Understanding ligand-receptor recognition by means of high-throughput molecular dynamics

A perspective for drug discovery

Noelia Ferruz Capapey

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI
Prof. Gianni De Fabritiis
Prof. Ferran Sanz Carreras

Departament de Ciències Experimentals i de la Salut

Universitat
Pompeu Fabra
Barcelona

‘‘How often have rage and pain made you cry? How often has exhaustion made you lose your memory, voice, and common sense? And how often in this state have you exclaimed, with a broad smile on your face, ‘The final stage! Two more hours! Go, onward, upward!’

That pain only exists inside your head. Control it, destroy it, eliminate it, and keep on’

- Kilian Jornet. The skyrunner manifesto.

Acknowledgements

It’s been more than three years and a half since I started this journey called PhD. Although not always easy and pleasant, (it is a PhD after all) I am happy that I finally and unexpectedly managed to wrap up. And I say unexpectedly because I could have not managed without the key influence of a few people.

I would like first to thank Gianni, my supervisor, who dealt with my alternating phases of overwhelming excitement, boredom and continuous indecision. Special thanks for having given me the opportunity to go to Boston. I would like to thank Stefan and Matt, for being always available and willing to provide computational support, and the other labmates and colleagues in Acellera who assisted me. I want also to thank the crunchers in GPUGRID, and the donors in AGAUR, for both making my research possible.

Thanks to my parents and brother, who always supported me.

And finally, many thanks to you, Ferran.

Abstract

Understanding how receptor-ligand interactions occur is a first step towards designing new drugs. The complete reconstruction of the binding process in a drug-receptor system provides all the physical-chemistry variables for rational design of inhibitors of a chosen target, an important step in drug discovery. Although very powerful, direct experimental observation of full binding processes is very hard to perform. In this thesis, by using high-throughput molecular dynamics in the distributed computing project GPUGRID.net and analysing the resulting data by Markov state models (MSM), we successfully estimated kinetics, thermodynamics and binding modes for different molecular systems. In the initial works, we focused on estimating the potency of inhibitor-protein complexes. In subsequent studies, we described more complex pictures of binding, taking into account the receptor dynamics or other binding molecules. The results are promising and establish the methodology as a very powerful tool in the first stages of the drug discovery pipeline.

Resumen

Comprender las interacciones entre proteína y ligando es el primer paso para diseñar nuevos medicamentos. Llegar a reconstruir completamente este proceso de unión proporciona todas las variables físico-químicas para una optimización racional, un paso muy importante en el descubrimiento de fármacos. Pese a que esto ofrece muchas ventajas, todavía es complicado observar estos procesos experimentalmente. En esta tesis, utilizando simulaciones moleculares de alto rendimiento (HTMD) mediante el proyecto distribuido GPUGRID.net y análisis por Markov *state models* (MSM), hemos obtenido datos cinéticos, termodinámicos y modos de unión para varios sistemas. En los primeros trabajos nos centramos en estimar la afinidad entre complejos inhibidor-proteína. En trabajos posteriores, logramos caracterizar completamente rutas de unión del ligando teniendo en cuenta los confórmeros de la proteína u otros ligandos presentes. Los resultados son prometedores y establecen la utilidad de HTMD en las primeras fases de descubrimiento de fármacos.

Preface

There has been a fair amount of publications, awards, and advances in the field of molecular dynamics (MD) in the last five years. Starting from the improvement and design of special-purpose machinery, publications in high-impact factor journals and ending with the Nobel Prize in Chemistry won in 2013 by Martin Karplus, Michael Levitt and Arieh Warshel for developing the method. All of them, and this latter in particular, exemplify how pervasive MD has become in solving all kinds of chemical problems.

Also five years ago was published the first quantitative reconstruction of a binding process for an enzyme–inhibitor system.[1] The publication was based on a benchmark system, with a small fragment binding to a rigid protein; however, it set the entire basis for an application to a broader scenario and became a hallmark study in the field. This thesis has focused with this particular matter, the application of this methodology to current problems in drug discovery.

Concretely, we have proved the feasibility of HTMD at three different stages of the pipeline. First, in compound screening, by means of approximated methods. Second, in hit identification, performing an *in-silico* binding assay of a focused library of 42 fragments. And finally, in lead optimization, by applying the method to current problems encountered in real drug discovery projects. Specifically, this last part was performed in conjunction with three big pharmaceutical companies, Janssen Pharmaceuticals, Boehringer Ingelheim and Pfizer, and turned out in three works that will be published in the next couple of months.

Of course, this does not mean that we consider all the problems to have been solved, there are a lot of challenges to overcome: the forcefield and parameterization issues, discretization methods and the endless problem of sampling. In spite of this, after the experience of this thesis, we strongly believe that the methods and applications developed have the potential to becoming useful tools in drug discovery in the near future.

PUBLICATIONS

This section lists the publications that were carried out during the period of this thesis. Publications 1, 2, 7 and 8 are published. Publications 3, 4 are currently submitted. Publications 5 and 6 are pre-printed. However, in publication 6, performed in collaboration with Pfizer, we could not reach a disclosure agreement by the time of writing and will not be entirely included in this thesis. The numbering here used does not apply in following sections.

First author

1. Lauro G, Ferruz N, Fulle S, Harvey MJ, Finn PW, De Fabritiis G. Reranking docking poses using molecular simulations and approximate free energy methods. J Chem Inf Model. 2014 Aug 25;54(8):2185-9. doi: 10.1021/ci500309a.
2. Ferruz N, Harvey MJ, Mestres J, De Fabritiis G. Insights from Fragment Hit Binding Assays by Molecular Simulations. J Chem Inf Model. 2015 Oct 26;55(10):2200-5. doi: 10.1021/acs.jcim.5b00453.
3. Ferruz N, De Fabritiis G. Binding kinetics in drug discovery. Mol Inform. 2016 Jul;35(6-7):216-26.
4. Ferruz N, Tresadern G, Pineda-Lucena A, De Fabritiis G. Multibody cofactor and substrate molecular recognition in the *myo*-inositol monophosphate enzyme. Sci Rep. 2016 Jul 21;6:30275.
5. Ferruz N, De Fabritiis G. Insights from in-silico binding assay of drug-like molecules, preprint.
6. Ferruz N, De Fabritiis G. Potent selective D3 antagonist reveals a unique binding mode in aminergic GPCRs.

Other publications

7. 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 23;53(12):3123-6. doi: 10.1021/ci400456m.

8. Arena S, Bellosillo B, Siravegna G, Martínez A, Cañadas I, Lazzari L, Ferruz N, Russo M, Misale S, González I, Iglesias M, Gavilan E, Corti G, Hobor S, Crisafulli G, Salido M, Sánchez J, Dalmases A, Bellmunt J, De Fabritiis G, Rovira A, Di Nicolantonio F, Albanell J, Bardelli A, Montagut C. Emergence of Multiple EGFR Extracellular Mutations during Cetuximab Treatment in Colorectal Cancer. Clin Cancer Res. 2015 May 1;21(9):2157-66. doi: 10.1158/1078-0432.CCR-14-2821.

Table of Contents

Chapter 1

INTRODUCTION

1.1 Protein-ligand interactions

1.1.1 Theory and context

Drug action begins with the interaction of a molecule against a receptor, triggering a series of events that ultimately promote a safe, pharmacological response that corrects a specific pathology. The molecular details of the recognition between the two partners are the ultimate responsible of the duration and magnitude of the drug effect. Designing and improving this recognition requires an understanding of the specific interactions and quantitative measures of their strength and duration. Therefore, the details of the response are fully determined by the thermodynamics, kinetics and receptor’s modulation the drug promotes in the receptor upon the process.

For the purpose of understanding the drug-receptor complex formation from a theoretical point of view, we will consider the binding process in a simplistic single-step model in which the ligand (L) reversibly binds to a unique pocket on its receptor (R), forming a complex (LR) (**eq. 1**). The rates at which this complex forms and dissolves are the so-called on (k_{on}) and off-rates (k_{off}):

$$+ R \ \Leftrightarrow LR \qquad (\ 1\ )$$

$$\frac{\mathrm{d[LR]}}{\mathrm{dt}} = \mathrm{k_{on}[L][R]} - \mathrm{k_{off}[LR]} \qquad (\ 2\ )$$

The speed at which the complex forms depends on the rate at which it is made from association of the reactants (k_{on}[L][R]) and the rate at which the complex dissolves (-k_{off}[LR]) (**eq. 2**). [L], [R] and [LR], which represent the molar concentrations of ligand, receptor and complex, respectively, do not change once the system is in equilibrium. The equilibrium association constant (K_{eq}) measures the extent to which the ligand is bound in the equilibrium (**eq. 3**):

$$K_{eq} = \frac{[LR]_{eq}}{[R]_{eq}[L]_{eq}} \qquad (3)$$

Binding affinities are more usually expressed in term of the equilibrium dissociation constant (K_D), which also has the units of concentration:

$$K_D = \frac{1}{K_{eq}} \qquad (4)$$

More intuitively, K_D is the concentration at which 50% of the receptors are occupied at a particular site. The K_D is directly related the concentration-independent standard Gibbs binding free energy (ΔG^0)[2] :

$$\Delta G^0 = -RTln(\frac{1}{K_D}) \qquad (5)$$

Where R is the ideal gas constant and T is the temperature in Kelvin degrees. **Fig. 1** presents the one-dimensional projection of reaction energy landscape along the reaction coordinate. As appreciated, the Gibbs free-energy ΔG only depends on the relative stability between free and bound states, initial and end points of the reaction coordinate regardless the pathway of binding.

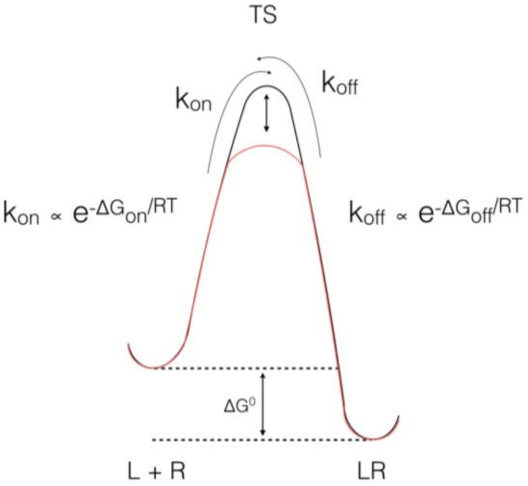

TS
k_{off}
k_{on}
$k_{on} \propto e^{-\Delta G_{on}/RT}$
$k_{off} \propto e^{-\Delta G_{off}/RT}$
ΔG^0
L + R
LR

Figure 1: Energy profile of a ligand (L) binding to its receptor (R) assuming a simple two-state model. The energy difference between unbound (L + R) and bound (LR) states is the binding affinity of the process (ΔG^0). The kinetics is governed by the energy of the transition state (TS), namely, the association rate (k_{on}) depends on the energy difference between unbound and the TS, (ΔG_{on}) and the dissociation rate on the ΔG_{off}. The curve in red represents a reaction occurring with the same affinity, but at faster timescales, as would happen under the effect of a catalyst.

When less than zero, ΔG^0 is an indicative of reaction's spontaneity at conditions of constant temperature and pressure such as the case of biological systems. However, affinity does not determine the reaction's rate. Instead, the kinetics of binding depend on the interactions along the binding pathway, and shape the energy profile of the binding reaction. This way, stabilization or destabilization of the highest point in the energy barrier (the transition state, TS) would modify both on and off-rates in the same direction without changing the affinity of the complex. Still, there is a link between the kinetics and thermodynamics of the reaction in equilibrium:

$$K_D = \frac{k_{off}}{k_{on}} \qquad (6)$$

Of course, more realistic reactions do not follow a single-step model, but occur instead through more complex mechanisms of binding. **Fig. 2** summarizes the three most accepted mechanisms of interactions. The first one refers to the single-step model represented in **Fig. 1**, where free and bound receptor conformations are similar. In the second process, the so-called induced-fit,[3] the ligand promotes changes in the receptor upon binding, shifting the conformation towards one energetically more favourable for the complex. The third mechanism corresponds to the conformational selection process,[4] where the receptor presents certain plasticity in solution and the ligand binds each conformer to different extents. The least populated protein states could actually be those the ligand presents highest affinities, thus slowing down the formation of the complex. In addition, ligands and receptors are not isolated at physiological conditions, and both might be interacting with other molecules in the same environment.

There has been debate for some years on which model represents better the ligand-receptor recognition.[5] The reality is that the three mechanisms probably occur to some extent in all binding processes. Proteins often present many long-lived conformations in solution, to which the ligand exposes different degrees of affinity. The proportion of the different states after ligand binding ultimately depends on the relative affinity of the ligand against each of the conformers, the specific kinetics of each individual binding event and its ability to produce an induced-fit towards other states upon binding. In the same way, there is a crescent acceptation that this complex equilibrium network is not purely a two-body problem, because membrane, water molecules, ions, cofactors or other specific molecular entities may guide or hamper the binding events.[6] Most of the work presented in this thesis has dealt with this particular issues, with the first works assuming small fast molecules binding in single-step processes to rigid receptors, and increasingly accounting for more realistic processes like protein plasticity, cofactor interactions and slower binders.

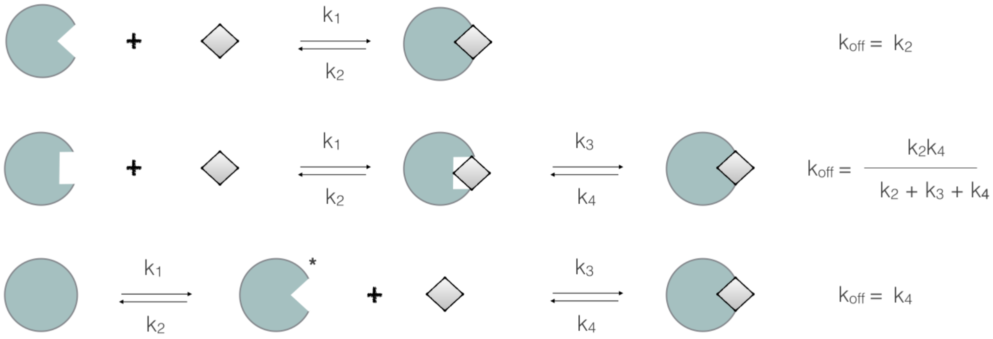


Figure 2: The three common mechanisms of inhibition and their respective dissociation rates: single-step or ‘lock and key’, induced-fit and conformational selection mechanisms. The macroscopic off-rate is related to the microscopic rate constants of each of the steps.

1.1.2 The drug discovery pipeline

Developing a new drug takes usually more than 12 years and more than $1 billion average costs.[7] The pipeline shown in **Fig. 3** presents the process of drug development following a target-based approach. There are two main stages involved in the process: the pre-clinical and clinical stages. The first one, in summary, involves finding active hit molecules, lead optimization and testing in animals. The clinical stage tests the developed compound in humans and evaluates its safety and efficacy.

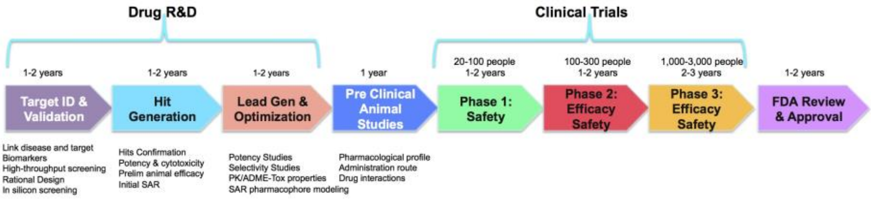


Figure 3: The drug discovery pipeline: duration and tasks of each phase.

Despite many drugs enter the clinical phase every year; the attrition rate is very high. For instance, up to 95% of the anticancer drugs entering the clinical trials are not finally approved.[8] Given the costs and failures of the clinical stage it is clear that any novel technique that helps to improve the success rate is welcome in the pre-clinical stage. This thesis has focused on testing and establishing methodology for these initial phases. Before entering into the core of the work, next sections will lay the background on the theory, context and methodology.

1.2 Current experimental methods

Binding thermodynamics influences the extent to which a ligand binds its receptor, but also its selectivity and drug-like properties. In a single high-throughput screening (HTS), many hits can be obtained presenting similar potencies. Maximizing this potency prior to the other preclinical phases, although necessary, is not the unique condition to produce a lead compound with the appropriate drug-like character, since many of this hits may present similar values. The first step of the drug discovery pipeline, not only focus on potency maximization, rather on a potency optimization. ΔG°, is composed of enthalpic (ΔH) and entropic (TΔS) contributions (**eq. 5**), the first related to the ligand-receptor interactions, and the second to the solvation/desolvation process and conformational disorder.

$$\Delta G^0 = \Delta H^0 - T\Delta S^0 \qquad (7)$$

Most often, the discovered compounds on the first screenings present low micromolar affinity, that is to say, values around -8kcal/mol in ΔG^0. Fragments are frequently found in the low-millimolar range. Typically, the aim is designing a compound with sub-nanomolar affinity, that is, below -13kcal/mol. Thus, a hit molecule must be optimized at least 5kcal/mol during the lead optimization phase, and this process can occur through an infinite combinations of different enthalpic and entropic contributions. Hence, a simultaneous representation of the three state functions – termed as thermodynamic signature- provides a useful visual representation of each of the variables. The thermodynamic signature is often measured by isothermal titration calorimetry.

Isothermal titration calorimetry (ITC) is a quantitative technique that provides affinity constant, enthalpy and stoichiometry of a reaction. Thus, all thermodynamical parameters can be obtained in a single experiment by the use of formulas (5) and (7). It works by directly measuring the heat that is released or absorbed during the course of a reaction where the ligand is gradually titrated into a cell containing the receptor. The advantages of this technique are its ability to provide all thermodynamic parameters using modest sample sizes and experiment times.

While determining binding thermodynamics is crucial in the binding process characterization, it is only one side of the coin. The process has its dynamic perspective as well; having the ligand off-rate in some cases more impact in the *in-vivo* activities that the thermodynamics. Thermodynamic measurements are very often performed in *in-vitro* settings, thus conforming what is termed as closed-system conditions. Ligand and receptor concentrations stay invariant through the course of the experiment and the equilibrium is ultimately reached in the long-term.[9]

However, in the *in-vivo* scenario the ligand concentration is determined by the time between doses, the interaction with other targets or the extracellular diffusion. This setting is termed as an open system. Because the drug's concentration is in continuous variation, sometimes the equilibrium is not reached and thermodynamic variables might not be adequate descriptors of the *in vivo* efficacy. In 2006, Robert Copeland defined the term residence time as the period of time the ligand is bound to its receptor.[9,10] This study showed how the residence time can determine temporal selectivity for the target receptor despite other targets having more affinity, and following works successfully put this theory into practice.[11,12]

Mathematically, the residence time is quantified as the reciprocal of the (k_{off}):

$$\tau = \frac{1}{k_{off}} \tag{8}$$

Characterizing the kinetic binding profile of a ligand of interest can be of utmost importance. First, it provides a more complete understanding of the ligand action. Second, the association and dissociation constants are linked to thermodynamics and offer new approaches to tune potency (**Fig. 1**). And last, optimization of a fast or slow binding and unbinding profile could be very advantageous depending on each specific case. Fast ligand binding increases the opportunities to capture short-lived receptor conformations. And while fast unbinding might confer safety advantages in target-mediated toxicity systems, slow unbinding ligands lead to long-lasting effects that prolong the therapeutic efficacy.

Increased focus on the kinetics of binding has been supported by an improvement of the related instrumentation, frequently divided in techniques using a label for fluorescence (radioisotopes of fluorescence), label-free techniques (biosensors) and enzymatic experiments.

