7 exhibits a greater variability score in GR relative to all other receptors. Interestingly, GR also has a segment of higher variability near the C-terminal end of H11. This region sits directly across from the N-terminal end of H7 (Figure 1C), and it is conceivable that amino acid sequences of these regions may well co-vary with each other. Figure S7A shows the variability score for the individual amino acids in the H6-H7 region for the full set of GR species. It is clear that the high variability score of the region resides in discrete positions (primarily in residues 631, 632, 635, 638, and 640). These residues are all located on the outside of the receptor in both the GR:Dexa and GR:dibC structures (Figure S7B).

## DISCUSSION

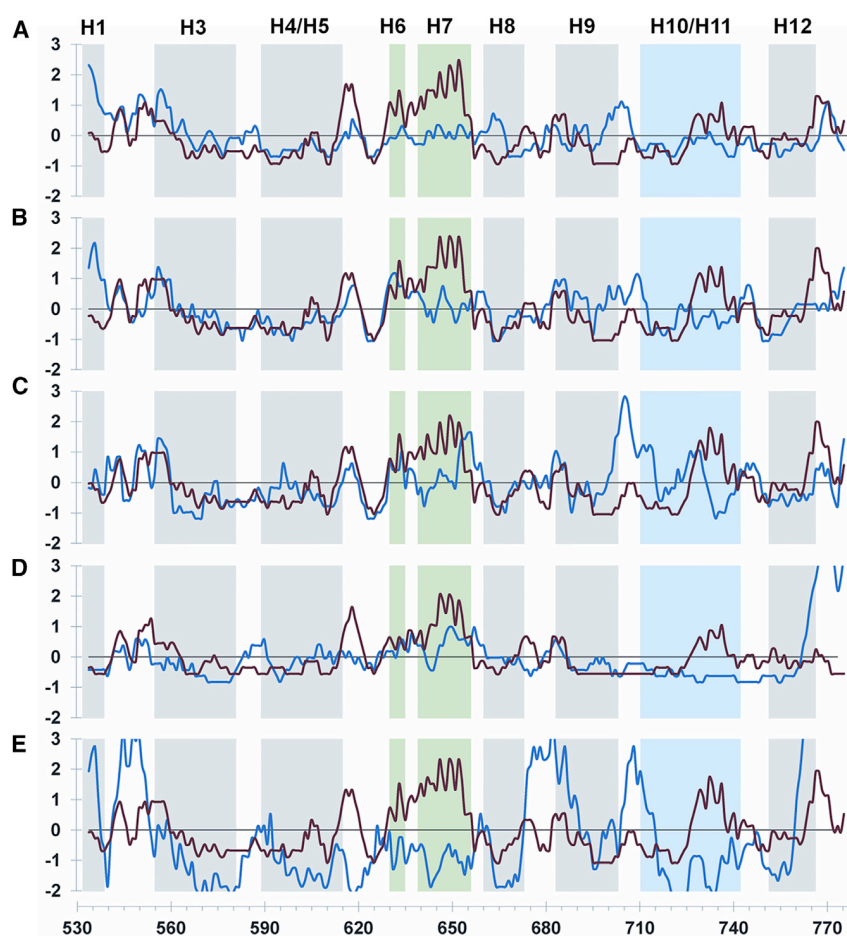
The fundamental role and mechanism of action of steroid receptors have been studied extensively, yet details of the ligand bind-

**Table 2. Measurement of  $k_{\text{off}}$  Using SPR and SMM**

| Ligand (Method)  | GR ( $\text{s}^{-1}$ ) | MR ( $\text{s}^{-1}$ ) |
|------------------|------------------------|------------------------|
| Dexa (SPR, 10°C) | 0.0034                 | 0.0011                 |
| dibC (SPR, 10°C) | 0.0010                 | <0.0001                |
| Dexa (SMM, 20°C) | 0.0070                 | 0.0025                 |
| dibC (SMM, 20°C) | 0.0029                 | 0.0012                 |

ing mechanisms have remained unclear. By comparing the structures of MR and GR in complex with dexamethasone and dibC, we confirmed the intrinsic capacity to open up the H6-H7 region. While the GR:Dexa structure adopts an open conformation compared with the MR:Dexa complex, the MR:dibC structure is able to extend the ligand binding pocket significantly and adopt a structural state akin to the GR:Dexa arrangement. Studies of ancestral corticoid receptor (AncCR), the common predecessor of MR and GR, revealed that the Ser106<sup>AncCR</sup> (corresponding to Ser843<sup>MR</sup>) to Pro637<sup>GR</sup> switch was a permissive mutation that facilitated a subsequent Leu111<sup>AncCR</sup> (corresponding to Leu848<sup>MR</sup>) to Gln642<sup>GR</sup> mutation (Bridgham et al., 2006). This is an example of conformational epistasis and has played an important role in the evolution of GR hormone selectivity (Ortlund et al., 2007). We show that GR and MR demonstrate a similar capacity to form an open conformation, and it is likely that the AncCR also exhibited the same flexibility. Hence, as GR evolved from AncCR, the Ser106<sup>AncCR</sup> to Pro637<sup>GR</sup> mutation would primarily serve to select a subset of preexisting structural states, rather than creating a completely new arrangement. The importance of conformational selection over induced fit has provided mechanistic insights for several biological systems (Changeux, 2013), and it is plausible that evolution through mutation often operates in an analogous way.

Extensive ligand binding simulations revealed that the entry and exit trajectories all pass through the H3-H7-H11 junction. As the ligands cross the receptor surface, the outward bending motion of the H6-H7 region is qualitatively similar to the observed perturbations caused by the large 17 $\alpha$  cyclohexyl substituent in the dibC complex structures, linking the observed H6-H7 plasticity to the ligand binding mechanism. Interestingly, H7 has also been shown to be important for dimerization of several nuclear receptors (Osz et al., 2012). This suggests that the two functions could be linked for these receptors, but the strength of this relationship remains to be determined. The results from the ligand binding simulations indicate that large-amplitude protein motions of helix 12, as suggested by apo and holo crystallographic nuclear hormone receptors (Moras and Gronemeyer, 1998; Yen, 2001; Brzozowski et al., 1997), are not required for ligand entry. Instead, the conformation of the LBD is likely to resemble the ligand bound agonistic conformations of the receptors during the ligand entry step (Capelli et al., 2013; Batista and Martínez, 2013). We show that small-scale vibrations combined with a structural rearrangement of H6-H7 region are enough to identify an energetically favorable pathway to allow the ligands to diffuse into the binding pocket. In contrast to other modeling studies using biased protocols, we do not observe multiple ligand entry or exit pathways (Capelli et al., 2013; Sonoda et al., 2008; Aci-Sèche et al., 2011). Finally, careful analysis of the binding energies along the entry trajectory revealed a



**Figure 8. Evolutionary Conservation of the LBD for the Steroid Receptors**

The graphs show normalized amino acid variability score for pairwise comparisons of MR (A), PR (B), AR (C), ER $\alpha$  (D), and ER $\beta$  (E) in blue versus GR in red plotted against the GR amino acid sequence. The variability score was average normalized and smoothed using a five-amino-acid sliding window. Helices 1–12 are annotated using vertical bars (green: H6–7; blue: H10–11; gray: all others). High variability scores indicate less conservation.

While the dibC complex structures show that both corticoid receptors can adopt an open conformation, they also highlight that the plasticity in the H6–H7 region is different. For MR, the challenge from a large  $17\alpha$  substituent results in a complete rearrangement of the H6–H7 structure. In contrast, GR responds with a rigid shift of the region. A closer inspection of the simulations revealed ensuing differences as MR require larger rearrangements in the gatekeeper residues for productive ligand binding and unbinding. This is in agreement with the kinetic measurements revealing that both dexamethasone and dibC exhibit longer receptor residence times in MR than in GR. However, these observations do not necessarily result in differences in ligand affinity per se, as both ligand entry and exit will be governed by the same plas-

potential peripheral binding site. While it requires further characterization, the function of such a site on the surface of the receptor could serve to capture the ligands and increase the chances for productive binding events.

It is firmly established that steroid receptors depend on a number of chaperone and co-chaperone proteins for correct folding that is capable of high-affinity hormone binding (Grad and Picard, 2007). Although the ligand entry function is likely to have evolved before the synergies with chaperone proteins, these proteins will nevertheless limit the access to the receptors and thereby form boundary conditions for any ligand entry hypothesis. Mutation and peptide competition studies suggest that Hsp90 interacts at the AF-2 surface (Ricketson et al., 2007; Fang et al., 2006). In addition, co-chaperones have been mapped to interact with regions surrounding the C-terminal end of H1 and the N-terminal end of H3 (Caamaño et al., 1998), and with the loop that connects them (Cluning et al., 2013). Neither of these areas overlap with the entry site proposed here. However, previous studies have shown that the chaperone complex promotes the ligand binding process (Grad and Picard, 2007). Interestingly, the simulations whereby we removed the co-regulator peptides resulted in greater fluctuations in both the H3–H7–H11 junction and H12. These results suggest that the presence of chaperone proteins at remote sites can allosterically influence the ligand entry process proposed here.

ticity, potentially affecting on and off rates equally. Nevertheless, it is important to note that ligand binding and unbinding are asymmetric events. While ligand binding occurs with the receptor in the chaperone complex in the cytoplasm, unbinding will likely occur in the different protein complex. As such, it is tempting to speculate that the relative stabilization of the open versus the closed conformation may differ for the two states. This could increase the apparent ligand affinity and potentially add another layer of differentiation. To resolve this, detailed structural information on the relevant protein complexes would be required.

The distinct receptor blueprints also appear to have evolutionary consequences. By comparing the amino acid sequence for different species across all steroid receptors, we found that GR exhibits a higher mutational frequency in the H6–H7 region. We propose that as GR evolved a cortisol selectivity profile, the change in the dynamic profile of the H6–H7 region, through the Ser106<sup>AncCR</sup> to Pro637<sup>GR</sup> mutation, altered the boundary conditions for the ligand entry mechanism. While for MR, residues need to be compatible with two distinct structural states during ligand entry, for GR the equivalent residues will be exposed to the solvent throughout the process. As a result the selection pressure was relaxed for specific positions in this region for GR, which explains why subsequent mutations could build.



The tremendous growth in the number of available X-ray structures from increasingly more advanced protein classes and complexes provides a plethora of snapshots of molecular mechanisms in action. However, to bridge the gap to detailed mechanistic insights and to establish evolutionary relationships, orthogonal data from biochemical experiments and *in silico* modeling are required. By combining information from several X-ray structures, extensive simulations, kinetic measurements, and bioinformatics analyses, we have uncovered the ligand binding mechanism into the occluded binding pocket of steroid hormone receptors. Ligand binding to the steroid receptors marks the first step in a chain of events that in the end triggers both broad genomic and nongenomic mechanisms. Understanding the details of ligand association and dissociation may facilitate the rational design of molecules that exploit the plasticity of the entry and exit processes to a greater extent. This could yield ligands with different modes of action, such as antagonists that block nuclear translocation or agonists with extended receptor occupancy and a prolonged pharmacological response.

## EXPERIMENTAL PROCEDURES

### Protein Expression, Purification, Crystallization, Structure Determination, and Analyses

The detailed protocols are described in the [Supplemental Experimental Procedures](#). For structure, the following protein constructs were used: GR:Dexa, GR-LBD (amino acids 500–777) N517D, F602S, C638D; GR:dibC, GR-LBD (amino acids 500–777) N517D, V571M F602S, C638D; MR:Dexa, MR-LBD (amino acids 735–984) C808S, C910S; MR:dibC, MR-LBD (amino acids 735–984) C808S, S810L, C910S. For the kinetic measurements, the following constructs were used: GR, GR-LBD (NR3C1; amino acids 529–777); MR, MR-LBD (amino acids 712–984) C808S.

### Mineralocorticoid Receptor Ligand Competition Binding Assay

A scintillation proximity-based radioligand binding assay was used to measure the ligand displacement of aldosterone to human MR-LBD. The detailed protocol is presented in the [Supplemental Experimental Procedures](#).

### PELE Simulations

#### Systems Setup

Initial coordinates for GR and MR were obtained from the crystals presented here. Three different receptor models were prepared: (1) the crystallographic structures, (2) the wild-type receptors generated by reverting the crystallographic mutations with the Schrödinger package (Schrödinger, 2013), and (3) the wild-type receptors in absence of the peptide co-factor. All structures were preprocessed with the protein preparation wizard (Madhavi Sastry et al., 2013) available in the Schrödinger package, adding hydrogen atoms and optimizing the hydrogen bond network, followed by a final visual inspection.

#### PELE Sampling

PELE combines a Monte Carlo approach with protein structure prediction methods, allowing exploration of long-timescale atomic biophysical processes (Borrelli et al., 2005; Cossins et al., 2012). Three main steps define the algorithm: (1) protein backbone and ligand perturbation, (2) specific side-chain sampling, and (3) global minimization (for more details see, for example, Kotev et al., 2015). The program uses an OPLS (Optimized Potentials for Liquid Simulations) all-atom force field with an implicit SGB (surface-generalized Born) continuum solvent model.

### Ligand Exit Simulations

From the crystallographically prepared models, the exit protocol included random ligand's translations of 0.8 Å and rotation of 0.2 radians. The backbone perturbation included the lowest six anisotropic network model modes with maximum displacements of each  $\alpha$  carbon up to 1 Å. A spawning criteria of 4 Å was used: any ligand whose center of mass is 4 Å behind the structure

with the center of mass farthest coordinates (with respect to the initial position), in any direction, will abandon its position and continue the execution with the coordinates from the leading (farthest) one. Thus, all processors search collectively, with no bias in direction, for an effective escape path. Simulations were finished after the ligand's solvent-accessible area was larger than 0.5, with typical simulation times of 10–20 CPU hr.

### Ligand Entrance Simulations

Starting from 20 conformations where the ligand is randomly distributed over the protein surface, free search simulations were performed with runs of 64 independent simulations (no spawning criteria were used) for 48 CPU hr. Ligand perturbation included equally probable translations of 3.0 Å/1.0 Å and rotation of 0.25/0.05 radians. Ligand displacement direction was randomly updated every six steps, thus ensuring that trajectories explored the entire surface. Furthermore, keeping the perturbation direction for six steps is necessary to observe entrance events in difficult cases.

### Residence Time Determination

Residence time measurements of GR/MR:dexamethasone and dibC were determined using SMM and SPR (Biacore). In brief, GR/MR was pre-equilibrated with dexamethasone/dibC. Directly after addition of budesonide/aldosterone, the rate of receptor binding to the surface-immobilized co-factor peptide, caused by the ligand-induced change in affinity, was monitored continuously over ~15 min with SMM or by consecutive injection cycles (typically six) in SPR. See the [Supplemental Information](#) for details on surface preparation and experimental procedures. The dissociation rate was determined by exponential fits to the change in binding rate as a function of time.

### Sequence Homology Analysis

Sequence clusters for each receptor were downloaded from the OrthoDB database (Waterhouse et al., 2013) by searching for the human ENS gene ID and selecting the vertebrate subset. For each receptor, sequences with a length two SDs below average length or that contained more than 100 "X" (unknown amino acids) were removed. The sequences for each receptor were aligned using ClustalX version 2.0 (Larkin et al., 2007), then further filtered to only keep sequences with an intact H6-H7 region (maximum 1 indel or "X" and  $\geq 50\%$  identity to the human H6-H7 region; sequences with large indels in H6-H7 were removed followed by realignment and re-filtering to correct for alignment errors around indels). The filtered sets were scored using custom perl scripts; for each position in the alignment, a variability score was calculated by counting the number of different types of amino acids (i.e. if a position contained 5F, 3Y, and 9L, then the score is 3). To remove bias stemming from the inclusion of sequences from different species across the various receptors, we generated subsets wherein the same species were included for pairs of GR with either of (MR, PR, AR, ER $\alpha$ , ER $\beta$ ). The paired subsets were realigned for each receptor, and the resulting alignments were analyzed and scored as previously described. Finally, the scores were normalized (variability score minus average variability score for LBD) and smoothed using a sliding window of five amino acids, and plotted against the GR protein sequence.

### Phylogenetic Analysis of the Human LBD Region

Human sequences for the studied nuclear receptors (AR, ER $\alpha$ , ER $\beta$ , GR, MR, and PR) were extracted from the aforementioned dataset. Sequences were trimmed so that only the LBD region remained, aligned using ClustalX, and manually edited based on the structure (minor adjustments). The tree was calculated using ClustalX (bootstrap 1,000 iterations) and visualized using NJplot version 2.3 (Perrière and Gouy, 1996).

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures and one movie and can be found with this article online at <http://dx.doi.org/10.1016/j.str.2015.09.012>.

## AUTHOR CONTRIBUTIONS

K.E., A.C.H., M.L., and V.G. designed the research. U.K. performed binding experiments. S.B., C.K., T.J.J., A.C., and E.N. expressed and purified protein.

A.A. and L.W. crystallized the proteins. K.E. performed the structural determination and analyzed the data. C.G. performed the principal component analysis. A.H. performed the exit and entry simulations. D.L. and R.T. performed MSM analysis. A.G., T.K., and S.G. performed the kinetic experiments. M.K.B. carried out the bioinformatics analysis.

## ACKNOWLEDGMENTS

This study was supported by European Research Council 2014-PoC-eDrug to V.G. and by the SEV-2011-00067 grant of the Severo Ochoa Program. We would like to acknowledge our AstraZeneca colleagues J. Hartleib, R. Unwin, and R. Knöll for helpful discussions. We also thank N. Blomberg (ELIXIR) and R. Neutze (University of Gothenburg) for careful reading of the manuscript.

Received: June 25, 2015

Revised: September 3, 2015

Accepted: September 4, 2015

Published: October 22, 2015

## REFERENCES

- Aci-Sèche, S., Genest, M., and Garnier, N. (2011). Ligand entry pathways in the ligand binding domain of PPAR $\gamma$  receptor. *FEBS Lett.* **585**, 2599–2603.
- Alexander, S.P.H., Benson, H.E., Faccenda, E., Pawson, A.J., Sharman, J.L., Spedding, M., Peters, J.A., and Harmar, A.J.; CGTP Collaborators (2013). The concise guide to pharmacology 2013/14: nuclear hormone receptors. *Br. J. Pharmacol.* **170**, 1652–1675.
- Andrieu, T., Mani, O., Goepfert, C., Bertolini, R., Guettinger, A., Setoud, R., Uh, K.Y., Baker, M.E., Frey, F.J., and Frey, B.M. (2015). Detection and functional portrayal of a novel class of dihydrotestosterone derived selective progesterone receptor modulators (SPRM). *J. Steroid Biochem. Mol. Biol.* **147**, 111–123.
- Batista, M.R., and Martínez, L. (2013). Dynamics of nuclear receptor helix-12 switch of transcription activation by modeling time-resolved fluorescence anisotropy decays. *Biophys. J.* **105**, 1670–1680.
- Bertocchio, J., Warnock, D.G., and Jaisser, F. (2011). Mineralocorticoid receptor activation and blockade: an emerging paradigm in chronic kidney disease. *Kidney Int.* **10**, 1051–1060.
- Bhabha, G., Ekiert, D.C., Jennewein, M., Zmasek, C.M., Tuttle, L.M., Kroon, G., Dyson, H.J., Godzik, A., Wilson, I.A., and Wright, P.E. (2013). Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat. Struct. Mol. Biol.* **11**, 1243–1249.
- Bledsoe, R.K., Montana, V.G., Stanley, T.B., Delves, C.J., Apolito, C.J., McKee, D.D., Consler, T.G., Parks, D.J., Stewart, E.L., Willson, T.M., et al. (2002). Crystal structure of the glucocorticoid receptor ligand binding domain reveals a novel mode of receptor dimerization and coactivator recognition. *Cell* **110**, 93–105.
- Borrelli, K.W., Vitalis, A., Alcantara, R., and Guallar, V. (2005). Protein energy landscape exploration. A novel Monte Carlo technique. *J. Chem. Theor. Comput.* **6**, 1304–1311.
- Bridgham, J.T., Carroll, S.M., and Thornton, J.W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. *Science* **312**, 97–101.
- Bridgham, J.T., Ortlund, E.A., and Thornton, J.W. (2009). An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515–520.
- Brzozowski, A.M., Pike, A.C., Dauter, Z., Hubbard, R.E., Bonn, T., Engström, O., Ohman, L., Greene, G.L., Gustafsson, J.A., and Carlquist, M. (1997). Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **389**, 753–758.
- Buch, I., Giorgino, T., and De Fabritiis, G. (2011). Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **108**, 10184–10189.
- Caamaño, C.A., Morano, M.I., Dalman, F.C., Pratt, W.B., and Akil, H. (1998). A conserved proline in the hsp90 binding region of the glucocorticoid receptor is required for hsp90 heterocomplex stabilization and receptor signaling. *J. Biol. Chem.* **273**, 20473–20480.
- Capelli, A.M., Bruno, A., Guadix, A.E., and Costantino, G. (2013). Unbinding pathways from the glucocorticoid receptor shed light on the reduced sensitivity of glucocorticoid ligands to a naturally occurring, clinically relevant mutant receptor. *J. Med. Chem.* **56**, 7003–7014.
- Changeux, J.P. (2013). 50 years of allosteric interactions: the twists and turns of the models. *Nat. Rev. Mol. Cell Biol.* **14**, 819–829.
- Cluning, C., Ward, B.K., Rea, S.L., Arulpragasam, A., Fuller, P.J., and Ratajczak, T. (2013). The helix 1-3 loop in the glucocorticoid receptor LBD is a regulatory element for FKBP cochaperones. *Mol. Endocrinol.* **27**, 1020–1035.
- Cole, T.J. (2006). Glucocorticoid action and the development of selective glucocorticoid receptor ligands. *Biotechnol. Annu. Rev.* **12**, 269–300.
- Cossins, P.B., Hosseini, A., and Guallar, V. (2012). Exploration of protein conformational change with PELE and meta-dynamics. *J. Chem. Theor. Comput.* **8**, 959–965.
- Dror, R.O., Pan, A.C., Arlow, D.H., Borhani, D.W., Maragakis, P., Shan, Y.B., Xu, H.F., and Shaw, D.E. (2011). Pathway and mechanism of drug binding to G-protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **108**, 13118–13123.
- Evans, R.M. (1988). The steroid and thyroid hormone receptor superfamily. *Science* **240**, 889–895.
- Fagart, J., Huyet, J., Pinon, G.M., Rochel, M., Mayer, C., and Rafestin-Oblin, M.E. (2005). Crystal structure of a mutant mineralocorticoid receptor responsible for hypertension. *Nat. Struct. Mol. Biol.* **12**, 554–555.
- Fang, L., Ricketson, D., Getubig, L., and Darimont, B. (2006). Unliganded and hormone-bound glucocorticoid receptors interact with distinct hydrophobic sites in the Hsp90 C-terminal domain. *Proc. Natl. Acad. Sci. USA* **103**, 18487–18492.
- Grad, I., and Picard, D. (2007). The glucocorticoid responses are shaped by molecular chaperones. *Mol. Cell Endocrinol.* **275**, 2–12.
- Gravez, B., Tarjus, A., and Jaisser, F. (2013). Mineralocorticoid receptor and cardiac arrhythmia. *Clin. Exp. Pharmacol. Physiol.* **40**, 910–915.
- Gronemeyer, H., Gustafsson, J.A., and Laudet, V. (2004). Principles for modulation of the nuclear receptor superfamily. *Nat. Rev. Drug Discov.* **3**, 950–964.
- Gunnarsson, A., Snijder, A., Hicks, J., Gunnarsson, J., Höök, F., and Geschwindner, S. (2015). Drug discovery at the single molecule level: inhibition-in-solution assay of membrane-reconstituted  $\beta$ -secretase using single-molecule imaging. *Anal. Chem.* **87**, 4100–4103.
- Hasui, T., Matsunaga, N., Ora, T., Ohyabu, N., Nishigaki, N., Imura, Y., Igata, Y., Matsui, H., Motoyaji, T., Tanaka, T., et al. (2011). Identification of benzoxazin-3-one derivatives as novel, potent, and selective nonsteroidal mineralocorticoid receptor antagonists. *J. Med. Chem.* **54**, 8616–8631.
- Hughes, T.S., Chalmers, M.J., Novick, S., Kuruvilla, D.S., Chang, M.R., Kamenecak, T.M., Rance, M., Johnson, B.A., Burris, T.P., Griffin, P.R., et al. (2012). Ligand and receptor dynamics contribute to the mechanism of graded PPAR $\gamma$  agonism. *Structure* **20**, 139–150.
- Kohn, J.A., Deshpande, K., and Ortlund, E.A. (2012). Deciphering modern glucocorticoid cross-pharmacology using ancestral corticosteroid receptors. *J. Biol. Chem.* **287**, 16267–16275.
- Kotev, M., Lecina, D., Tarragó, T., Giralt, E., and Guallar, V. (2015). Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques. *Biophys. J.* **108**, 116–125.
- Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948.
- Li, Y., Suino, K., Daugherty, J., and Xu, H.E. (2005). Structural and biochemical mechanisms for the specificity of hormone binding and coactivator assembly by mineralocorticoid receptor. *Mol. Cell* **19**, 367–380.
- Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27**, 221–234.
- Mangelsdorf, D.J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P., and Evans, R.M. (1995). The nuclear receptor superfamily: the second decade. *Cell* **83**, 835–839.

- Matias, P.M., Donner, P., Coelho, R., Thomaz, M., Peixoto, C., Macedo, S., Otto, N., Joschko, S., Scholz, P., Wegg, A., et al. (2000). Structural evidence for ligand specificity in the binding domain of the human androgen receptor. Implications for pathogenic gene mutations. *J. Biol. Chem.* *275*, 26164–26171.
- Moras, D., and Gronemeyer, H. (1998). The nuclear receptor ligand-binding domain: structure and function. *Curr. Opin. Cell Biol.* *10*, 384–391.
- Nettles, K.W., Bruning, J.B., Gil, G., O'Neill, E.E., Nowak, J., Guo, Y., Kim, Y., DeSombre, E.R., Dilis, R., Hanson, R.N., et al. (2007). Structural plasticity in the oestrogen receptor ligand-binding domain. *EMBO Rep.* *8*, 563–568.
- Ortlund, E.A., Bridgham, J.T., Redinbo, M.R., and Thornton, J.W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* *317*, 1544–1548.
- Osz, J., Brélivet, Y., Peluso-Ittis, C., Cura, V., Eiler, S., Ruff, M., Bourguet, W., Rochel, N., and Moras, D. (2012). Structural basis for a molecular allosteric control mechanism of cofactor binding to nuclear receptors. *Proc. Natl. Acad. Sci. USA* *109*, E588–E594.
- Perrière, G., and Gouy, M. (1996). WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* *78*, 364–369.
- Pfaff, S.J., and Fletterick, R.J. (2010). Hormone binding and co-regulator binding to the glucocorticoid receptor are allosterically coupled. *J. Biol. Chem.* *285*, 15256–15267.
- Ricketson, D., Hostick, U., Fang, L., Yamamoto, K.R., and Darimont, B.D. (2007). A conformational switch in the ligand-binding domain regulates the dependence of the glucocorticoid receptor on Hsp90. *Mol. Biol.* *368*, 729–741.
- Rupprecht, R., Reul, J.M.H.M., van Steensel, B., Spengler, D., Söder, M., Berning, B., Holsboer, F., and Damm, K. (1993). Pharmacological and functional characterization of human mineralocorticoid and glucocorticoid receptor ligands. *Eur. J. Pharmacol.* *247*, 145–154.
- Schrödinger. (2013). Release 2013-1: MacroModel, Version 10.0 (Schrödinger, LLC).
- Shelley, M., Bennett, C., Nathan, D., and Sartor, O. (2008). Non-steroidal anti-androgen use as part of combined androgen blockade therapy for metastatic or locally advanced prostate cancer: a review of the evidence on efficacy and toxicity. In *Cancer metastasis-biology and treatment*, R.J. Ablin and M.D. Mason, eds. (Springer Science and Business Media), pp. 283–307.
- Shibata, S., Rinehart, J., Zhang, J., Moeckel, G., Castañeda-Bueno, M., Stiegler, A.L., Boggon, T.J., Gamba, G., and Lifton, R.P. (2013). Mineralocorticoid receptor phosphorylation regulates ligand binding and renal response to volume depletion and hyperkalemia. *Cell Metab.* *18*, 660–671.
- Sitruk-Ware, R., and Nath, A. (2010). The use of newer progestins for contraception. *Contraception* *82*, 410–417.
- Soisson, S.M., Parthasarathy, G., Adams, A.D., Sahoo, S., Sitlani, A., Sparrow, C., Cui, J., and Becker, J.W. (2008). Identification of a potent synthetic FXR agonist with an unexpected mode of binding and activation. *Proc. Natl. Acad. Sci. USA* *105*, 5337–5342.
- Sonoda, M.T., Martínez, L., Webb, P., Skaf, M.S., and Polikarpov, I. (2008). Ligand dissociation from estrogen receptor is mediated by receptor dimerization: evidence from molecular dynamics simulations. *Mol. Endocrinol.* *22*, 1565–1578.
- Takahashi, R., Gil, V.A., and Guallar, V. (2014). Monte Carlo free ligand diffusion with Markov state model analysis and absolute binding free energy calculations. *J. Chem. Theor. Comput.* *10*, 282–288.
- Waterhouse, R.M., Tegenfeldt, F., Li, J., Zdobnov, E.M., and Kriventseva, E.V. (2013). OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.* *41*, D358–D365.
- Williams, S.P., and Sigler, P.B. (1998). Atomic structure of progesterone complexed with its receptor. *Nature* *393*, 392–396.
- Yen, P.M. (2001). Physiological and molecular basis of thyroid hormone action. *Physiol. Rev.* *81*, 1097–1142.

## **Publication 3 - The unravelling of the complex pattern of tyrosinase inhibition**

**Authors:** Batel Deri, Margarita Kanteev, Mor Goldfeder, [Daniel Lecina](#), Victor Guallar, Noam Adir, Ayelet Fishman

**Journal:** Scientific Reports 6, 34993 (2016)

### **Summary:**


In this study we show for the first time the inhibition mechanisms of kojic acid (KA) and hydroquinone (HQ) in a tyrosinase, combining experimental and computational techniques. Experimental techniques consisted of crystallization, binding constant analysis and kinetic experiments, whereas *in-silico* tools involved running sets of unbiased all-atom simulations for both inhibitors and a posterior MSM analysis. Our findings suggest that KA acts as a mixed inhibitor; when it is bound in the active site, it is not accessible to substrate molecules, but when it is in the peripheral binding site, it restricts the entrance and efflux, and impedes reaching the maximum catalytic velocity. On the contrary, HQ can act both as a substrate and an inhibitor, suggested by its binding heterogeneity.

### **Author contribution:**

On one side, my tasks were running and analyzing all the simulations by means of MSM, and on the other side, writing the manuscript.



# SCIENTIFIC REPORTS



OPEN

## The unravelling of the complex pattern of tyrosinase inhibition

Batel Deri<sup>1,\*</sup>, Margarita Kanteev<sup>1,\*</sup>, Mor Goldfeder<sup>1</sup>, Daniel Lecina<sup>2</sup>, Victor Guallar<sup>2,3</sup>, Noam Adir<sup>4</sup> & Ayelet Fishman<sup>1</sup>

Received: 27 June 2016

Accepted: 22 September 2016

Published: 11 October 2016

Tyrosinases are responsible for melanin formation in all life domains. Tyrosinase inhibitors are used for the prevention of severe skin diseases, in skin-whitening creams and to avoid fruit browning, however continued use of many such inhibitors is considered unsafe. In this study we provide conclusive evidence of the inhibition mechanism of two well studied tyrosinase inhibitors, KA (kojic acid) and HQ (hydroquinone), which are extensively used in hyperpigmentation treatment. KA is reported in the literature with contradicting inhibition mechanisms, while HQ is described as both a tyrosinase inhibitor and a substrate. By visualization of KA and HQ in the active site of TyrBm crystals, together with molecular modeling, binding constant analysis and kinetic experiments, we have elucidated their mechanisms of inhibition, which was ambiguous for both inhibitors. We confirm that while KA acts as a mixed inhibitor, HQ can act both as a TyrBm substrate and as an inhibitor.

Tyrosinases belong to the type 3 copper-containing protein family together with hemocyanins that serve as oxygen carriers<sup>1,2</sup>, and catechol oxidases that are strict diphenolases<sup>3,4</sup>. The two copper ions in the conserved active site, CuA and CuB, are coordinated by six histidine residues<sup>5–7</sup>. Tyrosinases hydroxylate monophenols to form *ortho*-diphenols (monophenolase activity) and subsequently oxidize the *o*-diphenols to *o*-quinones (diphenolase activity). Melanin is formed rapidly by the spontaneous polymerization of the quinones<sup>5,8</sup>. Monophenols can react only with the *oxy* state of tyrosinase, which represents about 15% of the enzyme molecules in solution<sup>9</sup>. In the presence of *o*-diphenols such as L-dopa (L-3,4-dihydroxyphenylalanine), both the *oxy* and *met* forms react enabling the production of *o*-quinones<sup>4,9</sup>.

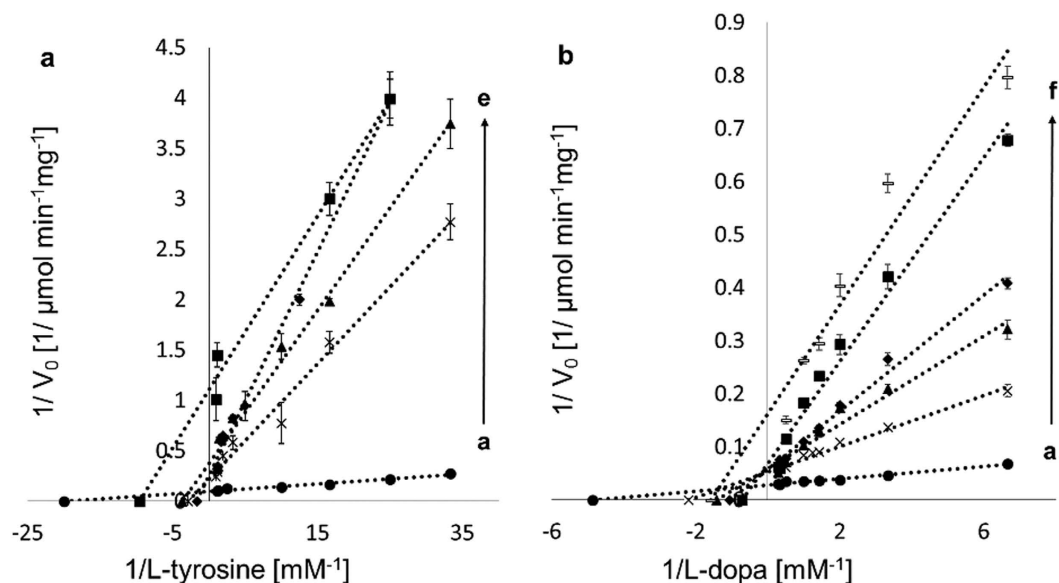
Disorder in melanin formation has been found to cause a variety of skin diseases in humans such as hyperpigmentation, lentigo, vitiligo and skin cancer<sup>10</sup>. Furthermore, appearance of brown pigments in fruits and vegetables due to tyrosinase activity is a leading cause for postharvest losses<sup>9</sup>. Therefore, tyrosinase inhibitors are highly warranted by the pharmaceutical, cosmetics and food industries<sup>11–15</sup>.

Kojic acid (KA), a fungal metabolite, is the most widely used skin-whitening agent with possible side effects being dermatitis, sensitization and erythema<sup>9,16</sup>. Animal experiments suggested possible tumor promotion and weak carcinogenicity, and thus concentrations of 1% are recommended for safe human use<sup>17</sup>. Numerous contradicting mechanisms are described in the literature for KA as either a competitive or mixed inhibitor for mushroom tyrosinase<sup>18–20</sup>, possibly by chelating copper in the active site<sup>18,20–22</sup>. Previously, KA was found bound at the entrance to the active site of TyrBm (tyrosinase from *Bacillus megaterium*), suggesting one significant intermediate binding site. However, the full mechanism of KA inhibition still remains unclear<sup>22</sup>.

Hydroquinone (HQ), another well-studied whitening agent, has been used clinically in leading cosmetic hyperpigmentation treatment<sup>23</sup>, however, it was also found to cause serious problems by generating reactive oxygen species leading to oxidative damage of lipids and permanent loss of melanocytes. Subsequently, HQ has been banned for the general use by the European Committee and can be prescribed by dermatologists only<sup>13,16</sup>. Previous studies suggested that HQ is a competitive inhibitor of tyrosinase<sup>24,25</sup>, while others demonstrated the potency of HQ as a tyrosinase substrate<sup>26,27</sup>. Garcia-Canovas and co-workers suggested that the enzymatic activity is not evident since HQ is not able to transform *met*-tyrosinase to the *oxy*-form<sup>26</sup>, and that the transformation may be substantial by addition of an *o*-diphenol or H<sub>2</sub>O<sub>2</sub><sup>26,28</sup>. However, to date, no structural data is available in order to elucidate the orientation of HQ in the active site of tyrosinase.

<sup>1</sup>Department of Biotechnology and Food Engineering, Technion-Israel Institute of Technology, Haifa, 3200003, Israel.

<sup>2</sup>Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Jordi Girona 29, 08034 Barcelona, Spain. <sup>3</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, 08010 Barcelona, Spain. <sup>4</sup>Schulich Faculty of Chemistry, Technion-Israel Institute of Technology, Haifa, 3200003, Israel. \*These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.F. (email: afishman@tx.technion.ac.il)



**Figure 1.** Lineweaver–Burk plots for the inhibition of TyrBm by KA. (a) L-tyrosine (0.03–1.4 mM) in the presence of KA concentrations (mM): (a ●) 0, (b ×) 0.025, (c ▲) 0.05, (d ◆) 0.075, (e ■) 0.1 and (b) L-dopa (0.15–2.0 mM) in the presence of KA concentrations (mM): (a ●) 0, (b ×) 0.025, (c ▲) 0.04, (d ◆) 0.05, (e ■) 0.075, (f -) 0.1. All measurements were performed in heptaplates.

|    | Substrate  | $K_m$ (mM)       | $V_{max}$ ( $\mu\text{mol min}^{-1} \text{mg}^{-1}$ ) | $k_{cat}$ ( $\text{s}^{-1}$ ) | $IC_{50}$ ( $\mu\text{M}$ ) | $K_i$ ( $\mu\text{M}$ ) | $K_{IS}$ ( $\mu\text{M}$ ) | Inhibition mode |
|----|------------|------------------|---|-------------------------------|-----------------------------|-------------------------|----------------------------|-----------------|
| KA | L-tyrosine | $0.04 \pm 0.007$ | $9.7 \pm 0.4$   | 5.7                           | $26.8 \pm 0.8$              | $1.1 \pm 0.3$           | $61 \pm 20$                | mixed           |
|    | L-dopa     | $0.18 \pm 0.03$  | $34 \pm 1$  | 20.1                          | $52 \pm 3$                  | $3.5 \pm 0.6$           | $150 \pm 12$               | mixed           |
| HQ | L-tyrosine | $0.07 \pm 0.01$  | $9.0 \pm 0.5$   | 5.3                           | $32 \pm 2$                  | $40 \pm 10$             | —                          | competitive     |

**Table 1.** Kinetic and inhibition constants of TyrBm by KA and HQ. Data was extracted from Figs 1, 2 and Supplementary Fig. S1. Each value represents the mean  $\pm$  SD of five independent experiments.

Most mechanistic studies on tyrosinase inhibitors use KA or HQ as the comparative benchmark compound. Therefore, in depth analysis of their mechanism and inhibition mode are crucial for further development of potent inhibitors.

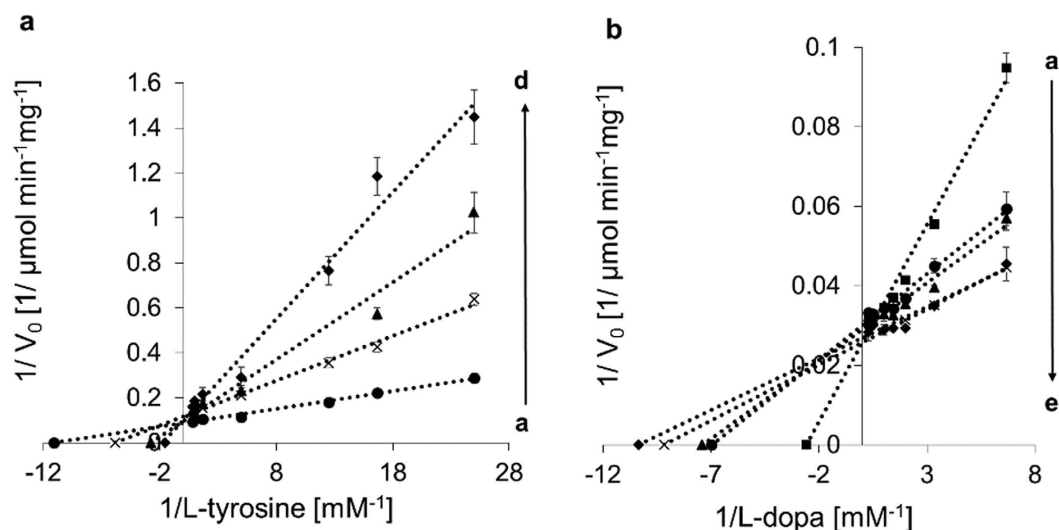
In this study, we elucidate the inhibition mechanism of these inhibitors by crystal structure determination of TyrBm with bound KA and HQ in the active site, along with biochemical characterizations, binding constants determination and molecular modeling.

## Results

**Inhibition mode of TyrBm activity.** The most widely used and effective tyrosinase inhibitors, HQ and KA<sup>9,29</sup>, were tested for their inhibitory effect on TyrBm. Overall, our results clearly show that KA and HQ have different inhibition modes on TyrBm monophenolase (L-tyrosine) and diphenolase (L-dopa) activities. While KA displays a mixed inhibition mode on both activities (Fig. 1 and Table 1), HQ is a competitive inhibitor of monophenols, and shows no inhibition of diphenols (Fig. 2 and Table 1), in contrast to previous reports that define HQ as a competitive inhibitor for both activities<sup>24,25</sup>. Our kinetic study showed no inhibition by HQ in the presence of L-dopa since with rising concentrations of HQ,  $K_m$  values decreased (Fig. 2b and Supplementary Fig. S1a).

The  $IC_{50}$  values representing inhibitor concentrations in which TyrBm activity was reduced by 50% were obtained from dose-response curves (Supplementary Fig. S2). The  $IC_{50}$  values for KA and HQ on the monophenol were 26.8 and 32.0  $\mu\text{M}$  respectively, while the  $IC_{50}$  value for KA on the diphenol was 52  $\mu\text{M}$  (Table 1). Similar results were obtained in previous reports for mushroom tyrosinase, with values for KA inhibition on monophenols and diphenols of 5.7 and 30.1  $\mu\text{M}$ , respectively<sup>30,31</sup>, and the value for HQ inhibition on monophenols of 33.5  $\mu\text{M}$ <sup>32</sup>. Since we observed that HQ does not inhibit tyrosinase in the presence of L-dopa, the value of  $IC_{50}$  was not determined. García-Canovas and co-workers suggested that HQ is a tyrosinase substrate and not an inhibitor although activity is not evident under conventional conditions since HQ cannot transform *met*-tyrosinase into *oxy*-tyrosinase<sup>26</sup>. However, in the presence of an *o*-diphenol (e.g. L-dopa) or  $\text{H}_2\text{O}_2$ , *oxy*-tyrosinase is generated and HQ becomes a substrate which is hydroxylated to 2-hydroxyhydroquinone and subsequently to 2-hydroxy-*p*-benzoquinone (HPB) that can be measured spectrophotometrically<sup>26</sup>. Our spectrophotometric measurements confirm their results, since with the addition of  $\text{H}_2\text{O}_2$  or L-dopa, product formation by TyrBm increased in the presence of HQ (Supplementary Fig. S1). Furthermore, the same trend was observed





**Figure 2. Lineweaver–Burk plots for the inhibition of TyrBm by HQ.** (a) L-tyrosine (0.03–1.4 mM) in the presence of HQ concentrations (mM): (a ●) 0, (b ×) 0.025, (c ▲) 0.075, (d ◆) 0.5 and (b) L-dopa (0.15–2.0 mM) in the presence of HQ concentrations (mM): (a ■) 0, (b ●) 0.025, (c ▲) 0.05, (d ×) 0.075, (e ◆) 0.1. All measurements were performed in heptaplicates.

| Substrate  | $K_m$ (mM)        | $V_{max}$ ( $\mu\text{mole min}^{-1} \text{mg}^{-1}$ ) | $k_{cat}$ ( $\text{s}^{-1}$ ) | $k_{cat}/K_m$ ( $\text{s}^{-1} \text{mM}^{-1}$ ) |
|------------|-------------------|--|-------------------------------|--|
| L-tyrosine | $0.082 \pm 0.006$ | $3.62 \pm 0.06$  | 2.1                           | 25.6   |
| L-dopa     | $0.24 \pm 0.02$   | $30.3 \pm 0.6$   | 17.8                          | 74.2   |
| HQ         | $0.27 \pm 0.05$   | $19 \pm 1$   | 11.3                          | 41.9   |

**Table 2. Kinetic constants of TyrBm on its natural substrates and HQ.** Data was extracted from Figs 1, 2 and Supplementary Fig. S1. Each value represents the mean  $\pm$  SD of five independent experiments.

by direct measurement of TyrBm activity through recording of the oxygen consumption during the reaction (Supplementary Fig. S3). The activity of TyrBm in the presence of L-dopa and HQ was 8% higher than with L-dopa alone for oxygen consumption measurements, and 17% higher as determined by absorbance readings.

In addition, when an activity test was performed for several hours with HQ as a sole substrate in comparison to a control without enzyme, low activity was visually observed, even without the addition of  $\text{H}_2\text{O}_2$  or L-dopa, and a light brown color was detected after 3 hours of incubation (the control remained colorless). Moreover, the low inhibitory effect of HQ was also evident in another kinetic study with mushroom tyrosinase that exhibited an  $\text{IC}_{50}$  value 80-fold higher when L-dopa was used, in comparison to L-tyrosine<sup>24</sup>. Since HQ requires a reducing agent in order to become a TyrBm substrate, the kinetic parameters of the monophenolase activity were determined in the presence of  $\text{H}_2\text{O}_2$ . The  $K_m$  and  $V_{max}$  values were 0.27 mM and  $19 \mu\text{mole min}^{-1} \text{mg}^{-1}$ , respectively, which are similar to the kinetic parameters of L-dopa (Table 2). A similar  $K_m$  value of 0.25 mM was reported by Garcia-Canovas and co-workers for *Agaricus bisporus* tyrosinase<sup>26</sup>.

**Effects of KA and HQ on the kinetic parameters of TyrBm.** The kinetic constants of TyrBm monophenolase and diphenolase activities were determined for L-tyrosine and L-dopa (Table 2). The values obtained in the presence of KA and HQ were calculated from Lineweaver–Burk plots (Figs 1 and 2; Table 1). With rising concentrations of KA, the  $K_m$  values of the monophenolase activity increased and the  $V_{max}$  values decreased, an indication of a mixed mode of inhibition, with an inhibition constant  $K_i$  of 1.1  $\mu\text{M}$  and  $K_{IS}$  of 61  $\mu\text{M}$ . The apparent  $K_m$  and  $V_{max}$  were 0.04 mM and  $9.7 \mu\text{mole min}^{-1} \text{mg}^{-1}$ , respectively (Table 1). When increasing concentrations of KA were added in the presence of L-dopa as the substrate, a similar mode of mixed inhibition was observed, with  $K_i$  of 3.5  $\mu\text{M}$  and  $K_{IS}$  of 150  $\mu\text{M}$ , similar to previous studies that reported  $K_i$  values of 3.4, 5 and 4.7  $\mu\text{M}$  for mushroom tyrosinase<sup>18,19,33</sup>. The apparent  $K_m$  and  $V_{max}$  of the diphenolase reaction were 0.18 mM and  $34 \mu\text{mole min}^{-1} \text{mg}^{-1}$ , respectively (Table 1). The mixed inhibition mode implies that KA binding is not limited to the active site. In a previous study we have already experimentally demonstrated that a peripheral KA binding site exists in TyrBm<sup>22</sup>.

Increasing concentrations of HQ in the presence of L-tyrosine, resulted in an increase in the  $K_m$  value while the  $V_{max}$  remained constant, an indication of a competitive inhibition mode on the monophenolase activity, with a  $K_i$  of 40  $\mu\text{M}$  and apparent  $K_m$  and  $V_{max}$  values of 0.07 mM and  $9.0 \mu\text{mole min}^{-1} \text{mg}^{-1}$ , respectively (Fig. 2a and Table 1). A similar inhibition mechanism was also reported by Chawla *et al.*, with  $K_i$  of 83  $\mu\text{M}$  for mushroom tyrosinase<sup>25</sup>.



| TyrBm ligand | $K_D$ ( $\mu\text{M}$ ) |
|--------------|-------------------------|
| KA           | $377 \pm 4$             |
| HQ           | $9 \pm 1$               |
| L-tyrosine   | $0.10 \pm 0.03$         |

**Table 3. Dissociation constants of TyrBm-ligand complexes.** Each value represents the mean  $\pm$  SD.

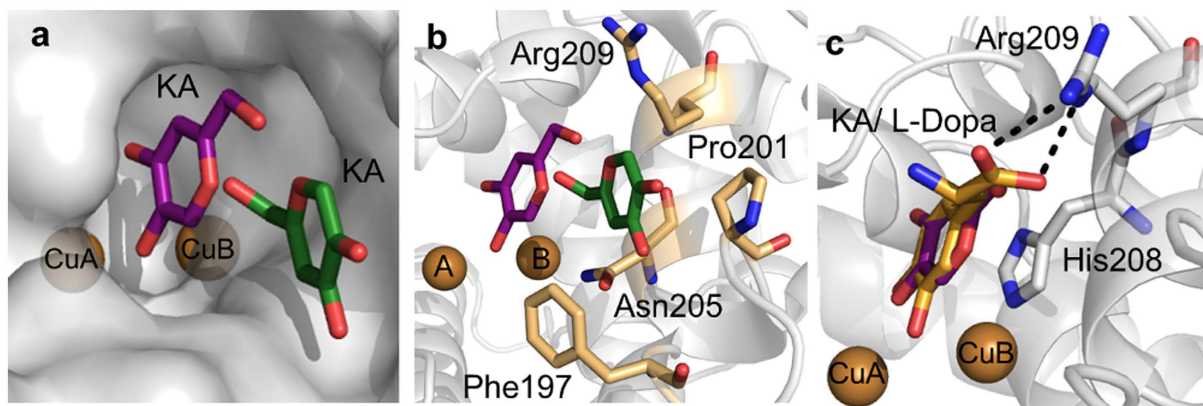
**Binding affinity of L-tyrosine vs. inhibitors.** In order to obtain a clearer understanding of the fashion by which substrates and inhibitors bind to tyrosinase, we have determined dissociation constants ( $K_D$ ) between TyrBm and its substrates or inhibitors using MicroScale Thermophoresis<sup>34</sup>. Surprisingly, such  $K_D$  values had not been previously measured. These experiments were performed by titrating fluorescently-labeled TyrBm with increasing concentrations of KA, HQ or L-tyrosine as the unlabeled ligands. According to the thermophoretic data points obtained with increasing concentrations of the ligands, the dissociation constants were evaluated. The  $K_D$  values of the TyrBm-KA, TyrBm-HQ and TyrBm-L-tyrosine interactions were determined as 377, 9 and 0.1  $\mu\text{M}$ , respectively (Table 3 and Supplementary Fig. S4). According to these results, L-tyrosine, the natural substrate, showed the highest affinity to TyrBm in comparison with KA and HQ. While HQ showed a dissociation constant 90-fold higher than L-tyrosine, KA exhibited a value of nearly 4000-fold higher than the natural substrate.

**Structure of TyrBm with KA in the active site.** In addition to the peripheral binding site of KA (PDB 3NQ1), we have recently determined the structures of TyrBm with L-tyrosine, L-dopa and the substrate analog *p*-tyrosol<sup>35</sup>, all found within the active site. We present here the crystal structure of TyrBm with KA bound in the active site at 2.6 Å resolution (Fig. 3, Supplementary Fig. S5 and Table 4). The possible movement of KA within the active site can be envisioned from Fig. 3a in which KA is shown in two positions: the peripheral site we reported earlier<sup>22</sup>, and in the active site. At the entrance to the active site, KA is stabilized by interactions with Phe197, Pro201, Asn205, and Arg209 (Fig. 3b)<sup>22</sup>. In the active site, KA is stabilized by  $\pi$ - $\pi$  interactions with His208 that coordinates CuB, similar to tyrosinase substrates (Fig. 3c), as presented by Goldfeder *et al.* and suggested in other studies<sup>35–37</sup>. The hydroxyl group of KA is oriented towards CuA with a distance of 3.3 Å, while the distance of the carbonyl group to CuA is 5.5 Å. These results are supported by a recent docking study of Lima *et al.*<sup>18</sup>, and contradict a previously proposed inhibition mechanism of KA by copper chelation<sup>20–22</sup>.

