Applications of molecular dynamics in drug discovery and technology transfer via a web-based platform

Gerard Martínez

TESI DOCTORAL UPF / ANY 2017

DIRECTOR DE LA TESI
Prof. Gianni de Fabritiis
Departament de Ciències Experimentals i de la Salut

Era un padre que valoraba excepcionalmente la riqueza interior que puede hallar el ser humano. Por eso envió a sus hijos a recibir instrucción y ejercitamiento espirituales de un gran maestro. Los muchachos estuvieron un año recibiendo la instrucción para la evolución interior y después regresaron junto a su padre.

-¿Habéis tenido la experiencia de lo Sublime? - les preguntó.

Uno de los hijos comenzó a extenderse sobre esa experiencia utilizando toda clase de conceptos, palabras y retóricas filosóficas. Cuando dejo de hablar, el padre preguntó al otro muchacho, pero éste se limitó a guardar silencio. Entonces el padre dijo:

-Hijo mío, tu sí has obtenido una experiencia de lo Sublime.

- Ramiro A. Calle

Acknowledgements

At the light at the end of this bittersweet tunnel called PhD, I’d like to thank my supervisor Gianni, for his strong and always patient role model, to all the lab mates that helped me and eased the journey, to my parents for believing this day could arrive and, specially, to all the anonymous heroes in GPUGRID, the contribution of which, byte by byte, is helping us push the limits of science.

Abstract

High-throughput molecular dynamics (MD) simulation is a valuable computational tool to study protein-ligand interactions and protein conformational plasticity at an atomic resolution. In this doctoral thesis we applied it to drug discovery by (1) running the first MD-driven 150-fragment screening against the chemokine CXCL12 with a total simulation time of 8.2ms, (2) developing an application to detect cryptic binding sites based on simulations of protein in a mixed solution of water/benzene and (3) studying the molecular basis of functional selectivity by simulating the μ-opioid receptor bound to two different ligands for 500 μs. Additionally, we have developed a web platform called *PlayMolecule* where we shared with the scientific community some of the applications developed during this thesis, including a tool for protein preparation before running molecular simulations.

Resum

La simulació de dinàmica molecular (MD) d'alt rendimient és una valuosa eina per estudiar les interaccions proteina-lligand amb resolució atòmica. En aquesta tesi doctoral, l'hem aplicat al camp de desenvolupament de fàrmacs mitjançant (1) l'execució del primer cribat de 150 fragments contra la quimiocina CXCL12 usant exclusivament dinàmica molecular amb un total de 8.2ms de temps de simulació, (2) el desevolupament d'una aplicació per trobar cavitats d'unió críptiques utilizant simulacions de proteina en un solvent mixte d'aigua/benzè i (3) l'estudi de la base molecular de la selectivitat funcional realitzant 500 μs de simulacions del receptor μ-opioid unit a dos fàrmacs diferents. A més a més, hem desenvolupat una plataforma web anomenada *PlayMolecule* on hem compartit amb la comunitat científica algunes de les aplicacions desenvolupades durant aquesta tesi, incloent una eina per preparar proteines abans d'executar simulacions moleculars.

Preface

The voice of my grandmother still echoes inside me: "you will be a great scientist", she used to tell me. And I, truth be told, always aimed to be an inventor since I can remember. You know, one of those crazy-haired, carefree lunatics with more papers stacked on his desk than clean socks in his wardrobe. So, after realizing that the *Invention Faculty* and the *Royal Inventors Guild* were just a result of my overexcited imagination, I slowly became tantalized by *Science*, which finally seduced my brain and stole all my economic pretensions in exchange for having my curiosity needs fulfilled.

The first years of our unique relationship were phenomenal. *Her* ability to surprise me and the richness of biological details *she* would offer me amazed each of my neurons. For instance, I never got to learn so many names of viruses, or bacteria, or diseases, or bones in the human body. I never got to pronounce words longer than the name of that one muscle called *sternocleidomastoid* or cultivate the patience Avogadro needed to heroically count $6.022 \cdot 10^{23}$ tiny little particles in his spare time.

Our relationship started to change after we had finished the Biology chapter of our lives. Instead, Bioinformatics looked a much more mature approach. Less *wet*, perhaps, but much more computationally intensive. *Science* surprised me once again by changing my own existential paradigms: I suddenly started counting from 0, repeating "Hello world" like a possessed creature and speaking languages that I thought only Harry Potter would speak. It was a fun time.

Later on, things started to get pretty serious. *Science* started to ask me much more commitment, require more of my invaluable time and really started to give me a hard time in terms of communication. At some point I even played a joke on *her* by suggesting I needed a PhD to understand *her*! But overall, it was an instructive time: *she* taught me how to philosophize about scientific reproducibility, how to *cook* data, how to solve problems simply by pressing a *restart* button and how not to get desperate when things seem not to work out...

Luckily, I seem to have made it through. I must confess, however,

that I ended with a bittersweet taste in my mouth. I drank a dose of reality, if you will. Let me explain. We all look up to our heroes in the media thinking how lucky are they to hold this or that position, or to earn that much of a salary, or to have such a good wife, or to hold such an amazing intellect... Well. The fact is that there is no shortcut for success, there are no magical recipes or luck enough in the world to get something worthwhile for free. Behind each masterful action, there are hours, and hours, and hours of training and the patience of a stoic. Let this PhD thesis be a humble proof.

Publications

This section lists the publications that were carried out during the period of this thesis. Publications 1, 4, 5, 6 and 8 are published. Publications 2, 3 and 7 are currently submitted or under review. The numbering of the list does not apply in following sections. However, the papers under "First author" and "Co-author" categories are contained integrally in Section 3 of this thesis (Publications). The specific subsection of each paper is written in bold at the end of each reference.

First author

1. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. Martínez-Rosell G, Giorgino T, De Fabritiis G. *J. Chem. Inf Model.* **2017** Jul 24;57(7):1511-1516. doi: 10.1021/acs.jcim.7b00190. **Pub. 3.1**.

2. PlayMolecule CryptoScout: predicting protein cryptic sites using mixed-solvent molecular simulations and mutual information. Martínez-Rosell G, de Fabritiis G. Submitted to *J. Chem. Theory Comput.* **Pub. 3.2**.

3. Molecular simulation-driven fragment screening for the discovery of new CXCL12 inhibitors. Martínez-Rosell G, Harvey MJ, de Fabritiis G. Submitted to *J. Chem. Inf Model.* **Pub. 3.3**.

4. Dynamic and Kinetic Elements of μ-Opioid Receptor Functional Selectivity. Kapoor A, Martinez-Rosell G, Provasi D, de Fabritiis G, Filizola M. *Sci. Rep.* **2017** Sep 12;7(1):11255. doi: 10.1038/s41598-017-11483-8. **Pub. 3.4**.

5. Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale. Martínez-Rosell G, Giorgino T, Harvey MJ, de Fabritiis G. *Curr. Top. Med. Chem.* **2017**;17(23):2617-2625. doi: 10.2174/1568026617666170414142549. **Pub. 3.5**.

Co-author

6. High-Throughput Automated Preparation and Simulation of Membrane Proteins with HTMD. Doerr S, Giorgino T, Martínez-Rosell G, Damas JM, De Fabritiis G. *J. Chem. Theory Comput.* **2017** Sep 12;13(9):4003-4011. doi: 10.1021/acs.jctc.7b00480. **Pub. 3.6**.

7. Data Augmentation and Predictions by Molecular Dynamics Simulations and Machine Learning. Pérez A, Martínez-Rosell G, de Fabritiis G. Under review in *Curr. Opin. Struct. Biol.* **Pub. 3.7**.

Other publications

8. DeepSite: Protein binding site predictor using 3D-convolutional neural networks. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. *Bioinformatics.* **2017** Oct 1;33(19):3036-3042. doi: 10.1093/bioinformatics/btx350. **Pub. 6.1**.

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Drug discovery: molecular recognition

The origin of drug discovery dates back to ancient times, when natural products, mainly extracted from plants, were used for medicinal purposes. In the early times, just like in any developing discipline, serendipity and empiricism would drive the discovery and application of new therapies. However, the idea of "chemoreceptors" by Paul Erlich in 1872 and the conceptual description of receptors by Langley in 1905 marked a point of inflection that augured the beginning of rational drug design [1]. In particular, Langley's description of the receptors as "switches" that receive and generate specific signals and that can be either blocked by antagonists or turned on by agonists [2] established the seeds of our current theories on pharmacology and mechanisms of action from a structural standpoint.

Different models of protein-ligand interaction have been formulated along the years, starting by the basic lock-and-key model [3] that states that a protein and its ligand possess geometric complementarity and that specificity is explained as a result of one fitting perfectly into the other. Half a century later, in 1958, once proteins started to be understood as dynamic and flexible structures, induced fit model superseded lock-and-key [4]. This new model suggested that the interaction of the ligand with the protein was able to induce and stabilize a particular protein conforma-

tion. Soon after, in 1965, conformational selection model was introduced [5] postulating that, differently from induced fit, the protein alone already fluctuates along a number of intermediate states among which, one particular conformation, is able to bind the ligand. Discussions and debates over which of the latter paradigms is the correct one have survived until the present day [6, 7, 8, 9]. In fact, is likely that both mechanisms may play a role in a system-dependent manner.

While most of the first drugs were discovered by serendipity [10], such as the emblematic penicillin [11], or even in absence of a tridimensional structure of the ligand and its receptor, such as in early steroid studies [12], nowadays drug discovery is driven by biological targets, genetic studies, transgenic animals models, molecular biology, gene technology or protein science, although serendipity still plays a role in late stages of drug development especially when we are still unable to efficiently predict drug activity and properties until they are tested on animal models.

When the tridimensional structure of a protein and a ligand is known, a particular type of drug discovery called structure-based drug design can be applied. In this type of drug design, the interactions between a protein and a ligand can be described or modeled with atomic resolution. In particular, one can study the binding of the ligand in terms of non-bonded interactions established between the protein and the ligand (Van der Waals repulsive and attractive forces, Hydrogen-bonds, salt-bridges, and mediation by water molecules and ions).

1.1.1 Fragment-based drug design (FBDD)

One of the steps in early drug discovery, once the protein target is defined, is discovering compounds with a high potential to bind the protein and becoming a marketable drug. These compounds, usually called leads, can be found by screening libraries of ligands, either experimentally, i.e. high-throughput screening (HTS) [13], or *in silico*, i.e. virtual screening (VS) [14]. A particular strategy to find leads is called fragment-based drug discovery (FBDD). Because of the relevance of FBDD in the current thesis and included publications, it is worth explaining what the charac-

teristics and advantages of FBDD are.

FBDD started to get popular in early 2000s as an alternative to high-throughput screening or virtual screening of drug-like molecules. This tendency has continued until our days to the point that FBDD has become a mainstream technique and the driver technology of more than 30 drug candidates [15]. The main characteristic that differentiates FBDD from a typical drug-like HTS approach is the size of the ligands employed in the screening phase. In particular, fragments are usually defined as having less than 20 non-hydrogen (or "heavy") atoms while drug-like molecules can go up to 30 heavy atoms or more [15]. Therefore, while the objective of a HTS technique is to find directly a drug-like lead, the approach used in FBDD is to discover small millimolar-binding fragments with high ligand efficiency [16, 17] (LE) that can later be extended or linked together to form a drug-sized lead [18].

Several advantages characterize FBDD. First, the smaller size of the ligands reduces the accessible chemical space. A study calculated that each heavy atom adds roughly one order of magnitude to the number of possible chemical combinations [19]. This implies that the chemical space of drug-like molecules is many orders of magnitude bigger than the fragment chemical space. A practical consequence of this fact is that a fragment library usually consists of only 1,000-5,000 compounds [20] while a drug-like library usually comprises between 0.5 and 3 million compounds [13]. Second, fragment libraries have been reported to yield higher hit rates than HTS [21, 22]. The rationale behind this observation is that, as molecules grow, there is more probability that a chemical group causing an unfavorable interaction is included in the molecule and that the introduction of this group ruins completely the affinity for the target. Conversely, fragments, due to their small size, establish less interactions with the target and should be able to bind to a greater number of sites. Moreover, the quality of interactions between a fragment and a protein is usually high, as supported by the conservation of the binding mode as the fragments are grown into larger molecules [23, 24]. These characteristics make FBDD specially appealing to tackle difficult targets such as allosteric sites or protein-protein interaction interfaces (PPIs).

1.1.2 Current methods in biophysics

Several experimental techniques are routinely used in structural biophysics and in particular in FBDD [15]. These techniques are constrained within a temporal and spatial resolution range (Fig. 1.1) and their use will depend on the system and the question at hand.

One of the most important techniques in structural biophysics is X-ray crystallography. This particular technique revolutionized the biophysics field since the first crystallization, performed on Mioglobin in 1958 [25], and allows us to describe the topology of proteins, and sometimes ligands, with atomic resolution. These crystal structures are static snapshots that represent either an ensemble of protein conformations in equilibrium or the most stable conformation. However, crystallography sometimes fail to resolve flexible regions such as loops. Often, these structures are released publicly in the PDB database [26, 27], where the number of protein structures raised from only 507 structures in 1990 to more than 130,000 structures in October 2017, more than 89% of which come from crystallography [28]. Cryo-electron microscopy (cryoEM) is a related technique which allows to observe bigger structures but with lower resolution. The application of crystallography in FBDD is exclusively to unravel the binding mode of fragments and does not yield any affinity information.

Another popular technique used in structural biology and that accounts for an approximate 9% of the total number of structures in PDB is nuclear magnetic resonance (NMR) spectroscopy. NMR can be used to determine not only the structure but also the dynamics of an array of biomolecules including proteins, nucleic acids, carbohydrates and many metabolites [29]. A major advantage of NMR is that it can quantitatively describe populations and exchange rates between various conformers. Furthermore, the application of NMR in drug discovery is quite straightforward, for instance: the addition of a ligand in a solution with our favorite protein will cause a change in the NMR observable spectra (e.g. chemical shift, NOEs, relaxation times, etc.) if the binding event occurs [30]. This way, one can leverage NMR in FBDD to detect the affinity of millimolar-binding fragments and, on top of this, have a rough approxi-

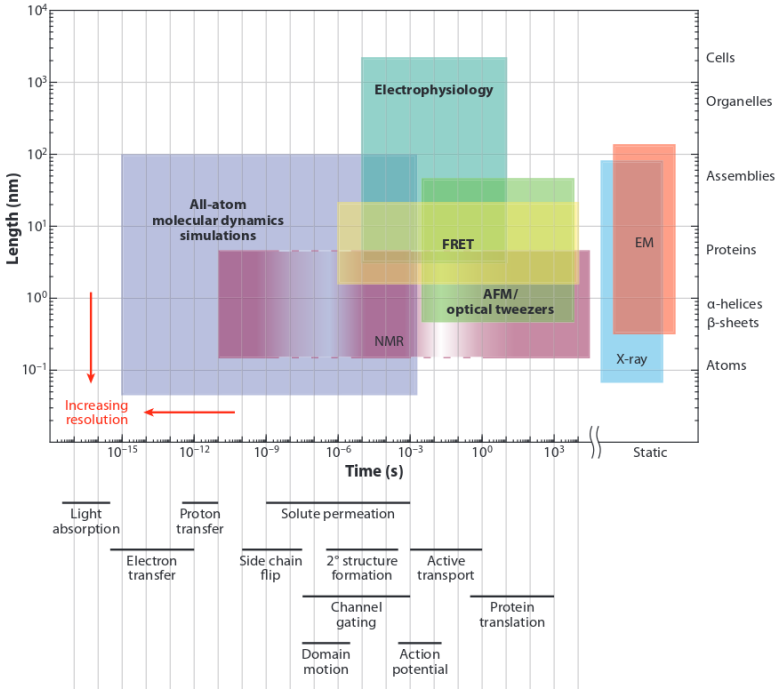


Figure 1.1: Plot showing the main investigative methods routinely used in biophysics in regard the spatial and temporal resolution they offer. Below the horizontal axis there are the timescales involved in many biological phenomena and in the vertical axis the size of different molecular constituents. Figure extracted from [31].

mation of the protein-ligand interaction location.

Several other experimental techniques exist that can be used to study structural and dynamic system-specific information. For instance, fluorescence resonance energy transfer (FRET) can be used to study protein folding [32] and surface plasmon resonance (SPR) can be used to detect the binding of low-affinity fragments by measuring changes in the refractive index [15].

Additionally, several computational methods have been developed over the years to tackle drug discovery and drug development computationally

[33]. From a protein structural point of view, homology modelling allows to infer the tridimensional structure of a protein by comparing its protein/DNA sequence to the sequence of proteins with known structure. Although cost-effective, problems associated with template identification, sequence alignment and refinement hinder its wider use in drug discovery [34]. Another method, docking, is usually employed in virtual screening. Docking works by fitting a ligand into a protein cavity and evaluating the fitness with a scoring function. Several algorithms and implementations exist, such as AutoDock VINA [35], Glide [36] or Gold [37]. Although computationally fast, docking suffers from a lack of protein flexibility, which some algorithms such as flexible docking have tried to mitigate to some extent [38]. It is worth noticing that, in general, the applicability of docking in FBDD has been quite limited to date, partially due to promiscuity of fragments binding mode [39, 40] and docking limited ability to correctly describe protein conformational plasticity [38, 41] and to score fragments [33].

Simulations also can be used to retrieve structural information. For instance, a particular type of simulations called Monte Carlo simulations work by introducing small random changes in the system, such as dihedral rotations, and evaluating the validity of the new structure by comparing its energy with the energy of previous structures. This allows to fold proteins *de novo* for relatively simple cases, being Rosetta [42] an emblematic application example of this type of simulations. Another type of simulations, quantum mechanics (QM), is used when quantum processes such as enzymatic reactions want to be studied. The high computational cost of these simulations limits the size of the system and the length of the systems investigated usually down to few atoms and maximum one nano-second, respectively. This makes it impractical to study processes happening in longer timescales. The development of hybrid QM/MM [43, 44] methods allowed us to simulate bigger systems by only running QM-level simulations on a small subset of atoms and the rest in a molecular mechanics (MM) fashion. In this thesis, we have used QM to optimize small ligand geometry and charge, as well as inferring drug dihedral parameters at MM-level by performing energetic scans along the dihedral

angles.

Another type of simulation, molecular dynamics (MD), models reality with atomic resolution and therefore is able to reach bigger time and spatial scales than QM. The reduction of computational costs and force-field improvements has made this technique especially valuable for drug discovery [45]. Although other computational techniques such as docking were also employed, MD was the main investigative tool used in this thesis.

1.2 MD applied to drug discovery

1.2.1 MD: Jigglings and wigglings

Richard Feynman in 1965 once described life as "jigglings and wigglins" of atoms [46] and is precisely the jigglings and wigglings of atoms what biophysics tries to understand and what molecular dynamics tries to model.

Molecular dynamics simulation is a computational method for studying the physical movements of atoms and molecules. In MD, atoms are treated as point masses and the bonded and non-bonded interactions between them are modelled by empiric force-fields [47]. Given a velocity for each of the atoms, which is usually initially randomized, one can generate trajectories of atom movements by using numerical integration schemes such as Verlet integration to solve Newton's equation of motion [48]. These trajectories or simulations produce "narratives" of the events occurring at a nanoscopic scale with an atomic resolution and have been valuable to describe a wide range of phenomena including protein-ligand binding, protein conformational changes, protein folding, etc. Section 1.2.7 further expands on the applicability scope and successful stories of MD in drug discovery.

1.2.2 Force-fields

Most of current force-field implementations differ little from the formula in Figure 1.2, typically including bond terms such as inter-atom bonds,

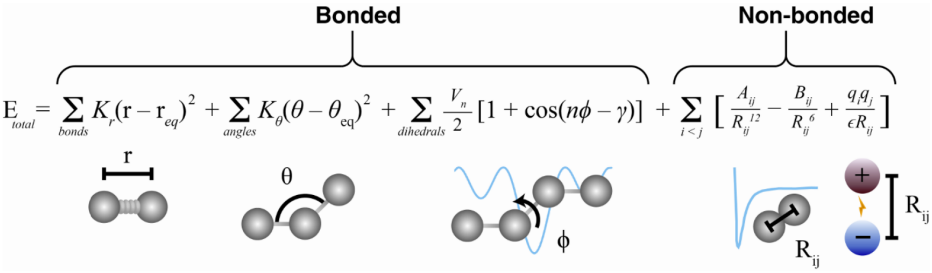


Figure 1.2: Basic equation, including bonded and non-bonded terms, for an MD force-field. Figure extracted from [61].

angles and dihedral angles, and non-bonded terms, typically a van der Waals term and a Coulomb term.

Main force-fields used in academia include CHARMM [49, 50], Amber [51, 52] and OPLS [53]. Parameters for their force-field terms are either derived from QM simulations or adjusted to match experimental observables. For instance, bonds lengths and angles can be extracted from crystallographic structures. While the official most recent version of these force-fields are CHARMM36 [54], ff14SB [55] and OPLS-AA/M [56], we use a modified version of CHARMM called CHARMM22* [57] that was modified by Piana *et al.* to solve overstabilization of helices and salt bridges. Furthermore, general force-fields for small organic molecules have also been developed such as GAFF [58, 59] for Amber and CGenFF [60] for CHARMM.

1.2.3 Software, hardware and future perspectives

Note: parts of this section were taken from my Publication 3.5.

Despite the low algorithmic complexity of MD in comparison to quantum chemistry methods, the computational cost is such that high performance computing (HPC) systems have been required to perform simulations of sufficient length to approach biologically relevant timescales [62]. The size and specialization of the parallel HPC systems required has made

MD sampling of even small biologically-interesting systems very costly in terms of Euro per simulated time. Consequently, much technical effort has been invested in developing specialized hardware, such as Anton supercomputer [63], and simulation software optimized to maximize performance on these machines.

In the latter half of last decade, developments in the computer graphics technology sector resulted in the introduction to the HPC field of a new class of processor with radically different characteristics to conventional CPUs. The characteristics of these processors, termed GPUs (graphics processing units), make them highly amenable to certain classes of scientific computation, in particular those such as MD which contain a high degree of intrinsic parallelism. The most efficient GPU MD codes are those such as ACEMD [64] and recent versions of PMEMD [65], OpenMM [66] and Desmond [67], all of which have been designed and optimized specifically for the architecture of GPUs. The computational cost reduction has been remarkable over the last years and one can expect single GPU simulation rates in the order of the μs/day by 2022 for systems of intermediate size (*circa* 50,000 atoms including solvent). When further coupled to a computing infrastructure that delivers access to large numbers of GPUs, such as GPU-based HPC machinery, or a distributed computing network like GPUGRID [68], we can extrapolate (such as done in in Publication 3.5) that by 2022, MD-based studies will employ aggregate sampling on the second timescale (Fig. 1.3). Interestingly, the aggregated simulation time of Publication 3.3 (currently under review) still correlates well with the predicted trend.

1.2.4 High-throughput molecular dynamics and MSMs

The MD field has experienced drastic improvements since the first MD simulation ever performed was produced by Karplus in 1977 on BPTI for 8.8ps [76]. In fact, since not so long ago, anecdotical simulations and single simulation studies have been superseded by more rigorous and extensive simulations following the principle that a single observation is not sufficient to answer hypothesis in a statistically significant way. In

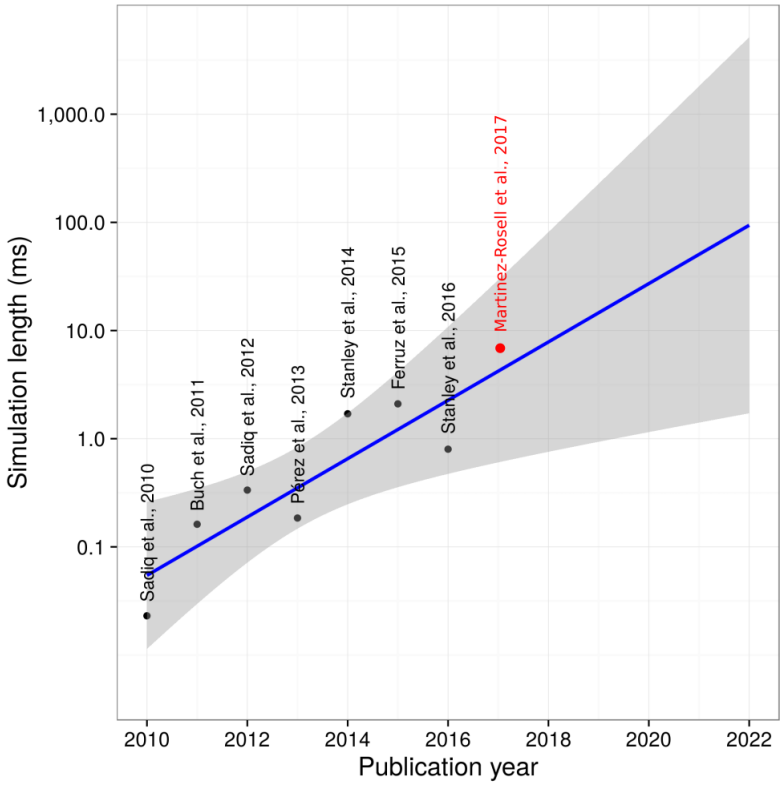


Figure 1.3: Approximate total aggregate sampled time for high-throughput all-atom molecular dynamics studies with maximum total simulation time per year using ACEMD software [64] published in years 2010-2017 (log scale). An exponential function (solid line with 95% confidence interval) was fit to the largest studies of each year (black dots). Red dot corresponds to Publication 3.3. The trend indicates that we will reach the second timescale by 2022. References used for this plot are, from left to right: [69], [70], [71], [72], [73], [74], [75] and Publication 3.3. Modified from Publication 3.5.

particular, the implementation of MD codes that can run on GPU, as well as the creation of specialized ASICs such as ANTON [63], combined with high-performance clusters (HPC) or distributed computing networks such as GPUGRID [68] have radically expanded the amount of simulation time we can have access to.

Although one could produce several very long simulations, in highly parallel clusters such as GPUGRID, the production of hundreds of short simulations results more cost-efficient. Beyond this technical limitation, long simulations may get trapped in metastable states or may produce unrealistic trajectories due to force-field errors.

Instead, a high number of short trajectories can be effectively produced and analyzed using a mathematical framework called Markov State Models (MSMs) [77] which are able to describe processes happening in longer timescales than a single simulation length. In Figure 1.4, we describe a very simple but visual example of how MSMs are built.