Radioligand binding is the preferred technique for G-protein coupled receptors (GPCRs).[13,14] There are two main ways to measure kinetics using radioligand binding: direct radiolabeling of the ligand of interest or indirectly by competition experiments.[15] In the direct method, the dissociation rates can be determined straightforwardly: the receptor is pre-incubated with a known concentration of the radioligand and the unbinding is measured in a washout phase by blocking the formation of new complexes. The binding decay can then be fitted by a non-linear regression analysis. The k_{on} is computed by performing association experiments at different radioligand concentrations, or by performing a single experiment at a concentration when k_{off} is known. Alternatively, the binding properties of a set of unlabelled drugs are computed by competition displacement by a radioligand of known affinity as was proposed by Motulsky and Mahan.[16] More recently, dual-point competition assays have also shown being a fast high-throughput method.[13] Although the method permits direct measurement of rates, radiolabeling is expensive, laborious, time consuming and generates radioactive waste. An alternative to the use of radioligands are spectroscopic labels, the so-called fluorescence methods, like time resolved fluorescence resonance energy transfer (TR-FRET), fluorescence anisotropy and intrinsic fluorescence.[17,18]

Label-free surface plasmon resonance (SPR) is a high sensitivity method which monitors refractive indexes changes when molecules absorb and desorb from a biosensor chip, dependent on the surface mass increase.[19] The receptor is immobilized on the solid surface and the drug (analyte) diluted in solution under continuous flow while the association is monitored in real-time. The method is particularly suitable for globular proteins.[20,21] The method needs short development time, little material and is parallelizable. However, the immobilization of the receptor could affect the binding properties,[22] it has relatively low throughput and there are a limited range of rates which can be sensitively determined.[6]

Enzymatic activity assays can also determine the binding kinetics from enzyme activity. This approach has successfully been used to measure binding kinetics for many enzymes.[23–27] Jump dilution assays provide a format to highlight the dissociation kinetics by pre-incubating with high ligand concentration and diluting a hundredfold afterwards.[28]

On the side of structure and dynamic determination, there are the X-ray crystallography and nuclear magnetic resonance (NMR). **X-ray crystallography** allows determining the position of atoms within a crystal, by striking a crystal with a beam of X-rays that is spread into different directions when it contacts the crystalline atoms. From the angles and intensities of the diffracted beam, a three-dimensional picture of the density of electrons can be produced, and from them, mean atom positions, their chemical bonds, disorder and various other information.[29] X-ray crystallography can also be used to study dynamics, especially of slow timescales. Recently, an application of the Hadamard time-resolved crystallographic (HATRX) method to the high-resolution measurement of processes occurring in the millisecond timescale was published.[30]

Nuclear magnetic resonance (NMR) spectroscopy is based on the property of certain isotopes that absorb and desorb electromagnetic radiation. The most commonly used isotopes in macromolecules are H^1, C^{13} or N^{15}. NMR is used to determine the structure and dynamics of many biological molecules. It is used to describe populations and exchange rates between protein conformers, and it can also be used to ligand binding. Methods for detecting protein-ligand interactions fall into two categories: those that monitor NMR signals from the protein, such as chemical shifts perturbation studies, and those that monitor the ligand. In the first ones, the ligand alters the chemical environment around the binding site, which will perturb the chemical shift around it. The ligand-based experiments (STD, waterLOGSY) do not require isotopic labeling but provide a yes-or-no binding answer, with no additional structural information. NMR one-dimensional ligand-based is also amenable to competition experiments, by performing competition assays with known inhibitors.[31]

1.3 Molecular dynamics and analysis

1.2.1 Sampling, forcefield and ligand parameterization

Molecular dynamics (MD) is a technique that models the physical movement of the atoms in a system by following the Newton's law. In MD, each atom in the system is treated as a point particle with a specific mass that moves following classical forces.[32] As a more specific example of this thesis, we will focus in a ligand-protein simulation box. The initial coordinates of the receptor can be obtained by X-Ray or NMR experiments or either modelled through homology modelling.[33] The ligand coordinates are usually sketched with any of the many available software.[34,35] The evolution of the system occurs through iterations of short time steps (Δt), at each of which the Newton's law ($\vec{F} = m\vec{a}$) is evaluated and velocities and coordinates updated. The sum of these forces is derived from a set of potentials termed as molecular forcefields, parameterized to capture the environment of all particles. An example of a class I forcefield equation is shown in **eq. 9**.[36] Typical simulation steps are usually less than 5fs,[37] and therefore millions of steps must be produced to reach biologically relevant timescales. **Fig. 4** presents approximate timescales for some biological events involving proteins or drugs. While a simulation timestep is of the order of 10^{-15} s, the fastest protein motions and ligand binding events occur in the microsecond timescales and protein folding and ligand unbinding usually take milliseconds.

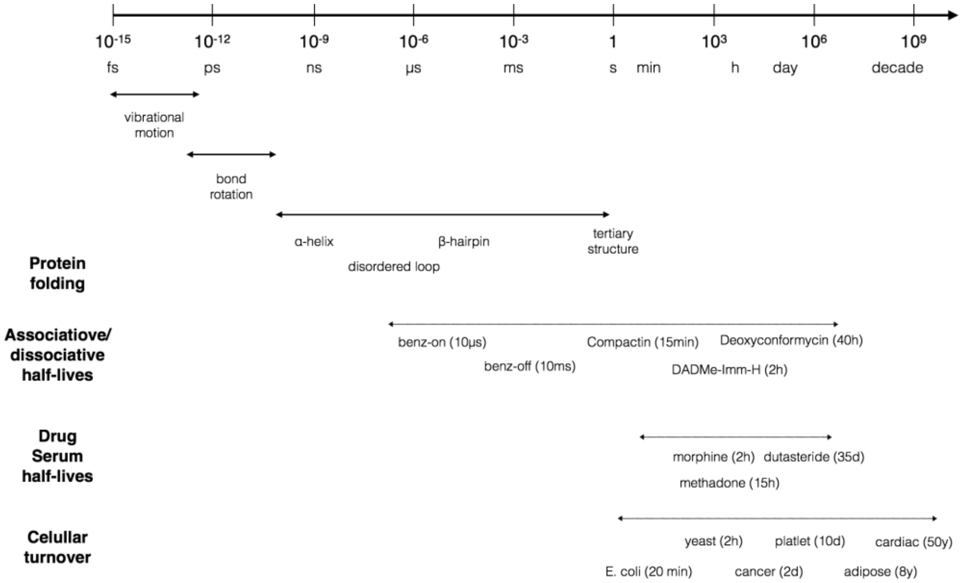


Figure 4: Approximate timescales for biological half-lives and protein motions. Some of the biological processes inside the cell occur in timescales slower than the second (such as cellular turnover or drug serum half-lives). Molecular dynamics is currently able to sample processes in the low millisecond processes, although intelligent schemes and analysis methods can recover processes occurring in the second scale.

The first MD simulations produced back in 1977 for the motions of trypsin in vacuum were just a few picoseconds long,[38] but modern simulations (or ensembles of them) are able to reach the low-millisecond timescales. Among the reasons of such a dramatic increase –surpassing Moore’s Law[39]- are the algorithm improvements, parallelization of codes to run in high-performance supercomputers,[40] designing of specialized hardware[41] and development of intelligent MD protocols.[42]

Along with an increase in sampling capabilities, the forcefields are also continuously being refined and updated. While sampling times are the responsible of the precision in the simulation outcomes, molecular mechanics forcefields are in charge of the accuracy to correctly reproduce the binding events. Most commonly used forcefields for ligand-receptor systems are AMBER,[43] CHARMM[44] and OPLS,[45] the first two used throughout this thesis. Although their general mathematical description of the forcefield is quite similar, they differ in their parameters and the methods to obtain those. For a nice description comparing different forcefields for a set of protein systems refer to ref. 46.

Eq. 9 represents a class I potential energy function used in the forcefields for biomolecular simulations.

$$Epair =$$

$$\sum_{bonds} k_r\,(r - r_{eq})^2 +$$

$$\sum_{angles} k_\theta\,(\theta - \theta_{eq})^2 \;\; +$$

$$\sum_{dihedrals} k_\varphi\,(1 + \cos(n\varphi - \;\delta)\,) +$$

$$\sum_{i<j}\left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}\right) + \frac{q_i q_j}{\varepsilon r_{ij}} \qquad (9)$$

The forces acting on each atom in the system arise from bonded and non-bonded terms. The first ones, also called internal terms, relate to bond, angle and dihedral parts, which are modelled using virtual springs and sinusoidal functions, respectively. The second term, arise from Van der Waals and Coulombic (electrostatic) interactions. Additionally, CHARMM has the Urey-Bradley term, which parameterizes the 1,3 interactions, and another term for impropers. In the latest CHARMM versions, Eq. 9 also includes a CMAP term that corrects backbone terms.[47]

In drug design the forcefields should accurately represent both proteins and ligands interacting with them. Both Amber and

CHARMM have general purpose forcefields which present a set of atom types with defined parameters to which the new molecules are mapped by similarity. Amber has the generalized Amber forcefield (GAFF)[48] while CHARMM has developed the general forcefield CGenFF.[36]

GAFF and CGenFF function by a similar approach. The first step is the atomtype assignation. Atom types are generic atom definitions that encode specific chemical environments. GAFF presents 35 basic atomtypes and 22 special ones, while CGenFF has more than 150 at the time of writing, coming from a set of diverse model compounds which describe a wide spectrum of the chemical space, and were firstly parameterized by QM. While the list of atomtypes in the CGenFF continues to grow, the number of atomtypes in GAFF is fixed. For further explanation on how the bonded and non-bonded parameters are obtained after the atom type assignment, works in refs. 48 and 49, 50 provide the specific details for GAFF and CGenFF, respectively.

The previous general (organic) forcefields, although permit the parameterization of large sets of molecules in seconds,[50] lack the accuracy needed in later stages of the drug discovery pipeline, when fewer compounds are tested and their specific binding modes and properties need to be reliable, even at the expenses of requiring larger times.

1.2.2 Thermodynamics, kinetics, and pathway reconstruction by molecular dynamics

We have previously introduced the importance of an accurate determination of thermodynamics, kinetics, pathway reconstruction and receptor conformational characterization in the context of drug design. The computational methods presented in this section go from the fastest but less accurate methods to those more accurate but expensive.

In this section, we limit our description to the methods that we have used in this thesis. Any comprehensive description of methods dealing with reconstructing physical chemistry properties from molecular dynamics simulations would be so extensive that it is out of scope for an introduction. In particular, we focus on docking, linear interaction methods, umbrella sampling, Markov state models, high-throughput simulations and adaptive sampling.

Scoring functions

Scoring functions estimate binding affinities as the sum of a set of approximated terms -like hydrogen bonding, salt bridges, van der Waals interactions and protein conformations- whose parameters are fitted to experimental data.[51] Although very approximated, scoring functions are widely used in molecular docking in the first stages of the drug discovery pipeline, providing high-throughput ranked poses and affinities for protein-ligand complexes in a fast manner.[52] However, although the ligand is considered as flexible, the protein counterpart is usually rigid. Solvent molecules usually do not take part in the computation.

Linear interaction energy (LIE)

The LIE method considers the two endpoints of the reversible binding cycle: the free state, where the ligand is solvated in water, and the bound state, where the ligand is in complex with its receptor. Therefore, the binding energy is estimated as the process of transferring the ligand from the water to the protein environment.[53] LIE predicts the binding free energy of compound-protein complexes using the following linear approximation:

$$\Delta G_{bind} = \alpha\left(\langle U_{c-s}^{vdW}\rangle_b - \langle U_{c-s}^{vdW}\rangle_f\right) + \beta\left(\langle U_{c-s}^{el}\rangle_b - \langle U_{c-s}^{el}\rangle_f\right) + \gamma \qquad (10)$$

Where the brackets indicate averages along the simulations for the compound surrounding interaction energies (c-s), *el* stands for electrostatic, *vdw* for nonpolar, *b* for bound and *f* for free. α and β are the related scaling factors that could vary depending on the specific system considered, and γ is an additive factor generally weighted to fit the experimental binding affinities.

The polar contribution comes from the linear response theory[54] and the nonpolar term from the linear dependency between solvation free and potential energies with the size of the compound.[53] Unlike more elaborated methods such FEP or IT where unphysical states must also be simulated, the equation reduces the protocol to virtually simulate two states: ligand in water and protein surroundings.

Since the firsts applications of the LIE method in the 90s a wealth of works have been published,[55–57] mostly applied to the screening of a low-medium number of compounds. A big part of the research has focused on the derivation of the optimal scaling factors. Ref. 58 focused a first set of values for the β parameter taken from deviations of the linear response theory. Almlof et al [59] proposed a detailed set of values based on FEP calculations, allowing for greater flexibility. The values used in this work were the basis for those in Publication 3.2. Regarding the nonpolar term, α, was estimated based on a set of compounds, giving a value of 0.18.[60] Finally, the γ value is mostly used for the estimation of absolute binding energies, and has been related to the hydrophobicity of the binding pocket.

Potential of mean force and umbrella sampling

Free energies of binding can be computed by simulating the process in which the ligand moves from an infinite separation of the protein to the binding site through a reaction coordinate, also called collective variable. Collective variables are usually geometric parameters that change over the course of the reaction, such as angles between three points in the protein, or the one-dimensional projection on the z-coordinate of the bulk-pocket distance. The forces affecting the course of the reaction are then described as an effective potential of mean force (PMF), that is, the free energy profile along the reaction coordinate.[61,62]

Specifically, the PMF, W(z), is defined as the negative logarithm of the probability of being at a certain state in this reaction coordinate (z):

$$W(z) = k_b T ln \ln p(z) \qquad (11)$$

The PMF is the base for the computation of the pathway-based free energy calculation methods as the ones previously mentioned. All thermodynamical properties can be expressed in terms of W(z), and therefore is a key variable in macromolecular computational studies. However, PMF calculations would need intractable computational times to complete the reaction coordinate: along z there might be high barriers taking microsecond or millisecond times. This problem has been circumvented using biasing protocols.[63,64] For instance, The Jarzinsky equation and the Crooks fluctuation theorem allow to recover the PMF from non-equilibrium simulations which use pulling forces, termed steered molecular simulations (SMD).[65,66] Other mostly used protocol to compute PMFs is the umbrella sampling (US) method.

In umbrella sampling the ligand is moved by stratification in successive steps or windows ($z_0 \ldots . z_n$) along the collective variable. At each step, a potential function is applied to the ligand such that it stays in the surroundings of z_i. Usually, this potential is represented by a harmonic function of the form:

$$v_i(z) = 0.5k(z - z_i)^2 \qquad (12)$$

Successive windows along the z coordinate are needed to complete the binding pathway. The weighted histogram analysis method (WHAM)[67] is used to sum up the free energy differences corrected for the potentials. In the case of unbiased one-dimensional umbrella sampling methods -as the one performed in publication 3.4-, the standard free energy can be straightforwardly calculated.[68,69] The choice of the correct collective variables is the most crucial parameter for obtaining accurate free energy estimations and the main drawback of pathway-based methods. An incorrect election of the reaction coordinate would lead to poor free energy estimates.[63,68,70]

High throughput molecular dynamics and Markov state modeling

All previously mentioned techniques, despite being fast approaches with reasonable success, modify the system by applying forces that might alter the real dynamics to different extents. In the most recent years, the advent of new computing infrastructure has promoted a wealth of studies making use of unbiased brute force simulations. In unbiased MD, ligand and receptor freely move in a solvated system. Often many replicas of the same processes are run, and these ensembles -which could vary from a few very long simulations or ensembles of thousands of short ones- contain all the needed information for reconstruction of the binding process. In 2008, a massively parallel supercomputer (ANTON)[71] was designed and built at D. E Shaw Research in New York, able to produce millisecond-long simulations. The ANTON2 chip was released in 2014 and is currently capable of running microsecond-per-day simulations in multi-million atom systems.[41] MDGRAPE is another petaflop special-purpose supercomputer devoted to MD simulations, currently in its fourth generation.[72]

However, this highly specialized computing machinery, although have promoted an impressive contribution to the field and have settled MD as a well-recognized technique, is economically inaccessible to most researchers. Fortunately, graphical processor units (GPUs) have also been demonstrated to be very efficient in MD. With the introduction of generalized GPU architecture like CUDA or OpenCL a GPU workstation is capable now of performing microsecond-length simulations. Still, most biological processes we are interested in occur in high-microsecond or millisecond timescales (**Fig. 3**), unaffordable times for a single GPU. However, it is possible to run multiple parallel simulations in GPU clusters, and posteriorly analyse them with probabilistic models. This thesis has mainly focused in the application of this methodology, namely, the production of high-throughput molecular dynamics (HTMD) ensembles and their posterior analysis with Markov state models.

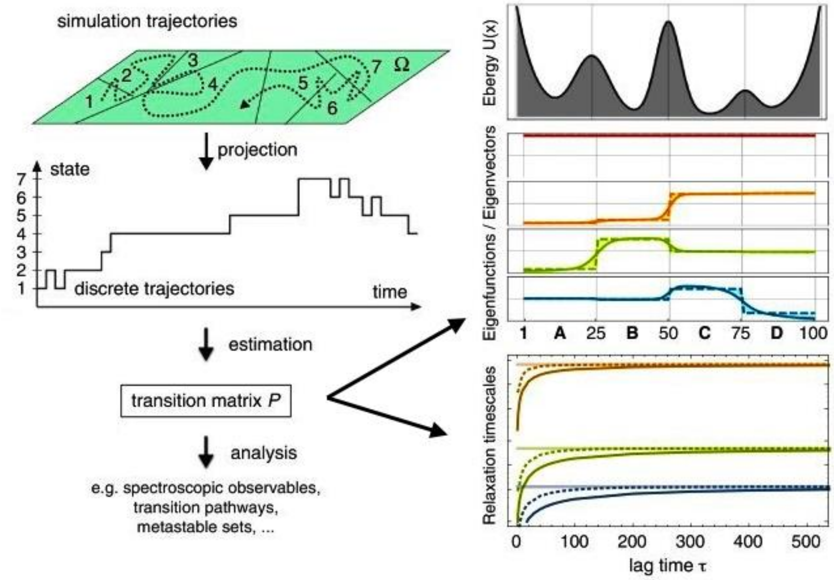
Nowadays, a single GPU is able to run around 125ns of simulation time for a system sized 50000 atoms in a benchmark GPU like the GTX980. Assuming the case of a prototypical fast fragment binding event as benzamidine binding to trypsin protease - which occurs with an on mean first passage time of 6.9μs at a concentration ~5mM- we would need around 20μs of aggregated simulation time to sample a more than an anecdotal binding event and obtain enough statistics for the construction of our model. Having an *in-house* cluster of 10 GPUs running uninterruptedly - currently affordable for many research groups- we could obtain our brute-force MD ensemble in about two weeks.

This specific example was a breakthrough work 5 years ago, when using GPUs volunteered from all over the work (GPUGRID.net),[73] it was possible to completely reconstruct this binding event.[1] 495 simulations of 100ns each were performed leading to 197 binding events within 2Ǻ RMSD of the crystal structure. By using a Markov state model (MSM) analysis, the simulations reconstructed binding intermediates affinities and kinetic estimates that occur at longer timescales, from a microsecond ensemble, with remarkable accuracy. A couple of years later, the protocol was applied to the binding of carboxythiophene to AmpC β-lactamase,[74] by performing 148μs of total aggregate time. In this work, it was possible to observe other secondary poses in agreement with the X-ray results. More interestingly, this study permitted to characterize the role of a loop in the vicinities of the pocket during the binding pathway: the loop was able to explore both open and closed conformations in absence of the ligand, however, the ligand stabilized the open conformer while entering the pocket, and the closed one once inside it. This thesis, following the line of these studies, has focused on pushing the limits of this methodology for more complicated scenarios, outlined in the Objectives section.

Markov State Models (MSMs)

MSMs are probabilistic models able to map the ligand binding energy landscape and thus provide thermodynamic and kinetic estimates.[75] The construction of a MSM for a ligand binding process passes roughly through the following phases. First, the data is discretized into a lower dimensional space for its posterior clustering, usually by means of distance or ligand contact maps against the protein or its alpha carbons. This data still presents a high dimensionality that can be further projected by the time-lagged independent component analysis (TICA).[76] TICA projects the data into the slower order parameters and thus separates well metastable minima placing clusters in the transition regions.

Then the data is geometrically clustered using any of the available algorithms (k-centers, k-means, regular clustering, etc.) and from them then construction of the transition matrices at the different lag times. This step allows for the visualization of the implied timescales and the subsequent selection of the appropriate lag time for the MSM construction, taken at the first point where the timescales are convergent to ensure Markovianity while keeping enough statistical sampling. The first eigenvector of the transition matrix at the chosen lag time, correspond to the global stationary distribution of the system, and subsequent eigevectors approximate the intrinsic relaxation times of the system. Therefore, the quality of the MSM highly depends on the convergence of these timescales, and the chosen lag time. At this stage, the model represents the dynamics of the system, but it can contain thousands of states that need some sort of coarsing before human visualization. Therefore, we lump this microstates to the number of macrostates of our choice



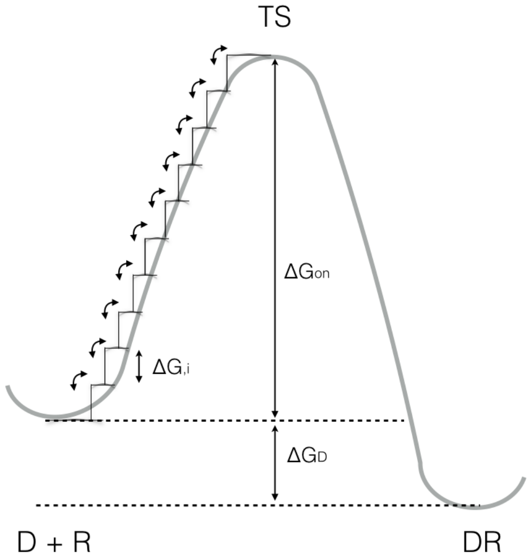
with the Perron Cluster Cluster Analysis (PCCA), although there are many other methods available.[75] The number of macrostates is usually hindered from the number of processes above the largest gap in the timescales plot, which provides an appropriate separation of the slowest processes in the system. Usually a range between 4 and 8 macrostates represents fairly all the basins in the ligand binding landscape. Refs. 75,77,78 are suggested reading for a detailed description and applications of MSM analyses. **Fig. 5** shows a schematic view of the process.

Figure 5: Overview of the MSM process. The simulations are geometrically discretised into a set of states from which the transition matrices can be built at a given lag times. The stationary distribution and slowest relaxation times of the processes occurring in the system can be approximated from its eigenvectors. In this case, a four-well potential model presents three slowest processes, as the three jumps of the kinetic barriers. Figure adapted from ref. 79.

Adaptive ligand sampling

In all the ligand-binding studies mentioned in this section the reconstruction of the MSM was performed out of brute-force simulation ensembles. Intuitively, it implies that the ligand is inefficiently spending considerable time in previously sampled states, such as the bulk. However, for the construction of MSM, the simulations do not need to start from equilibrium, and can actually start from interesting states previously sampled in successive batches of simulations. This intelligent approach is called adaptive sampling, and has recently been used for protein folding[80] and for ligand binding,[42,81] and in some of the publications in this thesis. Adaptive methods decrease in one order of magnitude the required sampling times.[82–85]

The theoretical reasons for this decrease in sampling time are sketched in **Fig. 6.** While standard brute-force schemes would need tens of millisecond to jump the activation barrier of a slow binding ligand (k_{on} ~ 10^4), finely discretizing the barrier into smaller ones would exponentially decrease the sampling time, provided a correct discretization of the high-dimensionality binding process.



$\Delta G_{on} = 10 \cdot \Delta G_{,i}$

Figure 6: Simplified one-dimensional energy profile for a ligand binding process. The large kinetic barrier between the two basins can be discretized into successive steps.

