ROLE OF KINETIC MECHANISMS IN DRUG DESIGN

Badaoui, Magd

Awarding institution:
King's College London

King's College London

SCHOOL OF NATURAL & MATHEMATICAL SCIENCES

Department of Chemistry

ROLE OF KINETIC MECHANISMS IN DRUG DESIGN

Magd Badaoui

A thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy

First Supervisor:
Dr. Edina Rosta

Second Supervisor:
Prof. Carla Molteni

December 2020

Abstract

In recent years, the application of Molecular Dynamic (MD) simulations has become a widespread tool in biological and medical research. This type of simulations provide atomistic information and estimate thermodynamics and kinetics event associated with physical and biological processes. Due to improvement in simulation speed, accuracy and accessibility, MD simulations have become a routine protocol applied in different subjects, such as the modelling of biomolecules or during the drug discovery process. However, it is not always possible to properly sample biological processes that happen in long-timescale through simple MD. For this reason, an important effort has been put to provide new methodologies in MD, to accelerate the timescale of simulations, and to obtain results in agreement with experimental data.

This thesis contributes to these efforts by presenting a new method to sample rare events and by understanding biological mechanisms. I start by presenting a new method to predict the free energy of protein-ligand unbinding and demonstrate the efficacy of the method by applying it to a system with experimental kinetic data. Furthermore, I describe what information MD simulations can provide by applying it to different biological systems. Specifically, I provide insights into the inhibition mechanism of integrase inhibitors used for the treatment of HIV. I then unravel the mechanism behind the formation of the D-Ala peptide through the D-Ala-D-Ala ligase, an essential enzyme for the formation of the peptidoglycan wall in the bacterial cell of Mycobacterium tuberculosis. Next, I show how, through MD simulations and homology modelling, we can

predict the holoprotein Sars-Covid19 Helicase structure. Lastly, I include work where I study the behaviour of known drugs while crossing a generic membrane layer, providing kinetic and structural information.

Overall, this thesis demonstrates the potential of applying MD simulations to provide insights into diverse biological events. However, the application of MD simulations cannot entirely replace experimental procedures but should be a complementary method, to be applied in different biological fields, such as drug discovery.

Acknowledgements

There are many people whom I would like to thank for their help and support.

First of all, I would like to thank my first supervisor Edina Rosta, and my second supervisor Carla Molteni for their unlimited support and guidance, allowing me to submit my PhD successfully.

I am also grateful to Luiz Pedro Carvalho and Peter Cherepanov and the people of their group for their help. I would also like to thank Callum Dickson, Victor Hornak, Kato Mitsunori from Novartis and Antti Poso and Thales Kronenberger for hosting me during my stay in Finland.

I would also like to thank all the former and present member of Rosta lab for their support during these years of work, including Dénes Berta, Silvia Gómez-Coca, Adam Kells, Daniel Groom, Pedro Boigues, Fahim Faizi, Tamas Foldes, Sam Martino, Teodora Mateeva, Vladimir Koskin and Francois Sicard.

I would also like to thank all the people from the Chemistry Department of King's, including Rossana Fanelli, Laura Bryant, and Ecaterina Burevschi for the laughs and support during numerous tea breaks.

Thanks to all the friends I have encountered here in London, including Riccardo Ronzoni, Anwen Brown, Emma Elliston, Alistair Jagger, James Irving, Cesira De Chiara, Roberto Bellelli, Giuseppe Nicastro and Monica Moriggi, as well as my friends from Italy, Diana Cavallina, Elisa Seregni, Stefania Mattavelli, Claudia Ziboni, Federica Casati, Diego Lorito, Andrea Grassi, Pietro Tadini and Davide Del Fiol.

Eternal gratitude goes to my family, for their unconditional support and love, particularly Mohammad Badaoui, Ghada Dahhan and Omar Badaoui.

To conclude, a special thanks goes to Marta Wojciechowska, for all her support and love, and patience during my last years of my PhD.

Table of Contents

Table of Figures

List of abbreviations

ADME	Absorption, Distribution, Metabolism and Excretion
BIC	Bictegravir
CDK2	Cycling Dependent Kinase 2
CV	Collective Variable
D-ala	D-Alanine
DCS	D-Cycloserine
Ddl	D-Ala-D-Ala Ligase
DFT	Density Functional Theory
DHAM	Dynamic Histogram Analysis Method
DTG	Dolutegravir
EM	Electron Microscopy
FEP	Free Energy Profile
FF	Force Field
FR	Feature Reduction
HIV	Human Immunodeficiency Viruses
IC	Internal Coordinates

INSTI	Integrase Strand Transfer Inhibitors
LC-MS	Liquid Chromatography-mass spectrometry
MD	Molecular Dynamics
MLP	Multi-layer Perceptron
MLTSA	Machine Learning Transition State Analysis
MM	Molecular Mechanics
MSMs	Markov State Models
Mtb	*Mycobacterium tuberculosis*
NMR	Nuclei Magnetic Resonance
NSP	Non-structural Protein
PBC	Periodic Boundary Condition
PDB	Protein Data Bank
PIMD	Path Integral Molecular Dynamics
PME	Particle Mesh Ewald
PMF	Potential of Mean Force
QM/MM	Quantum Mechanics/Molecular Mechanics
RMSD	Root Mean Square Deviation
TB	Tuberculosis
TST	Transition State Theory

US	Umbrella Sampling
VdW	Van der Waals
WHAM	Weighted Histogram Analysis Method

Chapter 1 Introduction to the Thesis

"… if we were to name the most powerful assumption of all, which lead one on and on in an attempt to understand life, is that all things are made of atoms and that everything that living things do can be understood in terms of jigglings and wigglings of atoms"

Richard P. Feynman, 1963 [1]

The quote above implies that proteins, the *living things*, are not rigid bodies but are dynamic objects. Fourteen years later, the first molecular dynamics (MD) simulations were performed. From this point forward, the number of studies using MD has increased exponentially, especially in the last two decades, thanks to the ability to access faster computers and programs to perform the simulations. Within the last 12 months alone, there have been ~6,000 publications that involve MD simulations (research conducted from Webofscience.com, using as a filter molecular dynamics and protein).

Today, MD has becomes a powerful technique that provides thermodynamic and kinetic information about biological processes. In particular, in drug discovery, MD is a fundamental tool for providing information on a biological molecule's flexibility and dynamics. By looking at the movements of atoms in proteins during a specific time, it is possible to obtain information about the structure and its function. Understanding the structure and the functional activity of proteins provides key information in uncovering the mechanism

behind diseases and provides information to address the design and optimisation of molecules for the treatment of these diseases.

Following the criteria of Karplus and McCammon [2] we can generally classify the application of MD simulations applied to macromolecules into three groups: (1) sampling of conformational space to refine structures obtained from experimental techniques (2) description of a system in equilibrium to obtain thermodynamic properties (3) use of MD to examine the dynamics of the system of interest. When examining the dynamics of a system, it is particularly important to broadly sample all states and configurational space. Unfortunately, this can be challenging as rare events not always can be sampled by running long time-consuming simulations.

This thesis provides examples of how MD simulations can be a resourceful tool in understanding the activity of a protein involved in different diseases. I focus my attention on understanding the mechanism of interactions between proteins and small molecules, particularly focusing on these events' kinetics. Furthermore, I also present new methodological tools to improve the sampling of rare events. In all the works presented in this thesis, MD play a central role.

Chapter 2 will provide a brief review of all the computational techniques that I have used in this thesis, providing a general introduction to MD, docking and QM/MM simulations.

Chapter 3 presents one of my main works, in which I present a novel method to predict kinetic information of protein-ligand unbinding; I have applied the method to Cyclin-Dependent Kinase 2 (CDK2) and compared the results obtained from this method to experimental data. This work is about to be submitted to a peer-review journal.

Chapter 4 describe the work from the collaboration with Cherepanov group at the Francis Crick Institute. The project provides a comparison of the binding affinity of two different HIV-integrase inhibitors using MD and Quantum Mechanic simulations; using as initial coordinate the recently deposited cryo-EM structures generated from Cherepanov lab. This research was published in 2020 in Science.

Chapter 5 presents another piece of work resulting from the collaboration with the Carvalho group at the Francis Crick. I analyse through experimental and computational techniques the mechanism of D-Ala-D-Ala ligase in Mycobacterium Tuberculosis. Through a combination of spectrophotometer experiments and liquid chromatography-mass spectroscopy analysis, I provide evidence on which residues are important for the enzyme activity, using the wild type protein and three single point mutations. Those results are then compared with QM/MM calculations. Additionally, we perform a set of docking calculations to provide new interesting candidates to inhibit the activity of MtDdl, and from the results obtained, we tested them experimentally.

Chapter 6 uses MD simulation and homology modelling results to provide structural insights into Sars-Cov-2 helicase; an important enzyme highly conserved in the coronavidae family. Understanding the mechanism and function of this enzyme will aid the development of potent inhibitors for this disease. The project started as a response to the Covid19 pandemic. The work has been deposited into a BioRxiv in November 2020.

Chapter 7 is dedicated to one of the first projects of my PhD, which I analysed the behaviour of several known drugs across a lipid layer lipid membrane, and I will provide kinetic information in agreement with experimental results. The work was published in 2018 in The Journal of Physical Chemistry B.

Chapter 2 Methods

2.1 Molecular Dynamics

Molecular dynamics (MD) is a method used in computational chemistry to simulate the atoms' motion by integrating Newton's law of motion. In biology, MD simulations can provide important atomistic information about the mechanism of biological processes, such as ligand-target interaction, conformational change, protein folding/unfolding. This is achieved by providing the position of all the atoms represented in the simulation at femtosecond time resolution [3].

MD was for the first time introduced in 1957; it was used to study the collision between particles in a rectangular box to calculate the system's equilibrium properties [4]. It was only in 1977 that McCammon et al. [2] applied MD for the first time to a biological system to study the dynamics of the Bovine Pancreatic Trypsin Inhibitor (BPTI) protein for a total time of 9.2 picoseconds on a system with 58 residues. Thanks to the development of MD codes, force fields, and faster computers, MD's applicability has increased drastically in the last decades, becoming a powerful tool to understand biological and chemical processes [4-5].

The method consists of generating a time-dependent set of atomic coordinates by iteratively integrating Newton's law of motion, through the following equation:

$$\frac{d^2 r_i(t)}{dt^2} = \frac{F_i(t)}{m_i} \quad (2.1)$$

Where $F_i(t)$ represents the force applied on atom *i* of mass *m* at time *t*, and $r_i(t)$ represent the vector position of the atom *i*.

Practically, to perform MD simulations, we require several steps summarized as follow:

1 – Definition of the initial coordinates of the system to simulate. This is a key factor for the success of the MD simulation, especially for biological systems, like proteins. Usually, the atoms' initial position comes from experimental results (such as X-ray crystallography, NMR, or Cryo-EM). Unfortunately not always a good initial structure is provided, however using homology modelling, docking, or quantum mechanical approaches it is possible to obtain a reliable structure.

2 – Choice of the method to calculate the potential energy. Different methods have been proposed, *ab initio* quantum chemical, Density Functional Theory (DFT), hybrid quantum mechanical methods, or quantum mechanics. The last is the most common used, but it requires the definition of a Force Field (FF), a set of terms (bond stretching, angle bending, dihedral torsion, partial charges, VdW interaction) that takes into account the distortion of the values from an ideal position (see 2.1.1 for further details).

3 – Energy minimisation and assignment of an initial velocity to all the atoms before initiating the molecular dynamics. As mentioned in the first step, the initial structures might present steric clashes and overlaps between atoms. An

initial energy minimization allows the sampling of the conformational place towards the closest local minimum. The most common method is the steepest descend gradient that uses a first-order derivative equation [7]. For the initial velocity, important is the definition of the temperature at which to perform the MD simulations. It can be calculated through the following equation of the kinetic energy:

$$\langle E_{kin} \rangle = \langle E_{pot} \rangle = \sum_{i=1}^{N} \frac{m_i |v_i|^2}{2} = \frac{3}{2} N K_b T \tag{2.2}$$

Where N represents the total number of atoms of the ensemble and k_b is the Boltzmann constant. During the simulation, the assigned velocity is assigned to each atom with a randomized factor, according to a predetermined distribution and rescaled by a constant, to ensure that the overall kinetics matches equation 2.2.

4 – Simulation of the system to a required number of time-step n_{step}. The application of Newton's law of motion is made through different integration methods; the most common of the Verlet family is the Velocity-Verlet algorithm [8]:

$$\begin{cases} R(t + \Delta t) = R(t) + \boldsymbol{v}(t)\Delta t + \frac{1}{2}\boldsymbol{a}(t)\Delta t^2 \\ v(t + \Delta t) = v(t) + \frac{1}{2}\{a(t) + a(t + \Delta t)\}\Delta t \end{cases} \tag{2.3}$$

Where **v** and ***a*** are the abbreviation for the first (velocity) and second (acceleration) time derivatives of the position vector *R*. The first term provides the position of each time *t* at the next time step, and the second term gives the velocity of each atom. The method is a variant from the original Verlet integration, with a main difference that in the Velocity-Verlet algorithm the velocity is directly calculated in each step, and not derived from a mean value approach that generates additional errors from an approximate equation.