In order to create an MSM, first we need to project the high dimensional space contained in a MD simulation (N^3 where N is the number of particles) into a lower dimensional space, for instance contact maps of a ligand with each residue of the protein (contact map with N dimensions). We can even reduce the dimensionality further by projecting the aforementioned projection (e.g. contact maps) into tICA space [72], which is similar in concept to PCA but instead of placing the axes along the highest variant coordinates it places them along the slowest processes coordinates, which are usually the biologically relevant. Then, we have to discretize a continuous space into a discrete space, this is we have to cluster tICA coordinates or contact maps into a number of states (note that contact maps are already discrete, but there is a need to reduce dimensionality even further into few clusters, e.g. 1000). We can achieve this by using a clustering algorithm such as Kmeans [78]. Then, we calculate a N to N transition matrix (where N is the number of clusters) by counting how many times a trajectory in a state x jumps to state y or remains in the same state for a given lag time. From the transition matrix one can extract the equilibrium probability of each microstate and the microstates can be clustered together into few macrostates in order to ease human

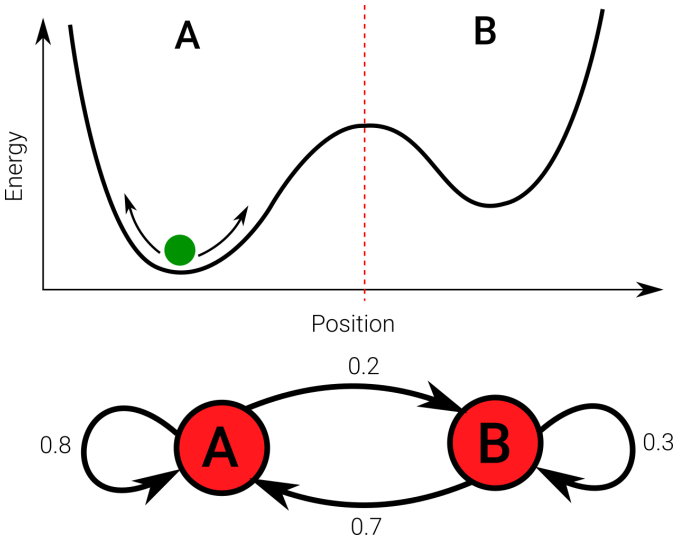


Figure 1.4: Visual example of how an MSM works. Lets imagine a given green ball that moves in the energetic landscape due to thermal fluctuations. The landscape is a continuous space but we could discretize it by defining a well A and a well B separated by the saddle point. One could expect that the ball would move around in well A and sometimes it would jump to well B. Due to the depths of the wells, one would also expect that the ball would spend more time in well A than in well B. If one records the ball trajectory (the equivalent to a MD trajectory) and writes down the state of the ball at regular intervals (e.g. AAABBAAAAA) one could create a transition matrix at a specific lag time by counting how many times the ball in A goes to B or stays in A, and the same for well B. From this transition matrix one can obtain the probability of the ball being in state A and B. This simple two-state model could represent the binding of a ligand to a protein (A being bound state and B being unbound state) or a folding process (A being folded state and B unfolded state).

visualization using an algorithm such as Perron-cluster cluster analysis (PCCA) [79]. The visualization of the macrostate can help us to identify, for instance, the binding pose of a ligand. Several metrics can be extracted from the transition matrix and from the equilibrium distribution of the macrostates. For instance, a metric widely used throughout this thesis is the protein-ligand binding free energy (i.e. binding affinity) that can be calculated from the probabilities of a bulk state and a bound state by using the Boltzmann distribution:

$$\Delta G = -K_{\mathrm{B}} T \ln \left(\frac{P_{\mathrm{sink}}}{P_{\mathrm{bulk}} c} \right), \tag{1.1}$$

where ΔG is the Gibbs free energy, K_B i s the Boltzmann constant in kcal/(mol·K), T is the temperature (300K), P_{sink} is the equilibrium probability of the sink or bound state, P_{bulk} is the equilibrium probability of the bulk or unbound state and c is the concentration of the ligand.

Other metrics such as kinetics (k_{on} and k_{off}) between two macrostates or mean first passage time can also be obtained. This complex mathematical framework can be easily applied to analyze our MD simulations by using software such as HTMD [80], implemented in python language, which leverages pyEMMA [81] to build MSM.

1.2.5 Adaptive sampling

While the computational power has increased over the years and has allowed us to access to longer timescales, the amount of simulations necessary to converge MSM statistics still remains very high. One way to reduce the computational cost and help the statistics to converge is to increase the sampling of rare events and unexplored configurational space. In practice, this means that we can re-spawn simulations from undervisited states in an MSM. This is, if a state has only been visited by very few simulations, our statistics about whether the state is stable or not will have a high associated error. In order to solve this, we can re-spawn simulations from that state and see if the system "likes" to stay there or, on the contrary, easily jumps to other states. This way we can also avoid sam-

pling states for which we already have lots of statistics. For instance, in a protein-ligand system, ideally we would like to reduce the sampling of the bulk (i.e. unbound) state and increase the sampling of sink (i.e. bound) or quasi-sink states. These simulation schemes are popularly known as adaptive sampling schemes. For the publications of this thesis, we have used the adaptive sampling scheme implemented in HTMD [80]. This particular implementation has proven to reduce the computational cost at least one order of magnitude in the benzamidine-trypsin system [82].

1.2.6 MD limitations

The validity of MD simulations is highly dependent on several factors: (i) whether the starting macromolecule structure and coordinates are correct, (ii) whether the protonation states of the protein residues are correctly assessed, (iii) whether the force-field of use approximate well the atom-atom interaction forces acting in nature.

In order to address the first two points, we developed the ProteinPrepare application (Publication 3.1) in which the hydrogen-bond network of a protein is optimized and the residues protonation is assessed by titrating protein residues at a given pH. However, solving force-field limitations can prove much more complex. For instance, in case of small organic ligands, we used parameterization tools such as GAAMP [83], where quantum mechanics simulations are performed to scan dihedral angles of the ligands and then a fitting procedure is applied to fit the force-field parameters to the energetic profile obtained for those dihedrals (Figure 1.5).

Quantum phenomena such as polarizability and protonation changes are not usually regarded by classical MD force-fields and the extent in which the lack of these terms may affect the results is probably system-dependent. In order to solve these issues, polarizable force-fields [84] such as AMOEBA [85], constant-pH simulations [86] and hybrid QM/MM simulations [87] have been developed. However, these improvements always come in exchange for a higher computational cost and this is one of the reasons why their use in research and industry is yet to become mainstream.

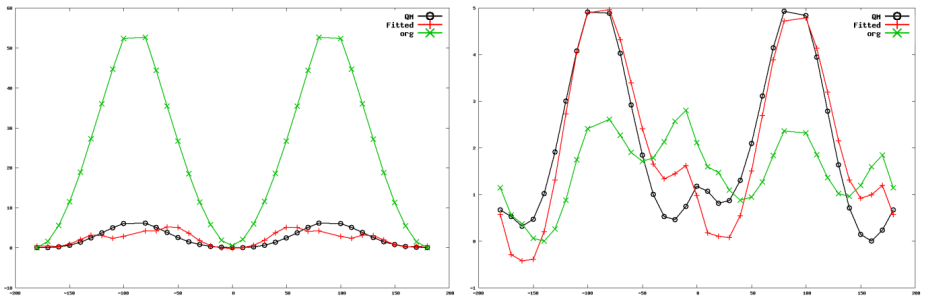


Figure 1.5: Examples of force-field dihedral energies fitted from QM data using GAAMP [83]. In green is depicted the energy profile for the dihedral angle before the fitting procedure (original force-field parameters, in this case CGenFF). In black is the energy profile obtained from 1D QM scans. In red is the energy profile of the force-field dihedral parameters after fitting the QM data.

1.2.7 Evolution of MD applications in drug discovery

In Publication 3.5 we followed the historical trajectory of the oldest MD code implementation for GPU: ACEMD. The objective of the publication was to describe the evolution of MD applied to drug discovery and how the scientific community was able to tackle increasingly more complex tasks. We outline here some of the milestones achieved in the MD field applied to drug discovery (Figure 1.6).

ACEMD was released in 2009 and already in 2010 the first publications appeared focusing peptide-protein, ligand-protein and ion-protein binding using ACEMD as simulation software. However, the limited sampling time available at that moment (partially due to slower GPUs) motivated the use of biased simulation techniques such as umbrella sampling [88] or metadynamics [89], which rely on the knowledge of reaction coordinates (i.e. collective variables) along which sampling is enhanced. Note, however, that unbiased simulations were also used to observe single isolated binding events of ion-protein, whose binding kinetics are extremely fast and therefore observable within few nanoseconds simulation

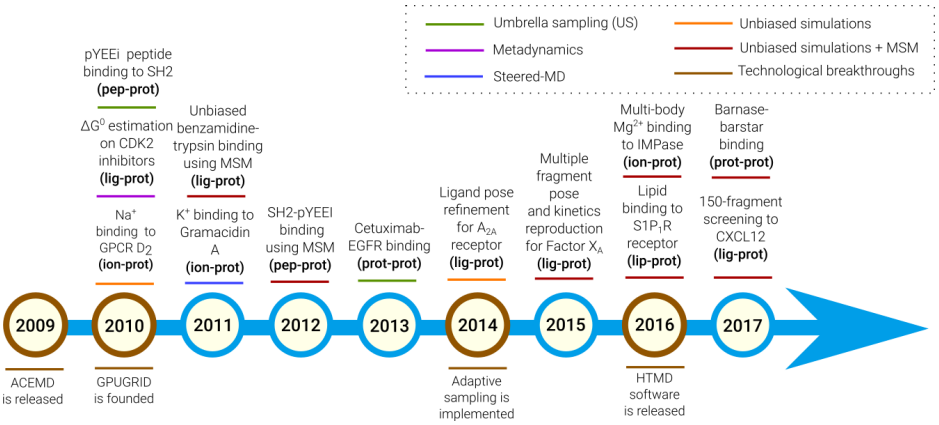


Figure 1.6: Evolution of the MD field from 2009 to 2017 in terms of applications in drug discovery using the software ACEMD. Publications listed from left to right and from up to down are: [64], [68], [90], [91], [68], [70], [92], [93], [94], [95], [82], [74], [96], [75], [80], [97] and Publication 3.3.

time.

The creation of distributed computing networks such as GPUGRID in 2010 [68] allowed us to launch and simulate hundreds of parallel simulations which could be later analyzed using Markov State Models (MSMs) to obtain binding states (i.e. binding poses), kinetics, state equilibrium distribution and therefore binding free energies. This particular setup was employed in the landmark study of benzamidine-trypsin binding [70], which was one of the first efforts to demonstrate the utility of the so-called high-throughput molecular dynamics [62], a new paradigm that, opposed to the single simulation studies, leverages hundreds of short simulations to describe biological processes with timescales longer than a single simulation time. By 2012, full pathway reconstruction of peptide-protein binding processes were produced using distributed computing and MSMs [93]. In 2015, a similar technology that reproduced trypsin-benzamidine binding was applied in the first multiple ligand binding reconstruction with a total of 15 ligands against the protein factor Xa [74]. In 2016,

the first unbiased multi-body [96] and lipid-protein binding [75] studies were published using ACEMD. Finally, in 2017, latest MSM innovations plus a decrease in computational cost allowed us to tackle full pathway reconstruction of protein-protein binding events with the Barnase-Barstar system [97]. Additionally, in Publication 3.3 we fully leveraged adaptive sampling scheme developed in 2014 [82] to produce the first large-scale fragment screening exclusively using MD and MSM.

The aforementioned studies help us to draw a general picture of how the field is steadily pushing the limits to reach harder and harder milestones: from single simulation ion-protein binding events to converged multi-body high-throughput studies, from binding free energy prediction using metadynamics simulations to converged free energy calculations for tenths of ligands using high-throughput unbiased simulations. This scientific advancement has been the fruit of an interdisciplinary effort: implementation of faster MD codes, improvement of hardware specifications, enhancement of analysis tools, acquisition of a better scientific understanding of the biologic systems and phenomena, adjustment of the force-fields, collaboration of hundreds of computing time donors, etc.

1.3 *PlayMolecule*: the computerization of the drug discovery pipeline

One of the main contributions of the current thesis is the transfer of scientific knowledge to a web platform that we have named *PlayMolecule*. The aim of the platform is offering all the generated know-how and tools to the scientific community for the better and faster advancement of basic research and drug design. As we have seen in the previous section, our experience in drug discovery and the continuous advancement towards more sophisticated methods has yielded the creation of a number of applications, some of which have been packaged into accessible web apps (apps in red in Figure 1.7) in the *PlayMolecule* platform (*www.playmolecule.org*).

In Figure 1.7 we show a typical *in silico* drug discovery pipeline and

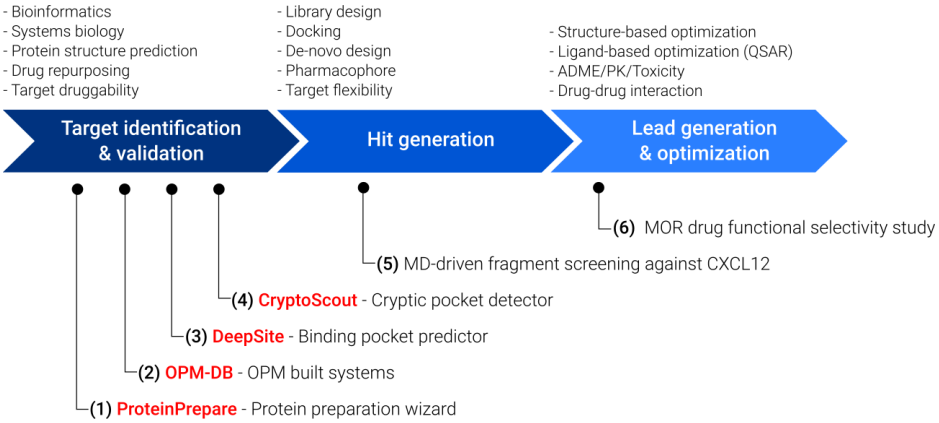


Figure 1.7: *PlayMolecule* applications and Publications in the drug discovery pipeline. (1) to (6) refer to Publications 3.1, 3.6, 6.1, 3.2, 3.3 and 3.4.

the publications included in this thesis annotated below. Interestingly, publications span all along the drug discovery pipeline, starting from protein preparation (Publication 3.1) where the crystallographic structure of a protein is titrated, protonated and optimized in terms of H-bond network. Then, in Publication 3.6 we offer a database of built membrane systems ready for simulation. In Publication 6.1 we describe a method that allows to detect binding pockets based on convolutional neural networks trained on scPDB database [98]. Publication 3.2 offers a method for cryptic pocket detection using mixed-solvent simulations of the protein solvated in water and benzene. In Publication 3.3 we perform the first MD-driven fragment screening. Finally, in Publication 3.4 we study the conformational changes of a GPCR based on ligand-mediated modulation.

PlayMolecule is the fruit of several prototypes and iterations starting as early as 2014 (Figure 1.8). The final result is a modular platform that leverages latest technologies such as AngularJS, Angular Material and NGL [99] protein viewer for the client-side and Flask server, slurm queue, HTMD [80] analysis package, ACEMD [64] and the invaluable

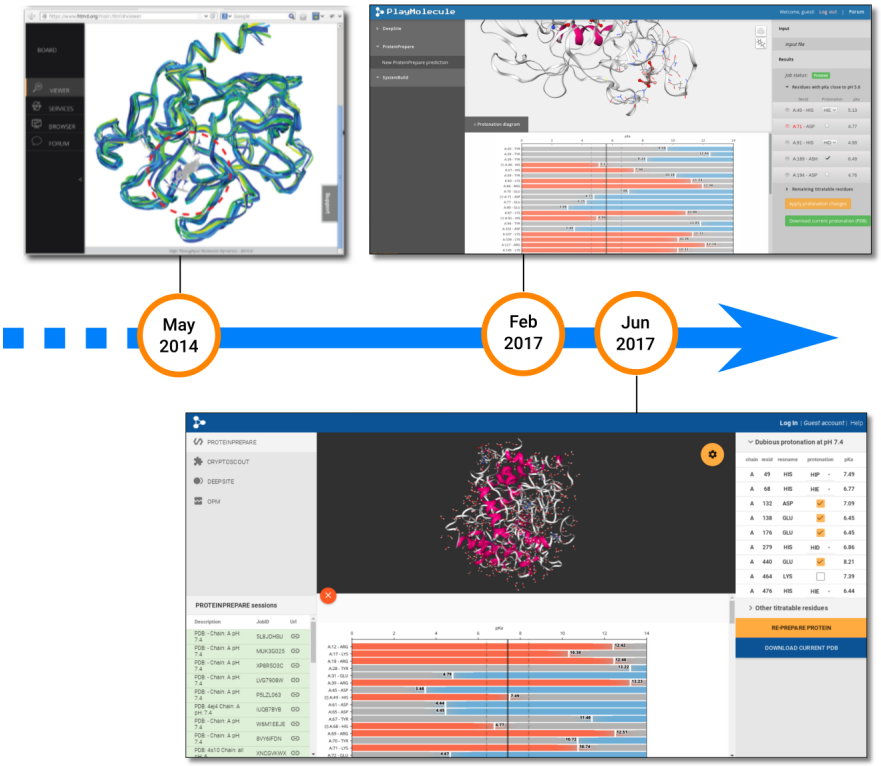


Figure 1.8: Evolution of the *PlayMolecule* web platform in terms of prototyping and design.

collaboration of thousands of GPU contributors in GPUGRID [68] in the server-side.

1.4 Biological systems investigated

Note: parts of this section were taken or adapted from my Publications 3.3 and 3.6.

This section will give an overview of the systems studied in this doctorate.

In particular, CXCL12/SDF-1 and μ-opioid receptor systems have been extensively investigated in Publications 3.3 and 3.4, respectively. Note that in Publications 3.2 and 3.6, a total of 18 and approximately 700 systems have been built and simulated, respectively. However, these systems were built and simulated in an automatic manner with little or no manual work involved and therefore were not thoroughly studied further than in the particular scope of the paper.

1.4.1 CXCL12/SDF-1

CXCL12 (stromal cell-derived factor-1/SDF-1) is a chemokine, a small dimerizable soluble protein that stimulates chemotactic cell migration via activation of a G-protein coupled receptor (GPCR) [100]. Its structure consists of a C-terminal α-helix, three anti-parallel β-sheets and a N-terminal flexible loop (Fig. 1.9A). CXCL12 and its receptor CXCR4 are particularly well studied and their participation in physiological processes [101] (e.g. embriogenesis, wound healing, stell cem homing) as well as morbid processes (e.g. autoimmune diseases [102], cancer [103, 104, 105], HIV [106, 107]) is known.

In particular, the significant role of the CXCR4/CXCL12 axis in metastasis, tumor survival and tumor angiogenesis has raised the interest in developing targeted drug therapies [103]. While most attempts have been focused in inhibiting the receptor CXCR4 [108], which presents a clear druggable cavity where the chemokine CXCL12 docks, targeting CXCL12 has traditionally deemed "undruggable" due to its shallow surface [109].

However, recent studies have shown that CXCL12 surface is not completely flat. In fact, scientists have learned about CXCL12 druggability by studying the interaction between the chemokine and its receptor. This protein-protein interface has been resolved via NMR in several cases, one of which is displayed in Figure 1.9B. From the inspection of these structures and mutation studies, we learned the CXCL12-CXCR4 interaction and affinity is mediated by key CXCR4 residues [110, 111, 112], some of the most important being tyrosines 7, 12, 21 (Fig. 1.9B) and isoleucines 4 and 6. The o-sulfation of the aforementioned tyrosines (Fig. 1.9B) in the

Golgi apparatus seem to selectively enhance the affinity of CXCR4 for CXCL12 [112]. Furthermore, CXCL12 has also been resolved bound to heparin [113] (Fig. 1.9C). The binding of all these residues to CXCL12 involve the formation of small pockets and therefore reveal potential binding hot spots (Fig. 1.9D and 1.9E) that can be leveraged to design and dock specific inhibitors. Consistently with this hypothesis, recent studies report small molecules binding to sY21 [109, 114, 115], sY12 [116] and I4/I6 [116] binding pockets.

1.4.2 μ-opioid receptor (MOR)

μ-opioid receptor (MOR) is a member of the family of G protein-coupled receptors (GPCRs). Opioid therapeutics that target the main (orthosteric) MOR binding site remain the preferred treatment for chronic pain, which is known to affect more individuals than those impacted by cancer, heart disease, and diabetes combined [117]. However, these classical opioid drugs (e.g., morphine) produce a number of dangerous side effects (e.g., respiratory depression), which have captured the public's attention due to an increased number of opioid overdose deaths in the last decade [118, 119, 120].

MOR undergoes specific conformational changes upon ligand binding, leading to the activation of G protein and/or β-arrestin signaling pathways. Notably, suppression of morphine's analgesic efficacy in MOR knockout mice suggested that this receptor is absolutely necessary to mediate morphine action on pain pathways [121]. While mice lacking β-arrestin2 exhibited enhanced morphine analgesia suggesting that the drug's beneficial effect is mediated by G proteins, MOR-dependent β-arrestin recruitment appeared to contribute to some of the side effects of classical opioids [122, 123, 124].

Although the majority of known opioid analgesics activate both G protein and β-arrestin pathways, a few MOR ligands have recently been shown to have an improved pharmacological profile in vivo by virtue of their G protein-biased agonism. Among them is TRV-130, a potent analgesic exhibiting less respiratory depression and constipation than mor-

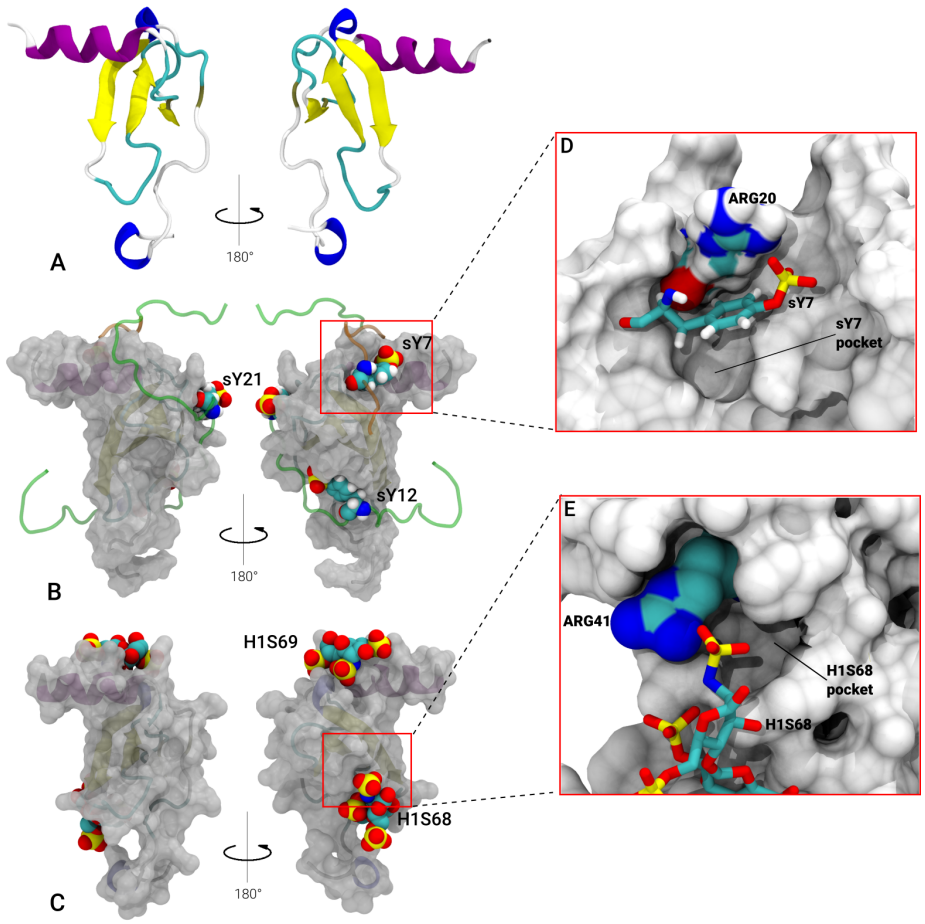


Figure 1.9: CXCL12 system overview. **A**. CXCL12 monomer (chain A of PDB 4UAI) in cartoon-style colored by secondary structure. **B**. CXCL12 monomer (chain A of PDB 2K05) represented as grey surface bound to CXCR4 (green and orange chains) with sulfo-tyrosines depicted in VDW-style. **C**. CXCL12 monomer (chain A of PDB 2NWG) bound to two heparin molecules depicted in VDW-style. **D**. Detail of the sY7 pocket. **E**. Detail of the H1S168 pocket.

phine [125, 126], which is currently being evaluated in human clinical trials for acute pain management [127, 128, 129].

Comparison between the high-resolution crystal structures of inactive [130] and activated MOR [131] bound to the morphinans β-funaltrexamine (β-FNA) and BU72, respectively, suggests very small structural differences in the extracellular region of the receptor with larger conformational changes occurring at its cytoplasmic side as the result of a large outward movement of transmembrane helix (TM) TM6 relative to TM3 (Fig. 1.10A) and smaller inward movements of TM5 and TM7 (Fig. 1.10A). Notably, the classical $R^{3.50}$-D/$E^{6.30}$ salt bridge (superscript numbers refer to the Ballesteros and Weinstein's generic numbering scheme [**?**]) that stabilizes the inactive conformation of TM6 in a number of inactive GPCR crystal structures (e.g., refs [132, 133, 134, 135, 136]) does not form in MOR because of the lack of an acidic amino acid at position 6.30. This salt bridge is replaced by a hydrogen bond between $R165^{3.50}$ and $T279^{6.34}$ in the MOR inactive crystal structure and a hydrogen bond between $R165^{3.50}$ MOR and $Y252^{5.58}$ in the MOR activated crystal structure. Together with residues $N332^{7.49}$, $Y336^{7.53}$, $L158^{3.43}$, and $V285^{6.40}$, $Y252^{5.58}$ is also involved in a hydrogen bonding network that stabilizes the inward movement of the so-called $N^{7.49}PxxY^{7.53}$ (Fig. 1.10B) motif towards TM5 in the MOR activated crystal structure.