1.4 Biological systems studied

1.4.1 Serine proteases

Between 2 and 4% of the genome encodes proteolytic enzymes, a set known as the degradome.[86] Almost one third of all proteases are serine proteases,[87] grouped in 4 clans and 13 families.[88] In these peptidases, the serine in the catalytic pocket acts as nucleophile attacking the carbonyl in the peptide backbone forming an acyl-enzyme intermediate. The other accompanying residues in this catalysis are Asp and His (**Fig. 7**), which confer the catalytic triad.[89] Four different folds present the same enzymatic mechanism driven by this triad: trypsin-like, subtilisin-like, prolyl oligopeptidases, and ClpP peptidases.[90] Many serine peptidases employ a simpler dyad mechanism where Lys or His is paired with the catalytic Ser, or a pair of His combined with Ser.[90] Activation of serine proteases requires the cleavage of an inactive zymogen precursor.[91] The enzymatic mechanism starts with the oxygen in the serine attacking the carbonyl in the substrate peptide backbone by using histidine as the base. It then forms a tetrahedral intermediate termed as oxyanion, creating the positively charged pocket oxyanion hole. Collapse of the intermediate gives the acyl intermediate that is finally released after the attack of a nucleophilic water molecule.

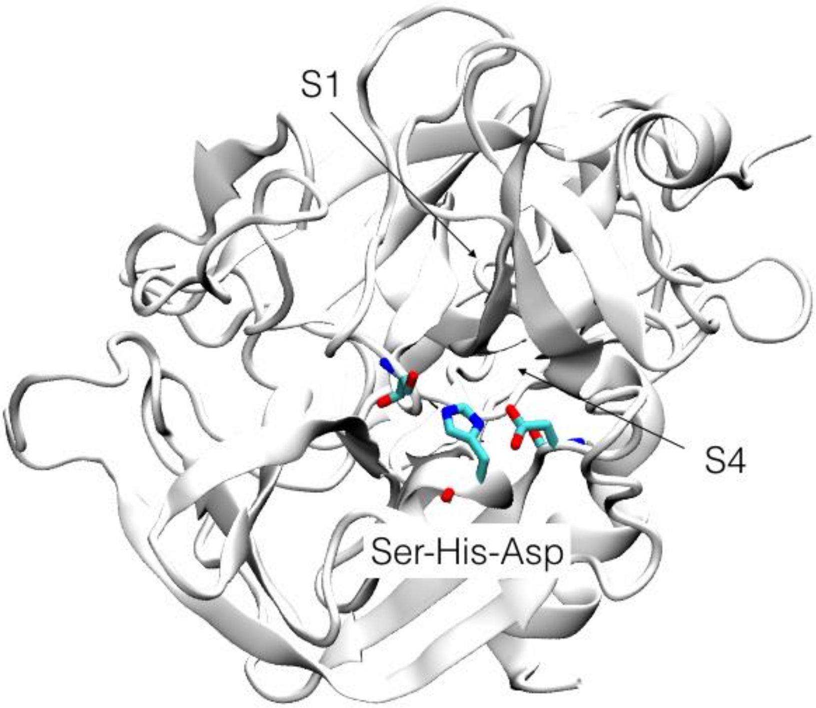


Figure 7: Overview of a serine protease (factor Xa). The catalytic triad (Ser-His-Asp) is shown along with the S1 and S4 subpockets.

As a consequence of the great diversity of serine proteases in structure and function, serine proteases are important targets for different diseases, ranging from diabetes to coagulation problems.[92] In nearly all cases serine proteases can be inactivated by blocking the nucleophile serine with generic inhibitors as diisopropylfluorophosphate and phenylmethanesulfonyl fluoride.[90]

In this thesis, three different works have focused on serine proteases. Two of them, trypsin and coagulation factor Xa, are Clan PA peptidases with a very similar fold. Dipeptidyl peptidase IV is a clan SC exopeptidase.[90]

Trypsin

We used trypsin protease as a receptor for a comparison of the ligand-ranking efficiency between the LIE method and molecular docking[93] in Publication 3.2.

The three residues where the catalysis occurs are specifically residues His57, Asp102 and Ser195. Trypsin cleaves peptide bonds that follow a positively charged aminoacid (Lys or Arg), as these residues favourably bind the Asp aminoacid located at the S1 pocket (**Fig. 7**). Trypsin, like elastase, acts in the digestive system, breaking polypeptides into shorter chains.[94] Trypsin has been characterized with many methodologies and trypsin-like serine proteases are perhaps the best studied group of enzymes.[90] It was one of the first proteins being crystallized by X-ray crystallography[95] and is currently also co-crystallized with many small inhibitors.[96] Trypsin has been used as a methodological work for understanding binding contributions and as a toy model for numerous computational works.

Factor Xa

We used the optimized MSM protocol for ligand binding processes using contact maps,[1,97] previously applied to single ligand analysis, for a library of 42 fragments against the serine protease factor Xa[98] in Publication 3.3. The library contained extensive experimental data annotated and was suitable for testing and optimizing the method for larger libraries.

Factor Xa converts prothrombin to thrombin in the coagulation cascade. As such, the protein has been an attractive target for the search of bioavailable anticoagulants. Structurally, it presents a heavy and a light chain held together by a disulphide bond. The protein active site contains four subpockets, located at the heavy chain defined from S1 to S4, following the convention for proteases and peptide cleavage.[99] Inhibitors usually only exploit pockets S1 to S4 (**Fig. 8**), which display high similarity with related proteases. Specifically, the S1 pocket is defined by A190, D189 and Q192 and recognizes charged moieties and aryl halogens,[100] only differing from Trypsin’s S1 in A190. (For a nice comparison between the pockets in trypsin, factor Xa and thrombin see ref. 101) The S4 pocket favours aromatic moieties, and it is also a cation recognition pocket.[102]

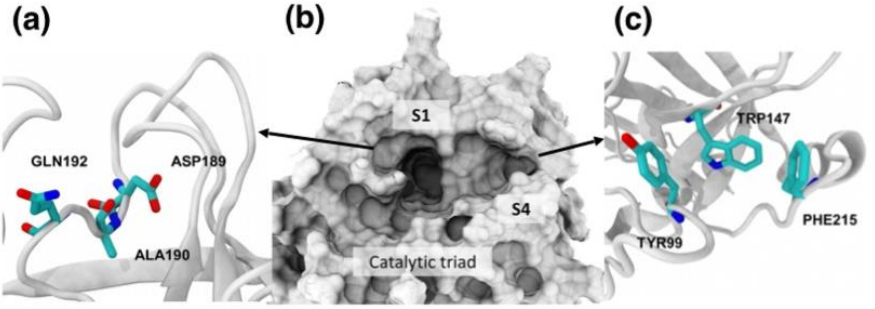


Figure 8: Overview of the factor Xa catalytic pocket. **(a)** Acidic S1 pocket formed by the residues Q192, D189, and A190. **(b)** Overview of the entire catalytic site. **(c)** The aromatic residues Y99, W147 and F215 form the S4 pocket.

Dipeptidyl peptidase IV (dppIV)

We have characterized the binding event of a slow inhibitor of dipeptidyl peptidase (dppIV) in Publication 3.6. DppIV modulates the activity of specific chemokines, hormones, cytokines and neuropeptides by cleaving dipeptides after a penultimate N-terminal alanine or proline.[103] dppIV specifically attenuates incretins GLP-1 and GIP, the reason why its modulation has been used to treat diabetes.[104]

DppIV is usually found as a dimer, a state likely to be highly populated in solution. In fact, it has been shown that the monomer has only residual enzymatic activity compared to the dimer, but both monomeric and dimeric forms bind with similar affinity to adenosine deaminase.[105] The co-crystal structures of dppIV with different inhibitors in the catalytic pocket are available[106,107] and show an interaction as the one depicted in **Fig. 9**. Specifically, the inhibitors span through the S1 and S2 sites close the catalytic triad (Ser630, Asp708, His740), blocking the peptide cleavage.

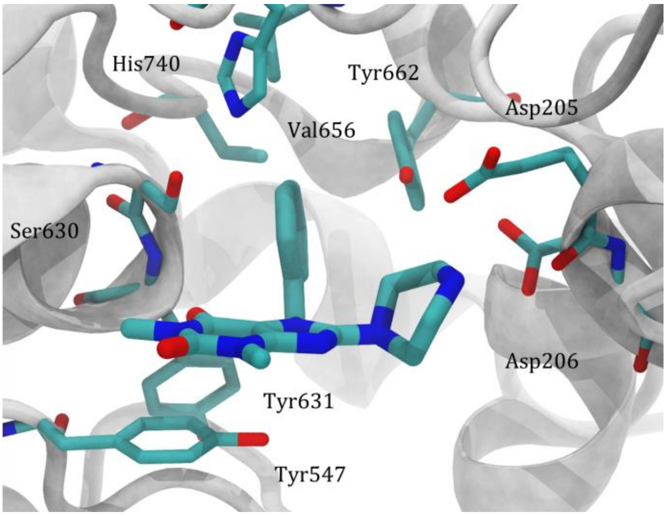


Figure 9: Overview of the catalytic pocket of dppIV. Ligand (BDPX)[107] is shown along with its protein surroundings depicted. The xanthine scaffold of the compound stacks against Tyr547 by π-π interactions. The phenyl extends towards Val656, in the S1 subsite. The piperazine ring interacts via two hydrogen bonds with E205 and E206 in the S2 subsite. Residues His740, Ser630 and Asp708 form the catalytic triad.

1.4.2 Epidermal growth factor receptor (EGFR)

The umbrella sampling protocol (section 1.2.2 and ref. 69) was applied to understand the binding differences between wild-type and mutant receptor for endogenous ligand and drug cetuximab in Publication 3.1.[108]

Epidermal growth factor receptor (EGFR) is one of the four members of the Her1-4 family of receptors (ErbB), which are involved in multiple cellular processes such as growth, migration, differentiation and apoptosis.[109] Concretely, mutations in the EGFR, Her1, alter its signalling transduction pathway and can trigger the development of a subset of epithelial tumours. EGFR overexpression has shown to be highly correlated with tumour progression in colorectal cancer.[110] Different targeting agents have been developed in the recent years, with the two monoclonal antibodies cetuximab and panitumumab presenting the best responses.[111,112] However, the response to the treatment is halted when secondary resistance occurs, usually between 3 and 12 months.[113] In this regard, although the two antibodies bind the same epitope, a missense mutation S468R in the extracellular domain III of EGFR promotes resistance against cetuximab but not for panitumumab[114] (**Fig. 10**).

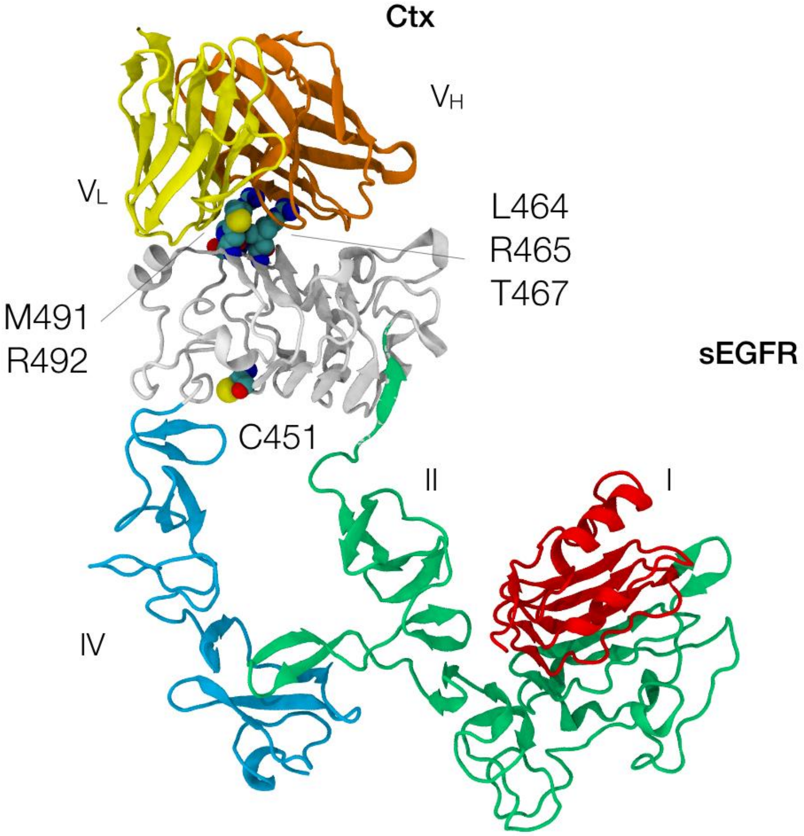


Figure 10: Overview of the the extracellular part of the EGFR receptor (sEGFR) and their mutations known to affect cetuximab (Ctx) inhibition. sEGFR is composed by four different sub-domains, namely I (red), II (green), III (grey) and IV (cyan). The binding of cetuximab prevents the dimerization and the posterior cross-phosphorylation in the intracellular part.

1.4.3 Myo-Inositol monophosphatase (IMPase)

IMPase is a homodimeric enzyme that plays a critical role in the phosphatydilinositol signalling pathway, hydrolizing myo-inositol monophosphate (IP)[115] (**Fig. 11**). Patients suffering from bipolar disorder present overactive IMPase levels and thus higher inositol concentration than under normal conditions. This physiopathology is treated with low concentration of Li^+ (0.5mM-1mM) that directly inhibits IMPase and depletes the inositol levels in neurons.[116]

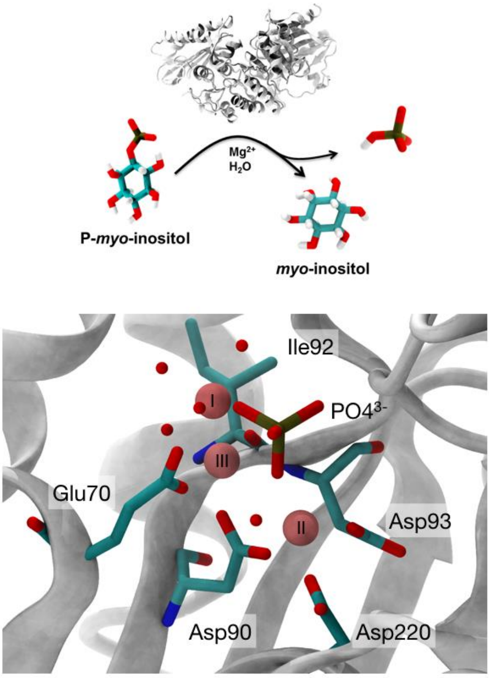


Figure 11: Scheme showing the dephosphorylating process performed by the IMPase dimer. Three Mg^{2+} ions act as cofactor in each subunit in the course of the reaction. IMPase's catalytic pocket is a highly polar environment formed by four acidic residues and the three metals.

A highly polar pocket, and a low understanding of IMPase mode of action prior to the pre-catalytic complex formation are some of the reasons for the failure in finding bioavailable inhibitors.

More concretely, IMPase's catalytic site presents four acidic residues in close vicinity which can adopt the binding of three Mg^{2+} ions acting as cofactors in the hydrolysis of IP.[117] The three metals present affinities in the high micromolar and milimolar range, and given the low neuronal concentrations of Mg^{2+} it is not clear if the three ions occupy the pocket prior the reaction occurs.[118] In the same manner, IP's mechanism of binding and its possible cooperation with cofactor during binding remains unclear.[119,120] In Publication 3.5 we attempted to shed some light into these issues.

1.4.4 D3 Dopamine receptor (D3R)

Dopamine receptors are a class of G-protein coupled receptors (GPCR) located in the central nervous system (CNS) and are specifically activated by the neurotransmitter dopamine. Dopamine receptors have been classified into two different subfamilies, D1 and D2-like, according to their structural similarities.[121] D1-like receptors (D1R and D5R) are coupled with stimulatory G-protein α subunits (Gs/olf) activating adenyl cyclase whereas the D2-like receptors (D2R, D3R, and D4R) couple to inhibitory G-protein α subunits ($G_{i/o}$), inhibiting adenyl cyclase.[122,123]

D3R and D2R receptors share a 78% sequential similarity,[124] being specially conserved the residues of the binding pocket, and have therefore settled a major challenge for therapeutic selectivity.[125] Antipsychotic drugs able to block both D2 and D3 receptors are used to treat schizophrenia although have usually been considered as intolerable due their side-effects. It was hypothesized that designing D3R-binding specific drugs would reduce these side-effects. After years of major industrial and academic research D3R-preferential antagonists and partial agonists (e.g. 7-OH-DPAT, pramipexole, and rotigotine)[126] were developed. The antagonist showed to attenuate drug-seeking behaviour in rodent models and act as antidepressant, supporting D3R blockade as a plausible target for therapeutic discovery[127–131] particularly for substance abuse.[132]

These D3R specific compounds are, however, very lipophilic and have shown poor bioavailability in clinical trials. Eticlopride is a D2R and D3R potent antagonist which has recently crystallized in D3R providing invaluable structural insight to better design more specific compounds within the D2-like receptors.[133]

Chapter 2

OBJECTIVES

The main objective of this thesis was to apply high-throughput molecular dynamics simulations to drug discovery projects. Specifically, we have proved the feasibility of HTMD at three different stages of the drug discovery pipeline: in compound screening, by means of approximated methods (LIE), in hit fragment identification, performing an *in-silico* binding assay of a focused library of fragments, and in lead optimization, comprehensively understanding binding pathways for cases of single ligand-receptor recognition. This last part was performed in conjunction with pharmaceutical companies.

2.1 Testing the capabilities of the linear interaction energy method in drug discovery screenings.

The first steps of the drug discovery pipeline –once the target validation is performed- pass through the systematic screening of a large number of compounds, from which several hits will be identified. This stage of the pipeline must be accurate enough to provide very few false positives while providing results in fast timescales. Molecular docking is usually the preferred method for these purposes. However, it does not account for the receptor flexibility or solvent interactions that may be crucial in specific binding modes. The LIE method had been extensively tested in the past for smaller libraries of compounds providing accurate results. With the emergence of computing infrastructures like GPUGRID, we wanted to extent the scope of the LIE method to larger libraries and test it against molecular docking.

In Publication 3.2 we focused on the benchmark case of trypsin and 1500 ligands and decays of the DUD database. We compared three different scoring functions against the LIE method. The results show that although LIE is effective at reranking ligand and decoys it does not significantly improve current molecular docking software at predicting ranking positions.

2.2 Establishment of HTMD *in-silico* binding assays for focused libraries of compounds.

Fragment-based drug discovery has been widely used in drug discovery projects since the concept was first designed in the 90s.[134] The approach has generated several drug candidates,[135] and offers advantages with regard to the screening of larger compounds: more efficient chemical space exploration, high ligand efficiency and step-wise growth of the ligands. The identification of fragment hits involves the characterization of the small compounds using sensitive techniques, although sometimes different assays will return different hits.[136,137]

High-throughput molecular dynamics in combination with Markov state models analysis have made an impact in the understanding of ligand-receptor recognition processes since the first case was published five years ago.[108] The approach is able to simultaneously provide binding poses, kinetics, and thermodynamics for the most probably binding mode, but also for secondary poses. Posterior examples in literature have focused on small set of compounds, very often on just one ligand. We wanted to extend the methodology to a larger library of fragments. In publication 3.3, we focused on a set of 42 fragments with annotated experimental data.

2.3 Collaboration with pharmaceutical companies

As mentioned, the last five years have witnessed an incredible amount of excellent work in ligand binding processes by high-throughput molecular dynamics. Both promoted by the use of GPUs (we previously mentioned the cases of one or various ligands binding to trypsin,[108,138] beta lactamase,[74] factor Xa,[98] and FBP12[81]) and specialized hardware.[71,139–142]

Therefore, the approach is gaining a lot of attention in the pharmaceutical sector and will perhaps become a standard method in the next couple of years. In these previous works, the methodology was applied to monitor the binding of single ligands to rigid proteins, but it also has a lot of potential in understanding allostery, the dynamics of the receptor or finding new druggable pockets. It would certainly complement the techniques used by the pharmaceutical industry.

In this thesis we put into practice the methodology by applying to cases of real drug discovery pipelines in enterprises. Publications 3.5, 3.6 and 7.2 have been a part of collaborative projects with Janssen pharmaceutical, Boehringer Ingelheim and Pfizer, respectively.

Chapter 3

PUBLICATIONS

3.1 Computational modeling of an epidermal growth factor receptor single-mutation resistance to cetuximab in colorectal cancer treatment.

Here, we applied an optimized protocol for binding affinity calculations by umbrella sampling to provide a molecular structure-based explanation for the S468R acquired mutation in EGFR. The mutation is known to cause resistance to treatment with cetuximab of colorectal cancer. By inspecting the bound structures of cetuximab, alternative antibody necitumumab, and three EGFR ligands, we determined the putative impact of the mutation in their bindings. To confirm the structural analysis, we performed binding free energy calculations using one-dimensional potential of mean force sampled using umbrella sampling. The method was applied to cetuximab and endogenous ligand (EGF) binding wild type and S468R mutant variants of EGFR. We predict a loss of affinity for cetuximab of at least 1kcal/mol and an increase in affinity for EGF of about 1.1kcal/mol. Although in need of experimental validation, we can propose a mechanism of inhibition where cetuximab is outcompeted by EGF leading to the treatment in this mutant EGFR ineffective. This work served as an example of the applicability of molecular modeling to rationalize drug usage in the context of personalized medicine.