Similarly to the Velocity-Verlet algorithm, the Leap-Frog algorithm determine the new positions of the atoms using half-integer time steps:

$$\begin{cases} R(t+\Delta t) = R(t) + \boldsymbol{v}\left(+\frac{\Delta t}{2}\right)\Delta t + 0\Delta t^3 \\ v\left(t+\frac{\Delta t}{2}\right) = v\left(t - \frac{\Delta t}{2}\right) + a(t)\Delta t + 0\Delta t^3 \end{cases} \quad (2.4)$$

The kinetic and the potential energy are not defined at the same time, but the position and the forces are calculated at interleaved time points [9]. The choice of the integration results to be important while considering: the long-term energy conservation, the conservation of the phase-space volume and time reversibility.

Important is the choice of the time-step since the stability of the simulations is dependent on this step. Ideally, we need to choose a long time step to reduce the time of the calculations; however, the side effect is that motions occurring faster than the selected time-step are not sampled, resulting in a loss of information [7], [10], [11]. Typically for biological processes, the time-step is set at two femtoseconds per step, thanks to the introduction of different techniques (SHAKE [12] or RATTLE [9]), where the fastest motions (usually all X-H bonding) are kept frozen, and we can obtain a good compromise of sampling over the time step.

5 – Analysis of the simulations. From the simulation, it is possible to extract each atom's position and the relative energy of the system at specific time steps. For example, to check if an important interaction is kept during the MD simulations, we extract the interatomic internal distances between the atoms involved in the interaction and seeing if the distance increases, reduces, or stays constant during the trajectory. Dihedral and angle values of the intramolecular molecules might give information about the flexibility of the last. The Root Mean Square

Deviation (RMSD), is a simple but straightforward way to overall analyse the fluctuation of a selected set of atoms from a reference position.

2.1.1 Force Field

As mentioned in the previous section, an important set of parameters that needs to be defined is the force field (FF). FF is a parametric function of atomic coordinates, that provides the potential energy of the system. The potential energy is represented by the sum of the different terms used to model the bonded and non-bonded interactions. The bonded interactions contain the contribution from bond stretching, torsional angle, rotation around a dihedral, while non-bonded interactions come from Van der Waals and electrostatic interaction. The value for each of these parameters can be derived from experimental analysis or *ab initio* calculations. Atoms are grouped by similarity, and for each of these contributions (bond distances, angle torsion, dihedral), the ideal value for the specific parameter is provided including the amplitude of the distortion. Popular FF algorithms include AMBER [13], GROMOS [14], OPLS [15], and CHARMM [16]. In most of the work presented in this thesis, I have employed the CHARMM force field, where the potential energy is expressed as follow:

$$U^{ff}{}_{CHARMM} = \sum_{bonds} K_b(r - r_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\chi(1 + \cos(n\chi - \delta)) + \sum_{Urey-Bradley} K_{UB}(S - S_0)^2 + \sum_{impropers} K_\phi(\phi - \phi_0)^2 \quad (2.5)$$

$$+ \sum_{cmap} u_{cmap}(\phi, \psi)$$

$$+ \sum_{Non-Bonded} \left(\varepsilon_{ij} \left[\left(\frac{R_{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)$$

$r_0, \theta_0, S_0, and\ \phi_0$ represents the equilibrium value for the distance (2-3), angle (1-2-3), distance (1-3), and dihedral of improper angle (1-2-3-4) respectively (see Figure 2.1). Urey-Bradley (UB) represents the cross-potential factor for 1-3 bending. $n, \phi, and\ \delta$ are respectively the number of barriers, angle, and phase that represent the torsional potential.

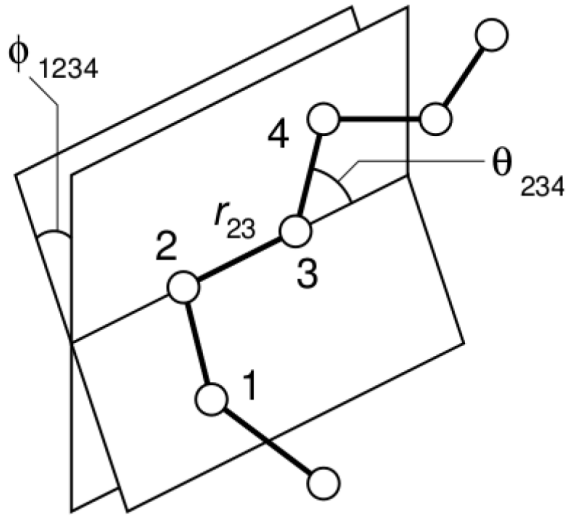


Figure 2.1. Illustration of the geometry in a simple chain molecule, with distance r_{23}, bending angle θ_{234} and torsional angle ϕ_{1234}. (From ref [11])

R_{min} is the arithmetic average of the minimum value of $r_i + r_j$. The last term corresponds to the non-bonded interaction, where the first part is the Lennard-Jones contribution, which includes both the excluded volume repulsion (r_{ij}^{-12}) and the VdW attraction component (r_{ij}^{-6}), while the second part is the Coulomb potential. The last parameter, the Coulomb potential, can be achieved by

assigning a partial charge q to each atom of the system. However, this will usually lead to an important approximation, as the partial charge of each atom depends on the adjacent atoms attached and is also affected by the surrounding environment. In general, each of these parameters can be defined specifically for each atom interactions, but this is unfeasible in practice. For this reason, atoms that share similar interactions are grouped within the same *atom types*, and the parameters applied to those atoms will be the same. For example, in proteins, the Cα carbon of the backbone is called CA, the nitrogen atom of the backbone is N et cetera. Grouping atoms by *atom type* simplify the parametrization process consistently, making it simpler. However, this type of clustering in some cases can lead to a big approximation, mining the quality of the results in molecular dynamics. For small molecules, the automatic parametrization becomes more challenging, due to the variability of the atoms interaction, and even with using popular force filed parametrization software, such as CHARMM General Force Field (CGenFF) [17] and the Generalized AMBER Force Field (GAFF) [18], not always it provides accurate results. The main problem is often related to the assignment of the partial charges for each atom of the molecule, which is usually done by using a semi-empirical method with bond charge correction (AM1-BCC) or based on atom types and connectivity. In this thesis's several works, we deploy a more exhaustive calculation to obtain the small molecules' partial charges based on quantum mechanical approaches (see chapter 2.3).

2.1.2 Periodic Boundary Condition

Additional parameters need to be considered and set before submitting the simulations. Because we sample a closed system, in MD simulations we

introduce periodic boundary condition (PBC). In a PBC system, the entire simulation volume is surrounded by replicas of itself; if an atom leaves the simulation box, the same atom's image is then created from the opposite side of the box to interact with the nearby particles or their replica image. Using PBC, you include long-range interaction between the atoms at the edge of the box in order to approximate a bulk environment. Typical algorithms used to consider the PBC is the Ewald summation [19], where the Coulomb term is divided into a short-range component, treated in the real space, a long-distance component, treated as reciprocal and a correction for when the particle is seeing its own image:

$$U_{el} = U_{real} + U_{reciprocal} + U_{correction} \quad (2.6)$$

Due to the costly computational limit while calculating U_{el} in a $O(N^2)$ system, most MD codes introduce an efficient grid-base projected O(NlogN) Ewald summation method, where $U_{reciprocal}$ is calculated using Fast Fourier Transform (FFT) called Particle Mesh Ewald (PME) [20]. Additionally to Ewald summation, in the PME the partial charges of the system are divided into a grid located on the surface of the simulation cell of the system.

2.1.3 Temperature and Pressure Control

Another aspect to consider is the type of statistical ensemble to apply; a typical ensemble used in MD simulations is the microcanonical ensemble (NVE), where the number of atoms (N), volume (V), and the total energy (E) are kept constant along the trajectory. However, NVE is a closed state, not representing real condition of experiments. For this reason, MD simulations are usually coupled with *thermostats* or *barostats*. When using a thermostat, the volume (V) and the

temperature (T) are kept constant (NVT, also referred to a canonical ensemble), while with a barostat, the pressure (P) is maintained constant and V is allowed to change (NPT). Several thermostats have been implemented in MD simulations [21]: velocity rescaling [22], Andersen [23], Nosé-Hoover [24], [25], Berendsen [26] and Langevin [27] *thermostat*. In the present thesis I applied the Langevin thermostat as implemented in NAMD software. An external bath of virtual particle is applied to the system, influencing the solute with stochastic collision. A constant friction γ_i, and an additional Gaussian random force ***R(t)*** acting on all the particles is applied following the differential equation:

$$\frac{\partial \boldsymbol{P_i}}{\partial t} = -\frac{\partial U}{\partial r_i} - \gamma P_i + \boldsymbol{R}(t) \tag{2.7}$$

The benefit of using the Langevin *thermostat* is the ability to reproduce correctly canonical ensemble while maintaining a stable dynamics. Experiments of chemical and biological reactions are normally done in an open atmosphere, for this reason NPT become a better choice. Also here, several *barostats* has been applied to MD simulations: Berendsen [28], Andersen [23], Parrinello-Rahman [29], and Nosé-Hoover [30] and Martyna-Tuckerman-Tobias-Klein [30] barostat. NAMD use Nosé-Hoover-Langevin pressure control, where the Nosé-Hoover *barostat* is coupled with Langevin dynamics. An additional pressure bath is included in the system to simulate virtually a real piston. The piston degree of freedom V is coupled to a thermal reservoir, and the dynamic equation of the velocity, pressure gradient and modified Langevin equations can be expressed as:

$$\dot{\underline{r}}_i(t) = \frac{\underline{p}_i(t)}{m_i} + \frac{1}{3}\frac{\dot{V}}{V}\underline{r}_i(t) \tag{2.8}$$

$$\ddot{V} = \frac{1}{W}[P(t) - P_{\text{ext}}] - \gamma\dot{V} + \underline{R_i}(t)$$

P(t) is the instantaneous system pressure at time *t*, and P_{ext} represent the external pressure. The mass of the piston degree of freedom is represented by W and R_i is the solvent random force.

2.1.4 Free Energy Landscapes

On of the benefit of MD, and in general of computational chemistry, is the ability to characterize chemical reactions. Theoretically, relative free energies can be calculated by counting the relative populations at different states (or configurations) in long equilibrated simulations. We can obtain thermodynamic and kinetic information approximated by summing multiple states through the application of statistical mechanics. While in QM, the relative free energy can be calculated as a sum of rotational, transitional, and vibrational energy contribution, in MD, that usually deals with a large system, this becomes impractical. In molecular mechanic approaches, we consider the possible position **q** and their momenta **r** adopted by the system's atoms. For a set of **N** atoms we can apply a Hamiltonian operator **H(p,r)** to calculate the partition function, where the atom has the form:

$$H(\boldsymbol{p}, \boldsymbol{r}) = \sum_{i=1}^{N} \frac{\boldsymbol{p_i^2}}{2m_i} + V(r_1, r_2, \ldots, r_N) \qquad (2.9)$$

Where m_i represent the mass of the atom *i* and V is the potential energy function. Because the system is usually on a NPT ensemble, the free energy F is calculated through:

$$F(N,V,T) = -K_b T \ln \left[h^{-3N} \iint \exp\left\{ -\frac{H(p,r)}{K_b T} \right\} d\boldsymbol{p} d\boldsymbol{r} \right] \quad (2.10)$$

With h as Planck's constant, this factor can often be omitted when calculating relative free energy as the number of atoms is not changing along with the simulation [31]. In a system with thousands of atoms, the exact solution of equation 2.10 becomes impractical, as there are too many sets of coordinate **r** to consider. However, MD assumes that the probability with which each configuration is visited during the trajectory is proportional to the one expressed by equation2.10. To obtain representative states for each population using MD, we either perform long-term MD simulations or introduce an enhanced sampling technique (see chapter 2.1.5).

2.1.4.1 Transition State Theory

Rate constant between two states can be calculated, using statistical thermodynamics, by applying the Eyring equation:

$$k \cong k_{TST} = \frac{k_B T}{h} \times \frac{1}{c_{std}^{\Delta n}} \times \exp\left(\frac{-\Delta G^{\ddagger}}{RT} \right) \quad (2.11)$$

c_{std} is the standard concentration assumed when calculating the translational partition function, n is equal to zero when there is no change in the number of molecules between states otherwise is one. $\Delta G^{\ddagger}$ is the Gibbs energy of activation, that is the minimum amount of energy required to achieve the transition state. Therefore, this reaction rate is called transition state theory (TST), due to the qualitative value on how chemical reactions occur. Because in classical MD, in its simplified nature, we do not have the breaking and forming

of a covalent bond, what we can see is the energy difference between configurational states. In several chapters of this thesis, we analyse the transition state between protein-ligand complexes. Such information provides key atomistic insight about the type and strength of the interaction between these two states.