1.4.3 Eukaryotic membrane proteins from the OPM

For Publication 3.6, we built and equilibrated all the eukaryotic membrane proteins of the OPM database [137] using AMBER and CHARMM force-fields, with the exception of 9 systems in both force-fields due to the presence of non-standard residues and additional 8 systems in AMBER due to the presence of deprotonated arginines, which are not supported by AMBER. The total number of systems automatically prepared, built and equilibrated was 699 for the CHARMM force-field and 691 for the AMBER system. Overall, the built database contains membrane proteins with a variable size between 10 and 4,898 residues (1.1 to 410 kDa), the 90% of which ranged from 3.0 to 175.5 kDa with a median of 31.1 kDa. Addi-

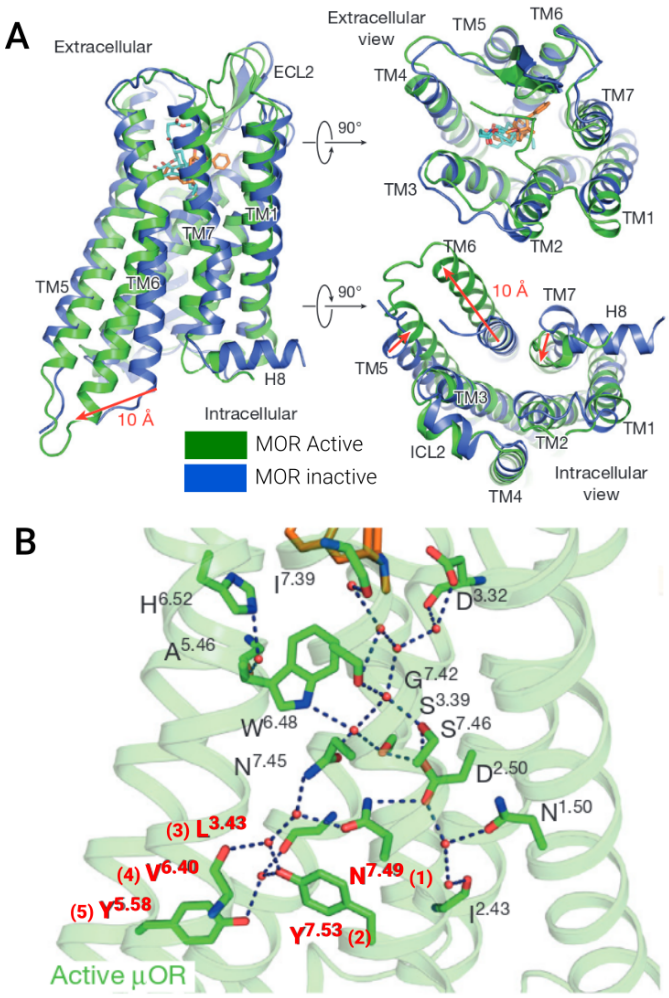


Figure 1.10: μ-opioid receptor (MOR) overview. **A**. Structural differences between the active and inactive conformation of MOR. Biggest rearrangements include mainly TM6 and TM5 and TM7 to a lesser extent. **B**. Internal H-bond network of the active MOR. The amino-acids forming part of the NPxxY motif are written in red and numbered from (1) to (5).

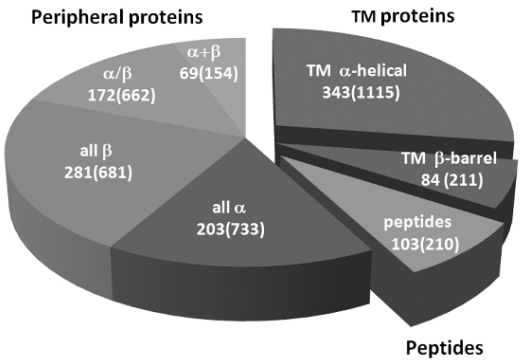


Figure 1.11: Distribution of the different OPM entry types (as of 20 July 2011). Note that the entire OPM is featured, therefore not only contains eucaryotic membranes but also prokaryotic. Image extracted from 1.11.

tionally, the OPM database classifies the proteins in a four-level hierarchy, the first one of which is the *type* according to which each protein is described as a (1) transmembrane (TM) protein, a (2) peripheral/monotopic protein or a (3) peptide (Fig. 1.11).

1.4.4 Cryptic pocket-containing protein test set

In Publication 3.2 we developed an application to detect cryptic cavities based on mixed-solvent MD simulations. In order to test our application, we aimed to find pairs of structures, where one of the structure had the cryptic pocket closed (*apo*) and the other structure had the cryptic pocket opened by the binding of a ligand (*holo*). Hence, we assembled a comprehensive set of 18 cryptic pocket-containing proteins including classic systems such as interleukine-2 (IL-2), Polo-like kinase 1 (PLK1) and β-lactamase (TEM-1), whose cryptic pockets are throughly studied in the literature, and additional 15 systems obtained from a recent publication [138]. These extra 15 systems were chosen based on the following characteristics: globularity, small to medium size (i.e. less than 250 residues), structure-completeness (i.e. no missing loops) and absence of non-standard residues. All the systems are summarized in Table 1.1.

Protein name	Apo PDB	Apo chain	Holo PDB	Holo chain
Interleukine-2	1M47	A	1PY2	C
Beta-lactamase	1JWP	A	1PZO	A
Polo-like kinase 1	1Q4K	A	3P37	C
Niemann-Pick C2 protein	1NEP	A	2HKA	C
Staphylococcal nuclease	1TQO	A	1TR5	A
Toluene-4-monooxygenase	2BF3	A	3DHH	E
Adipocit Lipid-Binding protein	1ALB	A	1LIC	A
Calcium-Bound Domain VI	1ALV	A	1NX3	A
Guanylate Kinase	1EX6	A	1GKY	A
Pyrophosphokinase	1HKA	A	3IP0	A
Heme oxygenase	1NI6	D	3HOK	B
Ribonuclease A	1RHB	A	2W5K	B
RhoA protein	1XCG	B	1OW3	B
Chymotrypsinogen A	2CGA	B	1AFQ	C
HSP90	2QFO	B	2WI7	A
LFA I domain	3F74	C	3BQM	C
NM23-H1	3L7U	C	2HVD	C
Adenylate kinase	4AKE	B	1ANK	B

Table 1.1: Systems used as a test set for the *CryptoScout* application. PDB entries for the *apo-holo* pairs are listed, as well as chains used in the MD simulations. Adapted from Publication 3.2.

Chapter 2

OBJECTIVES

The exhaustion of the so-called "low hanging fruits" in the pharma industry requires the development of novel methods to tackle the complex drug discovery cases and expanding the *druggable* protein space. One source of this innovation must come from computational efforts focused in understanding better the behavior of proteins and the nature of protein-ligand interactions. Therefore, the main objective of this doctorate has been to support the computerization of the drug discovery pipeline by developing innovative and state-of-the-art applications, some of which were presented to the scientific community via a web-based platform. Hence, the aims of this thesis can be formulated as below:

2.1 Computerize the drug discovery pipeline by means of MD simulations

In structure-based drug discovery (SBDD) the protein-ligand binding mode is the basic unit of knowledge. From the atomic description of the protein-ligand interactions we are able to build models that allow us to modify or extend a lead to enhance its affinity and produce a potential drug. Therefore, there is a need for the accurate detection and description of protein-ligand interactions.

These interactions, however, are difficult to study at an atomic scale using experimental techniques. In this sense, MD simulations can fulfill the gap. In particular, their ability to reconstruct full protein-ligand binding pathways and its combination with MSM in high-throughput molecular dynamics, makes them an interesting tool to the detect and predict protein-ligand binding *de novo*.

We have applied these principles in three different pioneer applications documented in Publications 3.2, 3.3 and 3.4. Specifically, we have studied the binding of benzene molecules to 18 protein systems as a proxy to detect cryptic cavities (Pub. 3.2). We have also performed the first 150-fragment screening against the chemokine CXCL12 fully driven by high-throughput MD simulations in combination with an MSM analysis framework (Pub. 3.3). Finally, we have studied the differential conformational plasticity of the μ-opioid receptor (MOR) bound to a classical opioid drug (morphine) and a G protein-biased agonist such as TRV-130 (Pub. 3.4).

2.2 Transfer know-how and applications to the web-based platform *PlayMolecule*

One of the greatest barriers for the adoption of a new technology, further than the computational or economical cost, is the access format to this technology. In the case of bioinformatics tools, packages such as HTMD have made a great progress in reducing the learning curve and entry barrier to technology such as the production and analysis of MD simulations. However, for these tools to work, medium to high informatics expertise is needed to configure a working environment and to use the existing API (Application Programming Interface) to fit one's custom needs.

A step further than a command-line and programming language-based environment, is the implementation of an intuitive graphic user interface (GUI) that allows the user to access the technology without any programming skills required. For instance, a web-based platform could disseminate the technology to a higher segment of users, with no pro-

gramming skills restrictions and by virtually solving any configurational and computing hardware issues by relying on *Cloud* computing. This is the philosophy that drove the creation of the *PlayMolecule* platform (*www.playmolecule.org*).

As part of this doctorate, we have fostered the access to state-of-the-art technology by developing and deploying four novel web applications and associated publications. First, *ProteinPrepare* (Pub. 3.1) allows to prepare a protein structure for MD simulations by optimizing the H-bond network and titrating the residues at a given pH. Second, *CryptoScout* (Pub. 3.2) allows to detect protein cryptic pockets by running simulations of the protein in presence of benzene, a probe that binds to hydrophobic regions of the protein including potential ligand binding sites and cryptic cavities. Third, we have developed *OPM-DB* (Pub. 3.6) a database of OPM membrane systems built and ready to run MD simulations. Finally, we have developed *DeepSite* (Pub. 6.1) a state-of-the-art ligand binding pocket detector based on convolutional neural networks.

Chapter 3

PUBLICATIONS

3.1 PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations

Martínez-Rosell G, Giorgino T, De Fabritiis G. PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations. J Chem Inf Model. 2017 Jul 24;57(7):1511–6. DOI: 10.1021/acs.jcim.7b00190

Summary
ProteinPrepare is a web application leveraging *PROPKA 3.1* and *PDB2PQR 2.1* software to prepare a protein extracted from the PDB database or uploaded by the user to be run in MD simulations. First, residues are titrated and the most likely protonation state is assessed. Second, the missing hydrogens are added to the structure based on the titration. Third, the H-bond network is optimized. The application allows the user to override the default protonation and inspect the predicted protonation in a user-friendly webGL protein viewer. The application is part of the *PlayMolecule* suit of apps and available at *www.playmolecule.org*.

3.2 PlayMolecule CryptoScout: predicting protein cryptic sites using mixed-solvent molecular simulations and mutual information

Gerard Martínez-Rosell and Gianni de Fabritiis. Submitted to *J. Chem. Theory Comput.*

Summary
CryptoScout is a novel method and web application leveraging MD simulations of protein solvated in a mixed-solvent of water and benzene to detect the presence of cryptic pockets (i.e. binding pockets invisible in available crystal structures) and structural insight of the opening mechanism. In the simulations, benzene binding to the protein surface is used as an indicator of binding pockets and cryptic sites. In order to detect them, first we calculate the occupancy of benzene and define binding hot spots. Additionally, we detect communities of residues with correlated fluctuation of solvent-accessible surface area (SASA) and calculate a likelihood of containing a cryptic pocket based in a pre-trained model. We test our protocol on 18 different cryptic pocket-containing systems, being the largest validation study of this kind. Finally, we present the method to the scientific community in a web application available at the *PlayMolecule* platform (*www.playmolecule.org/cryptoScout/*).

PlayMolecule CryptoScout: predicting protein cryptic sites using mixed-solvent molecular simulations and mutual information

Gerard Martinez-Rosell[†] and Gianni De Fabritiis[*,‡]

†*Computational Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain*

‡*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona 08010, Spain*

E-mail: gianni.defabritiis@upf.edu

Abstract

Cryptic pockets are protein cavities that remain hidden in resolved *apo* structures and generally require the presence of a co-crystallized ligand to become visible. Finding new cryptic pockets is crucial for structure-based drug discovery (SBDD) to expand the druggable space and to modulate protein activity via allosterism. We present here an application leveraging mixed-solvent molecular dynamics (MD) simulations using benzene as a hydrophobic probe to detect cryptic pockets and mutual information for the analysis. Our all-atom MD-based workflow was systematically tested on 18 different systems, being the largest validation study of this kind. CryptoScout first identifies benzene binding hot-spots, which correlate well with known cryptic pockets; second, it detects communities of residues with coordinated fluctuation of solvent-accessible surface area (SASA) and evaluates the likelihood of containing a cryptic pocket. CryptoScout also provides structures extracted from the MD simulations which may serve as starting structures for SBDD. The method is presented to the scientific community in a web application available at *www.playmolecule.org* using distributed and cloud computing infrastructures for the simulations.

Introduction

Knowing the tridimensional structure of a binding pocket is fundamental in structure-based drug discovery[1] (SBDD). While many cavities are already visible in crystal structures, for instance in the case of an orthosteric site bound to a natural ligand, other cavities remain closed and generally invisible in *apo* crystal structures.[2] These hidden cavities, known as cryptic sites,[3–6] may offer a number of advantages in comparison to conventional binding pockets. For instance, they are known to play a role in protein-protein interactions[7] or in allosteric modulation.[3,5] Therefore, detecting and understanding their dynamics can open the way for the development of novel highly selective drugs, i.e. involved in therapies based on allosteric modulation,[8] or targeting proteins previously considered undruggable.[9]

Cryptic sites are often discovered serendipitously[6] when they become co-crystallized with a ligand. Although is not clear if these pockets already open in solution by conformational selection, via an induced-fit mechanism or a mix of both,[10] the presence of a ligand seems to stabilize their opening as suggested by crys-

tal structures. Experimental techniques such as tethering associated with fragments[11–13] leverage from this phenomenon in order to reveal potential cryptic sites.

Different computational approaches to detect cryptic pockets or druggable hot-spots in general have been developed over the years.[14] In broad terms, these could be classified into: (a) molecular simulation-based, (b) bioinformatics-based[15,16] and (c) docking-based.[17,18] While the last two are computationally cheaper to perform, molecular simulations are the only ones able to resolve a molecular mechanism for pocket opening and the only one to provide structures eligible for SBDD. Several methods leveraging molecular simulations have been reported and they could be generally grouped into (a) protein-alone simulations and (b) mixed-solvent simulations. While the first type of methods assume that cryptic pockets can open in equilibrium in the absence of a ligand[3,19,20] (i.e. via conformational selection), the second one includes a ligand or co-solvent in the simulation whose binding to the protein is expected to reveal the presence of druggable pockets (i.e. via induced fit or a combination of both). Several successful applications of mixed-solvent methods have been published in the literature. A recent review[21] describes them extensively, giving special mention to MDmix,[22] SILCS[23] and MixMD.[24]

In this work, we present a new approach that leverages a well-established protocol of mixed-solvent MD simulations, similar in nature to the recently published work of Kimura et al.,[25] but with a novel analysis framework consisting on the identification of protein regions with correlated solvent exposed surface area (SASA) that may contain a cryptic pocket. The co-solvent used in our study is benzene, a simple yet generic and versatile hydrophobic probe that has already proved useful in detecting new cryptic sites.[26] Additionally, binding hot-spots of benzene on the protein surface are mapped and their correspondence with known cryptic sites is assessed. Our protocol has been tested on the largest dataset known to date for a mixed-solvent MD technique, consisting on 18 systems including 3 classic cases and other 15 systems extracted from the work of Cimermancic et al.,[15] where a collection of cryptic pocket-containing *apo-holo* protein pairs was reported.

In our study we aim to assess: (a) whether a mixed-solvent MD simulation protocol including benzene as a probe is able to identify cryptic pockets, (b) whether the novel analysis framework can improve the prediction performance and (c) whether benzene binding can actually sample cryptic pocket opening and conformations valid for SBDD.

The described protocol is wrapped up and made available at *www.playmolecule.org* to the scientific community in a web application that leverages latest web technologies to enable users to prospect cryptic pockets and druggable hot-spots on their protein of interest.

Results and discussion

Benzene as an ubiquitous bulky and unspecific hydrophobic probe

Druggable binding sites are known to have a higher average hydrophobicity than non-druggable binding sites,[27–30] although, at the same time, a recent study concludes that cryptic pockets are less hydrophobic than conventional binding pockets.[15] Furthermore, cryptic sites opening seems to be specially dependent on induced-fit mechanisms, as these pockets are usually discovered in *holo* structures. Therefore, the predominantly hydrophobic nature of these cryptic pockets and the fact that induced fit may play an important role in their opening suggest that a technique using unspecifically-binding hydrophobic probes should be able to correctly increase the sampling of pocket opening. A recently reported computational technique called SWISH,[20] precisely leverages from this fact in MD simulations by scaling down the non-bonded interactions of water molecules, which turn them into "ligand-like" molecules with higher affinity for apolar cavities such as cryptic pockets.

In a study by Wang et al.[31] all the ligands in the PDB[32] database were fragmented and a list of the most repeated chemical groups was

produced. Benzene turned out to be the most common of all by far. This aromatic molecule constitutes a 6-carbon ring, bulkier than linear carbon chains such as propane or butane. Its bulkiness, added to the fact that is an ubiquitous chemical group in the drug chemical space, makes it an excellent hydrophobic probe to induce a substantial opening of cryptic pockets while also sampling, when possible, pocket regions with relatively high affinity to aromatic rings, as long as these interactions are captured by the forcefield of use.

Our protocol

Several protocols involving mixed-solvent simulations have been proposed over the last years. One of the most notorious, SILCS, has been reported using high concentrations (around 1M) of benzene as a co-solvent and has shown relative success in the identification of binding hot-spots in various targets. However, the approach followed by the original SILCS algorithm included atom constraints and an interligand repulsion potential to avoid denaturation of the protein and probe aggregation, respectively. These protocol peculiarities, although were claimed to reduce the computational cost necessary to reach convergence, introduce clear protein flexibility restraints and potential artifacts that can hinder the results. In another study, the creators of mixMD discarded benzene as a good co-solvent arguing the existence of aggregation at the high concentrations (50% water/50% co-solvent) used in their repulsion-free protocol.[33]

In our study, we have chosen a restraint-free protocol and the use of lower benzene co-solvent concentrations, which allowed us to avoid aggregation without introducing repulsion potentials. In particular, we have tested our protocol in three different concentrations of benzene (0.2M, 0.1M and 0.05M) to assess the influence of the co-solvent concentration on the convergence of results and prediction power. At lower concentrations (0.2M) and using our benzene charmm-derived parameters, aggregation does not occur as shown by the radial distribution function (RDF) tending to 1 calculated from water+benzene simulations at 0.2M (S5.1).

Test set of labelled cryptic pocket-containing proteins

In order to test our protocol we have assembled a comprehensive set of 18 cryptic pocket-containing proteins (table 1) including classic systems such as IL-2,[3,34–38] PLK1[39,40] and TEM-1,[3,6,41] whose cryptic pockets are throughly studied and have been used as benchmarks in the past, and additional 15 systems extracted from a dataset recently published by Cimermancic et al..[15] The criteria we followed to select the later systems were: globularity, small to medium size (i.e. less than 250 residues), sequence-completeness (i.e. no missing loops) and absence of non-standard residues.

Summary of results

In the present work, we used mixed-solvent MD simulations in the presence of three concentrations of co-solvent benzene to unravel cryptic pockets on a set of 18 test proteins. For the analysis, we used two parallel methods for cryptic pocket discovery. The first method is the detection of co-solvent binding hot-spots, which are obtained following a protocol that has been widely used in similar applications. Results show that our algorithm detects the cryptic cavity within the first 3 hot-spots in 15 out of 18 proteins and with an average rank position of 2.3 in the 0.1M benzene condition. The second method we propose, designed to be used in parallel to the hot-spot-based one, leverages mutual information analysis to define communities of residues with coordinated SASA fluctuation. Cryptic pockets appear contained predominantly in one of these communities. A set of descriptors for each community is calculated (e.g. binding free energy score, SASA amplitude) and used to train a logistic regression model. Given a set of descriptors for a community, our regression model predicts the likelihood for a community of containing a cryptic pocket. In order to assess the predictive power

Table 1: Systems used as a test set for CryptoScout. PDB entries for the *apo-holo* pairs are listed, as well as chains used in the MD simulations, names of the ligands used to define the cryptic pockets and references from where the proteins were obtained. Additionally, RMSD of the backbone for each *apo-holo* pair is reported. The PDBID of the *apo* form is used throughout the present work as a unique identifier for each system.The asterisk (*) denotes that the ligand for 1Q4K system was a peptide chain instead of a small molecule.

	Ref	Protein name	Apo PDB	Apo chain	Holo PDB	Holo chain	Holo ligand	Apo-Holo RMSD
1M47	38	Interleukine-2	1M47	A	1PY2	C	FRH	1.07
1JWP	41	Beta-lactamase	1JWP	A	1PZO	A	CBT	0.91
1Q4K	40	Polo-like kinase 1	1Q4K	A	3P37	C	Chain F*	0.79
1NEP	15	Niemann-Pick C2 protein	1NEP	A	2HKA	C	C3S	1.11
1TQO	15	Staphylococcal nuclease	1TQO	A	1TR5	A	THP	0.64
2BF3	15	Toluene-4-monooxygenase	2BF3	A	3DHH	E	BML	0.69
1ALB	15	Adipocit Lipid-Binding protein	1ALB	A	1LIC	A	HDS	0.52
1ALV	15	Calcium-Bound Domain VI	1ALV	A	1NX3	A	ISA	0.63
1EX6	15	Guanylate Kinase	1EX6	A	1GKY	A	5GP	3.64
1HKA	15	Pyrophosphokinase	1HKA	A	3IP0	A	HHR	1.84
1NI6	15	Heme oxygenase	1NI6	D	3HOK	B	Q80	1.75
1RHB	15	Ribonuclease A	1RHB	A	2W5K	B	NDP	0.57
1XCG	15	RhoA protein	1XCG	B	1OW3	B	GDP	1.90
2CGA	15	Chymotrypsinogen A	2CGA	B	1AFQ	C	0FG	5.36
2QFO	15	HSP90	2QFO	B	2WI7	A	2KL	1.00
3F74	15	LFA I domain	3F74	C	3BQM	C	BQM	1.58
3L7U	15	NM23-H1	3L7U	C	2HVD	C	ADP	0.73
4AKE	15	Adenylate kinase	4AKE	B	1ANK	B	ANP	6.91

of our model, we use a leave-one-out cross validation scheme which yields an average AUC of 0.86 in the 0.2M benzene condition.

Additionally, we study the dynamics of the cryptic pockets in presence and absence of the co-solvent benzene to assess whether simulations starting from the *apo* conformation ever reach the *holo* conformation. Results based on a PCA dimensionality reduction show that, in some systems, the binding of benzene seems to trigger conformational changes towards the *holo* conformation. Additionally, we use AutoDock VINA[42] to assess whether docking is able to reconstruct the *holo* ligand binding pose using structures extracted from the MD simulations. We are able to reconstruct the pose of 5 out 17 ligands within 3 RMSD when using 10 representative conformers in comparison to recovering only 1 out 17 when using the *apo* structure as docking structure.

Mutual information analysis identifies communities of residues with correlated SASA fluctuations

Most of mixed-solvent simulation approaches for druggability assessment focus on the detection of binding hot-spots by mapping the co-solvent affinity on the protein surface. In the present study we approached the question slightly differently: additionally to binding hot-spots, can we determine communities of residues that fluctuate in a coordinated way and, if so, can we detect and rank cryptic pockets enclosed in these communities? To answer this question we used a mutual information framework, usually employed in allosterism studies, with few modifications to account only for short-range residue-residue interactions. We measured the SASA for every residue along our simulations and calculated how its fluctuation was correlated with the residues around them. This allowed us to define "patches" of protein with correlated SASA. See Methods section for further details.

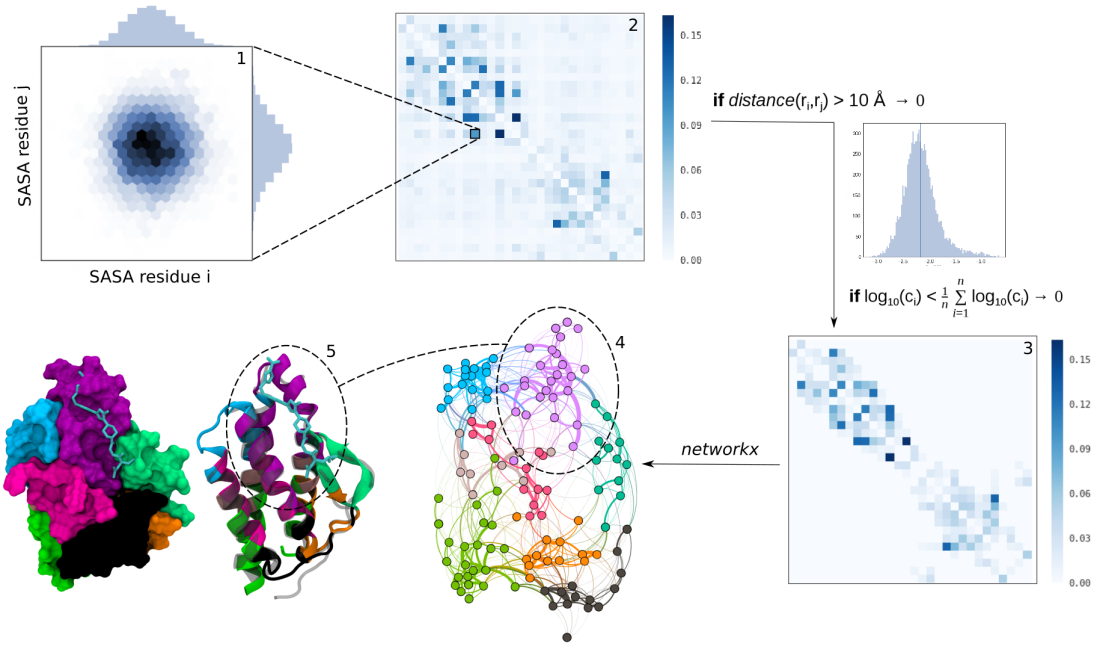


Figure 1: Workflow to define communities of residues with co-variant fluctuation of SASA. **1**. For each protein residue pair, a SASA co-variation is calculated by building a contingency table (a 2D histogram with 20 equally spaced bins). **2**. A co-variance matrix for all residue pairs is calculated. **3**. The co-variance matrix is filtered for (a) values lower than the mean of the logarithm of the correlation in order to keep only strong co-variants and (b) all those co-variances between residues further than 10 Å to keep only short-range co-variations. **4**. Communities of residues are clustered together using *networkx*.[43] **5**. Note the similarity between the residue communities in the SASA short-range co-variation graph and the 3D structural position of those residues.

The majority of residues involved in a cryptic pocket are contained within one or two CryptoScout communities

In order to assess if our framework was able to identify and differentiate regions with cryptic pockets from non-cryptic pocket regions, we created an automated protocol to calculate the overlap between CryptoScout-defined residue communities and the experimental cryptic pocket residues (i.e. residues in contact with the ligand in the experimental *holo* form). To do so, first we assigned each residue of the *apo* structure to a community of residues using our mutual information protocol based on SASA (figure 1.4). Then, we aligned the *apo* and *holo* structures and calculated the contacts within 3 Å between the *apo* structure and the aligned ligand present in the *holo* form (figure 1.5). We added up the number of "real" contacts per CryptoScout community. Finally, we labelled those communities with at least 30% of the total contacts as "positive" (i.e. they contain a cryptic pocket) and labelled all others as "negative". The results show that, with a degree of variability, our model assigns the majority of the *apo*-ligand contacts to one, sometimes two, main communities (column *PRCP* in table 2 which stands for percentage of residues in positive community). These results suggest that different regions of the protein fluctuate or "breath" in a coordinated way, including the cryptic pockets. Furthermore, while in most cases one single community includes most of the residues of the cryptic pocket, in some cases two different communities may independently fluctuate and contribute to pocket opening.

A logistic regression model correctly identifies cryptic pockets in leave-one-out cross-validation using the community-based CryptoScout score

A number of features was calculated for each community (see Methods for further details). Then, a logistic regression model was trained using those features to categorize "positive" communities (i.e. containing cryptic pockets) and "negative" communities. Stepwise regression analysis was performed to select the most meaningful and correlated features and discard the noisy ones from our model. We used the Akaike information criterion[44] implemented in R.[45] For each benzene concentration condition (0.2M, 0.1M and 0.05M) different metrics turned out significant or explicative. To solve this issue we ended up picking a consensus set of 4 descriptors: SASA max, SASA mean, FEG relative mean score and FEG amplitude. The features selected, p-values and logistic model correlation for 0.1M benzene condition can be found in table 3.