***In silico* simulations of KA and HQ in the active site.** TyrBm structure with KA at the entrance to the active site was used as an initial model to run an extensive non-biased ligand migration simulations with PELE (Protein Energy Landscape Exploration) in a constrained sphere of 20 Å (from the initial ligand center of mass)<sup>38</sup>. By means of 128 processors and 24 hours, ~200,000 different ligand conformations were obtained that allowed to evaluate the absolute binding free energy ( $\Delta G$ ) by Markov State Models (MSM) analysis<sup>39</sup>. Briefly, MSM first involves clustering all conformations (a total number of 100 clusters were used) in metastable states and building the transition matrix between them. The obtained clusters overlap mostly with the two positions of KA, at the peripheral site and in the active site (Fig. 4). Integration of these cluster centers (with respect to the bulk solvent) allowed determining the binding free energy for the active site structure of  $-5.5$  kcal/mol, whereas the surface bound complex was only of  $-1.4$  kcal/mol. Therefore, the transition from the surface bound complex to the active site is exothermic and likely to occur.

An analogous simulation was also performed for HQ. In contrast to KA, HQ showed a significant larger mobility in the active site, where multiple orientations are frequently visited. This is clearly seen when analyzing the metastable states (after MSM clustering) accessible within 1 kcal/mol from the best-bound minima (Supplementary Fig. S6). While KA presents mainly two orientations (that occupy similar volume), HQ adopts multiple orientations exploring a larger area of the active site. Interestingly, for HQ we found structures (within the lowest 1 kcal/mol) involving the peripheral site, which for KA is about 4.1 kcal/mol above the best-bound minima (Fig. 4).

**Structure of TyrBm with HQ in the active site.** The kinetic measurements with HQ indicated that it is a poor substrate of TyrBm under natural conditions, and a good substrate in conditions favoring *oxy*-tyrosinase. In order to trap HQ in the active site of TyrBm, mature crystals were soaked with zinc instead of copper ions to prevent enzymatic activity<sup>22,35,40</sup>. We have obtained two different structures of TyrBm with HQ bound in the active site (Supplementary Figs S7 and S8) at 2.2 Å resolution (Table 4). The HQ hydroxyl group is oriented towards ZnA, and its benzyl ring is stabilized through hydrophobic interactions with His208, similar to tyrosine substrates (Fig. 5a)<sup>35</sup>. HQ was observed to bind in three different orientations in total in the active site of TyrBm (orientations 1, 2 and 3) (Fig. 5, Supplementary Figs S7 and S8). It seems that HQ binding is rather flexible in the active site, agreeing with the *in silico* simulations shown above, and does not have one specific orientation in contrast to L-tyrosine and L-dopa (Fig. 5)<sup>35</sup>. In orientation 1, a polar interaction between HQ and Asn205 is observed (Fig. 5b). In orientation 2, the HQ molecule is oriented similarly to tyrosinase substrates (and KA) in the active site, supporting our kinetic experiments showing that HQ can act as a TyrBm substrate (Fig. 5a). In addition, when TyrBm crystals were soaked with copper and HQ for 16 hours, the crystals turned brown, in comparison to crystals that were soaked with zinc that did not show a change in color (data not shown). Brown TyrBm crystals indicate on substrate oxidation as was previously shown by Sendovski *et al.* and provide additional confirmation on the role of HQ as a substrate of TyrBm<sup>22</sup>.



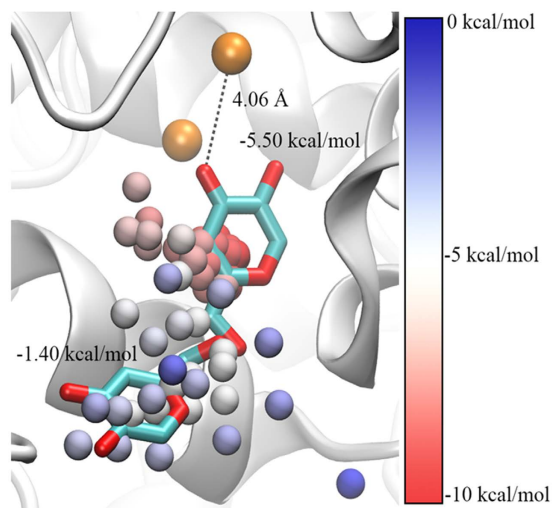
**Figure 3. Structures of KA bound to TyrBm.** (a) KA is observed inside the active site (purple) and at the entrance to the active site (green) (3NQ1). Copper ions are presented as brown spheres. (b) KA at the entrance to the active site (green) (3NQ1) is stabilized by second shell residues (light brown sticks). (c) Superposition with TyrBm structures contain KA (purple) and L-Dopa (orange, 4P6S) oriented through hydrophobic interactions with His208. All the structures presented in this work were generated using PyMOL.

| Structure name (PDB code)  | TyrBm:KA (5I38)                               | TyrBm:HQA (5I3A)                              | TyrBm:HQB (5I3B)                              |
|--|---|---|---|
| <i>Data collection</i>   |   |   |   |
| Space group  | P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> | P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> | P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> |
| Cell dimensions  |   |   |   |
| <i>a, b, c</i> (Å)   | 70.24, 74.97, 121.70                          | 69.62, 74.38, 120.78                          | 69.62, 74.42, 119.69                          |
| $\alpha, \beta, \gamma$ (°)                                      | 90, 90, 90                                    | 90, 90, 90                                    | 90, 90, 90                                    |
| Resolution (Å)   | 51.26–2.5                                     | 35.54–2.2                                     | 33.43–2.2                                     |
| <i>R</i> <sub>merge</sub> <sup>†</sup>                           | 0.12(0.25)                                    | 0.082(0.387)                                  | 0.08(0.326)                                   |
| <i>I</i> / $\sigma$ <i>I</i> <sup>†</sup>                        | 9.7(5.9)                                      | 18.2(6.2)                                     | 15.4(5.8)                                     |
| Completeness <sup>‡</sup>  | 90.4(99.3)                                    | 99.9(100)                                     | 99.8(99.9)                                    |
| Redundancy <sup>‡</sup>  | 6.2(5.8)                                      | 12.6(13.1)                                    | 6.6(6.9)                                      |
| <i>Refinement</i>  |   |   |   |
| Resolution (Å)   | 51.26–2.5                                     | 35.54–2.2                                     | 33.43–2.2                                     |
| No. of reflections   | 127,679                                       | 558,159                                       | 214,516                                       |
| <i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub> <sup>‡</sup> | 20.81/23.77                                   | 20.11/22.72                                   | 18.84/21.69                                   |
| No. of atoms   |   |   |   |
| Protein  | 4,687   | 4,697   | 4,696   |
| Ligand/ion   | 36  | 33  | 33  |
| Water  | 84  | 226   | 272   |
| B-factors (Å <sup>2</sup> )                                      |   |   |   |
| Protein  | 32.05   | 35.02   | 32.75   |
| Ligand/ion   | 33.51   | 39.86   | 29.39   |
| Water  | 30.74   | 37.73   | 35.42   |
| Root mean square deviations                                      |   |   |   |
| Bond length (Å)  | 0.004   | 0.007   | 0.006   |
| Bond angle (°)   | 0.59  | 0.88  | 0.80  |

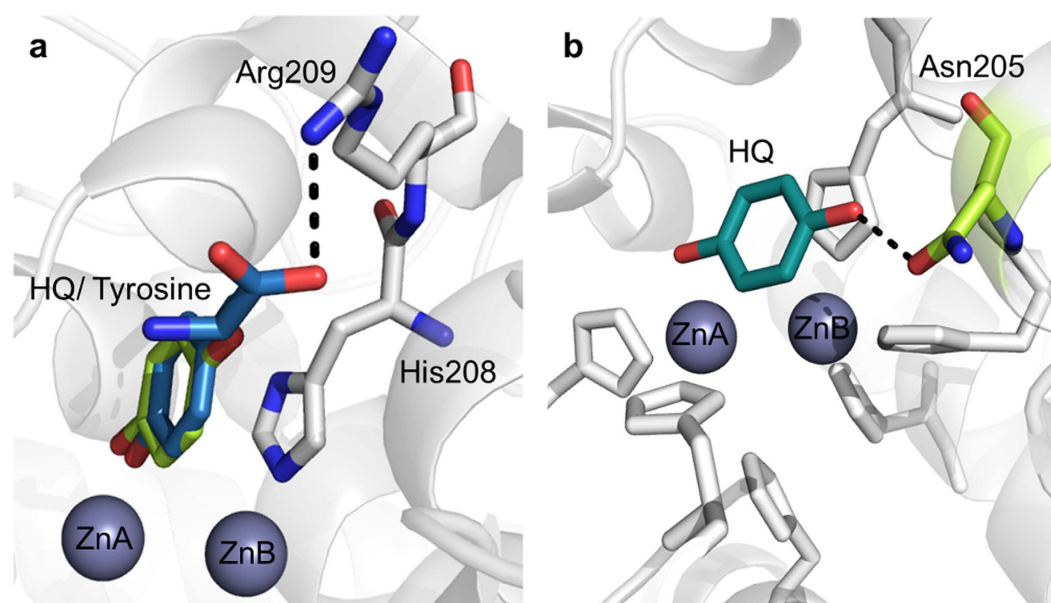
**Table 4. Data collection and refinement statistics.** <sup>†</sup>Values in parentheses are for the last shell.  $R_{\text{merge}} = \frac{\sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle|}{\sum_{\text{hkl}} \sum_i I_i(\text{hkl})}$ , where *I* is the observed intensity, and  $\langle I \rangle$  is the mean value of *I*.  $R/R_{\text{free}} = \frac{\sum_{\text{hkl}} |F_{\text{obs}}| - |F_{\text{calc}}|}{\sum_{\text{hkl}} |F_{\text{obs}}|}$  where *R* and *R*<sub>free</sub> are calculated using the test reflections respectively. The test reflections (5%) were held aside and not used during the entire refinement process.

## Discussion

Disorders in melanin formation have been linked to various skin diseases in humans such as hyperpigmentation and skin cancer. KA and HQ, are frequently used as inhibitors of tyrosinase, and have been used as skin-whitening agents in leading cosmetic hyperpigmentation treatment<sup>9,13,14,16</sup>. Over the past few years, numerous docking studies and molecular dynamic simulations were performed in an attempt to elucidate the binding modes of



**Figure 4.** KA's center of mass cluster analysis along the PELE simulation. Clusters are presented as spheres and colors indicate the potential of mean field  $\Delta G$ . Absolute standard binding free energies (with volume corrections) are shown for the active site and the surface bound minima, along with the ligand crystallographic complexes (cyan sticks). The two copper ions are presented as brown spheres.



**Figure 5.** Structures of HQ bound in the active site of TyrBm. (a) Superposition with TyrBm structures contain HQ in orientation 2 (green) and L-tyrosine (blue, 4P6R), which forms a hydrogen bond with Arg209. Zinc ions are presented as grey spheres. (b) HQ, in orientation 1 (teal), forms a hydrogen bond with Asn205 His residues are in white.

tyrosinase inhibitors. In this work we demonstrate for the first time the true binding orientations of KA and HQ in the active site of TyrBm which explain the biochemical characterization.

Previously, we had determined a crystal structure of TyrBm with KA bound at the entrance to the active site<sup>22</sup>. Here, by modifying our protocol for ligand binding *in crystal*, we have visualized the structure of TyrBm with KA bound in the active site similar to tyrosinase substrates (Fig. 3c). This position of KA might lead to the false assumption of competitive inhibition mechanism. However, the two orientations of KA, which are demonstrated by crystallography and *in silico* simulations (Figs 3 and 4), support the mixed inhibition mechanism, which is confirmed by our kinetic experiments (Fig. 1 and Table 1).

In contrast to previous studies<sup>18–20</sup>, in this work we unequivocally display mixed inhibition mode of KA on both monophenolase and diphenolase activities and undermine the hypothesis regarding copper chelation by KA. Since the  $K_{IS}$  value is significantly greater than  $K_I$  for the oxidation of both L-tyrosine and L-dopa (Table 1), KA is able to bind more strongly to the free enzyme than to the enzyme–substrate complex (at the peripheral

site). We suggest that when KA is bound strongly in the active site, the binding pocket is not accessible to substrate molecules, subsequently TyrBm is not active. However, when KA is oriented at the entrance of the active site, it restricts substrate entrance and product efflux, consequently, TyrBm cannot reach its maximum velocity (Table 1). Tropolone, another tyrosinase inhibitor that has been studied, was also found at the entrance of the active site of mushroom tyrosinase and exhibited mixed inhibition mode similar to KA<sup>41,42</sup>. It is quite possible, that KA may also bind to this site or has a different peripheral binding site yet to be elucidated. Together with the fact that KA showed a dissociation constant 3-orders of magnitude higher than L-tyrosine, these findings support the existence of a significant intermediate binding site in TyrBm and explains the mechanism of mixed inhibition.

Numerous studies have raised questions regarding the behavior of HQ as a tyrosinase inhibitor<sup>26,27</sup>, and most of them characterized HQ as a competitive inhibitor<sup>24,43–45</sup>. Our results support this inhibition mode of HQ on L-tyrosine (Fig. 2a and Table 1). On the other hand, other studies demonstrated the potency of HQ as a tyrosinase substrate<sup>26,27</sup>. Stratford and co-workers suggested that HQ is neither a substrate nor an inhibitor of tyrosinase<sup>27</sup> while del Mar García-Molina *et al.* suggested that HQ is a tyrosinase substrate with a poor activity due to the inability to transform *met*-tyrosinase to *oxy*-tyrosinase on its own<sup>26</sup>. The transformation is achieved by addition of an *o*-diphenol (such as L-dopa) or H<sub>2</sub>O<sub>2</sub>, which promotes the activity on HQ<sup>26,28</sup>. Our results corroborate this argument, since in the presence of H<sub>2</sub>O<sub>2</sub> or L-dopa, TyrBm was indeed active on HQ as measured by two unrelated methods (Fig. 2b and Supplementary Figs S1 and S3). These findings led us to determine the kinetic constants of TyrBm with HQ as a substrate, which resulted in a similar  $K_m$  value for HQ and L-dopa, and a  $V_{max}$  value higher than that of L-tyrosine (Table 2).

Additional conclusive evidence for the action of TyrBm on HQ lies in the formation of brown crystals soaked in HQ that indicate on substrate oxidation as was previously shown by Sendovski *et al.* with L-tyrosine<sup>22</sup>. We assume that small amounts of *oxy*-TyrBm molecules present in the crystals enabled the activity on HQ within 16 hours that resulted in brown pigmentation<sup>9</sup>.

In order to elucidate the inhibition mechanism of HQ, we solved two crystal structures of TyrBm with HQ in the active site. It was discovered that HQ is bound less strongly than L-tyrosine (Table 3), and its binding heterogeneity is evident from the several different orientations observed in the active site (Fig. 5 and Supplementary Fig. S7). It is presumed that the polar amine and carboxyl groups of L-tyrosine and L-dopa, which are not present in HQ, help to stabilize the substrates through polar interactions with Arg209 in a productive mode (Fig. 5a)<sup>35</sup>. In orientation 1, a hydrogen bond between HQ and Asn205 was observed (Fig. 5b). Asn205 was suggested to be crucial for tyrosinase activity through the activation of a conserved water molecule<sup>35,46</sup>. The interaction of Asn205 with HQ might prevent this activation, and thus inhibit tyrosinase activity<sup>18,47,48</sup>. Furthermore, in the structures of TyrBm with KA at the entrance of the active site this interaction was also found to be important for KA stabilization<sup>22</sup>. Thus, we propose that the polar bond between Asn205 and the hydroxyl group of HQ indicates on an inhibitory effect on TyrBm. In contrast, in orientation 2, HQ is positioned similarly to L-tyrosine in the active site (Fig. 5a), supporting the role of HQ as TyrBm substrate. The flexibility of HQ in the active site of TyrBm was also demonstrated by our PELE simulations, which provided visualization of the numerous energetically feasible orientations of HQ in the active site. Whereas for KA it takes some time and energy to go from the peripheral docking site to the inner active site, for HQ there is constant interconversion between the two of them suggesting that the pre-docking site is very transient.

It seems that the combination of both the orientation of HQ in the active site and the oxidative state of tyrosinase will define the behavior of HQ.

## Methods

**Expression, purification and crystallization of tyrosinase from *B. megaterium*.** The gene encoding tyrosinase from *Bacillus megaterium* (TyrBm) was cloned into *Escherichia coli* BL21, purified and crystallized as previously described<sup>49,50</sup>.

**Tyrosinase inhibition assay.** Tyrosinase inhibitory activity was determined spectrophotometrically in 96-well plates with a final volume of 200  $\mu$ l. First, 50 mM PBS buffer pH 7.4 and 0.01 mM CuSO<sub>4</sub> were mixed with 6  $\mu$ g ml<sup>-1</sup> of purified enzyme. Then, the mixture was incubated at 40 °C for 2 minutes. Finally, various concentrations of inhibitor were mixed with 1.2 mM L-tyrosine or 2 mM L-dopa and were added to the pre-incubated mixture. KA was studied in the range of 0.025–0.1 mM and HQ in the range of 0.025–0.5 mM. The reaction mixture was then monitored for L-dopachrome formation ( $\epsilon = 3600 \text{ M}^{-1} \text{ cm}^{-1}$ ) by measuring the absorbance at 475 nm. Specific activity was calculated as the ratio of the conversion rate and the total protein content as determined by the Bradford analysis method (Bio-Rad, Israel). All measurements were performed in seven replicates. The inhibitor concentration necessary for 50% inhibition (IC<sub>50</sub>) was determined with respect to a control (no inhibitor).

**Kinetic analysis of tyrosinase.** The mode of inhibition and inhibition parameters, i.e. the Michaelis–Menten constant ( $K_m$ ), maximal velocity ( $V_{max}$ ), turnover number ( $k_{cat}$ ) of TyrBm and the inhibition constants ( $K_i$ ,  $K_{is}$ ) of each inhibitor were determined by Lineweaver–Burk plot analysis using various concentrations of L-tyrosine (0.03–1.4 mM) and L-dopa (0.15–2.0 mM) as substrates. The inhibitor concentrations were mentioned above. The inhibition kinetics module of Sigma Plot 13.0 software was used (Systat Software, Inc., Richmond, CA, USA). All measurements were performed in 5-replicates.

**Tyrosinase activity assay on HQ as a substrate.** TyrBm activity was determined by measuring the formation of 2-hydroxy-*p*-benzoquinone (HPB) from HQ, in the presence of H<sub>2</sub>O<sub>2</sub> or L-dopa. TyrBm activity was determined as explained in the inhibition assay above with varying concentrations of H<sub>2</sub>O<sub>2</sub>, ranging from 0 to 90 mM, while maintaining the concentration of HQ constant. The formation of HBP was monitored by measuring the absorbance at 475 nm<sup>26</sup>.



The values of  $K_m$ ,  $V_{max}$  and  $k_{cat}$  of TyrBm in the presence of HQ as a substrate were determined with the following conditions: 50 mM PBS buffer pH 7.4, 0.01 mM  $CuSO_4$ ,  $6 \mu g ml^{-1}$  of purified enzyme, various concentrations of HQ (0.1–2.0 mM) in the presence of saturating concentration of hydrogen peroxide (100 mM)<sup>51</sup>.

For further verification of TyrBm activity on HQ, the activity was determined by recording the oxygen consumption in the presence of L-dopa and HQ. Measurements were carried out using a Hansatech Oxygraph+ electrode (Norfolk, UK) in a reaction volume of 1000  $\mu L$ . The reaction contained  $4 \mu g ml^{-1}$  of purified TyrBm, 50 mM PBS buffer pH 7.4, 0.01 mM  $CuSO_4$ , 1 mM L-dopa and 0.1 mM HQ.

**Dissociation constants using MicroScale Thermophoresis (MST).** TyrBm was labeled fluorescently with a RED dye (NT-647-NHS) according to the manufacturer's protocol (NanoTemper Technologies, Munich, Germany). Non-bound dye was removed by purification of the enzyme on a Sephadex G-25 column with buffers provided in the commercial kit. Then, serial dilutions of unlabeled binding partner samples (inhibitor or substrate) were mixed with  $0.377 \mu M$  of dye-labeled TyrBm in 50 mM PBS buffer pH 7.4 and incubated for 5 minutes. Approximately 10  $\mu l$  of sample was loaded into hydrophilic monolith NT capillaries and the measurement was performed in a NanoTemper Monolith NT.015T instrument. The emission of the red fluorescence was recorded at a focused location of the capillary. In the same location, a microscopic temperature gradient was created using an infrared laser and the fluorescence depletion was measured. According to changes in the fluorescent thermophoresis signal and the concentrations of unlabeled inhibitor, the dissociation constant values were determined by the NanoTemper analysis software. The unlabeled binding partners tested for  $K_D$  determination were KA (0–4 mM), HQ (0–1 mM) and L-tyrosine (0–2 mM).

**Statistical analysis.** All experiments were performed in duplicates or triplicates in order to ensure the reproducibility of the results. Statistical analysis was performed using Student's t-test: \* $P < 0.05$  compared with the control. Data is summarized as mean  $\pm$  SD.

**Substrate binding in crystals.** In order to trap ligands in the active site, mature crystals were soaked overnight in 1 mM of either  $CuSO_4$  or  $ZnCl_2$  and subsequently in 10 mM of the appropriate ligand (KA and HQ) before crystal freezing.

**Data collection and structure determination.** X-ray diffraction data was collected at the European Synchrotron Radiation Facility, Grenoble, France, at beamlines ID14-4 and ID 29. All data were indexed, integrated, scaled and merged using Mosflm and Scala<sup>52</sup>. The structures of TyrBm with bound inhibitors were solved by molecular replacement using Phaser<sup>53</sup> and the coordinates of earlier determined TyrBm structure (PDB code 4P6R). Refinement was performed using Phenix<sup>54</sup> and Refmac5<sup>55,56</sup>, coupled with rounds of manual model building, real-space refinement and structure validation performed using COOT<sup>57</sup>. Data collection, phasing and refinement statistics are presented in Table 3.

**In silico simulations.** Ligand migration sampling with Protein Energy Landscape Energy (PELE). PELE has widely been used to study ligand-protein interactions and protein dynamics at a fraction of the cost compared to other sampling methods such as molecular dynamics<sup>58–60</sup>. This algorithm is composed of a perturbation and a relaxation stage, and uses a mixture of random moves with protein structure prediction algorithms. The resulting structure is accepted or rejected following the Metropolis criterion.

Binding free energy with Markov State Models (MSM). MSM are coarse grain statistical models that allow extracting equilibrium properties such as the binding free energy<sup>61</sup>. In order to build our MSM, we split the conformational space using the Voronoi decomposition, clustering the ligand's center of mass and using the cluster centers as seeds. Hence, each microstate will contain all possible ligand, protein and solvent arrangements compatible with having the ligand's center of mass within the cell. In order to study the different metastable minima, microstates are kinetically clustered utilizing Perron Cluster Analysis (PCCA+). The absolute binding free energy,  $\Delta G$ , is obtained integrating the potential of mean force (Gpmf) in the whole bound region<sup>39</sup>.

## References

- Coates, C. J. & Nairn, J. Diverse immune functions of hemocyanins. *Dev. Comp. Immunol.* **45**, 43–55 (2014).
- Olianas, A., Sanjust, E., Pellegrini, M. & Rescigno, A. Tyrosinase activity and hemocyanin in the hemolymph of the slipper lobster *Scyllarides latus*. *Journal of Comparative Physiology B* **175**, 405–411 (2005).
- Decker, H. & Tuczek, F. Tyrosinase/catecholoxidase activity of hemocyanins: structural basis and molecular mechanism. *Trends Biochem. Sci.* **25**, 392–397 (2000).
- Kaintz, C., Mauracher, S. G. & Rompel, A. In *Advances in Protein Chemistry and Structural Biology* Vol. 97 (ed C. Z. Christov) 1–35 (Academic Press, 2014).
- Claus, H. & Decker, H. Bacterial tyrosinases. *Syst. Appl. Microbiol.* **29**, 3–14 (2006).
- Decker, H. *et al.* Similar enzyme activation and catalysis in hemocyanins and tyrosinases. *Gene* **398**, 183–191 (2007).
- Kanteev, M., Goldfeder, M. & Fishman, A. Structure–function correlations in tyrosinases. *Protein Sci.* **24**, 1360–1369 (2015).
- Halaouli, S., Asther, M., Sigoillot, J. C., Hamdi, M. & Lomascolo, A. Fungal tyrosinases: new prospects in molecular characteristics, bioengineering and biotechnological applications. *J. Appl. Microbiol.* **100**, 219–232 (2006).
- Chang, T.-S. An updated review of tyrosinase inhibitors. *Int. J. Mol. Sci.* **10**, 2440–2475 (2009).
- Rao, A. R. *et al.* Effective inhibition of skin cancer, tyrosinase, and antioxidative properties by astaxanthin and astaxanthin esters from the green alga *Haematococcus pluvialis*. *J. Agric. Food Chem.* **61**, 3842–3851 (2013).
- Erdogan Orhan, I. & Tareq Hassan Khan, M. Flavonoid derivatives as potent tyrosinase inhibitors—a survey of recent findings between 2008–2013. *Curr. Top. Med. Chem.* **14**, 1486–1493 (2014).
- Abu Ubeid, A. & Hantash, B. M. Minireview: peptide analogs and short sequence oligopeptides as modulators of skin pigmentation. *Curr. Top. Med. Chem.* **14**, 1418–1424 (2014).
- Solano, F., Briganti, S., Picardo, M. & Ghanem, G. Hypopigmenting agents: an updated review on biological, chemical and clinical aspects. *Pigment Cell Res.* **19**, 550–571 (2006).

14. Parvez, S. *et al.* Survey and mechanism of skin depigmenting and lightening agents. *Phytother. Res.* **20**, 921–934 (2006).
15. Bagherzadeh, K. *et al.* A new insight into mushroom tyrosinase inhibitors: docking, pharmacophore-based virtual screening, and molecular modeling studies. *J. Biomol. Struct. Dyn.* **33**, 487–501 (2015).
16. Gillbro, J. & Olsson, M. The melanogenesis and mechanisms of skin-lightening agents—existing and new approaches. *Int. J. Cosmetic Sci.* **33**, 210–221 (2011).
17. Burnett, C. L. *et al.* Final report of the safety assessment of kojic acid as used in cosmetics. *Int. J. Toxicol.* **29**, 244S–273S (2010).
18. Lima, C. R. *et al.* Combined kinetic studies and computational analysis on kojic acid analogs as tyrosinase inhibitors. *Molecules* **19**, 9591–9605 (2014).
19. Bochot, C. *et al.* Probing kojic acid binding to tyrosinase enzyme: insights from a model complex and QM/MM calculations. *Chem. Commun.* **50**, 308–310 (2014).
20. Noh, J.-M. *et al.* Kojic acid–amino acid conjugates as tyrosinase inhibitors. *Bioorg. Med. Chem. Lett.* **19**, 5586–5589 (2009).
21. Battaini, G., Monzani, E., Casella, L., Santagostini, L. & Pagliarini, R. Inhibition of the catecholase activity of biomimetic dinuclear copper complexes by kojic acid. *J. Biol. Inorg. Chem.* **5**, 262–268 (2000).
22. Sendovski, M., Kanteev, M., Ben-Yosef, V. S., Adir, N. & Fishman, A. First structures of an active bacterial tyrosinase reveal copper plasticity. *J. Mol. Biol.* **405**, 227–237 (2011).
23. Ramsden, C. A. & Riley, P. A. Mechanistic aspects of the tyrosinase oxidation of hydroquinone. *Bioorg. Med. Chem. Lett.* **24**, 2463–2464 (2014).
24. Chiari, M. E., Vera, D. M. A., Palacios, S. M. & Carpinella, M. C. Tyrosinase inhibitory activity of a 6-isoprenoid-substituted flavanone isolated from *Dalea elegans*. *Bioorg. Med. Chem.* **19**, 3474–3482 (2011).
25. Chawla, S. *et al.* Mechanism of tyrosinase inhibition by deoxyarbutin and its second-generation derivatives. *Br. J. Dermatol.* **159**, 1267–1274 (2008).
26. del Mar García-Molina, M. *et al.* Tyrosinase-catalyzed hydroxylation of hydroquinone, a depigmenting agent, to hydroxyhydroquinone: a kinetic study. *Bioorg. Med. Chem.* **22**, 3360–3369 (2014).
27. Stratford, M. R., Ramsden, C. A. & Riley, P. A. The influence of hydroquinone on tyrosinase kinetics. *Bioorg. Med. Chem.* **20**, 4364–4370 (2012).
28. Ramsden, C. A. & Riley, P. A. Tyrosinase: the four oxidation states of the active site and their relevance to enzymatic activation, oxidation and inactivation. *Bioorg. Med. Chem.* **22**, 2388–2395 (2014).
29. Mendes, E., Perry, M. d. J. & Francisco, A. P. Design and discovery of mushroom tyrosinase inhibitors and their therapeutic applications. *Expert Opin. Drug Discov.* **9**, 533–554 (2014).
30. Komori, Y., Imai, M., Yamauchi, T., Higashiyama, K. & Takahashi, N. Effect of p-aminophenols on tyrosinase activity. *Bioorg. Med. Chem.* **22**, 3994–4000 (2014).
31. Neeley, E. *et al.* Variations in IC<sub>50</sub> values with purity of mushroom tyrosinase. *Int. J. Mol. Sci.* **10**, 3811–3823 (2009).
32. Ha, Y. M. *et al.* 4-(6-Hydroxy-2-naphthyl)-1,3-benzodiol: a potent, new tyrosinase inhibitor. *Biol. Pharm. Bull.* **30**, 1711–1715 (2007).
33. Choi, J., Choi, K.-E., Park, S.-J., Kim, S. Y. & Jee, J. Ensemble-based virtual screening led to the discovery of new classes of potent tyrosinase inhibitors. *J. Chem. Inf. Model.* **56**, 354–367 (2016).
34. Jerabek-Willemsen, M. *et al.* Microscale thermophoresis: interaction analysis and beyond. *J. Mol. Struct.* **1077**, 101–113 (2014).
35. Goldfeder, M., Kanteev, M., Isaschar-Ovdat, S., Adir, N. & Fishman, A. Determination of tyrosinase substrate-binding modes reveals mechanistic differences between type-3 copper proteins. *Nat. Commun.* **5** (2014).
36. Nithitanakool, S., Pithayanukul, P., Bavovada, R. & Saparpakorn, P. Molecular docking studies and anti-tyrosinase activity of Thai mango seed kernel extract. *Molecules* **14**, 257–265 (2009).
37. Deeth, R. J. & Diedrich, C. Structural and mechanistic insights into the oxy form of tyrosinase from molecular dynamics simulations. *J. Biol. Inorg. Chem.* **15**, 117–129 (2010).
38. Borrelli, K. W., Cossins, B. & Guallar, V. Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J. Comput. Chem.* **31**, 1224–1235 (2010).
39. Takahashi, R., Gil, V. A. & Guallar, V. Monte Carlo free ligand diffusion with markov state model analysis and absolute binding free energy calculations. *J. Chem. Theory Comput.* **10**, 282–288 (2013).
40. Han, H.-Y. *et al.* The inhibition kinetics and thermodynamic changes of tyrosinase via the zinc ion. *BBA-Proteins Proteom.* **1774**, 822–827 (2007).
41. Ismaya, W. T. *et al.* Crystal structure of *Agaricus bisporus* mushroom tyrosinase: identity of the tetramer subunits and interaction with tropolone. *Biochemistry* **50**, 5477–5486 (2011).
42. Kahn, V. & Andrawis, A. Inhibition of mushroom tyrosinase by tropolone. *Phytochemistry* **24**, 905–908 (1985).
43. Inoue, Y. *et al.* Analysis of the effects of hydroquinone and arbutin on the differentiation of melanocytes. *Biol. Pharm. Bull.* **36**, 1722–1730 (2013).
44. Chen, Y.-S. *et al.* Kinetic study on the tyrosinase and melanin formation inhibitory activities of carthamus yellow isolated from *Carthamus tinctorius* L. *J. Biosci. Bioeng.* **115**, 242–245 (2013).
45. Yang, C.-H., Chen, Y.-S., Lai, J.-S., Hong, W. W. & Lin, C.-C. Determination of the thermodegradation of deoxyArbutin in aqueous solution by high performance liquid chromatography. *Int. J. Mol. Sci.* **11**, 3977–3987 (2010).
46. Solem, E., Tuzcek, F. & Decker, H. Tyrosinase versus Catechol Oxidase: one Asparagine Makes the Difference. *Angew. Chem. Int. Ed.* **55**, 2884–2888 (2016).
47. Asadzadeh, A., Fassih, A., Yaghmaei, P. & Pourfarzam, M. Docking studies of some novel kojic acid derivatives as possible tyrosinase inhibitors. *Biomed. Pharmacol. J.* **8**, 535–545 (2015).
48. Hu, Y.-H. *et al.* 4-Hydroxy cinnamic acid as mushroom preservation: anti-tyrosinase activity kinetics and application. *Int. J. Biol. Macromol.* **86**, 489–495 (2016).
49. Shuster, V. & Fishman, A. Isolation, cloning and characterization of a tyrosinase with improved activity in organic solvents from *Bacillus megaterium*. *J. Mol. Microbiol. Biotechnol.* **17**, 188–200 (2009).
50. Sendovski, M., Kanteev, M., Shuster Ben-Yosef, V., Adir, N. & Fishman, A. Crystallization and preliminary x-ray crystallographic analysis of a bacterial tyrosinase from *Bacillus megaterium*. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **66**, 1101–1103 (2010).
51. Ortiz-Ruiz, C. V., Berna, J., Rodriguez-Lopez, J. N., Tomas, V. & Garcia-Canovas, F. Tyrosinase-catalyzed hydroxylation of 4-hexylresorcinol, an antibrowning and depigmenting agent: a kinetic study. *J. Agric. Food Chem.* **63**, 7032–7040 (2015).
52. Leslie, A. G. W. *joint CCP4+ ESF-EAMCB Newsletter on Protein Crystallography* **26** (1992).
53. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. Sect. D Struct. Biol. Cryst. Commun.* **63**, 32–41 (2007).
54. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Struct. Biol. Cryst. Commun.* **66**, 213–221 (2010).
55. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. Sect. D Struct. Biol. Cryst. Commun.* **53**, 240–255 (1997).
56. Skubak, P., Murshudov, G. N. & Pannu, N. S. Direct incorporation of experimental phase information in model refinement. *Acta Crystallogr. Sect. D Struct. Biol. Cryst. Commun.* **60**, 2196–2201 (2004).

57. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Struct. Biol. Cryst. Commun.* **60**, 2126–2132 (2004).
58. Kotev, M., Lecina, D., Tarragó, T., Giralt, E. & Guallar, V. Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques. *Biophys. J.* **108**, 116–125 (2015).
59. Babot, E. D. *et al.* Steroid hydroxylation by basidiomycete peroxygenases: a combined experimental and computational study. *Appl. Environ. Microbiol.* **81**, 4130–4142 (2015).
60. Cossins, B. P., Hosseini, A. & Guallar, V. Exploration of protein conformational change with PELE and meta-dynamics. *J. Chem. Theory Comput.* **8**, 959–965 (2012).
61. Bowman, G. R., Pande, V. S. & Noé, F. *An introduction to markov state models and their application to long timescale molecular simulation*. Vol. 797 (Springer Science & Business Media, 2013).

## Acknowledgements

This work was supported by the Israel Science Foundation founded by the Israel Academy of Sciences and Humanities, grant number 419/15 and by the Gurwin Fund for Scientific Research. We also acknowledge the Russell-Berrie Nanotechnology Institute (RBNI) at the Technion for supporting this research. This research benefited from use of the Technion Center of Structural Biology facility of the Lorry I. Lokey Center for Life Sciences and Engineering. We thank the staff of the European Synchrotron Radiation Facility (beamlines ID14-4 and ID 29) for provision of synchrotron radiation facilities and assistance. V.G. would like to thank the OxiDesign Spanish project (CTQ2013-48287-R) and D.L. acknowledges support from the Spanish Severo Ochoa Program (SEV-2011-00067).

## Author Contributions

A.F. and N.A. conceived and designed the study, analyzed the results and wrote the manuscript. B.D. and M.K. performed and analyzed the structural and biochemical experiments, and wrote the manuscript. M.G. performed the crystallization experiments and analysis of kojic acid structures. D.L. and V.G. performed the molecular dynamic simulations and wrote the manuscript. All authors discussed and approved the manuscript.

## Additional Information

**Accession codes:** The coordinates and structure factors of TyrBm in different states have been deposited in the RCSB PDB under accession codes 5I3B (TyrBm with configuration B of HQ), 5I3A (TyrBm with configuration A of HQ) and 5I38 (TyrBm with KA).

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Deri, B. *et al.* The unravelling of the complex pattern of tyrosinase inhibition. *Sci. Rep.* **6**, 34993; doi: 10.1038/srep34993 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016

## **Publication 4 - Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions**

**Authors:** Christoph Grebner\*, [Daniel Lecina\\*](#), Victor Gil, Johan Ulander, Pia Hansson, Anita Dellsen, Christian Tyrchan, Karl Edman, Anders Hogner, Victor Guallar. (*\* equally contributing authors*)

**Journal:** Biophysical Journal, 112, 1147-1156 (2017)

### **Summary:**

In this work, we extended the binding mechanism study to all the members of the steroid nuclear hormone receptor family and their endogenous ligand. We found a shared entry path through the helix 3, 7, and 11 region, and identified two different folds of the helix 6-7 regions that had an impact in the number of observed binding events in unbiased simulations. We also saw that adding *X-ray* information into the protein perturbation promoted the plasticity of the helix 6-7 regions, and thus enhanced the sampling of binding events compared to the anisotropic network model. These PCA-based modes in combination with a path sampling can be used to improve the convergence of MSM simulations. Our absolute binding free energy estimations were in very good agreement with experimental results. The binding mechanisms analysis highlighted the importance of a previously reported peripheral binding site, and reported the influence of ligand hydrophobicity into the transition of the peripheral binding site into the active binding site.

### **Author contribution:**

My tasks involved launching the simulations, analyzing results and writing the final manuscript.





# Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions

Christoph Grebner,<sup>1,\*</sup> Daniel Lecina,<sup>2</sup> Victor Gil,<sup>2</sup> Johan Ulander,<sup>1</sup> Pia Hansson,<sup>3</sup> Anita Dellsen,<sup>3</sup> Christian Tyrchan,<sup>4</sup> Karl Edman,<sup>3,\*</sup> Anders Hogner,<sup>1</sup> and Victor Guallar<sup>2,5,\*</sup>

<sup>1</sup>Cardiovascular & Metabolic Disease (CVMD), AstraZeneca, Mölndal, Sweden; <sup>2</sup>Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain; <sup>3</sup>Discovery Sciences, AstraZeneca, Mölndal, Sweden; <sup>4</sup>Respiratory, Inflammation, and Autoimmunity (RIA), AstraZeneca, Mölndal, Sweden; and <sup>5</sup>Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

**ABSTRACT** In this study, we performed an extensive exploration of the ligand entry mechanism for members of the steroid nuclear hormone receptor family (androgen receptor, estrogen receptor  $\alpha$ , glucocorticoid receptor, mineralocorticoid receptor, and progesterone receptor) and their endogenous ligands. The exploration revealed a shared entry path through the helix 3, 7, and 11 regions. Examination of the x-ray structures of the receptor-ligand complexes further showed two distinct folds of the helix 6–7 region, classified as “open” and “closed”, which could potentially affect ligand binding. To improve sampling of the helix 6–7 loop, we incorporated motion modes based on principal component analysis of existing crystal structures of the receptors and applied them to the protein-ligand sampling. A detailed comparison with the anisotropic network model (an elastic network model) highlights the importance of flexibility in the entrance region. While the binding (interaction) energy of individual simulations can be used to score different ligands, extensive sampling further allows us to predict absolute binding free energies and analyze reaction kinetics using Markov state models and Perron-cluster cluster analysis, respectively. The predicted relative binding free energies for three ligands binding to the progesterone receptor are in very good agreement with experimental results and the Perron-cluster cluster analysis highlighted the importance of a peripheral binding site. Our analysis revealed that the flexibility of the helix 3, 7, and 11 regions represents the most important factor for ligand binding. Furthermore, the hydrophobicity of the ligand influences the transition between the peripheral and the active binding site.

## INTRODUCTION

Understanding the underlying processes in protein-ligand binding events is a key parameter in computer-aided drug design (1), often requiring a proper description of the induced-fit (1–3) and/or the conformational ensembles (1,3,4). Such analysis mandates a thorough study of the flexibility and accessible conformational states of the proteins, which could play a fundamental role in the biological function and response to modulators (1).

The intrinsic conformational flexibility specifically plays a crucial role in the steroid nuclear hormone receptor family, which belongs to the nuclear hormone receptor super fam-

ily. The family consists of five members: the androgen receptor (AR), estrogen receptor (ER $\alpha$ ,  $\beta$ ), glucocorticoid receptor (GR), mineralocorticoid receptor (MR), and progesterone receptor (PR). All steroid receptors regulate gene expression upon binding to cholesterol derivatives (5) and are involved in various physiological functions ranging from embryonic development to cell differentiation or homeostasis. Due to their critical roles in diverse biological processes, the receptors have received a lot of attention from the pharmaceutical industry, which has resulted in several medicines with application in diabetes (6,7), cancer (8,9), heart diseases (10), or COPD and asthma (11,12).

The nuclear hormone receptors share a common architecture with three separate domains: the variable N-terminal domain, the highly conserved DNA binding domain, and the ligand binding domain (LBD) (13). The endogenous ligands bind within the LBD, thereby triggering specific receptor conformational changes that determine the biological function of the complex. For instance, ligand binding

Submitted October 25, 2016, and accepted for publication February 1, 2017.

\*Correspondence: [christoph.grebner@astrazeneca.com](mailto:christoph.grebner@astrazeneca.com) or [karl.edman@astrazeneca.com](mailto:karl.edman@astrazeneca.com) or [victor.guallar@bsc.es](mailto:victor.guallar@bsc.es)

Christoph Grebner and Daniel Lecina contributed equally to this work.

Editor: Bert de Groot.

<http://dx.doi.org/10.1016/j.bpj.2017.02.004>

© 2017 Biophysical Society.



allosterically controls the coregulator-binding site called “activation function-2”, located at the surface of the LBD, which allows the receptor to interact with transcriptional cofactors (13).

The x-ray structures of the LBD for the five aligned steroid receptors in complex with their endogenous ligands (AR, testosterone; ER $\alpha$ , estradiol; GR, cortisol; MR, aldosterone; and PR, progesterone) are shown in Fig. 1.

The general conserved structural motif of the LBD for all steroid receptors encompasses 12 helices. It is built up by a three-layered  $\alpha$ -helical sandwich fold enveloping the ligand binding pocket in between the helices (13), where the endogenous ligands show similar binding modes in a fully occluded binding pocket for all the receptors. The conformations of the receptors' LBD do overlay very well (the overall backbone root mean square deviation (RMSD) for the different receptors ranges from 1.0 to 1.8 Å; see Fig. 1).

Although the five receptors share the overall similar fold, there are notable local variations in specific parts of the x-ray structures including conformational differences in helix 3, helix 6–7, and helix 11. Helix 3 has already been discussed for playing a role in ligand entry (14,15). The large flexibility of the loop connecting helix 6 and helix 7 is of special importance when studying ligand binding events, due to its close vicinity to the ligand binding site. The plasticity of this region as well as its integral part during ligand entry was recently discussed in respect to MR and GR ligand entry mechanisms (16).

In this study, we extend the investigation of protein flexibility and ligand entry mechanism to the entire steroid receptor family. One key question unresolved so far is if all steroid receptors share the same entry pathway. As in our previous work (16), we use the state-of-the-art computational tool Protein Energy Landscape Exploration (PELE),

to efficiently sample the conformational space and protein-ligand binding events. This method (17) has been shown to be accurate and efficient in locating the binding site and ligand induced-fit mechanisms even for deeply buried and complex binding pockets (16,18–21). To model the binding event even more accurately, we tuned PELE to incorporate experimentally observed receptor flexibility using principal component analysis (PCA). Various x-ray structures of the steroid receptors in complex with different ligands indicate significant differences in the receptor conformations, which might aid in modeling the intrinsic plasticity of the receptors upon ligand binding. Therefore, we hypothesize that incorporation of the experimentally available information (e.g., PCA analysis of x-ray structures) could improve sampling of backbone flexibility and simulation of the ligand binding events.

Our results provide a detailed analysis of the ligands' entry mechanism and the influence of different factors like protein dynamics and conformational states. In addition, for PR, the simulation results are further used to predict absolute binding free energies of the three ligands progesterone, cortisol, and aldosterone and to investigate the reaction profiles and kinetics using Markov state modeling (MSM) (22,23).

## MATERIALS AND METHODS

### PELE algorithm

PELE is a Monte Carlo (MC)-based technique that uses protein structure prediction algorithms (17,24). Each MC move consists of three main steps, i.e., ligand and protein perturbation; side chain sampling; and minimization. Ligand perturbation is based on rotation and translation, whereas the protein perturbation is based on an all-atom minimization with constrained displacements along the C $\alpha$ -atoms following a set of given modes.

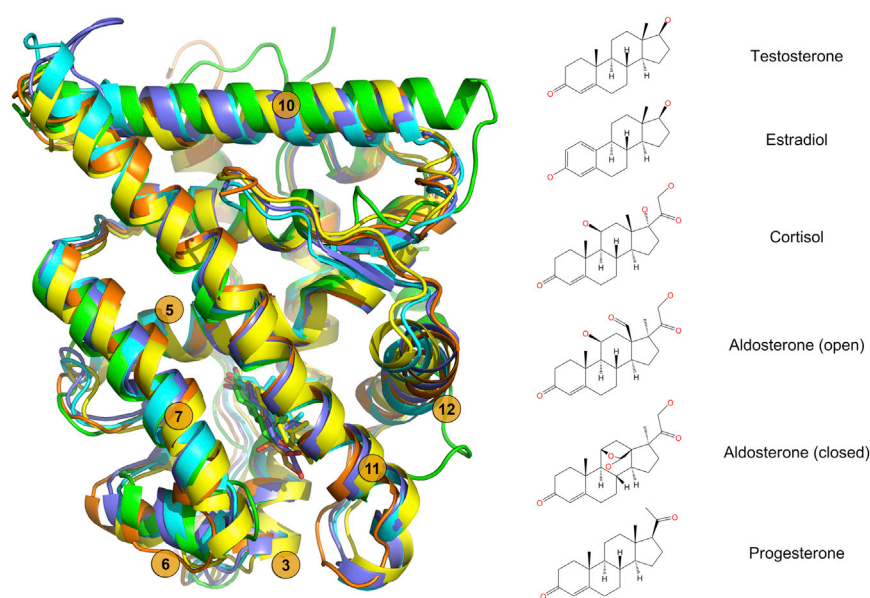


FIGURE 1 Overlay of the five steroid receptors AR (cyan), ER $\alpha$  (green), GR (orange), MR (ice blue), and PR (yellow) in complex with the endogenous ligand (testosterone, estradiol, cortisol, aldosterone, and progesterone, respectively; chemical structures of the hormones on the right side). Helices numbered according to Bourguet et al. (42). To see this figure in color, go online.

In the original PELE approach, the calculation of these modes is performed employing the anisotropic network model (ANM) method (25,26). The development of a new, to our knowledge, protein perturbation based on x-ray structure information is described in this article. The side chain sampling step includes all side chains for residues with at least one atom within 6 Å of the ligand's heavy atoms. The last step involves a complete minimization of the system. The resulting structure is accepted or rejected by applying a Metropolis criterion. In these simulations, an optimized-potentials-for-liquid-simulations (OPLS-2005) (27) all-atom force field with an Onufriev-Bashford-Case continuum solvent model (28) was used.

## Structure preparation

All simulations for the investigated steroid receptor with the endogenous ligand were started from the following x-ray structure Protein Data Bank IDs: AR, 2Q7J; ER $\alpha$ , 1QKT; GR, 4P6X; MR, 2AA2; and PR, 1A28. Structures were prepared using the Protein Preparation Wizard (29) of Maestro (30), adding hydrogen atoms, checking the protonation states of side chains (at pH = 7.0), and optimizing the hydrogen-bond network. The resulting structures were checked by visual inspection. If necessary, loops were closed using Prime (AR, C844–K845; MR, K909–N913). For all systems, no coactivator peptide was included in the simulations (mirroring experimental conditions).

## Binding site exploration

Two different binding site exploration simulations were performed. First, a complete space search was performed by placing the ligand in the bulk solvent and using 400 independent trajectories (one trajectory per computing core, using Intel Xeon CPU E5620 processors; Intel, Santa Clara, CA) for 48 h. Then, a local refinement search was performed, where the ligand was manually placed at the surface in the proximity of the proposed entry, ~20 Å away from the active site. Local search simulations were run using the same number of trajectories for 24 h, where the ligand moves freely within a 20 Å sphere around the central point of the binding site (defined by the center of mass of the bound ligand).

Ligand perturbation depended on the ligand's solvent-accessible surface area (SASA, indicating the percentage [0:1] of the ligand's surface available to the solvent). For SASA values <0.2, the translation was 0.75 Å and rotation 0.1 rad; otherwise, translation was set to 2.0 Å and rotations to 0.15 rad. Rotation was increased to 0.45 rad within 4 Å to the binding site, allowing potential reorientations of the ligand. The number of steering steps, i.e., the number of steps that the ligand perturbation direction is kept, was eight, to enhance the entrance into cavities.

A combination of the six main modes (the lowest in ANM or those with larger variance in PCA) was used for perturbing the protein backbone. The main mode was mixed 50/50 with a random mixture of the five remaining ones, which was updated every three steps. A maximal displacement for the C $\alpha$ -atoms constrained potential of 1.5 Å was used.

PELE results are analyzed by plots combining the binding energy and the ligand heavy atom root RMSD from the bound x-ray structure; the binding energy is computed as the internal energy difference between the complex and the free receptor and ligand:  $E_{\text{bind}} = E_{\text{AB}} - (E_{\text{A}} + E_{\text{B}})$ , where the energy function takes into account solvation terms.

## PELE settings for comparing ANM and PCA

Different backbone perturbation settings were used to compare the ANM and PCA perturbation methods. The mode with highest amplitude for the helix 6–7 movement was selected as the main mode (PCA, AR, ER $\alpha$ , as mode 4; PR, GR, MR, as mode 2; ANM, AR, GR, as mode 5; ER, as mode 2; MR, as mode 4; and PR, as mode 3). Because we aim at quantitatively discriminating the binding events for the two different techniques,

ligand translation was reduced to 0.5 and 1.5 Å with rotational steps of 0.05 and 0.15 rad (all variables equally distributed). Due to the smaller ligand translation, the number of steering translation steps was sequentially increased for each receptor until entries were found: AR, 9; AR with MR-modes, 7; ER $\alpha$ , 1, GR, 9; MR, 7; and PR, 2. Simulations have been performed for 500 MC steps using 264 processors.

## PELE settings for calculating binding free energies with MSM

Absolute binding free energy simulations involved 600 independent trajectories for 24 h using the local restricted space described above. To improve convergence, we used six different initial structures with SASA ~1.0, 0.8, 0.6, 0.4, 0.2, and 0.0 (bound complex), extracted from a binding trajectory; we thus had 100 trajectories starting in each of the initial structures. To avoid nonphysical transition between states, ligand translation and rotation adopted the small values described above in PELE Settings for Comparing ANM and PCA, and the number of steering steps was reduced to three.

## PCA modes based on x-ray structures

PCA analysis was performed in a Python program ([www.python.org](http://www.python.org)) employing the ProDy library, version 1.5.1 ([prody.csb.pitt.edu](http://prody.csb.pitt.edu)) (31). For each system, the x-ray complex of the receptors with its endogenous ligand was taken as the reference structure, and public structures (monomer A) with a sequence identity of at least 98% and a maximum of one missing residue (gap) were taken; all structures are listed in Table S1. Structures were superimposed (C $\alpha$ -atoms) to the reference structure using ProDy's iterative superposition algorithm. The ensemble was used for calculation of the covariance matrix, which, after diagonalization, yielded the principal components. These were saved using ProDy's nmd file format (31). Only the C $\alpha$ -atoms were taken into account and gaps were treated with a weight of 0.0.