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 23;53(12):3123-6. doi:10.1021/ci400456m

3.2 Reranking docking poses using molecular simulations and approximate free energy methods

Molecular docking software is commonly used in the first stages of the pipeline due to their relative good accuracy with reduced computational times. By contrast, the LIE method was usually employed in subsequent phases as it is thought to present better accuracy at the expenses of being more time-consuming. With the advances in distributed infrastructures like GPUGRID, a methodology like LIE can be carried out in hundreds of GPUs simultaneously, thus being suitable for virtual screening. In this work, we wanted to establish a protocol for LIE in large libraries of compounds and test its validity against the most used current docking methods. We found that LIE is effective in re-ranking ligands and compounds but is not significantly better than current molecular docking methods.

Lauro G, Ferruz N, Fulle S, Harvey MJ, Finn PW, De Fabritiis G. Reranking docking poses using molecular simulations and approximate free energy methods. J Chem Inf Model. 2014 Aug 25;54(8):2185-9. doi: 10.1021/ci500309a

3.3 Insights from fragment hit binding assays by molecular simulations

The use of large-scale unbiased MD simulations to obtain accurate descriptions of binding events has been demonstrated in several studies so far. However, the method has been usually restricted to studies of less than then ligands and very often only one. In this study, we wanted to test the method into the context of fragment-hit identification, where focused libraries of 30-50 compounds are screened by biophysical techniques in the search of sub-millimolar starting points. Here, we proved that in-silico binding assays (ISBAs) are very powerful for drug discovery, being able to recover binding poses, affinities, kinetics and pathways simultaneously and in an unsupervised fashion. In this case we used a target and library with experimental data in order to test our results. The simulations also provide insights into the dynamics of the receptor and its kinetic fingerprint.

Ferruz N, Harvey MJ, Mestres J, De Fabritiis G. Insights from Fragment Hit Binding Assays by Molecular Simulations. J Chem Inf Model. 2015 Oct 26;55(10):2200-5. doi: 10.1021/acs.jcim.5b00453

Ferruz N, Harvey MJ, Mestres J, De Fabritiis G. Correction to Insights from Fragment Hit Binding Assays by Molecular Simulations. J Chem Inf Model. 2016 Oct 24;56(10):2123. doi:10.102/acs.jcim.6b00557

3.4 Binding kinetics in drug discovery

In this review we summarize current computational works describing methods to obtain kinetic estimates. The current state of the art, challenges and the use of adaptive sampling methods is discussed.

Ferruz N, De Fabritiis G. Binding Kinetics in Drug Discovery. Mol Inform. 2016 Jul;35(6-7):216-26. doi: 10.1002/minf.201501018

3.5 Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme

In this work we collaborated with Janssen pharmaceuticals that wanted to understand the binding mechanism of myo-inositol monophosphatase, the target for bipolar disorder. The target, a homodimer, contains a very polar pocket and depends upon the binding of three Mg^{2+} ions for its activity. Although a lot of research focused in this target in the last decades, the inhibitors discovered in the 1990s were not CNS drug-like and research became to a halt. Most inhibitors are highly polar and contain substrate-like phosphate of inositol mimics that results in problems in cell permeation and brain penetration. In order to look for a new series of more bioavailable molecules, we first attempted to characterize the IMPase's behaviour under the presence and absence of substrate. The results are robust and also show how the HTMD methodology can be applied for the case of multibody binding mechanisms.

Ferruz N, Tresadern G, Pineda-Lucena A, De Fabritiis G. Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme. Sci Rep. 2016 Jul 21;6:30275. doi: 10.1038/srep30275

3.6 Insights from in-silico binding assays of drug-like molecules.

Ferruz N, De Fabritiis G. In-silico binding assays of drug-like molecules. Preprint.

Once assessed the application of HTMD in fragment libraries and multi-body mechanism, we collaborated with another pharmaceutical company, Boehringer Ingelheim, in order to test the method in more realistic scenarios.

Concretely, we tested the capabilities of the in-silico binding assays when performed to drug-like compounds in larger proteins. In order to do so, we were provided with the kinetic data of a derivative of the linagliptin drug binding to dipeptidyl peptidase IV, a target of diabetes type II. The work shows what is the current stage of the methodology, being accurate at predicting poses and on-rates, but having limitations when estimating residence times.

Additionally, we were also able to map the route of binding of the ligand, which was hypothesized, to occur via a large opening. We show in this work that entrance and exit via the smaller opening might also be possible at longer timescales.

Insights from *in-silico* binding assay of drug-like molecules

Supporting Information Placeholder

ABSTRACT: **Accurate reconstruction of ligand binding processes by computational means has become a standard practice. However, it is usually performed in ideal systems with small fragments binding to rigid proteins, due to its considerable computational requirements. Here, we have recovered binding pathway, rates and poses for a drug-like compound binding to a large protease by means of high-throughput molecular dynamics.**

INTRODUCTION

The thermodynamics of binding are not the only important factor for drug selectivity and efficacy, but also the lifetime of the drug-receptor complex.[1–3] In this regard, the residence time of a drug -defined as the inverse of its dissociation rate, i.e off-rate or k_{off}- is known to be one of the most important factors determining safety and efficacy, and has also proven to regulate target selectivity *in vivo*.[4,5] The increased focus on kinetics in drug discovery has come also thanks to improvements in the related experimental instrumentation, surface plasmon resonance (SPR), radio-ligand binding and enzymatic assays increasingly becoming routine measurements in the pipeline.[6] However, despite several works have shown links between structural modifications and kinetic rates there is not yet a clear understanding of how structural changes affect the binding rates.[7] The latter are bounded to the height of the energy barriers between meta-stable states, in which dewetting processes,[7] shielded hydrogen bonds[8] or transient interactions[9] might be playing a critical role and pass unnoticed in our frequent 'free-or-bound' view. Thus, being able to characterize binding intermediates and the slowest transitions among them could help to rationalize a drug optimization based on kinetics. [9–13]

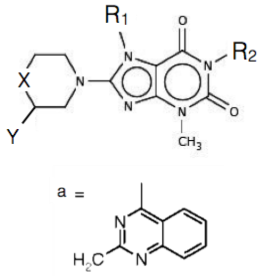
With the advent of new computing infrastructure,[14,15] unbiased simulations are becoming an affordable technique for observing drug-receptor recognitions. In unbiased MD simulations, the ligand freely moves in a solvated box and is able to identify different receptor pockets provided the simulation times are sufficient. In this sense, one of the first reconstruction of a ligand binding processes were already published few years ago.[2,9,16] In one of these works, running parallel short simulations in GPUs volunteered from all over the world in the distributed computing project GPUGRID,[17] it was possible to reconstruct a complete binding process of a small molecule (benzamidine) to its receptor (trypsin). A construction of a Markov State Model (MSM)[18,19] analysis of the aggregated 50μs of simulation time, allowed to accurately characterize metastable states affinities, and transition rates that occur in slower timescales.[9] The work showed the first complete reconstruction of a binding process by these means and provided an estimate of the dissociation rate within two orders of magnitude from the experimental value. Some other works followed this one, such as the binding of carboxythiophene to AmpC ß-lactamase,[20] where secondary X-ray poses were also recovered, or the binding of a larger library of 42 fragments against factor Xa.[21] A more recent study reconstructed the complete ligand binding framework for the trypsin-benzamidine case, taking also into account the different protein conformers in solution and the rates of inter-conversion among them.[22]

All aforementioned works most often represented toy models of drug binding, with fast small molecules probes and rigid globular proteins. For instance, benzamidine binds and unbinds with an average time of 6.9µs (assuming a concentration of 5mM) and 1.7ms to trypsin, respectively. Other studies, of course, have focused on more realistic scenarios, with drug-like size molecules binding in slower timescales. Two works reported accurate off-rate prediction for multi-millisecond residence time ligands,[23,24] although were using biasing techniques. In other investigations, using microsecond length unbiased simulations, Shaw *et. al.* successfully determined the on-rate of slow binding ligands,[2,16,25] although the residence times were not provided.

Here, we attempted to study the capabilities and limitations of the state-of-the-art methodology to reproduce residence time for a drug-like compound binding to a larger receptor. As a proof of concept, we first tested our protocol in this target by validating its binding pose and association rate, with successful outcomes. Then, we extrapolated the methodology to the more challenging case of pathway reconstruction and residence time estimation. For this test, we selected the serine exopeptidase dipeptidyl peptidase IV (dppIV). dppIV is a large protease that modulates the biological activity of specific chemokines, hormones, cytokines and neuropeptides by cleaving dipeptides after a penultimate N-terminal alanine or proline.[26] dppIV attenuates the incretins GLP-1 and GIP[27] and thus can be used in the treatment of diabetes type-II.[28–30] In May 2011, the FDA approved the use of linagliptin,[31] a xanthine derivative developed by Boehringer Ingelheim. While optimizing the drug, structural analogues were tested using different biophysical techniques, among them SPR, providing constants of a series of compounds.[32]

The reconstruction of the protein-ligand binding process was performed using our high-throughput molecular dynamics (HTMD) software,[33] which employs Markov state modelling for analysis (MSM).[34] The simulations were produced in a fully solvated system with the ACEMD[14] molecular dynamics software on the distributed computing project GPUGRID.[15] Recently, MSM analyses have been successfully used in a wide range of problems, extending from to the characterization of protein folding,[34,35] intrinsically disordered proteins,[36] and ligand binding.[37,38] In this last group, MSM methods are able to produce quantitative estimations of k_{on}, k_{off} and ΔG^0 for multiple binding poses on the protein (see **Methods)**.

The ligand employed in this study is presented in **Table 1**, along with other two structurally similar molecules whose binding mode is available in crystal structures.[32,39]



	PDB code	R_1	X/Y	R_2
1	2AJ8	CH_2Ph	NH/H	Me
2	2RGU	CH_2CCMe	CH_2/NH_2	a
3	-	CH_2Ph	CH_2/NH_2	Me

	MW (g/mol)	k_{off} (s^{-1})	k_{on} ($M^{-1}\cdot s^{-1}$)	K_D (nM)
1	355.41	-	-	2800 (IC_{50})
2	472.54	$3.0 \cdot 10^{-5}$		1nM (IC_{50})
3	368.44	$8.0 \cdot 10^{-2}$	$6.3 \cdot 10^{6}$	12.71 82 (IC_{50})

Table 1: Binding data for the ligands referred in this work and their structures.

Structures, molecular weight, kinetic and thermodynamic data when available are presented in **Table 1.** Compound 1, also known as BDPX, was discovered in a high-throughput screening campaign after showing low-micromolar inhibitory activity.[40] It was the starting point for systematic structural modifications and optimizations, from which compound 2, linagliptin, was discovered.[32] Compound 3 was present among these optimizations, and was characterized by SPR. The three compounds contain the same xanthine-based scaffold. Compound 1 and 2 and their interaction into dppIV catalytic site are represented in **Fig. 1.** The two compounds recover the same interactions, with identical parts perfectly superimposing.

Specifically, the xanthine scaffolds are placed such that the uracil moiety undergoes π-π stacking with Tyr547, thus pushing its sidechain from its relaxed position.[41] The phenyl and butynil substituents (R_1) are allocated in the hydrophobic pocket formed by Val656 and Tyr662, also termed S1 subsite following the convention for proteases.[42] Main differences between the two compounds attribute to the substituent at the C-8 of the xanthine scaffold, which occupies the S2 subsite. In compound 1, the piperazine needs to adopt an unfavourable twist conformation, and its secondary nitrogen atom donates two hydrogen bonds to the carboxylates E205 and E206. The great improvement in affinity in compound 2 comes from the replacement of the piperazine with an aminopiperidine, which can instead form 3 hydrogen bonds and adopt a low-energy chair conformation.[32] Both compounds act close to the catalytic triad (Ser630, Asp708, His740), and block the peptide cleavage. Compound 3, presents similarities with the other compounds. The xanthine and benzyl group are maintained without modifications from compound 1 and should capture the same interactions with dppIV residues. Conversely, the substituent at the C-8 position is identical to compound 2 and should adopt the favourable chair conformation.

Regarding the quaternary structure of the protein, it is usually found as a dimer or tetramer in crystal structures (**Fig. S1**).[41,43–45] However, both monomeric and dimeric forms are known to bind with similar affinity to adenosine deaminase.[46] Each of the monomers contains a β-propeller and a catalytic centre, which together encircle a large cavity. This cavity, can be accessed through two openings, more concretely, the propeller opening, also present (although slightly narrower) in the sequentially related prolyl oligopeptidase (POP),[47] and the relatively large side opening. It is been hypothesized for long whether the access of substrates and products occurs via the propeller or the side opening.[39,48–50]

For the complete characterization of the binding event, we focused on compound 3, for which full kinetic and thermodynamic data are available. After simulation of 487μs and analysis by MSM (see **Methods** for details on system setup and analysis), we obtained a most probable state that overlaps with the crystal structure of compound 1, with an RMSD of 3.5Å accounting heavy atoms in the xanthine and benzyl moieties (**Fig. 2**). The overall precision for this pose is quite good, showing an RMSD of 2.1Å. Note that no information of the bound modes was provided to the simulations in any way. Among the main differences with the crystal are the angle between the two planes formed by the xanthine moieties, and the consequent displacement of the benzyl ring from the X-ray position. More concretely, the xanthine plane bends up to 90 degrees compared the X-ray bound reference structure. Tyr547, preserving the π-π stacking with the xanthine moiety, is accordingly displaced from its initial coordinates. Interestingly, the Tyr547 phenol resembles the conformation of the *apo*-enzyme, where it is displaced 70° towards the Ser552 sidechain. (**Fig. S2**).[41,51] As a consequence of the overall displacement, the aminopiperidine ring is only able to form the two hydrogen bonds with E206 and E205 being around 5Å away from Tyr662.

The MSM model produced quantitative on-rate estimation: we obtained a value of the order of 10^6 s^{-1} M^{-1}, remarkably close to the experimental reference ($6.3 \cdot 10^6$ s^{-1} M^{-1}). When computing the on-rate by binding frequency, we observe that one binding trajectory in the set completed a binding event within an RMSD of 2Å to the crystal structure (see **Video S1**). Taking into account the total simulation time, it computes a binding frequency of $2{\cdot}10^3$ s^{-1}, which transforms to an on-rate $1{\cdot}10^6$ $s^{-1}{\cdot}M^{-1}$, in the same line as the MSM estimative. The off-rate, however, is several orders of magnitude above of the experimental value.[19] Why this is the case it is not clear, but probably due to limitations in current MSM analysis, projections and discretization used. Video S1 shows the steps the one simulation that completed a binding event performed before entering the pocket. The route of entry of substrates to dppIV has remained an answered question for a while. Our analysis confirm the substrate enters through the side opening, showing the molecule first performing some short interactions at the entrance, and then fast identifying the catalytic pocket.

In order to provide off-rate estimations and observe other unbinding/binding events that could confirm the observed pathway, we tried to produce a second multi microsecond-long simulation ensemble. In this set, taking dppIV monomer as the protein coordinates, we docked compound 3 into the catalytic pocket, using ligand 2's coordinates in 2RGU[32] as the reference for the alignment. Compound 3's off-rate, as presented in Table 1, is 0.08s^{-1}. The off-rate is a zero-order constant independent of the concentration, and thus it directly computes a residence time of 12.5s. Using a brute-force simulation scheme, as the one performed for the estimation of binding events would require a multi-second simulation trajectory, far beyond any MD ensemble simulated to date. For this reason, and with the advent of more efficient protocols we performed an adaptive sampling scheme[52] focused on the distance between the ligand and the pocket's most characteristic residues. Adaptive sampling methods, without biasing, attempt to enhance the sampling by spawning simulations into successive epochs along the binding pathways. This way, estimation of off-rates compared to brute force sampling can be achieved one or a few orders of magnitude faster.[52–54] A total of 7 epochs and 108μs simulation time were needed until the ligand freely diffused in bulk. We then produced an MSM analysis as previously done. The MSM provided a residence time estimate of the order of 10^{6}ns (or 0.001s), which compared to the experimental value of 12.5, is still 4 orders of magnitude faster.

From the set of 435 simulations, 13 explored sites with an RMSD of at least 30Å with regard to the bound position. Their evolution is summarized in **Fig. 3**. All the simulations come from four original simulations in the first epoch (1-4). From these four, which ended up in states 2, 2' and 2'', another four were respawned, one of them shifting to state 3, in the propeller. Twelve simulations were respawned from the latter, at different points. In overall, from the 13 simulations reaching states significantly distinct to bound, 11 of them reached bulk through the side opening, while two of them partially exited through the propeller opening. We then searched for a possible complete entrance or exit through the propeller opening in the first set. To our surprise, we found that 6 out of the 2000 came into the proximity of the propeller, but only one of them performed long interactions. This simulation did not perform a complete passage through the propeller, staying invariantly at the propeller side during its length.

In conclusion, we have presented an example of a ligand-protein system whose recognition was characterized by atomistic molecular dynamics. Running parallel 200ns-length simulations in the distributed computer project GPUGRID we produced two ensembles totalling 0.5ms. Owing to the advances in sampling protocols and improvements in the analysis by MSM we were able to obtain the binding on-rate with remarkable accuracy. The most probable pose reproduced the expected binding mode as observed in available crystal structures from structurally very similar compounds, with the main difference being the xanthine scaffold twist. Tyr547, which plays a critical role in the enzymatic mechanism and runs parallel to the xanthine moiety, was found in the same conformation as in the *apo*-dppIV form. The aminopiperidine ring adopts the expected low-energy chair conformation H-bonding with E205 and E206 via hydrogen bonds although the third bond with Tyr662 is not formed. We found the substrate egresses from the binding site via the side opening. From the thirteen simulations that reached distant states from the pocket, 11 simulations transferred to bulk. The other two, instead, interacted in the propeller opening and partially exited through it from the inner cavity. We identified another simulation in which the ligand accesses from bulk, but neither performed a complete entrance event. In overall, these results suggest the route of entry occurs via the side opening as anticipated by previous hypothesis, although we have observed partial entrances and escapes through the propeller opening. We conclude this secondary pathway, is indeed a possible route, although highly depends on the size of the ligand or substrate.

The estimated residence time, although certainly shorter than the experimental -by 4 orders of magnitude- still the best approximation feasible with the current analysis and computational techniques. It is very good that one hundred microseconds of unbiased adaptive sampling managed to unbind a compound that is supposed to have a residence time of 12.5 seconds. Yet we still fail to produce a good approximation of the off-rate even having observed several unbinding events. It is not clear why this is the case but one possibility is a poor projection space, in this case contact maps, and a poor clustering over this space. We are currently investigating optimal dimensionality reduction methods in order to be able to tackle drug-like compounds in future. We argue that with the fast advance in the proper construction of MSM occurring in the recent years, overcoming the barrier of the second will occur in the near future. Finally, this work attempts to push the current computational and analysis limits to characterize realistic processes in drug discovery projects.

ASSOCIATED CONTENT

Methods, supporting text and figures are included in the Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org

AUTHOR INFORMATION

Corresponding Author

gianni.defabritiis@upf.edu

Author Contributions

NF built the systems, analysed the simulations and wrote the manuscript. GF wrote part of the software, designed and supervised the project.

Funding Sources

No competing financial interests have been declared.

ACKNOWLEDGMENTS

NF acknowledges support from Generalitat de Catalunya (FI-Agaur). GDF acknowledges support from MINECO (BIO2014-53095-P) and FEDER.We also thank all the volunteers of GPUGRID who donate GPU computing time to the project.

REFERENCES

(1) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular Determinants of Drug–receptor Binding Kinetics. *Drug Discov. Today*.

(2) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133* (24), 9181–9183.

(3) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug–target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* **2006**, *5* (9), 730–739.

(4) Moulton, B. C.; Fryer, A. D. Muscarinic Receptor Antagonists, from Folklore to Pharmacology; Finding Drugs That Actually Work in Asthma and COPD. *Br. J. Pharmacol.* **2011**, *163* (1), 44–52.

(5) Gavaldà, A.; Ramos, I.; Carcasona, C.; Calama, E.; Otal, R.; Montero, J. L.;

Sentellas, S.; Aparici, M.; Vilella, D.; Alberti, J.; Beleta, J.; Miralpeix, M. The in Vitro and in Vivo Profile of Aclidinium Bromide in Comparison with Glycopyrronium Bromide. *Pulm. Pharmacol. Ther.* **2014**, *28* (2), 114–121.

(6) Keserü, G.; Swinney, D. C. *Thermodynamics and Kinetics of Drug Binding*; John Wiley & Sons, 2015.

(7) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133* (24), 9181–9183.

(8) Schmidtke, P.; Luque, F. J.; Murray, J. B.; Barril, X. Shielded Hydrogen Bonds as Structural Determinants of Binding Kinetics. Application in Drug Design. *J Am Chem Soc* **2011**.

(9) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* **2011**, *108* (25), 10184–10189.

(10) Miller, D. C.; Klute, W.; Brown, A. D. Discovery of Potent, Metabolically Stable Purine CRF-1 Antagonists with Differentiated Binding Kinetic Profiles. *Bioorg. Med. Chem. Lett.* **2011**, *21* (20), 6108–6111.

(11) Basavapathruni, A.; Jin, L.; Daigle, S. R.; Majer, C. R. A.; Therkelsen, C. A.; Wigle, T. J.; Kuntz, K. W.; Chesworth, R.; Pollock, R. M.; Scott, M. P.; Moyer, M. P.; Richon, V. M.; Copeland, R. A.; Olhava, E. J. Conformational Adaptation Drives Potent, Selective and Durable Inhibition of the Human Protein Methyltransferase DOT1L. *Chem. Biol. Drug Des.* **2012**, *80* (6), 971–980.

(12) Jin, M.; Petronella, B. A.; Cooke, A.; Kadalbajoo, M.; Siu, K. W.; Kleinberg, A.; May, E. W.; Gokhale, P. C.; Schulz, R.; Kahler, J.; Bittner, M. A.; Foreman, K.; Pachter, J. A.; Wild, R.; Epstein, D.; Mulvihill, M. J. Discovery of Novel Insulin-Like Growth Factor-1 Receptor Inhibitors with Unique Time-Dependent Binding Kinetics. *ACS Med. Chem. Lett.* **2013**, *4* (7), 627–631.