2.1.5 Enhanced Sampling Techniques

Biological events, such as ligand-protein binding or unbinding, protein folding or unfolding, happen on a second to hour scale, rendering hardto sample through simple MD simulations, even using the most sophisticated hardware. To sample and observe such biological events, MD needs to be run for a long time. Only in the last 20 years, we can perform μs long time simulations in few days [32], this was possible only thanks to advances in computer technology and the possibility to run simulations in high-performance computing. In 2008, D.E. Shaw built in New York a parallel supercomputer designed specifically to run MD simulations, and in one of his firsts work he was able to perform 100 μs to characterize the folding of the FiP35 domain. [33] More affordable progress has been achieved by using a GPU processor to run MD simulations using CUDA architecture. Nowadays, is it possible to obtain hundreds of ns in a day only using a personal desktop accessorized by a good GPU.

From MD simulations, we obtain a time series set of coordinates along with a user-defined time range. If the system is trapped to a local minimum, defined as a low energy state along the free energy surface, a jump to another state requires overcoming a high energy barrier, which is unlikely but,can be achieved through long simulations. However, to sample rare events, we need to enhance the sampling simulation and ensure that the relevant regions of the

configurational space are visited well. For this purpose, a valid option is either to temper, usually done by increasing the temperature of the system, such as parallel tempering where coordinates extracted along the trajectory are exchanged between replicas run at different temperature, or to modify (bias) the potential energy of the system [34]. The latter is done by applying an external bias potential; this allows the system to sample different areas of the potential energy surface, including rare states, by elevating the energy of the reactant, though lowering the energy barriers of the potential energy surface [35]. However, because we bias with an additional potential the system, the distribution over the structures cannot allow direct calculation of the relative free energy. These sets of methods are called enhanced sampling techniques [36]. Overall, all these methods can be generally divided into different families: **CV based methods** such as Umbrella Sampling (US) [37] or metadynamics [38], **multiple replica based methods** such as Hamiltonian Replica exchange [39], **path based methods**, such as finite temperature string method [40], PathCV and **Markov State Models (MSM)** [41] **methods**, such as WHAM [42] and DHAM [43]. Among these methods, in this thesis I have used Umbrella sampling combined with DHAM to analyse the passive permeation of 7 drugs crossing a membrane layer and finite temperature string method associated with a binless version of WHAM to predict the kinetics of protein ligand unbinding. Hereafter, I will briefly describe these methods.

2.1.5.1 Umbrella Sampling (US)

Suggested by Torrie and Vallau in 1977 [37], is one of the first enhanced sampling technique applied to MD. The method consists of sampling the reaction coordinate by applying a harmonic bias potential to a reaction coordinate (ξ) and a potential strength k:

$$V(\xi) = \frac{k}{2}(\xi - \xi_{ref})^2 \quad (2.12)$$

The method consists of dividing the reaction coordinate into multiple windows, where for each window a specific bias to that CV is applied. During the simulations, the trajectory is constrained along with that specific phase space. If the position of the windows are close enough, and there is overlap along each trajectory along with the windows, then it is possible to apply a post-process method to obtain the free energy profile, such as WHAM [42] and DHAM [43] (described at chapter 2.1.5.3 and 2.1.5.4 respectively).

2.1.5.2 Finite Temperature String Method

Finite Temperature String Method sample adaptively the energy landscape, and through weighting the equilibrium probability distribution, it is possible to construct the transition path. [40] After defining a set of collective variable, an N-dimensional space string is built, representing the reaction pathway. Thebeginning and the end of the strings represent respectively the reactant and the product of the reaction. Multiple constrained simulations are performed along these hyperplanes to sample the conformational space. In the end, the data coming from each string window can be processed using WHAM or DHAM.

2.1.5.3 WHAM

Several methods have been proposed for the calculation of the free energy profile through MD simulations. The Weighted Histogram Analysis Method (WHAM) [42] is one of the first methods that include all the intermediate state, along with the potential of mean force (PMF). Coming from the Multiple Histogram equation developed by Ferrenberg and Swendsen [44], the method works with the idea that all the states of your system can be discretized into a

defined number of bins, that describes the reaction coordinate. From the probability of being on each of these bins, it is then possible to calculate the free energy profile. To obtain the free energy profile (FEP) using WHAM, we need to solve iteratively two equations; one to calculate the probability distribution and then the free energy function:

$$p_j = \frac{\sum_{i=1}^{S} n_{ij}}{\sum_{i=1}^{S} N_i f_i c_{ij}}$$
$$F_i = \sum_{j=1}^{M} c_{ij}\, p_j \qquad (2.13)$$

The total number of configurations is represented by S, p_j is the probability for each j bin, for a total of M bin. $N_i f_i$ are the number of configuration and normalizing factor respectively and c_{ij} is the biasing factor.

2.1.5.4 DHAM

Unlike WHAM, the dynamic histogram analysis method (DHAM) [43] calculates the free energy profile by using a Markov model. The data points obtained, usually from the umbrella sampling simulations, are unbiased through, and the Markov matrix is constructed by calculating the transition count. The advantage of this method is that, compared to WHAM, there is no need to calculate the iterative solution applied in the first equation of WHAM. Please refer to Chapter 7.4.1 for more detail on the equation and the relative method.

2.2 Docking

Docking is a molecular modelling technique that consists of fitting one or multiple molecular structures into a target (for example, a protein) and predicts the complex's binding affinity. The method fits a ligand to a binding site, analyse each poses for steric, hydrophobic, and electrostatic interactions and rank the results according to a scoring function between different molecules and configurations of the same molecule. Docking has been widely used in the drug discovery process; the benefit of this method is that you can screen millions of compounds against a specific target within hours. The process is divided into two stages: configurational sampling and the evaluation of the scoring function. In the configurational sampling, the ligand, which can be a small molecule, a peptide, or another protein, is fitted to the target, following the force field potential. To speed up the process, the sampling can be: (1) rigid, where both the ligand and the receptor are not moving (often this method is compared as key-lock imposition), (2) semi-flexible, where the ligand has some degree of freedom, or (3) flexible, where also the protein is free to move. The rigid and semi-flexible methods assume that the atom coordinates of the receptor are the ones able to interact with the ligand. The application of constraints to the atom positions of the system is made to speed-up the process itself; however, if the structure used is not reliable, the results obtained can be misleading. The sampling of the ligand poses during the docking can be done through a systematic approach, such as exhaustive search, or through a stochastic approach, for example, using either the Monte Carlo algorithm or more recently through the application of Machine Learning techniques. According to how many and which components of the force field are considered, multiple scoring

technique can be suggested. Some scoring functions implement a knowledge-based function, using a database of structures.

2.3 Quantum Mechanics and QM/MM

While in classical mechanics, atoms are treated as spherical particles connected by springs that follows the rules of parametrized force fields, in quantum mechanics QM, the system is described as a function of the particle coordinates, called wavefunction. The potential energy of the system is a function of the atomic coordinates and the total energy can be expressed by the Hamiltonian equation:

$$\hat{H}_{\mathrm{tot}} = \hat{T}_{\mathrm{e}} + \hat{T}_{\mathrm{n}} + \hat{U}_{\mathrm{en}} + \hat{U}_{\mathrm{ee}} + \hat{U}_{\mathrm{mn}} \tag{2.14}$$

Where $\hat{T}_{\mathrm{e}}$ and $\hat{T}_{\mathrm{n}}$ are the sum of the kinetic energy operators for the electrons and nuclei respectively. $\hat{U}_{\mathrm{en}}$ is the attractive electrostatic potential between the electron and the nuclei and $\hat{U}_{\mathrm{ee}}$ and $\hat{U}_{\mathrm{mn}}$ represents the electrostatic repulsive interaction between electron-electron and nuclei-nuclei respectively. Using the time-dependent Schrodinger equation, it is possible to predict the evolution in time of a system based on its wavefunction:

$$\hat{H}|\Psi\rangle = E|\Psi\rangle \tag{2.15}$$

Here the wavefunction Ψ describe all the electron (r) and nuclear coordinates (R). Through approximation, the total wavefunction can be decoupled into an electronic ($\Phi(\mathbf{r};\mathbf{R})$) and a nuclear wavefunction ($X(\mathbf{R})$):

$$\Psi(\mathbf{r},\mathbf{R}) = \Phi(\mathbf{r};\mathbf{R})X(\mathbf{R}) \tag{2.16}$$

Since the size of the electrons are ~2000 times smaller than the size of the protons and neutrons we can implement the Born-Oppenheimer (BO) approximation. In the BO approximation the nuclei are considered as fixed and the kinetic energy of the nuclei can be excluded, simplifying the Hamiltonian as:

$$H_{ele} = -\frac{1}{2}\sum_i \nabla_i^2 + \sum_{i<j} \frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_{I,i} \frac{Z_I}{|\vec{R}_I - \vec{r}_i|} + \sum_{I<j} \frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|}. \tag{2.17}$$

The electronic Hamiltonian (H_{ele}) is given as the sum of the kinetic energy of the electrons, the repulsive potential of between electrons, the attractive potential between nuclei and electrons, and the nuclear energy due to the repulsion between nuclei.

A different computational approach is the density functional theory, which uses the electron density to describe the system instead of directly solving the electronic wavefunction to obtain the orbitals. In their seminal work [45], Hohenberg and Kohn suggested that the ground-state properties of a many-electron system are uniquely determined by its density. However, the exact expression of the density functional is not known, instead in practice, an approximation of the functional is applied:

$$E[\rho(r)] = \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}}\, dr_1 dr_2 + T_s[\rho(r)] - \sum_J \left(\int \frac{\rho(r)Z_J}{r_J} dr\right) + V_{XC}[\rho(r)] \tag{2.18}$$

The first term corresponds to the Coulombic repulsions in a non-interacting electron gas, the second term to the kinetic energy, the third term is the Coulombic energy between the orbital and the nuclei, and the fourth term accounts for the electronic exchange and correlation interaction energies. The first and third terms can be derived analytically, while the exact form of the

second term is only known for non-interacting electron gas systems [46]. The form of the remaining fourth term is unknown, and numerical approximants were developed to complete the equation. These exchange correlation functionals vary from relatively simple local density approximations to composite hybrid functionals, which can even include ab initio calculated components. For example, B3LYP is a commonly applied member of the hybrid functionals [47], parametrized by Becke, and includes a combination of Hartree-Fock exchange and DFT exchange-correlation. The first term corresponds to the Coulombic repulsions accounting for the electronic density, the second term to the orbital kinetics, the third term is the Coulombic energy between the orbital and the nuclei and the fourth term is the exchange-correlation energy functional.

In this thesis, quantum mechanics calculations have been performed to calculate the partial charges of small molecules. I refer the reader to see chapter (4.4.2 or 5.3.2) to obtain information about the method and level of theory used.

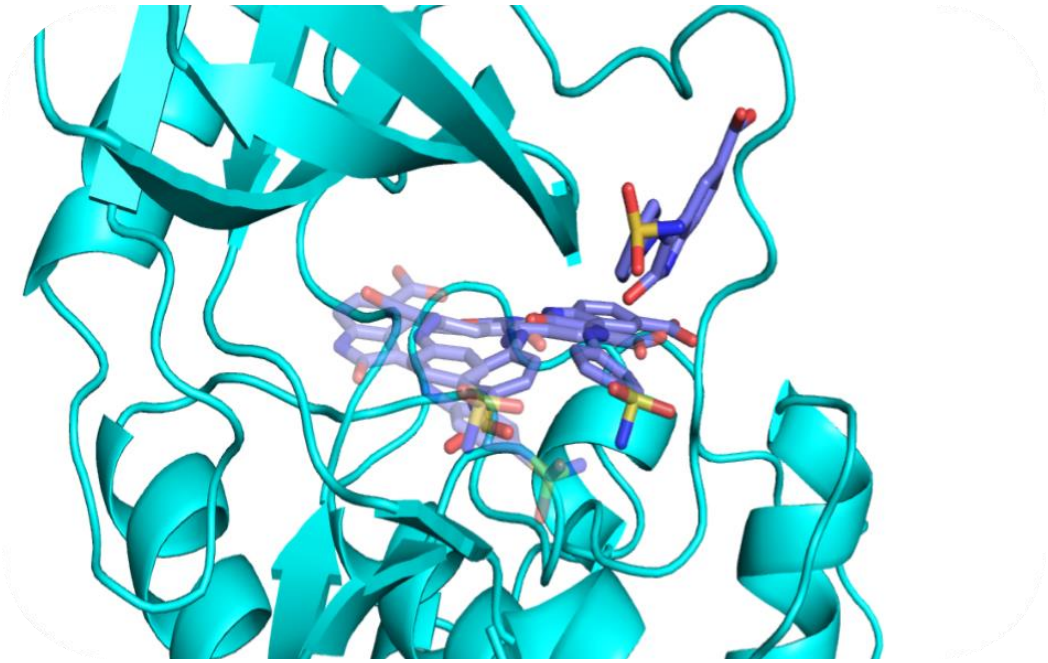
In chapter (DDL), we applied Quantum Mechanic/Molecular Mechanics (QM/MM) calculations to understand the reaction mechanism of the D-Ala-D-Ala ligases in *Mycobacterium Tuberculosis*. The principle behind QM/MM relies on an idea first presented by Warshel and Levitt in 1976 [48]. The idea behind QM/MM is to treat at a high level of theory a small number of atoms, for example, the atoms involved in chemical reactions, using quantum mechanic approaches, and treat with a lower level of theory, molecular mechanics, the surrounding environment. Using QM as a level of theory for the active site, we can compute the electronic rearrangement and sample the formation and breaks of bonds, which cannot be done by running simple MD simulations. The energy extracted from QM/MM can be of two types: additive, where $E_{tot} = E_{QM(QM)} + E_{MM(MM)} + E_{QM-MM}$ where E_{QM-MM} is the energy between the two

interfaces; or however less common, subtractive, where $E_{tot} = E_{QM(QM)} + E_{MM(MM)} - E_{QM-MM}$. Like MD, in QM/MM, it is possible to apply enhanced techniques such as umbrella sampling and the finite temperature string method. From the data obtained, we can unbias the data using for example, WHAM and determine a minimum free energy path.