PELE code was then modified to load modes externally from files in the nmd-file format, replacing the ANM calculation step in PELE. This integration made it possible to use the same common code for both methods.

## Binding affinities

### Prediction with MSMs

The MSM method (23) is a powerful technique for describing the equilibrium properties of a system, based on the concept that conformational changes can be modeled as Markov chains. To build them with our MC procedure, we used the following steps: First, we ran 600 unbiased PELE trajectories during 24 h, starting from different conformations along the binding pathway (as described above in PELE Settings for Calculating Binding Free Energies with MSM). Afterwards, we divided the conformational space into 600 coarse-grained states (often called microstates), by clustering the different ligand's centers of mass using *k*-means. Then, the transition probabilities and the stationary distribution were estimated at a lag time that ensured Markovian behavior (i.e., memoryless), so that the MSM framework was valid. The potential of mean force,  $G_{\text{pmf}}$ , was obtained for the *i*th state by using  $G_{\text{pmf}}(i) = -k_{\text{B}}T \times \log(\pi_i)$ , where  $k_{\text{B}}$  is the Boltzmann constant, *T* is the temperature, and  $\pi_i$  is the stationary distribution of the *i*th state. We obtained the binding free energy,  $\Delta G$ , as seen in Takahashi et al. (32).

One can extend the MSM analysis to study the binding mechanism. To do so, it is convenient to lump the microstates into larger states (often called metastable states, or macrostates), allowing us to further coarse-grain the original MSM and obtain a simpler picture of the binding process. We used Perron-cluster cluster analysis (PCCA+) (33), implemented in the software EMMA 1.3 (Free University of Berlin, Berlin, Germany) (34), and lumped the original 600 clusters into seven metastable states. To

highlight differences between ligands, we studied the transition probabilities for the different metastable states, not counting internal cluster transitions. Then, transition path theory (35) was used to study the main binding pathways. We computed the normalized fluxes for all the main pathways (i.e., with a flux >4%), and the committer probability, which is the probability of reaching the binding site from the different metastable states, before going back to the bulk.

### Fluorescence polarization ligand binding assays (used for AR, ER $\alpha$ , GR, and PR)

Competition binding studies were performed using assay kits (PanVera, Madison, WI) for AR, ER $\alpha$ , GR, and PR, and a PolarScreen Competitor Assay kit (Life Technologies/Thermo Fisher Scientific, Carlsbad, CA) for PR. The receptor and fluorophore in the different assays forms a complex that gives a high polarization value, while the presence of a competitor prevents the formation of a complex, resulting in a decrease of the polarization. The shift in polarization value in the presence of test compound is used to determine relative affinity for the receptor. The proteins used in the assay kits were rat recombinant AR ligand binding domain (AR-LBD), human recombinant ER $\alpha$ , human recombinant GR, and human PR ligand-binding domain (PR-LBD). Compounds dissolved in DMSO were tested in black 384-well low-volume, nonbinding-surface glass plates (Corning, Corning, NY; PanVera AR and PR assays) or black 384-well small-volume, medium-binding Greiner plates (PanVera ER $\alpha$  and GR assays) in five-point concentration response, five-times dilution steps. In the Life Technologies PR assay, compounds were tested in 10-point concentration response, 1/2 log serial dilution. In the PanVera AR, ER $\alpha$ , and PR assays 15  $\mu$ L of either AR-LBD/Fluormone AL Green, ER $\alpha$ /Fluormone EL Red, or PR-LBD/Fluormone PL Red (Thermo Fisher Scientific, Carlsbad, CA) was added to 200 nL test or control compound already present in the well. This was followed by a 4–6 h (AR), 1–5 h (ER $\alpha$ ), or 1–6 h (PR) incubation in darkness at room temperature (RT). Final assay concentrations were 1.3% DMSO, 12.5 nM AR-LBD/0.5 nM Fluormone AL Green, 7 nM ER $\alpha$ /0.5 nM Fluormone EL Red, or 18 nM PR-LBD/1 nM Fluormone PL Red (Thermo Fisher Scientific). In the GR assay, 7  $\mu$ L GR/Stabilizing peptide mix was added to assay plates with 1  $\mu$ L test or control compound followed by 7  $\mu$ L Fluormone GS Red. Plates were incubated in darkness for 2 h at RT. Final assay concentrations were 6.7% DMSO, 4 nM GR, 1  $\times$  stabilizing peptide, and 1 nM Fluormone GS Red.

In the Life Technologies PR assay (used for comparing to MSM results), 5  $\mu$ L of PR Fluormone Red was added to assay plates with 50 nL test or control compound, followed by 5  $\mu$ L PR-LBD and a 2 h incubation in darkness at RT. Final concentrations were 2 nM PR Fluormone Red/150 nM PR. The PanVera assay plates were read on an Analyst AD plate reader (LJL Biosystems, Sunnyvale, CA) with Ex 485/Em 530 (AR) or Ex 530/Em 590 (ER $\alpha$ , GR, and PR), while the Life Technologies PR assay plates were read on a PHERAstar Plus (BMG Labtech, Ortenberg, Germany) using a fluorescence polarization (FP) optic module (540/590/590). Data from PanVera assays was analyzed in the software ActivityBase (ID Business Solutions, Guildford, UK) and with the Life Technologies PR assay in Genedata Screener (Genedata, Basel, Switzerland). IC<sub>50</sub> values were calculated using a four-parameter logistic fit.  $K_i$  values for the Life Technologies PR assay were calculated using the Cheng-Prusoff equation ( $K_i = IC_{50}/(1 + c(\text{ligand})/K_d)$ ) with  $K_i(\text{PR}) = 0.019 \mu\text{M}$  and  $c(\text{ligand}) = 0.002 \mu\text{M}$ .

### Scintillation proximity assay ligand-binding assay (used for MR)

A 96-well format scintillation proximity assay (SPA) was used to identify compounds that show binding to the human mineralocorticoid receptor ligand binding domain (MR-LBD). The immobilization of the fusion protein to the scintillation beads was done via rabbit MBP (maltose-binding protein) antibodies that were captured by the anti-rabbit SPA PVT (Polyvi-

nyltoluene) beads. The inhibition of the scintillation signal by displacement of <sup>3</sup>H-aldosterone with test compounds was measured on a MicroBeta TriLux instrument (Wallac/PerkinElmer, Waltham, MA). Compounds were tested in five-point concentration response (five-step dilution) and the assay was run in 96-well Corning white/clear NBS plates. Assay buffer contained 100 mM Tris pH 7.5, 0.1 mM EDTA, 20 mM NaMoO<sub>4</sub>, 10% Glycerol, and 0.1 mM DTT. A MR/<sup>3</sup>H-aldosterone solution was prepared by mixing 14.2  $\mu\text{g}/\text{mL}$  MBP-tagged human MR-LBD coinfecting with p23 lysate (MBP-MR-LBD/p23) with 10.1 nM <sup>3</sup>H-aldosterone (Amersham, Little Chalfont, UK). To assay plates with 1  $\mu$ L test or control compound dissolved in DMSO (final concentration 1%), 50  $\mu$ L of MR/<sup>3</sup>H-aldosterone solution was added and incubated for 1 h on a shaker (<100 rpm). SPA imaging beads (Amersham) were dissolved to a concentration of 5 mg/mL and mixed with 4  $\mu\text{g}/\text{mL}$  anti-MBP antibody (Abcam, Cambridge, UK) in assay buffer and 50  $\mu$ L of the mix was added to assay plates. Plates were incubated for 3–6 h at RT before being read on a MicroBeta TriLux instrument (Wallac/PerkinElmer). The raw data output was analyzed in ActivityBase (ID Business Solutions) using a four-parameter logistic fit to calculate IC<sub>50</sub> values.

## RESULTS AND DISCUSSION

### Exploiting experimental information: PCA of x-ray structures

Exploration of ligand binding pathways in MR and GR revealed an intrinsic plasticity of the helix 6–7 region (16). Protein perturbation in PELE is based on the ANM method, which should provide a good description of the intrinsic overall protein motion using low-frequency modes. However, small local rearrangements, such as the ones often derived from ligand-induced protein movement, might not always be well represented by low-frequency modes (25,26,36,37). In contrast to ANM, PCA provides the essential movements that describe the ensemble of (experimental) structures (38–41). If the required flexibility is included in the reference ensemble, in our case a large set of x-ray structures, this motion can be included by PCA modes. As can be seen in Figs. 2 and S1, for all the systems the general shape and magnitude of the displacement vectors are similar for the main modes of ANM and PCA. There are, however, key differences, especially in the helix 6–7 region highlighted in green. Further differences belong to helix 3 (760<sup>MR</sup>), helix 9–10 (915<sup>MR</sup>), and helix 11–12 (950<sup>MR</sup>).

To investigate the impact of experimentally derived motion modes, we applied PCA modes in PELE's protein perturbation step, and compared the results to those of using standard ANM, as shown below in Influence of Protein Dynamics and Helix 6–7 Fold on the PELE Simulation.

### Ligand binding site and ligand entry exploration

To address if all steroid receptors share the same entry pathway, we investigated the ligand entry mechanisms of their corresponding endogenous ligand (AR, testosterone; ER $\alpha$ , estradiol; GR, cortisol; MR, aldosterone; PR, progesterone). In the unbiased search, where simulations explored the complete protein surface (see Fig. 3 F, blue surface),



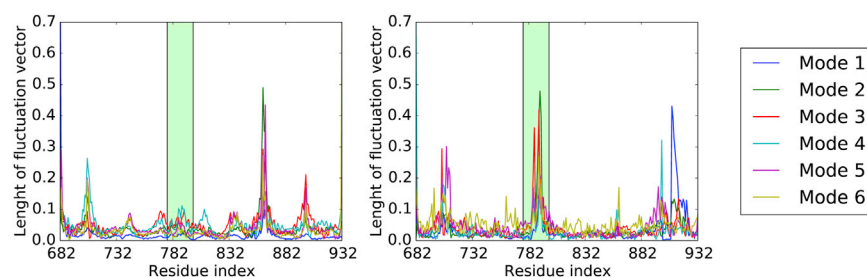


FIGURE 2 Visualization of the per-residue displacement magnitude of the lowest six ANM modes (*left*) and first six PCA modes (*right*) for PR. (*Green*) Helix 6–7 loop region (residues 833–853 for MR). To see this figure in color, go online.

entry events are only observed in the region where helices 3, 7, and 11 meet. This is in agreement with the previous studies of ligand entry for GR and MR (16). For ER $\alpha$ , GR, MR, and PR, a few trajectories completely enter the binding site and result in structures close to the experimentally observed structure. In contrast, a complete binding event is not observed for AR, even though the ligand explores the entry region in a similar manner as the other receptors.

To enhance the sampling of binding events and increase the statistics, additional simulations were performed where the search radius was restricted to a 20 Å radius around the identified entry site. For all receptors, the endogenous ligand enters the receptor at the same position between helices 3, 6–7, and 11, despite other potential entry pathways being possible, such as through helix 11–12. Representative entry paths are shown in Fig. 3. The covered search space (Fig. 3 F, *green surface*) shows that although the whole pro-

tein surface is no longer explored, the exploration would still allow different possible entry paths, e.g., through helix 11–12; but this is not observed. As it can be seen in Figs. 4 and S2, for each receptor the experimental ligand pose was reproduced in the simulations (ligand heavy atom RMSD values at  $\leq 1.0$  Å). In addition, the low RMSD structures generally also possess the lowest ligand binding energy, except for ER $\alpha$ , where the binding pose cannot be clearly identified with the binding energy. It is interesting to note that the ER $\alpha$  x-ray structure exhibits a specific helix-12 conformation that opens up the binding site (see below for a more complete analysis of the fold difference), and could potentially influence the binding energy. By employing the enhanced sampling around the helix 3, 6–7, and 11 regions, ligand entries are also observed for AR, analogous to the other receptors. However, in keeping with the results of the full protein exploration, AR shows fewer entry events than any other receptor (see Table 1).

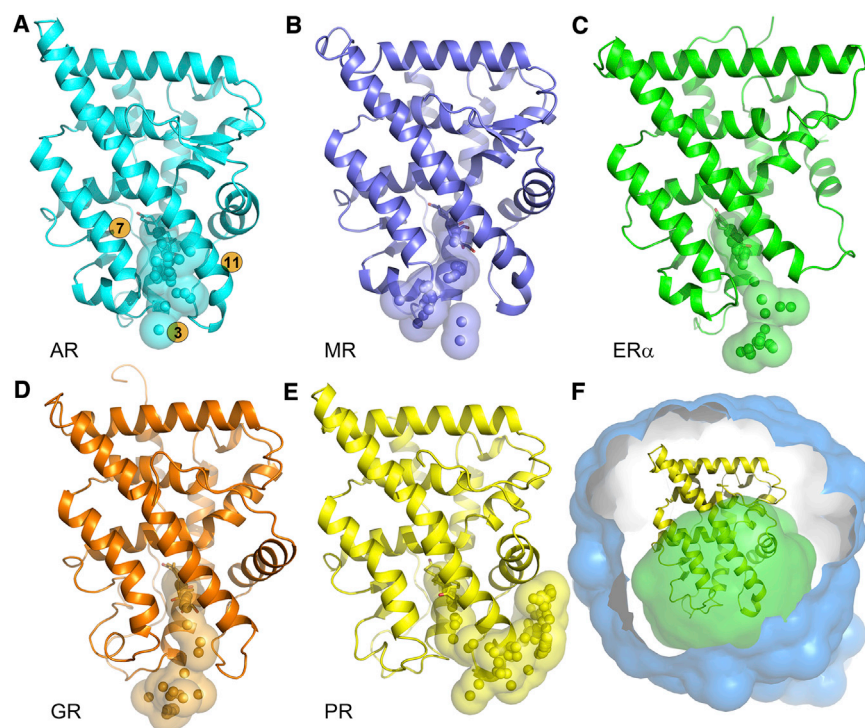


FIGURE 3 Entry paths obtained from PELE simulations. The spheres and surfaces represent ligand center-of-mass position in the entry trajectories. (A) AR, (B) MR, (C) ER $\alpha$ , (D) GR, and (E) PR. (F) Covered search space in the binding site exploration is shown as surface representation. (*Blue surface*) Full receptor exploration; (*green surface*) reduced local search space around the entry region. To see this figure in color, go online.

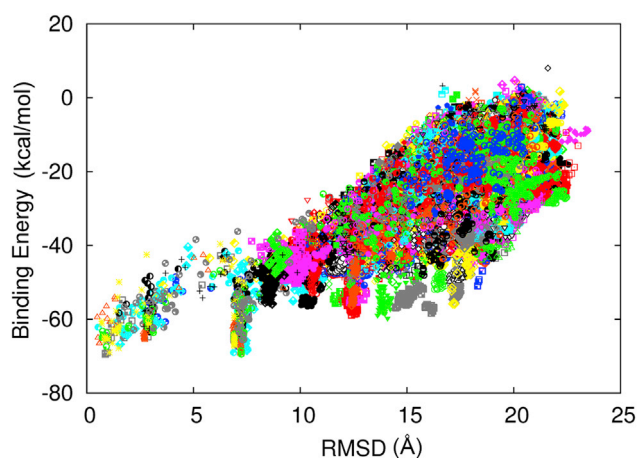


FIGURE 4 Results for PELE simulations using PCA modes for PR. It shows the correlation of the ligand heavy atom RMSD to the bound crystal (in Ångstroms), and the binding energy (in kilocalorie per mole). Each color and symbol corresponds to an independent trajectory from the PELE sampling. To see this figure in color, go online.

### Investigation of entry path and mechanism

While the entry pathways are consistent across all receptors, simulations yield different numbers of binding events (see Table 1). Furthermore, the PCA analysis reveals different degrees of flexibility (see Figs. 2 and S1). To build an understanding of how the number of entry events is influenced by the protein structure and dynamics, protein conformations are compared in more detail, and results from PCA-based simulations are directly compared to results from simulations using ANM.

#### Conformational pairing of helix 6–7 loop

When studying the steroid hormone complex structures in more detail, it is evident that the helix 6–7 loop differences largely pair-up into two different folds (see Fig. 5). In AR and MR, the helix 6–7 loop possesses a closed conformation with respect to the openness of the entry site. This can be defined by the distance between helix 11 and the helix 6–7 loop, in the range of 9–11 Å, which will be referred to as the “closed fold”. In contrast, the loop region is significantly more open in the PR and GR structures, in the ~15 Å range, showing a partially unstructured helix 6. Furthermore, helix 7 is extended and longer than in AR and MR. This will be referred to as the “open fold”. ER $\alpha$  belongs to the closed fold with respect to the distance of helix 6–7 and helix 11 (8 Å). However, helix 7 is slightly extended

TABLE 1 Number of Trajectories

| System                       | AR | MR | ER $\alpha$ | GR | PR |
|------------------------------|----|----|-------------|----|----|
| Number of entry trajectories | 2  | 5  | 19          | 13 | 14 |

Given for where the ligand entered the binding site in each set of 400 independent trajectories for free ligand binding-site exploration runs, with reduced search space of 20 Å.

and helix 6 is partially unfolded. Furthermore, helix 12 shows a completely different folding compared to the other receptors, which makes the binding site intrinsically more open. Looking at other x-ray structures of ER $\alpha$ , many of them have no structural information about helix 12 at all, pointing to a very flexible region, and few of these show a structure similar to those of the other receptors.

Thus, the separation between closed and open folds defines two structurally different gateways to the binding pockets: the binding pockets of GR, PR, and also ER $\alpha$  (due to the unfolded helix 12) should be easier to enter than those of AR and MR. This trend is largely represented in the number of entries highlighted in Table 1 as ER $\alpha$  and PR show many more entries than AR and MR. However, the influence of the fold of the receptors can only be directly compared when the same ligand is used for entry studies, thus eliminating effects coming from different molecular properties of the ligands.

Therefore, we further performed ligand entry simulations of progesterone to the five different receptors as being the least hydrophilic compound and thus minimizing the effect of polar interactions toward the receptors. The receptors belonging to the closed fold (AR and MR) should show fewer entries than the structures of the open fold (PR and GR) as well as ER $\alpha$  (as helix 12 is unfolded). Looking at the relative number of entries and sampled structures inside the receptor binding site (Fig. 6 A), it can clearly be seen that this trend is reproduced. When looking at the same ligand, the ligand binding energy can be compared as well. The correlation of the ligand binding energy to experimental pIC<sub>50</sub> values (Fig. 6 B) clearly shows that the binding energy obtained by PELE can be used to quickly score one ligand for different receptors.

#### Influence of protein dynamics and helix 6–7 fold on the PELE simulation

The use of PCA modes based on experimental x-ray structures allows us to include experimental information in the simulation protocol and to sample the observed protein dynamics. Furthermore, it allows us to investigate the importance of different conformational flexibility in specific regions being important for the entry mechanism, as ANM and PCA modes can differ (as shown above). Therefore, results from simulations with PCA modes are compared to those of simulations with ANM modes; Table 2 gives an overview of the results of the simulations.

ER $\alpha$  shows significantly more flexibility at the entry region in the PCA modes than in the ANM modes. Although helix 12 possesses a more open fold, which can intrinsically facilitate the entry, simulations with PCA modes clearly show more entrances than the ANM-based simulations. Furthermore, the entrance occurs faster when using PCA modes (fewer steps needed until entry).

For PR, the changes are expected to be larger as the difference between ANM and PCA is highest. Indeed, only a

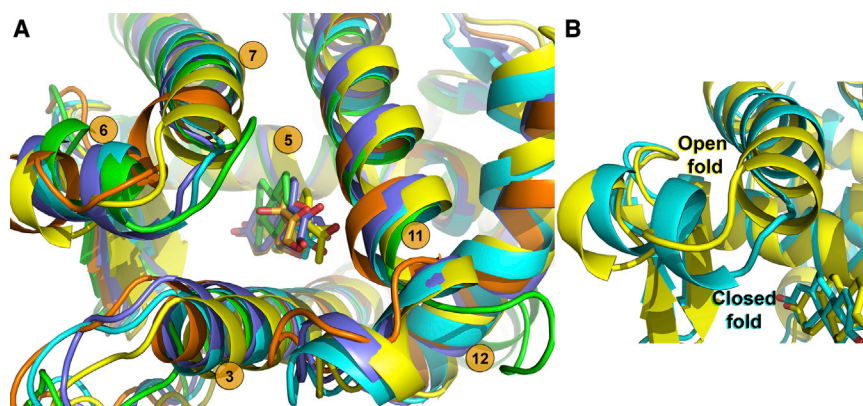


FIGURE 5 (A) Ligand binding pocket from the crystal structures of the AR (cyan), ER $\alpha$  (green), GR (orange), MR (ice blue), and PR (yellow) in complex with the endogenous ligand (testosterone, estradiol, cortisol, aldosterone, and progesterone, respectively). Helices numbered according to Bourguet et al. (42). (B) Close-up view of the helix 6 and 7 regions, showing the main differences between closed fold and open fold at AR (cyan) and PR (yellow). To see this figure in color, go online.

few entries were observed when using ANM modes while the ligand enters much more frequently in the PCA-based simulations (around five times more often). For both AR and MR, it was not possible to observe ligand entries with ANM when using small ligand translational steps. On the contrary, the PCA modes facilitate the entry. The effect is very small for AR, which can be expected from comparing the ANM and PCA modes where no significant flexibility is present for the entry region.

As the protein structures of MR and AR are very similar, but the PCA modes show significant differences (much higher flexibility in the helix 6–7 region for MR, Fig. S1), we applied the PCA-motion modes of MR to AR, thus artificially introducing the helix 6–7 flexibility observed in MR into AR. Using these modes increases the number of entries to values similar to those observed for MR, and allowed much faster entries.

The difference of the ANM and PCA modes for GR is not as high as, e.g., PR. Nevertheless, when using PCA modes (with a steering of nine steps), many more entries can be observed. Although GR possesses an already more open conformation of the helix 6–7 loop, very few entries are observed for ANM. This indicates that the open fold of helix 6–7 is not enough to allow ligand entry; proper treatment of protein flexibility is required.

The results show that the employment of modes based on PCA of experimental x-ray structures can clearly improve

the performance of the ligand-protein sampling if important protein flexibility is included and covered by the mode set used.

### Binding free energies for PR

While PELE's binding energy (computed as a protein-ligand interaction energy) is very useful to discriminate poses (Fig. 4) and compare receptors (Fig. 6 B), the extensive but fast sampling obtained with PELE makes it possible to further predict absolute binding free energies using MSM (32). As a model system, we explored the binding free energies for progesterone, cortisol, and aldosterone to PR, as PR provides the highest number of entries and thus is expected to show the best performance in MSM. The resulting predicted values are compared to binding free energies obtained from experimental  $K_i$  values (obtained by the Life Technologies PR assay; see the Materials and Methods). An overview of the absolute and relative binding free energies is given in Table 3, and a graphical visualization of the results in Fig. 7. While absolute values are slightly reduced, possibly due to the limited solvent exploration (see the Materials and Methods), the relative binding free energies correlate very well, as all ligands share an entry point and binding site in PR. Detailed results including implied timescales, the Chapman-Kolmogorov test, and convergence tests for MSM, can be found in the Supporting Material.

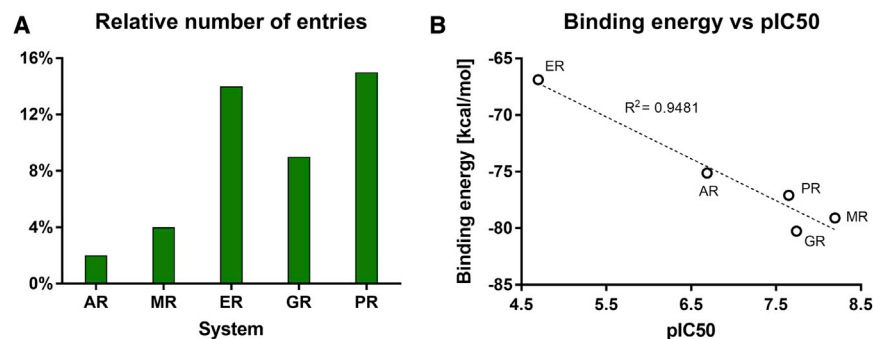


FIGURE 6 Results for progesterone binding to all five receptors. (A) Relative number of entries; (B) correlation plot for PELE ligand binding (interaction energy and pIC50 values). To see this figure in color, go online.



**TABLE 2** Percentage of Structures

|     | AR         | AR (MR-modes) | ER $\alpha$ | GR          | MR        | PR         |
|-----|------------|---------------|-------------|-------------|-----------|------------|
| ANM | —          | —             | 0.3% (219)  | 0.02% (196) | —         | 0.2% (189) |
| PCA | 0.1% (268) | 0.3% (162)    | 0.8% (117)  | 0.3% (38)   | 0.4% (56) | 1.0% (139) |

This is relative to all sampled structures, with SASA < 0.1 (i.e., inside the entry site or binding pocket) and averaged number of accepted Monte Carlo steps until entry, in parentheses, for ANM- or PCA-based protein motion modes (steering factors: AR, 9; AR with MR-modes, 7; ER $\alpha$ , 1; GR, 9; MR, 7; and PR, 2).

We should underline that in this study, to our knowledge, we introduced a new approach to improve convergence. First, we run a local exploration with larger ligand translations and rotation, aiming at finding a nonbiased binding event. Then, we selected six representative snapshots (covering the SASA space along the entrance event) and performed an additional PELE exploration where we place ~100 processors in each initial structure. This MSM exploration used small ligand translations and rotations, avoiding nonphysical transitions between states. Overall, convergence is significantly improved, allowing us to quantitatively score few ligands in a faster manner (convergence is already achieved after ~12 h and 300 cores for this system).

Finally, we used PCCA+ to analyze the ligand binding mechanisms. Fig. 8 shows the five main metastable states that are common to all three ligands: the A (red) cluster in the bulk solvent; the D (green), B (orange), and E (pink) clusters in the receptor surface; and the C (blue) cluster representative of the active site bound complex. Interestingly, the B cluster is located at the surface entrance site, and it largely resembles a peripheral binding site seen in a crystal structure of GR occupied by a steroid-like molecule that is part of the crystallization condition (16). Fig. 8 also shows the transition probabilities between each metastable state (excluding internal transitions) along the MSM simulations for the three ligands. Clearly, all ligands enter the binding cavity by the B peripheral binding site. Also, there is a clear correlation of the B  $\rightarrow$  C (binding) transition probability with the hydrophobicity of the ligand: 62% for progesterone (the most hydrophobic ligand), 29% for aldosterone, and only 0.1% for cortisol. In addition to the larger transition probability, we find the average residence time (MC steps) for progesterone in site B, 212 steps, to be significantly smaller than the values for aldosterone, 650 steps, and cortisol, 410 steps, indicating an overall faster binding

**TABLE 3** Experimental and Predicted Binding Free Energies for Ligand Binding to PR

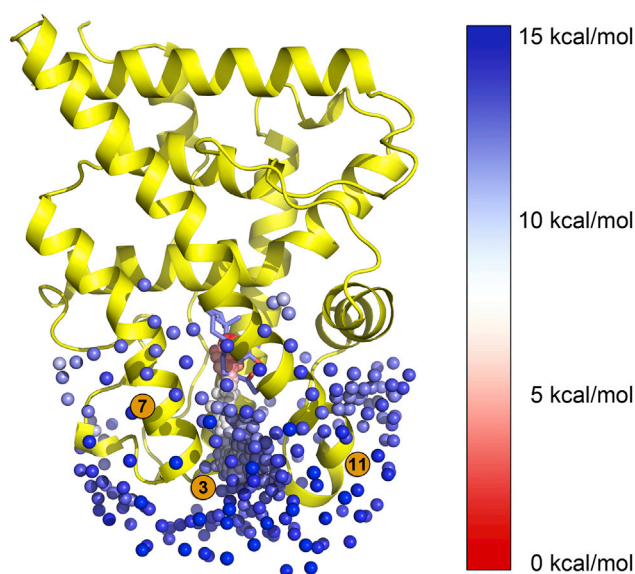
| [kcal/mol]             | Progesterone   | Aldosterone    | Cortisol       |
|------------------------|----------------|----------------|----------------|
| $\Delta G$ (exp)       | -10.0, -9.7    | -8.9, -8.8     | -7.2, -7.1     |
| $\Delta G$ (MSM)       | $-9.1 \pm 0.4$ | $-8.4 \pm 0.4$ | $-7.3 \pm 0.7$ |
| $\Delta\Delta G$ (exp) | 0.0            | 1.0            | 2.7            |
| $\Delta\Delta G$ (MSM) | 0.0            | $0.7 \pm 0.6$  | $1.8 \pm 0.8$  |

Experimental results are calculated from (two different)  $K_i$  values obtained with FP ligand binding assays, and calculated using the Cheng-Prusoff equation at 294 K ( $K_i$ (PR-progesterone), 36, 64 nM;  $K_i$ (PR-aldosterone), 264, 233 nM; and  $K_i$ (PR-cortisol), 4939, 4060 nM). Results for MSM simulations are averaged using 600 trajectories and 600 clusters.

mechanism for progesterone. Besides, the larger hydrophobicity of this ligand significantly increases transitions from the bulk to the protein surface site E (a significantly apolar site). In the other two (more polar) ligands, we observe more transitions to the hydrophilic D site (with charged residues such as Glu-126) and the dominance of transitions from the bulk to the peripheral binding site. Finally, the overall flux from the bulk solvent to the bound C state, together with the committor probability for each state, is also shown in Fig. S3, where we observe again the importance of the E state in reaching the peripheral binding site for progesterone.

## CONCLUSION

In this study, we performed a comprehensive analysis of the ligand entry mechanism in the steroid nuclear hormone receptor family. An initial full (receptor) exploration revealed one shared entry path through the helix 3, 7, and 11 regions. A refined exploration concentrating on the discovered entry region allowed us to localize ligand conformations with ligand heavy atom RMSD toward the x-ray structure of <1.0 Å, which correlated well to the binding energy. Furthermore, binding energies largely correlated with



**FIGURE 7** Clustering of the PELE trajectories used for MSM. Each sphere represents one ligand cluster, colored according to its binding free energy value. The image depicts aldosterone in complex to PR. To see this figure in color, go online.

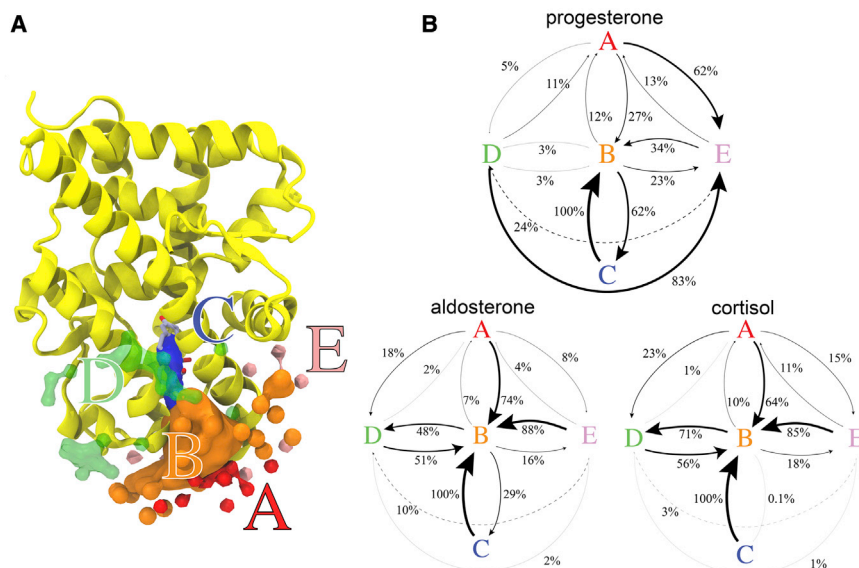


FIGURE 8 (A) Main metastable states (common to all ligands) obtained after PCCA+. (B) Transition probabilities between each metastable state where we excluded internal transitions. (Arrow sizes are proportional to the transition value.) To see this figure in color, go online.

experimental  $pIC_{50}$  values when comparing all five receptors and one ligand (progesterone).

Our findings, however, reveal differences in the number of entries for the different receptors. Analysis of the protein conformations shows two distinct loop foldings of the helix 6–7 region that can be classified as open and closed. Moreover, in line with previous studies, we observe that ligand entry in steroid nuclear hormone receptors is mainly driven by the flexibility observed at the helix 3, 7, and 11 regions. To better model the protein dynamics in these regions, we performed a PCA analysis on the existing bound crystal structures. Application of PCA-based motion modes can clearly improve the sampling performance and description of protein dynamics in cases where significant information is present in the underlying x-ray structure ensemble. The comparison of ANM and PCA modes, both quick and computationally inexpensive, may reveal important induced-fit movements of the protein. Therefore, the employment of experimentally based modes represents a reasonable and straightforward approach for exploiting experimental information in protein-ligand sampling.

The extensive sampling provided by PELE allowed the prediction of absolute binding free energies and binding mechanism using MSM. The predicted relative binding free energies for aldosterone, progesterone, and cortisol binding to PR show a very good correlation to experimental relative binding free energies and describes the correct trend of the three ligands. PCCA+ revealed the importance of a peripheral binding site and the hydrophobic nature of the ligand. Overall, our study suggests the important combination of two factors: 1) the flexibility at the helix 3, 7, and 11 regions, necessary for the ligand to enter the binding site cavity; and 2) the hydrophobic nature of the ligand, increasing the transitions between the peripheral and the active binding sites.

## SUPPORTING MATERIAL

Six figures and one table are available at [http://www.biophysj.org/biophysj/supplemental/S0006-3495\(17\)30160-1](http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30160-1).

## AUTHOR CONTRIBUTIONS

C.G., J.U., C.T., K.E., A.H., and V. Guallar designed research; C.G., D.L., V. Gil, J.U., C.T., K.E., A.H., and V. Guallar wrote the article; C.G. and V. Guallar performed simulations; V. Gil and C.G. performed PCA analysis; D.L. performed MSM analysis; and P.H. and A.D. performed binding affinity experiments.

## ACKNOWLEDGMENTS

This work was supported under grant No. SEV-2011-00067 of the Severo Ochoa Program, awarded by the Spanish Government.

## REFERENCES

- Sinko, W., S. Lindert, and J. A. McCammon. 2013. Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design. *Chem. Biol. Drug Des.* 81:41–49.
- Sotriffer, C. A. 2011. Accounting for induced-fit effects in docking: what is possible and what is not? *Curr. Top. Med. Chem.* 11:179–191.
- Csermely, P., R. Palotai, and R. Nussinov. 2010. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* 35:539–546.
- Boehr, D. D., R. Nussinov, and P. E. Wright. 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5:789–796.
- Beato, M., and J. Klug. 2000. Steroid hormone receptors: an update. *Hum. Reprod. Update.* 6:225–236.
- Wang, X. X., T. Jiang, and M. Levi. 2010. Nuclear hormone receptors in diabetic nephropathy. *Nat. Rev. Nephrol.* 6:342–351.
- Chuang, J.-C., J.-Y. Cha, ..., J. J. Repa. 2008. Research resource: nuclear hormone receptor expression in the endocrine pancreas. *Mol. Endocrinol.* 22:2353–2363.

8. Miyamoto, H., Y. Zheng, and K. Izumi. 2012. Nuclear hormone receptor signals as new therapeutic targets for urothelial carcinoma. *Curr. Cancer Drug Targets*. 12:14–22.
9. Conzen, S. D. 2008. Minireview: nuclear receptors and breast cancer. *Mol. Endocrinol.* 22:2215–2228.
10. Smith, A. G., and G. E. O. Muscat. 2005. Skeletal muscle and nuclear hormone receptors: implications for cardiovascular and metabolic disease. *Int. J. Biochem. Cell Biol.* 37:2047–2063.
11. Cheng, P. T. W., E. Kick, ..., D. J. Abraham. 2010. Nuclear hormone receptor medicinal chemistry. In *Burger's Medicinal Chemistry and Drug Discovery*. John Wiley, Hoboken, NJ, pp. 77–188.
12. Niewoehner, D. E., M. L. Erbland, ..., N. A. Morgan; Department of Veterans Affairs Cooperative Study Group. 1999. Effect of systemic glucocorticoids on exacerbations of chronic obstructive pulmonary disease. *N. Engl. J. Med.* 340:1941–1947.
13. Edwards, D. P. 2000. The role of coactivators and corepressors in the biology and mechanism of action of steroid hormone receptors. *J. Mammary Gland Biol. Neoplasia.* 5:307–324.
14. Tsuji, M. 2015. A ligand-entry surface of the nuclear receptor superfamily consists of the helix H3 of the ligand-binding domain. *J. Mol. Graph. Model.* 62:262–275.
15. Wagner, R. L., J. W. Apriletti, ..., R. J. Fletterick. 1995. A structural role for hormone in the thyroid hormone receptor. *Nature.* 378:690–697.
16. Edman, K., A. Hosseini, ..., V. Guallar. 2015. Ligand binding mechanism in steroid receptors: from conserved plasticity to differential evolutionary constraints. *Structure.* 23:2280–2290.
17. Borrelli, K. W., A. Vitalis, ..., V. Guallar. 2005. PELE: protein energy landscape exploration. A novel Monte Carlo based technique. *J. Chem. Theory Comput.* 1:1304–1311.
18. Grebner, C., J. Iegre, ..., C. Tyrchan. 2016. Binding mode and induced fit predictions for prospective computational drug design. *J. Chem. Inf. Model.* 56:774–787.
19. Kopečná, J., I. Cabeza de Vaca, ..., R. Sobotka. 2015. Porphyrin binding to Gun4 protein, facilitated by a flexible loop, controls metabolite flow through the chlorophyll biosynthetic pathway. *J. Biol. Chem.* 290:28477–28488.
20. Kotev, M., D. Lecina, ..., V. Guallar. 2015. Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques. *Biophys. J.* 108:116–125.
21. Carlson, H. A., R. D. Smith, ..., J. B. Dunbar. 2016. CSAR 2014: a benchmark exercise using unpublished data from pharma. *J. Chem. Inf. Model.* 56:1063–1077.
22. Pande, V. S., K. Beauchamp, and G. R. Bowman. 2010. Everything you wanted to know about Markov state models but were afraid to ask. *Methods.* 52:99–105.
23. Bowman, G. R., V. S. Pande, and F. Noé. 2014. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer, New York.
24. Madadkar-Sobhani, A., and V. Guallar. 2013. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Res.* 41:W322–W328.
25. Atilgan, A. R., S. R. Durell, ..., I. Bahar. 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80:505–515.
26. Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2:173–181.
27. Banks, J. L., H. S. Beard, ..., R. M. Levy. 2005. Integrated modeling program, applied chemical theory (IMPACT). *J. Comput. Chem.* 26:1752–1780.
28. Onufriev, A., D. Bashford, and D. A. Case. 2004. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins.* 55:383–394.
29. Sastry, G. M., M. Adzhigirey, ..., W. Sherman. 2013. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* 27:221–234.
30. Schrödinger LLC. 2014. Maestro: Release 2014–2. Schrödinger, New York.
31. Bakan, A., L. M. Meireles, and I. Bahar. 2011. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics.* 27:1575–1577.
32. Takahashi, R., V. A. Gil, and V. Guallar. 2014. Monte Carlo free ligand diffusion with Markov state model analysis and absolute binding free energy calculations. *J. Chem. Theory Comput.* 10:282–288.
33. Deuffhard, P., and M. Weber. 2005. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* 398:161–184.
34. Senne, M., B. Trendelkamp-Schroer, ..., F. Noé. 2012. EMMA: a software package for Markov model building and analysis. *J. Chem. Theory Comput.* 8:2223–2238.
35. Noé, F., C. Schütte, ..., T. R. Weikl. 2009. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA.* 106:19011–19016.
36. Bakan, A., and I. Bahar. 2009. The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. *Proc. Natl. Acad. Sci. USA.* 106:14349–14354.
37. Xu, C., D. Tobi, and I. Bahar. 2003. Allosteric changes in protein structure computed by a simple mechanical model: hemoglobin T ↔ R2 transition. *J. Mol. Biol.* 333:153–168.
38. Balsera, M. A., W. Wriggers, ..., K. Schulten. 1996. Principal component analysis and long time protein dynamics. *J. Phys. Chem.* 100:6.
39. Maisuradze, G. G., A. Liwo, and H. A. Scheraga. 2009. Principal component analysis for protein folding dynamics. *J. Mol. Biol.* 385:312–329.
40. Meireles, L., M. Gur, ..., I. Bahar. 2011. Pre-existing soft modes of motion uniquely defined by native contact topology facilitate ligand binding to proteins. *Protein Sci.* 20:1645–1658.
41. Peng, J., and Z. Zhang. 2014. Simulating large-scale conformational changes of proteins by accelerating collective motions obtained from principal component analysis. *J. Chem. Theory Comput.* 10:3449–3458.
42. Bourguet, W., M. Ruff, ..., D. Moras. 1995. Crystal structure of the ligand-binding domain of the human nuclear receptor RXR- $\alpha$ . *Nature.* 375:377–382.

## **Publication 5 - Adaptive simulations, towards interactive protein-ligand modeling**

**Authors:** Daniel Lecina, Joan Francesc Gilabert, Victor Guallar

**Journal:** Scientific Reports (In Revision)

### **Summary:**

In this work, we propose a novel procedure to overcome the sampling limitations caused by metastability. In particular, our methodology combines an adaptive machine learning protocol with PELE in the frame of modern multi-core computational resources, and is able to map complex binding mechanisms with at least a speedup gain of one order of magnitude compared to non-adaptive executions of PELE. The methodology was tested with complex systems, where standard sampling protocols have not succeeded, such as GPCRs and NHR, which suggests the potential of the technique in screening and lead optimization studies.

### **Author contribution:**

My contribution comprised the design and analysis of the algorithm, running simulations as well as writing the manuscript.

## Adaptive simulations, towards interactive protein-ligand modeling

Daniel Lecina<sup>1</sup>, Joan Francesc Gilibert<sup>1</sup> and Victor Guallar<sup>1,2</sup>

<sup>1</sup>Barcelona Supercomputing Center, Joint BSC-CRG-IRB Research Program in Computational Biology, Jordi Girona 29, E-08034 Barcelona, Spain

<sup>2</sup>ICREA, Passeig Lluís Companys 23, E-08010 Barcelona, Spain

### Abstract

Modeling the dynamic nature of protein-ligand binding with atomistic simulations is one of the main challenges in computational biophysics, with important implications in the drug design process. Although in the past few years hardware and software advances have significantly revamped the use of molecular simulations, we still lack a fast and accurate *ab initio* description of the binding mechanism in complex systems, available only for up-to-date techniques and requiring several hours or days of heavy computation. Such delay is one of the main limiting factors for a larger penetration of protein dynamics modeling in the pharmaceutical industry. Here we present a game-changing technology, opening up the way for fast reliable simulations of protein dynamics by combining an adaptive machine learning procedure with Monte Carlo sampling in the frame of modern multi-core computational resources. We show remarkable performance in mapping the protein-ligand energy landscape, being able to reproduce the full binding mechanism in less than half an hour, or the active site induced fit in less than 5 minutes. We exemplify our method by studying diverse complex targets, including nuclear hormone receptors and GPCRs, demonstrating the potential of using the new adaptive technique in screening and lead optimization studies.

Accurately describing protein-ligand binding at a molecular level is one of the major challenges in biophysics, with important implications in applied and basic research in, for example, drug design and enzyme engineering. In order to achieve such a detailed knowledge, computer simulations and, in particular, molecular *in silico* tools are becoming increasingly popular<sup>1, 2</sup>. A clear trend, for example, is seen in the drug design industry: Sanofi signed a \$120M deal with Schrödinger, a molecular modeling software company, in 2015. Similarly, Nimbus sold for \$1,200M its therapeutic liver program (a computationally designed Acetyl-CoA Carboxylase inhibitor) in 2016. Clearly, breakthrough technologies in molecular modeling have great potential in the pharmaceutical and biotechnology fields.

Two main reasons are behind the revamp of molecular modeling: software and hardware developments, the combination of these two aspects providing a striking level of accuracy in predicting protein-ligand interactions<sup>1, 3, 4</sup>. A remarkable example constitutes the seminal work of Shaw's group, where a thorough optimization of hardware and software allowed a complete *ab initio* molecular dynamics (MD) study on a kinase protein<sup>5</sup>, demonstrating that computational techniques are capable of predicting the protein-ligand binding pose and, importantly, to distinguish it from less stable arrangements by using atomic force fields. Similar efforts have been reported using accelerated MD through the use of graphic processing units (GPUs)<sup>6</sup>, metadynamics<sup>7</sup>, replica exchange<sup>8</sup>, etc. Moreover, these advances in sampling capabilities, when combined with an optimized force field for ligands, introduced significant improvements in ranking relative binding free energies<sup>9</sup>.

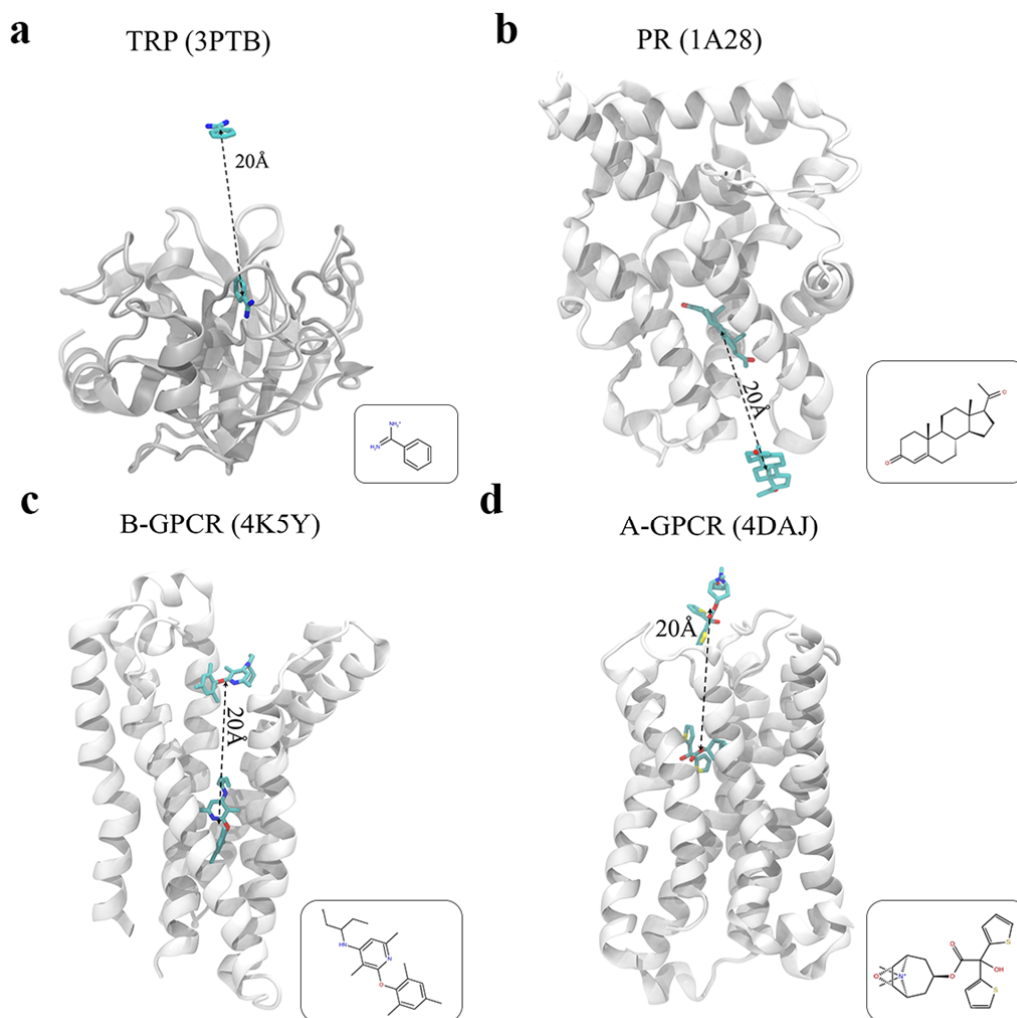
Despite these achievements, accurate (dynamical) modelling still requires several hours or days of dedicated heavy computation, being such a delay one of the main limiting factors for a larger penetration of these techniques in industrial applications. Moreover, this computational cost severely limits examining the binding mechanism of complex cases, as seen recently in another study from Shaw's group on GPCRs<sup>10</sup>. From a technical point, the conformational space has many degrees of freedom, and simulations often exhibit metastability: competing interactions result in a rugged energy landscape that obstructs the search, oversampling some regions whereas undersampling others<sup>11, 12</sup>. In MD techniques, where the exploration is driven by numerically integrating Newton's equations of motion, acceleration and biasing techniques aim at bypassing the highly correlated conformations in subsequent iterations<sup>13</sup>. In Monte Carlo (MC) algorithms, another main stream sampling method, stochastic proposals can, in theory, traverse the energy landscape more efficiently, but their performance is often hindered by the difficulty of generating uncorrelated protein-ligand poses with good acceptance probability<sup>14, 15</sup>. The Protein Energy Landscape Exploration (PELE) program<sup>16</sup> addresses the problem by making use of protein structure prediction algorithms, which introduces larger conformational changes<sup>17</sup> and, importantly, allows mapping complex protein-ligand binding mechanisms<sup>18, 19, 20</sup>. This technique, for example, has been underlined as an impressive accomplishment in the last Community Structure-Activity Resource (CSAR) blind competition<sup>21</sup>. Nonetheless, PELE simulations still show some degree of metastability, requiring several hours for solving the binding mechanism in complex systems, restricting its use in a drug design screening setup. For introducing large impact, we should aim for fast (minutes) and accurate simulations, allowing a drug design team to obtain accurate protein-ligand structures interactively, opening the possibility to combine their knowledge and expert intuition with *in silico* techniques on-the-fly. In this work, we present such a breakthrough tool: Adaptive-PELE, a combination of PELE with an adaptive machine learning procedure.

Of particular interest in our study are iterative methods making use of short simulations and deciding on-the-fly the most interesting regions to sample, such as adaptive sampling<sup>8</sup>, weighted ensemble<sup>22</sup>, the adaptive seeding method<sup>23</sup>, or the FAST<sup>24</sup> technique. The latter method rewrites the conformational exploration in terms of the well-studied multi-armed bandit (MAB) problem<sup>25</sup>, taking advantage of the gradient existing in measurables, such as the solvent-accessible surface area (SASA) or some energy components. We understand the ligand-protein exploration as an exploration-exploitation dilemma, since the phase space is highly dimensional



and sufficient sampling of relevant regions, not only of a few metastable states, is necessary for an accurate characterization. The exploration is a learning process where we acquire knowledge of the energy landscape as the simulation progresses, and we decide to focus on the most rewarding regions. We serve of the MAB as a theoretical framework, since it has been successfully applied in a wide range of problems such as protein folding<sup>24</sup>, on-line advertising or news recommendation<sup>26</sup>.

Adaptive-PELE is based on an iterative procedure where each iteration, referred as an epoch, involves three different steps: exploration, clustering and spawning (or seeding). Its landscape exploration capabilities, confronted with standard PELE executions (non-adaptive trajectories), are shown in four different protein-ligand complexes: i) the trypsin—benzamidine (TRP system); ii) a progesterone nuclear hormone receptor with its endogenous ligand (PR system), iii) the M3 muscarinic acetylcholine class A G-protein coupled receptor (GPCR) with an inverse agonist (A-GPCR system); iv) the corticotropin-releasing factor, a class B GPCR with an antagonist ligand (B-GPCR system). Our results demonstrate that the new adaptive technique is capable of mapping the binding energy landscape for complex systems in less than half an hour, or the active site induced fit process in less than 5 minutes.



**Figure 1 | Protein-ligand complexes studied.** (a) Trypsin with benzamidine as a ligand (TRP, PDB ID: 3PTB). (b) Progesterone nuclear hormone receptor with progesterone as a ligand (PR, PDB ID: 1A28). (c) Corticotropin-releasing factor GPCR with CP-376395 as a



ligand (B-GPCR, PDB ID: 4KY5). **(d)** M3 muscarinic acetylcholine GPCR with tiotropium as a ligand (A-GPCR, PDB ID: 4DAJ). The initial structures for the protein-ligand exploration, with the ligand  $\sim 20$  Å away from the binding site, are shown. The square inset in each panel depicts a 2D scheme of each ligand.