Using these features, we followed a leave-one-out cross-validation scheme to calculate an average AUC (area under the curve). This scheme consists in using as training dataset n-1 proteins and predicting on the test set, consisting of one protein (leave-one-out). The output of our logistic regression model is, for each test community, a probability of containing a cryptic pocket. Therefore, for each test protein and for each community in that protein defined by SASA mutual information, we obtained the probability of containing a cryptic pocket. Based on our predictions we obtained an AUC which reflects how correctly our prediction score could separate the true positive communities from the true negative communities. The average AUC is 0.86, 0.80 and 0.74 for 0.2M, 0.1M and 0.05M conditions respectively (table 2). The standard deviation was 0.18, 0.19 and 0.22, respectively, which reflects the system-specific variability, i.e. our model ranked very well the positive communities of some proteins (e.g. AUC=1) but worked considerably worse on others (e.g., AUC=0.4).

Benzene binding affinity (FEG score) is the most relevant feature in the prediction of communities containing cryptic sites

FEG score is the most significant descriptor in our regression model. The higher the relative

Table 2: Community-based and hot-spot-based score prediction results for each of the three CryptoScout benzene conditions (0.2M, 0.1M, 0.05M) and results for fpocket[46] and DeepSite[47] in pocket detection. **AUC** stands for *area under the curve* and measures how well the community-based score separates the true positives from the true negatives in a leave-one-out cross-validation scheme. **Hotspot** represents the rank position of the best benzene binding hot-spot found within 5 Å of the ligand in the *holo* form. The first value represents the rank and the second represents the total number of hot-spots found. *NA* means no hot-spot was found within 5 Å. **PRPC** stands for *percentage of residues in positive community* which reflects the ability of the mutual information framework to include cryptic pockets inside one or two main communities. The number in parentheses in the **average** row represents the number of *NA* predictions. The asterisk (*) in the DeepSite column denotes that the particular *holo* PDB was present in the DeepSite training set and therefore the predictive model "had already seen" where the pocket should be located.

	0.2M			**0.1M**			**0.05M**				
	AUC	**Hotspot**	**PRPC**	**AUC**	**Hotspot**	**PRPC**	**AUC**	**Hotspot**	**PRPC**	**fpocket**	**DeepSite**
1Q4K	1.0	1/13	0.562	0.89	1/20	0.809	0.79	1/23	0.745	1/30	1/20
1ALB	1.0	2/12	0.609	1.0	1/13	0.652	1.0	1/25	0.652	1/25	1/20
3F74	1.0	1/12	0.959	0.86	1/18	0.959	0.86	3/21	0.581	7/30	NA/6*
2QFO	1.0	2/19	0.348	0.67	2/22	0.348	0.17	15/24	0.478	1/22	1/20
1JWP	1.0	1/22	0.929	1.0	1/22	0.955	1.0	2/37	0.536	1/28	NA/20
1M47	1.0	8/14	0.768	0.43	7/12	0.821	0.67	3/17	0.982	NA/41	NA/1
1RHB	1.0	2/8	0.75	0.4	1/10	0.688	0.75	3/15	1.0	4/29	NA/20
1NI6	1.0	2/19	0.867	0.88	1/22	0.583	0.75	1/32	0.617	1/35	1/20
1NEP	1.0	1/9	0.851	1.0	1/13	0.836	0.8	1/18	0.94	NA/27	1/1*
2BF3	1.0	1/7	0.923	1.0	1/9	0.923	1.0	1/15	0.923	5/28	NA/2
1EX6	1.0	1/12	0.938	0.86	3/17	0.938	0.75	5/22	0.938	1/33	11/17*
1HKA	0.75	3/13	0.769	0.86	1/18	1.0	0.5	7/15	0.769	1/29	1/20
3L7U	0.57	NA/13	0.917	0.75	NA/13	0.917	0.9	12/23	0.917	5/37	2/20*
2CGA	0.64	5/13	1.0	1.0	2/22	1.0	0.89	3/30	0.75	3/28	NA/20
1TQO	0.67	2/8	1.0	0.58	3/11	1.0	0.7	6/17	1.0	3/30	1/11
4AKE	0.5	3/19	0.896	0.86	2/21	0.979	0.83	1/26	0.875	1/35	4/20*
1XCG	0.57	12/13	0.898	0.56	11/16	0.51	0.64	4/28	0.776	2/31	NA/20*
1ALV	0.75	14/19	0.75	0.88	1/22	0.75	0.25	1/23	0.75	4/30	1/20*
average	**0.86**	3.59(1)	0.82	0.80	**2.35(1)**	0.81	0.74	3.89(0)	0.79	2.56(2)	2.27(7)

mean and amplitude are for a community, the more likely the community contains a cryptic pocket. This point is supported by the positive signs of the FEG score relative mean and FEG score amplitude coefficients (third column of table 3). Although the deletion of FEG score relative mean and amplitude from the model separately is responsible for a small loss of AUC mean (fourth column of table 3), the reason for these results resides in the fact that both FEG score measurements share common and redundant information. As such, models using these descriptors as a single feature yield AUC means of 0.79 and 0.77 for the 0.1M benzene condition (table 3). On the other hand, SASA measurements perform poorly in single feature models but seem to be significant and may add predictive power when combined with FEG score descriptors. Interestingly, SASA mean and max are inversely correlated with cryptic pocket-containing communities. This may be explained by the fact that cryptic pockets are usually non-terminal regions and remain partially closed for long time in the simulation. This fact is reflected in the density plots in figure S6.

Benzene binding hot-spots can be recovered and correlate well with the known cryptic pockets

In order to detect binding hot-spots, we have followed a similar approach as previous studies by generating a free energy grid (FEG) and clustering together free energy minima to define what we call "binding hot-spots". These hot-spots were compared to the cryptic cavity by aligning *apo* and *holo* structures and defining "cryptic hot-spots" as those hot-spots closer than 5 Angstrom to the aligned *holo* ligand. Hot-spots were ordered in increasing free energy order and the position where the best cryptic hot-spot was ranked was annotated and shown in table 2. CryptoScout was able to rank cryptic hot-spots with an average rank position of 2.3 (table 2, 0.1M condition). Specifically, it was able to identify the cryptic hot-spot in the first 3 positions in 15 out of 18 systems. Some examples of detected hot-spots can be found in figure 2.

Autodock VINA applied to structures extracted from MD simulations are able to reconstruct original *holo* ligand binding poses

We tried to recover the *holo* ligand binding pose using a AutoDock VINA applied to conformations extracted from the simulations, considering that we started the simulations from the *apo* structure, which had the cryptic pocket closed. In order to sample different conformations, we used Kmeans clustering algorithm on SASA fluctuation data for the community of residues containing the cryptic pocket to obtain up to 10 clusters with variable degree of SASA (i.e. pocket opening) and sampled one representative per cluster. Autodock VINA applied to these conformations recovered 10 out of 17 *holo* ligand binding poses within 5Å in RMSD and 5 out of 17 within 3Å RMSD (S4). For comparison, docking of the ligands onto the *apo* form recovered 5 out 17 within 5Å RMSD and only 1 out of 17 within 3Å RMSD (S4). Note that the reported poses were specifically selected as those with minimum RMSD to the original *holo* pose. The number of binding poses generated by AutoDock VINA were a maximum of 10 per each of the 10 automatically selected conformations. System 1Q4K was excluded from the docking due to the complexity of the peptide ligand. The majority of systems whose *holo* binding pose could not be recovered include the largest ligands, for which docking becomes extremely difficult due to the ligand rotameric variability and the challenge of sampling the exact protein conformation that allows a successful docking. In supplementary figure S3 we show the strong correlation between number of heavy atoms per ligand and RMSD of the docked pose closest to the *holo* pose. Two examples of docking can be found in figure 3 where major side-chain re-arrangements are well captured.

Table 3: Consensus of metrics used for the training of a logistic regression predictor model. p-values as reported by the Akaike Information Criterion (AIC) for each feature, Coefficients for the particular descriptor in the regression model, mean AUC loss on removing the descriptor from the model and mean AUC on building a regression model with the particular single feature.

Descriptor	p-value	Coefficient	AUC loss on removal	AUC single descriptor model
FEG score relative mean	0.0151*	0.7639	-0.0061	0.7962
FEG score amplitude	0.1973	0.0034	-0.0572	0.7709
SASA mean	0.0146*	-0.9596	-0.0175	0.4694
SASA max	0.9861	-1.0718	-0.0036	0.4896

CryptoScout performs better than fpocket and DeepSite in detecting cryptic pockets

In order to compare CryptoScout performance with other algorithms of non-cryptic pocket detection, we used fpocket[46] and DeepSite[47] to detect binding pocket centers on the apo structure (table 2). Results show that fpocket is reasonably good at detecting cryptic pockets using a cutoff of 5Å distance between predicted center and *holo* ligand, with only 2 proteins for which no center was predicted within that cutoff. On the other hand, DeepSite seems to find cryptic pockets on some proteins but is unable to detect them in 7 out of 18 systems. On average, CryptoScout was the best performer with an mean hot-spot rank of 2.35 and only 1 protein for which no hot-spot was detected within 5Å of the *holo* ligand. Although fpocket seems to be a very good option to detect cryptic sites at a extremely lower computational cost compared to CryptoScout, there's two main considerations that justify the extra computational cost of CryptoScout. The first is that fpocket algorithm works by detecting concave surfaces and most of the cryptic sites used in the current study are pre-formed in some way. Therefore, fpocket would probably fail to detect pockets in really complex cases. The second is that, although some part of the cryptic pockets may be pre-formed (for instance, the protein may present an internal cavity inaccessible from the exterior such as in 1ALB system), fpocket sheds no light on the pocket opening mechanism or how a ligand can access the pre-formed part of the pocket. On the other hand, CryptoScout, as it requires benzene to bind in order to detect the pocket, provides explicit hints of the pocket opening mechanism at an atomic resolution and provides additional information about the pocket conformation. For example, figure 3.A shows the case of 1ALB, where benzene triggers a sidechain rotation that reaches a pose similar to the *holo* conformation and that helps to expose the internal cavity.

Principal component analysis shows that some simulations starting from the *apo* conformation visit the *holo* conformation

Detecting cryptic cavities constitutes a great challenge by itself. However, in order to apply SBDD technologies efficiently, scientists need a structure of the open cryptic pocket with a spatial configuration of the lateral chains that allows the correct inference of the ligand binding pose.

In order to assess whether our simulations sampled valid configurations for SBDD, i.e. the *holo* protein conformation, we decomposed the geometry of the cryptic pocket into 2 principal components using a Principal Component Analysis (PCA). Cryptic residues were defined as those with an atom within 4Å from the ligand present in the *holo* form as in Kimura et al..[25] To do so, first we calculated an N-dimensional vector by applying PCA on the minimum distance between each N pair of cryptic residues for each frame of our simulations in presence of benzene. Later, we projected each frame of the simulations with or without benzene into this N-dimensional vector and selected

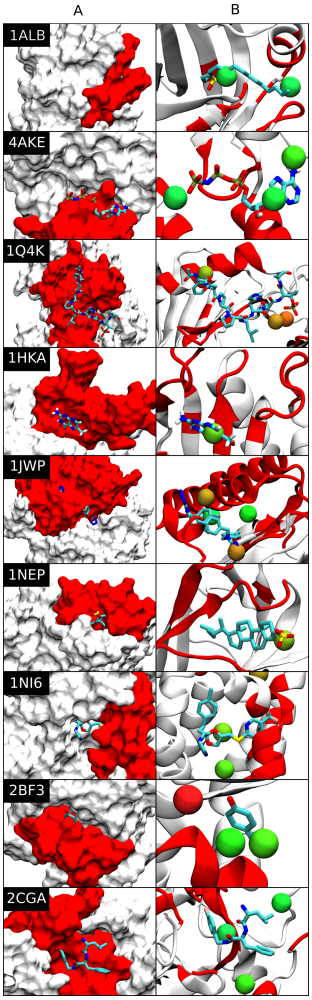


Figure 2: Benzene binding hot-spots found for each of 9 representative systems simulated in the 0.1M benzene condition. Column **A** shows the surface of the *apo* structure and the ligand present in the *holo* pose after aligning *apo* and *holo* structures by the backbone. The community of residues labelled as “positive” (i.e. contains a cryptic pocket) is colored in red. Note how most of the ligands present clashes with the *apo* surface due to the closure of the cryptic pocket. Column **B** shows the *apo* backbone, aligned with the *holo* ligand and the hot-spots detected by CryptoScout depicted as beads colored from red to green, being green the lowest free energy. Residues part of a community labelled as “positive” are colored in red.

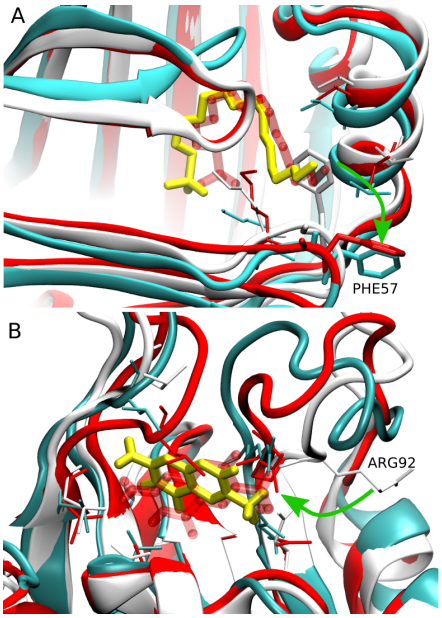


Figure 3: Examples of docking performed against 10 conformers extracted from the MD simulations. Docking for 1ALV system depicted on the top, and 1HKA on the bottom. Docked ligand colored in yellow licorice. Original ligand in the *holo* structure represented as transparent red licorice. *Holo* conformation backbone in red cartoon, *apo* conformation backbone in white cartoon and backbone conformation extracted from the simulations in cyan. Important residues are displayed in licorice style. Green arrows point out side-chain rotations from the *apo* position to an *holo*-like position adopted during the MD simulation in presence of benzene.

the first two principal components to produce a 2-D histogram plot (figure 4). This dimensionality reduction allowed us to assess visually whether our simulations reached and sampled extensively the *holo* lateral chain configuration (red crosses in fig. 4) starting from the *apo* conformation (green crosses in fig. 4) in presence or absence of benzene. The presence of benzene seems to shift cryptic pocket configuration to the *holo* form in some cases (e.g. as seen in 1HKA, 1NI6 and 1NEP to certain extent), although in some other cases there's little or no difference between benzene and water conditions (e.g. 1TQO) and sometimes water alone shifts to *holo* even better than with benzene (e.g. 1ALV).

Is also interesting to notice two points: first, the post-equilibration conformation of the protein (purple crosses in figure 4), this is the conformation of the protein after performing the equilibration of the *apo* structure, is different from the *apo* one and its effect on the exploration of the landscape can be very determinant. For instance, in 2CGA PCA analysis (figure 5) we can see that at 0.1M and 0.05M benzene concentrations the post-equilibration pose (purple cross) is close to a minimum below the *holo* pose (red cross) while in 0.2M and water conditions the simulations are spawned from a post-equilibration conformation located in a deep well to the right of the *holo* conformation, from which they were unable to exit.

Second, the presence of benzene seems to modify the configuration landscape. For instance, in the 1M47 system (figure 4), although in both benzene and water conditions the post-equilibration structure was very similar, the landscape in the water condition is a double well while in benzene condition is a single well.

Overall, seems that benzene modifies the configuration landscape, sometimes shifting it to the *holo* pose, but the effect of the post-equilibration pose on the exploration of conformational space has also to be taken into consideration. Note that a possible improvement of the current protocol would be to run two or more equilibrations in order to start simulations from a diverse conformational space which would lead to a better sampling of the landscape. Note also that reaching convergence was not the intention of the current study.

One of the aims of this work was to assess the effect of the co-solvent concentration in cryptic pocket discovery using mixed-solvent MD with our community-based or hot-spot-based scoring system. To our surprise, a second factor was introduced in the analysis, which is the post-equilibration conformation (i.e. the conformation achieved by the protein after the equilibration run). This last confounding factor may influence the extent in which the conformational space is explored, especially in such short amount of simulated time, and therefore can influence the results by providing conformations with a differential affinity for benzene, whose binding is crucial for cryptic pocket detection in our protocol. Unfortunately, the post-equilibration conformation factor is hardly separable from the concentration factor. However, we believe broad conclusions can still be drawn. For instance, seems that a higher concentration of benzene (0.2M) enhances the community-based score (+0.06 AUC mean in respect to 0.1M condition) and a lower concentration (0.1M) enhances the hot-spot-based scoring system (it improves more than 1 rank position in respect to 0.2M). One possible explanation is the fact that the community-based score depends on the benzene binding free energy to the whole community surface, where transient aggregates of benzene can be effectively detected as molecules stack up, while hot-spot-based score requires a more fine protein-ligand interaction, ideally unperturbed by competitive interactions of other benzene molecules. A low concentration like 0.05M, although still detects most of the cryptic sites, is probably very sensitive to lack of sampling as convergence time is highly dependent on the number of molecules available to interact with the protein.

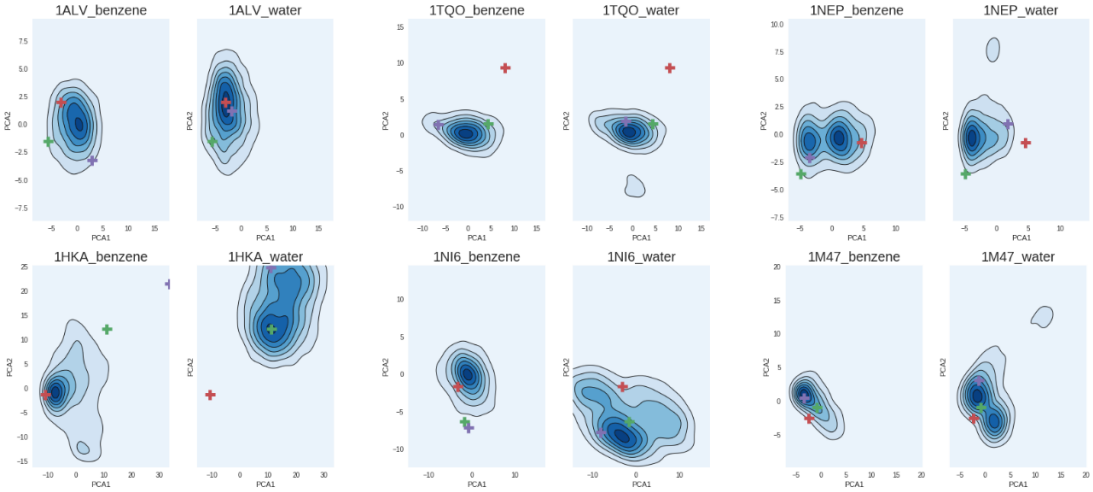


Figure 4: 2D histogram plot calculated from the first 2 PCA components of the minimum inter-residue distance between all cryptic pocket residues pairs for 6 representative systems in the 0.1M benzene and water-only conditions. Green crosses represent the *apo* conformation, red color represents the *holo* conformation and purple crosses represent post-equilibration conformation in the PCA space.

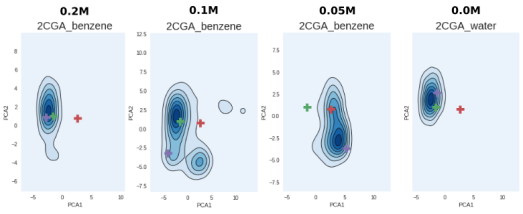


Figure 5: 2D histogram plot calculated from the first 2 PCA components in all benzene concentrations and water for the 2CGA system.

Total simulation time necessary for model convergence can be down-scaled in production

We analyzed the effect of sampling time over the hot-spot-based and community-based scoring methods for decreasing amounts of data by bootstrapping 5 times from 100% of data to only 30% of data in 10% data jumps, this is from 800ns total simulation time to only 240ns. Results in figure 6 show that while hot-spot detection convergence is quite dependent on the amount of data, the community-based score is much more independent to the point that 30% of the data is able to give equal or even better classification of communities containing a cryptic pocket than the full data (notice, however, the higher standard deviation). Based on this fact and the need to find a compromise between results and computational cost, we have set the amount of simulation time to half the maximum (400ns) in the production CryptoScout web application.

CryptoScout limitations

Although CryptoScout has proved effective in detecting and ranking the cryptic pockets for most of the systems, both community-based and hot-spot-based scoring methods failed on few systems. In this section we will mention some of the possible underlying causes.

First, benzene may not bind the cryptic cavity. It is possible that the cryptic cavity has low affinity for benzene, such as in the case of hydrophilic cryptic pockets. Other probe or co-solvent molecules could be added to expand the chemical space that could potentially bind and unravel cryptic pockets. The preference of a cryptic pocket for specific fragments has been

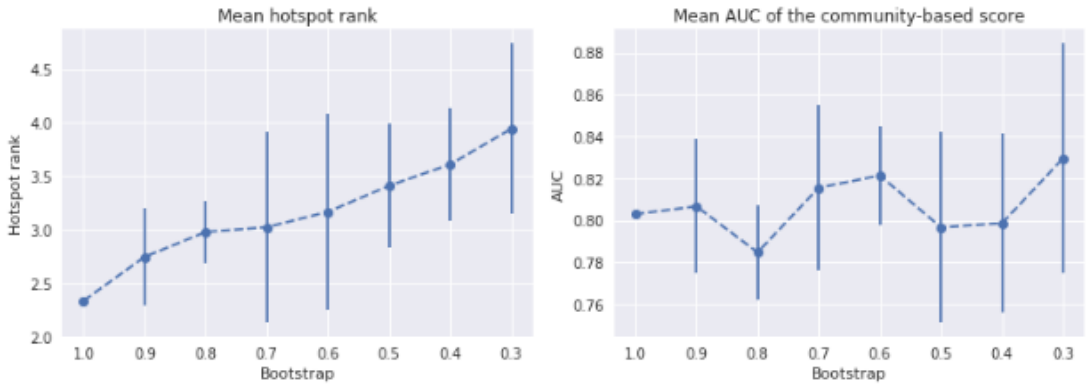


Figure 6: Prediction performance on data bootstraps using the community-based and hot-spot-based scores.

recently reported by Kimura et al.[25] where they show that hydrophobic and hydrophilic fragments tend to bind and reveal cryptic pockets in a differential way.

Second, convergence may not have been reached within the simulation time, 800 nanoseconds in our case. The timescales for cryptic pocket opening or the necessary conformational rearrangements for benzene binding may be orders of magnitude slower than the simulated time. One indication supporting this point can be found when measuring the RMSD between the *apo* and the *holo* forms for each system, which is an indicator of the amount of atomic rearrangements that the *apo* structure must undergo in order to reach the *holo* conformation, in some cases exceeding 5Å (table 1). Another indication is found in the inconsistencies between the PCA analysis of the same system in different benzene concentrations, whose cause could be possibly due to lack of convergence rather than a direct effect of the benzene concentration on the spatial rearrangement of the cryptic cavity residues. An example of the latter case is the 2CGA system, where only in the 0.05M condition the PCA landscape shifted towards the holo configuration (S2.1).

Finally, causes may also include bad detection of regions containing cryptic pockets (e.g. low *PRPC* value due to lack of SASA short-range covariation) such as in the case of system 1XCG. Another possible cause is the existence of alternative strong benzene binding hot-spots, some of which could be unknown cryptic pockets.

PlayMolecule CryptoScout web application details

The CryptoScout web application has been made available through the PlayMolecule web platform (see figure 7) and uses the GPUGRID volunteer infrastructure for the MD simulation calculations. The access to the app is open to the scientific community but a request needs to be made first by filling a quick form with identification details and purpose of the usage. This measure allows us to moderate the access to these intensive resources and provide a good service. See S7 for more details.

Conclusions

In this study we propose a new and completely automated approach using mixed-solvent MD simulations to discover cryptic pockets combining two predictor metrics consisting in (a) the assessment of benzene binding hot-spots and (b) the identification of communities of residues with correlated SASA associated with a probability of containing a cryptic pocket. Both prediction methods have shown success in detecting cryptic cavities in our test dataset with an average hot-spot rank position of 2.3 (0.1M)

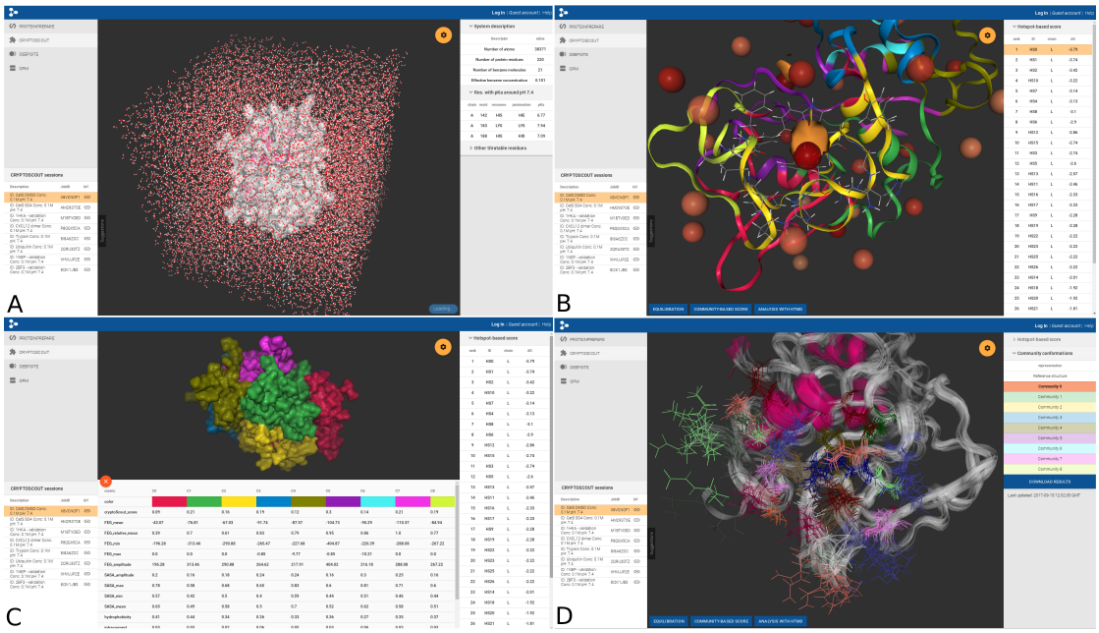


Figure 7: Screenshots of PlayMolecule CryptoScout web application. **A**. Report of the system after building. **B**. Visualization of a benzene binding hot-spot. **C**. Metrics and CryptoScout score for each community of residues detected, identified by a distinct color. **D**. Visualization of 10 representative frames per community extracted from the MD simulations colored by residue name.

and an average AUC of 0.86 (0.2M) for the hotspot-based and the community-based scores, respectively.