(13) Markgren, P.-O.; Schaal, W.; Hämäläinen, M.; Karlén, A.; Hallberg, A.; Samuelsson, B.; Danielson, U. H. Relationships between Structure and Interaction Kinetics for HIV-1 Protease Inhibitors. *J. Med. Chem.* **2002**, *45* (25), 5430–5439.

(14) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, *5* (6), 1632–1639.

(15) Fabritiis, G. D. The GPUGRID.org website.

(16) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci.* **2011**.

(17) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *J. Chem. Inf. Model.* **2010**, *50* (3), 397–403.

(18) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131* (12), 124101.

(19) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134* (17), 174105–174105 – 23.

(20) Bisignano, P.; Doerr, S.; Harvey, M. J.; Favia, A. D.; Cavalli, A.; De Fabritiis, G. Kinetic Characterization of Fragment Binding in AmpC Β-Lactamase by High-Throughput Molecular Simulations. *J. Chem. Inf. Model.* **2014**, *54* (2), 362–366.

(21) Ferruz, N.; Harvey, M. J.; Mestres, J.; De Fabritiis, G. Insights from Fragment Hit Binding Assays by Molecular Simulations. *J. Chem. Inf. Model.* **2015**.

(22) Plattner, N.; Noé, F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* **2015**, *6*, 7653.

(23) Mollica, L.; Decherchi, S.; Zia, S. R.; Gaspari, R.; Cavalli, A.; Rocchia, W. Kinetics of Protein-Ligand Unbinding via Smoothed Potential Molecular Dynamics Simulations. *Sci. Rep.* **2015**, *5*.

(24) Tiwary, P.; Limongelli, V.; Salvalaglio, M.; Parrinello, M. Kinetics of Protein–ligand Unbinding: Predicting Pathways, Rates, and Rate-Limiting Steps. *Proc. Natl. Acad. Sci.* **2015**, *112* (5), E386–E391.

(25) Dror, R. O.; Green, H. F.; Valant, C.; Borhani, D. W.; Valcourt, J. R.; Pan, A. C.; Arlow, D. H.; Canals, M.; Lane, J. R.; Rahmani, R.; Baell, J. B.; Sexton, P. M.; Christopoulos, A.; Shaw, D. E. Structural Basis for Modulation of a G-Protein-Coupled Receptor by Allosteric Drugs. *Nature* **2013**, *503* (7475), 295–299.

(26) Lambeir, A. M.; Proost, P.; Durinx, C.; Bal, G.; Senten, K.; Augustyns, K.; Scharpé, S.; Van Damme, J.; De Meester, I. Kinetic Investigation of Chemokine Truncation by CD26/dipeptidyl Peptidase IV Reveals a Striking Selectivity within the Chemokine Family. *J. Biol. Chem.* **2001**, *276* (32), 29839–29845.

(27) Mentlein, R.; Gallwitz, B.; Schmidt, W. E. Dipeptidyl-Peptidase IV Hydrolyses Gastric Inhibitory Polypeptide, Glucagon-like Peptide-1(7-36)amide, Peptide Histidine Methionine and Is Responsible for Their Degradation in Human Serum. *Eur. J. Biochem. FEBS* **1993**, *214* (3), 829–835.

(28) Villhauer, E. B.; Brinkman, J. A.; Naderi, G. B.; Burkey, B. F.; Dunning, B. E.; Prasad, K.; Mangold, B. L.; Russell, M. E.; Hughes, T. E. 1-[[(3-Hydroxy-1-Adamantyl)amino]acetyl]-2-Cyano-(S)-Pyrrolidine: A Potent, Selective, and Orally Bioavailable Dipeptidyl Peptidase IV Inhibitor with Antihyperglycemic Properties. *J. Med. Chem.* **2003**, *46* (13), 2774–2789.

(29) Kim, D.; Wang, L.; Beconi, M.; Eiermann, G. J.; Fisher, M. H.; He, H.; Hickey, G. J.; Kowalchick, J. E.; Leiting, B.; Lyons, K.; Marsilio, F.; McCann, M. E.; Patel, R. A.; Petrov, A.; Scapin, G.; Patel, S. B.; Roy, R. S.; Wu, J. K.; Wyvratt, M. J.; Zhang, B. B.; Zhu, L.; Thornberry, N. A.; Weber, A. E. (2R)-4-Oxo-4-[3-(trifluoromethyl)-5,6-dihydro[1,2,4]triazolo[4,3-A]pyrazin-7(8H)-Yl]-1-(2,4,5-Trifluorophenyl)butan-2-Amine: A Potent, Orally Active Dipeptidyl Peptidase IV Inhibitor for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2005**, *48* (1), 141–151.

(30) Augeri, D. J.; Robl, J. A.; Betebenner, D. A.; Magnin, D. R.; Khanna, A.; Robertson, J. G.; Wang, A.; Simpkins, L. M.; Taunk, P.; Huang, Q.; Han, S.-P.; Abboa-Offei, B.; Cap, M.; Xin, L.; Tao, L.; Tozzo, E.; Welzel, G. E.; Egan, D. M.; Marcinkeviciene, J.; Chang, S. Y.; Biller, S. A.; Kirby, M. S.;

Parker, R. A.; Hamann, L. G. Discovery and Preclinical Profile of Saxagliptin (BMS-477118): A Highly Potent, Long-Acting, Orally Active Dipeptidyl Peptidase IV Inhibitor for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2005**, *48* (15), 5025–5037.

(31) Allegrini, P.; ATTOLINO, E.; Artico, M. *Process for the Preparation of Linagliptin*; Google Patents, 2012.

(32) Frank Himmelsbach; Elke Langkopf; Matthias Eckhardt; Michael Mark; Roland Maier; Ralf Richard Hermann Lotz; Mohammad Tadayyon. 8-[3-Amino-Piperidin-1-Yl]-Xanthines, the Production Thereof and the Use of the Same as Medicaments.

(33) Stefan Doerr; Gianni De Fabritiis. www.htmd.org https://www.htmd.org/htmd/index.html (accessed Jan 11, 2016).

(34) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of Ab Initio Protein Folding for a Millisecond Folder NTL9(1-39). *J. Am. Chem. Soc.* **2010**, *132* (5), 1526–1528.

(35) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Atomistic Description of the Folding of a Dimeric Protein. *J. Phys. Chem. B* **2013**.

(36) Stanley, N.; Esteban-Martín, S.; De Fabritiis, G. Kinetic Modulation of a Disordered Protein Domain by Phosphorylation. *Nat. Commun.* **2014**, *5*.

(37) Held, M.; Noé, F. Calculating Kinetics and Pathways of Protein–ligand Association. *Eur. J. Cell Biol.* **2012**, *91* (4), 357–364.

(38) Lawrenz, M.; Shukla, D.; Pande, V. S. Cloud Computing Approaches for Prediction of Ligand Binding Poses and Pathways. *Sci. Rep.* **2015**, *5*, 7918.

(39) Engel, M.; Hoffmann, T.; Manhart, S.; Heiser, U.; Chambre, S.; Huber, R.; Demuth, H.-U.; Bode, W. Rigidity and Flexibility of Dipeptidyl Peptidase IV: Crystal Structures of and Docking Experiments with DPIV. *J. Mol. Biol.* **2006**, *355* (4), 768–783.

(40) Purine-2,6-Diones Which Are Inhibitors of the Enzyme Dipeptidyl Peptidase Iv (dpp-Iv).

(41) Thoma, R.; Löffler, B.; Stihle, M.; Huber, W.; Ruf, A.; Hennig, M. Structural Basis of Proline-Specific Exopeptidase Activity as Observed in Human Dipeptidyl Peptidase-IV. *Struct. Lond. Engl. 1993* **2003**, *11* (8), 947–959.

(42) Schechter, I.; Berger, A. On the Size of the Active Site in Proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **1967**, *27* (2), 157–162.

(43) Oefner, C.; D'Arcy, A.; Mac Sweeney, A.; Pierau, S.; Gardiner, R.; Dale, G. E. High-Resolution Structure of Human Apo Dipeptidyl Peptidase IV/CD26 and Its Complex with 1-[([2-[(5-Iodopyridin-2-Yl)amino]-Ethyl]amino)-Acetyl]-2-Cyano-(S)-Pyrrolidine. *Acta Crystallogr. D Biol. Crystallogr.* **2003**, *59* (Pt 7), 1206–1212.

(44) Rasmussen, H. B.; Branner, S.; Wiberg, F. C.; Wagtmann, N. Crystal Structure of Human Dipeptidyl Peptidase IV/CD26 in Complex with a Substrate Analog. *Nat. Struct. Biol.* **2003**, *10* (1), 19–25.

(45) Weihofen, W. A.; Liu, J.; Reutter, W.; Saenger, W.; Fan, H. Crystal Structure of CD26/dipeptidyl-Peptidase IV in Complex with Adenosine Deaminase Reveals a Highly Amphiphilic Interface. *J. Biol. Chem.* **2004**, *279* (41), 43330–43335.

(46) Chien, C.-H.; Tsai, C.-H.; Lin, C.-H.; Chou, C.-Y.; Chen, X. Identification of

Hydrophobic Residues Critical for DPP-IV Dimerization. *Biochemistry (Mosc.)* **2006**, *45* (23), 7006–7012.

(47) Fülöp, V.; Böcskei, Z.; Polgár, L. Prolyl Oligopeptidase: An Unusual Beta-Propeller Domain Regulates Proteolysis. *Cell* **1998**, *94* (2), 161–170.

(48) Engel, M.; Hoffmann, T.; Wagner, L.; Wermann, M.; Heiser, U.; Kiefersauer, R.; Huber, R.; Bode, W.; Demuth, H.-U.; Brandstetter, H. The Crystal Structure of Dipeptidyl Peptidase IV (CD26) Reveals Its Functional Regulation and Enzymatic Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (9), 5063–5068.

(49) Hiramatsu, H.; Kyono, K.; Higashiyama, Y.; Fukushima, C.; Shima, H.; Sugiyama, S.; Inaka, K.; Yamamoto, A.; Shimizu, R. The Structure and Function of Human Dipeptidyl Peptidase IV, Possessing a Unique Eight-Bladed Beta-Propeller Fold. *Biochem. Biophys. Res. Commun.* **2003**, *302* (4), 849–854.

(50) Ludwig, K.; Yan, S.; Fan, H.; Reutter, W.; Böttcher, C. The 3D Structure of Rat DPPIV/CD26 as Obtained by Cryo-TEM and Single Particle Analysis. *Biochem. Biophys. Res. Commun.* **2003**, *304* (1), 73–77.

(51) Bjelke, J. R.; Christensen, J.; Branner, S.; Wagtmann, N.; Olsen, C.; Kanstrup, A. B.; Rasmussen, H. B. Tyrosine 547 Constitutes an Essential Part of the Catalytic Mechanism of Dipeptidyl Peptidase IV. *J. Biol. Chem.* **2004**, *279* (33), 34691–34697.

(52) Doerr, S.; De Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **2014**.

(53) Bowman, G. R.; Huang, X.; Pande, V. S. Adaptive Seeding: A New Method for Simulating Biologically Relevant Timescales. *Biophys. J.* **2009**, *96* (3, Supplement 1), 575a.

(54) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6* (3), 787–794.

FIGURES

Figure 1: Overview of the catalytic pocket of dppIV. Compound 1 (BDPX) and 2 (linagliptin) from Table 1 are superimposed and their protein surroundings depicted. Both compounds share the same xanthine scaffold, which stacks against Tyr547 by π-π interactions. The phenyl and butynil groups in each case extend towards the S1 subsite, whereas the piperazine and aminopiperidine occupy the S2 subsite.

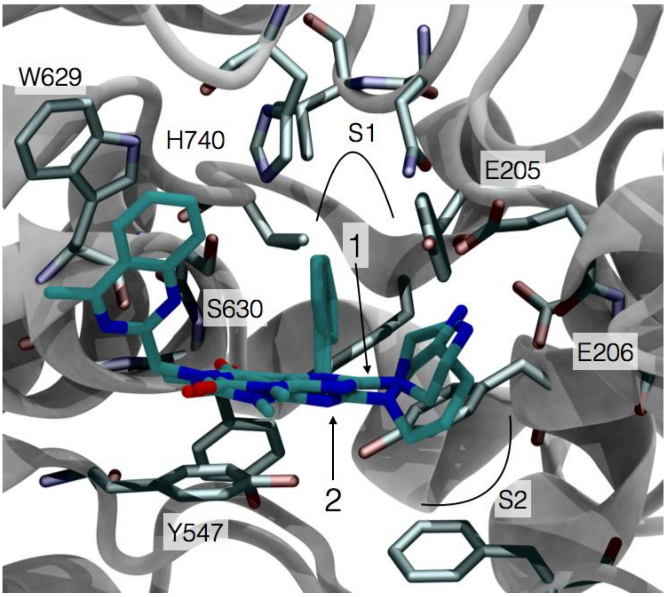


Figure 2: Binding pose obtained through MSM analysis in comparison with crystal bound pose of Compound 1 (BDPX) (pdb code 2AJ8). The main difference between the poses is the overall bending of the simulated pose, around 90 degrees for the planes of the two scaffolds. The aminopiperidine ring is expected to form three hydrogen bonds. As consequence of the overall bending of the structure it only forms two with E205 and E206.

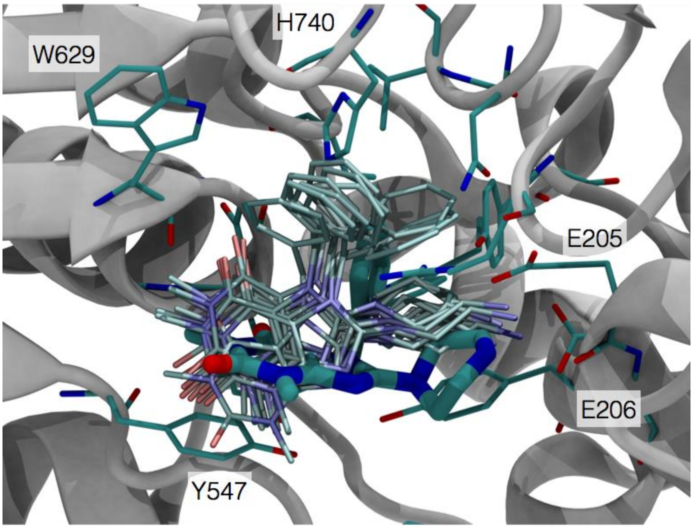
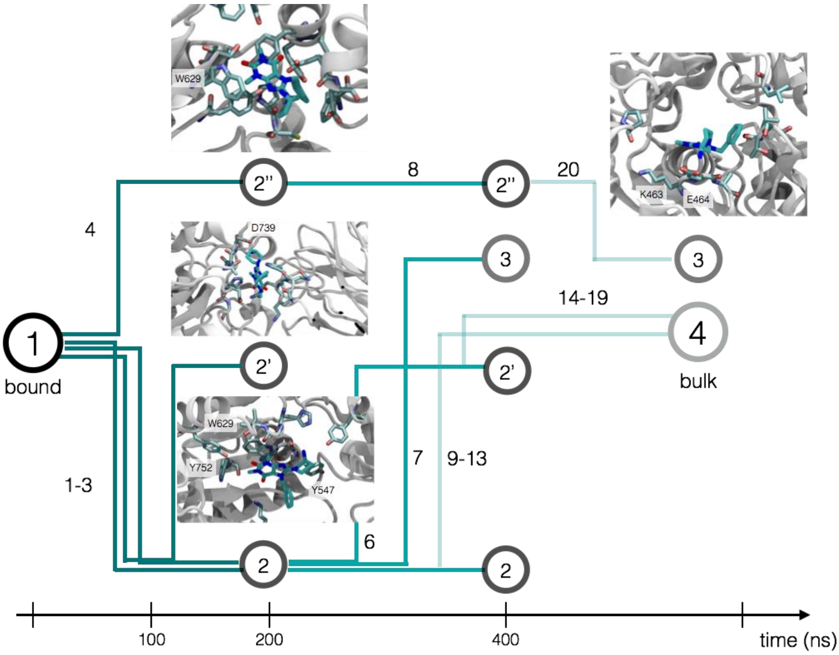


Figure 3: Temporal evolution of the 13 simulations that reached states distant away from bound. The trajectories were reconstructed from their parent simulations in previous epochs. In all cases except one reconstructed trajectories comprised three epochs. The colours indicate the different epochs, and the numbers refer to the order they appear in the main text. There are in total are 20 steps of 200ns. These simulations started from four initial simulations (1-4) that finished in states 2, 2' and 2''. Their interactions with dppIV are depicted on each picture, being states 2 and 2'' still in contact with the pocket and 2'' at the side opening. In the second epoch of simulations, two simulations stayed in their states while the other two shifted to others. In the last epochs, 5 and 6 simulations were respawned from states 2 and 2'', while one from state 2''. In overall, from the 13 trajectories, the ligand transferred to bulk in 11 via the side opening, while identified and remained at the propeller opening in the other two.



Supporting Information

TEXT

1. METHODS

Simulation system setup and simulation parameters

Input coordinates for DPP-IV monomer were based on the pdb code 2RGU, chain A.[1] The AMBER FF12SB[2,3] forcefield was used to describe all the protein parameters. Compound 2 was protonated with the OpenBabel software at pH 7.4[4] and parameterized by the Antechamber 12 tool.[3] All the complexes were explicitly solvated by the LEAP module of the AMBER 12 software package in a TIP3P[5] cubic water box with at least 15 Å distance around the complex and then electrically neutralized using Na+ and Cl- ions. All the systems always contained one ligand per box giving a final concentration of 0.0021 M. Final size systems were about 91000 atoms giving cubic boxes of 100 Å per side.

Each system was minimized and relaxed under NPT conditions for 1ns at 1atm and 298K using a time-step of 4 fs, rigid bonds, cut-off of 9Å and PME for long range electrostatics. Heavy protein and ligand atoms were constrained by a 10 kcal/mol/Å^2 spring constant. Production simulations were run using ACEMD[6] over GPUGRID[7] in the NVT ensemble using a Langevin thermostat with damping of 0.1 ps^{-1} and hydrogen mass repartitioning scheme to achieve timesteps of 4 fs.[8] For the analysis of the binding pathway, more than 2000 brute-force[9] simulations of 210 ns length were performed, giving an aggregate of 487 μs of simulation time. All the trajectories started with the ligand placed in different positions in solution, conforming an isoenergetic ensemble from which the on-rate can also be computed by binding frequency. An adaptive sampling method[10] was run used for the analysis of the off-rate, for which 7 epochs of 435 trajectories 200 ns length and 209 trajectories 100ns-length were produced. The sets used for analysis were 487 and 108 μs length, respectively.

Markov State Model

A Markov state model (MSM) for each of the systems was built from the molecular simulation trajectories. MSMs have been successfully used to reconstruct the equilibrium and kinetic properties in a large number of molecular systems.[9,11,12] By determining the frequency of transitions between conformational states we were able to construct a master equation which describes the dynamics between a set of conformational states. Relevant states are determined geometrically by clustering the simulation data onto a metric space (e.g. contact maps). In this case, a discrete description of the process was obtained by means of protein-ligand contact maps for the on rate estimation, using all heavy atoms of the ligand. Two atoms are in contact if their distance is less than 8 Å. The second set used distances between the ligand heavy atoms and the protein alpha carbons as the metric. The two analyses were performed as follows. First, one of the most important requirements for constructing Markov models is to be able to finely discretize the slowest order parameters. TICA[13] (time-lagged independent component analysis) is a method that projects the data on the slow order parameters, thus producing a very good discretization. After projecting the high- dimensionally protein-ligand contact maps onto the ten slowest processes found by TICA with a 2 ns lag-time, the 10-dimensional projected data was clustered using the k-means algorithm to produce a Markov model, producing around 3000 clusters in each case. The master equation is then built as:

$$\acute{P_\iota}(t) = \sum_{j=1}^{N}\left[k_{ij}P_j(t) - k_{ji}P_i(t)\right] = K_{ij}P_j(t) \qquad (1)$$

Where $P_i(t)$ is the probability of state i at time t, and k_{ij} are the transition rates from j to i, and $\mathbf{K} = (K_{ij})$ is the rate matrix with elements $K_{ij} = k_{ij}$ for $i \neq j$ and $K_{ii} = -\sum_{j\neq i} k_{ji}$. The master equation $d\mathbf{P}/dt = \mathbf{K}\ \mathbf{P}$ has solution with initial condition P(0) given by $\mathbf{P}(t) =$

T(t) **P**(0), where we defined the transition probability matrix $T_{ij}(t)$ = $(\exp[\mathbf{K}t])_{ij}$ = p(i,t|j,0), i.e. the probability of being in state i at time t, given that the system was in state j at time 0. In practical terms, $p_{ij}(\Delta t)$ is estimated from the simulation trajectories for a given lag time Δt using a maximum likelihood estimator compatible with detailed balance.[14] The eigenvector $\boldsymbol{\pi}$ with eigenvalue 1 of the matrix $T(\Delta t)$ corresponds to the stationary, equilibrium probability. Higher eigenvectors correspond to exponentially decaying relaxation modes for which the relaxation timescale is computed by the eigenvalue as $\tau_s = \frac{\Delta t}{log(\lambda_s)}$, where λ_s is to the largest eigenvalue above 1. For long enough lag times Δt the model will be Markovian, however every process faster than Δt is lost. Therefore the shortest lag is chosen for which the relaxation timescales do not show a dependence on the lag time Δt anymore. In our case, we chose a lag time of 100ns depending on the fragment as it showed the least dependence for the slowest processes the two cases. Furthermore, although this fine discretization provides very good Markov models, it is needed to reduce the amount of states to obtain a humanly interpretable model of the system in question. Therefore, the initial ~3000 microstates can be lumped together into x macrostates using kinetic information from the MSM eigenvector structure. Mean first passage times and commitor probabilities can also be calculated to obtain the relevant kinetics of the system.[15] Hence, the produced clusters were then lumped together into 7 macrostates using the PCCA algorithm, each consisting of a set of kinetically similar clusters. Electrostatic plots were performed with the PMEPot plugin from VMD.[16,17]

FIGURES

Figure S2: Overall structure of DPP-IV and simulation box. (a) DPP-IV crystallizes as a tetramer (pdb 2AJ8).[48] Each monomer presents two openings to a large cavity, the propeller opening, and the side opening. Substrates and products are hypothesized to enter through the latter. (b) Example of the setting for a random simulation box. The receptor was modelled as a monomer. Side and propeller openings are shown in the xy and yz orientations, respectively. Compound 2 was always placed at least 15Å apart from the protein surface in all the starting configurations. (Table 1, main text). The simulation boxes were 100 Å per side.