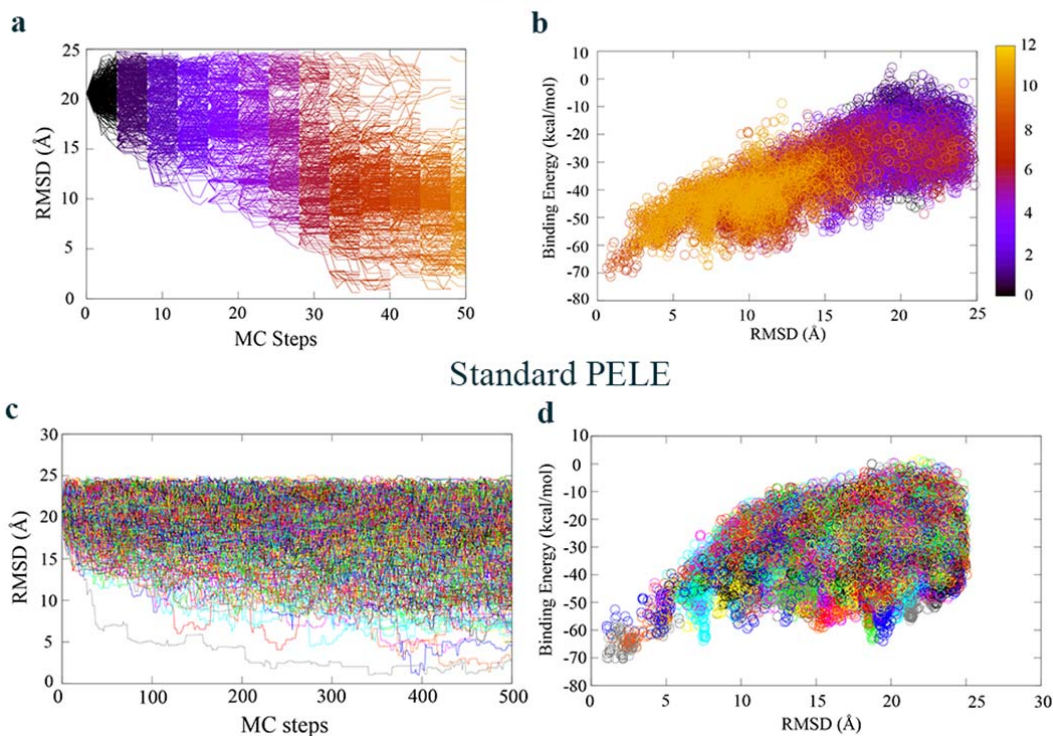
## Results

**Energy landscape exploration.** We first show the protein-ligand energy landscape exploration capabilities of Adaptive-PELE and compare them to that of a standard (non-adaptive) procedure. The evolution of the ligand root mean square deviation (RMSD) to the native bound structure along the simulation (MC steps), and the protein-ligand binding energy against the same ligand RMSD is shown in Fig. 2. We plot here the results for the B-GPCR system, using 512 trajectories (each trajectory runs in a computing core), but equivalent figures for the remaining systems are shown in the Supplementary Information. As seen in the RMSD evolution plots, both the adaptive (Fig. 2a) and standard (Fig. 2c) PELE methods succeed in sampling native-like conformations, with RMSD values  $\sim 1$  Å; analogous results are seen for all other systems (Supplementary Figs. 2 to 4). We should emphasize that the initial starting pose for the ligand is significantly away from the binding site ( $\sim 20$  Å, Fig. 1) and that there is no bias in the search: no information from the bound pose is used but for plotting purposes. Such a non-biased sampling performance, for example, has not been successful for MD techniques in complex systems such as the A-GPCR, only seeing binding to an extracellular site vestibule, approximately at 12 Å from the bound structure, when using 16  $\mu$ s of standard MD<sup>10</sup> or 1  $\mu$ s of accelerated MD<sup>27</sup>.

As we can see in Fig. 2a and 2b, the first phase of the adaptive simulation is devoted to explore the bulk and the vicinity of the initial pose. Significantly, as the adaptive epochs evolve few simulations enter deeper into the cavity, getting into an unexplored region. The MAB strategy uses this information to spawn several explorers there, increasing the possibilities of finding new unexplored areas. Towards the end of the sampling, we observe an almost complete shift of the explorers towards the binding site region. The standard PELE technique, however, keeps exploring the outer regions (Fig. 2c and 2d), with minimal excursions into the binding site, resulting in a much less efficient exploration (see below for a thorough comparison). A nice additional feature is that the exploration moves away from regions once they are sufficiently known, avoiding metastability. For example, the binding pose is found at around step 30, and the sampling is only kept there two more epochs, when exploration efforts are moved to more rewarding areas.

A noteworthy common aspect in both techniques is that we can easily identify the native-like pose using the binding energy. The potential of using PELE's binding energy, an all atom OPLS2005 protein-ligand interaction energy with an implicit solvent model, in pose discrimination was already shown in our initial induced-fit benchmark study<sup>28</sup>, being also the basis for our recent success in the CSAR blind competition. While this energy does not correlate with absolute experimental affinities (nor allows us to compare different ligands), it is very useful for pose discrimination; similar observations have emerged when using MD<sup>5</sup>. Importantly, introducing the adaptive procedure improves the binding energy landscape funnel shape, avoiding an unbalanced exploration of metastable regions, which eliminates the severe optimization on the energy by constantly minimizing over and over the same minimum. This can be seen, for example, when comparing the difference in "binding peaks" at 7.5 and 20 Å in Fig. 2b and 2d.

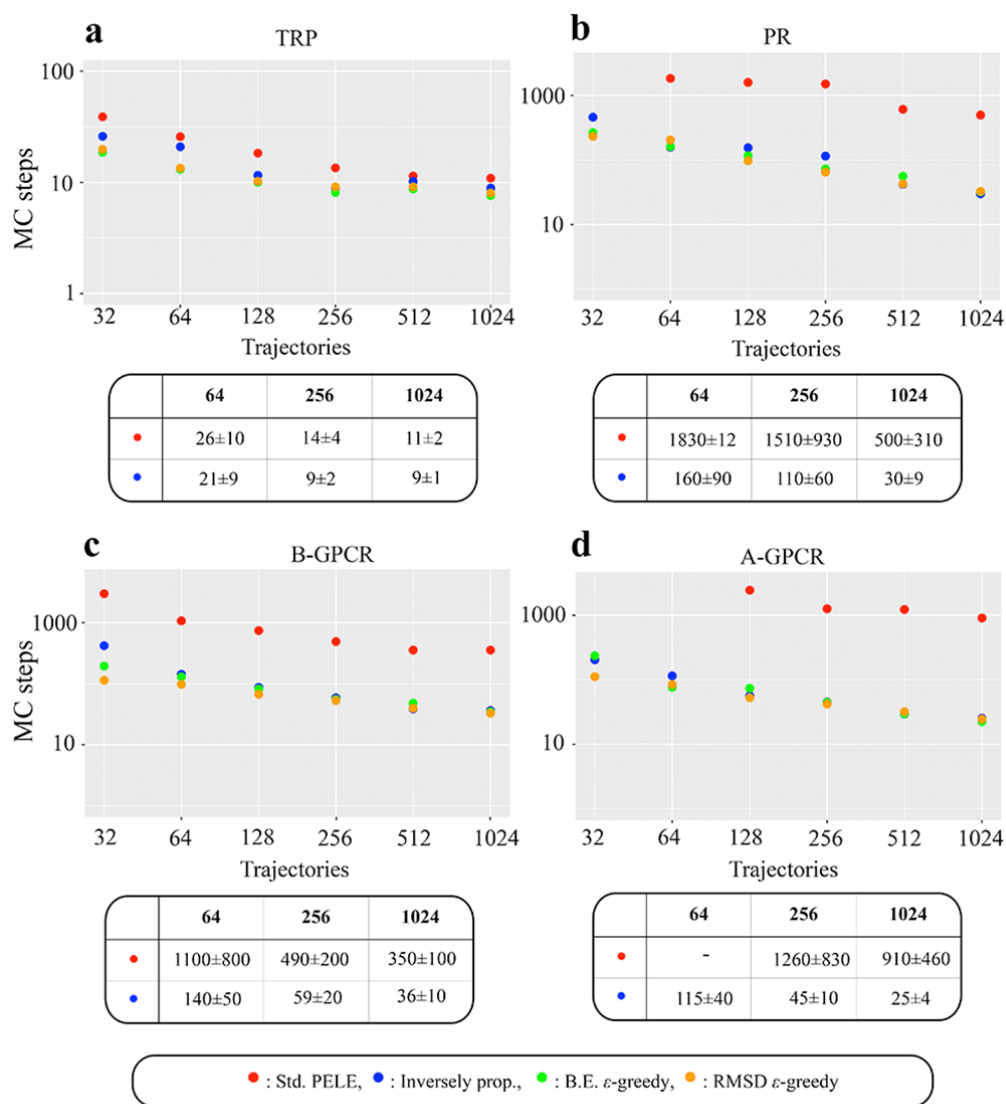
## Adaptive PELE



**Figure 2 | Energy landscape exploration of B-GPCR with 512 different explorers. (a,b)** The RMSD variation along MC steps and the binding energy against the RMSD for the adaptive results. Each color code corresponds to a different epoch number, for a total of 12 adaptive iterations. **(c,d)** Analogous plots for the standard executions. Each color corresponds to a different trajectory (performed in a different computing core). Notice the change in scale in the X-axis between **(a)** and **(c)**.

**Binding event observation - Binding time.** The ligand finds native-like poses in  $\sim 35$  MC steps when using the new adaptive approach (Fig 2a), the independent PELE simulation requiring approximately 10 more times,  $\sim 350$  steps (Fig. 2c). While standard PELE already represents a significant advance over other sampling techniques (microsecond MD simulations with the Anton computer, for example, could not observe a binding event for A-GPCR<sup>10</sup>), the adaptive scheme introduces a remarkable speed up. As a rule of thumb, each MC PELE step takes around 45 seconds on a SandyBridge-EP 2.6GHz computing core, and therefore, in this particular simulation the bound native structure can be predicted in under 30 minutes when using the adaptive approach.

To quantitatively assess our new algorithm's performance, we estimated the binding times by averaging over ten separate runs, considering that a binding event occurred when the ligand RMSD with the native bound structure was less than 2.5 Å. In addition, we checked the scalability by using an increasing number of trajectories (computing cores), from 32 to 1024, summing up to a total computing time of a quarter of million CPU hours. Moreover, different MAB strategies (see the Methods section) were used for the adaptive simulations, including the inversely proportional and  $\epsilon$ -greedy, guiding the exploration with two metrics: the protein-ligand interaction energy, where the native structure does not need to be known, and the ligand RMSD to the native, a biased strategy that allows us to estimate a lower bound for the binding time. Notice that when using a small number of explorers some standard PELE simulations did not produce binding events in 3000 MC steps. In those cases, we assigned the binding time to 3000 steps in order to set a lower bound for the comparison.



**Figure 3 | Binding times for all systems and MC techniques.** (a) Number of steps for observing a binding event against the number of trajectories (processors) for the TRP system, using the standard PELE (in red) and the adaptive-PELE with the inversely proportional (in blue) and the  $\epsilon$ -greedy guided strategies with binding energy (in green) and RMSD (in yellow). Actual data (MC steps) with their standard deviation for three different sets of processors is shown at the bottom table inset for the standard PELE and the inversely proportional adaptive-PELE methods. (b, c, and d) Analogous plots for PR, B-GPCR, and A-GPCR. A complete list of all data is shown in Supplementary Information.

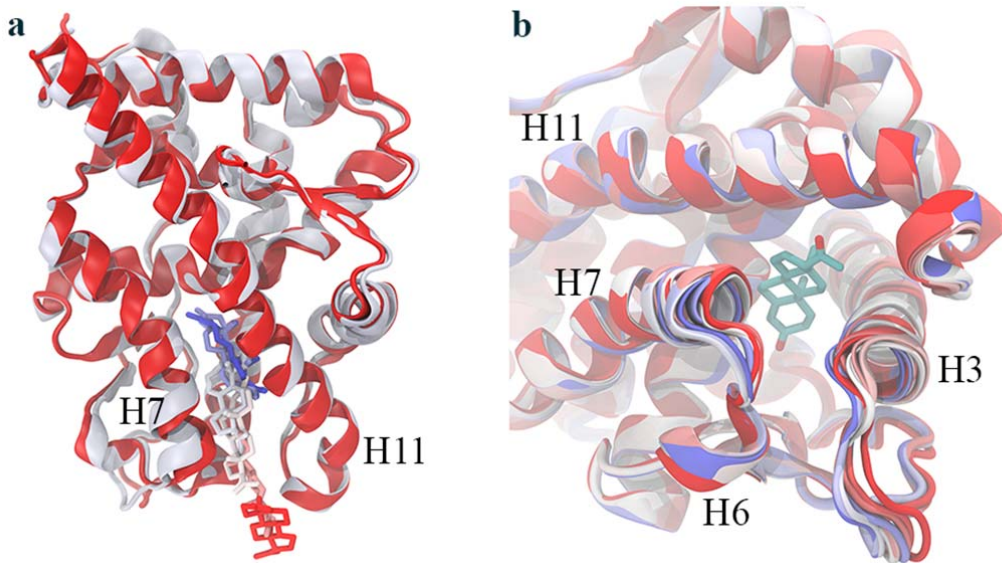
We observe that in general the binding time decreases with the number of processors for all systems and methods (Fig. 3). In TRP, however, we approach a plateau for 256 processors; adding up more explorers only yields minor improvements. TRP is a relatively rigid protein not requiring structural rearrangements to bind benzamidine, and using 256 processors we almost reach the minimum possible binding time, given the ligand translation range per MC step and the starting position. In the remaining three (more difficult) systems, however, the binding time keeps decreasing in the whole range, since we need a more exhaustive protein sampling, and ligand movements need to couple to protein rearrangements.

In agreement with the difficulties seen in MD simulations, the exploration in A-GPCR is especially poor for the standard PELE approach, not seeing a significant number of binding events with less than 128 trajectories. It is quite remarkable that by introducing the adaptive sampling we find the correct binding mode using 32 cores in only ~3 hours of simulation. The overall speed up achieved by adaptive-PELE for this system is approximately 40 times in the studied number of processors range, being at least one order of magnitude in the other two complex systems, PR and B-GPCR. As expected, TRP has the least speed up gain, since it is the least computationally demanding example. Importantly, for all studied systems the adaptive technique is capable of providing native-like poses in less than half an hour when a large number of computing cores is provided, a significant achievement.

Interestingly, the different MAB strategies perform quite similarly. Guiding the seeding with the protein-ligand binding energy does not require previous knowledge of the binding site and, as emphasized above, it correlates nicely with the native-like pose (although it has been reported that sometimes the SASA has been shown to perform better<sup>29</sup>). In addition, if one has available the bound crystal structure, one can use the RMSD to guide the binding, which serves as an estimation of the binding time limit that we could achieve; a similar strategy could be obtained by simply knowing the binding site and using its distance to the ligand's center of mass to guide the spawning. Surprisingly, when increasing the number of processors all these strategies yield similar results as our default option, the inversely proportional strategy, which seems to indicate that the choice of the reward function depending on the number of contacts (see Methods section) makes quite an optimal seeding.

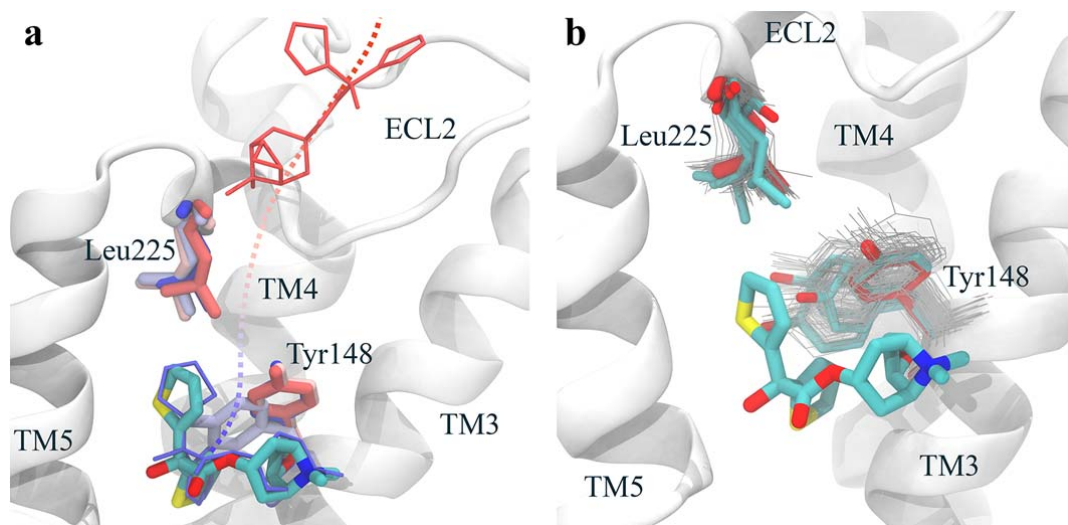
**Mechanistic studies: protein conformation exploration.** While we have shown that adaptive-PELE can provide native-like poses in complex systems in a fast manner, it is important to show that it also provides the proper binding mechanism. We show here the analysis for two of the more difficult systems, PR and A-GPCR.

PR. Recent crystallographic and computational studies in NHRs have underlined the conformational changes necessary for ligand delivery at the entry site: helices 3, 6, 7 and 11, along with the loops linked to them<sup>19, 30</sup>; with respect to this region, NHRs seem to adopt an open and a closed structure coupled to the ligand's entrance. The PR receptor, in particular, has the largest plasticity in this region, as shown in the PCA analysis on all available NHRs bound crystal structures<sup>30</sup>. Such conformational change is well captured by the adaptive technique. As seen in Fig. 4, the protein starts in the closed conformation (shown in red) and achieves its largest opening when the ligand starts entering the cavity from the peripheral binding site (shown in white), to progressively close again towards the native pose as it gets deemed bound (shown in blue).



**Figure 4 | PR binding mechanism.** Two different views of the ligand entrance and the plasticity upon progesterone binding in PR. **(a)** Different ligand snapshots along the binding with two protein structures highlighting the initial closed (red cartoon) and intermediate open states (white cartoon). **(b)** A closer zoom at the entrance region with the ligand shown in the native bound structure; same color-coding as in the (a) panel but for the ligand (shown with atom element colors).

A-GPCR. GPCRs represent a great challenge for the modeling community. On top to the difficulties in obtaining atomistic models for these membrane proteins, we have the large plasticity of their extracellular domain (involved in ligand delivery and binding), and the buried nature of most of their binding sites. For A-GPCR, in particular, the extracellular loop 2 (ECL2) mobility has been reported to be involved in ligand binding, where a movement of L225 away from the orthosteric site permits a transient opening (rotation) of Y148 towards TM4, allowing tiotropium to bind, which closes again to form a lid in the binding pose<sup>10</sup>. As shown in Fig. 5a, in our simulations, we see a movement of L225 that is accompanied by a dihedral rotation of Y148 towards TM4, which allows binding. Once the ligand is bound, the tyrosine and the leucine move back to generate the binding pose. In Fig. 5b, we show the plasticity of these two residues, grouping all the involved cluster center side chain structures (in grey lines) into four main clusters using the k-medoids (in colored licorice) implemented in pyProCT<sup>31</sup>.



**Figure 5 | A-GPCR binding mechanism.** (a) Different ligand snapshots showing the binding pathway from the initial structure (in red) to the bound pose (in blue), including Y148 and L225, which follow the same color-code. The white cartoon protein and the colored licorice ligand correspond to the bound crystal structure. (b) Side chain conformations for Y148 and L225, where the red licorice corresponds to the crystal structure. In grey lines, we show all the different conformations for those cluster centers along the adaptive process, and in colored licorice we show the resulting main conformations after a k-medoids clustering.

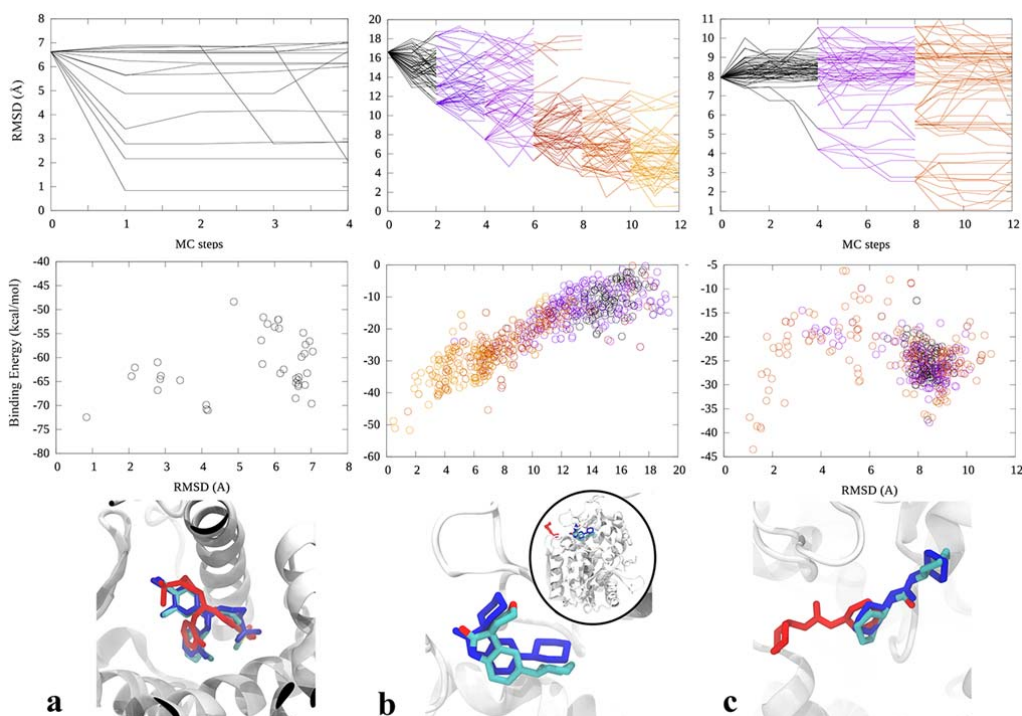
### Induced-Fit Docking

Predicting the non-biased binding mechanism is certainly a fancy computational effort, showing the capabilities of molecular modeling techniques. It aids in understanding the molecular mechanism of action, potentially finding, for example, alternative binding sites that might be used for rational inhibitor design. Another set of important simulations comprises docking refinement. Today, structure based design efforts ranging from virtual screening to fine tuning lead optimization activities, are hampered by having to properly handle the induced fit mechanisms. In this sense cross- and apo-docking studies, a significant less demanding modeling effort, constitute a better example. As seen in recent benchmark studies<sup>28, 29, 32</sup> (or in the CSAR exercise<sup>21</sup>), standard PELE is possibly the fastest technique providing accurate answers in cross- and apo-docking, requiring on the order of 30-60 minutes wall clock time using ~16/32 trajectories in average.

By introducing the adaptive sampling technique, we can now improve the simulation time to only few MC steps, as shown in Fig. 6, where we show the refinement of a wrong docked pose for the PR system and the application in cross docking for the soluble epoxide hydrolase (sEH), a tough benchmark system recently studied with standard PELE<sup>32</sup>. Notice that easy induced fit cases, such as PR requiring only a flip of the ligand, can be accomplished in one MC step, not representing any improvement from standard PELE. In difficult cases, such as for sEH, the adaptive scheme provides again significant improvement over standard simulations, shown in Supplementary Fig. 5. For example, notice in Supplementary Fig. 5a how standard PELE shows early non-productive low RMSD explorations (grey line achieving RMSD ~5 Å). This type of behavior motivated the development of the adaptive protocol.

Taking into account that the active site refinement MC steps require only 30 seconds (involving less protein perturbation and ligand translation, but more rotation), we can model the right pose in under 5 minutes using a modest computational cluster (32-64 processors), which allows refinement of a large number of docking poses or an interactive structural-guided optimization of a given lead.





**Figure 6 | Induced-fit docking studies.** (a) PR system: protein structure from PDB ID:1A28 and ligand structure from PDB ID:3KBA. (b) sHE system: protein structure from PDB ID:5AKE and ligand structure from PDB ID:5AM4. (c) sHE system: protein structure from PDB ID:5ALX and ligand structure from PDB ID:5AI5. In the upper panels we show the RMSD evolution along the simulation, in the middle ones the binding energy for the different RMSD values, and in the lower panels the native structure (atom-type colored), the lowest binding energy ligand structure (blue) and the starting ligand structure (red). Notice that in panel (b) the initial docking structure is slightly outside the active site (shown in the inset).

## Discussion

Breakthrough advances in software and hardware are shifting the development of complex design processes to computer modeling. Still, accurately modeling the protein-ligand structure requires several hours of heavy computation, even when using special purpose machines or large clusters of processors. We have introduced here a new method, combining a machine learning procedure with an all-atom molecular mechanics Monte Carlo technique, capable of providing non-biased accurate protein-ligand structures in minutes of CPU wall clock. This outstanding achievement opens the door for interactive usage, allowing to combine users' expertise and intuition with *in silico* predictions.

A nice feature of adaptive-PELE is its scalability with computational resources; adding more computing cores (more trajectories) significantly reduces the wall clock computing time. While interactive refinement of active site poses requires only few processors, addressing the full binding mechanism (from solvent to the active site) requires significant more resources. While accessibility to cheap HPC will certainly increase in the near future, access to large computational resources for researchers is already a reality. Most pharmaceutical and biotech companies account for in-house large computational clusters, with several thousands of computing cores. Moreover, cloud-computing access is drastically increasing while reducing its cost; an hour of 128 computing cores sells today for ~5\$ in *Amazon Cloud*. If associated security issues were a key negative aspect in the past, this has been largely solved: more and more companies have now developed cloud solutions (Schrödinger, Openeye, etc.)



In agreement with recent studies<sup>1, 2, 5</sup> we show how all-atom molecular mechanics force fields are mature enough to sample and distinguish native like poses in complex protein-ligand systems, providing excellent means for elucidating the atomic detailed binding mechanism. Our tests involved difficult protein-ligand systems, including diverse and pharmacological relevant targets, such as the PR receptor, and a GPCR receptor for which extensive MD simulations could not provide a native like pose.

Overall, we have developed a computational breakthrough with remarkable performance in mapping the protein-ligand energy landscape, being able to reproduce the full binding mechanism in complex systems in less than half an hour, or the active site induced fit in less than 5 minutes. While standard PELE already shows a competitive advantage as a sampling technique<sup>29, 32</sup>, combining it with machine learning techniques and high performance computing, provides a solid modeling technique to the drug-design community, with potential of being interactively used in computer aided drug design.

## Methods

**The Adaptive Algorithm.** The algorithm is composed of three main steps: sampling, clustering, and spawning, which run in an iterative approach. In the sampling phase, a swarm of trajectories, in this paper in the range from tens to one thousand, are independently run. Conformations are then clustered, and the final spawning step chooses the seeds for the next iteration. By stopping simulations and adaptively spawning them, we circumvent the problem of getting trapped due to metastability, avoiding the waste of computational resources in oversampled regions.

*Sampling.* The sampling is usually the computational bottleneck of the process, so it is desired to use a method that can generate uncorrelated poses in a relatively short time. We chose PELE since it can introduce moderate conformational changes in few minutes, providing robust protein-ligand exploration, even for complex systems, within few hours of a mid-range computing cluster (~100 commodity computing cores)<sup>18, 19, 33</sup>. PELE is a two-stage MC algorithm that uses protein structure prediction procedures to generate proposals. In the first stage, the ligand is randomly moved, and the protein is perturbed using a normal mode analysis method based on an anisotropic network model (ANM)<sup>17</sup>. In the second one, the structure is relaxed with a side chain prediction and a minimization (with constraints on alpha carbons and the ligand center of mass), and the resulting proposal is accepted or rejected with the Metropolis criterion.

We use rounds (epochs) of  $N$  simulations (trajectories) of length  $l$ , each one running on a computing core (using a MPI implementation). A larger  $N$  is expected to reduce the wall-clock time to see binding events, whereas  $l$  should be as small as possible to exploit the communication between explorers but long enough for new conformations to advance in the landscape exploration. While we use PELE in this work, one could use different sampling programs such as MD as well.

*Clustering.* We used the leader algorithm<sup>34</sup> based on the ligand RMSD, where each cluster has a central structure and a similarity RMSD threshold, so that a structure is said to belong to a cluster when its RMSD with the central structure is smaller than the threshold. The process is speeded up using the centroid distance as a lower bound for the RMSD (see Supplementary Information). When a structure does not belong to any existing cluster, it creates a new one being, in addition, the new cluster center. In the clustering process, the maximum number of comparisons is  $k \cdot n$ , where  $k$  is the number of clusters, and  $n$  is the number of explored conformations in the current epoch, which ensures scalability upon increasing number of epochs and clusters.

We assume that the ruggedness of the energy landscape grows with the number of protein-ligand contacts, so we make RMSD thresholds to decrease with them, ensuring a suitable discretization in regions that are more difficult to sample. This concentrates the sampling in interesting areas, and speeds up the clustering, as fewer clusters are built in the bulk.

*Spawning.* In this phase, we select the seeding (initial) structures for the next sampling iteration with the goal of improving the search in poorly sampled regions, or to optimize a user-defined metric; the emphasis in one or another will motivate the selection of the spawning strategy. Naively following the path that optimizes a quantity (e.g. starting simulations from the structure with the lowest SASA or best interaction energy) is not a sound choice, since it will easily lead to cul-de-sacs. Using MAB as a framework, we implemented different schemes and reward functions, and analyzed two of them to understand the effect of a simple diffusive exploration in opposition to a semi-guided one.

The first one, namely inversely proportional, aims to increase the knowledge of poorly sampled regions, especially if they are potentially metastable. Clusters are assigned a reward,  $r$ :

$$\square = \frac{\rho}{C} \quad (1)$$

where  $\rho$ , is a designated density and  $C$  is the number of times it has been visited. We choose  $\rho$  according to the ratio of protein-ligand contacts, again assumed as a measure of possible metastability, aiming to ensure sufficient sampling in the regions that are harder to simulate. The  $1/C$  factor guarantees that the ratio of populations between any two pairs of clusters tends to the ratio of densities in the long run (one if densities are equal). The number of trajectories that seed from a cluster is chosen to be proportional to its reward function, *i.e.* to the probability to be the best one, which is known as the Thompson sampling strategy<sup>35,36</sup>. The procedure generates a metric-independent diffusion.

The second strategy is a variant of the well-studied  $\varepsilon$ -greedy<sup>25</sup>, where a  $1-\varepsilon$  fraction of explorers are using Thompson sampling with a metric,  $m$ , that we want to optimize, and the rest follow the inversely proportional scheme. Metrics are typically used in PELE to extract information and to drive the system towards some determined actions. They include, for example, the binding energy, the SASA of the ligand, distances between atoms, etc. Depending on whether we want to maximize or minimize  $m$ ,  $r$  is respectively defined as:

$$r_i = m_{i,\min} - m_{\min} \quad (2)$$

$$r_i = m_{\max} - m_{i,\max}, \quad (3)$$

where  $m_{i,\max}$  and  $m_{i,\min}$  are the maximum and minimum metric values within the  $i$ -th cluster respectively, and  $m_{\min}$  and  $m_{\max}$  are the overall metric minimum and maximum.

### Benchmark Systems

We have chosen four systems with different levels of complexity: the trypsin-benzamidine, the PR nuclear hormone receptor with its endogenous ligand and two different GPCRs with a potent inverse agonist and an antagonist ligand respectively; these last three systems represent current pharmaceutical targets, allowing us to evaluate the viability of the protocol in real drug design processes.

The binding of trypsin with benzamidine (PDB ID: 3PTB) has been widely used as a benchmark system<sup>6,37,38</sup>. It is the smallest and least flexible receptor and ligand, being the system that requires the least computational time.

PR with its endogenous ligand (PDB ID: 1A28) belongs to the family of nuclear hormone receptors (NHR) and is an important pharmaceutical target. NHRs have been recently studied combining crystallography and PELE<sup>19</sup>, including studies with PR<sup>30</sup>, where it was found that protein plasticity was crucial for the ligand to enter the active site.

We also tested two different GPCRs with two different ligands, tiotropium (PDB ID: 4DAJ) and CP-376395 (PDB ID: 4K5Y). GPCRs are a class of transmembrane proteins involved in the signaling of a wide range of biological functions and key pharmaceutical targets. 4DAJ is an M3 muscarinic acetylcholine receptor belonging to class A GPCRs, for which extensive MD simulations have already been performed. Despite the use of the Anton supercomputer and of 16  $\mu$ s of MD production time<sup>10</sup>, binding of tiotropium, a bronchodilator drug, into the orthosteric site could not be reported, only seeing binding to an extracellular site vestibule. 4K5Y is a class B GPCR, involved in the treatment of anxiety and depression, whose bent transmembrane helix (TM) 7 produces a pronounced V-shape allowing the ligand to enter deeper into the channel<sup>39</sup>. While no binding simulations have been reported to our knowledge, the conformational changes between the apo and the holo structures have been recently studied running 100ns MD simulations, with and without the antagonist ligand<sup>40</sup>. In addition, binding dissociation pathways have been studied with random acceleration molecular dynamics<sup>41</sup>.

### Setup

#### System preparation

In order to test the potential of the new methodology in exploring the binding mechanism, we started simulations with a model where the ligand is placed 20Å from the bound pose (see Fig. 1), and constrained its movements to a sphere of 15Å, the center of which was placed in the middle point between the native and initial configurations. Structures were prepared with Schrödinger's Protein Wizard<sup>42</sup>. Simulations were run with the OPLS2005 force field and the OBC implicit solvent<sup>43</sup>. Ligands' atomic charges were parameterized with RESP quantum charges, obtained with Jaguar<sup>44</sup> optimizations at the DFT-B3LYP and 6-31G\*\*+ level of theory.

#### PELE control file

The same parameters were used for both adaptive and non-adaptive runs. The ligand translation was set to be dependent on its (relative) solvent accessible surface area (SASA), being 3Å for SASA > 0.6 whereas it otherwise ranged randomly from 0.75 to 1.5Å in the protein vicinity; the translation direction was kept for four consecutive steps. Ligand rotation was randomly set between 20° and 60°. For the

protein backbone perturbation, performed with a probability of 0.25, the lowest six ANM normal modes were randomly mixed with a maximum displacement of 1.5 Å. The same PELE control file has been used for all systems with except for the alpha carbon constraints in the relaxation step: since it was reported that the lipid bilayer was found not to play a significant role in the binding in the GPCR<sup>40</sup>, we speeded up simulations removing the membrane and adding constraints of 5 kcal/mol/Å<sup>2</sup> every 10-th alpha carbons in the TMs, setting it to 0.2 kcal/mol/Å<sup>2</sup> in TRP and PR.

#### ***Algorithm parameters***

Although a general set of parameters has been optimized and used in this work, users are encouraged to change them; limiting factors to consider are discussed in this section.

In the sampling phase, we use exploration rounds of  $l=4$  steps, which ensures epochs of less than four minutes with the current Marenostrum 3 processors at the Barcelona Supercomputing Center (SandyBridge-EP 2.6GHz processors). Protein conformational changes can already be captured with four steps, and longer simulations were leading to poorer performance.

The number of protein-ligand contacts is used as a measure of the sampling complexity, as more contacts lead to more competing interactions and, thus, more energy barriers and metastability. We consider that a pair of protein (alpha carbons only) and ligand atoms are in contact if their distance is less than 8Å, following Ref. 23. In our implementation, we use as a parameter the ratio of the number of contacts per ligand heavy atom,  $c$ , since it is less system dependent, and regard those conformations with  $c > 1$  as difficult to sample, which correspond to poses in the protein vicinity, and those with  $c \leq 0.5$  as easy, which correspond to largely solvent exposed poses.

We tried three different combinations of cluster threshold and density values, and summarized in the table of Supplementary Fig 6. Clusters need to be small enough so that one can distinguish (relevant) different conformations. We select the thresholds with a function composed of linearly decreasing step functions in  $c$ , from 5Å in the solvent ( $c \leq 0.5$ ) to 2Å in the protein frame ( $c > 1$ ). This ensures sufficient discretization in those regions that are difficult to sample, not spending too many resources in the bulk (Supplementary Fig. 6a). Using the same threshold everywhere, requires significant more sampling to reach native like poses (Supplementary Fig. 6b), since it introduces 3 times more clusters (Supplementary Fig. 6d).

In the spawning, the density value is chosen inversely proportional to the cluster volume ( $1/V$ ). We tried different density functions. For example,  $\rho=1$  allows seeing binding events, but it divides exploration efforts in the whole domain, as can be seen in (Supplementary Fig. 6c).

1. Abel R, *et al.* Accelerating drug discovery through tight integration of expert molecular design and predictive scoring. *Curr Opin Struct Biol* **43**, 38-44 (2017).
2. Plattner N, Noé F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nature Communications* **6**, 7653 (2015).
3. Orozco M. A theoretical view of protein dynamics. *Chem Soc Rev*, (2014).
4. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J Med Chem* **59**, 4035-4061 (2016).
5. Shan Y, Kim ET, Eastwood MP, Dror RO, Seeliger MA, Shaw DE. How Does a Drug Molecule Find Its Target Binding Site? *JACS* **133**, 9181-9183 (2011).
6. Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences* **108**, 10184-10189 (2011).
7. Gervasio FL, Laio A, Parrinello M. Flexible docking in solution using metadynamics. *JACS* **127**, 2600-2607 (2005).
8. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **314**, 141-151 (1999).
9. Wang L, *et al.* Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *JACS* **137**, 2695-2703 (2015).
10. Kruse AC, *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482**, 552-556 (2012).
11. Mobley DL. Let's get honest about sampling. *J Comput Aided Mol Des* **26**, 93-95 (2012).
12. Genheden S, Ryde U. Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys Chem Chem Phys* **14**, 8662-8677 (2012).
13. Bernardi RC, Melo MCR, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1850**, 872-877 (2015).
14. Jorgensen WL, Tirado-Rives J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J Comput Chem* **26**, 1689-1700 (2005).

15. Vitalis A, Pappu RV. Methods for Monte Carlo simulations of biomacromolecules. *Annu Rep Comput Chem* **5**, 49-76 (2009).
16. Madadkar-Sobhani A, Guallar V. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Res* **41**, W322-W328 (2013).
17. Cossins BP, Hosseini A, Guallar V. Exploration of Protein Conformational Change with PELE and Meta-Dynamics. *J Chem Theory Comput*, (2012).
18. Kotev M, Lecina D, Tarragó T, Giralt E, Guallar V. Unveiling Prolyl Oligopeptidase Ligand Migration by Comprehensive Computational Techniques. *Biophys J* **108**, 116-125 (2015).
19. Edman K, *et al.* Ligand binding mechanism in steroid receptors; from conserved plasticity to differential evolutionary constraints. *Structure* **23**, 2280-2290 (2015).
20. Kopečná J, *et al.* Porphyrin Binding to Gun4 Protein, Facilitated by a Flexible Loop, Controls Metabolite Flow through the Chlorophyll Biosynthetic Pathway. *J Biol Chem* **290**, 28477-28488 (2015).
21. Carlson HA, *et al.* CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J Chem Inf Model*, (2016).
22. Zwier MC, *et al.* WESTPA: An Interoperable, Highly Scalable Software Package for Weighted Ensemble Simulation and Analysis. *J Chem Theory Comput* **11**, 800-809 (2015).
23. Huang X, Bowman GR, Bacallado S, Pande VS. Rapid equilibrium sampling initiated from nonequilibrium data. *Proceedings of the National Academy of Sciences* **106**, 19765-19769 (2009).
24. Zimmerman MI, Bowman GR. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J Chem Theory Comput* **11**, 5747-5757 (2015).
25. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT Press (1998).
26. Chapelle O, Li L. An empirical evaluation of thompson sampling. In: *Advances in neural information processing systems* (2011).
27. Kappel K, Miao Y, McCammon JA. Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor. *Q Rev Biophys* **48**, 479-487 (2015).

28. Borrelli KW, Cossins B, Guallar V. Exploring hierarchical refinement techniques for induced fit docking with protein and ligand flexibility. *J Comput Chem* **31**, 1224-1235 (2010).
29. Grebner C, Iegre J, Ulander J, Edman K, Hogner A, Tyrchan C. Binding Mode and Induced Fit Predictions for Prospective Computational Drug Design. *J Chem Inf Model*, (2016).
30. Grebner C, *et al.* Exploration of binding mechanisms in nuclear hormone receptors by Monte Carlo simulations and X-ray derived motion modes. *Biophysical Journal* **ASAP**, (2017).
31. Gil VA, Guallar V. pyProCT: automated cluster analysis for structural bioinformatics. *J Chem Theory Comput* **10**, 3236-3243 (2014).
32. Kotev M, Soliva R, Orozco M. Challenges of docking in large, flexible and promiscuous binding sites. *Bioorg Med Chem* **24**, 4961-4969 (2016).
33. Acebes S, *et al.* Rational Enzyme Engineering Through Biophysical and Biochemical Modeling. *ACS Catalysis* **6**, 1624-1629 (2016).
34. Hartigan JA. *Clustering Algorithms*. John Wiley & Sons, Inc. (1975).
35. Thompson WR. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **25**, 285-294 (1933).
36. Abeille M, Lazaric A. Linear Thompson Sampling Revisited. *arXiv preprint arXiv:161106534*, (2016).
37. Takahashi R, Gil VA, Guallar V. Monte Carlo Free Ligand Diffusion with Markov State Model Analysis and Absolute Binding Free Energy Calculations. *J Chem Theory Comput* **10**, 282-288 (2014).
38. Guvench O, Price DJ, Brooks CL, 3rd. Receptor rigidity and ligand mobility in trypsin-ligand complexes. *Proteins* **58**, 407-417 (2005).
39. Hollenstein K, *et al.* Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **499**, 438-443 (2013).
40. Xu J, Wang Z, Liu P, Li D, Lin J. An insight into antagonist binding and induced conformational dynamics of class B GPCR corticotropin-releasing factor receptor 1. *Mol Biosyst* **11**, 2042-2050 (2015).
41. Bai Q, Shi D, Zhang Y, Liu H, Yao X. Exploration of the antagonist CP-376395 escape pathway for the corticotropin-releasing factor receptor 1 by random acceleration molecular dynamics simulations. *Mol Biosyst* **10**, 1958-1967 (2014).

42. Madhavi Sastry G, Adzhigirey M, Day T, Annabhimoju R, Sherman W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **27**, 221-234 (2013).
43. Onufriev A, Bashford D, Case DA. Exploring protein native states and large - scale conformational changes with a modified generalized born model. *Proteins: Structure, Function, and Bioinformatics* **55**, 383-394 (2004).
44. Jaguar 4.1. Schrödinger , Inc., Portland, Oregon, 2000.

#### **Acknowledgment**

We thank Drs Anders Hogner and Christoph Grebner (AstraZeneca) for fruitful discussions and feedback on the manuscript.

This work was supported by a grant from the Spanish Government CTQ2016-79138-R). D.L. acknowledges the support of *SEV-2011-00067*, awarded by the Spanish Government.

#### **Author Contributions**

All authors contributed in performing the simulations, analyzing the data and writing the manuscript.

#### **Competing financial interests**

The author(s) declare no competing financial interests.



## **Publication 6 – A Markov State Model benchmark using a Monte Carlo procedure as a propagator**

**Authors:** Daniel Lecina, Joan Francesc Gilabert, Christoph Grebner, Victor Guallar.

In Preparation

Publication 6 is in preparation, and we include as sections the main lines of work that will be added into it.

## Effects of minimizations in sampling

In this subsection, we analyze the effects of using minimizations in sampling and the consequent approximations to extract useful information to study protein-ligand binding with PELE.

PELE belongs to the *basin-hopping*<sup>219</sup> methods, which combine stochastic moves and minimizations to perform energy landscape explorations. These have been extensively applied in global optimization problems<sup>220</sup>. One example is the Monte Carlo-plus-minimization<sup>221</sup>, developed by Li and Scheraga in 1987. It has been used for finding the lowest energy structures of small molecules, showing better conformations and convergence than simulated annealing<sup>222</sup>; for crystal structure prediction of rigid and flexible small organic molecules<sup>223</sup>; and for small molecule docking<sup>224</sup>. A second instance is the activation-relaxation technique of Mousseau and colleagues, which has been applied to Lennard-Jones clusters<sup>225</sup>, relaxation of glasses<sup>226</sup>, amyloid fibril formation<sup>227,228</sup> and protein folding<sup>229,230</sup>. Finally, we especially underline the work of David Wales and coworkers, who have done exhaustive research on the field for over twenty years, establishing their theoretical underpinnings<sup>231</sup>.

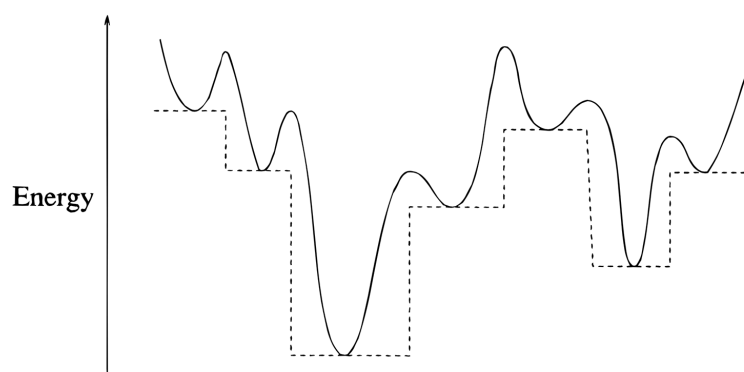


Figure 14. One-dimensional illustration of the transformed potential energy surface  $\tilde{E}(\mathbf{r})$  (dashed line) and the original one,  $E(\mathbf{r})$  (solid line). In the example,  $\tilde{E}(\mathbf{r})$  is built with a steepest-descent protocol, and local minima are partitioned in stationary points that correspond to maxima. Source: Ref. 219.

In *basin-hopping*, the potential energy surface is mapped at each point,  $\mathbf{r}$ , with a minimization:

$$\tilde{E}(\mathbf{r}) = \min (E(\mathbf{r})) \quad (20)$$

where  $\tilde{E}$  is the mapped potential energy and  $E$  the original one. Importantly, the relative energy differences between minima and the global minimum are preserved<sup>219</sup> (Fig. 14). A major consequence of the transformation is that the system can hop between basins more easily, as energy barriers between local minima are removed and thermodynamics are modified<sup>232</sup>. Note that there may exist barriers separating funnels, and this motivated the appearance of methods such as minima hopping<sup>233</sup>. The combination of the transformed energy landscape with protein structure prediction algorithms explains the great efficiency of PELE in search simulations: PELE has been able to reproduce the binding mechanism of challenging complexes on commodity computers

overnight, such as the nuclear hormone receptors<sup>19,20</sup> or GPCRs<sup>234</sup>, where special purpose supercomputers or enhanced sampling methods have not succeeded<sup>235,236</sup>.

But as we anticipated, the use of minimizations has important implications in sampling. In particular, minimizations break detailed balance and information on vibrational entropy<sup>142</sup> is lost. There are fortunately several workarounds to overcome this limitation. For example, the total density of states,  $\Omega(E)$ , and the partition function can be decomposed into a summation of all local minima, known as the superposition approximation<sup>231</sup>, and each local vibrational contribution can be approximated to harmonic terms<sup>198</sup>. Eventually, anharmonic corrections may be added<sup>237</sup>. In a different approach, *basin-sampling*<sup>238</sup>, these contributions are obtained sampling each local minimum. As opposed to the standard *basin-hopping*, where minimizations affect the proposal coordinates (it has been shown to perform better in global searches<sup>239</sup>), in *basin-sampling* minimizations are solely used to compute the proposal energy, and hence, detailed balance is satisfied. Still, vibrations within minima must still be recovered and are calculated using the Wang-Landau<sup>240</sup> sampling algorithm, restraining the sampling to the different local minima.

In PELE, current algorithmic limitations do not allow removing minimizations (*e.g.* see ANM), and we make the following approximations to recover detailed balance: 1) add harmonic constraints during minimizations to preserve stochastic proposals, and 2) choose a *sufficiently loose* convergence criterion in minimizations. Although these parameters, along with the temperature, have been fitted to reproduce experimental binding free energies, it is worth remarking that the approximations are heuristic and the Boltzmann distribution is not formally sampled. We illustrate the parameter choice first with a simplified harmonic potential model and then with different real-case examples.

*Harmonic potential model.* In this example, the energy well is described with a one-dimensional harmonic potential of force constant  $k_b$  (Fig. 15).

Following the *basin-hopping* procedure, the initial random sample  $x_o$  is minimized until a certain criterion is met, for example until the RMSG falls below a threshold value, yielding a final proposal  $x_f$ . Depending on  $x_o$ , two situations may occur. First, those  $x_o$  such that  $\left. \frac{dE}{dx} \right|_{x_o} := E'(x_o) \leq \text{RMSG}$  already meet the convergence criterion and are thus unaltered by minimizations, *i.e.*  $x_f = x_o$  (thick line in Fig. 15a). Note that the range of unaltered proposals  $[-x_b, x_b]$  such that  $E'(x_b) = \text{RMSG}$ , will increase proportionally to  $\text{RMSG}/k_b$  and in the limit of  $\text{RMSG}/k_b \gg 1$  we recover Boltzmann sampling. In the second case, the energy is minimized until  $E'(x_f) \leq \text{RMSG}$  and the specific  $x_f$  will depend on the minimization protocol. For example, minimizing the energy in small  $\Delta x$  following the gradient yields  $x_f \approx x_b$ .

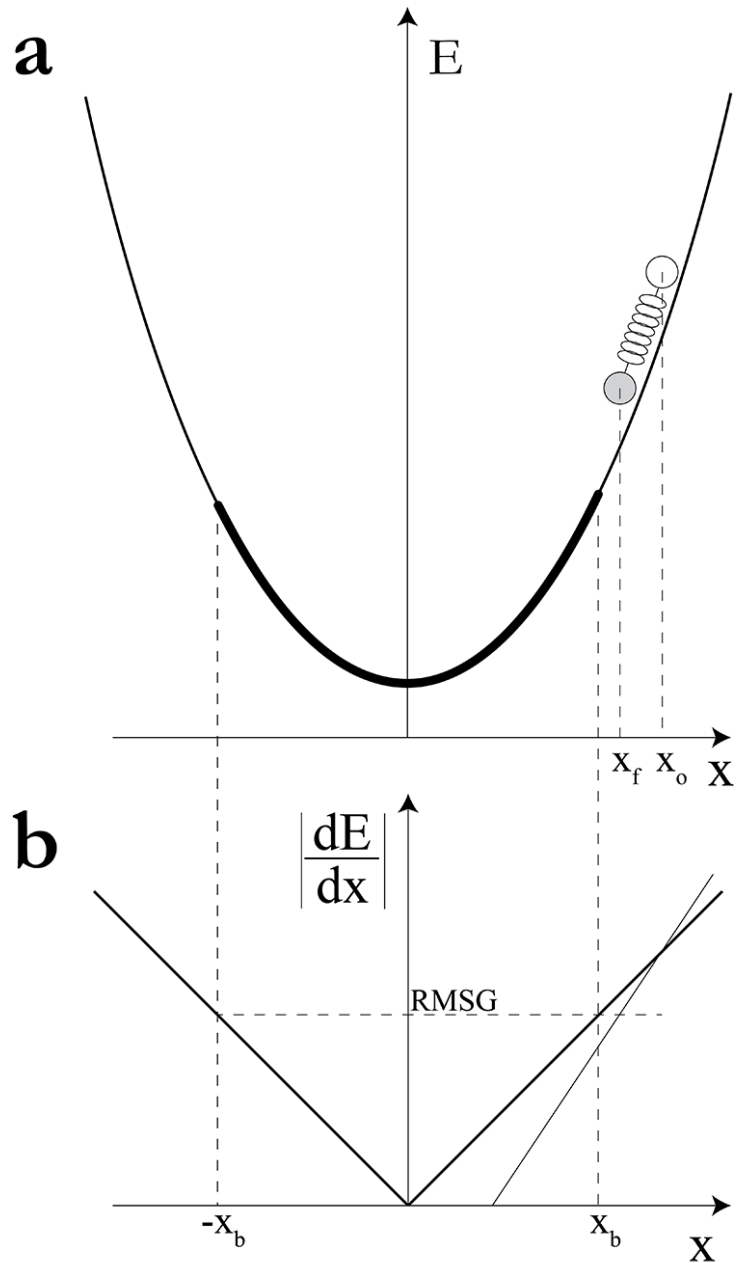


Figure 15. Sampling of a one-dimensional energy minimum. Panel **(a)** depicts the harmonic potential with the energy in the y-axis and coordinate in the x-axis (thin line).  $x=x_o$  corresponds to the random proposal, and  $x=x_f$  to the resulting coordinate after the minimization. Those random proposals such that  $|x_o| < x_b = \text{RMSG}/k_b$  are not affected by minimizations because they already meet the convergence criterion, *i.e.*  $x_o=x_f$  (thick line). Panel **(b)** shows the absolute value of the gradient (thick line). The RMSG limits the boundaries of unaltered proposals  $[-\text{RMSG}/k_b, \text{RMSG}/k_b]$  and by adding a harmonic constraint (thin line) we modify the gradient in such a way that  $|x_f| > |x_b|$ .