QM/MM have been exhaustively used in the last decades for biological systems, especially in enzymatic reactions, where the size of the protein and molecules involved in the reaction are often extremely big, making it impossible to be simulated using only QM approaches. However, using plain MD simulations, we cannot sample the reaction path, which involves breaking and forming bonds. Successful results have been achieved using QM/MM simulations. [49]. In a work done by me and colleagues we present a review where we show the importance of QM/MM calculations to understand enzymatic catalytic reactions highlighting the importance of Mg^{2+} and H^{+} ions in particular metalloenzymes catalysis:

Berta, D.; Buigues, P.; **Badaoui, M.**; Rosta, E. (2020) 'Cations in motion: QM/MM studies of the dynamic and electrostatic roles of H+ and Mg2+ ions in enzyme reactions', *Current Opinion in Structural Biology*, 61, pp. 198–206. doi: 10.1016/j.sbi.2020.01.002.

Chapter 3 Kinetics of Protein Ligand-Unbinding: How to Find the Right Collective Variable



3.1 Preface

This chapter details the main project of my doctorate, which involved a collaboration with Novartis US. The work aimed to present an alternative approach to predicting k_{off} from a protein-ligand complex in a simple and high throughput way. This research presents a state of the art method to predict the free energy profile of ligand-protein unbinding. My attention was focused on choosing the relative collective variable that best describes the unbinding event and how to generate an initial reliable unbinding trajectory. Additionally, after obtaining reliable unbinding paths, we used the unbinding trajectory data to apply supervised machine learning calculations to understand key interactions around the transition state. From the machine learning results, we obtain additional information on what the important interactions that best describe the directionality of a reaction at the transition state point. The work presented in this chapter was completely performed by me, with discussion and feedback from people within my group and Novartis. This work will soon be submitted for peer-review.

3.2 Abstract

Determining residence times of potential drugs, which define the time the inhibitor is in complex with its target, is fundamental in the drug discovery process. While several methods, e.g., surface plasmon resonance, are available experimentally, these are expensive and take a long time to perform. In this work, we aim to computationally identify drug residence times. We designed a new enhanced sampling technique to accurately predict the free energy profiles of the ligand unbinding process, focusing on the free energy barrier for unbinding. Our method first identifies unbinding paths determining a set of internal coordinates (IC) that form contacts with the ligand during the unbinding process. We iteratively identify the interactions between the ligand and the protein during a series of biased molecular-dynamics (MD) simulations to reveal these key ICs important for the unbinding process. Subsequently, we use them for accurate free energy calculations by performing finite temperature string simulations to obtain the free energy barrier for unbinding. Finally, we apply supervised machine learning calculations designed to identify key interactions driving the system through the transition state (TS).

We tested our method on the example of Cyclin Dependent Kinase 2 (CDK2) in complex with three different ligands. We demonstrate that the free energy barriers obtained from our calculations result in comparable kinetic unbinding rates as observed in available experimental data. Additionally, we identify key ligand-protein interactions that are determining structural factors in the TS structure and therefore, in the unbinding rates for the unbinding process. Our method provides a new tool to determine unbinding rates and point to key structural features of the ligands that provide starting points for novel design strategies in drug discovery to predict and optimize ligand unbinding kinetics.

3.3 Introduction

Two essential factors affect the interaction between a drug and its target: binding affinity and residence time [50]. While the binding affinity describes the intermolecular interaction between the ligand and the protein; the residence time defines the timescale of the interaction [51], [52]. Even if a drug interacts strongly with its target (high binding affinity), a short residence time can significantly reduce the efficacy of the drug [53]. The binding affinity arises from the thermodynamic relation between the stable bound and unbound states. However, the residence time will be determined by the path connecting those states, in particular, at the transition state of the unbinding pathway. Accordingly, promising hit candidates with high affinity have been discarded for the next step of the drug discovery process due to their low residence time [54]. Traditionally, drug discovery focused on finding compounds that interact with high binding affinity to a specific target. It is recently recognized that predicting pharmacokinetics properties is also vital in the drug design process [55], [56].

A major challenge in drug discovery is to find a fast and reliable method to predict the kinetics of ligand-protein interactions [57]. Different experimental methods have been used to obtain kinetics of ligand-receptor unbinding, such as radioligand binding assays, fluorescence methods, chromatography, isothermal titration calorimetry (ITC), surface plasmon resonance (SPR) spectroscopy, and nuclear magnetic resonance (NMR) spectroscopy [55]–[58].

Radioligand binding assays and fluorescence binding essays require binding essays of labelled ligands, where they exploit the physical-chemical characteristics of the ligand between their free and complexed forms with the target. Several successful essays have been used to predict ligand-protein

unbinding, for example, fluorescence resonance energy transfer (FRET)[59] or fluorescence correlation spectroscopy (FCS)[60]. These methods can suffer from interference (especially fluorescence), lack of accuracy for short residence times, and high cost/hazard in the case of radioligands. SPR is the most widely used essay to measure k_{on} and k_{off} of ligand-receptor unbinding. The receptors are immobilized to a sensor that can distinguish the protein from its ligand-free form and bound forms. This method is label-free; however, the attachment of the protein to the probe may influence the activity of the protein, due to conformational changes. In general, experimental techniques provide a direct measurement of the kinetic rate, however they do not provide mechanistic interpretation at the atomic level of the unbinding process. To offer a screening approach that alleviates these difficulties, various complementary computational techniques have been proposed to estimate the kinetics of unbinding events [52], [61].

Molecular dynamics (MD) is a powerful computational tool to understand at an atomistic level the behaviour of biological processes such as protein-ligand interactions [62]. Unbiased MD simulations were successfully used in the early drug discovery process, using either multiple independent relatively short simulations [63] or using specialized computer architecture, such as ANTON, where microsecond long simulations are easily accessible [64]. However, due to the limited timescales typically accessible via MD simulations, it is often challenging to obtain sufficient statistical sampling required to calculate kinetic and thermodynamic properties accurately. Drug-protein unbinding processes occur on long timescales, typically ranging from millisecond to hours, depending on the nature and the strength of the interaction between the two molecules. For example, some drugs, such as Telmisartan, Carbamazepine, Diazepam, or Meloxicam, have a residence time that reaches 24 hours,[65] requiring prohibitively long time scale simulations and highly demanding

computer resources, therefore enhanced sampling methods are required [66], [67].

To accelerate the simulations and sample rare events, different enhancement techniques have been proposed to predict free energy barriers and uncover the kinetics of biological events. Below, I provided a general overview of some common techniques used to predict kinetics of protein ligand unbinding:

Milestoning is a method that combine MD, Brownian dynamics and milestoning theory for the estimation of kinetic rates [68]. The reaction path is divided into predetermined intermediates, and the transition events between these intermediate states is defined as milestones. Multiple parallel simulations are run to obtain a proper sampling of these milestone events. **Metadynamics** (metaD) is a popular enhanced sampling method [38], [69], [70], where a history dependent potential is added to the overall potential energy of the system where the bias is applied to a user defined set of CVs. Because of the history dependency, the system is forced to escape from the local minima and sample the defined reaction coordinate. MetaD can be used with **path collective variables** (PCVs) [71] to enhance the sampling of the free energy surface. The advantage of this method is that it combines path-based methods, PCVs, with methods based on CVs, metaD. An extra CV that represents the path connecting two well defined state, for example bound an unbound, is defined in the space of other CVs. PCVs were successfully employed to determine the free energy profile of several ligands in complex with CDK2 [72]. In **Steered MD** (SMD) [73] we apply a scaling factor to the entire potential energy surface, reducing the hight of the energy barriers and the transition between energy minima are facilitated. In SMD there is no need to define the reaction coordinate, because the potential energy is applied to the entire system, the protein stability is affected, and to overcome this issue, positional restrained are applied to the protein. The residence time of Sunitinib and Sorafenib in complex with the

human endothelial growth factor receptor two has been calculated using SMD[74]. SMD was also used to calculate the unbinding free energy profile for TAK-632 and PLX4720 bound to B-RAF [75]. In both these works, the ligands could be distinguished qualitatively to assess shorter, or longer residence times, however, the predicted free energy barriers for the unbinding were lower than the experimental data. This methodology, was successfully tested to predict the ligand-protein unbinding of p38 MAP kinase bound to type II inhibitors,[76] where depending on the set of CVs chosen, they obtain different values for the k_{off}, and the closest k_{off} to the experimental data is still one order of magnitude lower. Relative residence time can be calculated by using **random accelerated MD** (tau-RAMD) [77]. In this technique a random force is applied to the system to promote the unbinding of the ligand from the protein. In tau-RAMD, there is no need to *a priori* select the CVs, the direction of the force is randomly reassigned if the defined time interval falls below a specified threshold distance. In a recent work of Kokh et al. [78] 70 compounds were correctly ranked by their relative residence time. Transition state-partial path transition interface sampling (TS-PPTIS) [79] uses the binding free energy obtained from a variety of enhanced technique, for example metaD simulations, by implementing interface crossing probabilities using a semi-Markovian approximation. A main advantage of this method consists into dividing the full path into windows, each of them independently sampled, to reduce the computational time. TS-PPTIS has been proven to predict in good agreement the k_{off} of imatinib in complex with the proto-oncogene c-Src [80]. A recent combination of enhanced technique and machine learning has been proposed by Evans et al. [81] and has been used to calculate the absolute free energy profile of 18 ligands in complex with the human soluble expoxide hydrolase.

Often during a simulation, the system is trapped to a local minimum, defined as a low energy state, and to jump to another state requires overcoming a high energy barrier. To produce a free-energy profile where all the biasing points are well-defined, we need to define an ideal set of CVs that map the full path of the reaction coordinate [82], [83]. Usually, these vectors that describe this manifold are selected based on *a priori* chemical/physical intuition, usually selected based on the ligand's initial binding pose. The same set of CVs are then kept constant and used for the full simulation. Considering only CVs from an initial structure is possibly neglecting essential interactions that occur during the unbinding process significantly affecting the free energy calculation. Path CV, introduced by Branduardi at all. [71] and its recent implementation by Hovan [84] are an ideal solution to overcome the problem of choosing a correct set of CVs, however the knowledge of the end states is still a requirement.

This work introduces a novel enhanced sampling method to obtain accurate free energy barriers for ligand-protein unbinding. Unlike existing methods, we suggest an iterative way of assigning our CVs during our unbinding trajectory and using these CVs as the driving force to pull the ligand out from the pocket and to perform the sampling for accurate free energy calculations. Through this method, we are able to build a reliable path of unbinding taking into consideration the flexibility and dynamics of the system; the path is used as a starting point for free energy calculations using the finite temperature string method [85]. In addition to determining unbinding rates, we also aim to identify key molecular descriptors that provide guidance for further design of drugs based on improved residence times. We propose a systematic approach to identify key low-dimensional sets of internal coordinates using machine learning (ML) approaches. Machine learning methods have been widely successful in multidimensional data driven problems, which are also applied to biomolecular simulations to determine key CVs [86]–[88]. Here, we develop a

novel approach making use of our obtained string unbinding pathway and, within that, the knowledge of the transition state (TS) ensemble. We generate unbiased “downhill” trajectories initiated at our TS, and use these to train a ML model which predicts the fate of binding or unbinding. Our results demonstrate that our novel ML analysis can identify the key features correlated to this selected double-well potential to define the end states and thus can be used for key feature selection successfully.