Furthermore, we assessed whether simulations starting from the *apo* structure visited the *holo* conformation in presence or absence of benzene. Our results suggest that in some cases benzene not only binds and triggers the opening of the cavity but also induces the rearrangement of the pocket residues towards the *holo* conformation. This is further confirmed using AutoDock VINA by reproducing 10 out of 18 binding poses within 5Å RMSD.

Knowing the exact *holo* conformation is crucial for all SBDD endeavors, such as in virtual screening or ligand optimization. While detecting a cryptic cavity is of a great interest, further development must be undergone to sample and correctly identify cavity conformations compatible with SBDD. Assuming the cryptic site had been correctly determined, one could leverage an adaptive sampling scheme based on SASA per residue or inter-residue distance to sample exhaustively all possible conformational states and pinpointing stable conformations and converged kinetics by using, for instance, Markov State Models. These models could be further employed to prioritize a conformation or ensemble of conformations for SBDD activities.

Methods

System building and simulation

Benzene parameters were calculated using the *Parameterize* module included in HTMD.[48] *Parameterize* performs charge fitting and rotamer scans using quantum mechanics calculations with PSI4[49] to fit parameters and optimize the topology.

Proteins were obtained from the PDB[32] database, with entry code specified in table 1, and simulated from the *apo* form. In case of 1M47, missing loops were modelled by Modeller[50] using the LoopModeller module included in HTMD. Protein protonation and hydrogen-bond network optimization was made using the ProteinPrepare module included in HTMD. Systems were built using HTMD, including the protein in the center of the water box and placing benzene molecules around the center to reach the target benzene concentration. Water padding was set to distance between the center and furthest atom of the protein plus a 6 Å water margin. In the case the 0.2M condition, padding was reduced in some cases to ensure a ratio protein residue/number of ligands higher than 4.5, which limited the amount of benzene molecules added to the system and prevented the formation of potential aggregates. Sodium and chloride ions were added to neutralize the system.

One equilibration was performed per system and 20 production runs of 40ns were performed (total of 0.8 μs per system). Equilibration protocol consisted of 500 system minimization steps, 500 NVT steps (4fs each) and the rest of the simulation for 40ns in NPT ensemble; restraints of 1 Kcal/mol were applied to heavy atoms and 0.1 kcal/mol to non-heavy atoms and were progressively switch off from the beginning of the simulation until half of the simulation, where the system was set restraint-free. The force-field used was charmm22*.[51,52] The simulations were run in our local cluster equipped with 16 GPUs using the simulation software ACEMD.[53]

Free Energy Grid (FEG)

In order to detect co-solvent binding we calculated a free energy grid (FEG) and clustered energy minima to define binding hotspots. Specifically, first, we computed the benzene occupancy by calculating a 3D histogram with 1Å cubic bins of the location of C1 of the benzene probes along our MD simulations. After, we transformed these occupancies into probabilities by dividing by the number of MD trajectory frames. Then, using the Boltzmann equation (eq. 1), we transformed the probabilities into free energies:

$$\Delta G = -K_{\mathrm{B}} T \ln\left(\frac{N}{N_0}\right), \qquad (1)$$

where T is the simulation temperature (300K), K_B is the Boltzmann constant in kcal/mol·K,

N is the co-solvent occupancy probability and N_0 is the standard occupancy in equilibrium, calculated as:

$$N_0 = \frac{V_B N_A [C]}{n_B} \quad (2)$$

where V_B is the total volume of our system box (in liter units); N_A is Avogadro's number ($6.022140857 \cdot 10^{23}$), $[C]$ is the co-solvent concentration in molarity (0.2M, 0.1M or 0.05M); and n_B is the number of boxes in our grid. In practice, the numerator calculates the number of ligands for a given concentration and the denominator divides the occupancy probability among the number of boxes present in the grid; this assumes that ligands do not aggregate and spatial distribution of benzene is similar to a noble gas. These last two assumptions are supported by the pair correlation function (also known as radial distribution function; RDF) converging to one in water+benzene simulations at 0.2M concentration (S5.1), which denotes the lack of benzene aggregation. RDF was also calculated for 0.1M condition (S5.2) and, for comparison, also for 1M condition (S5.3), in which benzene clearly aggregates in correspondence with results obtained by Lexa et al..[54] RDF was calculated using VMD.[55]

Hot-spot detection

Once we had a free energy grid, we proceeded to filter out all those free energies higher than a -1.75 kcal/mol cutoff, which is a bit lower than other reported cutoffs (-1 kcal/mol as in Bakan et al.[56] and -1.5 kcal/mol as in Kimura et al.[25]). This cutoff allowed us to reduce the number of hot-spots found and is probably force-field dependent as suggested in Kimura et al..[25] Then, we proceeded to find the minima in our grid and cluster them together. The algorithm we used to cluster the minima consists in joining clusters closer than 8Å and discarding those closer than 2.7Å (benzene diameter) starting from the lowest minima and moving to the next minima in growing order. The cutoff used to cluster is a bit higher than the 6.2Å cutoff reported by reference 56 but allowed us to generate fewer clusters and generally enhance the performance. Finally, for each cluster, instead of adding up the minima free energies as in reference 25, we followed a different approach, which is adding up the probabilities of both minima and then calculating the free energy of the joint probability by using the Boltzmann equation. This fundamental difference implicates that, for instance, two minima of -3 kcal/mol add up to a joint free energy of -3.4 kcal/mol instead of -6 kcal/mol. 3D structure plots were produced using VMD.[55]

Community definition by mutual information (MI) analysis on residue SASA

A matrix with SASA co-variance for every pair of protein residues was calculated. To do so we built a histogram based on 20 equally spaced bins along the SASA fluctuation per residue in our simulations and proceeded to calculate the correlation (figure 1.2) from the corresponding residue-residue pair contingency table (figure 1.1). We then applied two filters, setting to 0 the correlations that fulfill one of these conditions (figure 1.3): (a) the correlation weight is smaller than the mean of the weight logarithms, this way we only retain the most meaningful correlations; (b) the distance between two residues is bigger than 10 Angstroms, this way we remove long-range correlations and focus in the local correlations. Finally, we cluster the residues into hubs or communities of residues (figure 1.4 and 1.5) using the python module *networkx*.[43]

Free Energy Grid (FEG) score

For each carbon alpha of a community of residues, we created a box of 5Å padding centered on the CA (containing 5·5·5 $1Å^3$ subboxes), and calculated the average free energy in that box obtaining a measure in cal/A^3. The FEG score for a certain community was obtained by performing the mean of FEG score of all carbons alpha.

Other metrics and regression model

Maximum, minimum, mean, standard deviation and amplitude was obtained grouping all the SASA per residue in a community. Mutual information intra-correlation was calculated as the average of the MI network weights within a community. The logistic regression model was calculated using scikit-learn.[57]

Simulation bootstrap

In order to check the effect of decreasing the simulation time in the results convergence and robustness, we bootstrapped the data used for the analysis 5 times with decreasing amount of data, from 1 to 0.3 (over 1) with steps of 0.1 decrease. In practice this means that from a total of 800ns per system we decreased it to 240ns in 80ns steps and used our hot-spot-based and community-based scores to identify the cryptic pockets. For cryptic sites that were not detected during the analysis, instead of labelling them as "NA" (not available) we set the rank position to the total number of hot-spots found in order to penalize the average.

CryptoScout web application

CryptoScout application is part of the PlayMolecule bundle of web apps. It leverages webGL-powered protein viewer NGL,[58] as well as AngularJS and Angular Material for the client-side and Flask for the server-side. The computing infrastructure is currently GPUGRID.[59]

Acknowledgement GDF aknowledges support from MINECO (BIO2014-53095-P) and FEDER.

References

(1) Nisius, B.; Sha, F.; Gohlke, H. *J. Biotechnol.* **2012**, *159*, 123–134.

(2) Stank, A.; Kokh, D. B.; Fuller, J. C.; Wade, R. C. *Acc. Chem. Res.* **2016**, *49*, 809–815.

(3) Bowman, G. R.; Geissler, P. L. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 11681–11686.

(4) Diskin, R.; Engelberg, D.; Livnah, O. *J. Mol. Biol.* **2008**, *375*, 70–79.

(5) Durrant, J. D.; McCammon, J. A. *BMC Biol.* **2011**, *9*, 71.

(6) Horn, J. R.; Shoichet, B. K. *J. Mol. Biol.* **2004**, *336*, 1283–1291.

(7) Wells, J. A.; McClendon, C. L. *Nature* **2007**, *450*, 1001–1009.

(8) Hardy, J. A.; Wells, J. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 706–715.

(9) Dang, C. V.; Reddy, E. P.; Shokat, K. M.; Soucek, L. *Nat. Rev. Cancer* **2017**, *17*, 502–508.

(10) Csermely, P.; Palotai, R.; Nussinov, R. *Trends Biochem. Sci.* **2010**, *35*, 539–546.

(11) Sadowsky, J. D.; Burlingame, M. A.; Wolan, D. W.; McClendon, C. L.; Jacobson, M. P.; Wells, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6056–6061.

(12) Ostrem, J. M.; Peters, U.; Sos, M. L.; Wells, J. A.; Shokat, K. M. *Nature* **2013**, *503*, 548–551.

(13) Erlanson, D. A.; Braisted, A. C.; Raphael, D. R.; Randal, M.; Stroud, R. M.; Gordon, E. M.; Wells, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 9367–9372.

(14) Hall, D. R.; Enyedy, I. J. *Future Med. Chem.* **2015**, *7*, 337–353.

(15) Cimermancic, P.; Weinkam, P.; Rettenmaier, T. J.; Bichmann, L.; Keedy, D. A.; Woldeyes, R. A.; Schneidman-Duhovny, D.; Demerdash, O. N.; Mitchell, J. C.; Wells, J. A.; Fraser, J. S.; Sali, A. *J. Mol. Biol.* **2016**, *428*, 709–719.

(16) Hussein, H. A.; Borrel, A.; Geneix, C.; Petitjean, M.; Regad, L.; Camproux, A.-C. *Nucleic Acids Res.* **2015**, *43*, W436–442.

(17) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. *Nat. Protoc.* **2015**, *10*, 733–755.

(18) Ngan, C. H.; Bohnuud, T.; Mottarella, S. E.; Beglov, D.; Villar, E. A.; Hall, D. R.; Kozakov, D.; Vajda, S. *Nucleic Acids Res.* **2012**, *40*, W271–W275.

(19) Stank, A.; Kokh, D. B.; Horn, M.; Sizikova, E.; Neil, R.; Panecka, J.; Richter, S.; Wade, R. C. *Nucleic Acids Res.* **2017**, *45*, W325–W330.

(20) Oleinikovas, V.; Saladino, G.; Cossins, B. P.; Gervasio, F. L. *J. Am. Chem. Soc.* **2016**, *138*, 14257–14263.

(21) Ghanakota, P.; Carlson, H. A. *J. Med. Chem.* **2016**, *59*, 10383–10399.

(22) Alvarez-Garcia, D.; Barril, X. *J. Med. Chem.* **2014**, *57*, 8530–8539.

(23) Faller, C. E.; Raman, E. P.; MacKerell, A. D.; Guvench, O. *Methods Mol. Biol.* **2015**, *1289*, 75–87.

(24) Ghanakota, P.; Carlson, H. A. *J. Phys. Chem. B* **2016**, *120*, 8685–8695.

(25) Kimura, S. R.; Hu, H. P.; Ruvinsky, A. M.; Sherman, W.; Favia, A. D. *J. Chem. Inf. Model.* **2017**, *57*, 1388–1401.

(26) Lama, D.; Brown, C. J.; Lane, D. P.; Verma, C. S. *Biochemistry* **2015**, *54*, 6535–6544.

(27) Pérot, S.; Sperandio, O.; Miteva, M. A.; Camproux, A.-C.; Villoutreix, B. O. *Drug Discovery Today* **2010**, *15*, 656–667.

(28) Cheng, A. C.; Coleman, R. G.; Smyth, K. T.; Cao, Q.; Soulard, P.; Caffrey, D. R.; Salzberg, A. C.; Huang, E. S. *Nat. Biotechnol.* **2007**, *25*, 71–75.

(29) Volkamer, A.; Kuhn, D.; Grombacher, T.; Rippmann, F.; Rarey, M. *J. Chem. Inf. Model.* **2012**, *52*, 360–372.

(30) Schmidtke, P.; Barril, X. *J. Med. Chem.* **2010**, *53*, 5858–5867.

(31) Wang, L.; Xie, Z.; Wipf, P.; Xie, X.-Q. *J. Chem. Inf. Model.* **2011**, *51*, 807–815.

(32) Berman, H.; Henrick, K.; Nakamura, H. *Nat. Struct. Biol.* **2003**, *10*, 980.

(33) Lexa, K. W.; Carlson, H. A. *J. Am. Chem. Soc.* **2011**, *133*, 200–202.

(34) Foster, T. J.; MacKerell, A. D.; Guvench, O. *J. Comput. Chem.* **2012**, *33*, 1880–1891.

(35) Hyde, J.; Braisted, A. C.; Randal, M.; Arkin, M. R. *Biochemistry* **2003**, *42*, 6475–6483.

(36) Braisted, A. C.; Oslob, J. D.; Delano, W. L.; Hyde, J.; McDowell, R. S.; Waal, N.; Yu, C.; Arkin, M. R.; Raimundo, B. C. *J. Am. Chem. Soc.* **2003**, *125*, 3714–3715.

(37) Thanos, C. D.; Randal, M.; Wells, J. A. *J. Am. Chem. Soc.* **2003**, *125*, 15280–15281.

(38) Arkin, M. R.; Randal, M.; DeLano, W. L.; Hyde, J.; Luong, T. N.; Oslob, J. D.; Raphael, D. R.; Taylor, L.; Wang, J.; McDowell, R. S.; Wells, J. A.; Braisted, A. C. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 1603–1608.

(39) Tan, Y. S.; Śledź, P.; Lang, S.; Stubbs, C. J.; Spring, D. R.; Abell, C.; Best, R. B. *Angew. Chem., Int. Ed. Engl.* **2012**, *51*, 10078–10081.

(40) Cheng, K.-Y.; Lowe, E. D.; Sinclair, J.; Nigg, E. A.; Johnson, L. N. *EMBO J.* **2003**, *22*, 5757–5768.

(41) Wang, X.; Minasov, G.; Shoichet, B. K. *J. Mol. Biol.* **2002**, *320*, 85–95.

(42) Trott, O.; Olson, A. J. *J. Comput. Chem.* **2010**, *31*, 455–461.

(43) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA, 2008; pp 11 – 15.

(44) Sakamoto, Y.; Ishiguro, M.; Kitagawa, G. *Akaike information criterion statistics*; Tokyo : KTK Scientific Publishers ; Dordrecht ; Boston : D. Reidel ; Hingham, MA : Sold and distributed in the U.S.A. and Canada by Kluwer Academic Publishers, 1986.

(45) R Development Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2008.

(46) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. *BMC Bioinf.* **2009**, *10*, 168.

(47) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. *Bioinformatics* **2017**, *33*, 3036–3042.

(48) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. *J. Chem. Theory Comput.* **2016**, *12*, 1845–1852.

(49) Parrish, R. M. et al. *J. Chem. Theory Comput.* **2017**, *13*, 3185–3197.

(50) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.

(51) MacKerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(52) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. *Biophys. J.* **2011**, *100*, L47–L49.

(53) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.

(54) Lexa, K. W.; Goh, G. B.; Carlson, H. A. *J. Chem. Inf. Model.* **2014**, *54*, 2190–2199.

(55) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

(56) Bakan, A.; Nevins, N.; Lakdawala, A. S.; Bahar, I. *J. Chem. Theory Comput.* **2012**, *8*, 2435–2447.

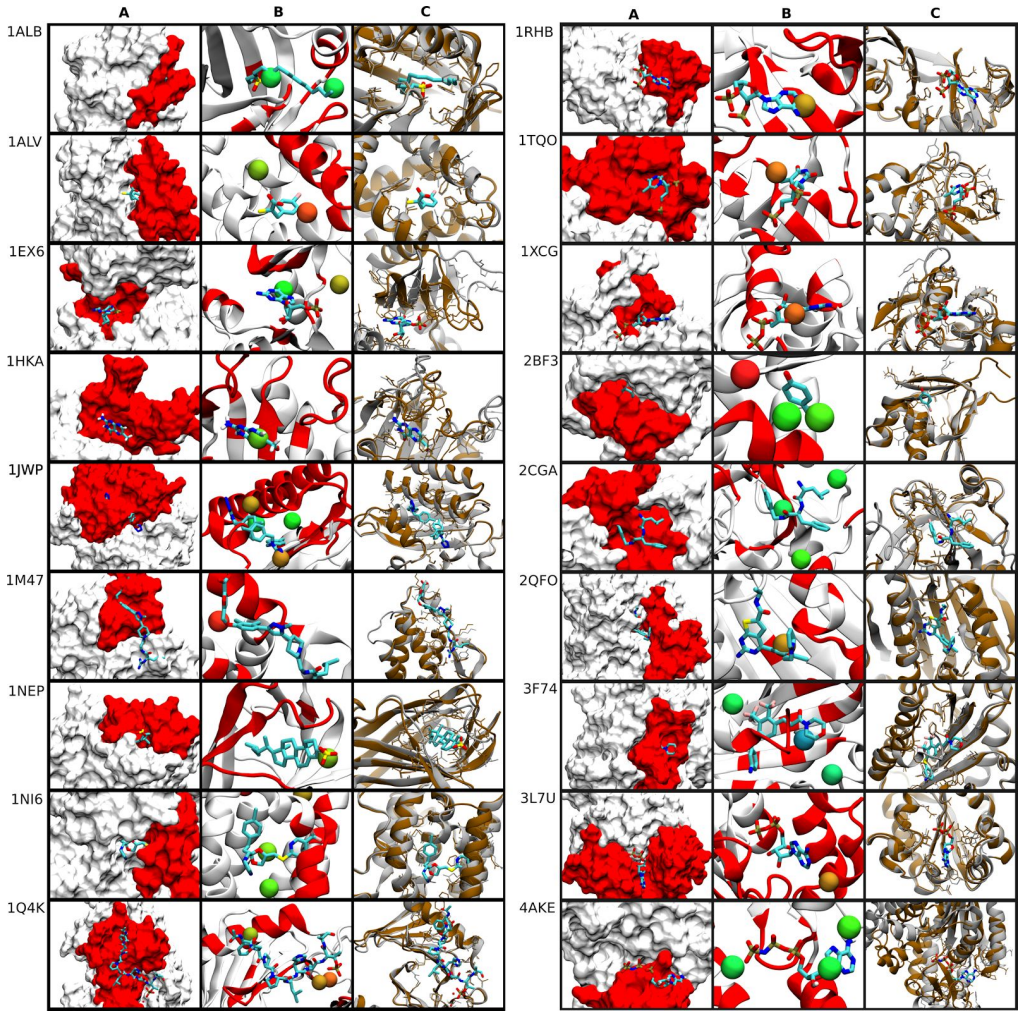
(57) Pedregosa, F. et al. *J. Mach. Learn. Res.* **2011**, *12*, 2825–¡–2830.

(58) Rose, A. S.; Hildebrand, P. W. *Nucleic Acids Res.* **2015**, *43*, W576–W579.

(59) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. *J. Chem. Inf. Model.* **2010**, *50*, 397–403.

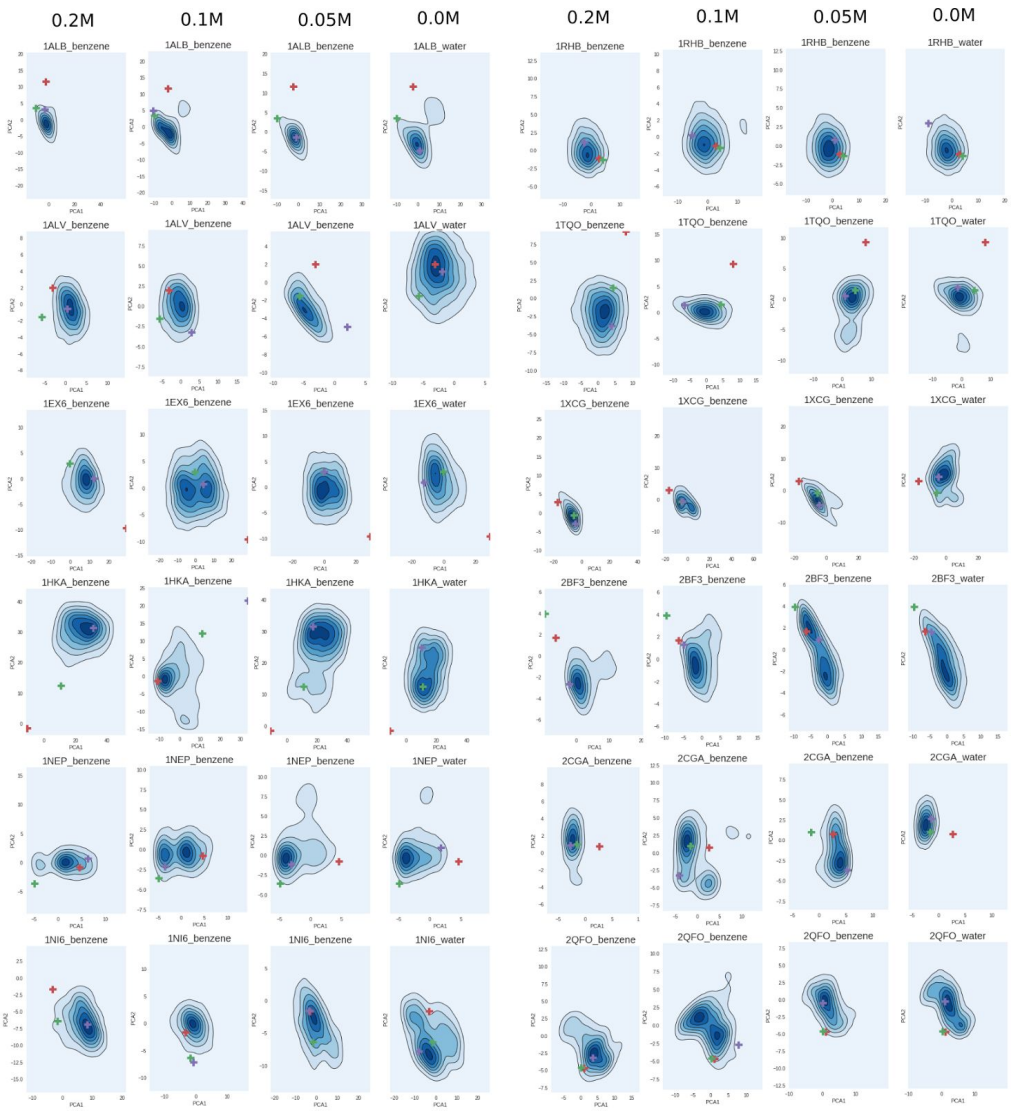
Supporting information for
PlayMolecule CryptoScout: predicting protein cryptic sites using mixed-solvent molecular simulations and mutual information. Gerard Martinez-Rosell and Gianni de Fabritiis

S1



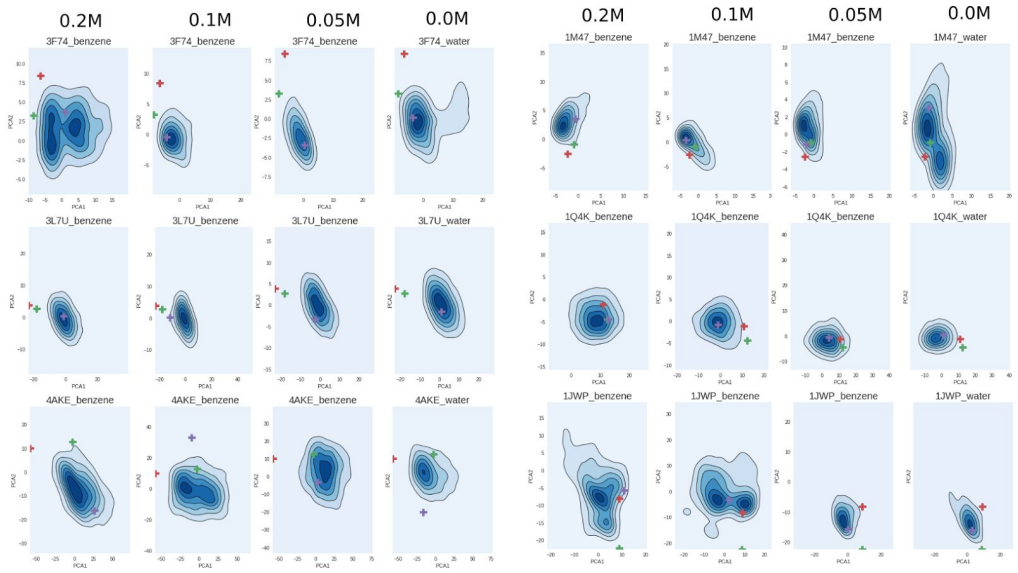
Benzene binding hot-spots found for each of the 18 systems in the 0.1M benzene condition. Column **A** shows the surface of the *apo* structure and the ligand present in the *holo* pose after aligning *apo* and *holo* structures by the backbone. The community of residues labelled as “positive” (i.e. contains a cryptic pocket) is colored in red. Note how most of the ligands present clashes with the *apo* surface due to the closure of the cryptic pocket. Column **B** shows the *apo* backbone, aligned with the *holo* ligand and the hot-spots detected by CryptoScout depicted as beads colored from red to green, being green the lowest free energy. Residues part of a community labelled as “positive” are colored in red. Column **C** shows the *holo* ligand environment with the *holo* conformation depicted in brown cartoon and the aligned *apo* conformation in grey cartoon.

S2.1



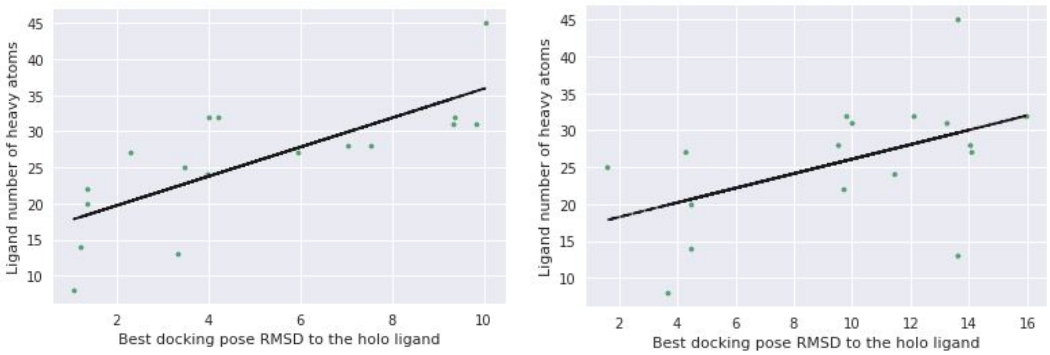
2D histogram plot calculated from the first 2 PCA components of the minimum inter-residue distance between all cryptic pocket residues pairs for 0.2M, 0.1M and 0.05M benzene and water-only conditions. Green crosses represent the *apo* conformation, red color represents the *holo* conformation and purple crosses represent post-equilibration conformation. Shown 12 out of 18 systems.