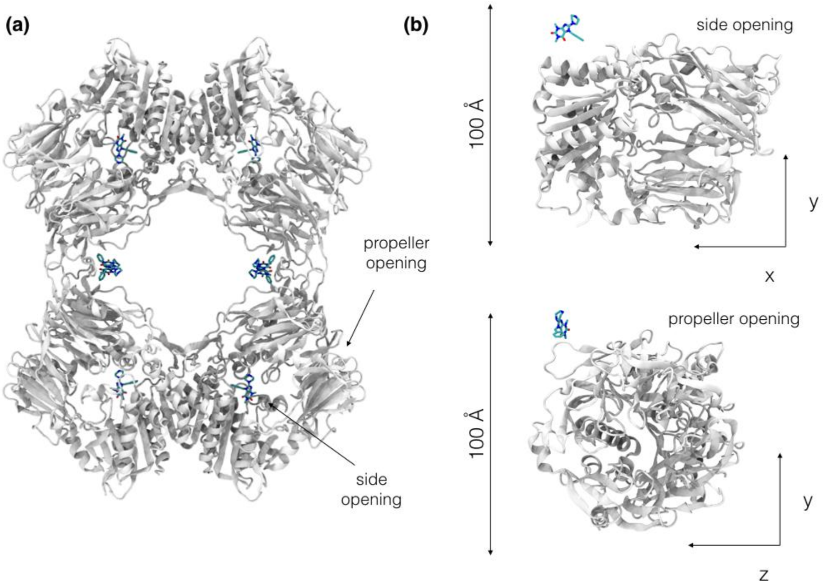
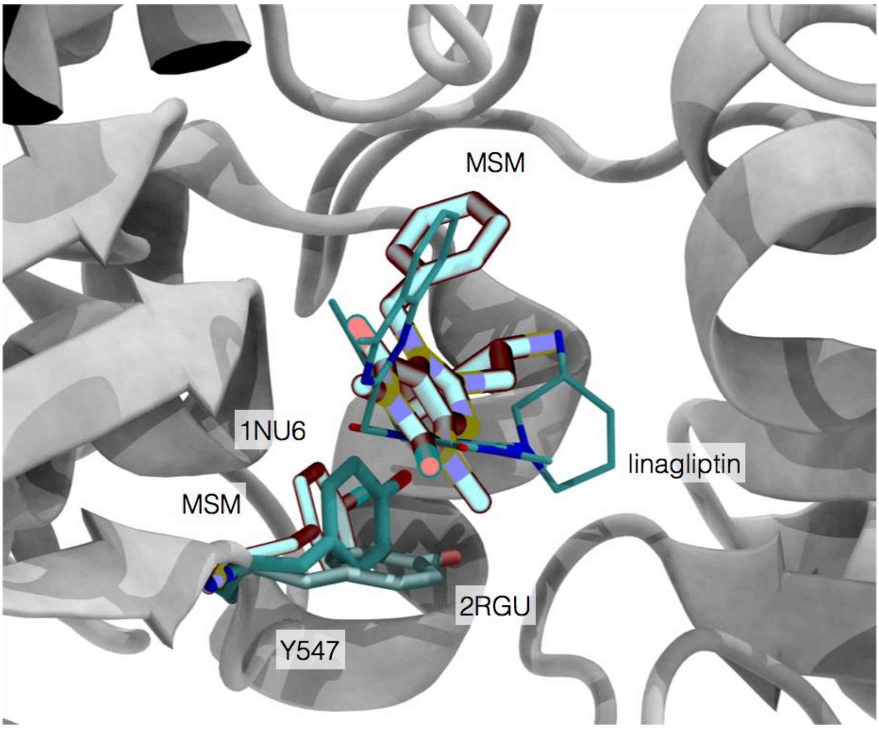


Figure S2: Tyr547 comparison for crystal and simulated binding mode. The conformation captured through MSM analysis, resembles that of the free enzyme (1NU6)[18] and compound 2 (linagliptin, 2RGU)[1]



References

(1) Eckhardt, M.; Langkopf, E.; Mark, M.; Tadayyon, M.; Thomas, L.; Nar, H.; Pfrengle, W.; Guth, B.; Lotz, R.; Sieger, P.; Fuchs, H.; Himmelsbach, F. 8-(3-(R)-Aminopiperidin-1-Yl)-7-but-2-Ynyl-3-Methyl-1-(4-Methyl-Quinazolin-2-Ylmethyl)-3,7-Dihydropurine-2,6-Dione (BI 1356), a Highly Potent, Selective, Long-Acting, and Orally Bioavailable DPP-4 Inhibitor for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2007**, *50* (26), 6450–6453.

(2) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(3) Amber Tools 12 manual. http://ambermd.org/doc12/AmberTools12.pdf.

(4) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3*, 33.

(5) Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J Phys Chem A* **2001**, *105* (43), 9954–9960.

(6) Harvey, M. J.; Giupponi, G.; Fabritiis, G. D. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, *5* (6), 1632–1639.

(7) Fabritiis, G. D. The GPUGRID.org website.

(8) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. Improving Efficiency of Large Time-Scale Molecular Dynamics

Simulations of Hydrogen-Rich Systems. *J. Comput. Chem.* **1999**, *20* (8), 786–798.

(9) Ferruz, N.; Harvey, M. J.; Mestres, J.; De Fabritiis, G. Insights from Fragment Hit Binding Assays by Molecular Simulations. *J. Chem. Inf. Model.* **2015**.

(10) Doerr, S.; De Fabritiis, G. On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations. *J. Chem. Theory Comput.* **2014**.

(11) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* **2011**, *108* (25), 10184–10189.

(12) Sadiq, S. K.; Noé, F.; Fabritiis, G. D. Kinetic Characterization of the Critical Step in HIV-1 Protease Maturation. *Proc. Natl. Acad. Sci.* **2012**.

(13) Pan, A. C.; Roux, B. Building Markov State Models along Pathways to Determine Free Energies and Rates of Transitions. *J. Chem. Phys.* **2008**, *129* (6), 064107.

(14) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1), 015102–015102 – 13.

(15) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134* (17), 174105–174105 – 23.

(16) Singhal, N.; Snow, C. D.; Pande, V. S. Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin. *J. Chem. Phys.* **2004**, *121* (1), 415.

(17) Aksimentiev, A.; Schulten, K. Imaging Alpha-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophys. J.* **2005**, *88* (6), 3745–3761.

(18) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38.

(19) Thoma, R.; Löffler, B.; Stihle, M.; Huber, W.; Ruf, A.; Hennig, M. Structural Basis of Proline-Specific Exopeptidase Activity as Observed in Human Dipeptidyl Peptidase-IV. *Struct. Lond. Engl. 1993* **2003**, *11* (8), 947–959.

Chapter 4

DISCUSSION

Previous works show the applicability of the methodology in different contexts. In this section, we discuss the implications of these results and future challenges to overcome.

4.1 Testing the capabilities of the linear interaction energy method in drug discovery screenings.

Scoring functions are widely used in the screening stages of the pipeline while the LIE method has usually been postponed to later phases, due to its computational requirements. However, with the advances in MD-related instrumentation, the simulation time required for the screening of a large library is now affordable. In publication 3.2 we tested the LIE method as a screening technique with ascribed better accuracy. We selected trypsin and a set of ligands and decoys from the DUD database,[143] giving a total of 1546 compounds, the largest application of the LIE method to date. We found in our study, by comparing the LIE method against three different molecular docking softwares (AutoDock Vina, Glide and GOLD) that its capabilities as a virtual screening tool are moderate. Considering the cost of setting up the simulations and running them (estimated computational time of 8 days in an in-house 100 GPU-sized cluster), the method does not provide any particular advantage versus known docking methods, which rank compounds with better accuracy in shorter times.

Although the take-home message seems somehow negative, the simulations are released to the scientific community and might be useful to the development of other methods or the optimization of the scaling factors. The results also further settle the docking methods –continuingly improving- as the best alternative for virtual screening. Subsequent to this publication, an improvement of the LIE method was published that enhances its accuracy and efficiency.[144,145] Although only a small set of systems were considered in these studies, perhaps that with these improvements and others the LIE method can successfully be applied as a virtual screening tool in the near future. Establishing MD-based methods like LIE or MM-PBSA/GBSA as a complementary tool in virtual screening may confer advantages, particularly when dealing with flexible receptors or water-mediated binding modes.

4.2 Establishment of HTMD *in-silico* binding assays for focused libraries of compounds.

In publication 3.3, we were able to implement and establish HTMD for *in-silico* binding assays (ISBAs) in a library of fragments. The method was fully developed and showed to agree with experimental results. 12 of 15 crystallographic poses were predicted with high accuracy, and affinity estimates in 4 out 6 cases. But, rather than in the successful cases, already highlighted in the publication, we will focus in the cases in which we failed to predict the expected value, or the problems we came across during the development of this project, which offer an invaluable input for improvement of the method and overview of current challenges.

Firstly, we will look at the fragments that failed to reproduce experimental data: fragments 27, and 36, which did not reproduce the affinity, and fragments 19, 35 and 38, which did not recover the pose (**Fig. 12**).

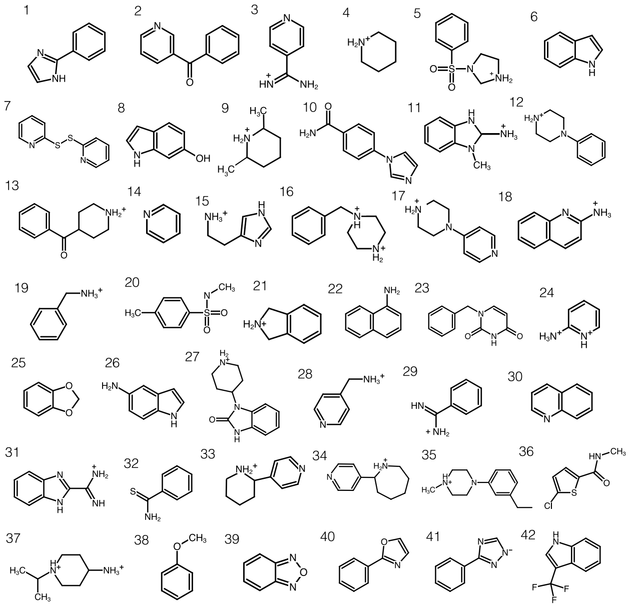
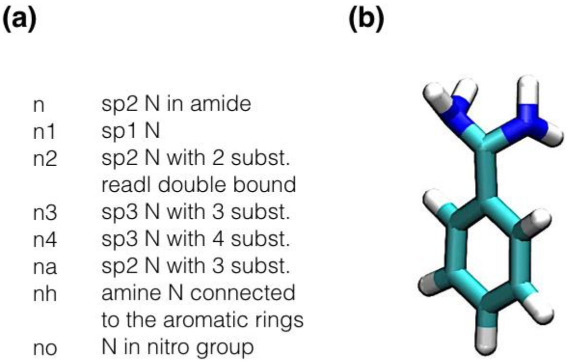


Figure 12: Library of fragments used in publication 3.3: fragments 19, 27, 35, 36 and 38 failed to reproduce expected values and provide the basis for improvement.

Fragment 27 is the largest-sized compound in this library. We run 63μs of simulation time for it, -the average was 50μs- and still we might have needed larger simulation times to sample a few binding events. For the purpose of understanding, assuming that compound size relates to association rates,[142] and seeing all other fragments were in the range 10^7-10^8 $M^{-1}s^{-1}$ in their k_{on}, we could hypothesize this ligand binds with a high 10^6 on-rate. The ISBA experiments were run with a concentration of 3.7mM, and therefore, 54μs of sampling time would be needed if $k_{on}=5 \cdot 10^6$ $M^{-1}s^{-1}$. Therefore, in our binding set, -which in this case was composed of 50μs of brute-force simulations and 13μs of adaptive sampling; we perhaps sampled the bound pose too few times. Of course, there might be other reasons for the lack in accuracy, but still serves an example to take into account for future projects. Estimation of the sampling times requirements, although roughly performed, could avoid encountering false negatives.

We also found another interesting case when looking into fragment 36. This compound (chlorotiophene) appears as a moiety in the rivaroxaban drug, an oral anticoagulant.[146] It is part of a new generation of bioavailable factor Xa neutral inhibitors, whose requirements for affinity are driven by the chlorine-Tyr228 interaction,[147–150] instead of the classic basic, poorly bioavailable, amidine interaction. The chemical nature of this interaction is the halogen bond: a non-covalent interaction which is driven between halogen atoms and partially negative molecules.[151] In the case of factor Xa, the interaction is produced between the chlorine and the negative π-cloud of the tyrosine phenyl ring. Although halogens are very electronegative, Cl, Br and I present an anisotropy in their charge distribution, with an external region of positive electrostatic potential termed sigma-hole. However, the current parameterization of the interaction is reduced nowadays to the modelling of a virtual massless particle with a point partial charge, with the disadvantage of the spatial constraints that small pockets might perform. Current implementations present different degrees of automatization, ranging from fast general virtual particle introduction to more elaborated methods requiring specific optimization.[152–154] At the moment, we are currently working on how to integrate the halogen bonding in combination with more accurate parameterization QM methods in an automatic fashion.

The halogen bond is only a particular example of the drawbacks of fast parameterization protocols. Concretely, we could say that the accurate fast parameterization of chemical entities is the current ‘Achilles’ heel’ of the MD focused on drug discovery. Previously mentioned in section 1.2.1, both mostly used AMBER and CHARMM forcefields offer the general-purpose forcefields GAFF and CGenFF, respectively. However, although very fast, these forcefields sometimes lack the accuracy required at the stages of hit identification or optimization as performed with HTMD. During the accomplishment of this work, we found one exemplifying case of the fast parameterization problems. When automatically parameterizing the set of fragments in **Fig. 13**, a library especially enriched with amidine compounds, we found the GAFF atomtype designation was faulty in the nitrogen. Different antechamber versions[155] assigned nh as the atomtype for the simplest case of benzamidine (**Fig. 13a**), which translated to non-planarity when in simulation (**Fig. 13b**). If we manually inserted n2 or na as the atomtype to keep planarity, the dihedrals did not behave as expected from experiments or QM calculations.[156] We finally parameterized this case by combining Gaussian[157] and Antechamber. However, this procedure requires unacceptable levels of human intervention.



n	sp2 N in amide
n1	sp1 N
n2	sp2 N with 2 subst. readl double bound
n3	sp3 N with 3 subst.
n4	sp3 N with 4 subst.
na	sp2 N with 3 subst.
nh	amine N connected to the aromatic rings
no	N in nitro group

Figure 13: (a) GAFF Atomtype designation for nitrogen atoms **(b)** Benzamidine’s hybridization is at the nitrogen atoms when automatically parameterized.

We mentioned two examples of problematic parameterization, the lack in accuracy for the concrete amidine case –note that there might be others- and the need to properly represent the halogen bond interactions –similarly, there are other determinants of ligand-receptor recognition currently not properly accounted: induced electronic polarizability,[158] changes in protonation states upon binding[159] or the existence of tautomers[160]-. The clear challenge for the next years is the implementation of an accurate fast protocol that incorporates these interactions while producing robust reliable parameterizations in an automated fashion.

The cases where we failed to reproduce the pose do not have so specific attributable reasons. Although, fragment 35 contains also more heavy atoms than the average in the set, we are also aware that fragments 19, 35 and 38 probably arise from a combination of all the current limitations of the methodology: accuracy of the forcefield and parameterization, need for larger sampling times and clustering methods used in the MSM production.[161]

Secondly, when performing the individual ISBAs, we observed the flipping of Trp215 towards the S1 site during the simulations where fragments performed short-lived interactions in the S4 pocket (**Fig. 8c**). We then simulated *apo*-factorXa in order to see if the conformation was also explored in the free enzyme, confirming the Trp215 shifted to other conformations regardless of the forcefield and the presence of ligands. By further literature search we found that Trp215 acts as the handover between two conformations in factor Xa that interconvert in the millisecond timescale (**Fig. 15**).

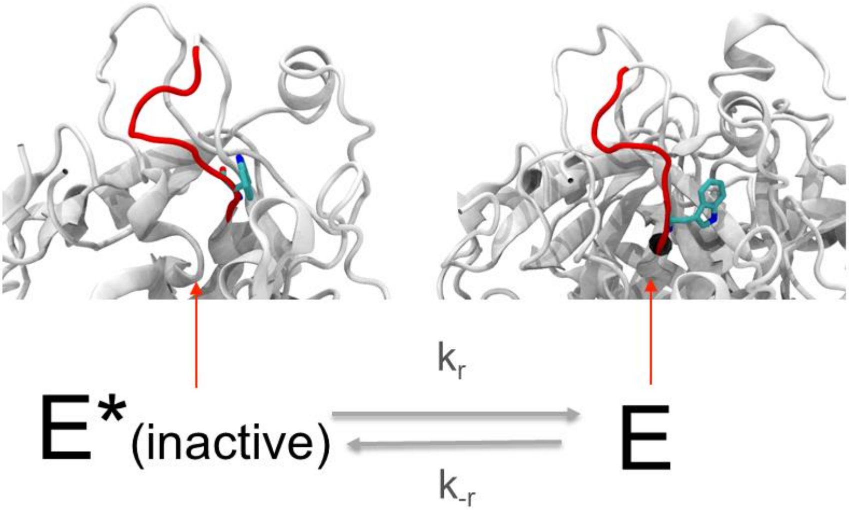


Figure 14: Different thrombin conformations. The two structures were obtained by X-ray crystallography[162] and their kinetic constants characterized,[162,163] k_r = 45 ± 2 and K_{-r} 70 ± 2 for factor Xa.

During these simulations, -and previously in trypsin,[108] also known to present this plasticity- we assumed a single-step process in which the protein, although flexible, was basically always the active conformer. The reality is that the protein presents also different conformers in solution interconverting at larger scales than our ISBAs, but for which the ligands present different affinities. Therefore, starting the simulations from one of the conformers biases our binding estimates to that single process, although in the bigger *in vivo* picture the ligand might find the receptor in other conformations as well. Very recently, and using the trypsin-benzamidine test case, Noé *et. al.*[138] characterized this protease conformational plasticity, and the relative affinities of the ligand for each of the conformers. It turned out that, for the case of trypsin-benzamidine (or factor Xa and these set of fragments), the main kinetic pathway is the direct binding to the active conformer, since it is mostly populated in solution. Therefore, our analyses were valid but could have impactful consequences when picking the least populated conformer. It is then necessary to perform a conformational analysis of the receptor and start the ISBAs from different receptor conformers following their equilibrium distribution.

Summarizing, the methodology was successfully established and applied into a medium-sized fragment library. With the reductions in sampling times provided by the adaptive sampling, the method is currently being applied to a library of 700 compounds in an unsupervised fashion. In the same way, we have now an awareness of the current limitations of the method that we can tackle, parameterization automation, more efficient sampling, clustering methods and *apo*-receptor conformational analysis. Some others problems, such as the forcefield issue, are out of our scope, but continuously under development. We hope that with further advances the technique becomes an accurate tool able to provide kinetics, poses, and thermodynamics for libraries of a few thousands of fragments in the next years.

4.3 Collaboration with pharmaceutical companies

We have successfully taken the HTMD methodology to the real world scenario, by collaborating with three big pharmaceutical companies in Publications 3.5, 3.6 and 3.7.

From the scientific point of view, the accomplishment of this works both closed and opened new questions in their fields. Myo-inositol monophosphatase is a very challenging target for the treatment of bipolar disorder and as such a good amount of research was centred on understanding the mechanism of binding of substrate of cofactors. From one side, the populations of the enzyme-cofactor complex in solution in the absence of substrate remained unknown. We concluded that the formation of the ternary complex is possible, although were not able to characterize its population due to the slow timescales. For other side, the concrete order of binding had disagreement between studies. We found that substrate binding can occur through two different pathways. It occurs in the low microsecond timescales to a ternary IMPase. It can also occur to a binary IMPase in the millisecond timescale, in complex with Mg^{2+} or alone, which quickly rearranges in the pocket after substrate binding. The real populations in solution of binary and ternary complex will tune the extents in which each of the pathways occurs, and it is a question that remains to be answered.

Dipeptidyl peptidase IV is serine exopeptidase targeted for diabetes. It cuts the penultimate aminoacid of polypeptides whose route of entry is thought to occur via the (large) side opening. We characterized the binding pathway of a drug-like compound and found that the exit and entrance through the smaller propeller opening might also be possible, and its validation remains to be addressed by other techniques. In this work we also concluded that the current methodology is able to accurately provide on-rates and binding poses, but the characterization of off-rates remains as a challenge for next years.