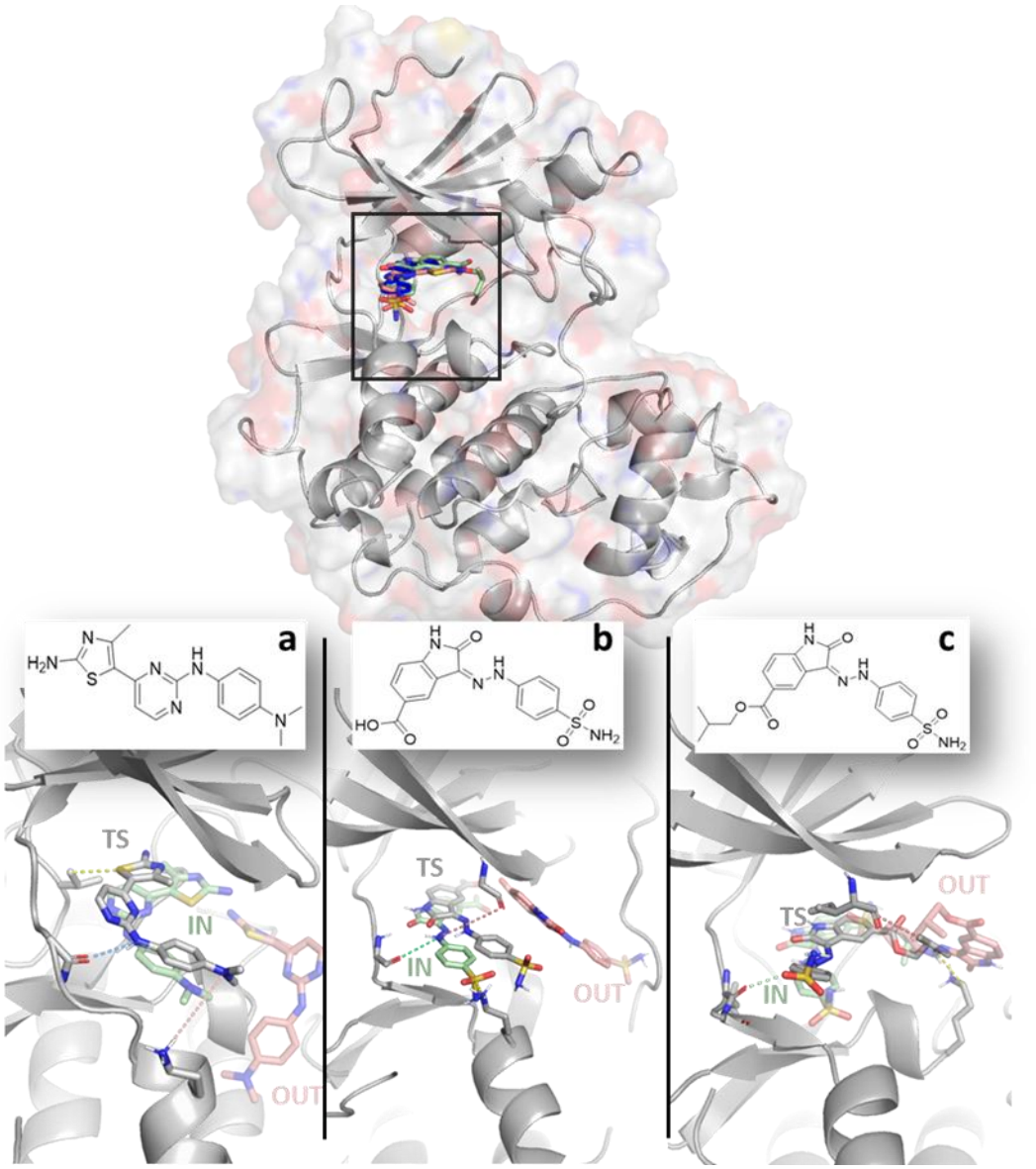


Figure 3.1. Graphical representation of CDK2 bound to three different ligands: **a** thiazolylpyrimidine derivative (18K), **b** oxindole carboxylic acid derivative (60K), and **c** carboxylate oxindole derivative (62K), originated from PDB structures 3sw4, 4fku, 4fkw, respectively. Structural details of the ATP pockets are shown for the three systems (bottom), with the ligands in the bound (green sticks), unbound (red sticks), and transition states (grey sticks). Dashed lines depict key interactions.

To demonstrate this approach's applicability and accuracy, we obtained free energy barriers for three ligands with PDB IDs of 18K, 60K, and 62K bound to CDK2 (Figure 3.1). CDK2 is an ideal benchmark system with its relatively small

size and well-documented kinetic data for the binding of a range of different molecules [89].

This method's novelty is that there is no need to *a priori* select CVs; these naturally arise from the unbinding trajectories and later get incorporated into the free energy calculations for better sampling. Additionally, we also perform a post-processing step using machine learning to identify key features that determine the outcome of trajectories near the TS and therefore are the key descriptors of the reaction coordinate describing the saddle point.

3.4 Method

3.4.1 Simulation Setup

The systems used to test our method comes from the following PDB: 3SW4, 4FKU, and 4FKW. The systems were modelled using AMBER ff14SB force field [90], and the ligands using the general Amber force field (GAFF) [18]. The ligand's atomic partial charges were obtained using density functional theory (DFT) ωB97X-D/def2TZVPP as implemented in Gaussian 09 Revision E [91]. The full system was neutralized with Na+ and Cl- and solvated with 12,000 -14,000 TIP3P water molecules. All the simulations were performed via the standard MD procedure using NAMD 2.12 [92].

The three systems were minimized using a standard protocol via the steepest descent algorithm for a total of 150,000 steps and equilibrated for ten ns with restrained heavy atoms in constant number pressure and temperature (NPT). All the production runs are then performed in constant number volume and

temperature (NVT) at 1 atm, 298 K using a time step of 2 fs, the non-bonded cut-off of 9 Å.

An initial unbiased simulation of 20 ns was performed for each ligand. This initial simulation allows the system to equilibrate and gives us an initial trajectory to calculate the first CV.

3.4.2 Unbinding Trajectory

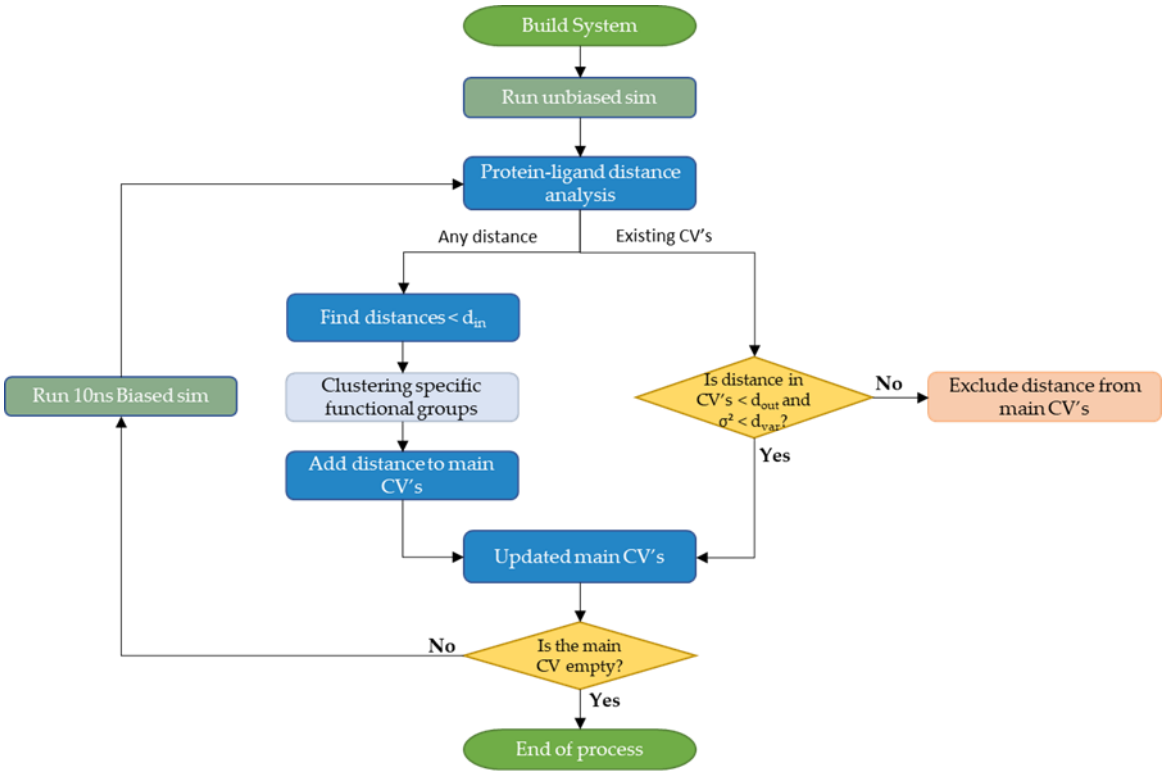


Figure 3.2. Flowchart illustrating the unbinding protocol.

Our unbinding method is summarized algorithmically in Figure 3.2. An explorational unbiased MD simulation of at least 20 ns was performed to identify the initial interactions between the protein and the ligand in the bound state. These initial simulations allow us to define the first set of CVs describing all distances between heavy atoms of the ligand and heavy atoms of the protein less than d_{in} = 3 Å, as our interaction cut-off. The identified interactions will

generate a single one-dimensional CV as the sum of M^n number of distances d_i^n and will be used for iteratively biasing the simulations to observe an unbinding trajectory (Figure 3.3).

At every iteration, we define our bias as a harmonic restraint: $V^n = \frac{1}{2}\, k \left(D^n - \sum_{i=1}^{M^n} d_i^n\right)^2$, where $D^n = D_0^n + M^n$. Here, we aim to reach the target value D^n for our 1D CV starting from the initial value at the beginning of the iteration n at D_0^n. The targeted D^n value will be reached progressively within the next 10 ns long MD simulation for every iteration. The force constant is set to 20 kcal/mol/Å^2.

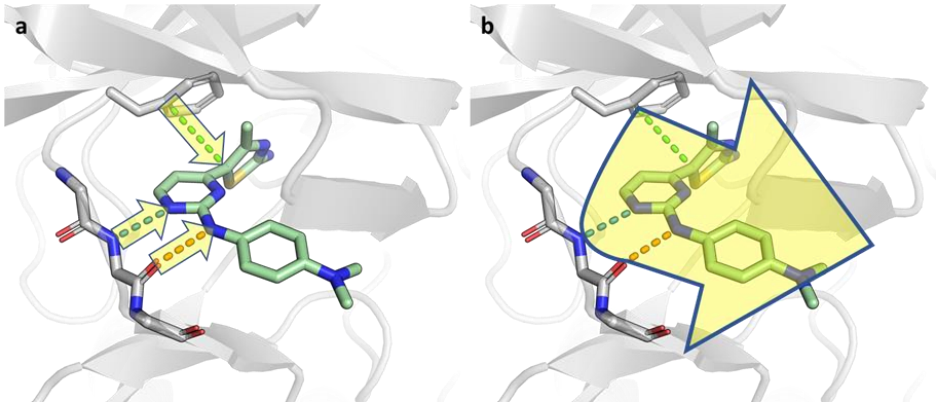


Figure 3.3. Graphical example between one strong interaction and one weak interaction. The yellow arrows represent the hypothetical spring force, in **a** the spring forces are selected for each distance of the CV, wherein **b** the spring forces is obtained as a sum of the individual interactions.

At the end of each iteration, the biased trajectory is analysed, and novel interactions are identified within d_{in} of the ligand that are present for more than half of the total simulation time (5 ns). These novel interactions are then added to the list of interactions that define the main CV for the next iteration. Additionally, we also re-evaluate existing interactions. If a distance during the last 5 ns of the trajectory exceeds d_{out} = 6 Å or its variance exceeds d_{var} = 1 Å, then the distance is removed from the main CV in the next iteration. This exclusion factor will ensure that once a protein-ligand atom pair distance has

exceeded d_{out}, and therefore there is no significant interaction between these atoms, we no longer bias this interaction. Similarly, such loosely interacting atom pairs have higher distance fluctuations, and thus this weak interaction does not need to be included in the bias. To reduce the number of interactions between the ligand and the protein and to remove redundancies, we combine atoms that are part of an equivalency group where a rotational degree of freedom can interconvert the atoms from one to the other (Figure 3.4). Here, we considered the center of the mass of that functional group and not the individual atoms. By using the center of the masses, we reduced the fluctuations caused by such bonds' rotations.