S2.2



2D histogram plot calculated from the first 2 PCA components of the minimum inter-residue distance between all cryptic pocket residues pairs for 0.2M, 0.1M and 0.05M benzene and water-only conditions. Green crosses represent the *apo* conformation, red color represents the *holo* conformation and purple crosses represent post-equilibration conformation. Shown the remaining 6 out of 18 systems.

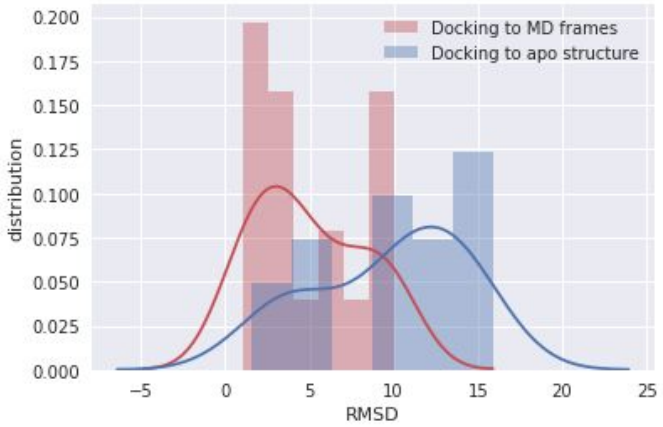
S3



Correlation between ligand number of heavy atoms and best docking pose RMSD (lowest RMSD to the *holo* ligand pose). Left: best docked posed using 10 structures extracted from MD simulations. P-value = 0.0004*; R-value=0.75; Right: best docked pose using the *apo* structure. P-value = 0.04*; R-value: 0.49

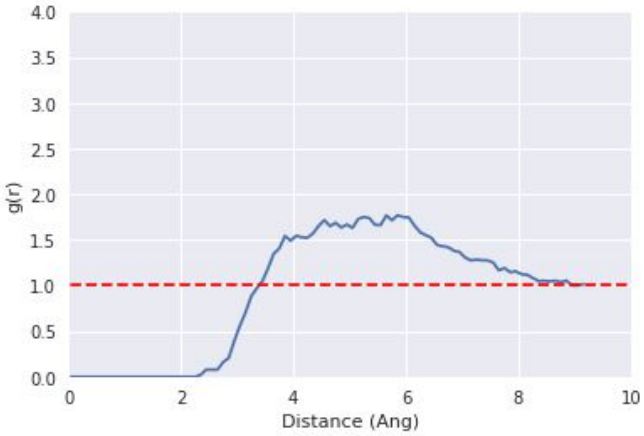
S4

Distribution plot of best RMSD from (a) docking to 10 representative frames extracted from MD (red) and (b) docking to the *apo* conformation (blue).



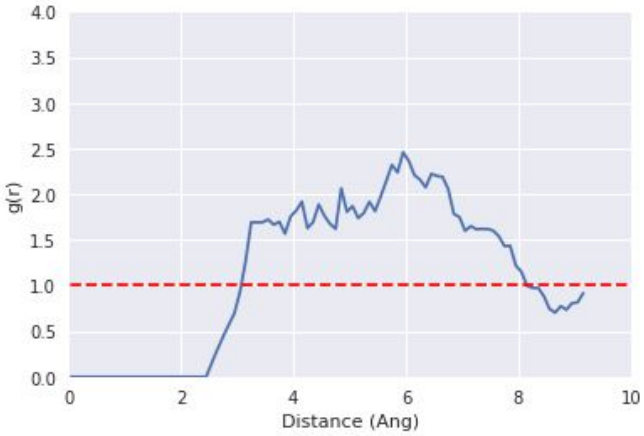
S5.1

RDF for benzene from 0.2M water+benzene simulations; calculated with VMD with 0.1 Ang sphere resolution plus a moving average smoothing of 10; based on 4 80ns-long simulation of water+benzene; notice it flattens at VDW cutoff, which is 9 Angstrom



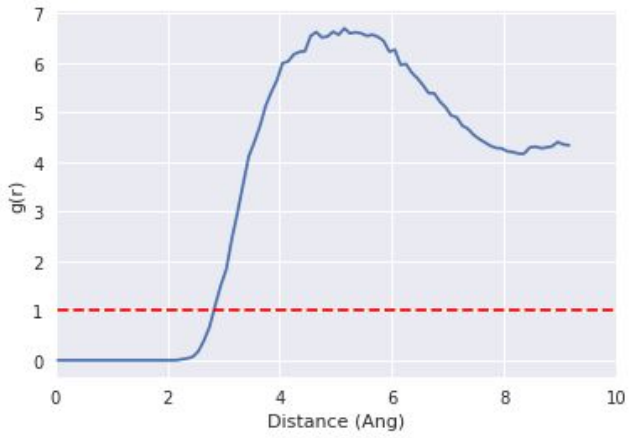
S5.2

RDF for benzene from 1 80 ns-long simulation of 0.1M water+benzene using moving average of 10. Resolution was 0.1 Ang.

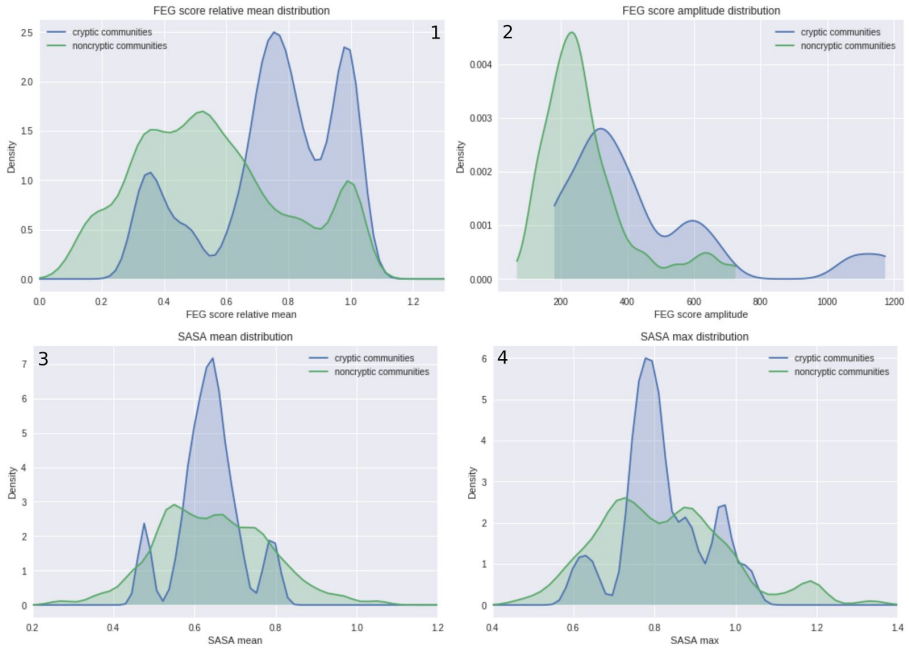


S5.3

RDF for benzene from 1 80 ns-long 1M water+benzene simulation using a moving average of 10. Resolution was 0.1 Ang.



S6



Distribution of the metrics (FEG score amplitude, FEG score relative mean, SASA mean, SASA max) used for the logistic regression classifier for cryptic communities (blue) and non-cryptic communities (green) calculated over the 0.2M MD simulations.

S.7 PlayMolecule CryptoScout web application details

The web application consists of the following 4 steps:

(1) **Job submission.** The user is able to launch a new cryptic pocket prospection starting from an id of the PDB database, a custom protein from a pdb file or a protein prepared with PlayMolecule ProteinPrepare[1]. The user can also choose a protein chain, a pH that is used for the residue tritration and a benzene concentration, although the recommended 0.1M concentration is pre-selected.

(2) **System building**. The web app reports to the user details about the built system (figure 7.A) or, in case something went wrong, an error message with suggestions to solve the issue.

(3) **Equilibration and production**. The equilibration and production simulations are run and the user is reported with the current stage of the progress.

(4) **Analysis and results**. The analysis is performed and the results are reported to the user. The results consist of: (a) a tab including each community of residues with an associated color, stats and CryptoScout score, i.e. likelihood of containing a cryptic/binding site (figure 7.C). (b) A central webGL-powered protein structure visualizer with the reference structure colored by CryptoScout community (figures 7.B and 7.C), hot-spots represented as beads with color from white to red proportional to decreasing free energy and an isosurface representing the benzene occupancy (figure 7.B). Several options are available such as showing/hiding surface representation. (c) A right accordion panel with each selectable hot-spot detected and associated free energy. (d) Another right panel with buttons for each of the community detected and its associated color. Each of these buttons allow the user to visualize 1 representative for each of the 10 clusters based on SASA fluctuation of the specific community (figure7.D). These structures are aligned by the backbone and lateral chains of the community are shown and colored by residue name to give a rough idea of the community dynamics and the potential cryptic sites. (e) Finally, a *download* button allows the user to download all result tables in *csv* format and the filtered raw simulations (i.e. without water to reduce the file size). (f) Additionally, a tab titled *Analysis with HTMD* offers to the user an example of analysis script using the HTMD module to encourage a further analysis of the trajectories.

[1] Martínez-Rosell, Gerard and Giorgino, Toni and De Fabritiis, Gianni. *PlayMolecule ProteinPrepare: A Web Application for Protein Preparation for Molecular Dynamics Simulations*. J. Chem. Inf. Model. 1511--1516

3.3 Molecular simulation-driven fragment screening for the discovery of new CXCL12 inhibitors

Martinez-Rosell G, Harvey MJ, De Fabritiis G. Molecular-Simulation-Driven Fragment Screening for the Discovery of New CXCL12 Inhibitors. J Chem Inf Model. 2018 Mar 26;58(3):683–91. DOI: 10.1021/acs.jcim.7b00625

Summary
In this work we produce the first 150-fragment screening exclusively driven by high-throughput molecular dynamics (MD) against CXCL12, a chemokine closely related to diseases such as cancer metastasis. As a result, we are able to predict the binding of 8 millimolar-affinity fragments to two CXCL12 cavities detected experimentally, namely sY7 and H1S68. The binding mode and the pharmacophoric properties of the fragment *hits* are consistent with the natural ligand-like binding moieties. The steady decrease in computational cost and the present study pave the way for the introduction of MD simulation as a screening tool in early phases of the mainstream drug discovery pipelines.

3.4 Dynamic and Kinetic Elements of μ-Opioid Receptor Functional Selectivity

Kapoor A, Martinez-Rosell G, Provasi D, de Fabritiis G, Filizola M. Dynamic and Kinetic Elements of μ-Opioid Receptor Functional Selectivity. Sci Rep. 2017 Sep 12;7(1):11255. DOI: 10.1038/s41598-017-11483-8

Summary
In this publication we study the effect of two drugs on the conformational plasticity of the μ-opioid receptor (MOR): (1) morphine, a classical opioid drug, and (2) TRV-130, a potent G protein-biased agonist. Particularly, we produced more than half millisecond of MD simulations of MOR bound to these two ligands to study the effect of the drugs on the dynamics and kinetics of MOR and to understand better the molecular basis of functional selectivity. As a result, we identify differential metastable states across the inactivation pathway, as well as differential deactivation pathways, kinetics and differential allosteric communication of the drugs across MOR.

Note: my contribution to this work has been the production of the simulations as well as the validation of analysis and manuscript.

3.5 Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale

Martínez-Rosell G, Giorgino T, Harvey MJ, de Fabritiis G. Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale. Curr Top Med Chem. 2017 Aug 8;17(23):2617–25. DOI: 10.2174/1568026617666170414142549

Summary
In this review we follow the steps of the oldest GPU MD code, ACEMD, since its inception in 2009 until 2016. In particular, we focus on publications focusing on drug discovery and we analyze the evolution of the field since its humble beginnings with studies of protein-ion binding until complex studies of multi-fragment binding. Furthemore, we analyze the evolution of GPU hardware performance and we predict that we will reach the second timescale by 2022 based on the observed trend.

3.6 High-Throughput Automated Preparation and Simulation of Membrane Proteins with HTMD

Doerr S, Giorgino T, Martínez-Rosell G, Damas JM, De Fabritiis G. High-Throughput Automated Preparation and Simulation of Membrane Proteins with HTMD. J Chem Theory Comput. 2017 Sep 12;13(9):4003–11. DOI: 10.1021/acs.jctc.7b00480

Summary
In previous work it was shown that HTMD [80] python module offered a powerful solution for the analysis of MD simulations and for running adaptive sampling schemes. In this publication, we extend the HTMD software functionality to include a module for building and running membrane protein systems. To test the reliability of our building protocol, we automatically built and equilibrated more than 640 membrane proteins from the OPM database for both CHARMM and AMBER forcefields. we then perform a short analysis to determine the quality of the built systems and their equilibration. Finally, we share all the built systems to the scientific community through the *PlayMolecule* web platform (*http://www.playmolecule.org/OPM/*).

Note: my contribution to this work has been the production of a web application that allows the users to access the data generated in this publication.

3.7 Data Augmentation and Predictions by Molecular Dynamics Simulations and Machine Learning

Adrià Pérez, Gerard Martínez-Rosell, Gianni de Fabritiis. Under review in *Curr. Opin. Struct. Biol.*

Summary
In this opinion article we envisage an upcoming scenario in which MD simulations will be used as data augmentation tools for machine learning algorithms such as deep learning and convolutional neural networks (CNN). With the current explosion of machine learning applications applied to chemistry and biophysics (such as Publication 6.1), we have realized that data scarcity is the main limiting factor for these algorithms to learn efficiently. Therefore, we propose that MD simulations can extend the amount of available data by, for instance, predicting protein-ligand binding affinity or protein folding for which no experimental data is available. These *in silico*-generated data points (e.g. K_d or protein folded structure) can be appended to the amount of experimental data already available and fed together into deep learning algorithms obtaining better generalizations and results than with experimental data alone.

Data Augmentation and Predictions by Molecular Dynamics Simulations and Machine Learning

Adrià Pérez[a], Gerard Martínez-Rosell[a], Gianni De Fabritiis[a,b]

[a] *Computational Biophysiscs Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003 Barcelona, Spain*
[b] *Institució Catalana de Recerca i Estudis Avanats (ICREA), Passeig Lluis Companys 23, Barcelona 08010, Spain*

Abstract

In the next five years, all-atom molecular dynamics (MD) simulations are expected to reach sampling within the second timescale, producing petabytes of simulation data. Notwithstanding this, MD will still be limited to low-throughput, high-latency predictions. To overcome this limitation, we envisage that MD simulations will also be used as a data augmentation tool integrating experimental data to train fast machine learning predictive models. The synergy between MD simulations and machine learning methods, such as artificial neural networks, has the potentiality to drastically reshape the way we make predictions in computational structural biology and drug discovery.

Highlights

- Within five years, MD will reach the second timescale and generate petabytes of data, yet MD is limited to low-throughput predictions.
- MD simulations can be used as a tool for data augmentation to train machine learning predictive models.
- Potential synergies exist between MD and machine learning, from forcefields to predictive models.

Email address: gianni.defabritiis@upf.edu (Gianni De Fabritiis)

Introduction

Molecular dynamics (MD) simulations are one of the predominant techniques to study protein dynamics. MD is often used to capture dynamical processes of proteins across different timescales with atomistic details, as a way to rationalize some biological phenomena. Despite the potential to become a surrogate model of real protein dynamics, some important issues still remain to be solved, mainly: i) forcefield accuracy and precision [1, 2, 3, 4], ii) high computational cost and sampling limitations. Classic MD simulations constitute a balance between accuracy and efficiency. Quantum-level phenomena such as enzymatic reactions and proton transfers are completely neglected in exchange for computational speed. The extent to which these limitations may affect the validity of the results depends on the system and the biological question at hand.

Nevertheless, MD has evolved from single simulation studies [5, 6, 7] to a high-throughput molecular dynamics [8, 9, 10, 11, 12] where hundreds of microseconds of simulations are performed in parallel to obtain converged statistics and new hypotheses about the underlying molecular phenomena of the given study. Although it is possible to generate a lot of data for a single system, the knowledge extracted from it is currently mainly used to rationalize a particular mechanism. Here, we envision an alternative use of the data generated from all these isolated studies in a way that general patterns can be learned and further predictions can be drawn by using machine learning approaches. By doing so, one could expand the system-specific knowledge to a much wider scope.

In this review, we first describe high-throughput simulation studies producing high quantity of data in the fields of protein folding and protein-ligand binding, to demonstrate the computer power currently available and future expectations in terms of data production. Secondly, we discuss state-of-the-art machine learning technology that has been recently applied to structural biology. Finally, we combine both ideas by hypothesizing the use of MD simulations to augment existing experimental data (structural and dynamical information about proteins, thermodynamics and kinetics estimations of recognition processes) and enhance the prediction power of machine learning approaches. We focus on two practical examples: the prediction of protein-ligand binding free energy and the prediction of protein structure.

Accelerated and high-throughput molecular dynamics

Software and hardware innovations, such as the implementation of MD codes for GPUs [13, 14, 15, 16], the appearance of distributed computing projects like Folding@home[17] or GPUGRID [18] and the development of special-purpose supercomputers like ANTON [19], are steadily decreasing the computational cost of molecular simulations. Additionally, the development of adaptive sampling schemes have introduced more efficient ways to explore the conformational space, decreasing the amount of simulations needed to obtain converged statistics [20, 21, 22]. Recently, new adaptive sampling algorithms have been proposed, aimed at improving even more the efficiency of the existing schemes [23, 24, 25]. Inspired by the multi-armed bandit problem, these algorithms balance between exploration of the conformational space and exploitation of a given metric, such as residue contacts or protein-ligand distances, to guide the sampling. They have proven to speed up convergence in different applications, such as ligand binding in GPCRs [23], protein folding [24, 25] and protein conformational exploration [24].

The introduction of GPU MD software made simulations of full protein-ligand binding processes faster and widely accessible, allowing for the prediction of thermodynamic and kinetic properties, e.g. binding free energy and binding rates [26]. MD simulations can efficiently reconstruct full protein-ligand binding events and kinetic properties. This has been demonstrated in several studies, such as the benzamidine-trypsin system [26], as well as in [8, 9], both using fragment-sized ligands, for which kinetics are known to be particularly fast and therefore computationally attainable. MD simulations have also proven to be a valuable approach in the field of protein folding, obtaining atomic-level descriptions of folding dynamics and shedding light on protein conformational plasticity. Several examples of atomistic folding simulations were performed with fast-folding proteins, like the Villin headpiece subdomain [5], the Trp-cage miniprotein [27, 28, 29] or the mutant Pin1 WW domain [30].

Specialized supercomputers and GPU computing made millisecond simulations possible and expanded the possibilities regarding protein folding. The first reported trajectory to surpass the millisecond barrier was a BPTI folding simulation, performed with the supercomputer ANTON [31]. One year later, ANTON was used again to perform a total of 8.2 ms of folding simulations for 12 fast-folding proteins in explicit solvent [32], generating between 0.1 and 1 ms for each protein and capturing several folding and unfolding

events. Besides fast-folding proteins, bigger systems with folding timescales in the order of milliseconds have also been successfully simulated, such as ubiquitin [33] and ACBP (in implicit solvent) [34].

Data augmentation for machine learning

In a recent review we estimated that MD will reach seconds of aggregated sampling using commodity hardware by 2022 [35], generating petabytes of simulation data. This amount of data constitutes a valuable source of potential information that can be exploited. For instance, instead of using MD as a way to learn about a particular protein or mechanism, simulation data coming from a diverse set of proteins can be combined and fed into machine learning algorithms to create predictive models based on MD training data. By following this approach, it is possible to exploit curated simulation datasets, so that the knowledge extracted from it can be applied to many more cases. In such way, MD would be used as a data augmentation tool to generate data and machine learning techniques would be used to exploit this data efficiently, creating faster and more accurate predictive models [Fig 1].

To illustrate the proposed approach, let us take as an example the case of binding affinity prediction. Currently, there are several ways to predict binding affinity of protein-ligand complexes. Although widely used, docking algorithms usually provide fast but inaccurate results. MD is a more expensive approach but with promising results throughout the literature. Unbiased MD simulations can sample the spontaneous binding of a ligand, but they are computationally very expensive to converge. To overcome this limitation, alternatives to unbiased methods have also arisen, such as free energy perturbation methods (FEP) [36, 37, 38], metadynamics [39, 40], umbrella sampling [41, 42] and steered MD [43, 44]. All these techniques are valuable to obtain binding free energies and literature provides plenty of successful cases. However, techniques and results seem to be relatively system-specific and low-throughput.

For illustrative purposes, a way to overcome these limitations could be to extend the PDBbind database [45]. This database includes 16.000 protein-ligand pairs and corresponding annotated affinity. The PDBbind database reviewed more than 50.000 protein-ligand structures from the PDB database in order to curate the general set of 16.000 structures. This reflects, once redundancy is excluded, the massive amount of ligand-protein structures with missing annotated affinity that could be computed using MD. While the

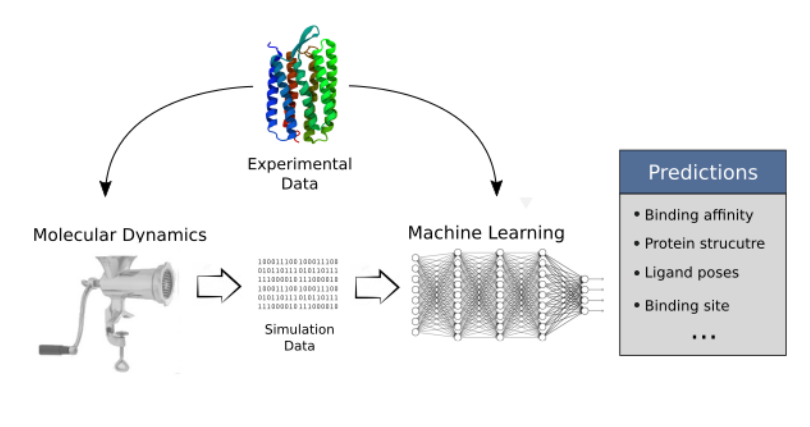


Figure 1: General scheme of the suggested approach. MD simulations and machine learning methods are both combined to obtain predictive models which can be used for high-throughput prediction studies. MD is used as a data augmentation tool to expand the training data available. Then, structural, thermodynamic and kinetic experimental data and simulation data are integrated together with machine learning methods to create predictive models. The synergy between MD and machine learning expands their current capabilities to perform fast and accurate predictions.

amount of computational work may be currently unattainable, our predictions foresee enough computational power in the near future to attempt this challenge. More interestingly, once an expanded structure-affinity dataset is created, one could use it to train machine learning models, such as deep neural networks. These algorithms are particularly sensitive to the amount of training data but, once extensively trained, they can yield predictions in an infinitesimal fraction of time compared to MD simulations.

Another example application where MD can be used to produce training data is protein structure prediction. One recent achievement showcased the use of evolutionary information and Rosetta to predict the unknown structure for 614 proteins [46]. While large-scale predictions, such as the aforementioned, are currently impossible using plain molecular dynamics due to sampling limitations, a combined approach of a protein folding simulation training dataset plus a machine learning algorithm that learns and generalizes could prove effective. Machine learning algorithms could leverage petabytes of folding simulation data of a representative group of proteins to learn about

the general mechanisms of folding. By doing so, the prediction of hundreds of protein structures could become feasible.

Current state of machine learning in computational structural biology

One of the main challenges then is how to efficiently analyze all the data generated to obtain knowledge of different biological events. Machine learning approaches are already being used to analyze MD trajectories, such as different clustering methods, signal processing methods (tICA [47, 48], PCA [49, 50, 51]) and Markov state models [52, 53, 54, 55]. These algorithms help to unravel the dynamic information contained in the simulation data. With MSMs, one can obtain a good representation of a protein's free energy landscape in a human-understandable way. Still, the current analysis methods used in MD have limited power when trying to leverage all simulation data in order to gain a generalized understanding of it.

To learn from simulation data, the analysis techniques should be data driven, being able to integrate the information from different simulations and learn from it, detecting the basic features inside the data and creating models for the general mechanisms of protein dynamics. This type of analysis is slowly rising in computational biology in the form of deep learning [56]. Several achievements have been accomplished using deep neural networks (DNN) in computational biology. For instance, the Merck molecular activity challenge demonstrated the potential of DNN-like models in the field [57]. Focusing on computational chemistry problems, variational autoencoders [58], a generative flavor of DNNs, were recently applied to convert discrete representations of molecules to and from a multidimensional continuous representation [59], allowing for efficient search and optimization through open-ended spaces of chemical compounds. DNN-like approaches consistently outperform previous existing models. For instance, DeepTox [60] won the Tox21 toxicology prediction challenge in 2014 by a large margin. The DeepChem software [61] and the MoleculeNet challenge [62] have recently helped by providing multiple featurization algorithms and access to relevant QSAR prediction datasets. Regarding the analysis of structural data, deep convolutional neural networks have become increasingly popular due to its extraordinary performance in machine vision [63, 64], and they have been used in problems such as virtual screening by classifying compounds as active or inactive [65], ligand binding site detection [66], ligand

pose prediction [67] and ligand affinity prediction [68].

A new interesting approach living in between machine-learning and MD is followed in [69, 70, 71], where a neural network is trained with QM simulation data to generate the potential energy surface and forces for a general system of atoms. In the same way as MD force-fields do, the forces are true derivatives of the interpolated potential energy surface using the gradients of the neural network and can be used to run dynamics. The QM simulation data is therefore learned by the model, with the accuracy of first-principle based methods at a computational cost several orders of magnitudes faster than the QM computational model, comparable to classical MD.

Discussion

Modern machine learning approaches can learn representative features from data obtained by MD simulations and could provide effective predictive models to apply in different structural biology problems, such as protein folding or protein-ligand binding. The synergy between MD simulations and state-of-the-art machine learning methods could bring the current prediction performance of MD substantially beyond its limits, in a more cost efficient way than with simulations alone. The main aim would be to develop predictive models, based on novel machine learning algorithms and trained on a unique datasets of petabytes of MD and QM simulation data. In this context, MD could be an *in silico* data generation method to expand experimental data.

From the point of view of *in silico* data generation, we can make a broad comparison between MD simulations and other machine learning applications. For instance AlphaGo, the artificial intelligence software designed to play the board game Go, which recently defeated the best human Go player [72], was trained with board positions from thousands of real-life games. In a second training phase, the computer was set to play against itself. While in the first phase the computer was trained with experimental data, the second phase used training data generated completely *in silico*.

The examples provided throughout this article illustrate that, soon, *in silico* generated data coming from MD simulations might be used as input for machine learning models to augment experimental data, in order to develop new integrated predictive models.

Acknowledgements

The authors thank Acellera Ltd. for funding. G.D.F. acknowledges support from MINECO (BIO2014-53095-P) and FEDER. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 675451 (CompBioMed project).