It may also happen that some coordinates become unreachable due to minimizations, acting the bottom of the energy basin as a “sink”. Adding a harmonic constraint to  $x_0$  with force constant  $k$  avoids collapsing all proposals to the lowest energy range. This has two consequences: it preserves the sampling at the bottom of the energy basin, which has the largest contribution to the partition function, and extends the number of accessible states within the basin. For  $k \gg k_b$ , we obtain  $x_f \approx x_0$ , and the Boltzmann distribution is recovered.

In real-case scenarios, the energy landscape is composed of energy wells with arbitrary shapes and curvatures, and it is hard to assess *a priori* the consequences of our approximations in free energy calculations. Below, we show the repercussions in free ligand diffusion examples, varying one parameter while keeping the other fixed in the fitted value.

*Minimization convergence.* The RMSG is again used to tune the convergence of minimizations; typically we set it to 0.05 kcal/mol/Å and 0.1 kcal/mol/Å in the ANM and final minimizations, respectively. Criteria are slightly different, as the sole purpose of the latter minimization is recovering a continuous side chain dihedral description after the use of discrete rotamer libraries. Note that other *basin-hopping* studies on free energy calculations of protein-ligand association have used  $\text{RMSG} = 0.001 \text{ kcal/mol/Å}^{198}$ , a much stronger criterion.

As we saw with the harmonic potential, the RMSG threshold has repercussions in detailed balance. For example, in the limit of  $\text{RMSG} \rightarrow \infty$ , minimizations are not applied, and detailed balance is fulfilled by construction. For finite values, we measure the magnitude of asymmetrical fluxes between states, with an eye on the future application to free energy calculations. First, we coarse grain the conformational space in clusters and compute the transition fluxes between any pair of states,  $i$ , and  $j$ :  $F_{ij}(\tau) := \pi_i p_{ij}(\tau)$ , being  $\pi$  the cluster population,  $p_{ij}$  the transition matrix, and  $\tau$  a lag time that ensures Markovianity in the coarse-grained space (see Markov State Models section). When detailed balance is satisfied, the flux matrix is symmetrical up to statistical fluctuations, which we assess with the ratio  $M(F)$ :

$$M(F) = \frac{\|F - F^T\|}{\|F + F^T\|} = \frac{\sqrt{\sum_{ij} |F_{ij} - F_{ji}|^2}}{\sqrt{\sum_{ij} |F_{ij} + F_{ji}|^2}} \quad (21)$$

We used the Frobenius norm<sup>241</sup>, and the dependence of  $M(F)$  on the lag time is implicit in the flux matrix. The value  $M(F)=0$  corresponds to perfectly symmetrical matrices, whereas  $M(F)=1$  corresponds to unidirectional fluxes, such as the example in Fig. 6a.

| RMSG<br>(kcal/mol/Å) | 0.001       | 0.01        | 0.05        | 0.1       | 1.0         |
|----------------------|-------------|-------------|-------------|-----------|-------------|
| $M(F)$               | 0.465±0.009 | 0.320±0.008 | 0.201±0.003 | 0.21±0.01 | 0.211±0.004 |

Table 1.  $M(F)$  for different RMSG values in free diffusions of CRA\_10655 with a Urokinase-type plasminogen activator (PDB ID: 1O3P). For each value, we used 512 different trajectories of one day on SandyBridge-EP 2.6GHz cores (*i.e.* from  $\sim 1000$  MC steps for  $\text{RMSG}=0.001 \text{ kcal/mol/Å}$  to  $\sim 4000$  MC steps for  $\text{RMSG}=1 \text{ kcal/mol/Å}$ ). Results are averaged over ten different clusterizations with 100 k-means clusters, bootstrapping the original data.

In Table 1, we show the results for five different sets of 512 free ligand diffusion simulations using  $\text{RMSG} = 0.001, 0.01, 0.05, 0.1$  and  $1.0$  kcal/mol/Å in both ANM and final minimizations. As we loosen the convergence criterion,  $M(F)$  becomes smaller, showing that detailed balance is more closely satisfied, in agreement with the harmonic potential model. For example, the choice of  $\text{RMSG} = 0.05$  kcal/mol/Å reduces the unsymmetrical fluxes more than one half compared to  $0.001$  kcal/mol/Å. The values in Table 1 are used as a rule of thumb to assess the quality of prospective simulations, but there is not a direct translation to free energy values.

*Harmonic constraints.* Harmonic constraints are added to preserve stochastic MC proposals during minimizations. More precisely, all ligand heavy atoms and protein  $C_\alpha$ 's are constrained after their respective perturbations with force constants  $k = 1$  kcal/mol/Å<sup>2</sup>. It is worth emphasizing they are solely applied during minimizations and do not affect the Metropolis criterion.

To select the appropriate  $k$  value, we chose a difficult scenario with numerous stabilizing interactions and study the effect of minimizations for different  $k$ 's. The interactions in the native pose of 5-iodo-2-(oxalylamino)-benzoic acid in a protein-tyrosine phosphatase (PDB ID: 1ECV) include (Fig. 16a): two salt bridges and six hydrogen bonds between three ligand oxygens and protein residues in the Cys215-Arg221 loop; a salt bridge between the ligand o-carboxylic acid group and Lys120, along with two possible additional hydrogen bonds with Tyr46 and Asp181; non-polar interactions with Ala217, Ile219, Val49... Overall, these favorable interactions make it a challenging system; the best interacting pose within our PELE simulations had an interaction energy of  $\sim -140$  kcal/mol. At the same time, it is a suitable test case because the ligand's low weight and rigidity permit extensive sampling.

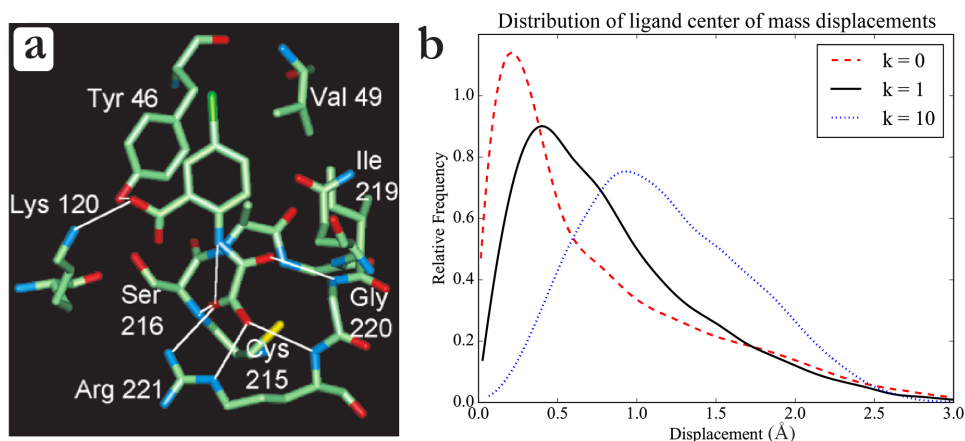


Figure 16. Panel (a): Native interactions of 5-iodo-2-(oxalylamino)-benzoic acid in a protein-tyrosine phosphatase (PDB ID: 1ECV). Adapted from Ref. 242. Panel (b): Distribution of ligand center of mass displacements over  $10^4$  PELE proposals for  $k=0, 1, 10$  kcal/mol/Å, respectively. The initial structure is that of panel (a) and corresponds to a difficult scenario, where minimizations oppose to the stochastic movements due to strongly favorable interactions. Ligand translations are drawn 50/50 from two uniform distributions:  $(0.75\text{Å}, 1.25\text{Å})$  or  $(1.5\text{Å}, 2.5\text{Å})$ , and have thus an average of  $1.5\text{Å}$ .

We launched three sets of 64 simulations for  $k=0, 1$  and  $10$  kcal/mol/Å<sup>2</sup> starting from the minimized crystal pose with the multiscale TN. We observed protein-

ligand dissociation events for all three  $k$  values, which importantly excludes the idea of absorbing states, from which the simulation cannot escape, in energetically favorable poses due to minimizations. This is probably due to the severe ruggedness of the energy landscape, which impedes completely undoing stochastic displacements with the sole use of minimizations.

To further analyze the effects of constraints, we modified PELE to output the ligand center of mass displacements for both accepted and rejected steps. In Fig. 16b, we show that larger  $k$  values have a better preservation of stochastic proposals, in agreement with the toy model. For example, the most likely displacements are respectively  $\sim 0.2$ ,  $0.5$  and  $1.0$  for each  $k$  value in increasing order. Finally, we decomposed ligand displacements within the different stages of a PELE step (Table 2), confirming that the ANM minimization has a stronger effect than the last one. This can be explained by its lower RMSG value (see above) and because the final minimization starts from a pose that has been already minimized.

| Algorithm    | Ligand Perturbation       | ANM                       | Minimization              | Total                     |
|--------------|---------------------------|---------------------------|---------------------------|---------------------------|
| Displacement | $1.5 \pm 0.5 \text{ \AA}$ | $1.2 \pm 0.6 \text{ \AA}$ | $0.2 \pm 0.2 \text{ \AA}$ | $0.8 \pm 0.6 \text{ \AA}$ |

Table 2. Average ligand center of mass displacements in the different stages of a PELE step for  $k=1$  kcal/mol/ $\text{\AA}^2$ . Results were averaged out over  $10^4$  PELE step proposals.

*Temperature.* Because of the modified thermodynamics, the *basin-hopping* Metropolis temperature,  $T$ , does not have a direct correspondence with that of the heat-reservoir. For example, in our simulations proteins maintain a compact shape and do not denature at  $T$  as high as 2000K. This effect has already been observed by Mosseau and coworkers<sup>227</sup>. The sampling temperature, 1000K, has also been fitted to reproduce binding free energies.

Launching 127 simulations of 3-Phenylpropylamine with trypsin (PDB ID: 1TNK) for different temperatures, we observe a maximum number of bound trajectories for  $T \sim 1000$ -1500K (Table 3). Although at large temperatures the system is able to cross entrance barriers more easily (*e.g.* see MC acceptance), it also has a tendency to explore a larger portion of the conformational space. On the contrary, lower temperatures tend to get trapped in local minima and we are not able to reproduce as many binding events in the same computational time.

| $T$ (K)            | 500 | 1000 | 1500 | 2000 |
|--------------------|-----|------|------|------|
| MC acceptance      | 15% | 30%  | 50%  | 60%  |
| Bound trajectories | 6   | 45   | 47   | 34   |

Table 3. Acceptance and bound trajectories in 127 free ligands diffusions of 3-Phenylpropylamine with trypsin (PDB ID: 1TNK) for different temperatures. We regard as bound those trajectories that at some point reproduced the crystal structure, *i.e.* root-mean-squared deviation (RMSD)  $\leq 3 \text{ \AA}$ . The acceptance increases with the sampling temperature and the number of bound trajectories.



## MSM sampling

### Initial seeding points

The convergence of a Markov model can be improved for example increasing the number of trajectories or their length. A third way is to enhance the sampling of certain regions. In a recent publication<sup>20</sup> (publication 4), we observed that convergence is achieved faster when the simulation initial structures are distributed along the binding pathway. In this work we aim to quantify the gain and show that both approaches converge to the same distribution. To this end, we compare the convergence to a well-converged gold model in two scenarios: 1) when an ensemble of trajectories starts from a single region in the bulk solvent, or 2) when the ensemble is split in different points along the binding pathway. To generate the binding pathway, we use an adaptive procedure that is able to map difficult binding mechanisms very efficiently (publication 5).

We assess the similarity of two Markov models with a measure based on the relative entropy, proposed by Pande and coworkers<sup>243</sup>:

$$D(P||Q) = \sum_{i,j} \pi_i P_{ij} \log \left( \frac{P_{ij}}{Q_{ij}} \right) \quad (22)$$

where P is the reference transition matrix, Q is the test distribution, and  $\pi$  is the reference stationary distribution. Note that it requires a common definition of microstates and lag time.

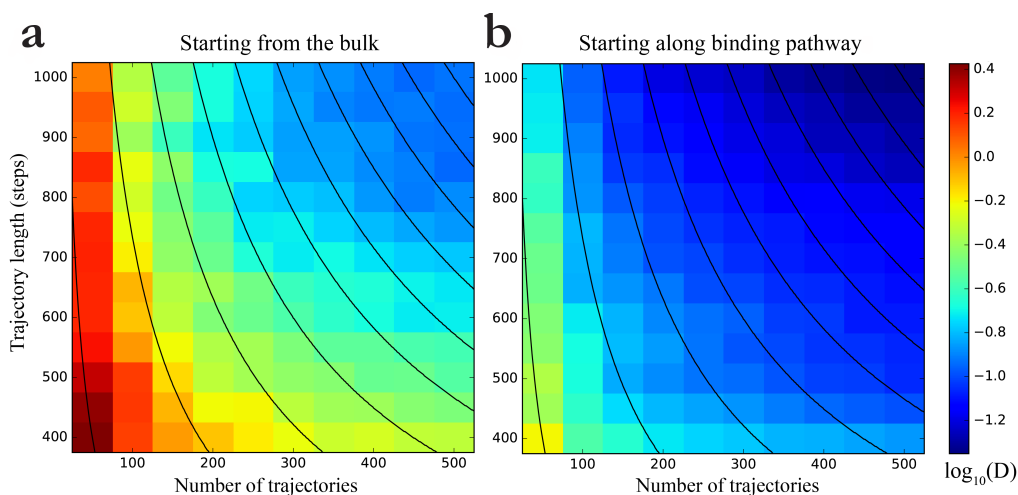


Figure 17. Relative entropy between a gold standard model and two different sampling schemes, in panel (a) all simulations start from the bulk solvent and in panel (b) in 100 points along the binding pathway. The second approach converges faster to the gold model. Each data point represents the average over 100 bootstrap iterations.

The test system is CRA\_10655 in a Urokinase-type plasminogen activator (PDB ID: 1O3P), since we are able to obtain a converged MSM using both sampling schemes. The sampling is obtained from 512 independent free ligand diffusions of 2000 steps, starting from the bulk solvent. The reference 100 microstates are constructed from all ligand C7 coordinates, using *k-means*. Finally, we measure D for an increasing number of trajectories and trajectory length using both sampling

schemes. We see that the second one converges faster to the gold model (Fig. 17). For example, the best achieved quality with the first scheme is  $\log_{10}(\mathbf{D}) \cong -0.9$ , which is obtained for 500 trajectories of 1000 steps, whereas the second is able to reach this quality with 200 trajectories of 500 steps, roughly 5 times less sampling. The second approach requires fewer trajectories to achieve the same convergence, since we proportionally see a larger number of binding events. Note that with this approach, data collection from occluded regions may start in the early stages of the simulations, as opposed to the first one, where binding events must occur first. For these reasons, the speedup is expected to be larger for receptors with more occluded binding sites, such as PR, and lower for systems with an easier binding mechanism, such as trypsin and benzamidine.

### Adaptive seeding points

In Markov models, the evolution of a system is approximated with a transition matrix that accounts for conditional transition probabilities<sup>207</sup>. A major consequence is that simulations may be much shorter than the global process under study, as they only need to be long enough to characterize local transitions. This opens a way for sampling schemes that exploit this idea. An example is adaptive sampling, where Markov models aid the choice of seeding points and has been shown to dramatically improve sampling convergence<sup>209–211,244</sup>. Note that the quality of the model may not be guaranteed, especially in the first adaptive rounds due to poor sampling. Also, as pointed out by Zimmerman and Bowman<sup>245</sup>, adaptive sampling may sample less relevant conformations, and guiding the exploration towards a goal would make a better use of computational resources.

We propose an iterative sampling scheme that: 1) solely uses Markov models for analytic purposes, and 2) focuses computational resources in *rewarding* regions. As seen in the previous section, choosing the seeding points along the binding pathway improves the convergence of the MSM, and the reward is hence chosen proportionally to the number of contacts. In order to balance the exploration, the reward is chosen inversely proportional to the cluster population (see publication 5). In order to gather statistics for the Markov model, the epoch length must be longer than the lag time, and we use epochs two times longer.

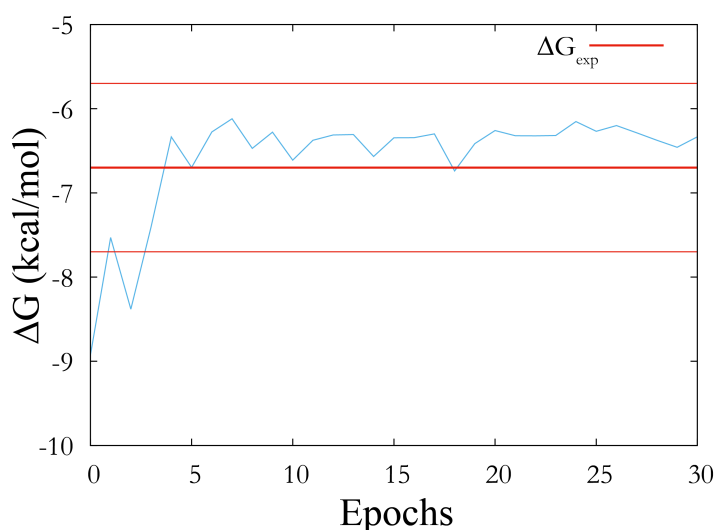


Figure 18. Evolution of the absolute binding free energy estimation (in blue) for trypsin and benzamidine (PDB ID: 3PTB) using 50 different trajectories and epochs of 50 steps. The experimental value is shown in a thick red line and in a thin red the interval  $\pm 1$  kcal/mol. The estimation converges in 4 epochs (200 MC steps), and the experimental result is reproduced.

To test the protocol, we use the adaptive simulations, and estimate the binding free energy at the end of each epoch. In the starting structure, the ligand is placed in the bulk solvent, and the exploration is limited to a sphere of radius  $20\text{\AA}$ , centered in the protein surface. In Fig. 18, we show results for trypsin with benzamidine, where it can be seen that convergence is reached after 4 epochs ( $\sim 200$  MC steps) using 50 processors. Additionally, the experimental result is correctly reproduced.

## Discussion

In this section, we summarize the results and discuss their impact.

### Development of PELE

The original version of PELE<sup>108,109</sup> was written in procedural Fortran and made use of PLOP (Protein Local Optimization Program) functions for tasks such as energy calculation, minimization, and side chain prediction. PLOP is an academic protein-modeling program<sup>71,133,134,139</sup> developed by Jacobson, Friesner, and collaborators with a commercial counterpart integrated into the Schrödinger suite under the PRIME package. PELE lacked of tests, and their different components were not entirely independent, which hampered the reliability and maintainability upon extension. This was a major concern, as competitive state-of-the-art computational tools must be able to take advantage of cutting-edge algorithms and techniques. For this reason, it was decided to rewrite the program from scratch, using the object-oriented programming paradigm.

Importantly, this first goal is a requirement to achieve the rest of objectives. For this reason, prior and during the first stage of this thesis, the author took part in the rewriting of PELE full-time during three years, and occasionally thereafter. In the recoding, we fostered good practices of software development<sup>246,247</sup>. These involved testing, improving the readability, modularity, encapsulation, maintenance and version control, just to name a few. **Testing** is one of the fundamental pillars, as it ensures that behaviors are not unexpectedly altered and guarantees reproducibility. They should cover all the code, and contain unit and integration tests. Testing floating-point numbers is particularly challenging, especially in end-to-end tests (see the Lyapunov instability in the introduction), and virtual environments are a possible solution. The lack of **readability** hampers the development, and some examples of this are bad variable namings, dead or duplicated code, misleading or outdated comments. **Modularity** is a fundamental pillar of object-oriented programming that should be addressed in the design. Different software components are more easily testable when they are divided independently, and therefore, better **maintainable** and less error prone. **Encapsulation** is linked to modularity, and hides the implementation from the user, and also improves maintainability (*e.g.* the code does not depend on a particular implementation but on the interface) and promotes **reusability**. **Version control** is crucial to keep track of the changes and trace back potential bugs in the code. All in all, it is essential to follow good practices in software development in order to obtain **reliable** and **maintainable** programs.

From a scientific point of view, the rewriting was also a turning point, and subsequent improvements involved the biomolecule model (*e.g.* FF or solvent)

and algorithms. Also, it is worth mentioning the efforts put in the development of a graphical user interface, which will ease simulation analysis and provide a more intuitive insight (Fig. 19).

Originally, PELE only supported OPLS-AA FFs and two surface area implicit solvents, SGBNP and VDGBNP. In the new version, we included AMBER FFs and the OBC model, which for example allowed studies on DNA<sup>248</sup>. The OBC represents an overall speedup of roughly  $\sim 1.5x$ , as it does not require the calculation of atomic surfaces. Regarding further improvements, in our experience, all of our implicit solvent models have the tendency to collapse the protein and work should be done to address this situation. On the other hand, new energy models, such as VSGB 2.0, improve the energy function with empirical corrections based on docking scoring functions (see biomolecular modeling section), which we have seen to dramatically improve the prediction of MM/GBSA free energy estimations. Moreover, FFs are evolving to account for the induced polarizations caused by heterogeneous environments, and future work will certainly involve including some of them (see biomolecular modeling section).

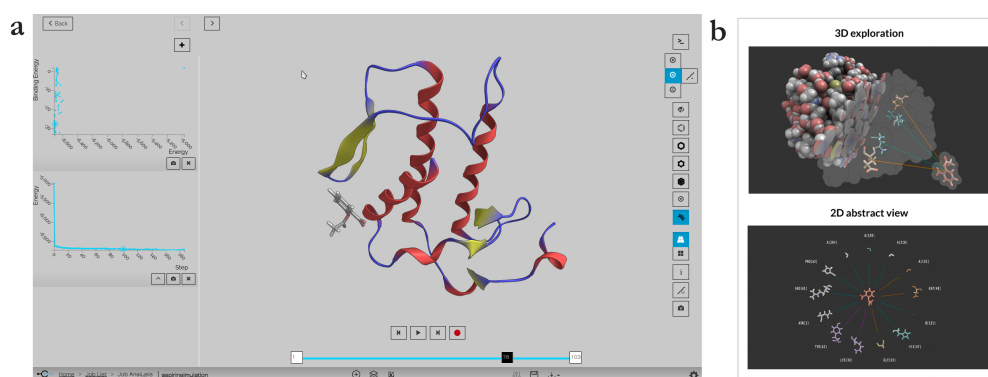


Figure 19. Panel (a): Screenshot of the beta version of the graphical user interface (courtesy of Jorge Estrada). Panel (b): Different plotting schemes and options will be available, for example highlighting the highest interacting ligand-residue pairs. Source: Ref. 249.

Regarding new algorithms, one of the biggest contributions has been the IC-NMA<sup>129</sup>, to which the author participated helping in the coding and writing the manuscript (publication 8, not included). It seems a promising tool to reproduce protein flexibility; it leads to a better coverage of the conformational space and protein perturbations incur in lower energy increments because of the collective motions of internal coordinates. Moreover, as opposed to our ANM implementation, it does not require of minimizations, and will affect the sampling. The implications of this development in binding free energy calculations will certainly need to be assessed once the integration to support ligands is complete. In a different project, efforts are currently put to include a consensus scoring function that is based on machine learning protocols and is aimed to be applied in virtual screening scenarios. Other contributions have been the addition of atom pulling algorithms to reproduce atomic force microscopy experiments<sup>250</sup>. Altogether, porting the code to a C++ modular language has helped in adding all these features.

The author of this thesis has also worked implementing algorithms for the replacement of minimizations in sampling. It is ongoing work and results have not been included in the thesis, but are briefly summarized in this paragraph. The first case is the use of biased MC<sup>54</sup>, where one can enhance the acceptance ratio with a bias in the proposals (*i.e.*  $\alpha_{in} \neq \alpha_{ni}$ ,  $\forall i,j$  in Eq. 13 of the introduction). In the particular procedure known as force-biased MC<sup>251–253</sup>, one can construct proposals such that  $\alpha_{ni}/\alpha_{in} \approx e^{\beta(E_n - E_i)}$  and the acceptance becomes one (or nearly one). We have successfully applied it to predict side chain conformations after mutations, where we computationally mutate one residue and expect to predict its crystal conformation. One such example is the mutation of Phe<sup>104</sup> to Met<sup>104</sup> in T4 Lysozyme (PDB ID: 1QTV to PDB ID: 1CV0, movie: <https://goo.gl/or5mhw>, the native protein is shown in orange, the initial conformation in blue licorice and in colored licorice the simulated side chain). Another test application has been protein-ligand sampling, as in the binding of 3-Phenylpropylamine to trypsin (PDB ID: 1TNK, movie: <https://goo.gl/QE7n7n> where the native protein and ligand are shown in orange, the sampled ligand in colored licorice and the protein in white cartoon). A second method that attracted our attention was the non-equilibrium candidate Monte Carlo (NCCMC) of Chodera and coworkers<sup>254</sup>. Monte Carlo proposals are obtained with a non-equilibrium procedure and accepted with a criterion that preserves Boltzmann sampling. It is aimed to improve the statistical efficiency, being able to generate more uncorrelated poses in subsequent steps while keeping a reasonable acceptance. The rationale behind our application is the elimination of potential side chain clashes resulting after the IC-NMA perturbation. It has already been implemented, and we are currently in the validation process.

We saw in the introduction that PELE has been used in many applications and is a recognized tool in the academic field. For example, in the CSAR 2014 blind docking contest it showed “an impressive success” cross-docking a set of ligands across an ensemble of different receptors<sup>174</sup>. In the forthcoming future, one of the objectives is having a major penetration in industrial drug design pipelines. For this reason, and in accordance with the following objectives of this thesis, we rewrote and extended PELE to build an efficient piece of software that can be easily maintained and extended with leading algorithms, according to the market needs.

## **Establishing a protocol to study protein-ligand binding**

After accomplishing the first objective, we were able to establish a protocol of unbiased all-atom simulations with PELE to study protein-ligand binding combining it with MSM to quantify the exploration, in view to be used in drug discovery pipelines. Comprehending protein-ligand recognition is one of the biggest challenges in molecular modeling, involving a large computational cost associated with sampling lots degrees of freedom in a rugged energy landscape<sup>255</sup>. The wide range of different methodologies devoted to its study gives an idea of its relevance (and difficulty!)<sup>256</sup>, and quantifying the association with computational tools gives a competitive advantage in drug design, reducing time and costs<sup>257</sup>.

PELE is used in the sampling phase due to its capacity in mapping protein-ligand binding in occluded binding sites, such as NHRs. As stated in the introduction, when using PELE, rather than using a single and long simulation, we often run

batches of 100s of simulations to characterize the energy landscape, which reduces the wall-clock time (we exploit this idea to perform collaborative explorations, see the discussion on the next objective below). On the other hand, MSM is a handful methodology to join all these independent simulations in a unique statistical model; one of the advantages is that individual simulations only need to characterize the energy landscape locally, and the global characterization can be extracted from the model. Also, with this approach we do not need to define reaction coordinates *a priori*, and they will be given by the MSM *a posteriori* in the analysis. Ryoji Takahashi, a former post-doc in our group, showed that the combination of PELE and MSM might be a suitable tool to study protein-ligand binding, correctly characterizing the association of four different small molecules related to benzamidine with trypsin<sup>115</sup>. Our objective has been establishing a protocol to study protein-ligand binding and developing a set of tools for routinely semi-automatic analysis.

The early stages of this objective were devoted to acquiring insight into PELE applications, which was accomplished with the help of post-docs in our laboratory, James J. Valdes and Martin Kotev. In publication 1, we studied the ligand migration of the inhibitor Z-pro-prolinal in POP, along with an undecapeptide substrate and its release product. Using PELE, we could map the ligand migration and binding pathway in such a complex system in less than 48 hours. In order to estimate the entropy loss upon binding, we used translational, rotational and vibrational terms<sup>258</sup> combined with a rotatable bond screening to account for flexibility. We found a significant loss of entropy due to flexibility, in agreement with previous studies<sup>259</sup>, and observed an apparent entropy-enthalpy compensation<sup>260,261</sup> upon binding, although words of caution have been issued lately on the topic<sup>262</sup>. The prohibitive costs hamper a routinely application of this technique, despite major approximations were made.

Posteriorly, efforts were devoted to designing the sampling protocol. In the first place, we assessed the effects of bounding the exploration to a sphere centered in the entrance and observed that the absolute binding free energy estimation was not compromised, likely due to the absence of alternative binding pockets in the studied systems. More importantly, we evaluated the effects of minimizations in sampling, which are discussed in the results section, and obtain a parameter set to use in free energy estimations. We also developed a validation protocol that we kept improving throughout the Ph.D., and in Fig. 20 we show an example of a typical validation pipeline (to appear in publication 6).

The binding of dexamethasone and dibC to MR is studied in publication 2. We should underline that application of PELE provides a full non-biased binding event for this challenging system, with a completely buried active site. We focused on the plasticity of the region where helices 3, 7 and 11 meet, being the helix 6-7 region the one with the largest mobility. The occluded nature of the binding site and the coupling of ligand displacements with backbone rearrangements make this system especially challenging. In order to obtain sufficient binding statistics, we placed the ligand in the peripheral binding site and perform simulations with a weak restrain that limited the exploration to a 10-15Å sphere<sup>b</sup>. With this technique we were not able to recover absolute binding free

---

<sup>b</sup> These simulations took place during the validation period of the new PELE, and were thus performed with the original version, which did not have the option to limit ligand moves within a box with no harmonic constraints.



energies, probably to the limited solvent sampling, but relative energies, presumably due to the shared entrance point.

In publication 3, we investigated the binding modes of KA and HQ in a tyrosinase to test the inhibition mechanism that was found experimentally. Given the lower complexity of the system compared to NHRs, we are able to obtain a converged MSM with 127 simulations during 24h ( $\sim 2 \cdot 10^5$  data points) and obtain the absolute binding free energy of KA. With a metastable state analysis (clustering utilizing the PCCA algorithm in EMMA 1.3<sup>215</sup>), we found two distinct binding poses, with a relative energy difference of  $\sim 4$  kcal/mol between them. Compared to KA, HQ presents a greater heterogeneity of binding poses (Fig. 21b), supporting experimental observations. We hypothesize that this is due to the lack of a carboxyl group that interacts with Arg<sup>209</sup>, which, in our results, stabilizes L-tyrosine (Fig. 5a in publication 3).

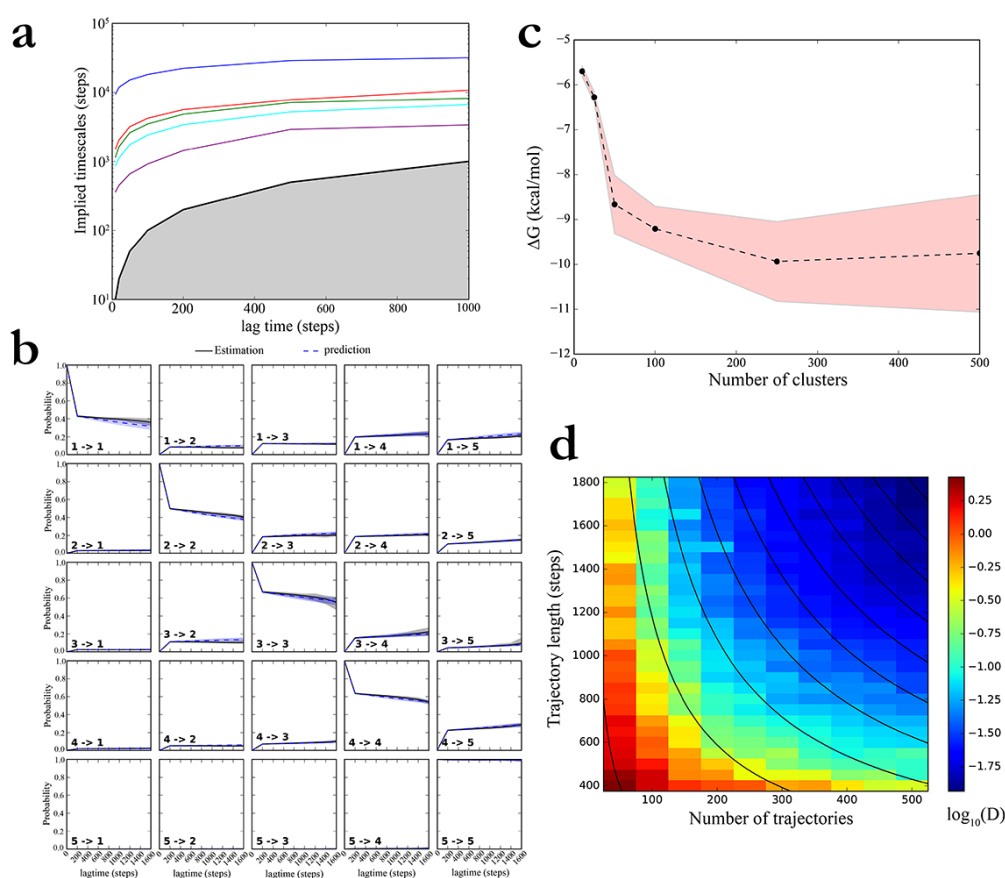


Figure 20. The validation process of our protocol, exemplified with 512 free diffusions of CRA\_10655 with a Urokinase-type plasminogen activator (PDB ID: 1O3P). In panel (a), we show the implied timescales for the four slowest decaying processes, which have reached the plateau at  $\tau = 200$  steps. In panel (b), we show the Chapman-Kolmogorov<sup>214</sup> test for the 5 metastable states built using the PCCA++ algorithm, where we can see that transition probabilities are estimated correctly. Panel (c) shows the evolution of  $\Delta G$  for a different number of clusters (dots), with their uncertainty (red shadowed region). Panel (d) shows the averaged values of  $D^{20}$  (a measure to assess convergence that is based on the relative entropy) over 100 different runs using bootstrap for a different number of trajectories and lengths. Isocost lines are shown.

Publication 4 is devoted to studying the binding of all NHRs with their endogenous ligands, focusing in more detail in the PR with three different NHR

ligands: progesterone, aldosterone, and cortisol. Notwithstanding the use of PCA-based modes in PELE's protein perturbation, obtaining sufficient sampling still remained elusive using unbiased diffusions from the bulk. For this reason, we selected six points along a binding pathway with different SASA values (0.0, 0.2, 0.4, 0.6, 0.8, 1.0), and launched 100 simulations from each point during 24 hours (a total of 600 simulations of  $\sim 2500$  steps) in a spherical box with radius  $20\text{\AA}$  centered in the peripheral binding site. Remarkably, with this technique, we were able to obtain absolute binding free energies in reasonable agreement with experiments. Aside from free energy estimations, we used MSMs to elucidate the binding mechanisms and observed that progesterone's highest hydrophobicity seems to play a role in the binding. For this publication, we developed different tools and procedures to analyze MSMs. Among these, we automated the calculation of uncertainties using bootstrap or devised a protocol for convergence assessment. The latter uses a measure based on the relative entropy that is computed for a varying number of trajectories and trajectory length (see SI7 in publication 4). This assessment can be made on the fly so that simulations can be automatically stopped when a convergence threshold is reached.

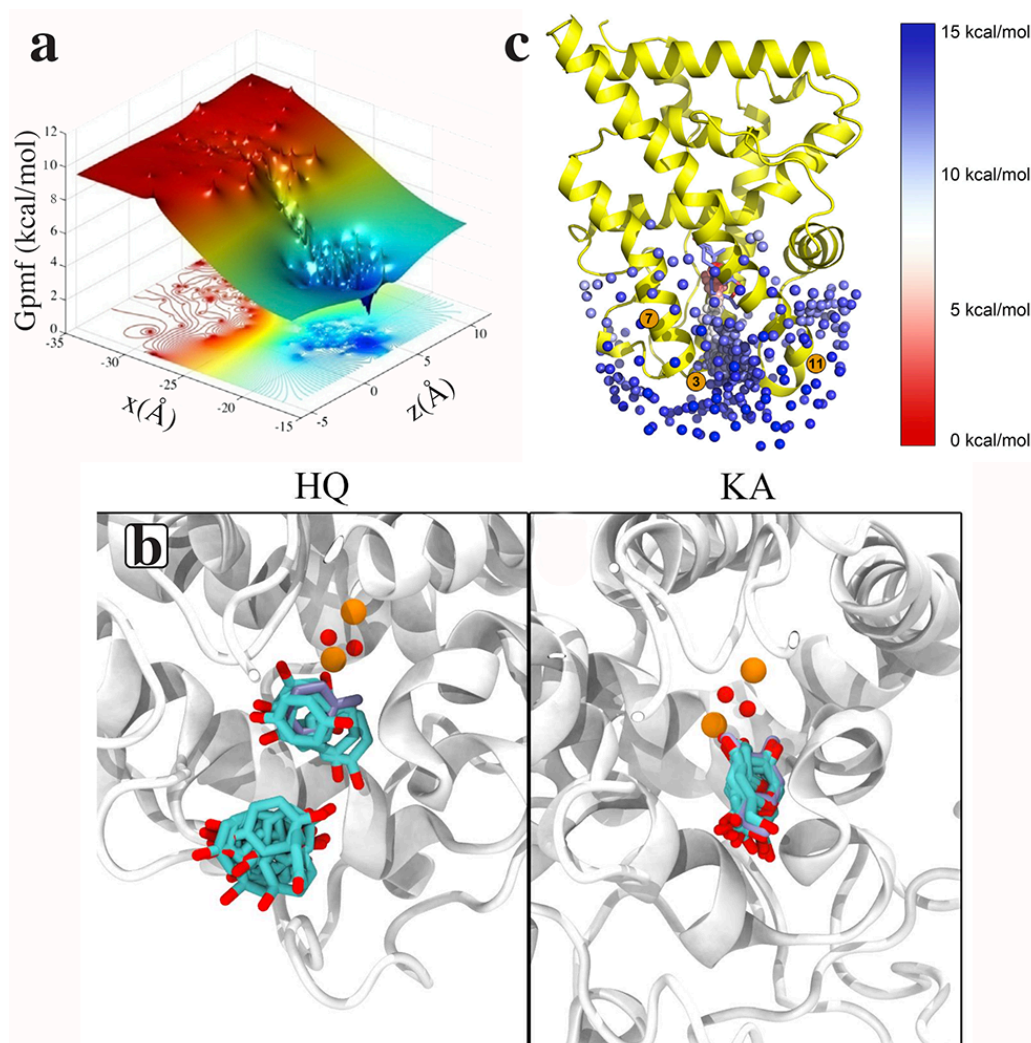


Figure 21. Different representations of  $G_{\text{pmf}}$  that are used in the thesis. In panel (a), we show a projection of  $G_{\text{pmf}}$  on the  $x$  and  $z$ -axis, used in publication 2. In panel (b) we show representative structures within 1 kcal/mol of the most likely pose (in blue), which is used in publication 3. In panel (c), we represent cluster centers as beads, with a color code representing the  $G_{\text{pmf}}$  value, utilized in publication 4. All figures are adapted from the corresponding publication.

In a thought-provoking opinion article<sup>158</sup>, David Mobley states that wrong binding free energy estimations can be usually traced back to three causes: 1) poor sampling, 2) wrong FFs or 3) different setups between experiments and simulations. According to his experience, he claims that in most cases where FFs are blamed for the inaccuracies, these can still be attributed to poor sampling. In our case, our protocol has a sampling advantage but we need to consider a fourth point, namely the deviations from the Boltzmann distribution due to minimizations. Throughout the thesis, we developed tools to perform a systematic analysis of convergence and also provide the basis to limit and evaluate the impact of minimizations on sampling, although further work should provide a more direct relationship with free energies. As anticipated, algorithmic improvements in PELE will certainly change the sampling protocol, yielding more reliable estimations but at a higher computational cost.

With no doubt, the quality of the model affects the estimation. For instance, the use of quantum charges calculated in the binding pose improves results in our experience (not shown). In the last SAMPL blind prediction challenge<sup>161</sup>, it was seen that all-atom explicit solvent models consistently give better results than other approaches, which suggests the addition into PELE of algorithms such as the grand canonical Monte Carlo-based technique proposed by Ross and coworkers<sup>263</sup>. However, as pointed out by Mobley and Gilson in a recent paper<sup>196</sup>, we need of a community-accepted benchmark that permits a systematic analysis and validation of free energy packages and methods. According to the authors, it should include sampling challenges (*e.g.* side chain or backbone rearrangements), system challenges (*e.g.* varying protonation state upon binding) and FF challenges (*e.g.* cases where treating polarization is necessary). This would allow straightforward comparisons; users could choose the methodology that better fits their needs, whilst developers could more easily find the weaknesses of their procedures and improve them.

We would like to give some final remarks in the contextualization of our binding free energy predictions. We have recently seen the appearance of methods such as FEP+ that have provided outstanding relative free energy estimations<sup>162</sup>. This methodology has been used to improve the force field parameterization in the new OPLS3<sup>49</sup>, and future developments will certainly improve its performance. However, seeing the modest results of rigorous free energy methods on the SAMPL contest, we would like to raise a word of caution on our binding affinity estimations. The study on minimizations in PELE has allowed its successful application in challenging systems with a fair success. Nonetheless, given the approximate nature of our calculations, we should not expect better accuracies than those rigorous methods on blind tests.

Overall, combining high-performance computing (HPC), PELE and MSMs we were able to study protein-ligand binding in real-case problems with industrial interest. An independent study of the pharmaceutical company AstraZeneca showed that PELE is a suitable tool for (industrial) drug design, due to its tradeoff between conformational sampling, accuracy, and speed<sup>264</sup>. In this thesis we have shown an ongoing collaboration with them in the study of NHRs, and will continue with another publication that is currently in preparation (publication 6). In context with the late renewed interest in computational techniques in drug discovery, we expect PELE to have a deeper dissemination in the pharmaceutical industry in the following years.

## Development of a procedure to overcome the sampling limitations associated to metastability

One of the main drawbacks of the previous protocol is the sampling limitation associated to metastability. MC simulations can theoretically traverse the energy landscape efficiently, but in practice, proposals are typically reduced to small variations of the initial structure due to the difficulty of generating uncorrelated poses with a non-negligible acceptance. While the combination of protein structure prediction algorithms and minimizations in PELE provides an improved search, it still suffers sampling problems, as simulations tend to get trapped in energy wells (Fig. 22c). The main effect of metastability is the long computational time associated with an adequate sampling of the energy landscape, as too short times make results depend on the initial structure, and thus ergodicity does not hold.

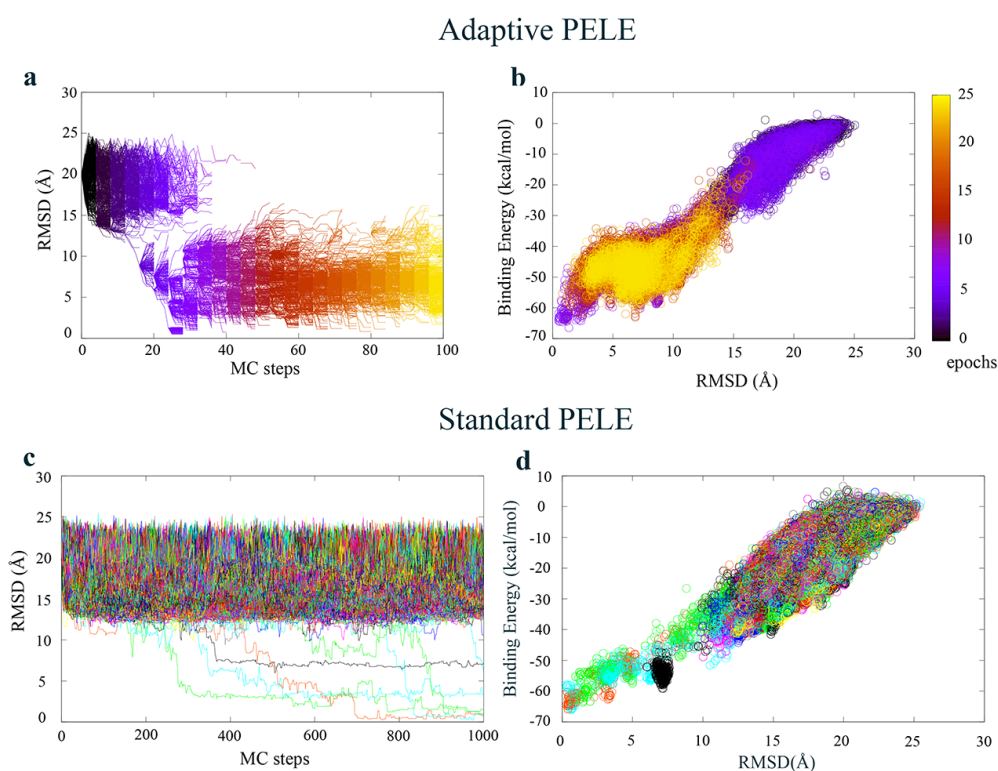


Figure 22. Energy landscape exploration of PR using 512 trajectories. Panels (a,b): RMSD and binding energy evolution for the adaptive PELE. The color code corresponds the epoch number. Panels (c,d): Analogous plots for standard PELE. The color code corresponds to the trajectory number. Source: publication 5.

To improve these sampling limitations, in publication 5 we presented an adaptive procedure using reinforcement learning ideas. It represents a significant advance in our laboratory. Compared to standard PELE simulations, with this technique we are able to map binding events between 10 and 50 times faster in challenging systems, which translates in performing mechanistic studies using non-biased simulations in less than 30 minutes with current computational multi-cores (*e.g.* 512 SandyBridge-EP 2.6GHz cores). The testing benchmark was composed of three complex systems with current application in the pharmaceutical industry. For example, one of them is the M3 muscarinic acetylcholine GPCR, where 16  $\mu$ s

performed with the special purpose supercomputer ANTON<sup>235</sup> or 1  $\mu$ s with accelerated MD<sup>236</sup> did not succeed to reproduce the binding of tiotropium. The adaptive exploration does not obey the Boltzmann distribution, and average measures cannot be extracted directly from them. However, transitions are unbiased within each epoch, and one could simply build a MSM to recover equilibrium properties, as stated by Zimmerman and Bowman<sup>245</sup> (see below). Hence, this technique opens the door to an intense screening of compounds.

We envision drug design as an iterative process where human expertise is combined with computational modeling to guide experiments. However, while fast docking methods have a limited conception of flexibility and cannot reproduce significant conformational rearrangements, the long execution times associated with (fully flexible) molecular simulations delay iterations to hours or days. The adaptive protocol dramatically reduces this delay, which not only makes PELE a more competitive tool to fit in real life industrial drug design settings but also possibly opens the way for human-computer interactive simulations when a large number of processors is available. Pharmaceutical companies usually possess large computational clusters, and the appearance of services such as Amazon has eased the access to HPC facilities. We believe that this sort of interactive interaction will become a routine in the near future.

The adaptive protocol also represents a simplification of the simulation setup and facilitates the use of PELE to non-expert users. For example, in NHRs, the binding site is deeply buried, and we have shown that the protein undergoes backbone rearrangements in the association process. This makes it a challenging system, and only a small fraction of the trajectories that start from the bulk solvent find the binding pose (*e.g.* Fig. 22c). For this reason, in publications 1 and 3 we use two kinds of simulations with different parameters to characterize the energy: a ligand migration simulation to study binding pathways, and a refinement simulation in the binding pocket to find the best binding pose. The adaptive protocol automatically balances the search and therefore only needs one simulation setup and makes the process transparent to the user at a fraction of the cost (*e.g.* Fig. 22a, 22b).

The technique is implemented as an object-oriented Python program, and we followed the good practices of software development discussed for PELE (testing, version control, maintainability...). It has been written with extensibility in mind, for example, to add the option to use MD<sup>265-267</sup> as a dynamics propagator. We also wrote a set of small programs to ease the simulation analysis, including metrics analysis (*e.g.* RMSD or binding energy evolution, binding energy correlation with RMSD...), clustering metrics extraction (*e.g.* evolution of the number of clusters, cluster representatives, population, transitions between clusters...) or building of binding pathways.

Further algorithm improvements presented in this paragraph will be discussed in Joan Francesc Gilabert's master thesis. One such example is the clustering. It is based on the ligand's RMSD and does not consider the protein. We hypothesize that taking it into account with a contact map (a binary matrix with ligand atoms as rows and protein atoms as columns) may improve the binding times. A second example is the spawning point within clusters. In publication 5, we solely use the representative (central) cluster structure, but it is certainly not the optimal choice. For instance, clusters with an RMSD threshold of 5Å limit the minimum epoch length to  $\sim$ 4 MC steps, as it is not feasible to discover new states with shorter

epochs. The use of alternative spawning points may enable the use of shorter epochs and possibly faster binding times.

In a publication 6, we show the application of adaptive PELE in binding free energy simulations. Choosing initial structures along the binding pathway provides better convergence compared to starting all the sampling from single pose in the bulk solvent, and adaptive simulations are efficient at building binding pathways. However, this idea can be taken one step further. Following the work of different research groups<sup>211,243,268</sup>, we built a sampling scheme to efficiently sample the energy landscape in view to construct MSMs. In Fig. 23, we show the evolution of the binding free energy estimation using the adaptive protocol, which shows converged binding free energy results for  $\sim 200$  MC steps, using 50 trajectories. As a matter of comparison, Takahashi<sup>115</sup> used 600 trajectories of 1000 MC steps to obtain a converged free energy value (note that in this case the exploration included the whole protein surface and the comparison is not fair).

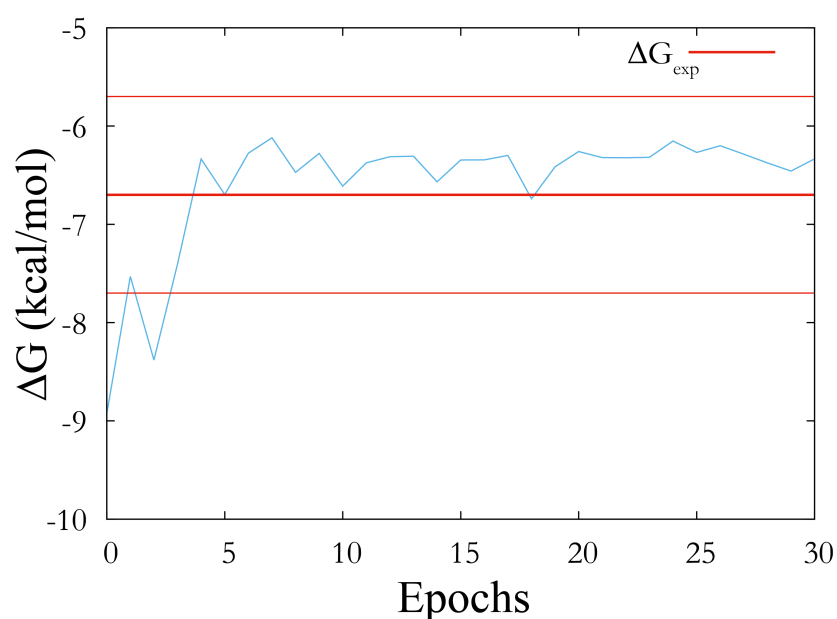


Figure 23. Evolution of the absolute binding free energy estimation (in blue) for trypsin and benzamidine (PDB ID: 3PTB) using 50 different trajectories and epochs of 50 steps. The experimental value is shown in a thick red line and in a thin red the interval  $\pm 1$  kcal/mol. The estimation converges in 4 epochs (200 MC steps), and the experimental result is reproduced.

## Conclusions

The final conclusions of this work are the following:

- The development of PELE resulted in a competitive cutting-edge software, integrating new algorithms and platforms, such as the OBC implicit solvent, the IC-NMA or the GUI, just to name a few. Importantly, the rest of the thesis is built upon it.
- The application of PELE in complex receptors, such as the prolyl oligopeptidase or nuclear hormone receptors, allows describing binding mechanisms in a fast atomically detailed manner.
- Computing the entropy variation upon binding requires a tremendous amount of computation, and despite major approximations, our proposed technique is not suitable for a routinely screening of compounds.
- Combining PELE with Markov models allows quantifying binding mechanisms. Importantly, using a limited sampling region, we are able to find relative binding free energies in nuclear hormone receptors and tyrosinases. However, this still represents a significant computational effort.
- The adaptive procedure represents a major improvement in our sampling protocol, and opens the door for an intensive screening of compounds.
- Coupling the adaptive sampling procedure with PELE and Markov models seems a promising tool for faster binding free energy predictions.
- Overall, due to its tradeoff between conformational sampling and speed, PELE is a suitable tool to study protein-ligand interactions.