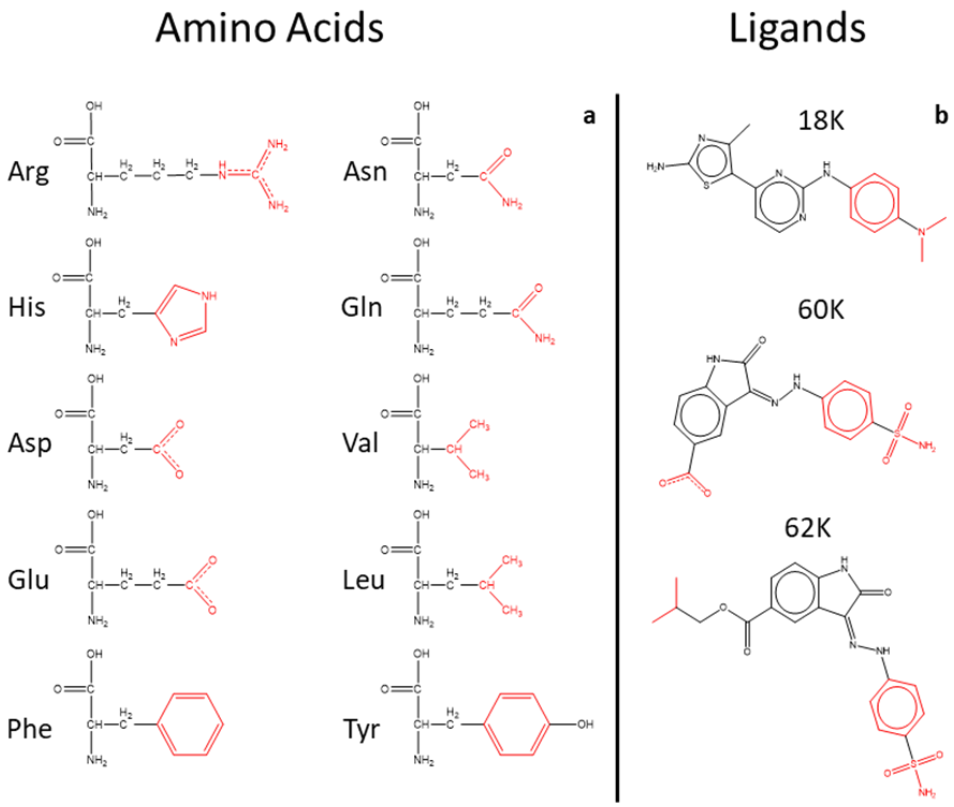


Figure 3.4. Chemical structures for **a** the amino acids and **b** the ligands residues where atoms represented in red are clustered.

The iterative process will end when no more distances are present in the main CV from the last iteration n, which can be associated with the fact that there are

no more stable interactions between the ligand and the protein, suggesting that the ligand is outside the pocket.

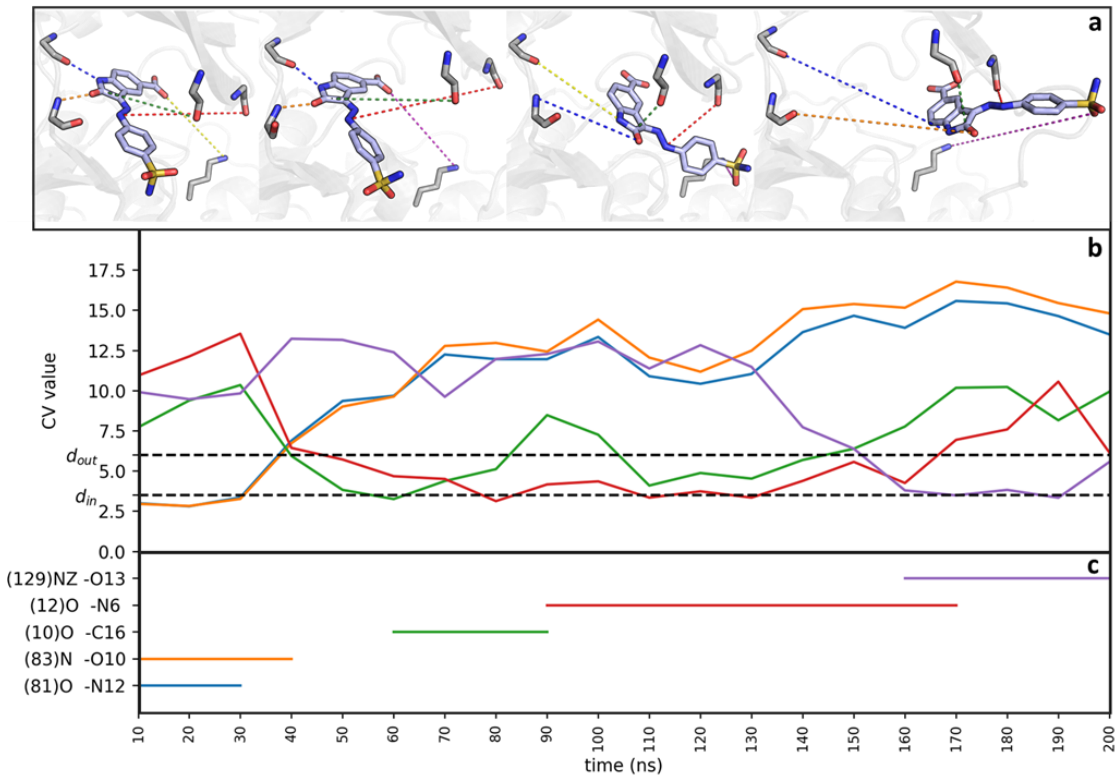


Figure 3.5. **a** Unbinding trajectory for the example of 60K represented as several snapshots along the trajectory. The relative distances used for the bias, **b** shows the same distances during the trajectory, the lower dashed line is the cut-off below which interaction is included in the main CV, the upper cut-off is the value above which the distance is excluded from the CV. **c** shows the same distances when they are included in the CV.

Figure 3.5 shows a representative result of the process for selected representative interactions, which illustrates that some distances (blue and orange) are identified from the initial unbinding trajectory, later in the unbinding process at 60 ns a new interaction is found (green line) and at 90, and 160 ns more distances are included in the main CV (red and purple respectively).

3.4.3 Free Energy Calculation

Once the ligand is outside the pocket, to determine the free energy path along with our set of CVs, we used a combination of finite-temperature string method and umbrella sampling method using as initial path and set of distances the ones obtained from the unbinding trajectory [40], [85], [93].

We collect all the distances found in the unbinding trajectory, and we extract the value of each interatomic distance along the unbinding path to construct the refined unbinding pathway, building a string of windows (100) of the coordinate space. For each window, and each distance we set the position to restrain equidistantly along the initial fitted string, using a force constant of 20 kcal/mol/Å^2 for a total time of 5 ns per window. From the obtained set of trajectories, a high-order (8) polynomial fitting is applied using the average collective coordinate to built the subsequent set of refined positions of the CVs.

We unbiased the simulations using the binless implementation [85] of the weighted histogram analysis method (WHAM) [42]. Temperature is set to 298K to be consistent with the experiments and simulations. The procedure is carried out until a convergence value is obtained by checking if all the CVs' changes along the different iterations are below a given threshold [85].

By adding multiple overlapping biasing potentials along the dissociation pathways which are parametrized as the multiple ICs, the string simulations can sufficiently sample all points of the ICs.

3.4.4 Transition State Analysis

To perform our novel Machine Learning Transition State Analysis (MLTSA), from the PMF obtained using WHAM, we then choose the closest five windows to the TS point of the PMF (Figure 3.6) as a starting coordinate for producing multiple unbiased MD simulations for each of the three ligands. From each of these five starting coordinates, we then run 50 independent unbiased MD simulations 5 ns long each. We then classify and label them by considering a combination of key distances, which simulations finish with the ligand in either a bound position (IN) or an unbound position (OUT).

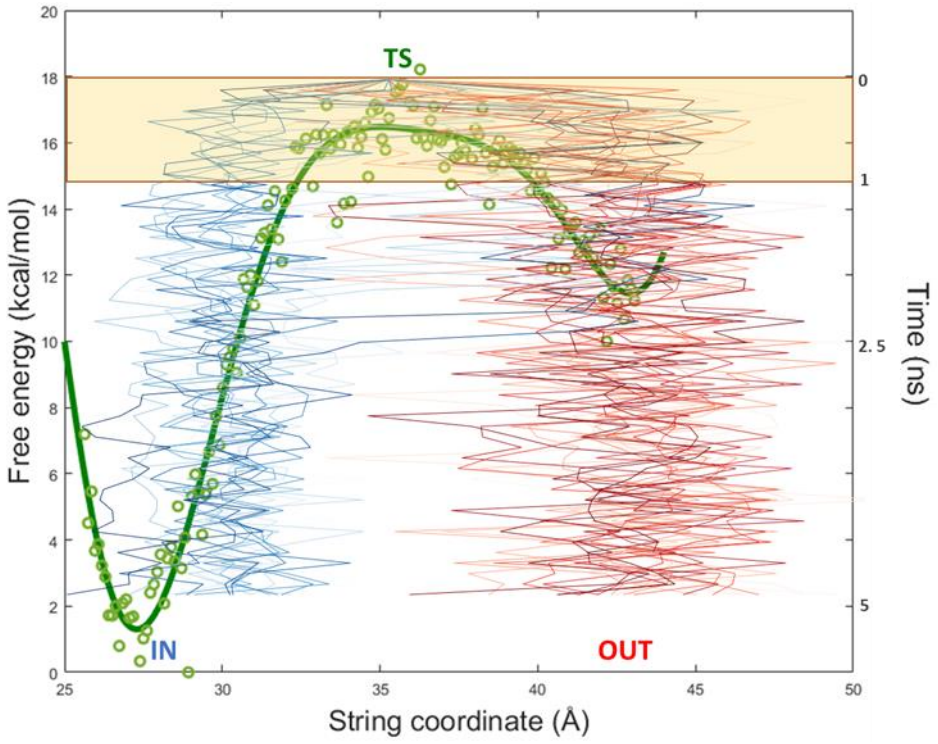


Figure 3.6. Representation of the TS along the PMF of 60K. From the TS coordinate as a starting point, a set of simulations leading to both an IN position (blue) and an OUT position (red) are represented as lines. The green dots represent the free energy profile obtained from the WHAM calculation using the string window as string coordinate, and as a green line, the fitting obtained

from the green dots. The area colored in yellow represents the simulation time used for analysis during our machine learning approach.

Once we found the window that provides the closest to a 1:1 ratio of IN and OUT events for these trajectories; we then run 200 additional unbiased MD simulations 5 ns long each, where we track all the interatomic distances initially (i.e., at the TS) within 5 Å from the ligand.

The next step was to train a Neural Network to analyze the downhill trajectories' dataset and predict their possible outcome from early on data, i.e. at 0.3 ns. The training was performed using Scikit-learn's implementation [94]. We trained a simple Multi-layer Perceptron (MLP) Classifier, made of three main layers (input, hidden, and output), with a hidden layer of 100 neurons, optimized using the Adam solver [95] and using the ReLu [96] function as an activation function, with a limit of 200 maximum iterations over data. Thus, using the frames coming from the multiple short, unbiased MD simulation trajectories starting from our TS, we provided a dataset of distances extracted along the trajectory, as well as the outcome of the IN or OUT events as the desired answer so that it would become a classification problem. We performed the training with trajectories of several different lengths (Figure 3.9), to observe the predicted accuracy at different time ranges along the simulations.

Once we obtained a trained model, the next step is to understand which features from this set of distances are important for the model to predict whether the simulation is going to bound (IN) or unbound (OUT) position. To do so, we then apply our feature reduction approach (FR), in which progressively every single distance is excluded from the analysis, and we check the drop in accuracy compared to the full set of distances. Differently from the standard approach, where the real value of each feature is replaced with a zero, here we replace each real value with the global mean of the selected features along with the simulations.

3.5 Results and Discussion

For each system, we perform three independent replicas starting from the crystal structure position. We perform the initial unbiased MD simulation for each replica, followed by our unbinding trajectory procedure and calculating the free energy profile from the finite temperature string method upon reaching convergence. The energy barrier extracted from the PMF of our simulations agrees with the experimental results (Table 3.1).

Table 3.1. Kinetics and thermodynamics values of the three systems from Dunbar et al. 2013 [89] and calculated results obtained from our computational simulations.

PDB	Ligand	$^{1}K_D$ (nM)	$^{1}k_{on}$ [$M^{-1}s^{-1}$]	$^{1}k_{off}$ [s^{-1}]	$^{1}\Delta g_{exp}$ (kcal/mol)	Δg_{calc} (kcal/mol)
3sw4	18K	9.61E-07	100030	0.0823	18.93	**18.04(±1.11)**
4fku	60K	9.86E-08	132366.7	0.0133	20.01	**15.61(±0.99)**
4fkw	62K	4.73E-08	64920	0.00261	20.97	**24.97(±3.19)**

For ligand 18K, we obtain the same energy barrier of the experimental data; however, both ligands 60K and 62K show a deviation of ~ 4 kcal/mol (Figure 3.7). The last could be due to the limitations of the current force field, the convergence of the sampling, and the sensitivity of the experimental data. We believe that our energy barriers results to be close to the experimental values, but also relatively high as we introduce in our CV distances that are not only taken from the initial coordinate but instead by introducing distances/interactions that are found along the unbinding path (Figure 3.6).