From the perspective of applying the HTMD in an industrial context, we believe that it will be gradually embedded in the assay routine. Molecular dynamics is being more and more introduced in the drug design industry.[164–166] Among the reasons, are the crescent notion that the receptor dynamics plays a critical role on the binding process and the possibility of computing microsecond ensembles at competitive prices. From the experience of these three works, we hope many other companies will increasingly integrate HTMD in the first steps of their pipelines (**Fig. 3**) as a complement to other established techniques.

Chapter 5

CONCLUSIONS

1. Molecular structure-based analysis coupled to binding free energy calculations in the determination of the impact of S468R mutation in EGFR in colorectal cancer therapy predicts that, resistance to cetuximab can be due to both a loss in cetuximab binding affinity and a gain in EGF affinity for the receptor. Structural analysis also suggests that alternative monoclonal antibody necitumumab might be less affected by the mutation.

2. The LIE method is not suitable for high-throughput screening purposes when compared to docking methods. Although showing satisfactory performance for predicting relative binding affinities in large databases of compounds, the cost of setting and performing simulations against turns unaffordable when compared with docking software.

3. HTMD is a suitable method for hit identification, by accurately determining poses, binding kinetics and thermodynamics simultaneously. With the awareness of its current limitations (forcefield and parameterization issues, lack of enough sampling, and conformational plasticity of receptors) and its continuous improvements, the method may become an accurate virtual screening tool in the near future.

4. IMPase is able to form binary and ternary complexes in neuronal conditions in the absence of inorganic molecules, substrates or inhibitors. The substrate mechanism is a three-body problem that follows two main pathways of binding: a fast, main pathway with *myo*-inositolphosphate binding to ternary IMPase and a slower binding to binary IMPase, in complex with Mg^{2+} cofactor or not.

5. The route of entry and egress of inhibitors to Dipeptidyl-peptidase pocket occurs mainly though the side opening, although, however, a slower, size-dependent, entry through the propeller site is also possible.

6. The estimation of on-rates and characterization of pathways of binding for drug-like compounds to large targets is currently achievable by HTMD with high accuracy. Accurate estimation of slow off-rates remains as a challenge.

Chapter 6

LIST OF COMMUNICATIONS

This section lists talks, international stays and posters that I carried out during this thesis. Publications were presented in a separated section.

Talks

- HTMD case: Comprehensively understanding inhibition, substrate and cofactor binding of myo-inositol monophosphatase. Workshop on High Throughput Molecular Dynamics, November 7-8th 2013, Barcelona, Spain.
- In silico binding assay, II Workshop on High Throughput Molecular Dynamics, November 26-27th 2015, Barcelona, Spain.
- Introduction to HTMD. Neuroscience department, Pfizer, Inc. June 13th, 2015, Cambridge, Massachusetts.

International stay

- Neuroscience department, Pfizer, Inc. May 14th – June 14th, 2015. Cambridge, Massachusetts.

Posters

- Quantitatively understanding protein-ligand interactions by high-throughput molecular simulations. EFS-EMBO research conference. On molecular perspectives on protein-protein interactions. May 25-30th, 2013, Pultusk, Poland.

- A novel method for characterization of binding kinetics, energetics and poses in fragment based drug design. Discovery Chemistry Congress conference. February 18-19th 2014, Barcelona, Spain.
- Fragment hit identification by molecular simulations. GRIB EXPO. The big data challenge. November 10th 2014, Barcelona, Spain
- HTMD integrated platform. The first automated simulation package. Spanish-Italian Medicinal Chemistry Congress (SIMCC-2015). July 12-15th, Barcelona, Spain.

Chapter 7

APPENDIX: OTHER PUBLICATIONS

This section summarizes a publication in which I contributed to a lesser extent than in the previous ones.

Publication 7.1: Emergence of Multiple EGFR Extracellular Mutations during Cetuximab Treatment in Colorectal Cancer.

Arena S, Bellosillo B, Siravegna G, Martínez A, Cañadas I, Lazzari L, Ferruz N, Russo M, Misale S, González I, Iglesias M, Gavilan E, Corti G, Hobor S, Crisafulli G, Salido M, Sánchez J, Dalmases A, Bellmunt J, De Fabritiis G, Rovira A, Di Nicolantonio F, Albanell J, Bardelli A, Montagut C. *Emergence of Multiple EGFR Extracellular Mutations during Cetuximab Treatment in Colorectal Cancer*. Clin Cancer Res. 2015 May 1;21(9):2157-66.

Publication 7.2: Potent selective D3 antagonist reveals a unique binding mode in GPCR

Once the methodology was fully validated for drug-liked compounds, we applied it to a patented drug binding to the Dopamine D3 receptor. We performed this work in collaboration with Pfizer, Inc. being able to characterize the concrete binding mode of the drug –which pushes residues in helices V and VI not previously seen- and conclude the reasons for selectivity. Unfortunately, we were not able to obtain permission for full disclosure of the data in time for this thesis, so we cannot include the manuscript that will be submitted for publications in the next months. The abstract is presented below:

Abstract

Characterizing the specific route of entry of known drugs to G-protein coupled receptors (GPCRs) and the binding mode in which they exert their therapeutic action is of inestimable value for the drug design process. Concretely, it can be particularly interesting when achieving subtype selectivity among high-sequence homology receptors becomes a challenging task. Here, by means large-scale molecular simulations we have captured this pharmaceutical process for a patented dopamine D3 receptor (D3R)-selective antagonist, whose binding mode remains unknown. Our results show a final pose that keeps some of the interactions performed by the crystallized antagonist eticlopride in the orthostheric site, but expands towards helix V and VI creating a novel pocket. Our continuous, detailed description of the binding process offers a dynamic rationale for subtype selectivity, and reveals a binding mode otherwise hardly to characterize with current experimental and computational methods.

References

(1) Buch, I.; Giorgino, T.; De Fabritiis, G. Complete Reconstruction of an Enzyme-Inhibitor Binding Process by Molecular Dynamics Simulations. *Proc. Natl. Acad. Sci.* **2011**, *108* (25), 10184–10189.

(2) General, I. J. A Note on the Standard State's Binding Free Energy. *J. Chem. Theory Comput.* **2010**, *6* (8), 2520–2524.

(3) Zhou, H.-X. From Induced Fit to Conformational Selection: A Continuum of Binding Mechanism Controlled by the Timescale of Conformational Transitions. *Biophys. J.* **2010**, *98* (6), L15–L17.

(4) Weikl, T. R.; Paul, F. Conformational Selection in Protein Binding and Function. *Protein Sci. Publ. Protein Soc.* **2014**, *23* (11), 1508–1518.

(5) Changeux, J.-P.; Edelstein, S. Conformational Selection or Induced-Fit? 50 Years of Debate Resolved. *F1000 Biol. Rep.* **2011**, *3*.

(6) Keserü, G.; Swinney, D. C. *Thermodynamics and Kinetics of Drug Binding*; John Wiley & Sons, 2015.

(7) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat Rev Drug Discov* **2010**, *9* (3), 203–214.

(8) Moreno, L.; Pearson, A. D. J. How Can Attrition Rates Be Reduced in Cancer Drug Discovery? *Expert Opin. Drug Discov.* **2013**, *8* (4), 363–368.

(9) Copeland, R. A.; Pompliano, D. L.; Meek, T. D. Drug–target Residence Time and Its Implications for Lead Optimization. *Nat. Rev. Drug Discov.* **2006**, *5* (9), 730–739.

(10) Copeland, R. A.; Harpel, M. R.; Tummino, P. J. Targeting Enzyme Inhibitors in Drug Discovery. *Expert Opin. Ther. Targets* **2007**, *11* (7), 967–978.

(11) Moulton, B. C.; Fryer, A. D. Muscarinic Receptor Antagonists, from Folklore to Pharmacology; Finding Drugs That Actually Work in Asthma and COPD. *Br. J. Pharmacol.* **2011**, *163* (1), 44–52.

(12) Gavaldà, A.; Ramos, I.; Carcasona, C.; Calama, E.; Otal, R.; Montero, J. L.; Sentellas, S.; Aparici, M.; Vilella, D.; Alberti, J.; Beleta, J.; Miralpeix, M. The in Vitro and in Vivo Profile of Aclidinium Bromide in Comparison with Glycopyrronium Bromide. *Pulm. Pharmacol. Ther.* **2014**, *28* (2), 114–121.

(13) Guo, D.; van Dorp, E. J. H.; Mulder-Krieger, T.; van Veldhoven, J. P. D.; Brussee, J.; Ijzerman, A. P.; Heitman, L. H. Dual-Point Competition Association Assay: A Fast and High-Throughput Kinetic Screening Method for Assessing Ligand-Receptor Binding Kinetics. *J. Biomol. Screen.* **2013**, *18* (3), 309–320.

(14) *Receptor Binding Techniques*; Davenport, A. P., Ed.; Methods in Molecular Biology$^{\text{TM}}$; Humana Press: Totowa, NJ, 2012; Vol. 897.

(15) Vauquelin, G. Determination of Drug–receptor Residence Times by Radioligand Binding and Functional Assays: Experimental Strategies and Physiological Relevance. *MedChemComm* **2012**, *3* (6), 645.

(16) Motulsky, H. J.; Mahan, L. C. The Kinetics of Competitive Radioligand Binding Predicted by the Law of Mass Action. *Mol. Pharmacol.* **1984**, *25* (1), 1–9.

(17) Fang, Y. Ligand-Receptor Interaction Platforms and Their Applications for Drug Discovery. *Expert Opin. Drug Discov.* **2012**, *7* (10), 969–988.

(18) Sridharan, R.; Zuber, J.; Connelly, S. M.; Mathew, E.; Dumont, M. E. Fluorescent Approaches for Understanding Interactions of Ligands with G Protein Coupled Receptors. *Biochim. Biophys. Acta* **2014**, *1838* (1 Pt A), 15–33.

(19) Giannetti, A. M. From Experimental Design to Validated Hits a Comprehensive Walk-through of Fragment Lead Identification Using Surface Plasmon Resonance. *Methods Enzymol.* **2011**, *493*, 169–218.

(20) Markgren, P.-O.; Schaal, W.; Hämäläinen, M.; Karlén, A.; Hallberg, A.; Samuelsson, B.; Danielson, U. H. Relationships between Structure and Interaction Kinetics for HIV-1 Protease Inhibitors. *J. Med. Chem.* **2002**, *45* (25), 5430–5439.

(21) Huber, W. A New Strategy for Improved Secondary Screening and Lead Optimization Using High-Resolution SPR Characterization of Compound–target Interactions. *J. Mol. Recognit.* **2005**, *18* (4), 273–281.

(22) Rich, R. L.; Myszka, D. G. Survey of the Year 2007 Commercial Optical Biosensor Literature. *J. Mol. Recognit. JMR* **2008**, *21* (6), 355–400.

(23) Anderson, K.; Lai, Z.; McDonald, O. B.; Stuart, J. D.; Nartey, E. N.; Hardwicke, M. A.; Newlander, K.; Dhanak, D.; Adams, J.; Patrick, D.; Copeland, R. A.; Tummino, P. J.; Yang, J. Biochemical Characterization of GSK1070916, a Potent and Selective Inhibitor of Aurora B and Aurora C Kinases with an Extremely Long Residence time1. *Biochem. J.* **2009**, *420* (2), 259–265.

(24) Baron, R. A.; Peterson, Y. K.; Otto, J. C.; Rudolph, J.; Casey, P. J. Time-Dependent Inhibition of Isoprenylcysteine

Carboxyl Methyltransferase by Indole-Based Small Molecules. *Biochemistry (Mosc.)* **2007**, *46* (2), 554–560.

(25) Case, A.; Stein, R. L. Kinetic Analysis of the Interaction of Tissue Transglutaminase with a Nonpeptidic Slow-Binding Inhibitor. *Biochemistry (Mosc.)* **2007**, *46* (4), 1106–1115.

(26) Frantom, P. A.; Coward, J. K.; Blanchard, J. S. UDP-(5F)-GlcNAc Acts as a Slow-Binding Inhibitor of MshA, a Retaining Glycosyltransferase. *J. Am. Chem. Soc.* **2010**, *132* (19), 6626–6627.

(27) Luckner, S. R.; Liu, N.; am Ende, C. W.; Tonge, P. J.; Kisker, C. A Slow, Tight Binding Inhibitor of InhA, the Enoyl-Acyl Carrier Protein Reductase from Mycobacterium Tuberculosis. *J. Biol. Chem.* **2010**, *285* (19), 14330–14337.

(28) Copeland, R. A.; Basavapathruni, A.; Moyer, M.; Scott, M. P. Impact of Enzyme Concentration and Residence Time on Apparent Activity Recovery in Jump Dilution Analysis. *Anal. Biochem.* **2011**, *416* (2), 206–210.

(29) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450* (7172), 964–972.

(30) Yorke, B. A.; Beddard, G. S.; Owen, R. L.; Pearson, A. R. Time-Resolved Crystallography Using the Hadamard Transform. *Nat. Methods* **2014**, *11* (11), 1131–1134.

(31) Fielding, L.; Fletcher, D.; Rutherford, S.; Kaur, J.; Mestres, J. Exploring the Active Site of Human Factor Xa Protein by NMR Screening of Small Molecule Probes. *Org. Biomol. Chem.* **2003**, *1* (23), 4235–4241.

(32) Karplus, M.; Petsko, G. A. Molecular Dynamics Simulations in Biology. *Nature* **1990**, *347* (6294), 631–639.

(33) Fiser, A.; Do, R. K.; Sali, A. Modeling of Loops in Protein Structures. *Protein Sci. Publ. Protein Soc.* **2000**, *9* (9), 1753–1773.

(34) Chemical Computing Group Inc. *Molecular Operating Environment (MOE)*; Chemical Computing Group Inc.: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2016.

(35) ChemAxon – cheminformatics platforms and desktop applications https://www.chemaxon.com/ (accessed Jan 10, 2016).

(36) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; MacKerell, A. D. 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**, *31* (4), 671–690.

(37) Harvey, M. J.; De Fabritiis, G. High-Throughput Molecular Dynamics: The Powerful New Tool for Drug Discovery. *Drug Discov. Today* **2012**, *17* (19–20), 1059–1062.

(38) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585–590.

(39) Moore, G. E. Cramming More Components onto Integrated Circuits, Reprinted from Electronics, Volume 38, Number 8, April 19, 1965, pp.114 Ff. *IEEE Solid-State Circuits Soc. Newsl.* **2006**, *11* (5), 33–35.

(40) Harvey, M. J.; Giupponi, G.; De Fabritiis, G. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, *5* (6), 1632–1639.

(41) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Ben Schafer, U.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*; SC ’14; IEEE Press: Piscataway, NJ, USA, 2014; pp 41–53.

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

(43) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(44) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102* (18), 3586–3616.

(45) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110* (6), 1657–1666.

(46) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS ONE* **2012**, *7* (2), e32131.

(47) Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations. *J. Comput. Chem.* **2004**, *25* (11), 1400–1415.

(48) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.

(49) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52* (12), 3144–3154.

(50) Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**.

(51) Jain, A. N. Scoring Functions for Protein-Ligand Docking. *Curr. Protein Pept. Sci.* **2006**, *7* (5), 407–420.

(52) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat Rev Drug Discov* **2004**, *3* (11), 935–949.

(53) Gutiérrez-de-Terán, H.; Åqvist, J. Linear Interaction Energy: Method and Applications in Drug Design. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Methods in Molecular Biology; Springer New York, 2012; pp 305–323.

(54) Lee, F. S.; Chu, Z. T.; Bolger, M. B.; Warshel, A. Calculations of Antibody-Antigen Interactions: Microscopic and Semi-Microscopic Evaluation of the Free Energies of Binding of Phosphorylcholine Analogs to McPC603. *Protein Eng.* **1992**, *5* (3), 215–228.

(55) Wang, J.; Dixon, R.; Kollman, P. A. Ranking Ligand Binding Affinities with Avidin: A Molecular Dynamics-Based Interaction Energy Study. *Proteins* **1999**, *34* (1), 69–81.

(56) Aqvist, J.; Marelius, J. The Linear Interaction Energy Method for Predicting Ligand Binding Free Energies. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 613–626.

(57) Almlöf, M.; Brandsdal, B. O.; Åqvist, J. Binding Affinity Prediction with Different Force Fields: Examination of the Linear Interaction Energy Method. *J. Comput. Chem.* **2004**, *25* (10), 1242–1254.

(58) Åqvist, J.; Hansson, T. On the Validity of Electrostatic Linear Response in Polar Solvents. *J. Phys. Chem.* **1996**, *100* (22), 9512–9521.

(59) Almlöf, M.; Carlsson, J.; Åqvist, J. Improving the Accuracy of the Linear Interaction Energy Method for Solvation Free Energies. *J. Chem. Theory Comput.* **2007**, *3* (6), 2162–2175.

(60) Aqvist J Samuelsson JE; Medina C. A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Eng* **1994**, *7* (3), 385–391.

(61) Aqvist, J.; Luzhkov, V. Ion Permeation Mechanism of the Potassium Channel. *Nature* **2000**, *404* (6780), 881–884.

(62) Treptow, W.; Tarek, M. Environment of the Gating Charges in the Kv1.2 Shaker Potassium Channel. *Biophys. J.* **2006**, *90* (9), L64–L66.

(63) Gervasio, F. L.; Laio, A.; Parrinello, M. Flexible Docking in Solution Using Metadynamics. *J. Am. Chem. Soc.* **2005**, *127* (8), 2600–2607.

(64) Zwier, M. C.; Chong, L. T. Reaching Biological Timescales with All-Atom Molecular Dynamics Simulations. *Curr. Opin. Pharmacol.* **2010**, *10* (6), 745–752.

(65) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78* (14), 2690–2693.

(66) Crooks, G. E. Path-Ensemble Averages in Systems Driven far from Equilibrium. *Phys. Rev. E* **2000**, *61* (3), 2361–2366.

(67) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011–1021.

(68) Doudou, S.; Burton, N. A.; Henchman, R. H. Standard Free Energy of Binding from a One-Dimensional Potential of Mean Force. *J. Chem. Theory Comput.* **2009**, *5* (4), 909–918.

(69) Buch, I.; Sadiq, S. K.; De Fabritiis, G. Optimized Potential of Mean Force Calculations for Standard Binding Free Energies. *J. Chem. Theory Comput.* **2011**, *7* (6), 1765–1772.

(70) Juraszek, J.; Bolhuis, P. G. Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water. *Biophys. J.* **2008**, *95* (9), 4246–4257.

(71) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.;

Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun ACM* **2008**, *51* (7), 91–97.

(72) Ohmura, I.; Morimoto, G.; Ohno, Y.; Hasegawa, A.; Taiji, M. MDGRAPE-4: A Special-Purpose Computer System for Molecular Dynamics Simulations. *Phil Trans R Soc A* **2014**, *372* (2021), 20130387.

(73) Buch, I.; Harvey, M. J.; Giorgino, T.; Anderson, D. P.; De Fabritiis, G. High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing. *J. Chem. Inf. Model.* **2010**, *50* (3), 397–403.

(74) Bisignano, P.; Doerr, S.; Harvey, M. J.; Favia, A. D.; Cavalli, A.; De Fabritiis, G. Kinetic Characterization of Fragment Binding in AmpC Β-Lactamase by High-Throughput Molecular Simulations. *J. Chem. Inf. Model.* **2014**, *54* (2), 362–366.

(75) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131* (12).

(76) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction. *J. Chem. Phys.* **2013**, *139* (1), 015102.

(77) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods San Diego Calif* **2010**.

(78) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*; Bowman, G. R., Pande, V. S., Noé, F., Eds.; Advances in

Experimental Medicine and Biology; Springer Netherlands: Dordrecht, 2014; Vol. 797.

(79) Chodera, J. D.; Noé, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135–144.

(80) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6* (3), 787–794.

(81) Lawrenz, M.; Shukla, D.; Pande, V. S. Cloud Computing Approaches for Prediction of Ligand Binding Poses and Pathways. *Sci. Rep.* **2015**, *5*, 7918.

(82) Hinrichs, N. S.; Pande, V. S. Calculation of the Distribution of Eigenvalues and Eigenvectors in Markovian State Models for Molecular Dynamics. *J. Chem. Phys.* **2007**, *126* (24), 244101.

(83) Pronk, S.; Larsson, P.; Pouya, I.; Bowman, G. R.; Haque, I. S.; Beauchamp, K.; Hess, B.; Pande, V. S.; Kasson, P. M.; Lindahl, E. Copernicus: A New Paradigm for Parallel Adaptive Molecular Dynamics. In *High Performance Computing, Networking, Storage and Analysis (SC), 2011 International Conference for*; 2011; pp 1–10.

(84) Weber, J. K.; Pande, V. S. Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies. *J. Chem. Theory Comput.* **2011**, *7* (10), 3405–3411.

(85) Bowman, G. R.; Huang, X.; Pande, V. S. Adaptive Seeding: A New Method for Simulating Biologically Relevant Timescales. *Biophys. J.* **2009**, *96* (3, Supplement 1), 575a.

(86) Lopez-Otin, C.; Overall, C. M. Protease Degradomics: A New Challenge for Proteomics. *Nat Rev Mol Cell Biol* **2002**, *3* (7), 509–519.

(87) Hedstrom, L. Serine Protease Mechanism and Specificity. *Chem. Rev.* **2002**, *102* (12), 4501–4524.

(88) Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J.; Barrett, A. J. MEROPS: The Peptidase Database. *Nucleic Acids Res.* **2008**, *36* (Database issue), D320–D325.

(89) Di Cera, E. Serine Proteases. *IUBMB Life* **2009**, *61* (5), 510–515.

(90) Page, M. J.; Di Cera, E. Serine Peptidases: Classification, Structure and Function. *Cell. Mol. Life Sci. CMLS* **2008**, *65* (7-8), 1220–1236.

(91) Neurath, H.; Dixon, G. H. Structure and Activation of Trypsinogen and Chymotrypsinogen. *Fed. Proc.* **1957**, *16* (3), 791–801.