References

[1] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ 1 and χ 2 Dihedral Angles," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3257–3273, 2012.

[2] K. A. Beauchamp, Y. S. Lin, R. Das, and V. S. Pande, "Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements," *J. Chem. Theory Comput.*, vol. 8, no. 4, pp. 1409–1414, 2012.

[3] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," *PLoS ONE*, vol. 7, no. 2, 2012.

[4] S. Piana, J. L. Klepeis, and D. E. Shaw, "Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations," *Curr. Opin. Struct. Biol.*, vol. 24, no. 1, pp. 98–105, 2014.

[5] Y. Duan, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science*, vol. 282, no. 5389, pp. 740–744, 1998.

[6] A. Grossfield, M. C. Pitman, S. E. Feller, O. Soubias, and K. Gawrisch, "Internal Hydration Increases during Activation of the G-Protein-Coupled Receptor Rhodopsin," *J. Mol. Biol.*, vol. 381, no. 2, pp. 478–486, 2008.

[7] R. O. Dror, D. H. Arlow, D. W. Borhani, M. O. Jensen, S. Piana, and D. E. Shaw, "Identification of two distinct inactive conformations of

the 2-adrenergic receptor reconciles structural and biochemical observations," *Proc. Natl. Acad. Sci.*, vol. 106, no. 12, pp. 4689–4694, 2009.

[8] N. Ferruz, M. J. Harvey, J. Mestres, and G. De Fabritiis, "Insights from Fragment Hit Binding Assays by Molecular Simulations," *J. Chem. Inf. Model.*, vol. 55, no. 10, pp. 2200–2205, 2015.

[9] A. C. Pan, H. Xu, T. Palpant, and D. E. Shaw, "Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations," *J. Chem. Theory Comput.*, vol. 13, no. 7, pp. 3372–3377, 2017.
• In this paper, Shaw and coworkers characterize the binding affinity of several drug fragments for the FKBP protein using unbiased MD simulations, and compare the results to FEP calculations to find the values agree within statistical error.

[10] N. Ferruz, G. Tresadern, A. Pineda-Lucena, and G. De Fabritiis, "Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme.," *Sci. Rep.*, vol. 6, p. 30275, 2016.

[11] N. Stanley, L. Pardo, and G. D. Fabritiis, "The pathway of ligand entry from the membrane bilayer to a lipid G protein-coupled receptor," *Sci. Rep.*, vol. 6, no. 1, p. 22639, 2016.

[12] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete proteinprotein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling," *Nat. Chem.*, 2017.
•• Plattner and coworkers managed to obtain the binding thermodynamic and kinetic properties of a protein-protein association between Barnase and Barstar.

[13] M. J. Harvey, G. Giupponi, and G. De Fabritiis, "ACEMD: Accelerating biomolecular dynamics in the microsecond time scale," *J. Chem. Theory Comput.*, vol. 5, no. 6, pp. 1632–1639, 2009.

[14] M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A. L. Beberg, D. L. Ensign, C. M. Bruns, and V. S. Pande, "Accelerating molecular dynamic simulation on graphics processing units," *J. Comput. Chem.*, vol. 30, no. 6, pp. 864–872, 2009.

[15] M. J. Harvey and G. De Fabritiis, "An implementation of the smooth particle mesh Ewald method on GPU hardware," *J. Chem. Theory Comput.*, vol. 5, no. 9, pp. 2371–2377, 2009.

[16] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L. P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, *et al.*, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," *PLoS Comput. Biol.*, vol. 13, no. 7, 2017.

[17] M. Shirts and V. S. Pande, "COMPUTING: Screen Savers of the World Unite!," *Science (New York, N.Y.)*, vol. 290, no. 5498, pp. 1903–4, 2000.

[18] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, "High-throughput all-atom molecular dynamics simulations using distributed computing," *J. Chem. Inf. Model.*, vol. 50, no. 3, pp. 397–403, 2010.

[19] D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Lerardi, I. Kolossváry, J. L. Klepeis, T. Layman, *et al.*, "Anton, a special-purpose machine for molecular dynamics simulation," *Commun. ACM*, vol. 51, no. 7, p. 91, 2008.

[20] N. Singhal and V. S. Pande, "Error analysis and efficient sampling in Markovian state models for molecular dynamics," *J. Chem. Phys.*, vol. 123, no. 20, 2005.

[21] N. S. Hinrichs and V. S. Pande, "Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics," *J. Chem. Phys.*, vol. 126, no. 24, 2007.

[22] S. Doerr and G. De Fabritiis, "On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations," *J. Chem. Theory Comput.*, vol. 10, no. 5, pp. 2064–2069, 2014.

[23] D. Sabbadin and S. Moro, "Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale," *J. Chem. Inf. Model.*, vol. 54, no. 2, pp. 372–376, 2014.

[24] M. I. Zimmerman and G. R. Bowman, "FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs," *J. Chem. Theory Comput.*, vol. 11, no. 12, pp. 5747–5757, 2015.
•• Zimmerman and Bowman describe the FAST adaptive scheme, which is a goal-oriented adaptive sampling method. The algorithm balances between the exploitation of promising conformations and exploration of new conformations.

[25] J. L. MacCallum, A. Perez, and K. A. Dill, "Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference," *Proc. Natl. Acad. Sci.*, vol. 112, no. 22, pp. 6985–6990, 2015.

[26] I. Buch, T. Giorgino, and G. De Fabritiis, "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations," *Proc. Natl. Acad. Sci.*, vol. 108, no. 25, pp. 10184–10189, 2011.

[27] C. Simmerling, B. Strockbine, and A. E. Roitberg, "All-atom structure prediction and folding simulations of a stable protein," *J. Am. Chem. Soc.*, vol. 124, no. 38, pp. 11258–11259, 2002.

[28] C. D. Snow, B. Zagrovic, and V. S. Pande, "The Trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations," *J. Am. Chem. Soc.*, vol. 124, no. 49, pp. 14548–14549, 2002.

[29] J. Juraszek and P. G. Bolhuis, "Sampling the multiple folding mechanisms of Trp-cage in explicit solvent," *Proc. Natl. Acad. Sci.*, vol. 103, no. 43, pp. 15859–15864, 2006.

[30] P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, "Ten-Microsecond Molecular Dynamics Simulation of a Fast-Folding WW Domain," *Biophys. J.*, vol. 94, no. 10, pp. L75–L77, 2008.

[31] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, *et al.*, "Millisecond-scale molecular dynamics simulations on anton," in *Proc. Conf. High Perform. Comput. Networking, Storage Anal.*, SC '09, (New York, NY, USA), pp. 39:1–39:11, ACM, 2009.

[32] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold.," *Science (New York, N.Y.)*, vol. 334, no. 6055, pp. 517–20, 2011.

[33] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "Atomic-level description of ubiquitin folding," *Proc. Natl. Acad. Sci.*, vol. 110, no. 15, pp. 5915–5920, 2013.

[34] V. A. Voelz, M. Jäger, S. Yao, Y. Chen, L. Zhu, S. A. Waldauer, G. R. Bowman, M. Friedrichs, O. Bakajin, L. J. Lapidus, S. Weiss, and V. S. Pande, "Slow unfolded-state structuring in acyl-CoA binding protein folding revealed by simulation and experiment," *J. Am. Chem. Soc.*, vol. 134, no. 30, pp. 12565–12577, 2012.

[35] G. Martínez-Rosell, T. Giorgino, M. J. Harvey, and G. de Fabritiis, "Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale," *Curr. Top. Med. Chem.*, 2017.

[36] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, *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.*, vol. 137, no. 7, pp. 2695–2703, 2015.
• In this study, they perform binding affinity predictions over 200 ligands and 10 proteins using FEP/REST to validate the high accuracy of FEP calculations.

[37] E. B. Lenselink, J. Louvel, A. F. Forti, J. P. D. van Veldhoven, H. de Vries, T. Mulder-Krieger, F. M. McRobb, A. Negri, J. Goose, R. Abel, *et al.*, "Predicting Binding Affinities for GPCR Ligands Using Free-Energy Perturbation," *ACS Omega*, vol. 1, no. 2, pp. 293–304, 2016.

[38] S. Wan, A. P. Bhati, S. Skerratt, K. Omoto, V. Shanmugasundaram, S. K. Bagal, and P. V. Coveney, "Evaluation and Characterization of Trk Kinase Inhibitors for the Treatment of Pain: Reliable Binding Affinity Predictions from Theory and Computation," *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 897–909, 2017.

[39] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proc. Natl. Acad. Sci.*, vol. 99, no. 20, pp. 12562–12566, 2002.

[40] F. S. Di Leva, E. Novellino, A. Cavalli, M. Parrinello, and V. Limongelli, "Mechanistic insight into ligand binding to G-quadruplex DNA," *Nucleic Acids Res.*, vol. 42, no. 9, pp. 5447–5455, 2014.

[41] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," *J. Comput. Phys.*, vol. 23, no. 2, pp. 187–199, 1977.

[42] W. Wojtas-Niziurski, Y. Meng, B. Roux, and S. Bernèche, "Self-learning adaptive umbrella sampling method for the determination of free energy landscapes in multiple dimensions," *J. Chem. Theory Comput.*, vol. 9, no. 4, pp. 1885–1895, 2013.

[43] J. S. Patel, A. Berteotti, S. Ronsisvalle, W. Rocchia, and A. Cavalli, "Steered molecular dynamics simulations for studying protein-ligand interaction in

cyclin-dependent kinase 5," *J. Chem. Inf. Model.*, vol. 54, no. 2, pp. 470–480, 2014.

[44] G. Palermo, E. Minniti, M. L. Greco, L. Riccardi, E. Simoni, M. Convertino, C. Marchetti, M. Rosini, C. Sissi, A. Minarini, and M. De Vivo, "An optimized polyamine moiety boosts the potency of human type II topoisomerase poisons as quantified by comparative analysis centered on the clinical candidate F14512," *Chem. Commun.*, vol. 51, no. 76, pp. 14310–14313, 2015.

[45] Z. Liu, M. Su, L. Han, J. Liu, Q. Yang, Y. Li, and R. Wang, "Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions," *Acc. Chem. Res.*, vol. 50, no. 2, pp. 302–309, 2017.

[46] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, and D. Baker, "Protein structure determination using metagenome sequence data," *Science*, vol. 355, no. 6322, pp. 294–298, 2017.
•• Using contact predictions based on metagenome sequencing data, Baker and colleagues use Rosetta to predict the structure of 614 protein families with currently unknown structure.

[47] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," *J. Chem. Phys.*, vol. 139, no. 1, 2013.

[48] C. R. Schwantes and V. S. Pande, "Improvements in Markov State Model construction reveal many non-native interactions in the folding of NTL9," *J. Chem. Theory Comput.*, vol. 9, no. 4, pp. 2000–2009, 2013.

[49] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, "Essential dynamics of proteins," *Proteins Struct. Funct. Bioinforma.*, vol. 17, no. 4, pp. 412–425, 1993.

[50] O. F. Lange and H. Grubmüller, "Can principal components yield a dimension reduced description of protein dynamics on long time scales?," *J. Phys. Chem. B*, vol. 110, no. 45, pp. 22842–22852, 2006.

[51] C. C. David and D. J. Jacobs, "Principal component analysis: A method for determining the essential dynamics of proteins," *Methods Mol. Biol.*, vol. 1084, pp. 193–226, 2014.

[52] J. H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schtte, and F. Noé, "Markov models of molecular kinetics: Generation and validation," *J. Chem. Phys.*, vol. 134, no. 17, 2011.

[53] N. Singhal, C. D. Snow, and V. S. Pande, "Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin," *J. Chem. Phys.*, vol. 121, no. 1, pp. 415–425, 2004.

[54] A. C. Pan and B. Roux, "Building Markov state models along pathways to determine free energies and rates of transitions," *J. Chem. Phys.*, vol. 129, no. 6, 2008.

[55] S. Olsson, H. Wu, F. Paul, C. Clementi, and F. Noé, "Combining experimental and simulation data of molecular processes via augmented Markov models," *Proc. Natl. Acad. Sci.*, vol. 114, no. 31, pp. 8265–8270, 2017.
• This paper introduces a statistically rigorous method to combine information from molecular simulations and experimental data into augmented Markov models.

[56] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Mol. Syst. Biol.*, vol. 12, no. 7, p. 878, 2016.

[57] G. E. Dahl, N. Jaitly, and R. Salakhutdinov, "Multi-task Neural Networks for QSAR Predictions," *arXiv*, pp. 1–21, 2014.

[58] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv*, pp. 1–14, 2013.

[59] R. Gómez-Bombarelli, D. Duvenaud, J. M. Hernández-Lobato, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," *arXiv*, pp. 1–28, 2016.

[60] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity Prediction using Deep Learning," *Front. Environ. Sci.*, vol. 3, 2016.
•• Mayr and coworkers present DeepTox, a deep learning approach to predict toxicity that they used to win the Tox21 data challenge.

[61] DeepChem, "Deepchem, a python library democratizing deep learning for science." URL: http://www.deepchem.io Accessed: 2017-09-21.

[62] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A Benchmark for Molecular Machine Learning," *arXiv*, pp. 1–39, 2017.

[63] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning.," *Nature*, vol. 521, no. 7553, pp. 436–44, 2015.

[64] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations," *2016 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016.

[65] C. N. Silla and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Min. Knowl. Discov.*, vol. 22, no. 1-2, pp. 31–72, 2011.

[66] J. Jiménez, S. Doerr, G. Martínez-Rosell, A. S. Rose, and G. De Fabritiis, "DeepSite: protein-binding site predictor using 3D-convolutional neural networks," *Bioinformatics*, vol. 33, no. 19, pp. 3036–3042, 2017.

[67] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, "Protein-Ligand Scoring with Convolutional Neural Networks," *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 942–957, 2017.
• The article presents a convolutional neural network to rank binding poses that uses a 3D representation of the protein-ligand structure as input.

[68] J. Gomes, B. Ramsundar, E. N. Feinberg, and V. S. Pande, "Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity," *arXiv*, pp. 1–17, 2017.

[69] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost," *Chem. Sci.*, vol. 8, no. 4, pp. 3192–3203, 2017.

[70] J. Han, L. Zhang, R. Car, and W. E, "Deep Potential: a general representation of a many-body potential energy surface," *arXiv*, pp. 1–13, 2017.
• The study introduces Deep Potential, a deep neural network trained with QM simulations in order to learn a representation of the potential energy surface in the same way empirical forcefields do.

[71] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, "Prediction

errors of molecular machine learning models lower than hybrid DFT error," *J. Chem. Theory Comput.*, 2017. **Just Accepted Manuscript**.

[72] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

Chapter 4

DISCUSSION

In this chapter we discuss some of the results obtained in this thesis, as well as challenges and future work to be performed.

4.1 MD-driven fragment screening

In Publication 3.3 we performed the first MD-driven screening of more than 150 fragments against the chemokine CXCL12. The analysis of hundreds of short MD simulations with a MSM framework allowed us to predict binding sites, kinetics and binding free energy. Overall, the MSM results were satisfying and demonstrate its utility to analyze MD simulations. However, there are several critical points in the protocol of Publication 3.3 that need to be regarded and likely corrected in upcoming work. It is important to note that these issues do not compromise the validity of the work but would enhance the quality of the results in future applications of the protocol.

First, there is a need for a correct and accurate selection of a fragment library for screening. Our lack of experience in the field of virtual screening made us take specific decisions, some of which were sub-optimal. For instance, while picking negatively-charged residues was a good decision, performing docking against the whole CXCL12 surface is a sub-optimal strategy. Instead, we should have picked a number of pockets, from exper-

imental structures or from MD simulations, and screen the library against each pocket individually. This would exclude the possibility of docking ligands in very flexible regions, which we already knew *a priori*, such as the N-terminal tail of CXCL12.

Second, the binding pose of the fragments was, for several fragments, blurry (i.e. was not defined, more like a cloud of structures) and future development should be focused on automatically identifying, inside the *bound* macrostate, the most stable and defined micro-states to ease the human visualization of the binding pose. However, we must take into consideration that small fragments are very promiscuous and prone to move if the protein structure that accommodates them is also flexible, such as in the case of arginine 20, in the sY7 pocket. In such cases, the observed cloudy binding pose would be consistent with the expected behavior of the fragment.

Finally, the parameterization of the fragments must be accurate and, possibly, automated. The enormous amount of ligands we had to parameterize for Publication 3.3 pushed us to look for a compromise solution between automatic parameterization using GAFF or CGENFF and the manual work of running QM simulations for the dihedral angles and manually fitting parameters. The solution we decided to use is GAAMP [83], an automatic tool for parameterization that runs QM scans along the dihedral angles and fits them automatically. The visual inspection of the *pre-* and *post-* QM parameters was generally satisfactory. However, GAAMP, as a black-box procedure, is really susceptible to introduce, if any, unnoticed errors. Therefore, work should focus on making these automatic tools as reliable as possible while requiring the least human intervention possible. Finally, just mentioning that even if automatic parameterization was perfect, there are phenomena like polarizability, protonation changes or tautomerization that are not regarded by standard force-fields and may have a big impact in the correct description of the protein-ligand interactions. Future adoption of polarizable force-fields or constant-pH simulations could address some of these issues.

4.2 Benzene binding as a proxy for cryptic pocket detection

In Publication 3.2 we used MD simulations of protein solvated in mixed-solvent of water and benzene to detect cryptic pockets based on benzene binding. We tested our protocol on 18 different systems and proved that the protocol is able to detect cryptic pockets based on (a) the detection of benzene binding hot-spots and (b) the results of a community-based score function. We compared our protocol to other non-cryptic pocket detector algorithms, *DeepSite* and *fpocket*, and, while our method was the most accurate, *fpocket* was suprisingly good for the metric used and at a fraction of computational time.

However, one of the advantages of our technique, *CryptoScout*, in comparison to methods such as *fpocket*, is the structural description of the pocket opening. In fact, the identification of a cryptic cavity alone results quite useless without a structural description of how the cavity looks when is open. The knowledge about the presence of a cryptic cavity can be leveraged by experimental techniques such as *tethering*, that can design protein mutants with cysteines on the predicted cryptic pocket and test a library of disulfide-containing fragments against that pocket expecting the formation of a fragment-cysteine disulfide bridge (Fig. 4.1). However, in order to apply additional *in silico* techniques, such as docking, is really important to decipher the structure of the open cryptic pocket and, in particular, in a conformation that allows the successful docking of the ligands.

For this reason, we assessed whether the binding of benzene could actually sample the pocket opening in a conformation compatible with docking by comparing conformations extracted from the MD simulations to the *holo* conformation, this is the conformation of the protein when a ligand is binding the cryptic cavity. We performed a PCA analysis of the distances between the residues forming the cryptic pocket and observed that, in some of the systems, the simulations reach pocket inter-residue distances corresponding with the *holo* conformation. This fact is further assessed by applying docking of the *holo* ligand to both the *apo* structure

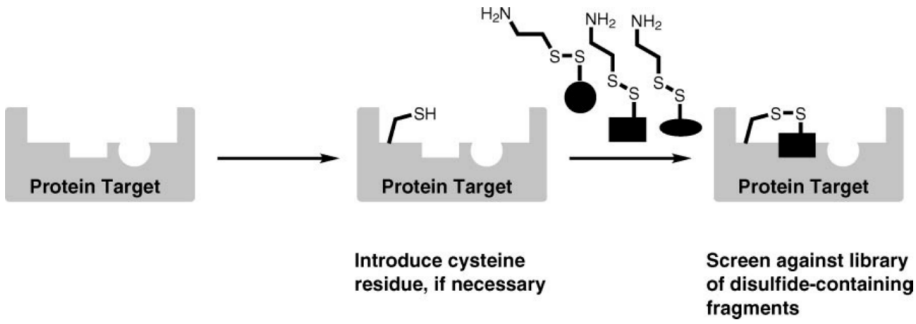


Figure 4.1: Overview of how tethering technique works. Extracted from [139].

and 10 representative frames extracted from the simulations. By doing so, we were able to reconstruct 10 out 17 ligands binding poses within 5$\mathring{A}$ RMSD when using the 10 frames as protein structure compared to only 4 out of 17 when using the *apo* conformation.

However, the question remains unanswered: can we know *a priori* which exact conformation should we use in docking? One could argue that knowing the position of the cryptic pocket one could run further MD simulations using an adaptive sampling scheme to explore thoroughly the conformational space of the pocket. However, should we run the simulations with or without benzene? The answer to this question lies in the eternal debate about whether the protein-ligand binding mechanism is due to the *conformational selection* paradigm or the *induced fit* paradigm. If the first paradigm is the case, then simulation of the protein alone should already be able to sample the pocket opening. If the case is the second paradigm, simulations with benzene may be a possible approximation although probably in most cases we would need the actual ligand or the ligand moiety responsible for the induction of the protein binding conformation. As it is likely that both paradigms may be true in a system-specific manner, the best approximation, arguably, would be to run the protein with and without benzene, sample pocket conformations from both simulations and dock the ligands to the conformation ensemble hoping that

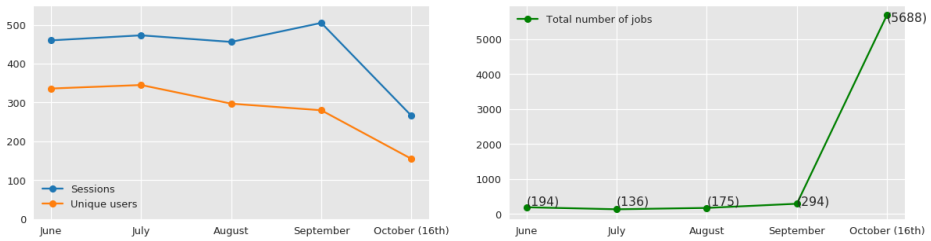


Figure 4.2: Statistics of *PlayMolecule* platform usage. Obtained using Google Analytics and data from the local database.

the ligand binding is mostly due to enthalpic energy (opposed to entropic) and the docking scoring function is good enough at approximating the interaction energy.

4.3 *PlayMolecule*: a web infrastructure for supporting drug discovery

One of the main contributions of this doctorate has been to transfer know-how and applications developed in the research group to a web platform that exposes those services to be used by the scientific community and, eventually, support the development of better drugs.

The platform, called *PlayMolecule*, was launched in June 2017 and, since then, it has had an affluence of an average of 300 unique users per month, as well as a continuous growth of number of jobs launched, which indicates the raising interest and trust by the user community (Fig. 4.2). Furthermore, at the time of writing, more than 40 users have registered into the platform, including at least 3 big pharma companies, several important principal investigators and PhD students from all over the world.

Chapter 5

CONCLUSIONS

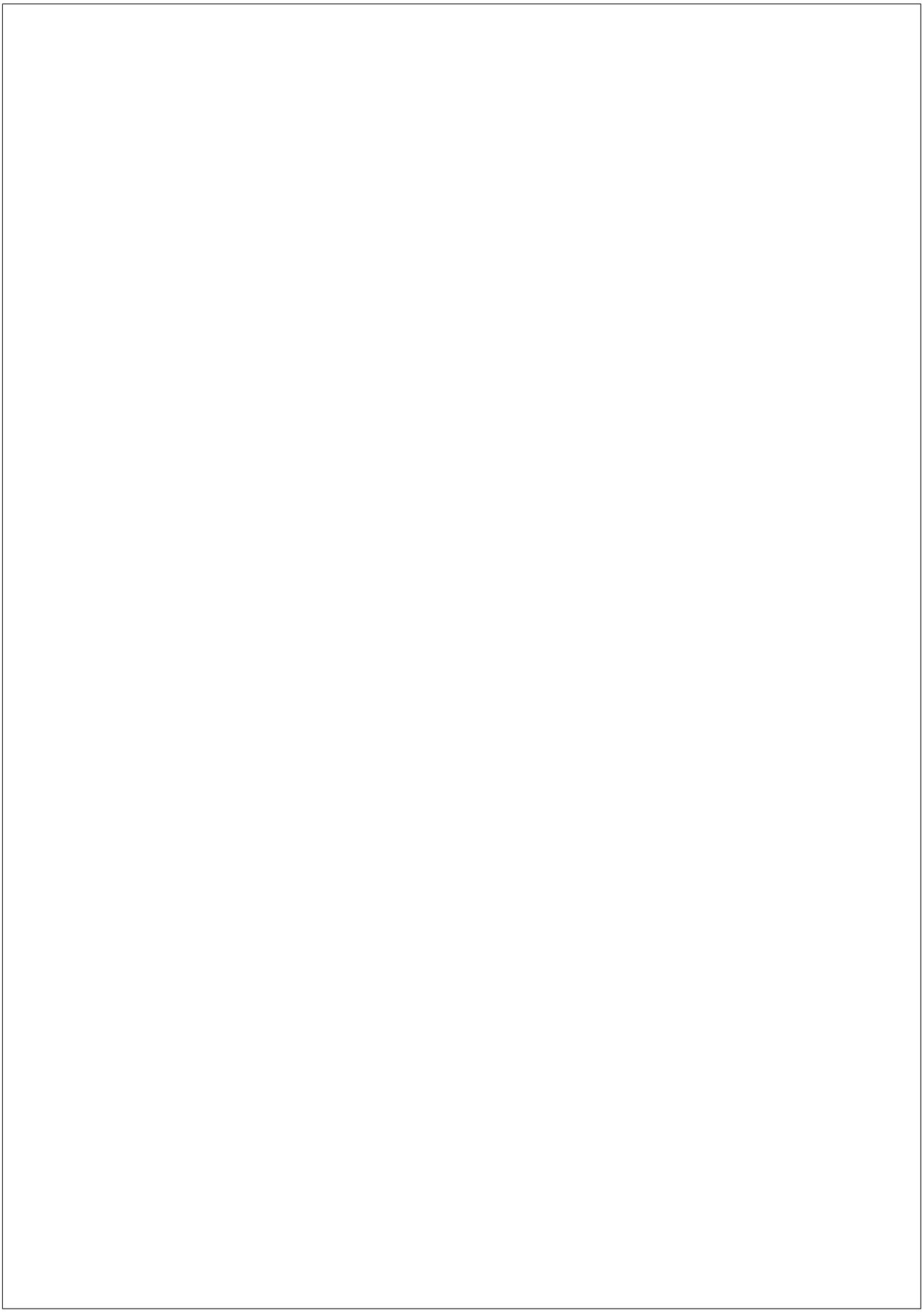
1. High-throughput MD simulations are able to capture protein-ligand binding events and can be applied in fragment-based drug discovery to screen a library of fragments

2. The chemokine CXCL12 possesses two binding pockets, termed sY7 and H1S68, where we predicted that small compounds consisting of a hydrophobic core and a negatively-charged group could bind

3. Mixed-solved simulations of a protein solvated in water and benzene are able to identify cryptic pockets and capture the molecular mechanism of pocket opening based on benzene binding

4. The molecular simulation of the μ-opioid receptor bound to a ligand is able to capture the dynamic and kinetic behavior of the receptor and can be used to rationalize GPCR functional selectivity

5. The transfer and implementation of applications in the web platform *PlayMolecule* is an approach to broaden the accessibility and applicability scope of the generated know-how and has been well received by the scientific community



Chapter 6

APPENDIX: OTHER PUBLICATIONS

6.1 DeepSite: Protein binding site predictor using 3D-convolutional neural networks

Jose Jiménez, Stefan Doerr, Gerard Martínez-Rosell, Alexander Rose, Gianni de Fabritiis. *Bioinformatics.* **2017** Oct 1;33(19):3036-3042.
doi: 10.1093/bioinformatics/btx350.