## Appendices

### Resum de la tesi

Les simulacions biomoleculares han sigut àmpliament emprades en l'estudi d'interaccions proteïna-lligand<sup>c</sup>. Comprendre els mecanismes involucrats en la predicció d'afinitats d'unió tindria una repercussió molt significativa en la indústria farmacèutica, ja que els processos de disseny de fàrmacs reduirien els costos i temps associats. Malgrat les llargues escales temporals i la dificultat de mostrejar l'espai de fases associades a aquests processos biofísics, les millores metodològiques i de *hardware* fan de les simulacions amb ordinador un candidat prometedor en la resolució de processos biofísics. En aquest context, l'objectiu de la tesi és el desenvolupament d'un protocol que permeti l'estudi d'interaccions proteïna-lligand amb vistes a ser aplicat en processos de disseny de fàrmacs.

L'autor ha contribuït a la reescriptura de PELE, el nostre programa de mostreig que està basat en la tècnica de Monte Carlo. Durant aquest procés hem fet servir bones praxis de desenvolupament de *software*, com per exemple el *testeig*, la millora de la llegibilitat, modularitat, encapsulació, manteniment i el control de versions, per anomenar-ne unes quantes. L'objectiu ha sigut produir un programa fiable i de fàcil manteniment, donant certes garanties que el seu comportament no es vegi inesperadament alterat amb la seva extensió, de manera que els experiments (computacionals) siguin reproduïbles. Aquestes característiques tenen una importància cabdal en un programa competitiu, ja que necessita la incorporació d'algorismes d'última generació per a satisfer les necessitats de la indústria. Això fa que el desenvolupament de PELE hagi sigut indispensable per tal de proposar una metodologia amb impacte en l'àmbit farmacèutic, l'objectiu principal d'aquesta tesi.

Hem demostrat que PELE és una eina capaç de reproduir mecanismes d'entrada de forma eficient i acurada. Una mostra n'és la publicació 1, on hem estudiat la migració de l'inhibidor Z-pro-prolinal en una prolyl oligopeptidasa, juntament amb el d'un pèptid d'onze residus i el producte de la reacció. Fent servir PELE, hem aconseguit caracteritzar la migració d'un sistema tan complex en menys de 48 hores. Paral·lelament, hem estimat l'entropia perduda en la unió mitjançant termes de translació, rotació i vibracionals, en conjunció amb un cribratge d'enllaços rotables per a tenir en compte la flexibilitat del lligand i del receptor. Hem vist una pèrdua significativa de l'entropia degut a la flexibilitat, suggerint una certa compensació entàlpica-entròpica en el procés d'associació, malgrat ser subjecte de debat actualment.

Un cops arribats a aquest punt, ens vam centrar en dissenyar el protocol per estudiar les interaccions proteïna-lligand. Aquest consisteix en combinar simulacions atomístiques de PELE, donada la seva capacitat per a reproduir

---

<sup>c</sup> Molècula petita que s'uneix (es "lliga") a una altra, en aquest cas una proteïna.

mecanismes d'unió en sistemes complexos com el detallat al paràgraf anterior, amb models de Markov, coneguts a la literatura com a *Markov state Models* (MSM), per tal de quantificar el paisatge energètic. En primer lloc, vàrem fer un estudi exhaustiu de l'efecte de les minimitzacions en el mostreig (detallat a la secció de resultats de la tesi), i vam avaluar l'error associat a mostrejar una part reduïda de la proteïna. Tanmateix, vàrem començar a establir un protocol de validació que hem anat perfeccionant al llarg del doctorat i que apareixerà en una publicació que es troba en preparació (publicació 6), i on a la Fig. 20 del cos principal de la tesi en podem veure un exemple.

La primera aplicació d'aquesta metodologia va ser l'estudi de la unió de dexamethasone i de desisobutyrylciclesonide a un receptor de mineralocorticoides (publicació 2). La naturalesa del lloc d'unió, enterrat dins de la proteïna, i de l'acoblament dels moviments del lligand amb canvis conformacionals de la proteïna fan d'aquest un sistema molt complex de simular. Per tal d'obtenir estadística suficient, vam col·locar el lligand a l'entrada del canal d'unió, i vam limitar la seva exploració a una esfera d'uns 15Å. Amb aquesta tècnica, no vam aconseguir calcular l'energia lliure absoluta, però sí la relativa entre ambdós lligands. Aquest estudi es va realitzar en col·laboració amb un grup de recerca de l'empresa farmacèutica AstraZeneca.

Seguidament, vam aplicar aquesta tècnica per a investigar els modes d'unió de KA (*kojic acid*) i de HQ (*hydroquinone*) en una tirosinasa, per tal de testejar mitjançant eines computacionals el mecanisme d'inhibició trobat experimentalment. En aquest estudi vam reproduir les energies d'enllaç absolutes per a KA i vam observar dues conformacions d'unió diferents amb una separació entre elles d'aproximadament 4kcal/mol. En canvi, vam observar més heterogeneïtat en les conformacions d'unió d'HQ, fet que corrobora les observacions experimentals. Presumim que és degut a la manca d'un grup carboxil que interactua amb Arg209, que, en els nostres resultats, estabilitza la L-tirosina (Fig. 5a a la publicació 3).

En la publicació 4, estudiem la unió de tots els receptors nuclears d'hormones i els seus respectius lligands endògens, focalitzant el nostre detall al receptor de progesterona amb tres lligands diferents: progesterona, aldosterona i cortisol. Malgrat que l'ús de modes normals de vibració basats en estructures experimentals amplifica el mostreig, no va ser suficient per tal d'obtenir resultats convergents d'energia lliure. Per aquesta raó, vam seleccionar sis punts al llarg d'un camí d'entrada i vam llançar 100 simulacions des de cadascun d'ells, durant 24 hores (un total de 600 simulacions i uns 2500 passos de Monte Carlo). Amb aquest procediment vam ser capaços d'estimar valors absoluts d'energia lliure equiparable als resultats experimentals. A més a més, vam fer servir MSMs per a descriure el mecanisme d'entrada i vam detectar que la hidrofobicitat de la progesterona pot tenir un rol en aquest mecanisme. En aquesta publicació, vam desenvolupar i millorar eines i procediments per analitzar models de Markov, com per exemple, per mesurar la convergència de les simulacions (SI publicació 4) o per estimar incerteses.

Com il·lustra el cas anterior, mostrejar l'espai conformacional és tot un repte a nivell computacional. Per això, a la publicació 5 fem una proposta de simulacions iteratives on el punt inicial varia de forma adaptativa i està basat en conceptes de *reinforcement learning*. Aquesta proposta representa una millora molt significativa al nostre grup de recerca. Per exemple, amb aquesta tècnica som capaços de reproduir mecanismes d'unió un ordre de magnitud més ràpid que amb el

protocol anterior. Com a conseqüència, som capaços de modelar mecanismes d'entrada en menys de 30 minuts quan disposem d'un nombre suficient de nuclis computacionals (per exemple, 512 nuclis SandyBridge-EP 2.6GHz). També hem mostrat que aquesta eina es pot fer servir pel càlcul d'energies lliures d'unió, fet que obre la porta a un cribratge molt més ràpid de compostos.

En resum, hem vist que combinant *high-performance computing*, PELE i MSM, som capaços d'estudiar la unió de proteïna-lligand en casos d'interès farmacèutic. Amb el conjunt d'eines presentades en aquesta tesi esperem que PELE tingui una major disseminació a la indústria farmacèutica i contribuir d'aquesta forma a millorar el procés de disseny de fàrmacs.



## Resum individual de cada publicació

### Resum de la primera publicació

En aquesta publicació estudiem la migració de l'inhibidor Z-pro-prolinal en una prolyl oligopeptidase (POP). POP és una proteasa que presenta un centre actiu enterrat profundament; per aquesta raó és un sistema que suposa un repte per a mètodes de mostreig convencional i motiva l'ús de PELE. En l'observació de múltiples esdeveniments d'unió, vam trobar que l'entrada es produeix a través d'un porus al domini d'hèlices enrotllades ( $\beta$ -propeller, en anglès). A més a més, vam modelar la unió d'un substrat, undecapèptid i l'alliberament d'un producte dipèptid mitjançant un protocol esbiaixat. La dissociació transcorre a través d'un *loop* flexible de 18 aminoàcids, un camí diferent al d'entrada.

### Resum de la segona publicació

En aquest article s'ha estudiat la flexibilitat dels receptors nuclears d'hormones (NHRs, en anglès). En particular, s'ha resolt les estructures de raig X dels receptors de glucocorticoides i mineralocorticoides (MR) per tal d'identificar la pasticitat conservada a la regió de les hèlices 6-7. Amb la finalitat de donar suport a la idea que constitueixen una part integral de la unió, s'han llançat simulacions d'entrada, de sortida i de refinament. Els NHRs presenten lloc d'unió enterrats i l'estudi de la migració del lligand està actualment fora de l'abast dels procediments atomístics estàndards, no esbiaixats. Per aquesta raó, es va utilitzar PELE per fer el mostreig d'aquest estudi. Es va desenvolupar un procediment de mostreig limitat al voltant del punt d'entrada compartit ( $\sim 10-15\text{\AA}$ ) i es va poder calcular la diferència d'afinitat de l'energia lliure entre la dexamestasona i la desisobutyrylciclesonida (dibC), ambdues amb MR. El temps de residència correlaciona amb la magnitud de les reorganitzacions estructurals requerides. En definitiva, mostrem que la natura ha conservat la capacitat d'obrir aquesta regió, la qual imposa diferents restriccions evolutives en els diferents receptors d'esteroides.

### Resum de la tercera publicació

En aquest estudi s'ha demostrat per primera vegada el mecanisme d'inhibició de l'àcid kòjic (KA) i de la hidroquinona (HQ) en una tirosinasa, combinant tècniques experimentals i computacionals. Les tècniques experimentals van consistir en cristal·litzacions, anàlisis de les constants d'unió i experiments cinètics. Les eines computacionals van involucrar el llançament de conjunts de simulacions atomístiques no esbiaixades per ambdós inhibidors, amb un posterior anàlisi amb models de Markov (Markov state models, MSM; en anglès). Els resultats mostren que el KA actua com a inhibidors mixt; quan aquest es troba al centre actiu, aquest no és accessible pel substrat, però quan està a la zona d'unió perifèrica, restringeix l'entrada i la sortida, i impedeix aconseguir la velocitat catalítica màxima. Al contrari, l'HQ pot actuar tant de substrat com d'inhibidor, suggerit per la seva heterogeneïtat d'unió.

### Resum de la quarta publicació

En aquest article extenem l'estudi del mecanisme d'unió a tots els membres dels receptors nuclears d'hormones amb els seus lligands endògens. Vam trobar un

camí d'entrada comú a la zona de les hèlices 3, 7 i 11, i vam identificar dos plegaments diferents de la zona de les hèlices 6 i 7, que tenia repercussió en el nombre d'esdeveniments d'unió a les simulacions sense biaix. També vàrem veure que incorporant informació de rajos x a la pertorbació de la proteïna, promovíem la plasticitat de la zona de les hèlices 6 i 7, i per tant milloraven el mostreig dels esdeveniments d'unió, comparat amb el model de xarxa neuronal anisotròpica. Aquestes nous modes de vibració, en combinació amb un mostreig més exhaustiu del camí d'entrada es pot fer servir per millorar la convergència de les simulacions que involucren MSMs. Les nostres estimacions de les energies lliures d'unió estan molt d'acord als resultats experimentals. El mecanisme d'entrada va posar en rellevància la importància d'una zona d'unió perifèrica, i la influència de la hidrofobicitat en la transició de la zona perifèrica a la zona activa.

### **Resum de la cinquena publicació**

En aquesta publicació proposem un nou procediment per millorar les limitacions de mostreig causades per la metastabilitat. En particular, la nostra metodologia combina un procediment d'aprenentatge automatitzat (machine learning, en anglès) amb PELE, en el marc de la supercomputació moderna, i és capaç de reproduir mecanismes d'unió complexes amb un guany d'un ordre de magnitud comparat amb execucions no adaptatives. La metodologia es va testejar amb sistemes complexos, on procediments estàndard no ho han aconseguit, com a receptors acoblats a proteïnes G (G-protein coupled receptor, GPCR; en anglès) i receptors nuclears d'hormones, que suggereix el potencial d'aquesta eina en estudis de cribratge i optimització de compostos.



## Supporting information paper 2

**Structure, Volume 23**

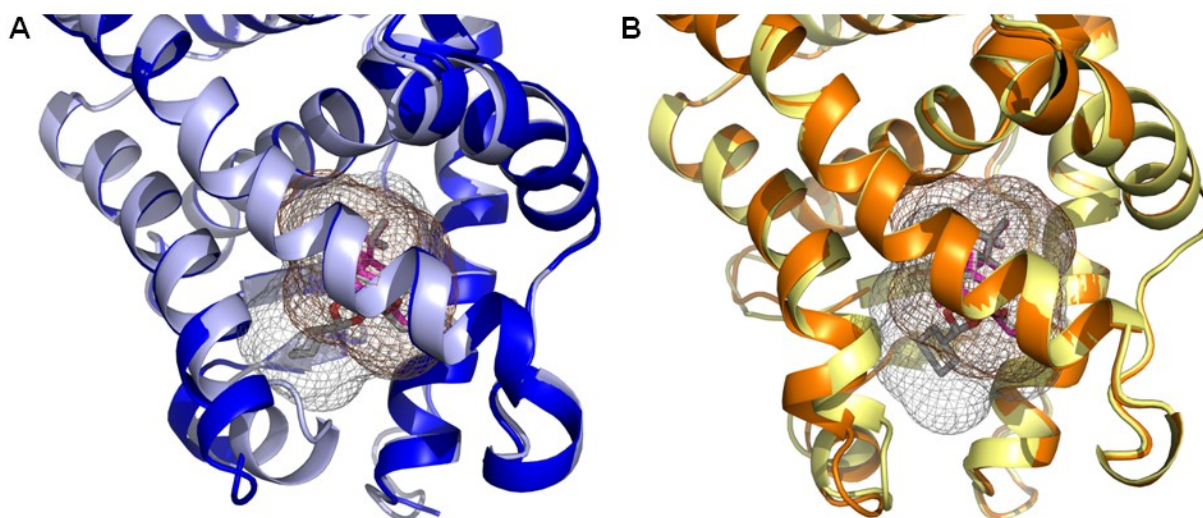
## **Supplemental Information**

**Ligand Binding Mechanism in Steroid Receptors:**

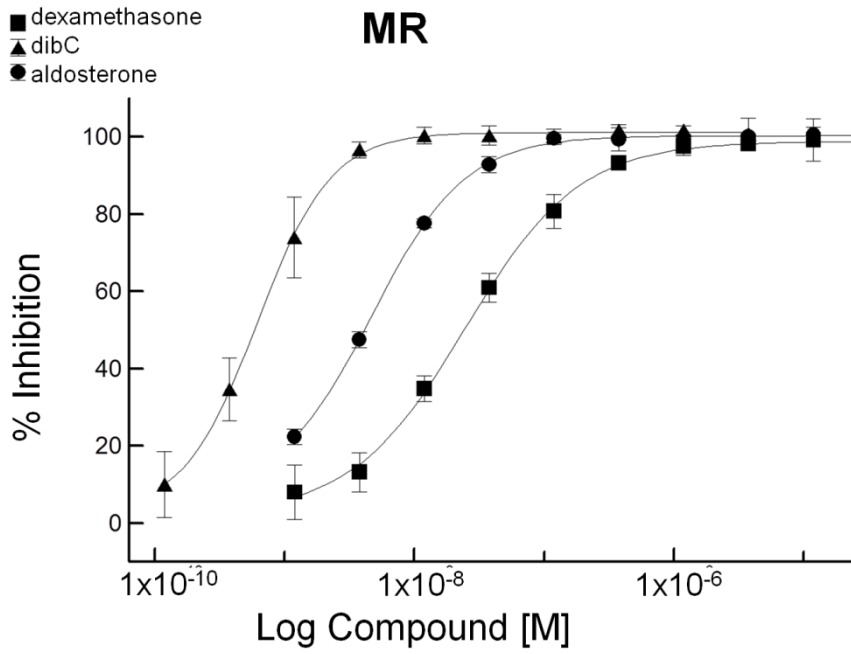
**From Conserved Plasticity to Differential**

**Evolutionary Constraints**

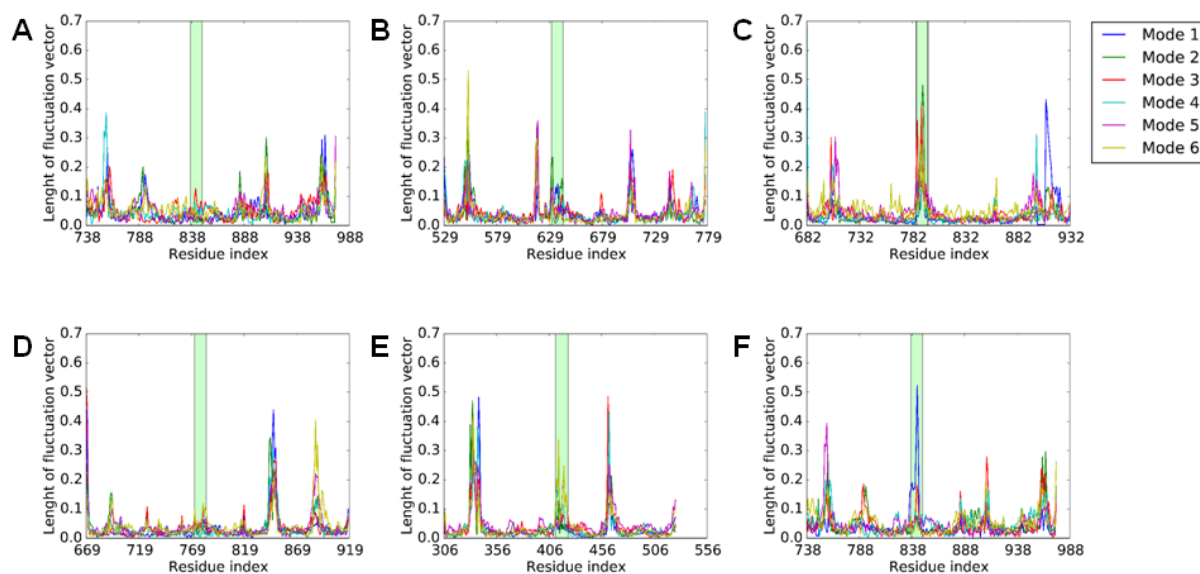
**Karl Edman, Ali Hosseini, Magnus K. Bjursell, Anna Aagaard, Lisa Wissler, Anders Gunnarsson, Tim Kaminski, Christian Köhler, Stefan Bäckström, Tina J. Jensen, Anders Cavallin, Ulla Karlsson, Ewa Nilsson, Daniel Lecina, Ryoji Takahashi, Christoph Grebner, Stefan Geschwindner, Matti Lepistö, Anders C. Hogner, and Victor Guallar**



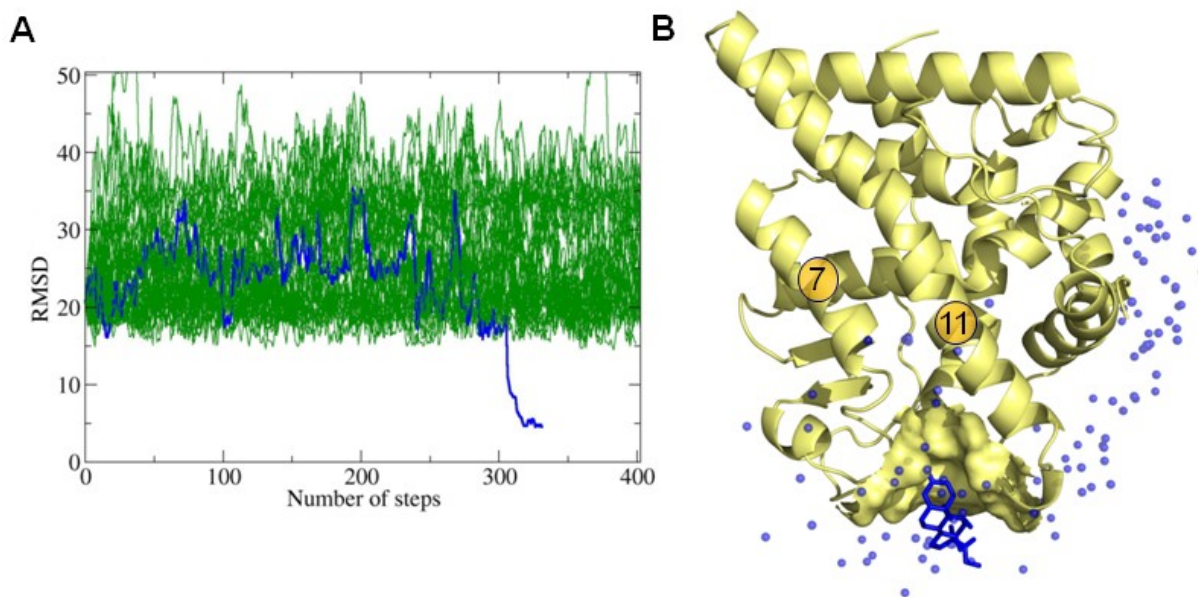
**Figure S1, Related to Figure 3.** Comparison of the volume of the ligand binding pocket in MR and GR in complex to Dexa and dibC. **(A)** The structure of MR (light blue) in complex with dexamethasone (magenta) overlaid on MR (dark blue) in complex with dibC (white). Ligand binding pockets are shown for MR:Dexa in brown (total volume  $543 \text{ \AA}^3$ ) and MR:dibC in gray (total volume  $714 \text{ \AA}^3$ ). **(B)** The structure of GR (pale yellow) in complex with dexamethasone (magenta) overlaid on the GR structure (orange) in complex with dibC (white). Ligand binding pockets are shown for GR:Dexa in brown (total volume  $572 \text{ \AA}^3$ ) and GR:dibC in gray (total volume  $661 \text{ \AA}^3$ ).



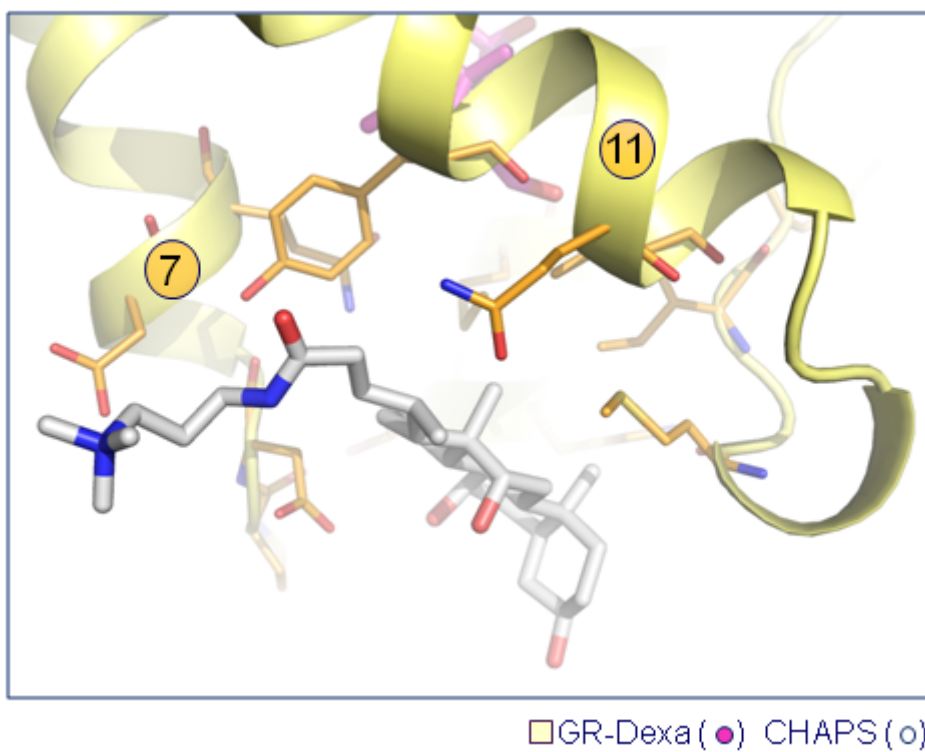
**Figure S2, Related to Figure 7C.** MR binding competition assay in the presence of dibC (triangle), aldosterone (circle), and dexamethasone (square). The corresponding  $IC_{50}$  values (mean  $\pm$  SD,  $n=3$ ) are:  $0.7 \pm 0.0$  nM (dibC);  $4.0 \pm 0.2$  nM (aldosterone);  $26.0 \pm 4.6$  nM (dexamethasone).



**Figure S3, Related to Figure 3.** Principal component analysis (PCA) for all X-ray structures of the steroid hormone receptors in the protein databank (PDB). The graphs show the amplitude of the top six modes from the PCA for MR (**A**), GR (**B**), PR (**C**), AR (**D**) and ER (**E**). The H6-H7 region which undergo the largest changes in the MR:dibC structure and the corresponding region in the other receptors are highlighted in green (MR: 837-848; GR: 631-642; PR: 786-797; AR: 772-783; and ER: 412-424). AR and MR exhibits the smallest variation in the H6-H7 region in the public domain structures. (**F**) The PCA of the MR public domain structures with MR:dibC added. In this analysis the mode describing the H6-H7 rearrangement becomes the dominant signal in the first mode.

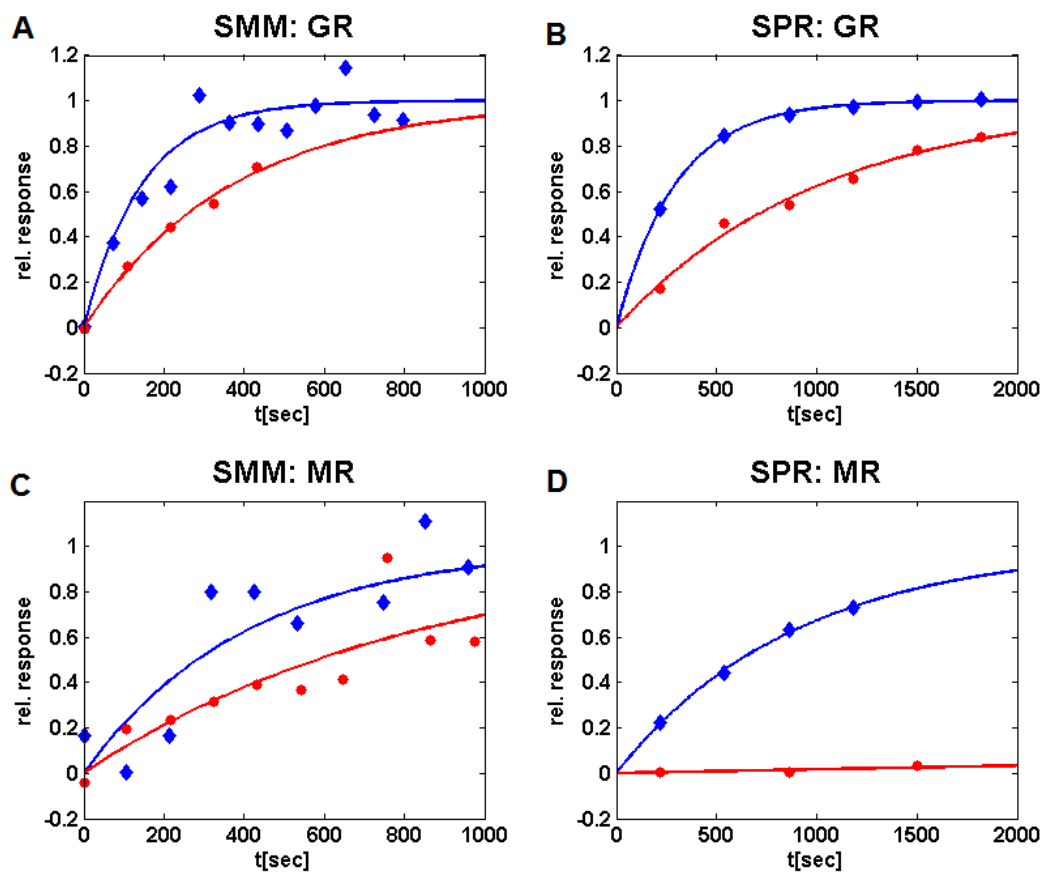


**Figure S4, Related to Figure 6.** Unbiased simulation of dexamethasone entering GR. **(A)** Each line represents the ligand heavy atom RMSD to the crystallographic structure for the total 64 trajectories. One of the trajectories represented by blue line enter the ligand binding pocket at step ~310. **(B)** The ligand's center of mass for the one trajectory that enter the binding pocket are shown as blue spheres. The region where the ligand enter the binding pocket is emphasized as a surface with the ligand shown in stick representation.

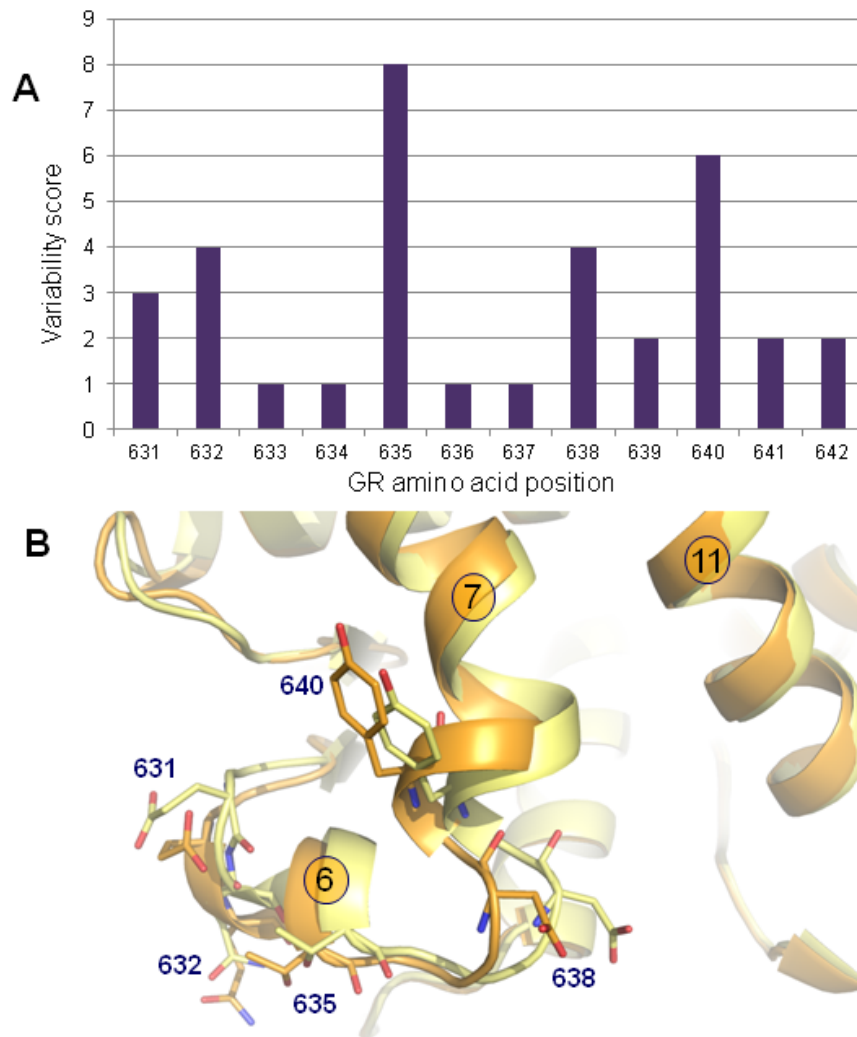


**Figure S5, Related to Figure 3.** The peripheral binding site. The structure of GR (yellow) in complex with dexamthasone (magenta) revealed that a CHAPS molecule (white) from the protein formulation is binding in between helices 7 and 11 about 12 Å away from the ligand binding pocket.





**Figure S6, Related to Table 2.** Residence time measurements of dexamethasone (blue diamonds) and dibC (red circles) bound to GR (**A, B**) and MR (**C, D**) using SMM (**A, C**) and SPR (**B, D**). Normalized change in receptor binding rate to surface-immobilized co-regulator peptide upon addition of >10-fold concentration excess of budesonide (GR) or aldosterone (MR). The extracted binding rates are fitted with  $k_+(t) = ae^{-k_{off}t} + c$  (colored solid lines). SMM and SPR experiments were conducted 20 °C and 10 °C, respectively.



**Figure S7, Related to Figure 8.** Variability score and structural arrangement of amino acids in the H6-H7 region in GR. (A) The GR variability scores plotted against the amino acid positions of the H6-H7 region. Higher scores indicate more variation at that position across the various species; a score of 1 indicate completely conservation. (B) Placement of amino acids with high variability score in the H6-H7 region in the GR:Dexa (yellow) and GR:dibC (orange) structures.

**Supplementary Movie, Related to Figure 4.** Unbiased simulation of dexamethasone entry into MR obtained with the PELE (Protein Energy Landscape Exploration) software. The simulated protein is shown in green, the NCOA1 peptide cofactor in yellow and dexamethasone ligand shown in light green. At the 0:27 timepoint, the MR:Dexa complex structure is overlaid onto the simulation for comparison with the protein in light blue and dexamethasone in magenta.

## **Supplemental experimental procedures**

### **Protein expression and purification for structure**

#### **GR:Dexa**

The cDNA sequence encoding the human GR-LBD (amino acids 500-777) with the mutations N517D, F602S and C638D and an N-terminal 6-histidine tag followed by a thrombin cleavage site was cloned into a pFastBac-HTb vector (Life Technologies). Recombinant baculovirus was generated using the Bac-to-Bac expression system (Life Technologies) and High Five cells (Life Technologies) were infected followed by suspension culture in Express Five medium (Gibco) for 48h at 27°C, the last 24h in the presence of 10  $\mu$ M dexamethasone, after which cells were collected by centrifugation. All protein purification steps were performed at 4°C. Cells were lysed in buffer A (50 mM Tris pH 8.0, 2.5 mM DTT, 1% CHAPS, 50  $\mu$ M dexamethasone and 10% glycerol) supplemented with Complete EDTA-free protease inhibitor cocktail (Roche) followed by affinity purification using Ni-NTA beads (Qiagen). Protein was eluted in buffer A supplemented with 150 mM NaCl and 300 mM imidazole, and subjected to size exclusion chromatography using a HiLoad 26/60 Superdex 200 gel filtration column equilibrated in buffer A. Five-fold molar excess of a TIF2 peptide, KENALLRYLLDK (Innovagen) was added, the N-terminal 6-histidine tag was removed using thrombin-agarose (Sigma) and subsequently the free 6-histidine tag was removed. The protein was thereafter passed over a Q Sepharose fast-flow ion-exchange column (GE Healthcare) equilibrated in buffer A and stored at -80 °C. Approximately 5.4 mg protein was obtained from 10 L High Five cells.

## **GR:dibC**

A pFastBac (Invitrogen) construct encoding human GR-LBD (amino acids 500-777) with the mutations N517D, V571M, F602S and C638D and an N-terminal, thrombin cleavable 6-His tag was used to generate baculoviruses in Sf9 cells (Invitrogen). GR-LBD encoding viruses were used to infect High Five cells (Invitrogen) at a density of  $2 \times 10^6$  cells/ml and a MOI of 3 in a Wave Bioreactor at 27°C. 24 hours post-infection, dexamethasone was added to a final concentration of 10 µM. The cells were harvested by centrifugation 48 hours post-infection, washed in PBS and stored at -80°C until lysis. Cells were resuspended in lysis buffer (50 mM Tris-HCl pH 8.0, 10% glycerol, 1% CHAPS, 2.5 mM DTT, Complete EDTA-free protease inhibitor cocktail (Roche) and 50 µM dexamethasone) and lysed by 5x1 min passes in a polytron homogeniser. The cell-lysate was clarified by centrifugation at 18500 g for 90 minutes and batch-bound to Ni-NTA Superflow (Qiagen) for 1.5 hours at 4°C. The IMAC resin was packed in a column, washed with wash buffer (50 mM Tris pH8.0, 60 mM NaCl, 30 mM imidazole, 10% glycerol, 1% CHAPS, 2.5 mM DTT and 50 µM dexamethasone) and GR-LBD was step eluted with elution buffer (50 mM Tris pH 8.0, 30 mM NaCl, 300 mM imidazole, 10% glycerol, 1% CHAPS, 2.5 mM DTT and 50 µM dexamethasone). The eluate was loaded on a HiLoad 26/60 Superdex 200 size exclusion column equilibrated in gel filtration buffer (50 mM Tris-HCl pH 8.0, 10% glycerol, 1% CHAPS, 2.5 mM DTT and 50 µM dexamethasone). GR containing fractions were pooled and a 3-fold excess of co-activator NR-box peptide (KENALLRYLLDK, human NCoA2, residues 740-751) was added. The His-tag was cleaved over night at 4°C with Thrombin-agarose (Sigma) and removed by negative IMAC using Ni-NTA. The protein was finally polished through Q Sepharose FF (GE

Healthcare) equilibrated in gel filtration buffer, flash-frozen in liquid nitrogen and stored at -80°C.

### **MR:Dexa and MR:dibC**

Human MR-LBD (amino acids 735-984) with the mutations C808S, C910S (and S810L in the case of dibC), an N-terminal, TEV cleavable 6-HN tag, and a C-terminal thrombin cleavable co-activator peptide PQAQQKSLQQLLQTE was cloned into pET24a(+). *Escherichia coli* BL21 Star™ (DE3) (Invitrogen) cells transformed with the expression vector were grown in terrific broth at 37°C until OD<sub>600</sub>=0.5-1.0, chilled on ice for 30 minutes and 100 µM of dexamethasone (Alfa Aesar) or dibC was added. Cells were shaken at 16°C for 30 minutes before protein production was induced using 0.1 mM isopropyl β-D-thiogalactopyranoside (IPTG) for an additional 24-48 hours. Cells were lysed in 30 mM Na-Hepes pH 7.5, 150 mM NaCl, 20 mM imidazole, 100 mM arginine-HCl, 10% glycerol, 1% CHAPS and 1 mM TCEP containing 20 µM of respective ligand, EDTA-free Complete protease inhibitor cocktail (Roche) and 0.05 g/ml of CellLytic™ Express (C1990, SIGMA), by rotation at room-temperature for 15 minutes. The lysate was cleared by centrifugation at 48000 g for 20 minutes and loaded onto Ni-Sepharose FF (GE Healthcare) equilibrated in lysis buffer. After washing, protein was step eluted by the addition of one column volume (CV) of lysis buffer containing 0.5 M Arginine-HCl followed by 5 CV of elution buffer (30 mM Na-Hepes, pH 7.5, 150 mM NaCl, 500 mM imidazole, 500 mM arginine-HCl, 10% glycerol, 1% CHAPS, 1 mM TCEP and 20 µM of respective ligand). Size exclusion chromatography was performed on a HiLoad Superdex 200 column (GE Healthcare) equilibrated in 20 mM Na-Hepes pH 6.7, 150 mM NaCl, 0.5 M arginine-HCl, 10% glycerol, 0.1% CHAPS, 1 mM TCEP and 2 µM dexamethasone or dibC. Finally, MR-LBD co-expressed with dexamethasone was diluted 10x in 20

mM Tris-HCl pH 8.0, 10 mM CaCl<sub>2</sub> and 20 μM dexamethasone, cleaved with TEV protease and Thrombin CleanCleave Kit (SIGMA), purified by reverse IMAC on Ni-Sepharose FF and concentrated to 15 mg/ml. MR-LBD co-expressed with dibC was diluted 15x in 10 mM Tris-HCl pH 8.5, 20 μM dibC and 1mM TCEP and concentrated to 7 mg/ml.

## **Protein expression and purification for biophysical characterization**

### **GR**

Human GR-LBD (amino acids 529-777) was cloned into the pET24a vector (Novagen) featuring an N-terminal His<sub>6</sub>-tag and a TEV protease cleavage site. The expression vector was transformed into E. coli BL21(DE3) STAR, followed by expression in PASM-5052 autoinduction medium. 100 μM dexamethasone was added after the cell culture reached an OD of 0.6 followed by expression over 48 hours at 16 °C. All purification buffers were degassed and contained 2 mM TCEP and 50 μM dexamethasone. The harvested cells were resuspended in lysis buffer (50 mM Tris pH 8, 10% glycerol, 1% CHAPS) supplemented by protease inhibitors (Complete, Roche) and DNase. Cells were lysed by sonication. The cleared lysate was applied to a nickel affinity column equilibrated with wash buffer (50 mM Tris pH 8, 10% glycerol, 1% CHAPS, 60 mM NaCl) and eluted by a 300 mM imidazole gradient. Remaining impurities were removed by an additional superdex 200 gelfiltration step using 50 mM Tris buffer at pH 9 as running buffer followed by storage at -80°C.

### **MR**

Human MR-LBD (amino acids 712-984) with the mutation C808S and an N-terminal, TEV cleavable 6-HN tag was cloned and expressed in the same way as the MR-LBD proteins used for structure determination. The cells were lysed in 50 mM Tris-



HCl, pH 8.0, 500 mM NaCl, 100 mM arginine-HCl, 1% CHAPS, 20 mM imidazole, 10% glycerol, 1mM TCEP, 50  $\mu$ M dexamethasone, EDTA-free Complete protease inhibitor cocktail (Roche) and 0.05 g/ml of CellLytic™ Express (C1990, SIGMA). The lysate was cleared by centrifugation at 48000 g for 20 minutes and loaded onto a HisTrap HP column (GE Healthcare). The protein was gradient eluted with 50 mM Tris-HCl, pH 8.0, 500 mM NaCl, 500 mM arginine-HCl, 1% CHAPS, 0- 300 mM imidazole, 10% glycerol, 1mM TCEP, 50  $\mu$ M dexamethasone.

## **Crystallization**

### **GR:Dexa**

A tube with 1.0 mg of GR(500-777) N517D, F602S and C638D was thawed and washed three times in the concentrator tube with 3.5 ml of 10 mM Tris pH 8.5, 2.5 mM DTT and 45 $\mu$ M dexamethasone. A fivefold molar excess of co-activator NR-box peptide (KENALLRYLLDKDD, human NCoA2, residues 740-753) was added and the complex was concentrated to 9 mg/ml.

Crystals were grown at 4°C in hanging drops using 1  $\mu$ l of protein and 1  $\mu$ l of well solution (10% PEG8000, 10% ethylene glycol and 0.1 M Hepes pH 7.5). Crystals were frozen in liquid nitrogen with 20% ethylene glycol as cryo protectant prior to data collection.

### **GR:dibC**

A tube with 5.0 mg's of GR(500-777) N517D, V571M, F602S and C638D was thawed and concentrated to about 1.5 ml. The protein was washed three times in the concentrator tube with 10 ml of 10 mM Tris pH 8.5, and 2.5 mM DTT (buffer B) to remove excess of dexamethasone and thereafter diluted to a final volume of 6 ml. dibC was added to a final concentration of 0.25 mM to boost ligand exchange prior to

dialysis. Dialysis was performed using two Slide-A-Lyzer dialysis cassettes in a beaker containing buffer B and 60  $\mu$ M of dibC. Dialysis solution was exchanged after 20, 28 and 46 hours before harvesting the sample. The protein was concentrated to 1 ml and buffer was exchanged to fresh buffer B using a NAP10 column. A twofold molar excess of co-activator NR-box peptide (KENALLRYLLDKDD, human NCoA2, residues 740-753) was added and the complex was concentrated to 9 mg/ml.

Crystals were grown at 4°C in hanging drops using 2  $\mu$ l of protein and 1  $\mu$ l of well solution (10% PEG8000, 20% ethylene glycol and 0.1 M Hepes pH 7.5). Crystals appeared as rod like crystals after 1-2 days but continued to grow for one to two weeks. Crystals were frozen in liquid nitrogen without any cryo protectant prior to data collection.

#### **MR:Dexa**

Crystals of MR(735-984) C808S and C910S co-expressed and purified with dexamethasone were grown by sitting drop vapor diffusion in 30% PEG4000, 0.1 M NaCl and 0.2 M Pipes pH 7.4. Crystals were cryo-protected in well solution supplemented with 20% glycerol and flash frozen in liquid nitrogen.

#### **MR:dibC**

Crystals of MR(735-984) C808S, C910S and S810L co-expressed and purified with dibC were grown by sitting drop vapor diffusion in 18% PEG4000, 0.14 M LiSO<sub>4</sub>, 85 mM Tris pH 8.5 and 15% glycerol. Crystals were flash frozen in liquid nitrogen.

#### **Data collection and structure determination.**

The MR:Dexa data were collected using an Rigaku FRE rotating anode (wavelength 1.54 Å). The GR:Dexa data were collected at the ID14:4 beam line at the ESRF

(wavelength 0.94 Å). The MR:dibC and GR:dibC data were collected at the ID29 beam line at the ESRF (wavelength 0.98 Å). All data sets were collected from a single crystal at 100K. The MR data sets were integrated with XDS (Kabsch et al., 2010) and the GR data sets were integrated with Mosflm (Leslie et al., 2007). All data sets were merged with SCALA (Evans et al., 2006) from the CCP4 suite (Collaborative Computational Project., 1994). The MR and GR structures were solved with PHASER (McCoy et al., 2007) using PDB entry 2AA2 and 1M2Z as starting models, respectively. The structures were refined using the BUSTER (Bricogne et al., 2011) and manual rebuilding using Coot (Emsley et al., 2004). The GR:Dexa structure had 1 (0.39%) Ramachandran outlier while the other structures did not have any outliers. All figures were prepared using PyMOL ([www.pymol.org](http://www.pymol.org)).

### **Structural analysis**

Cavity volumes were calculated with fpocket 2.0 (Le Guilloux et al. 2009). For a higher accuracy, the default number of Monte Carlo steps was increased from 2500 to 500000. The minimum size of alpha spheres was set to 3.5 Å to avoid connecting buried cavities (default value 3.0 Å).

PCA analysis was performed using ProDy 1.5.1 (Bakan et al. 2011) For each receptor, all public available structures were included in the analysis and one structure was selected as the reference structure (MR:Dexa, GR:Dexa, 1E3G (AR), 1A28 (PR), 1A52 (ER)). The sequence of monomer A from each protein was aligned to the sequence of the reference structure filtering out structures with less than 90% sequence identity and subsequently superimposed. The first six principal

components were plotted against the residue number by calculating the length of the x,y,z-fluctuation vector for each c-alpha atom.

### **Mineralocorticoid receptor ligand competition binding assay**

Human MR-LBD (729-984) with an N-terminal maltose binding protein (MBP) tag was expressed using the Bac-to-Bac expression system (Life Technologies). High Five cells were co-infected with recombinant P23 co-chaperone baculovirus followed by suspension culture in Express Five medium (Gibco) for 48h at 27°C. Cells were lysed in lysis buffer (10 mM Tris-HCl pH 7.4, 0.5 mM EDTA, 2.5 mM DTT, 10% glycerol, 20 mM Na<sub>2</sub>MoO<sub>4</sub> and Complete protease inhibitor (Roche)) followed by centrifugation. The supernatant was stored at -80°C. Compound binding was assessed using a ligand competition binding scintillation proximity assay (SPA). Compounds were incubated with MR-High Five cell lysate (7µg/ml) and 5 nM <sup>3</sup>H-aldosterone (Perkin Elmer NET419250UC) in assay buffer (10 mM Tris-HCl, 0.5 mM EDTA, 20 mM Sodium molybdate dehydrate, 10 % Glycerol and 0.1 mM DTT) for one hour before addition of 2.5 mg/ml anti-rabbit SPA PS beads (Perkin Elmer RPNQ0299) and 2 µg/ml rabbit anti-MBP antibodies (Abcam ab9084) followed by incubation at room temperature for 8 hours before detection of signal using a LeadSeeker imaging system (GE Healthcare). K<sub>i</sub> values were derived using the equation  $K_i = (IC_{50} - \text{receptor Concentration}/2) / (1 + \text{ligand Conc}/K_m)$ , where receptor concentration was set to zero, ligand concentration to 0.005 µM and K<sub>m</sub>-value to 0.0016 µM.

## **Biophysical characterization and residence time determination**

Residence time measurements of GR/MR:dexamethasone and dibC was determined using single molecule microscopy (SMM) and SPR (Biacore) by probing the time-resolved change in receptor binding to surface-immobilized co-regulator peptides (GR: Biotin-PRGC1\_130-155 / MR: PRGC2\_146-166). HBSP(+) buffer (10 mM HEPES, 150 mM NaCl, 0,005% P20, pH=7.4) was used for all measurements. For SPR, the two biotinylated peptides was immobilized on a streptavidin chip (GE healthcare) using a Biacore 3000 (GE healthcare) to 500-1000 RU. Budesonide/aldosterone was added to a final concentration of 25  $\mu$ M to a solution of 130 nM GR/MR, preequilibrated with 1  $\mu$ M dexamethasone/dibC. Directly after budesonide/aldosterone addition, receptor binding rate to the cofactor peptide was monitored by consecutive injection cycles (1 min injections). The peptide surface was regenerated with 0.005% SDS after each injection. To compensate for potential protein degradation over the time course of the measurement, the data was normalized to a reference sample containing only 1  $\mu$ M dexamethasone/dibC. For SMM, the respective NHR was bound via 6 $\times$ His-tag to liposomes containing POPC, DGS-NTA, lissamine rhodamine B sulfonyl in a ratio of 1:0.02:0.01. Liposomes were prepared as described by Gunnarsson Anal chem. 2015. The coregulator peptides were mixed with Neutravidin (NA) in a 1:1 molar ratio. Subsequently, the coregulator peptide-NA complex was incubated at 50  $\mu$ g/ml NA with TL1 cleaned PLL-g-PEG/ PLL-g-PEG-biotin (1:1, Surface Solutions) coated glass surfaces. Budesonide/aldosterone was added to a final concentration of 10  $\mu$ M to a 150 pM liposome-NHR solution containing 1  $\mu$ M dexamethasone/dibC. To compensate for potential protein degradation over time the data was normalized to a reference sample of 150 pM liposome-NHR solution containing only 1  $\mu$ M dexamethasone/dibC. Image data was collected on an inverted

microscope (Nikon Ti Eclipse) equipped with a 60x oil immersion objective (NA = 1.49), TRITC filter cube, perfect focus system and air cooled sCMOS (Orca Flash 4.0 v2 Hamamatsu). For imaging in an iterative fashion, 10 sec time series at 10Hz framerate were recorded for the competition and the reference well at two different positions continuously over ~15 minutes. Images were analyzed using custom made Matlab (Mathworks) routines to extract the liposome-NHR conjugate binding rate to the surface. The liposome-NHR binding rate during each time series (10 sec) was assumed to be constant and hence, the vesicle binding rate was extracted by linear regression to the cumulative number of binding liposomes as a function of time. To compensate for surface preparation inhomogeneities the data of the two different positions in each well were averaged. The extracted binding rates were plotted over time and fitted with  $k_+(t) = ae^{-k_{off}t} + c$ .

### **Supplemental references**

Bakan, A., Meireles, L.M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* 27, 1575–1577.

Bricogne, G., Blanc, E., Brandl, M., Flensburg, C., Keller, P., Paciorek, C., Roversi, P., Sharff, A., Smart, O. S., Vonrhein, C., et al. (2011). BUSTER version 2.11.5. Cambridge, United Kingdom; Global Phasing Limited.

Collaborative Computational Project, Number 4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 50, 760–763.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* *60*, 2126–2132.

Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.* *62*, 72–82.

Kabsch, W. (2010). XDS. *Acta Crystallogr. D Biol. Crystallogr.* *66*, 125–132.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinform.* *10*, 168.

Leslie, A.G.W., and Powell, H.R. In *Evolving Methods for Macromolecular Crystallography* 41–51 (Springer, Dordrecht, The Netherlands, 2007).

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. *J. Appl. Crystallogr.* *40*, 658–674.