(92) Drag, M.; Salvesen, G. S. Emerging Principles in Protease-Based Drug Discovery. *Nat. Rev. Drug Discov.* **2010**, *9* (9), 690–701.

(93) Lauro, G.; Ferruz, N.; Fulle, S.; Harvey, M. J.; Finn, P. W.; De Fabritiis, G. Reranking Docking Poses Using Molecular Simulations and Approximate Free Energy Methods. *J. Chem. Inf. Model.* **2014**, *54* (8), 2185–2189.

(94) Rothman, S. S. The Digestive Enzymes of the Pancreas: A Mixture of Inconstant Proportions. *Annu. Rev. Physiol.* **1977**, *39*, 373–389.

(95) Huber, R.; Kukla, D.; Bode, W.; Schwager, P.; Bartels, K.; Deisenhofer, J.; Steigemann, W. Structure of the Complex Formed by Bovine Trypsin and Bovine Pancreatic Trypsin

Inhibitor. II. Crystallographic Refinement at 1.9 A Resolution. *J. Mol. Biol.* **1974**, *89* (1), 73–101.

(96) Newman, J.; Dolezal, O.; Fazio, V.; Caradoc-Davies, T.; Peat, T. S. The DINGO Dataset: A Comprehensive Set of Data for the SAMPL Challenge. *J. Comput. Aided Mol. Des.* **2012**, *26* (5), 497–503.

(97) Bisignano, P.; Lambruschini, C.; Bicego, M.; Murino, V.; Favia, A. D.; Cavalli, A. In Silico Deconstruction of ATP-Competitive Inhibitors of Glycogen Synthase Kinase-3β. *J. Chem. Inf. Model.* **2012**, *52* (12), 3233–3244.

(98) Ferruz, N.; Harvey, M. J.; Mestres, J.; De Fabritiis, G. Insights from Fragment Hit Binding Assays by Molecular Simulations. *J. Chem. Inf. Model.* **2015**.

(99) Schechter, I.; Berger, A. On the Size of the Active Site in Proteases. I. Papain. *Biochem. Biophys. Res. Commun.* **1967**, *27* (2), 157–162.

(100) Pinto, D. J. P.; Smallheer, J. M.; Cheney, D. L.; Knabb, R. M.; Wexler, R. R. Factor Xa Inhibitors: Next-Generation Antithrombotic Agents. *J. Med. Chem.* **2010**, *53* (17), 6243–6274.

(101) Böhm, M.; St rzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure-Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis to Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J. Med. Chem.* **1999**, *42* (3), 458–477.

(102) Lin, Z.; Johnson, M. E. Proposed Cation-Pi Mediated Binding by Factor Xa: A Novel Enzymatic Mechanism for Molecular Recognition. *FEBS Lett.* **1995**, *370* (1-2), 1–5.

(103) Lambeir, A. M.; Proost, P.; Durinx, C.; Bal, G.; Senten, K.; Augustyns, K.; Scharpé, S.; Van Damme, J.; De Meester, I. Kinetic Investigation of Chemokine Truncation by CD26/dipeptidyl Peptidase IV Reveals a Striking Selectivity within the Chemokine Family. *J. Biol. Chem.* **2001**, *276* (32), 29839–29845.

(104) Sortino, M. A.; Sinagra, T.; Canonico, P. L. Linagliptin: A Thorough Characterization beyond Its Clinical Efficacy. *Front. Endocrinol.* **2013**, *4*.

(105) Weihofen, W. A.; Liu, J.; Reutter, W.; Saenger, W.; Fan, H. Crystal Structure of CD26/dipeptidyl-Peptidase IV in Complex with Adenosine Deaminase Reveals a Highly Amphiphilic Interface. *J. Biol. Chem.* **2004**, *279* (41), 43330–43335.

(106) Eckhardt, M.; Langkopf, E.; Mark, M.; Tadayyon, M.; Thomas, L.; Nar, H.; Pfrengle, W.; Guth, B.; Lotz, R.; Sieger, P.; Fuchs, H.; Himmelsbach, F. 8-(3-(R)-Aminopiperidin-1-Yl)-7-but-2-Ynyl-3-Methyl-1-(4-Methyl-Quinazolin-2-Ylmethyl)-3,7-Dihydropurine-2,6-Dione (BI 1356), a Highly Potent, Selective, Long-Acting, and Orally Bioavailable DPP-4 Inhibitor for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2007**, *50* (26), 6450–6453.

(107) Engel, M.; Hoffmann, T.; Manhart, S.; Heiser, U.; Chambre, S.; Huber, R.; Demuth, H.-U.; Bode, W. Rigidity and Flexibility of Dipeptidyl Peptidase IV: Crystal Structures of and Docking Experiments with DPIV. *J. Mol. Biol.* **2006**, *355* (4), 768–783.

(108) 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**, *53* (12), 3123–3126.

(109) McKay, J. A.; Murray, L. J.; Curran, S.; Ross, V. G.; Clark, C.; Murray, G. I.; Cassidy, J.; McLeod, H. L. Evaluation of

the Epidermal Growth Factor Receptor (EGFR) in Colorectal Tumours and Lymph Node Metastases. *Eur. J. Cancer Oxf. Engl. 1990* **2002**, *38* (17), 2258–2264.

(110) Porebska, I.; Harlozińska, A.; Bojarowski, T. Expression of the Tyrosine Kinase Activity Growth Factor Receptors (EGFR, ERB B2, ERB B3) in Colorectal Adenocarcinomas and Adenomas. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **2000**, *21* (2), 105–115.

(111) Hoy, S. M.; Wagstaff, A. J. Panitumumab: In the Treatment of Metastatic Colorectal Cancer. *Drugs* **2006**, *66* (15), 2005–2014; discussion 2015–2016.

(112) Li, S.; Schmitz, K. R.; Jeffrey, P. D.; Wiltzius, J. J. W.; Kussie, P.; Ferguson, K. M. Structural Basis for Inhibition of the Epidermal Growth Factor Receptor by Cetuximab. *Cancer Cell* **2005**, *7* (4), 301–311.

(113) Arena, S.; Bellosillo, B.; Siravegna, G.; Martínez, A.; Cañadas, I.; Lazzari, L.; Ferruz, N.; Russo, M.; Misale, S.; González, I.; Iglesias, M.; Gavilan, E.; Corti, G.; Hobor, S.; Crisafulli, G.; Salido, M.; Sánchez, J.; Dalmases, A.; Bellmunt, J.; De Fabritiis, G.; Rovira, A.; Di Nicolantonio, F.; Albanell, J.; Bardelli, A.; Montagut, C. Emergence of Multiple EGFR Extracellular Mutations during Cetuximab Treatment in Colorectal Cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **2015**, *21* (9), 2157–2166.

(114) Montagut, C.; Dalmases, A.; Bellosillo, B.; Crespo, M.; Pairet, S.; Iglesias, M.; Salido, M.; Gallen, M.; Marsters, S.; Tsai, S. P.; Minoche, A.; Seshagiri, S.; Somasekar, S.; Serrano, S.; Himmelbauer, H.; Bellmunt, J.; Rovira, A.; Settleman, J.; Bosch, F.; Albanell, J. Identification of a Mutation in the Extracellular Domain of the Epidermal Growth Factor Receptor Conferring Cetuximab Resistance in Colorectal Cancer. *Nat. Med.* **2012**, *18* (2), 221–223.

(115) Atack, J. R.; Broughton, H. B.; Pollack, S. J. Structure and Mechanism of Inositol Monophosphatase. *FEBS Lett.* **1995**, *361* (1), 1–7.

(116) Harwood, A. J. Lithium and Bipolar Mood Disorder: The Inositol-Depletion Hypothesis Revisited. *Mol. Psychiatry* **2004**, *10* (1), 117–126.

(117) Gill, R.; Mohammed, F.; Badyal, R.; Coates, L.; Erskine, P.; Thompson, D.; Cooper, J.; Gore, M.; Wood, S. High-Resolution Structure of Myo-Inositol Monophosphatase, the Putative Target of Lithium Therapy. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61* (Pt 5), 545–555.

(118) Rees-Milton, K.; Thorne, M.; Greasley, P.; Churchich, J.; Gore, M. G. Detection of Metal Binding to Bovine Inositol Monophosphatase by Changes in the near and Far Ultraviolet Regions of the CD Spectrum. *Eur. J. Biochem. FEBS* **1997**, *246* (1), 211–217.

(119) Ganzhorn, A. J.; Lepage, P.; Pelton, P. D.; Strasser, F.; Vincendon, P.; Rondeau, J. M. The Contribution of Lysine-36 to Catalysis by Human Myo-Inositol Monophosphatase. *Biochemistry (Mosc.)* **1996**, *35* (33), 10957–10966.

(120) Greasley, P. J.; Hunt, L. G.; Gore, M. G. Bovine Inositol Monophosphatase. Ligand Binding to Pyrene-Maleimide-Labelled Enzyme. *Eur. J. Biochem. FEBS* **1994**, *222* (2), 453–460.

(121) Awad, A. G. Psychopharmacology. *J. Psychiatry Neurosci.* **1996**, *21* (5), 350–351.

(122) Levant, B. The D3 Dopamine Receptor: Neurobiology and Potential Clinical Relevance. *Pharmacol. Rev.* **1997**, *49* (3), 231–252.

(123) Sokoloff, P.; Giros, B.; Martres, M. P.; Bouthenet, M. L.; Schwartz, J. C. Molecular Cloning and Characterization of a

Novel Dopamine Receptor (D3) as a Target for Neuroleptics. *Nature* **1990**, *347* (6289), 146–151.

(124) Shi, L.; Javitch, J. A. The Binding Site of Aminergic G Protein-Coupled Receptors: The Transmembrane Segments and Second Extracellular Loop. *Annu. Rev. Pharmacol. Toxicol.* **2002**, *42*, 437–467.

(125) Sokoloff, P.; Andrieux, M.; Besançon, R.; Pilon, C.; Martres, M. P.; Giros, B.; Schwartz, J. C. Pharmacology of Human Dopamine D3 Receptor Expressed in a Mammalian Cell Line: Comparison with D2 Receptor. *Eur. J. Pharmacol.* **1992**, *225* (4), 331–337.

(126) Breuer, M. E.; Groenink, L.; Oosting, R. S.; Buerger, E.; Korte, M.; Ferger, B.; Olivier, B. Antidepressant Effects of Pramipexole, a Dopamine D3/D2 Receptor Agonist, and 7-OH-DPAT, a Dopamine D3 Receptor Agonist, in Olfactory Bulbectomized Rats. *Eur. J. Pharmacol.* **2009**, *616* (1–3), 134–140.

(127) Gilbert, J. G.; Newman, A. H.; Gardner, E. L.; Ashby, C. R.; Heidbreder, C. A.; Pak, A. C.; Peng, X.-Q.; Xi, Z.-X. Acute Administration of SB-277011A, NGB 2904, or BP 897 Inhibits Cocaine Cue-Induced Reinstatement of Drug-Seeking Behavior in Rats: Role of Dopamine D3 Receptors. *Synap. N. Y. N* **2005**, *57* (1), 17–28.

(128) Pilla, M.; Perachon, S.; Sautel, F.; Garrido, F.; Mann, A.; Wermuth, C. G.; Schwartz, J. C.; Everitt, B. J.; Sokoloff, P. Selective Inhibition of Cocaine-Seeking Behaviour by a Partial Dopamine D3 Receptor Agonist. *Nature* **1999**, *400* (6742), 371–375.

(129) Spiller, K.; Xi, Z.-X.; Peng, X.-Q.; Newman, A. H.; Ashby, C. R.; Heidbreder, C.; Gaál, J.; Gardner, E. L. The Selective Dopamine D3 Receptor Antagonists SB-277011A and NGB 2904 and the Putative Partial D3 Receptor Agonist BP-897 Attenuate Methamphetamine-Enhanced Brain Stimulation

Reward in Rats. *Psychopharmacology (Berl.)* **2008**, *196* (4), 533–542.

(130) Vorel, S. R.; Ashby, C. R.; Paul, M.; Liu, X.; Hayes, R.; Hagan, J. J.; Middlemiss, D. N.; Stemp, G.; Gardner, E. L. Dopamine D3 Receptor Antagonism Inhibits Cocaine-Seeking and Cocaine-Enhanced Brain Reward in Rats. *J. Neurosci. Off. J. Soc. Neurosci.* **2002**, *22* (21), 9595–9603.

(131) Xi, Z.-X.; Newman, A. H.; Gilbert, J. G.; Pak, A. C.; Peng, X.-Q.; Ashby, C. R.; Gitajn, L.; Gardner, E. L. The Novel Dopamine D3 Receptor Antagonist NGB 2904 Inhibits Cocaine's Rewarding Effects and Cocaine-Induced Reinstatement of Drug-Seeking Behavior in Rats. *Neuropsychopharmacol. Off. Publ. Am. Coll. Neuropsychopharmacol.* **2006**, *31* (7), 1393–1405.

(132) Heidbreder, C. A.; Newman, A. H. Current Perspectives on Selective Dopamine D(3) Receptor Antagonists as Pharmacotherapeutics for Addictions and Related Disorders. *Ann. N. Y. Acad. Sci.* **2010**, *1187*, 4–34.

(133) Chien, E. Y. T.; Liu, W.; Zhao, Q.; Katritch, V.; Won Han, G.; Hanson, M. A.; Shi, L.; Newman, A. H.; Javitch, J. A.; Cherezov, V.; Stevens, R. C. Structure of the Human Dopamine D3 Receptor in Complex with a D2/D3 Selective Antagonist. *Science* **2010**, *330* (6007), 1091–1095.

(134) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531–1534.

(135) Erlanson, D. A. Introduction to Fragment-Based Drug Discovery. *Top. Curr. Chem.* **2012**, *317*, 1–32.

(136) Mashalidis, E. H.; Śledź, P.; Lang, S.; Abell, C. A Three-Stage Biophysical Screening Cascade for Fragment-Based Drug Discovery. *Nat. Protoc.* **2013**, *8* (11), 2309–2324.

(137) Dolezal, O.; Doughty, L.; Hattarki, M. K.; Fazio, V. J.; Caradoc-Davies, T. T.; Newman, J.; Peat, T. S. Fragment Screening for the Modelling Community: SPR, ITC, and Crystallography. *Aust. J. Chem.* **2013**, *66* (12), 1507–1517.

(138) Plattner, N.; Noé, F. Protein Conformational Plasticity and Complex Ligand-Binding Kinetics Explored by Atomistic Simulations and Markov Models. *Nat. Commun.* **2015**, *6*, 7653.

(139) Dror, R. O.; Pan, A. C.; Arlow, D. H.; Borhani, D. W.; Maragakis, P.; Shan, Y.; Xu, H.; Shaw, D. E. Pathway and Mechanism of Drug Binding to G-Protein-Coupled Receptors. *Proc. Natl. Acad. Sci.* **2011**, *108* (32), 13118–13123.

(140) Dror, R. O.; Green, H. F.; Valant, C.; Borhani, D. W.; Valcourt, J. R.; Pan, A. C.; Arlow, D. H.; Canals, M.; Lane, J. R.; Rahmani, R.; Baell, J. B.; Sexton, P. M.; Christopoulos, A.; Shaw, D. E. Structural Basis for Modulation of a G-Protein-Coupled Receptor by Allosteric Drugs. *Nature* **2013**, *503* (7475), 295–299.

(141) Shan, Y.; Kim, E. T.; Eastwood, M. P.; Dror, R. O.; Seeliger, M. A.; Shaw, D. E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133* (24), 9181–9183.

(142) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular Determinants of Drug–receptor Binding Kinetics. *Drug Discov. Today*.

(143) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.

(144) Vosmeer, C. R.; Pool, R.; Van Stee, M. F.; Peric-Hassler, L.; Vermeulen, N. P. E.; Geerke, D. P. Towards Automated Binding Affinity Prediction Using an Iterative Linear

Interaction Energy Approach. *Int. J. Mol. Sci.* **2014**, *15* (1), 798–816.

(145) Vosmeer, C. R.; Kooi, D. P.; Capoferri, L.; Terpstra, M. M.; Vermeulen, N. P. E.; Geerke, D. P. Improving the Iterative Linear Interaction Energy Approach Using Automated Recognition of Configurational Transitions. *J. Mol. Model.* **2016**, *22* (1), 1–8.

(146) Roehrig, S.; Straub, A.; Pohlmann, J.; Lampe, T.; Pernerstorfer, J.; Schlemmer, K.-H.; Reinemer, P.; Perzborn, E. Discovery of the Novel Antithrombotic Agent 5-Chloro-N-({(5S)-2-Oxo-3- [4-(3-Oxomorpholin-4-Yl)phenyl]-1,3-Oxazolidin-5-Yl}methyl)thiophene- 2-Carboxamide (BAY 59-7939): An Oral, Direct Factor Xa Inhibitor. *J. Med. Chem.* **2005**, *48* (19), 5900–5908.

(147) Stubbs, M. T.; Reyda, S.; Dullweber, F.; Möller, M.; Klebe, G.; Dorsch, D.; Mederski, W. W. K. R.; Wurziger, H. pH-Dependent Binding Modes Observed in Trypsin Crystals: Lessons for Structure-Based Drug Design. *Chembiochem Eur. J. Chem. Biol.* **2002**, *3* (2-3), 246–249.

(148) Tucker, T. J.; Brady, S. F.; Lumma, W. C.; Lewis, S. D.; Gardell, S. J.; Naylor-Olsen, A. M.; Yan, Y.; Sisko, J. T.; Stauffer, K. J.; Lucas, B. J.; Lynch, J. J.; Cook, J. J.; Stranieri, M. T.; Holahan, M. A.; Lyle, E. A.; Baskin, E. P.; Chen, I. W.; Dancheck, K. B.; Krueger, J. A.; Cooper, C. M.; Vacca, J. P. Design and Synthesis of a Series of Potent and Orally Bioavailable Noncovalent Thrombin Inhibitors That Utilize Nonbasic Groups in the P1 Position. *J. Med. Chem.* **1998**, *41* (17), 3210–3219.

(149) Choi-Sledeski, Y. M.; Kearney, R.; Poli, G.; Pauls, H.; Gardner, C.; Gong, Y.; Becker, M.; Davis, R.; Spada, A.; Liang, G.; Chu, V.; Brown, K.; Collussi, D.; Leadley, R.; Rebello, S.; Moxey, P.; Morgan, S.; Bentley, R.; Kasiewski, C.; Maignan, S.; Guilloteau, J.-P.; Mikol, V. Discovery of an Orally Efficacious Inhibitor of Coagulation Factor Xa Which

Incorporates a Neutral P1 Ligand. *J. Med. Chem.* **2003**, *46* (5), 681–684.

(150) Adler, M.; Kochanny, M. J.; Ye, B.; Rumennik, G.; Light, D. R.; Biancalana, S.; Whitlow, M. Crystal Structures of Two Potent Nonamidine Inhibitors Bound to Factor Xa†,‡. *Biochemistry (Mosc.)* **2002**, *41* (52), 15514–15523.

(151) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen Bonding: The Sigma-Hole. Proceedings of "Modeling Interactions in Biomolecules II", Prague, September 5th-9th, 2005. *J. Mol. Model.* **2007**, *13* (2), 291–296.

(152) Ibrahim, M. A. A. Molecular Mechanical Study of Halogen Bonding in Drug Discovery. *J. Comput. Chem.* **2011**, *32* (12), 2564–2574.

(153) Tan, Y. S.; Spring, D. R.; Abell, C.; Verma, C. The Use of Chlorobenzene as a Probe Molecule in Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2014**, *54* (7), 1821–1827.

(154) Jorgensen, W. L.; Schyman, P. Treatment of Halogen Bonding in the OPLS-AA Force Field: Application to Potent Anti-HIV Agents. *J. Chem. Theory Comput.* **2012**, *8* (10), 3895–3901.

(155) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25* (2), 247–260.

(156) Li, X.; He, X.; Wang, B.; Merz, K. Conformational Variability of Benzamidinium-Based Inhibitors. *J. Am. Chem. Soc.* **2009**, *131* (22), 7742–7754.

(157) *Gaussian 03*.

(158) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. Calculation of protein–ligand binding free energy by using a polarizable potential http://www.pnas.org (accessed Jan 20, 2016).

(159) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem. Int. Ed Engl.* **2002**, *41* (15), 2644–2676.

(160) Pospisil, P.; Ballmer, P.; Scapozza, L.; Folkers, G. Tautomerism in Computer-Aided Drug Design. *J. Recept. Signal Transduct. Res.* **2003**, *23* (4), 361–371.

(161) Noe, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139* (18), 184114.

(162) Niu, W.; Chen, Z.; Gandhi, P. S.; Vogt, A. D.; Pozzi, N.; Pelc, L. A.; Zapata, F.; Di Cera, E. Crystallographic and Kinetic Evidence of Allostery in a Trypsin-like Protease. *Biochemistry (Mosc.)* **2011**, *50* (29), 6301–6307.

(163) Vogt, A. D.; Bah, A.; Di Cera, E. Evidence of the E*-E Equilibrium from Rapid Kinetics of Na+ Binding to Activated Protein C and Factor Xa*. *J. Phys. Chem. B* **2010**, *114* (49), 16125–16130.

(164) Tautermann, C. S.; Seeliger, D.; Kriegl, J. M. What Can We Learn from Molecular Dynamics Simulations for GPCR Drug Design? *Comput. Struct. Biotechnol. J.* **2015**, *13*, 111–121.

(165) Durrant, J. D.; McCammon, J. A. Molecular Dynamics Simulations and Drug Discovery. *BMC Biol.* **2011**, *9*, 71.

(166) Zhao, H.; Caflisch, A. Molecular Dynamics in Drug Design. *Mol. Dyn. New Adv. Drug Discov.* **2015**, *91*, 4–14.