Summary
DeepSite is a novel method for binding pocket detection leveraging convolutional neural networks (CNN) trained with the scPDB database, a database of protein-ligand structure complexes. In order to train the CNN, the experimentally-resolved binding pockets are featurized by creating 3D maps of chemical properties such as atom occupancies, H-bond donors, H-bond acceptors and aromaticity. The feature maps of binding pockets are then used to train a CNN, that learns to differentiate feature maps of binding sites from non-binding sites. In prediction mode, a *query* protein structure is segmented into overlapping boxes and for each box the feature maps are calculated. Then, the feature maps are fed into the pre-trained CNN, which returns a probability of containing a binding site. Finally,

an iso-surface of the probability of containing a binding site is generated from the overlapping boxes all around the protein and binding site centers are calculated by clustering the probabilities. The application has been made available free of charge as part of the *PlayMolecule* suite of apps (www.playmolecule.org/deepsite/).

Bibliography

[1] Drews J. Drug Discovery: A Historical Perspective. Science. 2000 Mar;287(5460):1960–1964. Available from: http://science.sciencemag.org/content/287/5460/1960.

[2] Langley JN. On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curari. The Journal of Physiology. 1905 Dec;33(4-5):374–413. Available from: http://onlinelibrary.wiley.com/doi/10.1113/jphysiol.1905.sp001128/abstract.

[3] Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges. 1894 Oct;27(3):2985–2993. Available from: http://onlinelibrary.wiley.com/doi/10.1002/cber.18940270364/abstract.

[4] Koshland DE. Application of a Theory of Enzyme Specificity to Protein Synthesis. Proc Natl Acad Sci U S A. 1958 Feb;44(2):98–104. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC335371/.

[5] Monod J, Wyman J, Changeux JP. On the nature of allosteric transitions: a plausible model. J Mol Biol. 1965 May;12:88–118.

[6] Changeux JP, Edelstein S. Conformational selection or induced fit? 50 years of debate resolved. F1000 Biol Rep. 2011 Sep;3.

Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169905/.

[7] Vogt AD, Pozzi N, Chen Z, Di Cera E. Essential role of conformational selection in ligand binding. Biophys Chem. 2014 Feb;186:13–21.

[8] Csermely P, Palotai R, Nussinov R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci. 2010 Oct;35(10):539–546.

[9] Silva DA, Bowman GR, Sosa-Peinado A, Huang X. A Role for Both Conformational Selection and Induced Fit in Ligand Binding by the LAO Protein. PLOS Computational Biology. 2011 May;7(5):e1002054. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002054.

[10] Ban TA. The role of serendipity in drug discovery. Dialogues Clin Neurosci. 2006 Sep;8(3):335–344. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181823/.

[11] Fleming A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenza. Br J Exp Pathol. 1929 Jun;10(3):226–236. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048009/.

[12] Murray GR. The life-history of the first case of myxoedema treated by thyroid extract. Br Med J. 1920 Mar;1(3089):359–360.

[13] Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. Nat Rev Drug Discov. 2011;10(3):188–195.

[14] Cheng T, Li Q, Zhou Z, Wang Y, Bryant SH. Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review. AAPS J. 2012 Jan;14(1):133–141. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3282008/.

[15] Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H. Twenty years on: the impact of fragments on drug discovery. Nat Rev Drug Discov. 2016 Sep;15(9):605–619. Available from: http://www.nature.com/nrd/journal/v15/n9/abs/nrd.2016.109.html.

[16] Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. Drug Discov Today. 2004 May;9(10):430–431.

[17] Murray CW, Rees DC. The rise of fragment-based drug discovery. Nat Chem. 2009 Jun;1(3):187–192.

[18] Ferenczy GG, Keseru GM. How are fragments optimized? A retrospective analysis of 145 fragment optimizations. J Med Chem. 2013 Mar;56(6):2478–2486.

[19] Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model. 2012 Nov;52(11):2864–2875.

[20] Erlanson DA, McDowell RS, O'Brien T. Fragment-based drug discovery. J Med Chem. 2004 Jul;47(14):3463–3482.

[21] Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. J Chem Inf Model. 2001 Jun;41(3):856–864.

[22] Leach AR, Hann MM. Molecular complexity and fragment-based drug discovery: ten years on. Curr Opin Chem Biol. 2011 Aug;15(4):489–496.

[23] Murray CW, Verdonk ML, Rees DC. Experiences in fragment-based drug discovery. Trends Pharmacol Sci. 2012 May;33(5):224–232.

[24] Kozakov D, Hall DR, Jehle S, Jehle S, Luo L, Ochiana SO, et al. Ligand deconstruction: Why some fragment binding positions are conserved and others are not. Proc Natl Acad Sci U S A. 2015 May;112(20):E2585–2594.

[25] Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958 Mar;181(4610):662–666.

[26] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan;28(1):235–242.

[27] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. Arch Biochem Biophys. 1978 Jan;185(2):584–591.

[28] RCSB. PDB Current Holdings Report; 2017.;. Available from: https://www.rcsb.org/pdb/statistics/holdings.do.

[29] Cavanagh J, Fairbrother WJ, III AGP, Skelton NJ. Protein NMR Spectroscopy: Principles and Practice. Academic Press; 1995.

[30] Dias DM, Ciulli A. NMR approaches in structure-based lead discovery: Recent developments and new frontiers for targeting multi-protein complexes. Prog Biophys Mol Biol. 2014 Nov;116(2-3):101–112. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4261069/.

[31] Dror RO, Dirks RM, Grossman JP, Xu H, Shaw DE. Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys. 2012;41:429–452.

[32] Roy R, Hohng S, Ha T. A practical guide to single-molecule FRET. Nat Methods. 2008 Jun;5(6):507–516.

[33] Mortier J, Rakers C, Frederick R, Wolber G. Computational tools for in silico fragment-based drug design. Curr Top Med Chem. 2012;12(17):1935–1943.

[34] Cavasotto CN, Phatak SS. Homology modeling in drug discovery: current trends and applications. Drug Discov Today. 2009 Jul;14(13-14):676–683.

[35] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem. 2010 Jan;31(2):455–461. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041641/.

[36] Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. J Med Chem. 2004 Mar;47(7):1739–1749. Available from: http://dx.doi.org/10.1021/jm0306430.

[37] Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997 Apr;267(3):727–748.

[38] Fischer M, Coleman RG, Fraser JS, Shoichet BK. Incorporation of protein flexibility and conformational energy penalties in docking screens to improve ligand discovery. Nat Chem. 2014 Jul;6(7):575–583.

[39] Chen Y, Shoichet BK. Molecular docking and ligand specificity in fragment-based inhibitor discovery. Nat Chem Biol. 2009 May;5(5):358–364.

[40] Davis BJ, Erlanson DA. Learning from our mistakes: the 'unknown knowns' in fragment screening. Bioorg Med Chem Lett. 2013 May;23(10):2844–2852.

[41] Zoete V, Grosdidier A, Michielin O. Docking, virtual high throughput screening and in silico fragment-based drug design. J Cell Mol Med. 2009 Feb;13(2):238–248. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3823351/.

[42] Alford RF, Leaver-Fay A, Jeliazkov JR, O'Meara MJ, DiMaio FP, Park H, et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput. 2017 Jun;13(6):3031–3048.

[43] 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. 1976 May;103(2):227–249.

[44] Murphy RB, Philipp DM, Friesner RA. A mixed quantum mechanics/molecular mechanics (QM/MM) method for large-scale modeling of chemistry in protein environments. J Comput Chem. 2000 Dec;21(16):1442–1457. Available from: http://onlinelibrary.wiley.com/doi/10.1002/1096-987X(200012)21:16<1442::AID-JCC3>3.0.CO;2-O/abstract.

[45] Martinez-Rosell G, Giorgino T, Harvey MJ, de Fabritiis G. Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale. Curr Top Med Chem. 2017;17(23):2617–2625.

[46] Feynman RP, Leighton RB, Sands ML. The Feynman lectures on physics. Reading, Mass.: Addison-Wesley Pub. Co.; 1963. OCLC: 531535.

[47] Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol. 2002 Sep;9(9):646–652.

[48] Verlet L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. Phys Rev. 1967 Jul;159(1):98–103. Available from: https://link.aps.org/doi/10.1103/PhysRev.159.98.

[49] MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B. 1998 Apr;102(18):3586–3616.

[50] MacKerell AD, Banavali N, Foloppe N. Development and current status of the CHARMM force field for nucleic acids. Biopolymers. 2000;56(4):257–265.

[51] Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, et al. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc. 1995 May;117(19):5179–5197. Available from: http://dx.doi.org/10.1021/ja00124a002.

[52] Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, Zhang W, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem. 2003 Dec;24(16):1999–2012.

[53] Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J Am Chem Soc. 1996 Jan;118(45):11225–11236. Available from: http://dx.doi.org/10.1021/ja9621760.

[54] Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, et al. Optimization of the Additive CHARMM All-Atom Protein Force Field

Targeting Improved Sampling of the Backbone phi, psi and Side-Chain chi1 and chi2 Dihedral Angles. J Chem Theory Comput. 2012 Sep;8(9):3257–3273. Available from: http://dx.doi.org/10.1021/ct300400x.

[55] Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015 Aug;11(8):3696–3713. Available from: http://dx.doi.org/10.1021/acs.jctc.5b00255.

[56] Robertson MJ, Tirado-Rives J, Jorgensen WL. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. J Chem Theory Comput. 2015 Jul;11(7):3499–3509. Available from: http://dx.doi.org/10.1021/acs.jctc.5b00356.

[57] Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophys J. 2011 May;100(9):L47–49.

[58] Jakalian A, Jack DB, Bayly CI. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J Comput Chem. 2002 Dec;23(16):1623–1641.

[59] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 2004 Jul;25(9):1157–1174.

[60] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010 Mar;31(4):671–690. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2888302/.

[61] Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol. 2011 Oct;9:71.

[62] Harvey MJ, De Fabritiis G. High-throughput molecular dynamics: the powerful new tool for drug discovery. Drug Discov Today. 2012 Oct;17(19-20):1059–1062.

[63] Shaw DE, Chao JC, Eastwood MP, Gagliardo J, Grossman JP, Ho CR, et al. Anton, a special-purpose machine for molecular dynamics simulation. Communications of the ACM. 2008 Jul;51(7):91. Available from: http://portal.acm.org/citation.cfm?doid=1364782.1364802.

[64] Harvey MJ, Giupponi G, Fabritiis GD. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. J Chem Theory Comput. 2009 Jun;5(6):1632–1639.

[65] Salomon-Ferrer R, Gotz AW, Poole D, Le Grand S, Walker RC. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J Chem Theory Comput. 2013 Sep;9(9):3878–3888.

[66] Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput Biol. 2017 Jul;13(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5549999/.

[67] Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. SC '06. New York, NY, USA: ACM; 2006. Available from: http://doi.acm.org/10.1145/1188455.1188544.

[68] Buch I, Harvey MJ, Giorgino T, Anderson DP, De Fabritiis G. High-throughput all-atom molecular dynamics simulations using

distributed computing. J Chem Inf Model. 2010 Mar;50(3):397–403.

[69] Sadiq SK, De Fabritiis G. Explicit solvent dynamics and energetics of HIV-1 protease flap opening and closing. Proteins. 2010 Nov;78(14):2873–2885.

[70] Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. Proc Natl Acad Sci USA. 2011 Jun;108(25):10184–10189.

[71] Sadiq SK, Noe F, De Fabritiis G. Kinetic characterization of the critical step in HIV-1 protease maturation. Proc Natl Acad Sci U S A. 2012 Dec;109(50):20449–20454. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3528573/.

[72] Perez-Hernandez G, Paul F, Giorgino T, De Fabritiis G, Noe F. Identification of slow molecular order parameters for Markov model construction. J Chem Phys. 2013 Jul;139(1):015102.

[73] Stanley N, Esteban-Martin S, Fabritiis GD. Kinetic modulation of a disordered protein domain by phosphorylation. Nature Communications. 2014 Oct;5:ncomms6272. Available from: https://www.nature.com/articles/ncomms6272.

[74] Ferruz N, Harvey MJ, Mestres J, De Fabritiis G. Insights from Fragment Hit Binding Assays by Molecular Simulations. J Chem Inf Model. 2015 Oct;55(10):2200–2205.

[75] Stanley N, Pardo L, Fabritiis GD. The pathway of ligand entry from the membrane bilayer to a lipid G protein-coupled receptor. Sci Rep. 2016 Mar;6:22639.

[76] McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977 Jun;267(5612):585–590.

[77] Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. J Chem Phys. 2009 Sep;131(12):124101.

[78] Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982 Mar;28(2):129–137.

[79] Deuflhard P, Weber M. Robust Perron cluster analysis in conformation dynamics. Linear Algebra and its Applications. 2005 Mar;398(Supplement C):161–184. Available from: http://www.sciencedirect.com/science/article/pii/S0024379504004689.

[80] Doerr S, Harvey MJ, Noe F, De Fabritiis G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. J Chem Theory Comput. 2016 Apr;12(4):1845–1852. Available from: http://dx.doi.org/10.1021/acs.jctc.6b00049.

[81] Scherer MK, Trendelkamp-Schroer B, Paul F, Perez-Hernandez G, Hoffmann M, Plattner N, et al. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. J Chem Theory Comput. 2015 Nov;11(11):5525–5542. Available from: http://dx.doi.org/10.1021/acs.jctc.5b00743.

[82] Doerr S, De Fabritiis G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. J Chem Theory Comput. 2014 May;10(5):2064–2069.

[83] Huang L, Roux B. Automated force field parameterization for non-polarizable and polarizable atomic models based on ab initio target data. J Chem Theory Comput. 2013 Aug;9(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3819940/.

[84] Baker CM. Polarizable force fields for molecular dynamics simulations of biomolecules. WIREs Comput Mol Sci. 2015 Mar;5(2):241–254. Available from: http://onlinelibrary.wiley.com/doi/10.1002/wcms.1215/abstract.

[85] Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, et al. The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. J Chem Theory Comput. 2013;9(9):4046–4063.

[86] Donnini S, Ullmann RT, Groenhof G, Grubmuller H. Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer. J Chem Theory Comput. 2016 Mar;12(3):1040–1051.

[87] Senn HM, Thiel W. QM/MM methods for biomolecular systems. Angew Chem Int Ed Engl. 2009;48(7):1198–1229.

[88] Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation - Umbrella sampling. Journal of Computational Physics. 1977 Feb;23:187–199. Available from: http://adsabs.harvard.edu/abs/1977JCoPh..23..187T.

[89] Laio A, Gervasio FL. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep Prog Phys. 2008;71(12):126601. Available from: http://stacks.iop.org/0034-4885/71/i=12/a=126601.

[90] Fidelak J, Juraszek J, Branduardi D, Bianciotto M, Gervasio FL. Free-energy-based methods for binding profile determination in a congeneric series of CDK2 inhibitors. J Phys Chem B. 2010 Jul;114(29):9516–9524.

[91] Selent J, Sanz F, Pastor M, Fabritiis GD. Induced Effects of Sodium Ions on Dopaminergic G-Protein Coupled Receptors. PLOS

Computational Biology. 2010 Aug;6(8):e1000884. Available from: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000884.

[92] Giorgino T, De Fabritiis G. A High-Throughput Steered Molecular Dynamics Study on the Free Energy Profile of Ion Permeation through Gramicidin A. J Chem Theory Comput. 2011 Jun;7(6):1943–1950.

[93] Giorgino T, Buch I, De Fabritiis G. Visualizing the Induced Binding of SH2-Phosphopeptide. J Chem Theory Comput. 2012 Apr;8(4):1171–1175.

[94] Buch I, Ferruz N, De Fabritiis G. Computational Modeling of an Epidermal Growth Factor Receptor Single-Mutation Resistance to Cetuximab in Colorectal Cancer Treatment. J Chem Inf Model. 2013 Dec;53(12):3123–3126. Available from: http://pubs.acs.org/doi/abs/10.1021/ci400456m.

[95] Sabbadin D, Ciancetta A, Moro S. Bridging molecular docking to membrane molecular dynamics to investigate GPCR-ligand recognition: the human A2A adenosine receptor as a key study. J Chem Inf Model. 2014 Jan;54(1):169–183.

[96] Ferruz N, Tresadern G, Pineda-Lucena A, Fabritiis GD. Multibody cofactor and substrate molecular recognition in the *myo*-inositol monophosphatase enzyme. Sci Rep. 2016 Jul;6:srep30275. Available from: https://www.nature.com/articles/srep30275.

[97] Plattner N, Doerr S, De Fabritiis G, NoÃ© F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. Nat Chem. 2017 Oct;9(10):1005–1011. Available from: https://www.nature.com/nchem/journal/v9/n10/abs/nchem.2785.html.

[98] Desaphy J, Bret G, Rognan D, Kellenberger E. sc-PDB: a 3D-database of ligandable binding sites-10 years on. Nucleic Acids Res. 2015 Jan;43(Database issue):D399–D404. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384012/.

[99] Rose AS, Hildebrand PW. NGL Viewer: a web application for molecular visualization. Nucleic Acids Res. 2015 Jul;43(Web Server issue):W576–W579. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489237/.

[100] Zlotnik A, Yoshie O. The chemokine superfamily revisited. Immunity. 2012 May;36(5):705–716.

[101] Karpova D, Bonig H. Concise Review: CXCR4/CXCL12 Signaling in Immature Hematopoiesis–Lessons From Pharmacological and Genetic Models. Stem Cells. 2015 Aug;33(8):2391–2399.

[102] Karin N. The multiple faces of CXCL12 (SDF-1alpha) in the regulation of immunity during health and disease. J Leukoc Biol. 2010 Sep;88(3):463–473.

[103] Sun X, Cheng G, Hao M, Zheng J, Zhou X, Zhang J, et al. CXCL12 / CXCR4 / CXCR7 chemokine axis and cancer progression. Cancer Metastasis Rev. 2010 Dec;29(4):709–722.

[104] Domanska UM, Kruizinga RC, Nagengast WB, Timmer-Bosscha H, Huls G, de Vries EGE, et al. A review on CXCR4/CXCL12 axis in oncology: no place to hide. Eur J Cancer. 2013 Jan;49(1):219–230.

[105] Chatterjee S, Azad BB, Nimmagadda S. The Intricate Role of CXCR4 in Cancer. Adv Cancer Res. 2014;124:31–82. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4322894/.

[106] Patrussi L, Baldari CT. The CXCL12/CXCR4 axis as a therapeutic target in cancer and HIV-1 infection. Curr Med Chem. 2011;18(4):497–512.

[107] Arenzana-Seisdedos F. SDF-1/CXCL12: A Chemokine in the Life Cycle of HIV. Front Immunol. 2015 Jun;6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4456947/.

[108] Choi WT, Yang Y, Xu Y, An J. Targeting Chemokine Receptor CXCR4 for Treatment of HIV-1 Infection, Tumor Progression, and Metastasis. Curr Top Med Chem. 2014;14(13):1574–1589. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4372248/.

[109] Veldkamp CT, Ziarek JJ, Peterson FC, Chen Y, Volkman BF. Targeting SDF-1/CXCL12 with a ligand that prevents activation of CXCR4 through structure-based drug design. J Am Chem Soc. 2010 Jun;132(21):7242–7243.

[110] Veldkamp CT, Seibert C, Peterson FC, Sakmar TP, Volkman BF. Recognition of a CXCR4 sulfotyrosine by the chemokine stromal cell-derived factor-1alpha (SDF-1alpha/CXCL12). J Mol Biol. 2006 Jun;359(5):1400–1409.

[111] Veldkamp CT, Seibert C, Peterson FC, De la Cruz NB, Haugner JC, Basnet H, et al. Structural basis of CXCR4 sulfotyrosine recognition by the chemokine SDF-1/CXCL12. Sci Signaling. 2008 Sep;1(37):ra4.

[112] Ziarek JJ, Getschman AE, Butler SJ, Taleski D, Stephens B, Kufareva I, et al. Sulfopeptide probes of the CXCR4/CXCL12 interface reveal oligomer-specific contacts and chemokine allostery. ACS Chem Biol. 2013 Sep;8(9):1955–1963.

[113] Ziarek JJ, Veldkamp CT, Zhang F, Murray NJ, Kartz GA, Liang X, et al. Heparin oligosaccharides inhibit chemokine (CXC motif)

ligand 12 (CXCL12) cardioprotection by binding orthogonal to the dimerization interface, promoting oligomerization, and competing with the chemokine (CXC motif) receptor 4 (CXCR4) N terminus. J Biol Chem. 2013 Jan;288(1):737–746.

[114] Ziarek JJ, Liu Y, Smith E, Zhang G, Peterson FC, Chen J, et al. Fragment-based optimization of small molecule CXCL12 inhibitors for antagonizing the CXCL12/CXCR4 interaction. Curr Top Med Chem. 2012;12(24):2727–2740.

[115] Smith EW, Liu Y, Getschman AE, Peterson FC, Ziarek JJ, Li R, et al. Structural analysis of a novel small molecule ligand bound to the CXCL12 chemokine. J Med Chem. 2014 Nov;57(22):9693–9699.

[116] Smith EW, Nevins AM, Qiao Z, Liu Y, Getschman AE, Vankayala SL, et al. Structure-Based Identification of Novel Ligands Targeting Multiple Sites within a Chemokine-G-Protein-Coupled-Receptor Interface. J Med Chem. 2016 May;59(9):4342–4351.

[117] Steglitz J, Buscemi J, Jean Ferguson M. The future of pain research, education, and treatment: A summary of the IOM report "Relieving pain in America: A blueprint for transforming prevention, care, education, and research". Translational behavioral medicine. 2012 Mar;2:6–8.

[118] Corbett AD, Henderson G, McKnight AT, Paterson SJ. 75 years of opioid research: the exciting but vain quest for the Holy Grail. Br J Pharmacol. 2006 Jan;147 Suppl 1:S153–162.

[119] Pasternak GW, Pan YX. Mu Opioids and Their Receptors: Evolution of a Concept. Pharmacol Rev. 2013 Oct;65(4):1257–1317. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3799236/.

[120] Compton WM, Jones CM, Baldwin GT. Relationship between Nonmedical Prescription-Opioid Use and Heroin Use. N Engl J Med. 2016 Jan;374(2):154–163.

[121] Kieffer BL. Opioids: first lessons from knockout mice. Trends Pharmacol Sci. 1999 Jan;20(1):19–26.

[122] Raehal KM, Walker JKL, Bohn LM. Morphine side effects in beta-arrestin 2 knockout mice. J Pharmacol Exp Ther. 2005 Sep;314(3):1195–1201.

[123] Bohn LM, Lefkowitz RJ, Gainetdinov RR, Peppel K, Caron MG, Lin FT. Enhanced morphine analgesia in mice lacking beta-arrestin 2. Science. 1999 Dec;286(5449):2495–2498.

[124] Maguma HT, Dewey WL, Akbarali HI. Differences in the characteristics of tolerance to mu-opioid receptor agonists in the colon from wild type and beta-arrestin2 knockout mice. Eur J Pharmacol. 2012 Jun;685(1-3):133–140.

[125] Manglik A, Lin H, Aryal DK, McCorvy JD, Dengler D, Corder G, et al. Structure-based discovery of opioid analgesics with reduced side effects. Nature. 2016;537(7619):185–190.

[126] DeWire SM, Yamashita DS, Rominger DH, Liu G, Cowan CL, Graczyk TM, et al. A G protein-biased ligand at the mu-opioid receptor is potently analgesic with reduced gastrointestinal and respiratory dysfunction compared with morphine. J Pharmacol Exp Ther. 2013 Mar;344(3):708–717.

[127] Soergel DG, Subach RA, Burnham N, Lark MW, James IE, Sadler BM, et al. Biased agonism of the mu-opioid receptor by TRV130 increases analgesia and reduces on-target adverse effects versus morphine: A randomized, double-blind, placebo-controlled, crossover study in healthy volunteers. Pain. 2014 Sep;155(9):1829–1835.

[128] Soergel DG, Subach RA, Sadler B, Connell J, Marion AS, Cowan CL, et al. First clinical experience with TRV130: pharmacokinetics and pharmacodynamics in healthy volunteers. J Clin Pharmacol. 2014 Mar;54(3):351–357.

[129] Viscusi ER, Webster L, Kuss M, Daniels S, Bolognese JA, Zuckerman S, et al. A randomized, phase 2 study investigating TRV130, a biased ligand of the mu-opioid receptor, for the intravenous treatment of acute pain. Pain. 2016 Jan;157(1):264–272.

[130] Manglik A, Kruse AC, Kobilka TS, Thian FS, Mathiesen JM, Sunahara RK, et al. Crystal structure of the mu-opioid receptor bound to a morphinan antagonist. Nature. 2012 Mar;485(7398):321–326.

[131] Huang W, Manglik A, Venkatakrishnan AJ, Laeremans T, Feinberg EN, Sanborn AL, et al. Structural insights into mu-opioid receptor activation. Nature. 2015 Aug;524(7565):315–321. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4639397/.

[132] Vogel R, Mahalingam M, Ludeke S, Huber T, Siebert F, Sakmar TP. Functional role of the "ionic lock"–an interhelical hydrogen-bond network in family A heptahelical receptors. J Mol Biol. 2008 Jul;380(4):648–655.

[133] Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, et al. Crystal structure of rhodopsin: A G protein-coupled receptor. Science. 2000 Aug;289(5480):739–745.

[134] Ballesteros JA, Jensen AD, Liapakis G, Rasmussen SGF, Shi L, Gether U, et al. Activation of the beta2-Adrenergic Receptor Involves Disruption of an Ionic Lock between the Cytoplasmic Ends of Transmembrane Segments 3 and 6. J Biol Chem. 2001 Aug;276(31):29171–29177. Available from: http://www.jbc.org/content/276/31/29171.

[135] Dror RO, Arlow DH, Borhani DW, Jensen M, Piana S, Shaw DE. Identification of two distinct inactive conformations of the beta2-adrenergic receptor reconciles structural and biochemical observations. Proc Natl Acad Sci USA. 2009 Mar;106(12):4689–4694.

[136] Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SGF, Thian FS, Kobilka TS, et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science. 2007 Nov;318(5854):1258–1265.

[137] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. 2012 Jan;40(Database issue):D370–D376. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245162/.

[138] Cimermancic P, Weinkam P, Rettenmaier TJ, Bichmann L, Keedy DA, Woldeyes RA, et al. CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. J Mol Biol. 2016 Feb;428(4):709–719.

[139] Erlanson DA, Wells JA, Braisted AC. Tethering: Fragment-Based Drug Discovery. Annu Rev Biophys Biomol Struct. 2004;33(1):199–223. Available from: https://doi.org/10.1146/annurev.biophys.33.110502.140409.