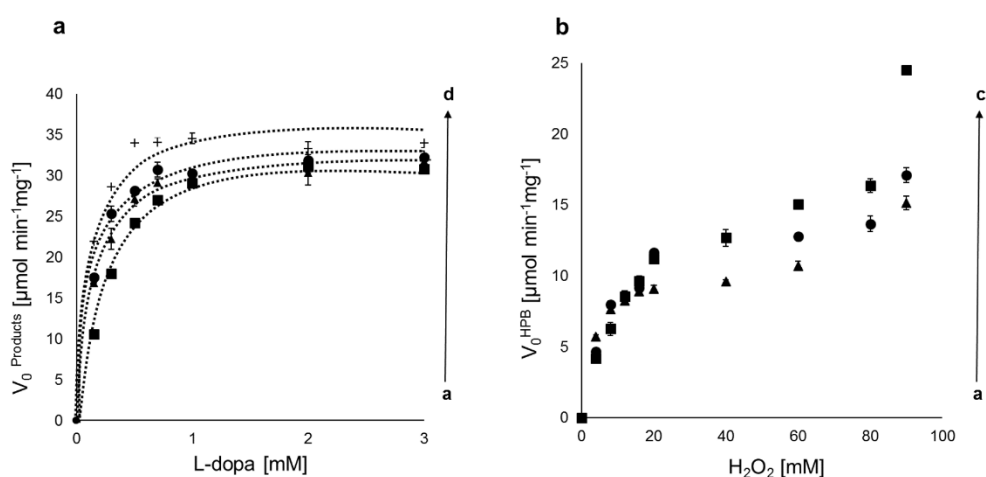


## Supporting information paper 3

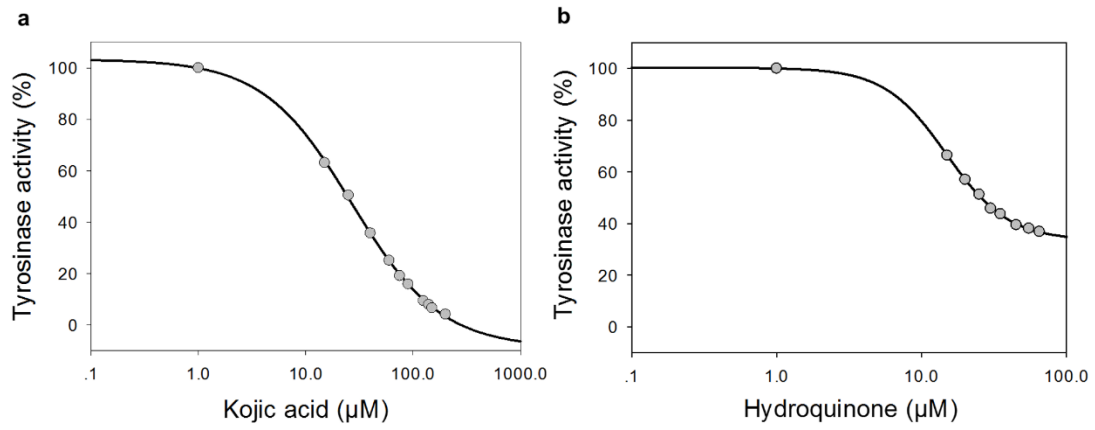
## Supplementary Data

### The unravelling of the complex pattern of tyrosinase inhibition

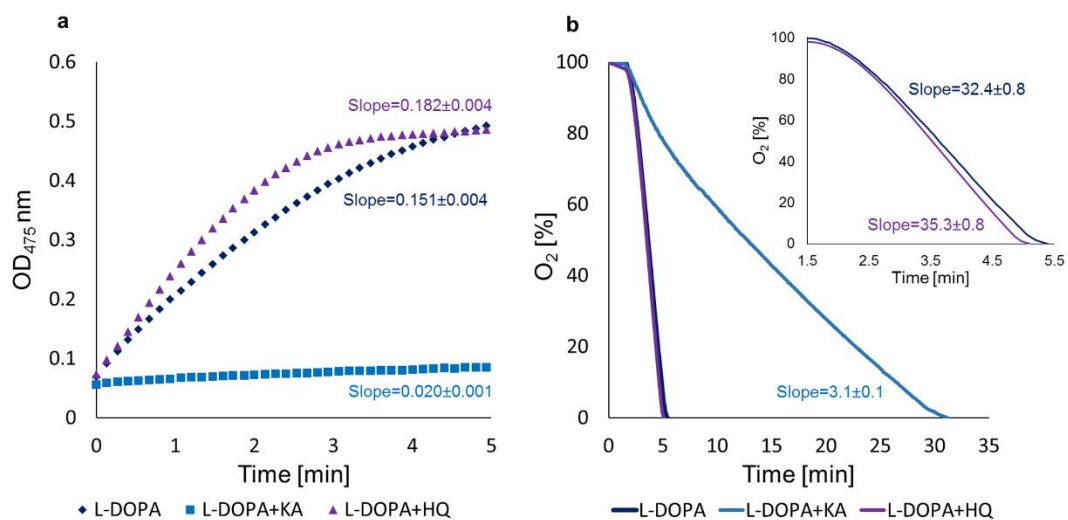
Batel Deri<sup>a,§</sup>, Margarita Kanteev<sup>a,§</sup>, Mor Goldfeder<sup>a</sup>, Daniel Lecina<sup>b</sup>, Victor Guallar<sup>b,c</sup>,  
Noam Adir<sup>d</sup>, and Ayelet Fishman<sup>a,†</sup>



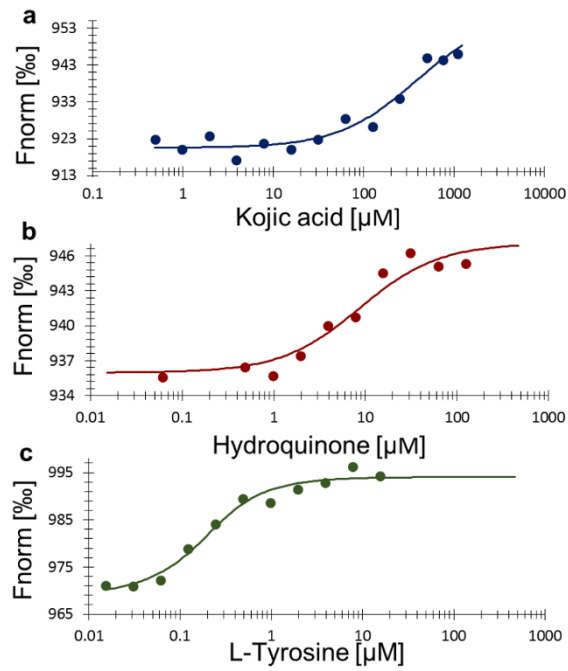
**Supplementary Figure S1:** Initial product formation rate of TyrBm activity on HQ in the presence of reducing agents measured at 475 nm. (a) HQ as the substrate in the presence of increasing concentrations of L-dopa (0-3 mM). HQ concentrations (mM) were: (a ■) 0, (b ◆) 0.025, (c ●) 0.075 and (d +) 0.5. The product may be composed of HPB (2-hydroxy-*p*-benzoquinone) and dopa-quinone since both HQ and L-dopa may serve as substrates. (b) HQ as the substrate in the presence of increasing concentrations of  $\text{H}_2\text{O}_2$  (0-90 mM). HQ concentrations (mM) were: (a ▲) 0.1, (b ●) 0.5 and (c ■) 1.5. The product is HPB only. The reactions contained  $6 \mu\text{g ml}^{-1}$  of purified TyrBm, 50 mM PBS buffer pH 7.4 and 0.01 mM  $\text{CuSO}_4$ .



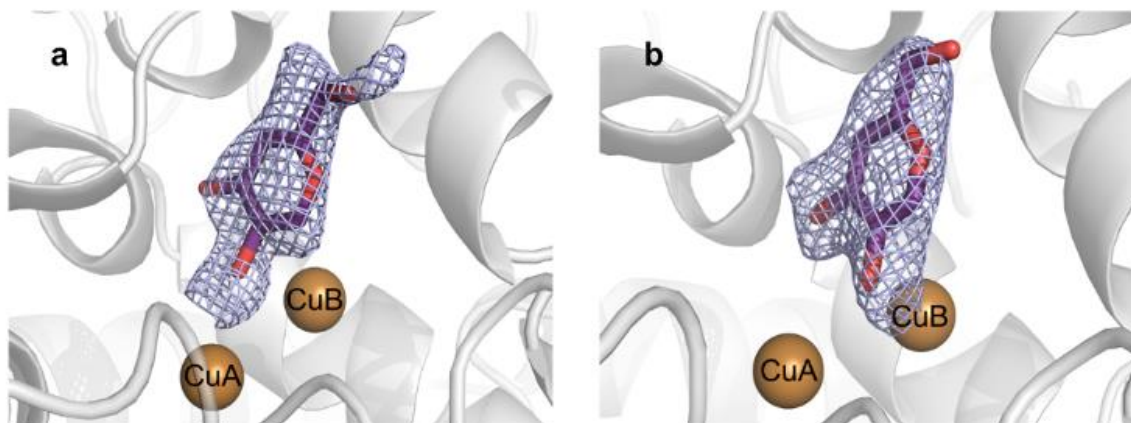
**Supplementary Figure S2:** Inhibition of TyrBm monophenolase activity by (a) KA and (b) HQ. IC50 values were determined at 50% activity.



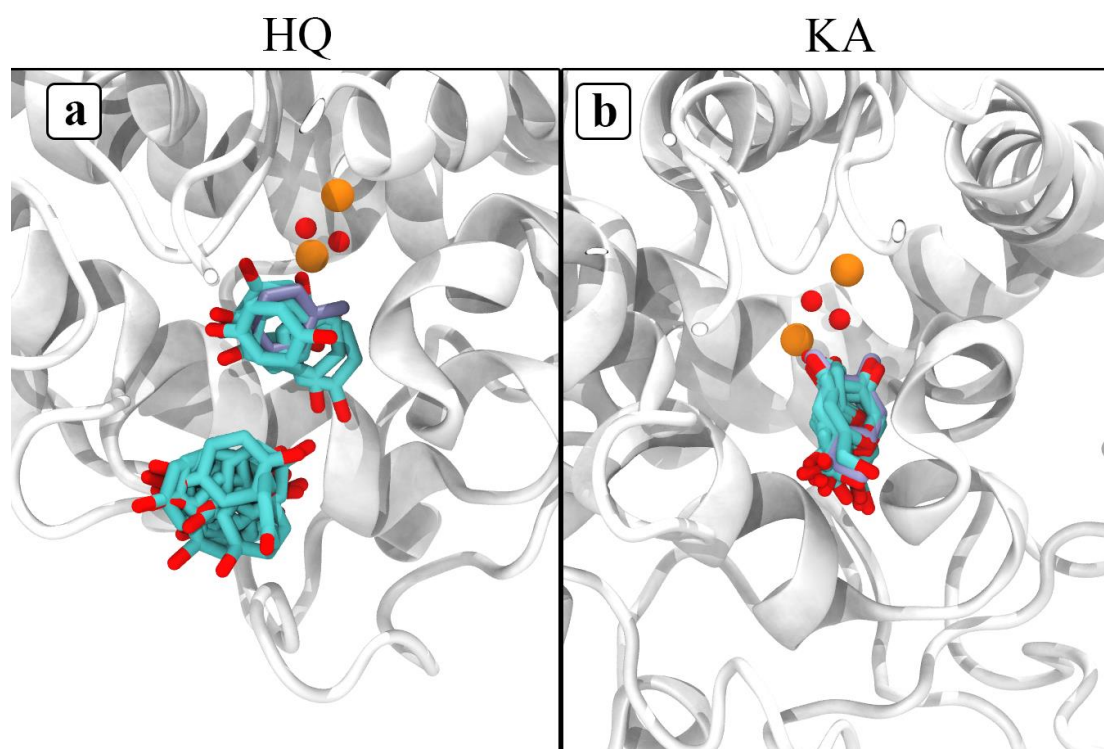
**Supplementary Figure S3:** Action of TyrBm on L-dopa alone or in the presence of KA and HQ. (a) Spectrophotometric recordings measured at 475 nm. All measurements were performed in heptaplicates. (b) Oxygen consumption recordings. The inset presents the oxygen consumption in the presence of L-dopa and L-dopa with HQ in the first 5 minutes. All measurements were performed in triplicates. The reactions contained 4  $\mu\text{g ml}^{-1}$  of purified TyrBm, 50 mM PBS buffer pH 7.4, 0.01 mM  $\text{CuSO}_4$ , 1 mM L-dopa, 0.1 mM HQ and 0.1 mM KA. The slopes representing the activity rate are given for each graph. KA clearly inhibits TyrBm activity, whereas HQ enhances activity.



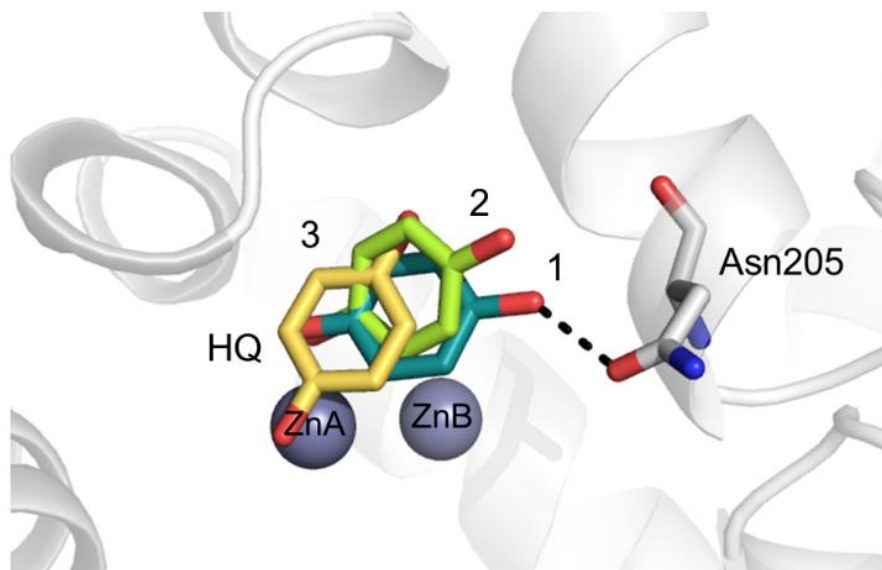
**Supplementary Figure S4:** Microscale thermophoresis (MST) analysis showing the binding of TyrBm-ligand by change in fluorescence. (a) Titration of rising concentrations of KA (0-4 mM) induces MST signal and yields  $K_D$  of  $377 \pm 4 \mu\text{M}$ . (b) Titration of rising concentrations of HQ (0-1mM) yields  $K_D$  of  $9 \pm 1 \mu\text{M}$ . (c) Titration of rising concentrations of L-tyrosine (0-2mM) yields  $K_D$  of  $0.10 \pm 0.03 \mu\text{M}$ . The reactions contained constant concentration of TyrBm ( $0.377 \mu\text{M}$ ) in 50 mM PBS buffer pH 7.4.



**Supplementary Figure S5:** KA with its mF<sub>o</sub>-DF<sub>c</sub> electron density omit map (light blue wire) contoured at 2σ. (a) KA in the active site of monomer A, structure 5I38. (b) KA in the active site of monomer B, structure 5I38. Copper ions are presented as brown spheres.

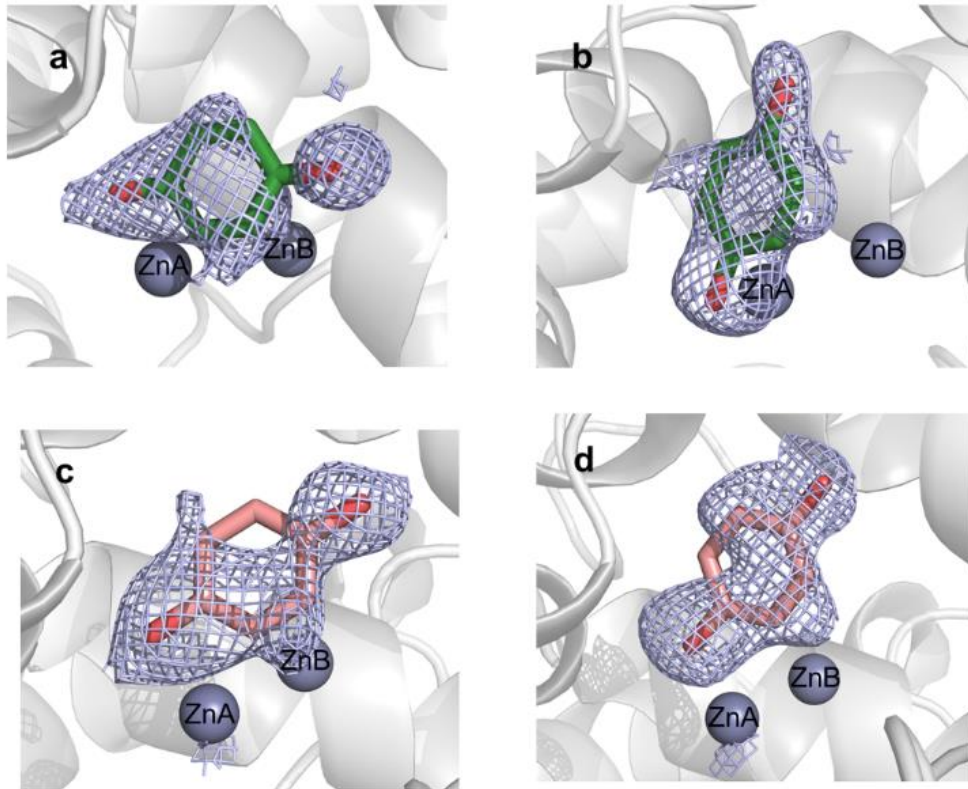


**Supplementary Figure S6:** Representative structures from the MSM clusters for HQ (panel a) and KA (panel b) within 1 kcal/mol of the best bound complex (highlighted in darker blue). Copper ions are presented as brown spheres, water molecules in red.



**Supplementary Figure S7:** Positions of HQ in the active site of TyrBm. Superposition of the active site of TyrBm monomers, obtained from two crystal structures with HQ in the active site. In orientation 1 of HQ (deep teal), the hydroxyl side chain forms a hydrogen bond with Asn205 (monomer A, 5I3B). In orientation 2 (green), HQ is oriented similar to tyrosinase substrates (monomer B, 5I3B). The different orientations of HQ (1, 2 and 3) represent flexibility in the active site. Zinc ions are presented as grey spheres. HQ in the active site of monomer A, 5I3A, is positioned in orientation 1 as well (data not shown).





**Supplementary Figure S8:** HQ with its  $mF_o-DF_c$  electron density omit map (light blue wire) contoured at  $2\sigma$ . (a) HQ (in orientation 1) in the active site of monomer A, structure 5I3A. (b) HQ (in orientation 2) in the active site of monomer B, structure 5I3A. (c) HQ (in orientation 1) in the active site of monomer A, structure 5I3B and (d) HQ (in orientation 3) in the active site of monomer B, structure 5I3B. Zinc ions are presented as grey spheres.

## Supporting information paper 4

**Biophysical Journal, Volume 112**

**Supplemental Information**

**Exploring Binding Mechanisms in Nuclear Hormone Receptors by  
Monte Carlo and X-ray-derived Motions**

**Christoph Grebner, Daniel Lecina, Victor Gil, Johan Ulander, Pia Hansson, Anita Dellsen, Christian Tyrchan, Karl Edman, Anders Hogner, and Victor Guallar**

|   |     |
|---|-----|
| PDB-ids of the receptor structures which were used to calculate the PCA modes | SI2 |
| Per residue displacement using ANM and PCA modes.                             | SI3 |
| Results for PELE simulations using PCA modes                                  | SI4 |
| Implied time scales for MSM   | SI5 |
| Chapman-Kolmogorov test for the MSM   | SI6 |
| Relative entropy plots for the MSM  | SI7 |
| Plots for committor probability for PCCA+ analysis                            | SI8 |
| Bibliography  | SI9 |

**Table S1. PDB-ids used for calculating the PCA modes**

|    |  |
|----|--|
| AR | 1E3G 1T63 1T79 1XJ7 1Z95 2AX6 2HVC 2PIP 2PIV 2Q7I 2YHD 3B5R 3G0W 3V49 1GS4 1T65<br>1T7F 1XNN 2AM9 2AX7 2IHQ 2PIQ 2PIW 2Q7J 2YLO 3B65 3L3X 3V4A 1I37 1T73 1T7M<br>1XOW 2AMA 2AX8 2NW4 2PIR 2PIX 2Q7K 2YLP 3B66 3L3Z 3ZQT 1I38 1T74 1T7R 1XQ2<br>2AMB 2AX9 2OZ7 2PIT 2PKL 2Q7L 2YLQ 3B67 3RLJ 4HLW 1T5Z 1T76 1T7T 1XQ3 2AO6 2AXA<br>2PIO 2PIU 2PNU 2QPY 2Z4J 3B68 3RLL |
| ER | 1G50 1R5K 1X7R 1ZKY 2B23 2G50 2JFA 2Q6J 2QAB 2QH6 2R6W 1A52 1L2I 1SJ0 1XPC 2B1V<br>2BJ4 2I0J 2OUZ 2Q70 2QE4 2QR9 2R6Y 1ERE 1PCG 1UOM 1XQC 2B1V 2FAI 2I0K 2P15 2QA6<br>2QGT 2QSE 3ERD 1ERR 1QKU 1QKT 1X7E 1YIN 2B1Z 2G44 2JF9 2POG 2QA8 2QGW 2QXM<br>3ERT   |
| GR | 1M2Z 1P93 2Q1V 3BQD 3E7C 3K22 3MNE 3MNP 4E2J 4P6X 2Q1H 2Q3Y 3CLD 3GN8 3K23<br>3MNO 3RY9  |
| MR | 1Y9R 1YA3 2A3I 2AA2 2AA5 2AA6 2AA7 2AAX 2AB2 2ABI 2OAX 3VHU 3VHV 5XHK  |
| PR | 1A28 1SQN 1ZUC 2OVM 3D90 3G8O 3KBA 3ZRA 4A2J 1E3K 1SR7 2OVH 2W8Y 3G8N 3HQ5<br>3ZR7 3ZRB 4APU   |

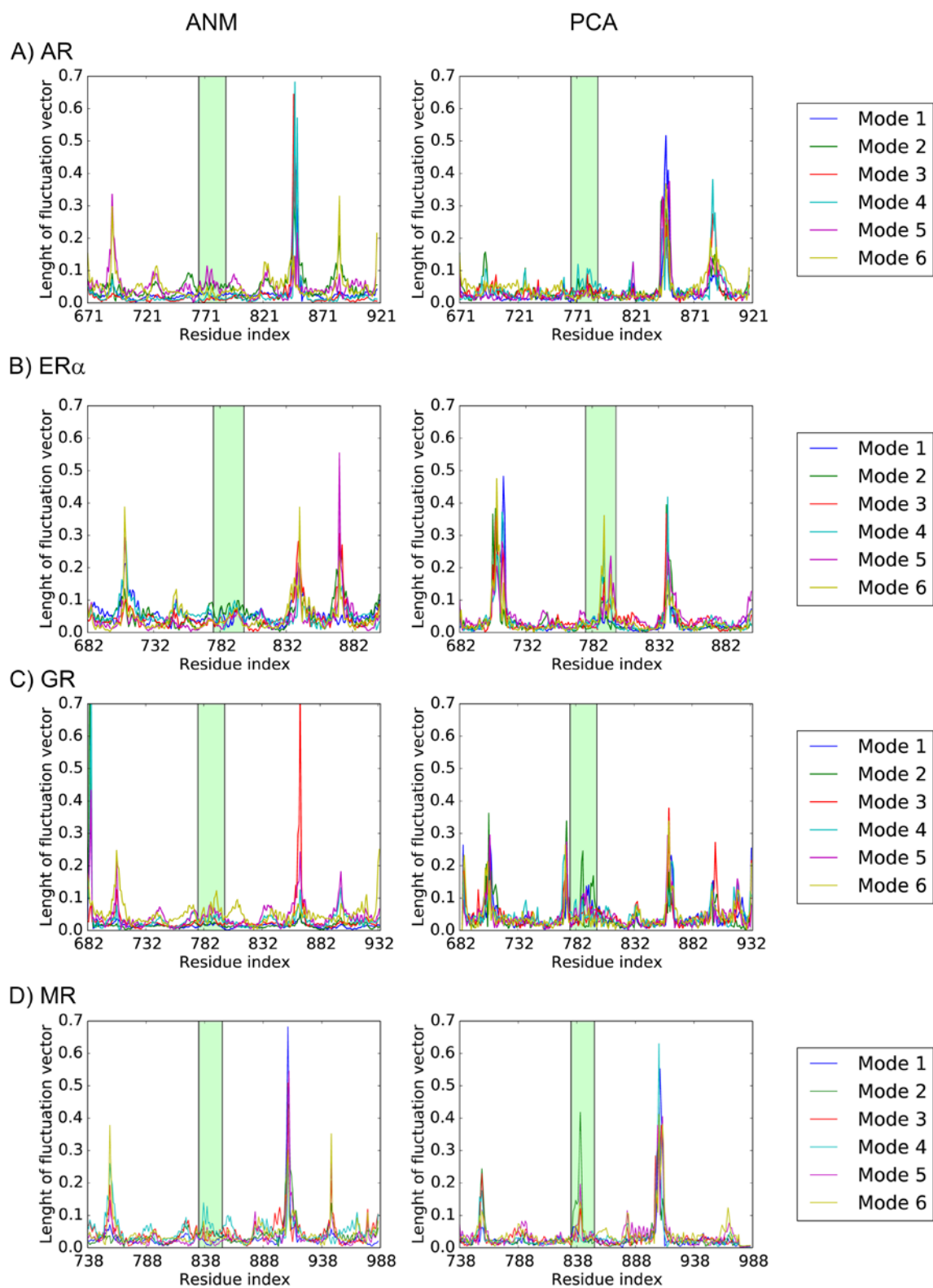


Figure S1 Visualization of the per residue displacement magnitude of the lowest 6 ANM and first 6 PCA modes, for the rest of systems. The helix 6-7 loop region (residue 833 to 853 for MR) is highlighted in green. A) Androgen receptor, B) Estrogen receptor  $\alpha$ , C) Glucocorticoid receptor, D) Mineralocorticoid receptor.

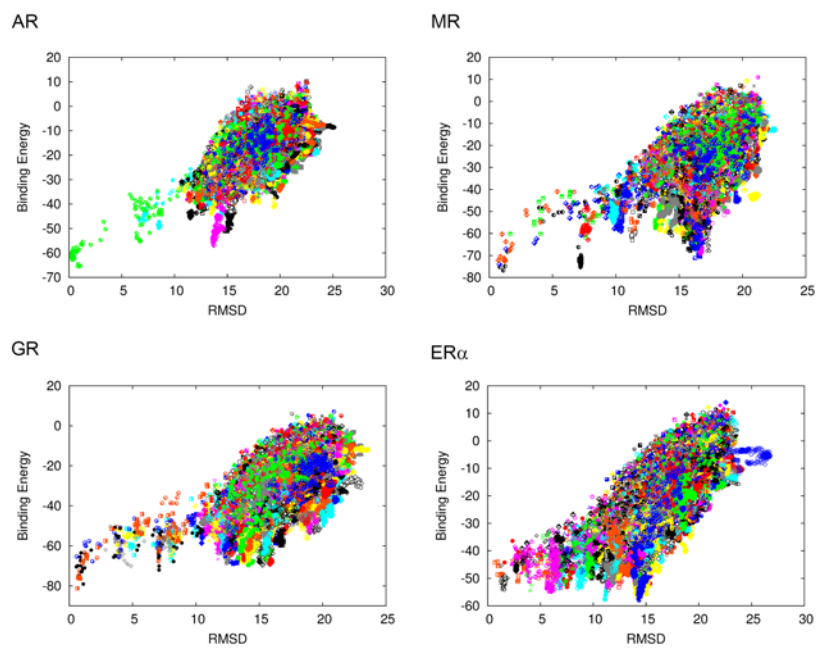


Figure S2 Results for PELE simulations using PCA modes for AR, ER $\alpha$ , GR, and MR. Plots show the correlation of the ligand heavy atom RMSD to the bound crystal (in angstroms), and the binding energy (in kcal/mol). Each color and symbol corresponds to an independent trajectory from the PELE sampling.

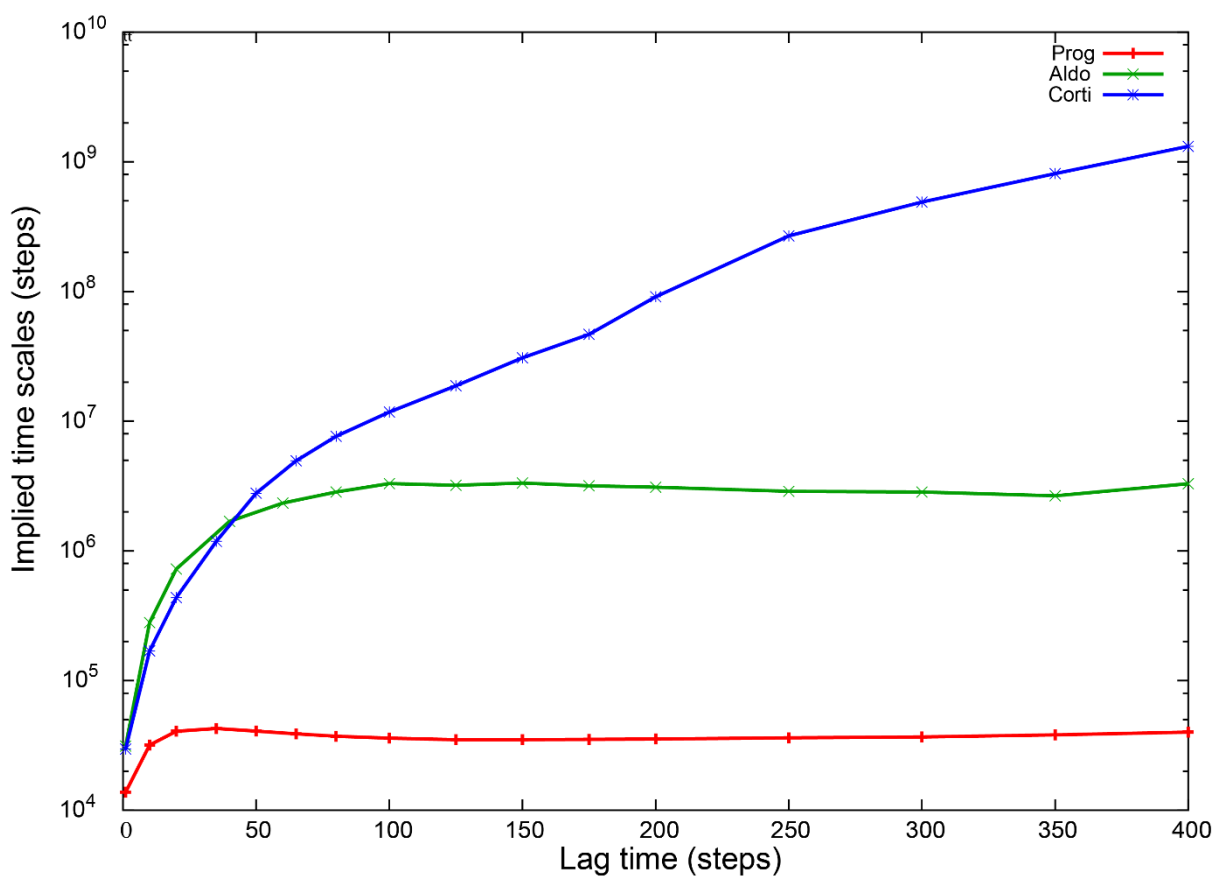


Figure S3 Implied time scales of the slowest relaxation mode for the three studied ligands. The MSM was generated with 600 independent 24h simulations, using k-means as the clustering method for the ligand center of mass. Convergence is achieved for all ligands at 300 steps.



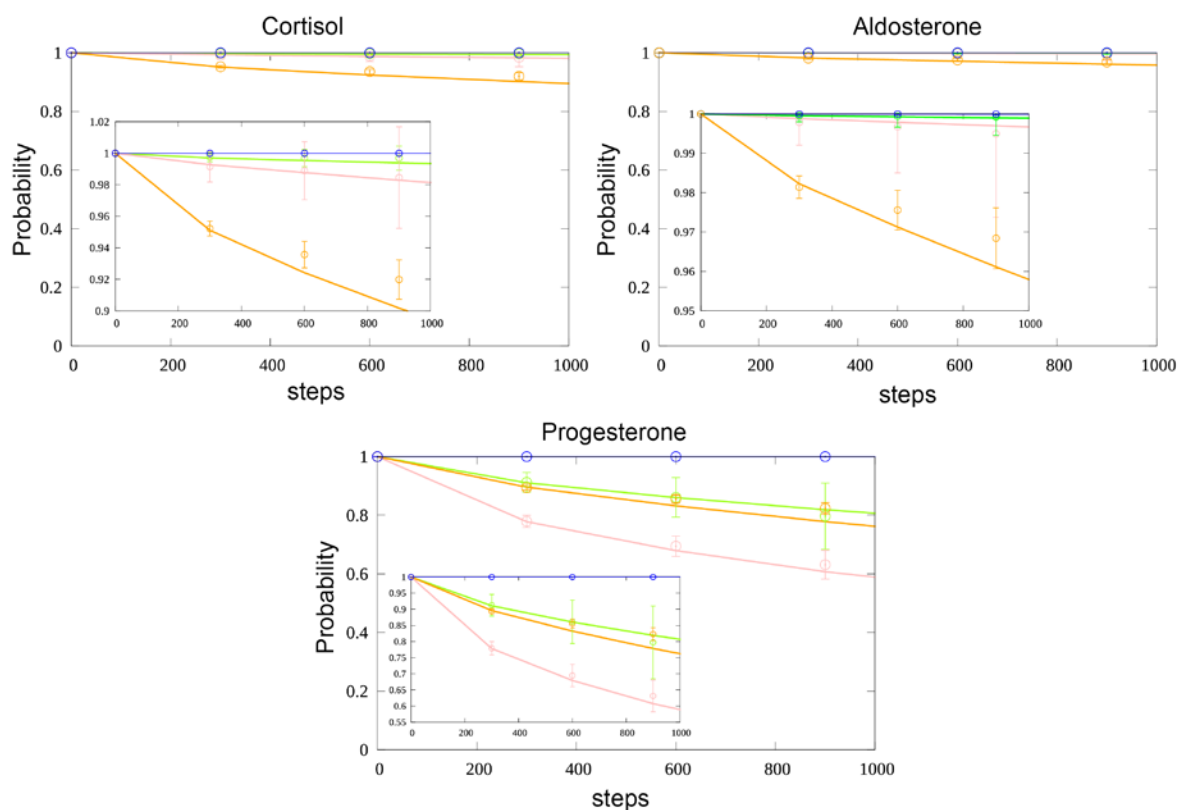


Figure S4. Chapman-Kolmogorov test, which checks the equality  $P(k\tau) \approx P^k(\tau)$  for the main PCCA+ clusters at multiple times of the lagtime. We followed the same color code as in the main text, and, within each plot, we show a close-up, given the slow transitions. We see that there is agreement between predicted and estimated probabilities. Again, this plot supports the idea that the hydrophobicity of progesterone is key in the binding. For progesterone, the orange (B), green (D), and pink (E) clusters are decaying faster in favor of the binding site cluster (blue, C), compared to both cortisol and aldosterone.

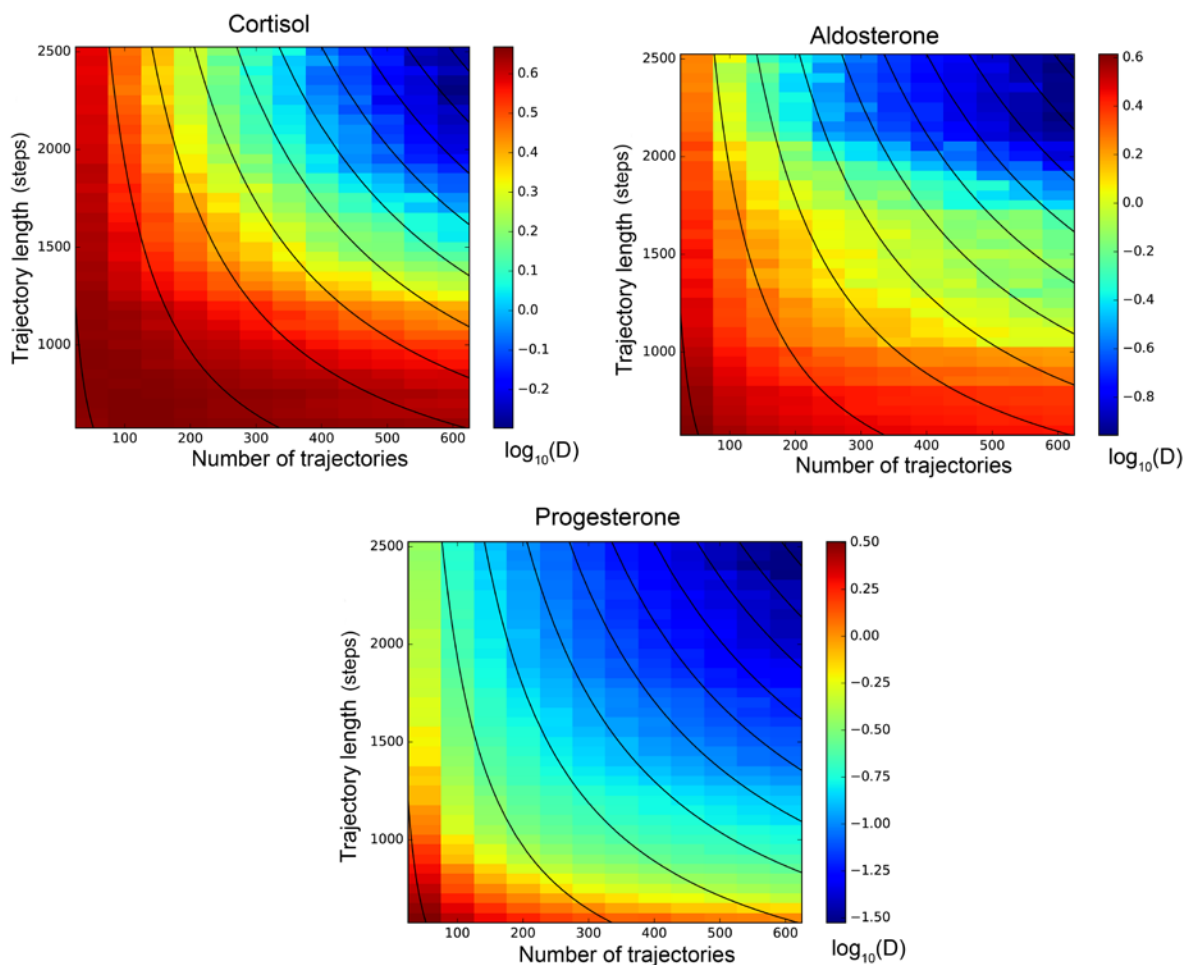


Figure S5. We test the convergence of the MSM using a metric,  $D$ , based on the relative entropy between normalized probability distributions. We compare the distribution of the transition probabilities for all different states, and weight the measures with the stationary distribution,  $\pi$ . Then, the metric  $D$  between the transition matrices  $P$  and  $Q$  is defined as in Ref. 1:  $D(P||Q) = \sum_{ij} \pi_i P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right)$ , where  $P$ , and  $Q$  correspond to the gold model and a trial one, respectively. We use the whole data set as the gold model, and analyze how the relative entropy varies for different numbers of number of trajectories and trajectory length. When the relative entropy does not change, we assume that we have enough sampling and, therefore, a converged MSM. Results show that convergence is first achieved for progesterone, then aldosterone, and finally cortisol. This is probably explained with the number of binding events: progesterone binds more easily than cortisol, and is able to start before collecting data and converging the transition matrix.

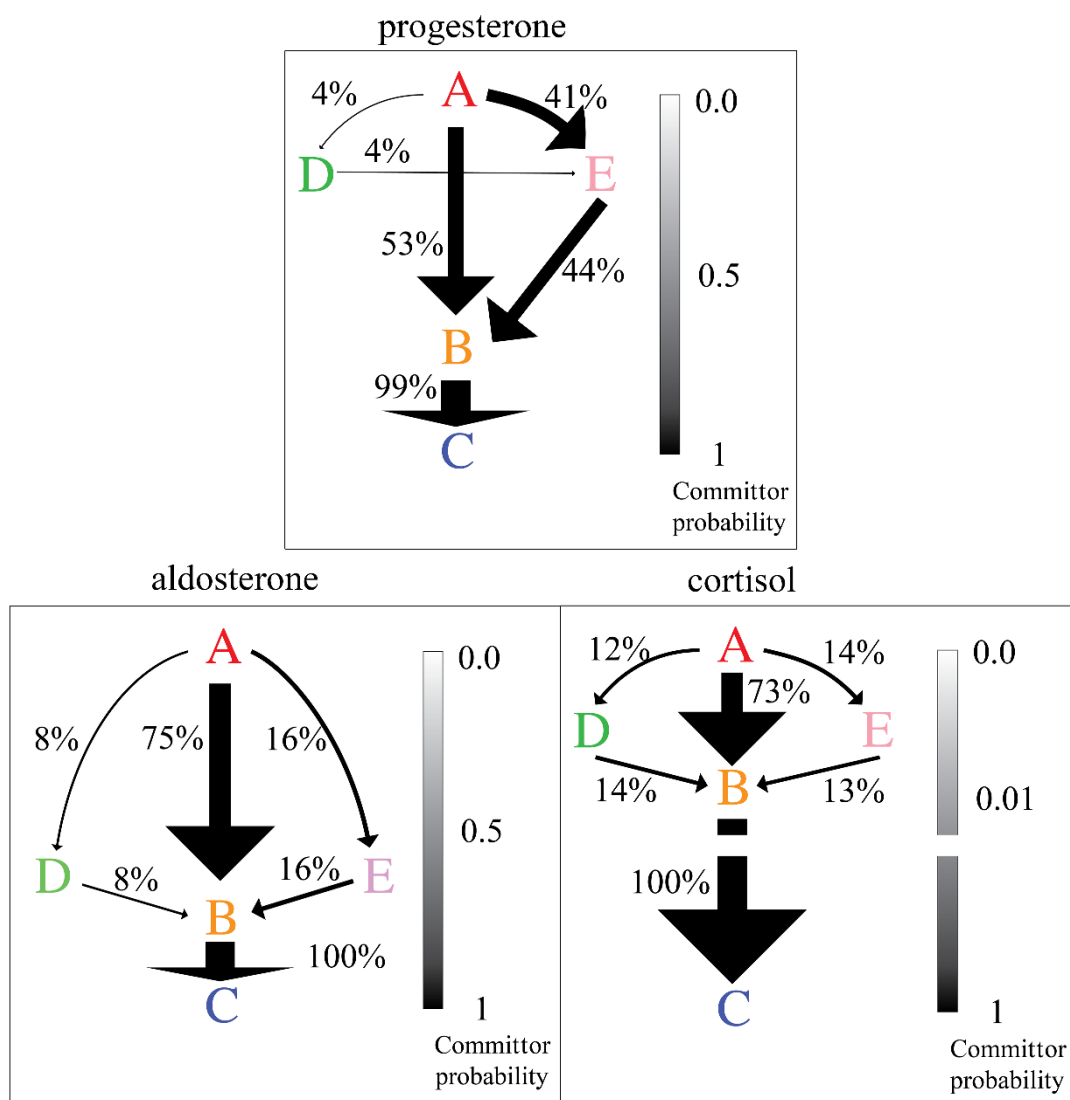


Figure S6. Main binding pathways (i.e. those with fluxes larger than 4%). The thickness of the arrow is proportional to the folding flux, shown in percentage. The committor probability, which is the probability of reaching the binding site from the different metastable states, before going back to the bulk is shown in the y axis.

## Bibliography

1. Bowman, G. R., Ensign, D. L. & Pande, V. S. Enhanced modeling via network theory: Adaptive sampling of markov state models. *J. Chem. Theory Comput.* **6**, 787–794 (2010).

## Supporting information paper 5

## Adaptive simulations, towards interactive protein-ligand modeling

Daniel Lecina<sup>1</sup>, Joan Francesc Gilabert<sup>1</sup> and Victor Guallar<sup>1,2</sup>

<sup>1</sup>Barcelona Supercomputing Center, Joint BSC-CRG-IRB Research Program in Computational Biology, Jordi Girona 29, E-08034 Barcelona, Spain

<sup>2</sup>ICREA, Passeig Lluís Companys 23, E-08010 Barcelona, Spain

### Supplementary Information

Supplementary information includes: i) methodological clustering details; ii) Clustering exploration results (Fig. 1); iii) Energy landscape exploration of TRP, A-GPCR and PR (Figs 2-4); iv) Complete table with all binding time statistics; v) Standard PELE induced fit simulations on sEH; vi) Clustering parameters configurations (Fig. 5)

### Centroid distance as a lower bound for the RMSD.

In the clustering, we have a ligand structure that we want to cluster, and a cluster center (the reference structure) with coordinates  $\mathbf{r}$  and  $\mathbf{r}_{\text{REF}}$ , respectively.

The distance vector between the  $i$ -th atom in both structures is defined as:

$$\mathbf{d}_i(\mathbf{r}_i, \mathbf{r}_{\text{REF},i}) = \mathbf{r}_i - \mathbf{r}_{\text{REF},i} \quad (1)$$

For the sake of brevity, we will avoid recalling explicitly the dependence on the two sets of coordinates:  $\mathbf{d}_i \equiv \mathbf{d}_i(\mathbf{r}_i, \mathbf{r}_{\text{REF},i})$ . The distance between a pair of atoms corresponds to its modulus:  $d_i = \|\mathbf{d}_i\|$ .

The centroid distance,  $c_d$ , between both structures is:

$$c_d(\mathbf{d}_i) = \left\| \sum_i \frac{\mathbf{d}_i}{N} \right\| \quad (2)$$

where the summation extends over all  $N$  ligand atoms.

After superposing the protein alpha carbons, the ligand RMSD is calculated with:

$$\text{RMSD}(d_i) = \sqrt{\sum_i \frac{d_i^2}{N}} \quad (3)$$

where the summation again extends over all ligand atoms.

The Cauchy-Schwarz inequality states that in an  $n$ -dimensional Euclidean space:  $|\langle \mathbf{u} \cdot \mathbf{v} \rangle| \leq \|\mathbf{u}\| \cdot \|\mathbf{v}\|$ , where  $\langle \cdot, \cdot \rangle$  is the inner product. Applying it to  $\mathbf{u}$  and  $\mathbf{v}$  such that  $u_i = \frac{d_i}{N}$  and  $v_i = 1, \forall i \mid i \in [1, N]$ :

$$\sum_i \frac{d_i}{N} \leq \sqrt{\sum_i \left(\frac{d_i}{N}\right)^2 \cdot N} = \sqrt{\sum_i \frac{d_i^2}{N}} = \text{RMSD}(d_i) \quad (4)$$

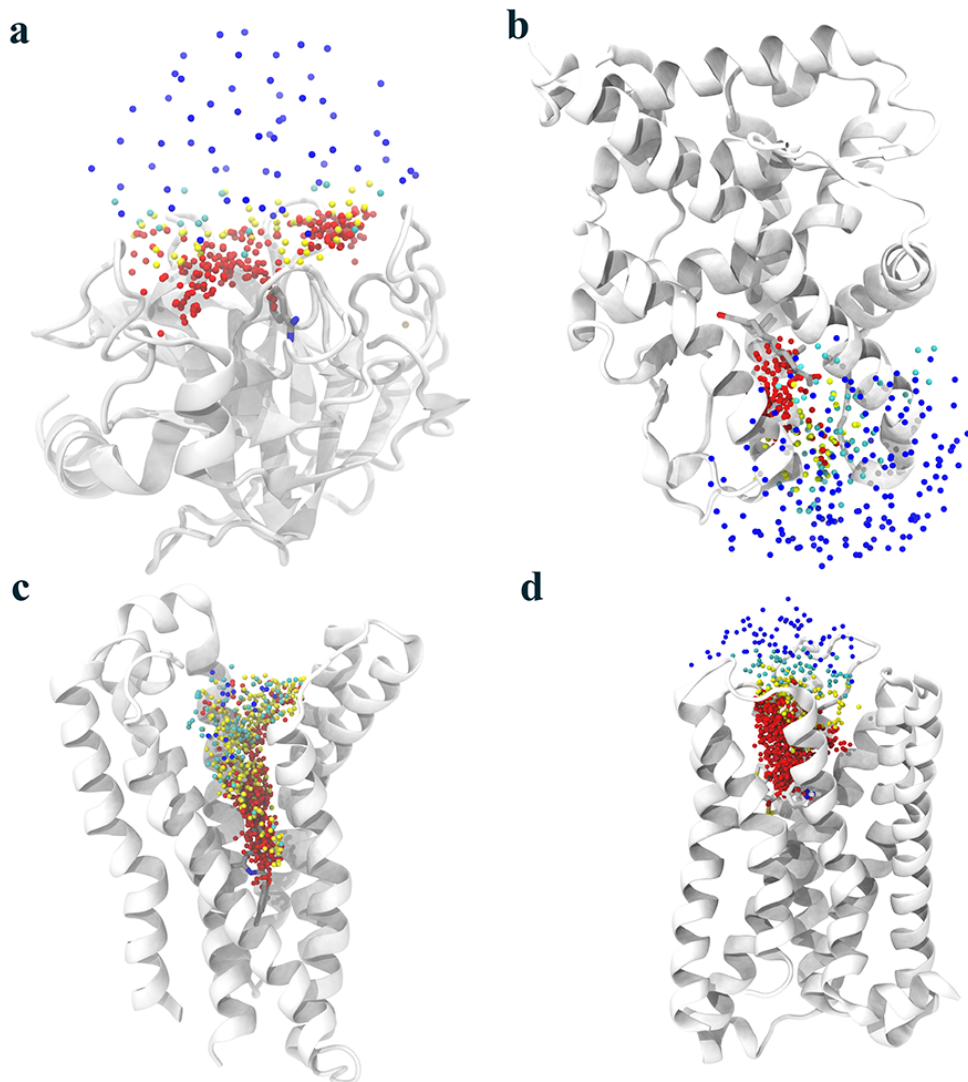
Using the triangle inequality for  $\mathbf{d}_i$ :  $\|\sum_i \mathbf{d}_i\| \leq \sum_i \|\mathbf{d}_i\| = \sum_i d_i$ , and dividing it by  $N$ , we obtain:

$$c_d(\mathbf{d}_i) = \left\| \sum_i \frac{\mathbf{d}_i}{N} \right\| \leq \sum_i \frac{d_i}{N} \quad (5)$$

Combining Eq. (4) and Eq. (5) we obtain:

$$c_d(\mathbf{d}_i) \leq \text{RMSD}(d_i), \quad (6)$$

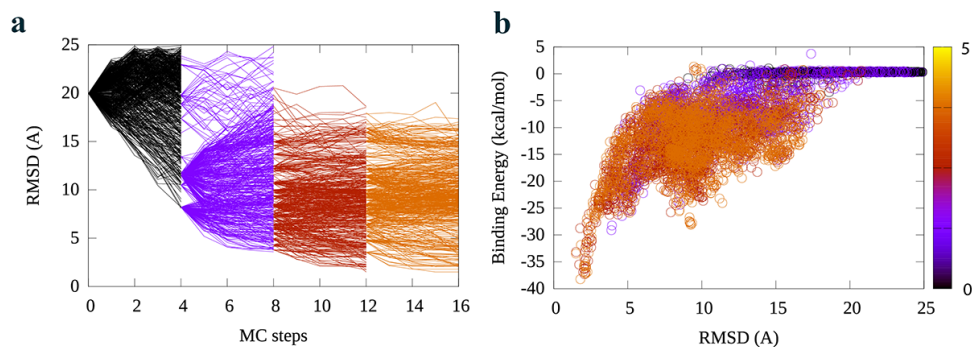
as we wanted to prove. Intuitively, the equality applies when the structure is a translation of the reference structure, and the inequality applies when there is a translation and rotation.



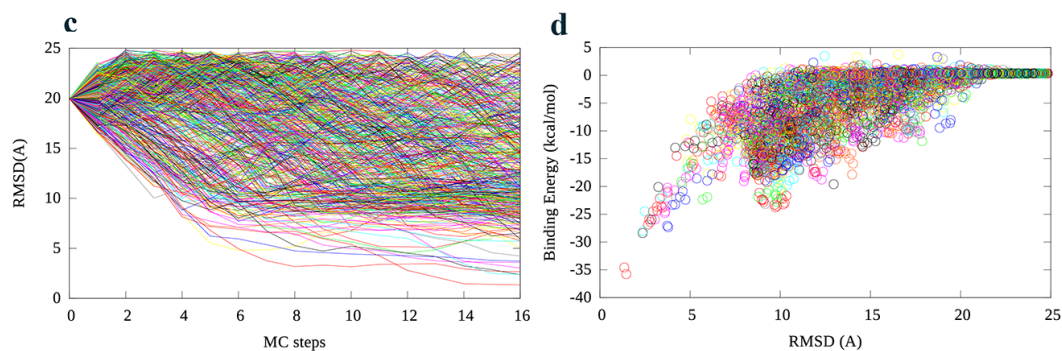
**Supplementary Figure 1 | Cluster exploration in a typical binding simulation with 1024 trajectories.** In white, we show the native structure, and each cluster is marked with the ligand's center of mass with a color that represents its number of contacts and RMSD threshold: in blue those with  $c \leq 0.5$  and a threshold of  $5\text{\AA}$ , in cyan those with  $0.5 < c \leq 0.75$  and a threshold of  $4\text{\AA}$ , in yellow those with  $0.75 < c \leq 1$  and a threshold of  $3\text{\AA}$ , and in red, those with  $c > 1$  and a threshold of  $2\text{\AA}$ . Panel **(a)** corresponds to TRP, panel **(b)** to PR, panel **(c)** to B-GPCR and panel **(d)** to A-GPCR.