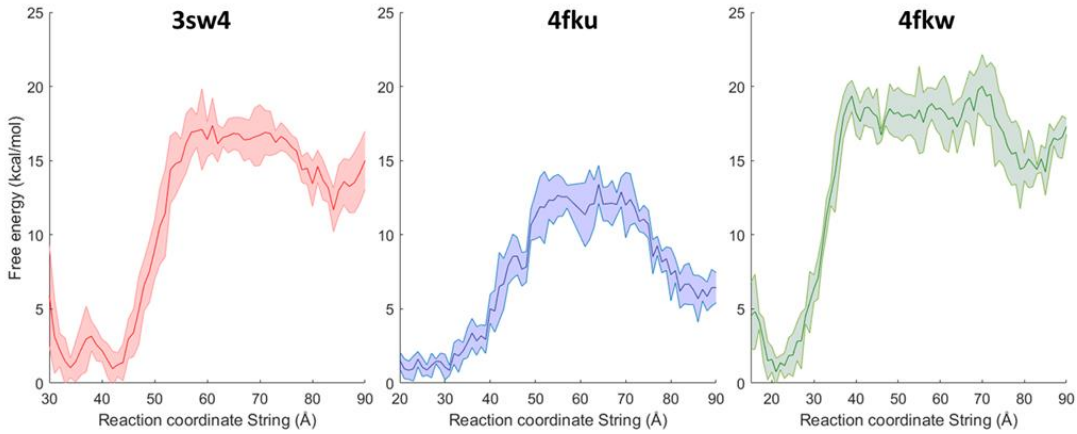


Figure 3.7. PMF of the unbinding path for 3sw4 (a) 4fkw (b), and 4fku (c). The free energy profile is obtained from a representative replica, the standard error, shown as shaded area are obtained by dividing the full dataset into four subgroups.

While different unbinding trajectories might lead to slightly different variations due to multiple local minima along the paths, we typically expect that the main transition state ensembles would be captured by all of these paths similarly after the convergence to the minimum free energy pathway. This is the main underlying assumption behind the finite temperature string method, which was proven to work very well even for complex systems [97], [98].. For all the three models, one important interaction is between the backbone carboxyl of His84 and a central N of the ligand (Figure 3.8). This H-bond has been reported as a key interaction in many ligands in complex with CDK2 [99]. These distances were found and already included from the initial unbinding simulation in each of the three systems. However, during the unbinding trajectory, once this important H-bond (His84(O)-N) is broken, new interactions are formed, for a varying time scale. For 3sw4, in all the three replicas, H-bonds are formed with the external amino group of the ligand (N5) and the backbone oxygen of Glu81 and later on with the backbone oxygen of His84. 60K and 62K molecules present a sulphamide terminal group, which, during the trajectory, interacts with Val163 and His84, for 4fku and 4fkw respectively.

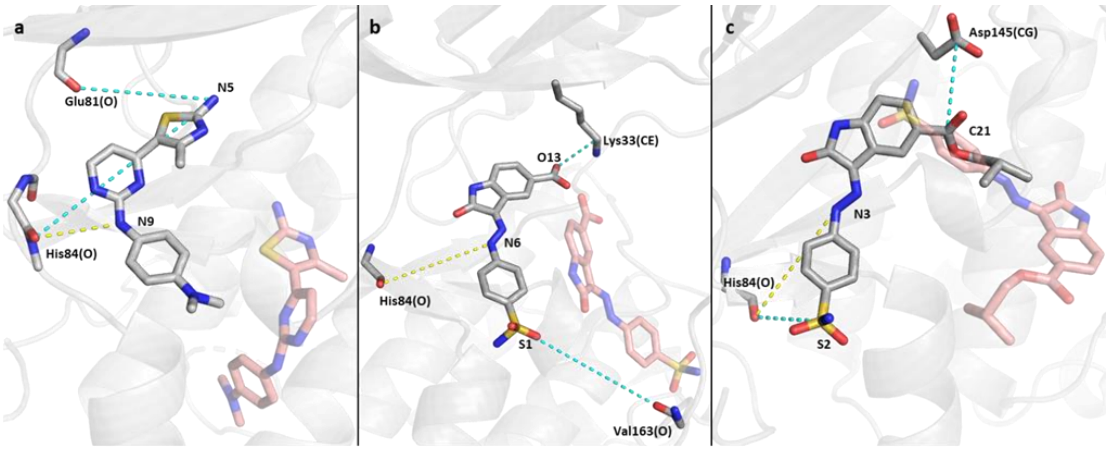


Figure 3.8. CV obtained from the unbinding of 18K (a), 60K (b), and 62K (c); representative distances represented in dashed lines (yellow: interaction from the initial coordinate, cyan: interaction found during the unbinding trajectory), colored in red represent the coordinate of the ligand when is outside the pocket.

These distances appear in each of the three replicas for each system.

From the MLTSA results, we obtain a list of distances for each system that are major determinants for predicting the bound or unbound states (Figure 3.11). By analyzing trajectory data up to 0.3 ns of each downhill simulation, the model can predict with high accuracy the IN or OUT directionality of the trajectories, more specifically: 80.11% for 18K, 90.44% for 60K and 93.83% for 62K respectively (Figure 3.9).

The ML training's effectiveness is confirmed by comparing the accuracy of predicting the trajectory outcomes using our original final free energy reaction coordinate. When analyzing the initial parts of the TS-initiated trajectories (0.3 ns from the total of 5 ns), we find that ML is able to predict much more accurately, about 80% versus ~60%, the final IN or OUT states as compared to using the string reaction coordinate value for the prediction (Figure 3.10).

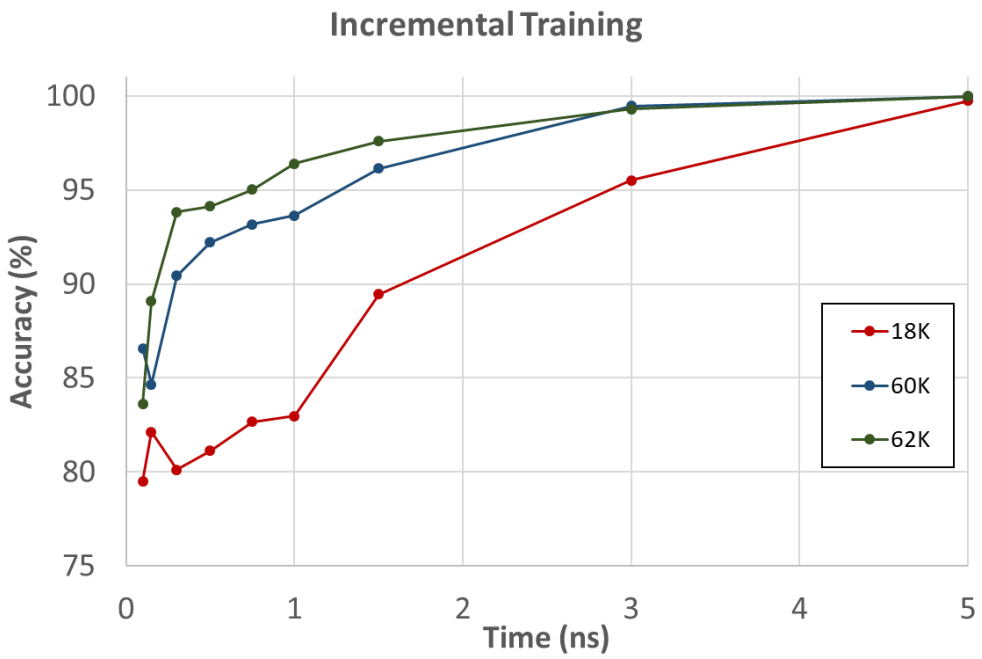


Figure 3.9. Accuracy prediction at a different time step of the simulations using the MLTSA for 18K in red, 60K in blue, and 62K in green.

Using the trained model, we then perform a feature reduction analysis to identify which CV features affect the most the overall prediction ability of the ML model. For all three molecules, we can select the most important structural features (Figure 3.11b,c,e), which leads to the significant reduction of the prediction accuracy, when this feature is eliminated (constant values fed to the ML). In contrast, other features do not affect the overall accuracy of the predictions.

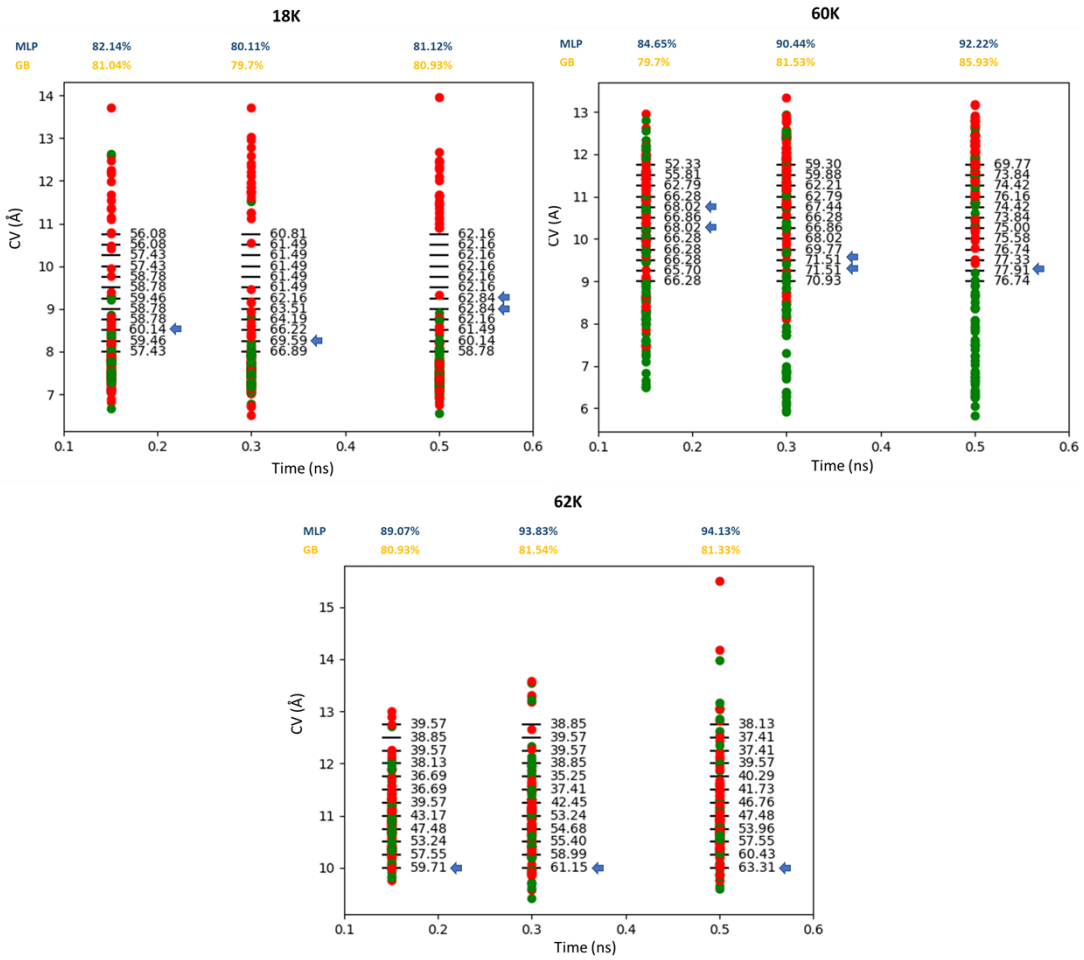


Figure 3.10. Manual check at different time points (0.15, 0.3 and 0.5 ns) for the three ligands compared with the results obtained from the ML.

We also compare the feature reduction approach's validity with a different machine learning algorithm named Gradient Boosting (GB). While the GB algorithm is inferior to the multilayer perceptron ML model, the results obtained show some similarity with our main MLTSA approach (see Figure 3.12). This suggests that alternative ML models may also be used successfully and further validates our results.