## Adaptive PELE

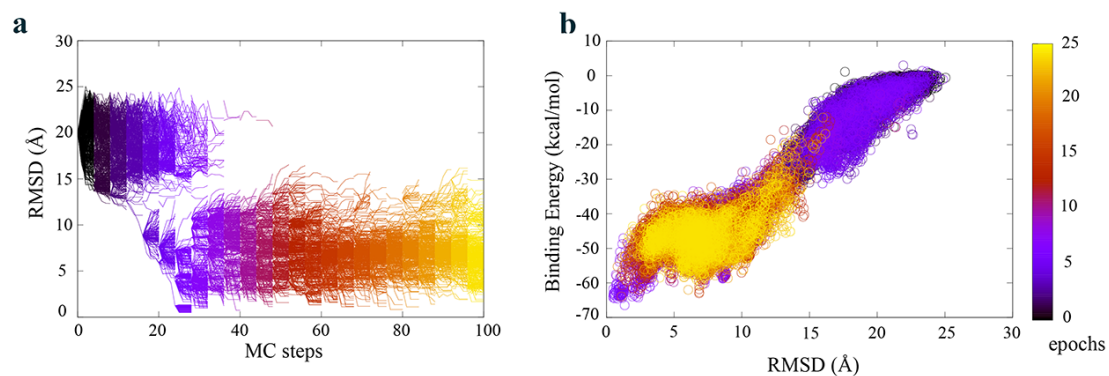


## Standard PELE

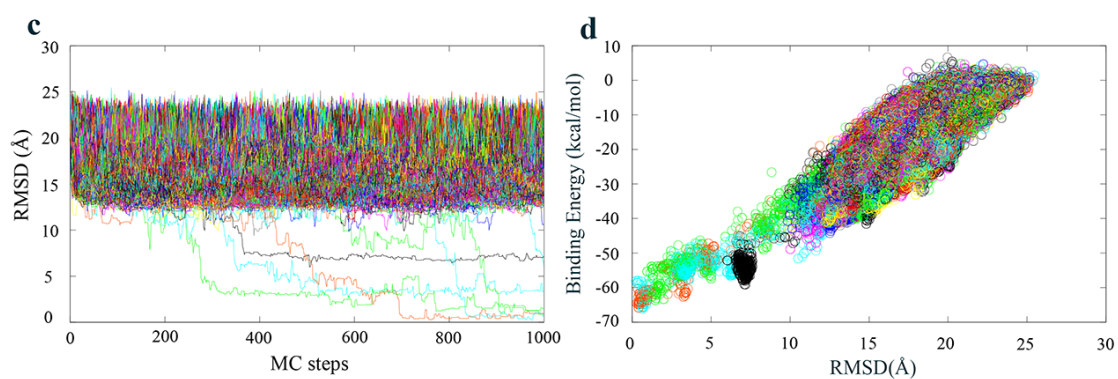


**Supplementary Figure 2 | Energy landscape exploration of TRP with 512 different explorers. (a,b)** The RMSD variation along MC steps and the binding energy against the RMSD for the adaptive results. Each color code corresponds to a different epoch number, for a total of 12 adaptive iterations. **(c,d)** Analogous plots for the standard executions. Each color corresponds to a different trajectory (performed in a different computing core).

## Adaptive PELE

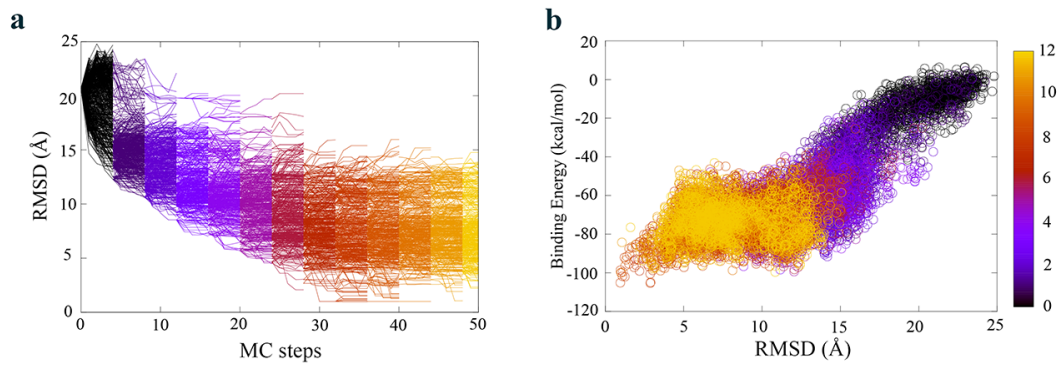


## Standard PELE

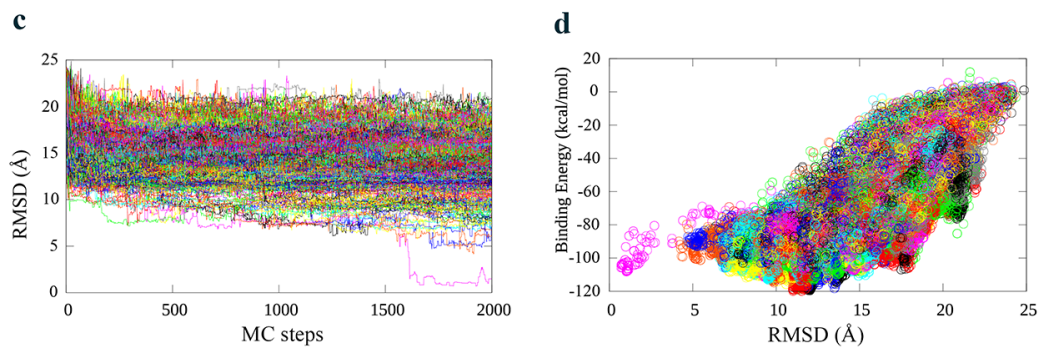


**Supplementary Figure 3 | Energy landscape exploration of PR with 512 different explorers. (a,b)** The RMSD variation along MC steps and the binding energy against the RMSD for the adaptive results. Each color code corresponds to a different epoch number, for a total of 12 adaptive iterations. **(c,d)** Analogous plots for the standard executions. Each color corresponds to a different trajectory (performed in a different computing core).

## Adaptive PELE



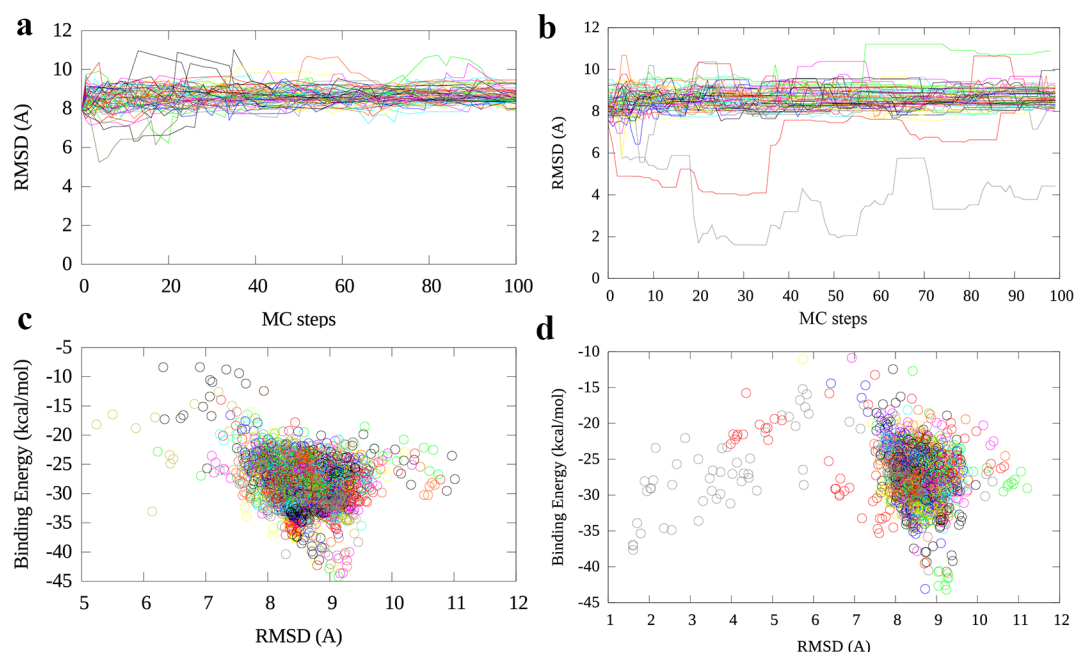
## Standard PELE



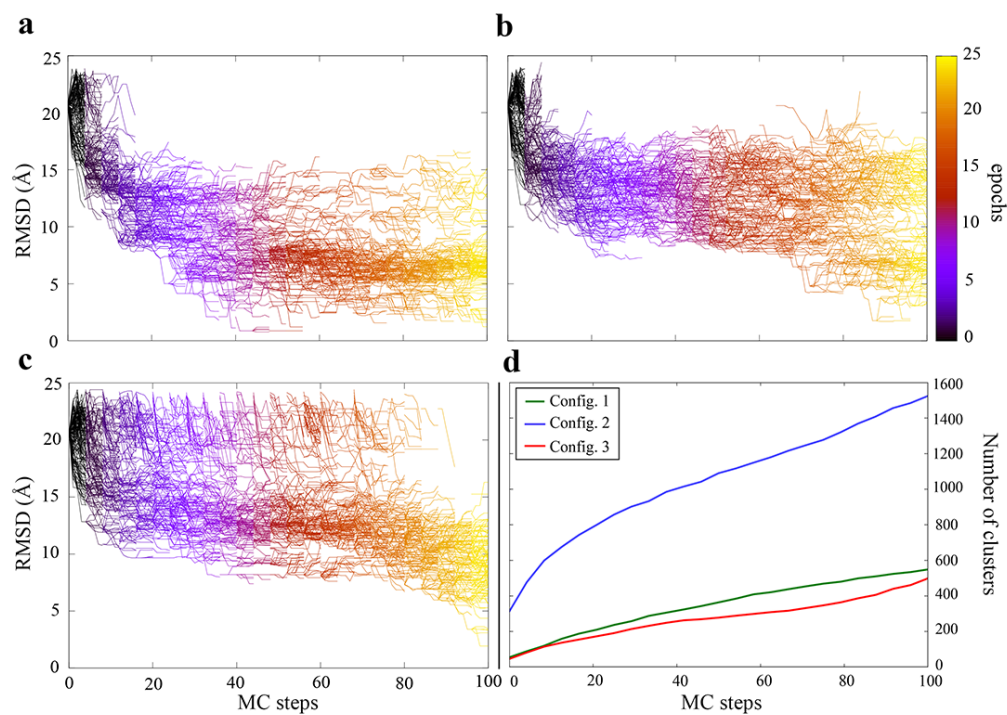
**Supplementary Figure 4 | Energy landscape exploration of A-GPCR with 512 different explorers. (a,b)** The RMSD variation along MC steps and the binding energy against the RMSD for the adaptive results. Each color code corresponds to a different epoch number, for a total of 12 adaptive iterations. **(c,d)** Analogous plots for the standard executions. Each color corresponds to a different trajectory (performed in a different computing core).

|   | 32  | 64            | 128           | 256      | 512      | 1024    |
|---|---|---------------|---------------|----------|----------|---------|
|   | TRP   |               |               |          |          |         |
| ● | 39±30   | 26±10         | 18±7          | 14±4     | 11±3     | 11±2    |
| ● | 26±20   | 21±9          | 12±3          | 9±2      | 10±2     | 9±1     |
| ● | 19±7  | 13±3          | 10±2          | 8±2      | 9±1      | 8±1     |
| ● | 19±9  | 13±4          | 10±2          | 9±3      | 9±2      | 8±2     |
|   | PR  |               |               |          |          |         |
| ● | -   | 1830±125<br>0 | 1590±115<br>0 | 1510±930 | 610±300  | 500±310 |
| ● | 460±270   | 160±90        | 160±130       | 110±60   | 42±20    | 30±9    |
| ● | 270±160   | 160±50        | 120±90        | 73±40    | 56±40    | 32±6    |
| ● | 230±110   | 200±90        | 97±60         | 65±30    | 43±10    | 33±10   |
|   | B-GPCR  |               |               |          |          |         |
| ● | 2200±100<br>0   | 1100±800      | 740±300       | 490±200  | 360±130  | 350±100 |
| ● | 420±500   | 140±50        | 87±20         | 59±20    | 39±10    | 36±10   |
| ● | 200±90  | 130±60        | 82±40         | 55±10    | 48±10    | 35±10   |
| ● | 110±40  | 98±30         | 67±20         | 53±7     | 40±6     | 33±8    |
|   | A-GPCR  |               |               |          |          |         |
| ● | -   | -             | 2440±700      | 1260±830 | 1230±640 | 910±460 |
| ● | 200±100   | 115±40        | 56±10         | 45±10    | 30±4     | 25±4    |
| ● | 230±100   | 76±22         | 74±30         | 45±10    | 30±5     | 23±4    |
| ● | 110±50  | 85±30         | 53±20         | 42±10    | 32±3     | 25±3    |
| ● | Std. PELE, ● : Inversely Proportional, ● B.E. $\epsilon$ -greedy, ● RMSD $\epsilon$ -greedy |               |               |          |          |         |

**Supplementary Table 1 | Binding times for all studied systems and strategies.** Results show the MC steps averaged over ten independent runs. For PR with 32 processors and A-GPCR for 32 and 64, we did not observe any binding event in more than half of the runs. The color code corresponds to the strategy; red for non-adaptive PELE, blue for the inversely proportional strategy, green for the binding energy  $\epsilon$ -greedy and orange for the RMSD  $\epsilon$ -greedy.



**Supplementary Figure 5 | Standard PELE induced-fit docking studies.** Two different cross docking simulations with 64 trajectories are shown for the sHE system: protein structure from PDB ID:5ALX and ligand structure from PDB ID:5AI5 **(a,b)** Evolution of the ligand RMSD to the bound crystal along the simulation. **(c,d)** Evolution of the binding energy for the different RMSD values.



|                 | <b>Density</b> | <b>RMSD thresholds</b> |
|-----------------|----------------|------------------------|
| Configuration 1 | $\propto 1/V$  | Linearly decreasing    |
| Configuration 2 | $\propto 1/V$  | Constant (2Å)          |
| Configuration 3 | Constant       | Linearly decreasing    |

**Supplementary Figure 6 | Clustering parameters configurations. (a,b,c)** Evolution of the ligand RMSD to the bound crystal along the simulation, corresponding to configuration 1, *i.e.* the one that is used throughout the paper, configuration 2 and 3, respectively. **(d)** Evolution of the number of clusters. Those results correspond to A-GPCR using 128 processors and 100 MC steps for three different parameter configurations.



## Bibliography

1. Top 100 Drugs for 2013 by Sales. Available at: <https://www.drugs.com/stats/top100/2013/sales>.
2. Schork, N. J. Personalized medicine: Time for one-person trials. *Nature* **520**, 609–611 (2015).
3. Schrödinger Announces Collaboration with Sanofi to Provide Computational Resources in Support of Its Drug Discovery Programs. Available at: <https://www.schrodinger.com/news/schrodinger-announces-collaboration-sanofi-provide-computational-resources-support-its-drug>.
4. Gilead Sciences Announces Acquisition of Nimbus Therapeutics' Acetyl-CoA Carboxylase (ACC) Program for NASH and Other Liver Diseases. Available at: <http://www.businesswire.com/news/home/20160404005324/en/>.
5. Scarpazza, D. P. *et al.* Extending the generality of molecular dynamics simulations on a special-purpose machine. *Proc. - IEEE 27th Int. Parallel Distrib. Process. Symp. IPDPS 2013* 933–945 (2013).
6. Susukita, R. *et al.* Hardware accelerator for molecular dynamics: MDGRAPE-2. *Comput. Phys. Commun.* **155**, 115–131 (2003).
7. Blue Waters User Portal. (2017).
8. Shirts, M. & Pande, V. S. COMPUTING: Screen Savers of the World Unite! *Science* **290**, 1903–4 (2000).
9. Bowman, G. R., Pande, V. S. & Noé, F. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*. Springer **797**, (2014).
10. Voelz, V. A., Bowman, G. R., Beauchamp, K. A. & Pande, V. S. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9 (1–39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
11. Buch, I., Giorgino, T. & De Fabritiis, G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 10184–9 (2011).
12. Buch, I., Harvey, M. J., Giorgino, T., Anderson, D. P. & De Fabritiis, G. High-throughput all-atom molecular dynamics simulations using distributed computing. *J. Chem. Inf. Model.* **50**, 397–403 (2010).
13. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–96 (2009).
14. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* **27**, 2985–2993 (1894).
15. Stein, A., Rueda, M., Panjkovich, A., Orozco, M. & Aloy, P. A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. *Structure* **19**, 881–889 (2011).
16. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci.* **44**, 98–104 (1958).
17. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their



- functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
18. Orellana, L. Protein dynamics studied by coarse-grained and atomistic theoretical approaches. (Universitat de Barcelona, 2014).
  19. Edman, K. *et al.* Ligand Binding Mechanism in Steroid Receptors: From Conserved Plasticity to Differential Evolutionary Constraints. *Structure* **23**, 2280–2291 (2015).
  20. Grebner, C. *et al.* Exploring Binding Mechanisms in Nuclear Hormone Receptors by Monte Carlo and X-ray-derived Motions. *Biophys. J.* **112**, 1147–1156 (2017).
  21. Williamson, M. P., Havel, T. F. & Wüthrich, K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by <sup>1</sup>H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **182**, 295–315 (1985).
  22. Wüthrich, K. *NMR of proteins and nucleic acids.* (Wiley, 1986).
  23. Fiaux, J., Bertelsen, E. B., Horwich, A. L. & Wüthrich, K. NMR analysis of a 900K GroEL GroES complex. *Nature* **418**, 207–211 (2002).
  24. Callaway, E. The Revolution Will Not Be Crystallized. *Nature* **525**, 172–174 (2015).
  25. Zhao, J., Benlekbir, S. & Rubinstein, J. Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. *Nature* **521**, 241–245 (2015).
  26. Glaeser, R. M. How good can cryo-EM become? *Nat Methods* **13**, 28–32 (2016).
  27. Bartesaghi, A. *et al.* 2.2 Å resolution cryo-EM structure of  $\beta$ -galactosidase in complex with a cell-permeant inhibitor. *Science (80-. )*. **348**, 1147–1151 (2015).
  28. Cozzini, P. *et al.* Target flexibility: An emerging consideration in drug discovery and design. *J. Med. Chem.* **51**, 6237–6255 (2008).
  29. Ryckbosch, S. M., Wender, P. A. & Pande, V. S. Molecular dynamics simulations reveal ligand-controlled positioning of a peripheral protein complex in membranes. *Nat. Commun.* **8**, 6 (2017).
  30. Tokuriki, N. & Tawfik, D. S. Protein dynamism and evolvability. *Science* **324**, 203–207 (2009).
  31. Orozco, M. & Eisenmesser. A theoretical view of protein dynamics. *Chem. Soc. Rev.* **43**, 5051 (2014).
  32. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
  33. Ryde, U. & Soderhjelm, P. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chem. Rev.* **116**, 5520–5566 (2016).
  34. Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. I. *J. Chem. Phys.* **24**, 966–978 (1956).
  35. Acebes, S. *et al.* Rational Enzyme Engineering Through Biophysical and Biochemical Modeling. *ACS Catal.* **6**, 1624–1629 (2016).
  36. Santiago, G. *et al.* Computer-Aided Laccase Engineering: Toward Biological Oxidation of Arylamines. *ACS Catal.* **6**, 5415–5423 (2016).
  37. Monza, E. *et al.* Insights into laccase engineering from molecular simulations: Toward a binding-focused strategy. *J. Phys. Chem. Lett.* **6**, 1447–1453 (2015).
  38. Szabo, A. & Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory.* (Courier Corporation, 1996). doi:10.1119/1.1973756
  39. Yilmazer, N. D. & Korth, M. Enhanced semiempirical QM methods for biomolecular interactions. *Comput. Struct. Biotechnol. J.* **13**, 169–175 (2015).
  40. Hohenberg, P. & Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, 864–871 (1964).
  41. Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249 (1976).

42. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648 (1993).
43. Bochevarov, A. D. *et al.* Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *Int. J. Quantum Chem.* **113**, 2110–2142 (2013).
44. Levitt, M. The birth of computational structural biology. *Nat. Struct. Biol.* **8**, 392–393 (2001).
45. Leach, A. *Molecular Modelling. Principles and Applications.* (Pearson, 2001).
46. Wang, J. M., Wolf, R. M., Caldwell, J. W., Kollman, P. a & Case, D. a. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
47. Yu, W., He, X., Vanommeslaeghe, K. & MacKerell, A. D. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* **33**, 2451–2468 (2012).
48. Huang, L. & Roux, B. Automated force field parameterization for nonpolarizable and polarizable atomic models based on ab initio target data. *J. Chem. Theory Comput.* **9**, 3543–3556 (2013).
49. Harder, E. *et al.* OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *J. Chem. Theory Comput.* **12**, 281–296 (2016).
50. Banks, J. A. Y. L. *et al.* Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **26**, 1752–1780 (2005).
51. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
52. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).
53. Reif, M. M., Hünenberger, P. H. & Oostenbrink, C. New interaction parameters for charged amino acid side chains in the GROMOS force field. *J. Chem. Theory Comput.* **8**, 3705–3723 (2012).
54. Frenkel, D. & Smit, B. Understanding molecular simulation: from algorithms to applications. (2001).
55. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N\*log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
56. Zhou, R. & Berne, B. J. A new molecular dynamics method combining the reference system propagator algorithm with a fast multipole method for simulating proteins and other complex systems. *J. Chem Phys.* **103**, 9444–9459 (1995).
57. Jiao, D., Golubkov, P. A., Darden, T. A. & Ren, P. Calculation of protein-ligand binding free energy by using a polarizable potential. *Proc. Natl. Acad. Sci.* **105**, 6290–6295 (2008).
58. Duan, L. L. *et al.* Large-scale molecular dynamics simulation: Effect of polarization on thrombin-ligand binding energy. *Sci. Rep.* **6**, 31488 (2016).
59. Luque, F. J., Dehez, F., Chipot, C. & Orozco, M. Polarization effects in molecular interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 844–854 (2011).
60. Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. Van & Hermans, J. Interaction models for water in relation to protein hydration. *Intermol. Forces* 331–342 (1981).
61. Stillinger, F. H. & Rahman, A. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.* **60**, 1545–1557 (1974).
62. Guillot, B. A reappraisal of what we have learnt during three decades of computer simulations on water. *J. Mol. Liq.* **101**, 219–260 (2002).

63. Liu, K. *et al.* Characterization of a cage form of the water hexamer. *Nature* **381**, 501–503 (1996).
64. Jorgensen, W. L. & Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6665–6670 (2005).
65. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926 (1983).
66. Fanourgakis, G. S. & Xantheas, S. S. Development of transferable interaction potentials for water. V. Extension of the flexible, polarizable, Thole-type model potential (TMM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water. *J. Chem. Phys.* **128**, 0–11 (2008).
67. Kumar, R., Wang, F. F., Jenness, G. R. & Jordan, K. D. A second generation distributed point polarizable water model. *J. Chem. Phys.* **132**, (2010).
68. Laury, M. L., Wang, L. P., Pande, V. S., Head-Gordon, T. & Ponder, J. W. Revised Parameters for the AMOEBA Polarizable Atomic Multipole Water Model. *J. Phys. Chem. B* **119**, 9423–9437 (2015).
69. Orozco, M. & Luque, F. J. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.* **100**, 4187–4225 (2000).
70. Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129 (1990).
71. Zhu, K., Shirts, M. R. & Friesner, R. A. Improved Methods for Side Chain and Loop Pred via the Prot Local Opt Prog Variable Diel Model for Implicitly Improving the Treatment of Polarization Effects. *J. Chem. Theory Comput.* **3**, 2108–2119 (2007).
72. Srinivasan, J., Trevathan, M. W., Beroza, P. & Case, D. A. Application of a pairwise generalized Born model to proteins and nucleic acids: inclusion of salt effects. *Theor. Chem. Acc.* **101**, 426–434 (1999).
73. Gallicchio, E., Zhang, L. Y. & Levy, R. M. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J. Comput. Chem.* **23**, 517–529 (2002).
74. Li, J. *et al.* The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins Struct. Funct. Bioinforma.* **79**, 2794–2812 (2011).
75. Schrödinger. Schrödinger Release 2015-2: Maestro. (2015).
76. Ghosh, A., Rapp, C. & Friesner, R. Generalized Born Model Based on a Surface Integral Formulation. *J. Phys. Chem. B* **102**, 10983–10990 (1998).
77. Hawkins, G., Cramer, C. & Truhlar, D. Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium. *J. Phys. Chem.* **100**, 19824–19839 (1996).
78. Onufriev, A., Bashford, D. & Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* **55**, 383–394 (2004).
79. Ponder, J. W. TINKER: Software tools for molecular design. (2004).
80. Schaefer, M., Bartels, C. & Karplus, M. Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model. *J. Mol. Biol.* **284**, 835–848 (1998).
81. Pan, A. C., Weinreich, T. M., Piana, S. & Shaw, D. E. Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems. *J. Chem. Theory Comput.* **12**, 1360–1367 (2016).

82. Minh, D. D. L. Protein-Ligand Binding Potential of Mean Force Calculations with Hamiltonian Replica Exchange on Alchemical Interaction Grids. *arXiv* (2015).
83. Bordoli, L. *et al.* Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Protoc.* **4**, 1–13 (2009).
84. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
85. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20**, 786–798 (1999).
86. Patrascioiu, A. The Ergodic Hypothesis. *Los Alamos Sci. Spec. Issue* (1987).
87. Shirts, M. R. Simple Quantitative Tests to Validate Sampling from Thermodynamic Ensembles. (2012).
88. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, a & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
89. Stanislaw Ulam 1909-1984. *Los Alamos Sci. Spec. Issue No. 15* (1987).
90. Abagyan, R. & Totrov, M. Biased Probability Monte Carlo Conformational Searches and Electrostatic Calculations for Peptides and Proteins Optimal Probability Distribution for a Random Step in the Monte Carlo Procedure Formula for the Modified Image Approximation of the Electrostat. *J. Mol. Biol.* **235**, 983–1002 (1994).
91. Zhang, Y. & Skolnick, J. Parallel-hat tempering: A Monte Carlo search scheme for the identification of low-energy structures. *J. Chem. Phys.* **115**, 5027–5032 (2001).
92. Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv* (2017).
93. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
94. Friesner, R. A. *et al.* Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem.* **49**, 6177–6196 (2006).
95. Rey, A. & Skolnick, J. Comparison of Lattice Monte-Carlo Dynamics and Brownian Dynamics Folding Pathways of Alpha-Helical Hairpins. *Chem. Phys.* **158**, 199–219 (1991).
96. Shimada, J., Kussell, E. L. & Shakhnovich, E. I. The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. *J. Mol. Biol.* **308**, 79–95 (2001).
97. Shimada, J. & Shakhnovich, E. I. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc. Natl. Acad. Sci.* **99**, 11175–11180 (2002).
98. Voter, A. in *Radiation Effects in Solids* (eds. Sickafus, K. E. & Kotomin, E. A.) (Springer, 2005).
99. Jorgensen, W. L. & TiradoRives, J. Monte Carlo vs molecular dynamics for conformational sampling. *J. Phys. Chem.* **100**, 14508–14513 (1996).
100. Favrin, G., Irbäck, A. & Sjunnesson, F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* **114**, 8154–8158 (2001).
101. Yun, M.-R., Lavery, R., Mousseau, N., Zakrzewska, K. & Derreumaux, P. ARTIST: An Activated Method in Internal Coordinate Space for Sampling Protein Energy Landscapes. *Proteins* **63**, 967–975 (2006).
102. Jorgensen, W. L. & Tirado-Rives, J. Molecular modeling of organic and

- biomolecular systems using BOSS and MCPRO. *J. Comput. Chem.* **26**, 1689–1700 (2005).
103. ProtoMS Documentation | Perturbation.
  104. Woods, C. J. Sire.
  105. Lee, W. G. *et al.* Picomolar inhibitors of HIV reverse transcriptase featuring bicyclic replacement of a cyanovinylphenyl group. *J. Am. Chem. Soc.* **135**, 16705–16713 (2013).
  106. Michel, J. & Essex, J. W. Hit Identification and Binding Mode Predictions by Rigorous Free Energy Simulations. *J. Med. Chem.* 6654–6664 (2008).
  107. Bahar, I., Lezon, T. R., Bakan, A. & Shrivastava, I. H. Normal mode analysis of biomolecular structures: Functional mechanisms of membrane proteins. *Chem. Rev.* **110**, 1463–1497 (2010).
  108. Madadkar-Sobhani, A. & Guallar, V. PELE web server: atomistic study of biomolecular systems at your fingertips. *Nucleic Acids Res.* **41**, 322–328 (2013).
  109. Borrelli, K. W., Vitalis, A., Alcantara, R. & Guallar, V. PELE: Protein energy landscape exploration. A novel Monte Carlo based technique. *J. Chem. Theory Comput.* **1**, 1304–1311 (2005).
  110. Hosseini, A. *et al.* Atomic picture of ligand migration in toluene 4-monooxygenase. *J. Phys. Chem. B* **119**, 671–678 (2015).
  111. Kotev, M., Lecina, D., Tarragó, T., Giralt, E. & Guallar, V. Unveiling prolyl oligopeptidase ligand migration by comprehensive computational techniques. *Biophys. J.* **108**, 116–125 (2015).
  112. Borrelli, K. W., Cossins, Benjamin, Guallar, V. Exploring Hierarchical Refinement Techniques for Induced Fit Docking with Protein and Ligand Flexibility. *J. Comput. Chem.* **31**, 1224–1235 (2010).
  113. Kilian, P., Valdes, J. J., Lecina-Casas, D., Chrudimský, T. & Růžek, D. The variability of the large genomic segment of Ťahyňa orthobunyavirus and an all-atom exploration of its anti-viral drug resistance. *Infect. Genet. Evol.* **20**, 304–311 (2013).
  114. Hosseini, A. *et al.* Computational Prediction of HIV-1 Resistance to Protease Inhibitors. *J. Chem. Inf. Model.* **56**, 915–923 (2016).
  115. Takahashi, R., Gil, V. A. & Guallar, V. Monte carlo free ligand diffusion with markov state model analysis and absolute binding free energy calculations. *J. Chem. Theory Comput.* **10**, 282–288 (2014).
  116. Deri, B. *et al.* The unravelling of the complex pattern of tyrosinase inhibition. *Sci. Rep.* **6**, 34993 (2016).
  117. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and Testing of the OLPS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).
  118. Kaminski, G. a, Friesner, R. a, Tirado-rives, J. & Jorgensen, W. L. Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **105**, 6474–6487 (2001).
  119. Wang, J., Cieplak, P. & Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **21**, 1049–1074 (2000).
  120. Hornak, V. *et al.* Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins* **65**, 712–725 (2006).
  121. Pérez, A. *et al.* Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92**, 3817–29 (2007).
  122. Schrödinger. MacroModel. (2017).
  123. Schrödinger. QSite. (2017).

124. Dunbrack, R. L. Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440 (2002).
125. Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **77**, 1905–1908 (1996).
126. Atilgan, A. R. *et al.* Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **80**, 505–515 (2001).
127. Kitao, A., Hirata, F. & Go, N. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* **158**, 447–472 (1991).
128. Rueda, M., Chacón, P. & Orozco, M. Thorough Validation of Protein Normal Mode Analysis: A Comparative Study with Essential Dynamics. *Structure* **15**, 565–575 (2007).
129. Gil, V. A., Lecina, D., Grebner, C. & Guallar, V. Enhancing backbone sampling in Monte Carlo simulations using internal coordinates normal mode analysis. *Bioorg. Med. Chem.* **24**, 4855–4866 (2016).
130. Gil Sepúlveda, V. A. Algorithmic and Technical Improvements for Next Generation Drug Design Software Tools Algorithmic and Technical Improvements for Next Generation Drug Design Software Tools. (Universitat de Barcelona, 2016).
131. Cossins, B. P., Hosseini, A. & Guallar, V. Exploration of protein conformational change with PELE and meta-dynamics. *J. Chem. Theory Comput.* **8**, 959–965 (2012).
132. Eyal, E., Yang, L.-W. & Bahar, I. Anisotropic network model: systematic evaluation and a new web interface. *Bioinformatics* **22**, 2619 (2006).
133. Jacobson, M. P., Kaminski, G. A., Friesner, R. A. & Rapp, C. S. Force field validation using protein side chain prediction. *J. Phys. Chem. B* **106**, 11673–11680 (2002).
134. Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **320**, 597–608 (2002).
135. Xiang, Z. & Honig, B. Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311**, 421–30 (2001).
136. Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. PROTEIN FOLDING: The Endgame. *Annu. Rev. Biochem.* **66**, 549–79 (1997).
137. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19**, 844–858 (2011).
138. Gainza, P., Roberts, K. E. & Donald, B. R. Protein design using continuous rotamers. *PLoS Comput. Biol.* **8**, (2012).
139. Zhu, K., Shirts, M. R., Friesner, R. A. & Jacobson, M. P. Multiscale optimization of a truncated newton minimization algorithm and application to proteins and protein-ligand complexes. *J. Chem. Theory Comput.* **3**, 640–648 (2007).
140. Xie, D. & Schlick, T. Efficient implementation of the truncated-Newton algorithm for large-scale chemistry applications. *siam j. optim.* **10**, 132–154 (1999).
141. Gilbert, J. & Nocedal, J. Global Convergence Properties of Conjugate Gradient Methods for Optimization. *SIAM J. Optim.* **2**, 21–42 (1992).
142. Zhou, H.-X. & Gilson, M. K. Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **109**, 4092–4107 (2009).
143. Biela, A. *et al.* Dissecting the hydrophobic effect on the molecular level: The role of water, enthalpy, and entropy in ligand binding to thermolysin. *Angew. Chemie - Int. Ed.* **52**, 1822–1828 (2013).
144. Zidek, L., Novotny, M. V & Stone, M. J. Increased protein backbone

- conformational entropy upon hydrophobic ligand binding. *Nat. Struct. Biol.* **6**, 1118–1121 (1999).
145. Chang, C. a, Chen, W. & Gilson, M. K. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1534–1539 (2007).
  146. Velazquez-Campoy, A., Todd, M. J. & Freire, E. HIV-1 protease inhibitors: Enthalpic versus entropic optimization of the binding affinity. *Biochemistry* **39**, 2201–2207 (2000).
  147. General, I. J. A Note on the Standard State 's Binding Free Energy. *Society* 2520–2524 (2010).
  148. Chaires, J. B. Calorimetry and thermodynamics in drug design. *Annu. Rev. Biophys.* **37**, 135–151 (2008).
  149. Ladbury, J. E. & Chowdhry, B. Z. Sensing the heat: the application of isothermal titration calorimetry to thermodynamic studies of biomolecular interactions. *Chem. Biol.* **3**, 791–801 (1996).
  150. Perozzo, R., Folkers, G. & Scapozza, L. Thermodynamics of Protein–Ligand Interactions: History, Presence, and Future Aspects. *J. Recept. Signal Transduct.* **24**, 1–52 (2004).
  151. Myszka, D. G. *et al.* The ABRF-MIRG'02 study: Assembly state, thermodynamic, and kinetic analysis of an enzyme/inhibitor interaction. *J. Biomol. Tech.* **14**, 247–269 (2003).
  152. Du, X. *et al.* Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. *Int. J. Mol. Sci.* **17**, 144 (2016).
  153. Olsson, T. S. G., Williams, M. A., Pitt, W. R. & Ladbury, J. E. The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* **384**, 1002–1017 (2008).
  154. Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein–ligand interactions and its potential for drug discovery. *Biochim. Biophys. Acta - Biomembr.* **1838**, 43–55 (2014).
  155. Zeng, S., Baillargeat, D., Ho, H.-P. & Yong, K.-T. Nanomaterials enhanced surface plasmon resonance for biological and chemical sensing applications. *Chem. Soc. Rev.* **43**, 3426–52 (2014).
  156. Rossi, A. M. & Taylor, C. W. Analysis of protein-ligand interactions by fluorescence polarization. *Nat. Protoc.* **6**, 365–387 (2011).
  157. Lea, W. A. & Simeonov, A. Fluorescence Polarization Assays in Small Molecule Screening. *Expert Opin Drug Discov.* **6**, 17–32 (2011).
  158. Mobley, D. L. Let's get honest about sampling. *J. Comput. Aided. Mol. Des.* **26**, 93–95 (2012).
  159. Michel, J. & Essex, J. W. Prediction of protein-ligand binding affinity by free energy simulations: Assumptions, pitfalls and expectations. *J. Comput. Aided. Mol. Des.* **24**, 639–658 (2010).
  160. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3**, 1–5 (2004).
  161. Yin, J. *et al.* Overview of the SAMPL5 host-guest challenge: Are we doing better? *J. Comput. Aided. Mol. Des.* **31**, 1–19 (2016).
  162. Wang, L. *et al.* Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**, 2695–2703 (2015).
  163. Gilson, M. K. & Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**, 21–42 (2007).
  164. Grinter, S. Z. & Zou, X. Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. *Molecules* **19**, 10150–10176 (2014).

165. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).
166. Trott, O. & Olson, A. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **31**, (2010).
167. Lang, P., Brozell, S., Mukherjee, S. & Pettersen, E. DOCK 6: Combining techniques to model RNA–small molecule complexes. *RNA* **15**, 1219–1230 (2009).
168. Ruiz-Carmona, S. *et al.* rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.* **10**, 1–7 (2014).
169. Huang, S.-Y., Grinter, S. Z. & Zou, X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **12**, 12899–908 (2010).
170. Huang, S.-Y. & Zou, X. Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins* **66**, 399–421 (2007).
171. Moal, I. H. & Bates, P. A. SwarmDock and the use of normal modes in protein-protein Docking. *Int. J. Mol. Sci.* **11**, 3623–3648 (2010).
172. Dobbins, S. E., Lesk, V. I. & Sternberg, M. J. E. Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10390–10395 (2008).
173. Kotev, M., Soliva, R. & Orozco, M. Challenges of docking in large, flexible and promiscuous binding sites. *Bioorganic Med. Chem.* **24**, 4961–4969 (2016).
174. Carlson, H. A. *et al.* CSAR 2014: A Benchmark Exercise Using Unpublished Data from Pharma. *J. Chem. Inf. Model.* **56**, 1063–1077 (2016).
175. Wang, W., Wang, J. & Kollman, P. A. What Determines the van der Waals Coefficient B in the LIE (Linear Interaction Energy) Method to Estimate Binding Free Energies Using Molecular Dynamics Simulations? *Proteins Struct. Funct. Genet.* **34**, 395–402 (1999).
176. Swanson, J. M. J., Henchman, R. H. & McCammon, J. A. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **86**, 67–74 (2004).
177. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **441**, 1–13 (2015).
178. Lee, M. S. & Olson, M. a. Calculation of absolute protein-ligand binding affinity using path and endpoint approaches. *Biophys. J.* **90**, 864–877 (2006).
179. Brown, S. P. & Muchmore, S. W. High-Throughput Calculation of Protein-Ligand Binding Affinities: Modification and Adaptation of the MM-PBSA Protocol to Enterprise Grid Computing. 999–1005 (2005). doi:10.1021/ci050488t
180. Genheden, S. & Ryde, U. Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys. Chem. Chem. Phys.* **14**, 8662–77 (2012).
181. Ochterski, J. W. Thermochemistry in Gaussian. *Gaussian Inc Pittsburgh PA* **264**, 1–19 (2000).
182. Nicolini, P., Frezzato, D., Gellini, C., Bizzarri, M. & Chelli, R. Toward quantitative estimates of binding affinities for protein-ligand systems involving large inhibitor compounds: A steered molecular dynamics simulation route. *J. Comput. Chem.* **34**, 1561–1576 (2013).
183. Chodera, J. D. *et al.* Alchemical free energy methods for drug discovery: Progress and challenges. *Curr. Opin. Struct. Biol.* **21**, 150–160 (2011).
184. Plattner, N. & Noé, F. Protein conformational plasticity and complex ligand-



- binding kinetics explored by atomistic simulations and Markov models. *Nat. Commun.* **6**, 7653 (2015).
185. Zuckerman, D. M. Equilibrium Sampling in Biomolecular Simulations. *Annu. Rev. Biophys.* **40**, 41–62 (2011).
  186. Roux, B. The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **91**, 275–282 (1995).
  187. Buch, I., Sadiq, S. K. & De Fabritiis, G. Optimized potential of mean force calculations for standard binding free energies. *J. Chem. Theory Comput.* **7**, 1765–1772 (2011).
  188. Doudou, S., Burton, N. A. & Henchman, R. H. Standard free energy of binding from a one-dimensional potential of mean force. *J. Chem. Theory Comput.* **5**, 909–918 (2009).
  189. Gervasio, F. L., Laio, A. & Parrinello, M. Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.* **127**, 2600–2607 (2005).
  190. Bussi, G., Gervasio, F. L., Laio, A. & Parrinello, M. Free-energy landscape for B hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.* **128**, 13435–13441 (2006).
  191. Laio, A. & Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports Prog. Phys.* **71**, 126601 (2008).
  192. Chen, L. Y. Hybrid steered molecular dynamics approach to computing absolute binding free energy of ligand-protein complexes: A brute force approach that is fast and accurate. *J. Chem. Theory Comput.* **11**, 1928–1938 (2015).
  193. Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **78**, 2690–2693 (1997).
  194. Darve, E. & Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **115**, 9169–9183 (2001).
  195. Liu, S. *et al.* Lead Optimization Mapper: Automating free energy calculations for lead optimization. *J. Comput. Aided. Mol. Des.* **27**, 755–770 (2013).
  196. Mobley, D. L. & Gilson, M. K. Predicting binding free energies: Frontiers and benchmarks. *bioRxiv* 74625 (2016). doi:10.1101/074625
  197. Chen, W., Chang, C.-E. & Gilson, M. K. Calculation of cyclodextrin binding affinities: energy, entropy, and implications for drug design. *Biophys. J.* **87**, 3035–3049 (2004).
  198. Mochizuki, K., Whittleston, C. S., Somani, S., Kusumaatmaja, H. & Wales, D. J. A conformational factorisation approach for estimating the binding free energies of macromolecules. *Phys. Chem. Chem. Phys.* **16**, 2842–53 (2014).
  199. Sutton, R. S. & Barto, A. G. *Reinforcement learning, an introduction*. Cambridge: MIT Press/Bradford Books (1998). doi:10.1109/MED.2013.6608833
  200. Prinz, J. H. *et al.* Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, (2011).
  201. Shukla, D., Hernández, C. X., Weber, J. K. & Pande, V. S. Markov state models provide insights into dynamic modulation of protein function. *Acc. Chem. Res.* **48**, 414–422 (2015).
  202. Lane, T., Bowman, G. R., Beauchamp, K. A., Voelz, V. A. & Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. **133**, 18413–18419 (2011).
  203. Singhal, N., Snow, C. D. & Pande, V. S. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* **121**, 415–425 (2004).
  204. Gu, S., Silva, D. A., Meng, L., Yue, A. & Huang, X. Quantitatively Characterizing

- the Ligand Binding Mechanisms of Choline Binding Protein Using Markov State Model Analysis. *PLoS Comput. Biol.* **10**, (2014).
205. Wu, H., Paul, F., Wehmeyer, C. & Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci.* **113**, E3221–E3230 (2016).
  206. Lawrenz, M., Shukla, D. & Pande, V. S. Cloud computing approaches for prediction of ligand binding poses and pathways. *Sci. Rep.* **5**, 7918 (2015).
  207. Noé, F., Schütte, C., Vanden-Eijnden, E., Reich, L. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19011–6 (2009).
  208. Kohlhoff, K. J. *et al.* Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nat. Chem.* **6**, 15–21 (2014).
  209. Singhal, N. & Pande, V. S. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.* **123**, (2005).
  210. Voelz, V. A., Elman, B., Razavi, A. M. & Zhou, G. Surprisal metrics for quantifying perturbed conformational dynamics in Markov state models. *J. Chem. Theory Comput.* **10**, 5716–5728 (2014).
  211. Doerr, S. & de Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **10**, 2064–2069 (2014).
  212. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, (2013).
  213. Beauchamp, K. A. *et al.* MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *J. Chem. Theory Comput.* **7**, 3412–3419 (2011).
  214. Scherer, M. K. *et al.* PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **11**, 5525–5542 (2015).
  215. Senne, M., Trendelkamp-schroer, B. & Noe, F. EMMA: A Software Package for Markov Model Building and Analysis. *J. Chem. Theory Comput.* **8**, 2223–2238 (2012).
  216. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
  217. Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **18**, 154–162 (2008).
  218. Noé, F., Horenko, I., Schütte, C. & Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.* **126**, (2007).
  219. Wales, D. J. & Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
  220. Wales, D. J. & Scheraga, H. a. Global optimization of clusters, crystals, and biomolecules. *Science* **285**, 1368–1372 (1999).
  221. Li, Z. & Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci.* **84**, 6611–6615 (1987).
  222. Nayeem, A., Vila, J. & Scheraga, H. A. A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin. *J. Comput. Chem.* **12**, 594–605 (1991).
  223. Pillardy, J., Arnautova, Y. A., Czaplewski, C., Gibson, K. D. & Scheraga, H. A. Conformation-family Monte Carlo: a new method for crystal structure prediction. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 12351–6 (2001).

224. Trosset, J. Y. & Scheraga, H. a. Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8011–8015 (1998).
225. Malek, R. & Mousseau, N. Dynamics of lennard-jones clusters: A characterization of the activation-relaxation technique. *Phys. Rev. E* **62**, 7723–8 (2000).
226. Barkema, G. T. & Mousseau, N. Event-Based Relaxation of Continuous Disordered Systems. *Phys. Rev. Lett.* **77**, 4358–4361 (1996).
227. Mousseau, N. & Derreumaux, P. Exploring the early steps of amyloid peptide aggregation by computers. *Acc. Chem. Res.* **38**, 885–891 (2005).
228. Melquiond, A., Mousseau, N. & Derreumaux, P. Structures of soluble amyloid oligomers from computer simulations. *Proteins* **65**, 180–191 (2006).
229. Wei, G., Mousseau, N. & Derreumaux, P. Complex folding pathways in a simple b-hairpin. *Proteins Struct. Funct. Genet.* **56**, 464–474 (2004).
230. Mousseau, N., Derreumaux, P. & Gilbert, G. Navigation and analysis of the energy landscape of small proteins using the activation-relaxation technique. *Phys. Biol.* **2**, S101–S107 (2005).
231. Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses.* (Cambridge University Press, 2003).
232. Doye, J. P. K. & Wales, D. J. Thermodynamics of Global Optimization. *Phys. Rev. Lett.* **80**, 1357–1360 (1998).
233. Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120**, 9911–9917 (2004).
234. Lecina, D., Gilabert, J. F. & Guallar, V. Adaptive simulations, towards interactive protein-ligand modeling. *Revis.*
235. Kruse, A. C. *et al.* Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **482**, 552–556 (2012).
236. Kappel, K., Miao, Y. & McCammon, J. A. Accelerated molecular dynamics simulations of ligand binding to a muscarinic G-protein-coupled receptor. *Q. Rev. Biophys.* **48**, 479–487 (2015).
237. Doye, J. P. K. & Wales, D. J. Calculation of thermodynamic properties of small Lennard-Jones clusters incorporating anharmonicity. *J. Chem. Phys.* **102**, 9659 (1995).
238. Bogdan, T. V., Wales, D. J. & Calvo, F. Equilibrium thermodynamics from basin-sampling. *J. Chem. Phys.* **124**, 1–13 (2006).
239. White, R. P. & Mayne, H. R. An investigation of two approaches to basin hopping minimization for atomic and molecular clusters. *Chem. Phys. Lett.* **289**, 463–468 (1998).
240. Landau, D. P., Tsai, S.-H. & Exler, M. A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling. *Am. J. Phys.* **72**, 1294 (2004).
241. Golub, G. H. & Van Loan, C. F. *Matrix Computations.* (The John Hopkins University Press, 1996). doi:10.1063/1.3060478
242. Andersen, H. S. *et al.* 2-(Oxalylamino)-Benzoic Acid Is a General, Competitive Inhibitor of Protein-Tyrosine Phosphatases. *J. Biol. Chem.* **275**, 7101–7108 (2000).
243. Bowman, G. R., Ensign, D. L. & Pande, V. S. Enhanced modeling via network theory: Adaptive sampling of markov state models. *J. Chem. Theory Comput.* **6**, 787–794 (2010).
244. Hinrichs, N. S. & Pande, V. S. Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J. Chem. Phys.* **126**, (2007).

245. Zimmerman, M. I. & Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).
246. Beck, K. *Implementation Patterns*. (Addison-Wesley Professional, 2006).
247. Wilson, G. *et al.* Best Practices for Scientific Computing. *PLoS Biol.* **12**, (2014).
248. Cabeza De Vaca, I., Lucas, M. F. & Guallar, V. New Monte Carlo Based Technique to Study DNA-Ligand Interactions. *J. Chem. Theory Comput.* **11**, 5598–5605 (2015).
249. Hermosilla, P. *et al.* Physics-based Visual Characterization of Molecular Interaction Forces. *IEEE Trans. Vis. Comput. Graph.* **2626**, 1–1 (2016).
250. Giannotti, M. I. *et al.* Direct Measurement of the Nanomechanical Stability of a Redox Protein Active Site and Its Dependence upon Metal Binding. *J. Phys. Chem. B* **119**, 12050–12058 (2015).
251. Ceperley, D., Chester, G. V. & Kalos, M. H. Monte Carlo simulation of a many-fermion study. *Phys. Rev. B* **16**, 3081–3099 (1977).
252. Pangali, C., Rao, M. & Berne, B. J. On a novel Monte Carlo scheme for simulating water and aqueous solutions. *Chem. Phys. Lett.* **55**, 413–417 (1978).
253. Neyts, E. C., Thijsse, B. J., Mees, M. J., Bal, K. M. & Pourtois, G. Establishing uniform acceptance in force biased Monte Carlo simulations. *J. Chem. Theory Comput.* **8**, 1865–1869 (2012).
254. Nilmeier, J., Crooks, G. E., Minh, D. D. L. & Chodera, J. D. Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation. *Proc. Natl. Acad. Sci.* **109**, 9665–9665 (2012).
255. Wereszczynski, J. & McCammon, J. A. Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition. *Q. Rev. Biophys.* **45**, 1–25 (2012).
256. Pohorille, A. & Chipot, C. *Free Energy Calculations. Theory and Applications in Chemistry and Biology*. (Springer-Verlag, 2007). doi:10.1007/978-3-540-68038-3-1
257. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303**, 1813–8 (2004).
258. Tidor, B. & Karplus, M. The contribution of vibrational entropy to molecular association. The dimerization of insulin. *Journal of molecular biology* **238**, 405–414 (1994).
259. Killian, B. J. *et al.* Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an HIV-Derived PTAP Nonapeptide. *J. Mol. Biol.* **389**, 315–335 (2009).
260. Gallicchio, E., Kubo, M. M. & Levy, R. M. Enthalpy-entropy and cavity decomposition of alkane hydration free energies: numerical results and implications for theories of hydrophobic solvation. *J. Phys. Chem. B* **104**, 6271–6285 (2000).
261. Baron, R. & McCammon, J. A. (Thermo)dynamic role of receptor flexibility, entropy, and motional correlation in protein-ligand binding. *ChemPhysChem* **9**, 983–988 (2008).
262. Chodera, J. D. & Mobley, D. L. Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* **42**, 121–42 (2013).
263. Ross, G. A., Bodnarchuk, M. S. & Essex, J. W. Water Sites, Networks, and Free Energies with Grand Canonical Monte Carlo. *J. Am. Chem. Soc.* **137**, 14930–14943 (2015).
264. Grebner, C. *et al.* Binding Mode and Induced Fit Predictions for Prospective Computational Drug Design. *J. Chem. Inf. Model.* **56**, 774–787 (2016).

265. Eastman, P., Friedrichs, M. & Chodera, J. OpenMM 4: a reusable, extensible, hardware independent library for high performance molecular simulation. *J. Chem.* (2012).
266. Doerr, S., Harvey, M. J., Noé, F. & De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **12**, 1845–1852 (2016).
267. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
268. Huang, X., Bowman, G. R., Bacallado, S. & Pande, V. S. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19765–19769 (2009).