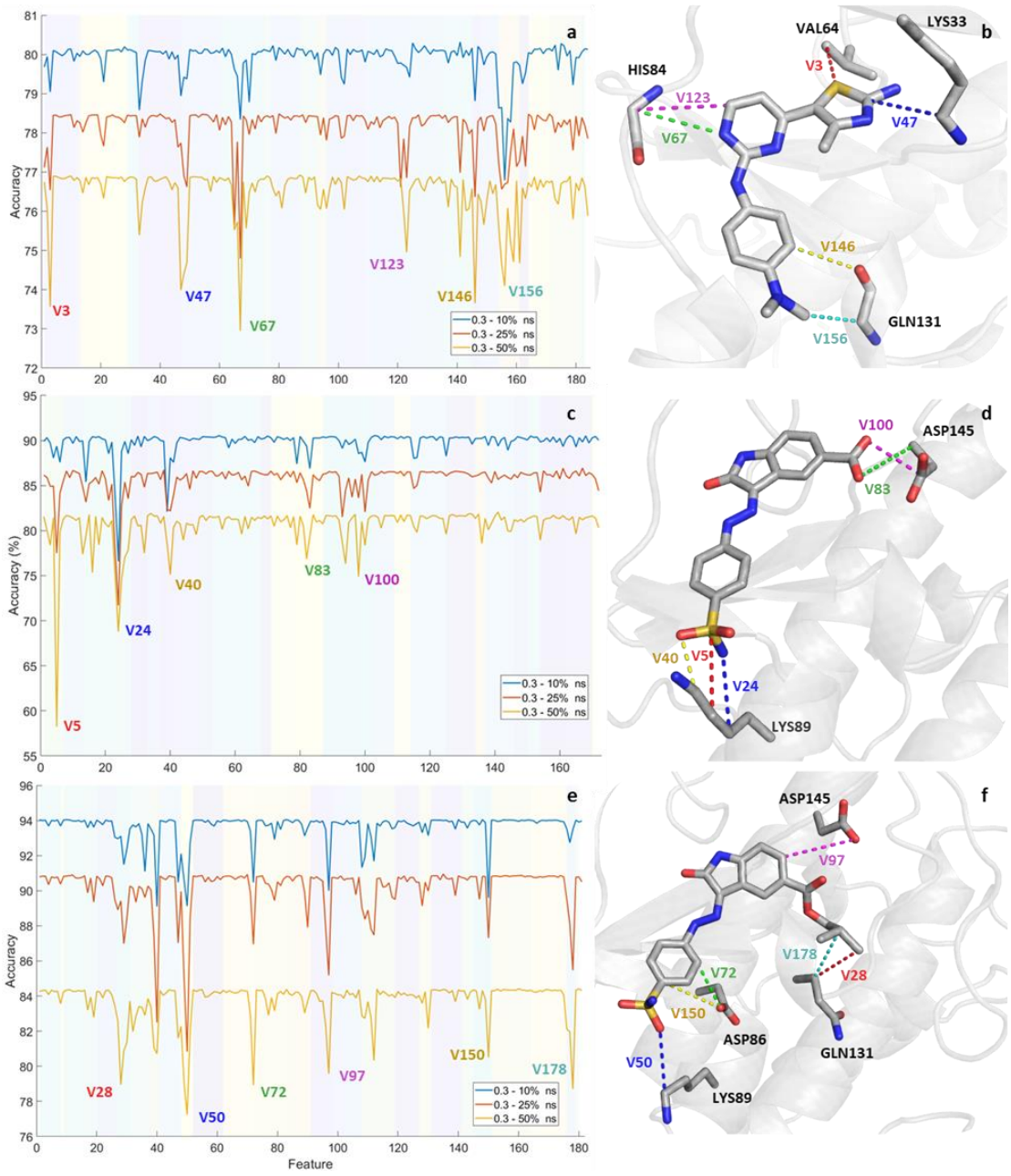


Figure 3.11. Identification of the essential distances from the feature reduction analysis at 0.3 ns using the last 50% (yellow), 25%(red), and 10%(blue) of data points for a: 18K, c: 60K, e: 62K. The different colours of the background groups the different features according to the atom of the ligand involved. Features presenting high accuracy drop are labelled and shown graphically in the right side of each plot: b:18K, d:60K, and f:62K.

Looking at the structures of the ligands, it appears that important interactions for the TS are between the protein and the extremities of the ligands. Interestingly, the main features used to define the IN/OUT outcome of the

ligand are not important according to the training. For ligand 60K and 62K, due to their similarities, important interactions are between the sulphidic group of the ligands and Lys89 from one extremity, and with the Asp145 for the other extremity, for ligand 60K is with the carboxylic group, and for ligand 62K is with the ester group (Figure 3.11d,f). Ligand 18K, similarly to the other two ligands, present important interactions between the protein and the external atoms of the ligand, however, because structurally different from the other ligands, the interaction is between different residues (Figure 3.11c).

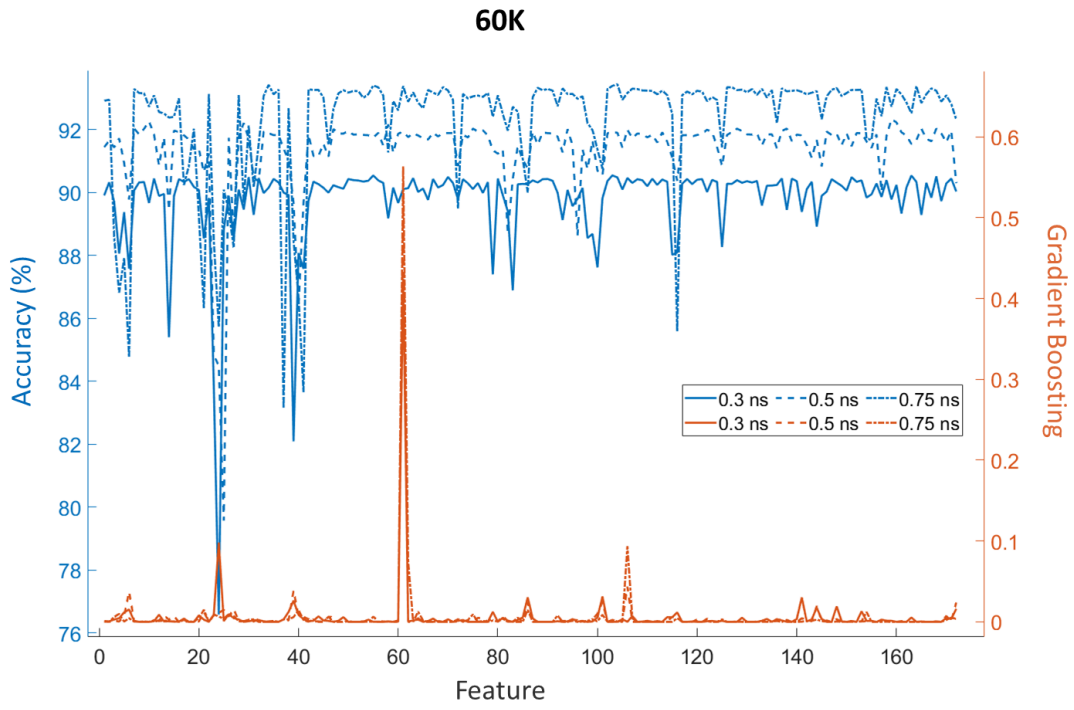


Figure 3.12. Comparison of the FR and GB approach for ligand 60K.

3.6 Conclusion

Optimizing ligand unbinding kinetics is a very challenging design problem for small molecule drug discovery that can lead to the development of drugs with superior efficacy. To tackle this, we have developed a new method, which allows

us to calculate the free energy barrier for the ligand unbinding step, therefore providing quantitative information about the residence time of a specific ligand. Our method involves first an exploration step, where a ligand unbinding path is determined together with key initial variables that describe this path. Subsequently, we perform accurate free energy calculations using the complete set of identified interactions as CVs along the unbinding path via the finite temperature string method. This provides us with the free energy barriers, and an ensemble of structures at the transition state of the ligand unbinding process. The novelty of the method lies in the combination of automated iterative addition and removal of the collective variables determining an unbinding trajectory, which allows us to discover novel interactions not available *a priori*, just based on the interactions from the ligand-bound structure. The combination of the unbinding path to find the CVs and the umbrella sampling-based string method with high dimensional reaction coordinates provides an efficient way to obtain quantitative kinetics of ligand unbinding.

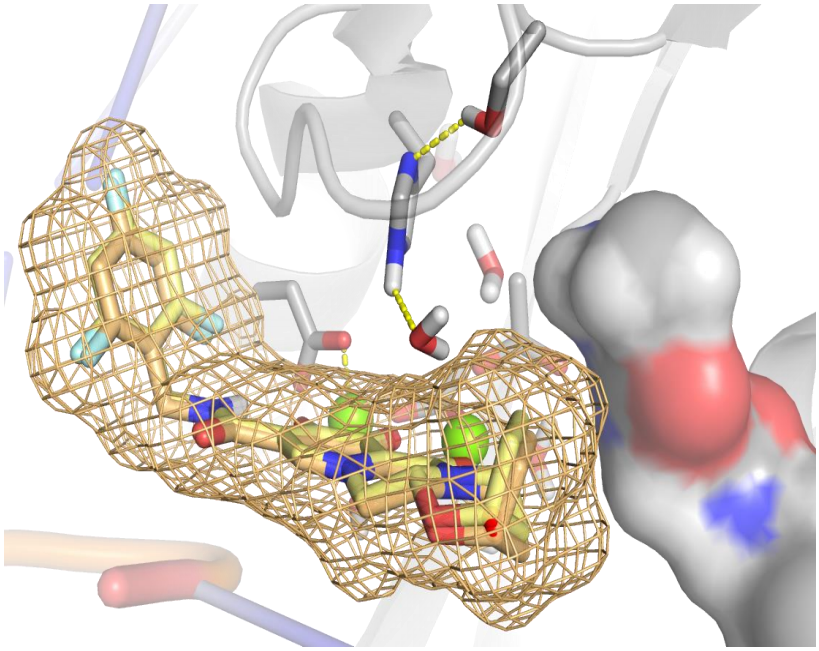
We tested this method using a well-studied cancer drug target, CDK2, using three drug molecules with known kinetic profiles. We obtained high energy barriers corresponding to experimental using our method, which demonstrates the fundamental importance of determining a well-selected, high-dimensional set of CVs obtain correct free energy profiles and kinetics results.

To aid the kinetics-based design of novel compounds, we also developed a novel method, MLTSA, that allows us to identify the most important features determining the transition state paths. Here, we generated multiple trajectories initiated at the TS of the unbinding process, which either terminated in the bound state or in the unbound state. We then trained a multilayer perceptron ML algorithm to predict the outcome of the trajectories by using a set of CVs and the initial segment of the trajectories only. By doing so, we demonstrated

that the ML was able to predict the trajectory outcomes with much more accuracy than it was available to us, using the original set of CVs used for the free energy calculations. A feature importance analysis was further developed to identify which key CVs and corresponding structural features determined the faith of the trajectories, and therefore are the most important descriptors of the TS.

Qualitatively, we identified novel interactions between the protein and specific parts of the ligands that were of major importance for the trajectories to pass the TS. These corresponded to protein interactions at the TS-bound poses with functional groups of the distal ends of the ligands. Importantly, to perform this analysis, we require the knowledge of the TS structures as well as the MLTSA analysis of a set of trajectories from these initial points. Our algorithms enable us to uncover novel design objectives for a kinetics-based drug discovery process.

Chapter 4 Structural Basis of Second-Generation HIV Integrase Inhibitor Action and Viral Resistance



4.1 Preface

The work presented in this chapter started from the collaboration between our group and the experimental work from the group of Cherepanov lab at the Francis Crick Institute. The project aimed to understand the interaction between known inhibitors used for the treatment of HIV with the integrase protein at the atomic level. The group of Cherepanov successfully provided cryo-electron microscopy structures of the integrase proteins, in its wild type form and Q148H/G140S mutant form in complex with Bictegravir and an analogue structure of Doluctegravir (analog 1). From the structure obtained, my contribution was to understand, through computational tools, the difference in the interaction between the two drugs, as despite both Dolutegravir analogue's and Bictegravir belongs to the second generation of inhibitors, the two molecules shows different inhibition activity. In this chapter, I set up, perform and analyse the results from the unbiased molecular dynamic simulations and analyse the data from the QM calculations.

This work was successfully published in Science:

Nicola J. Cook, Wen Li, Dénes Berta, **Magd Badaoui**, Allison Ballandras-Colas, Andrea Nans, Abhay Kotecha, Edina Rosta, Alan N. Engelman, and Peter Cherepanov. "Structural basis of second-generation HIV integrase inhibitor action and viral resistance." *Science* 367, no. 6479 (2020): 806-810.

4.2 Abstract

Despite worldwide prescription, the mechanistic basis for superiority of second-generation HIV integrase (IN) strand transfer inhibitors (INSTIs) is poorly understood. We used single-particle cryo-electron microscopy to visualize the mode of action of the advanced INSTIs Dolutegravir and Bictegravir at near-atomic resolution. Q148H/G140S amino acid substitutions in IN that pervade clinical INSTI failure perturb optimal magnesium ion coordination in the enzyme active site. The expanded chemical scaffolds of second-generation compounds mediate interactions with the protein backbone, which are critical for antagonizing Q148H/G140S mutant virus. Our results reveal that binding to magnesium ions underpins a fundamental weakness of the INSTI pharmacophore that is exploited by the virus to engender resistance and provide a structural framework for the development of this important class of anti-HIV/AIDS therapeutics.

4.3 Introduction

Despite the immediate clinical impact, the first-in-class INSTI Raltegravir (RAL) suffered setbacks from the emergence of viral resistance [100]. Although second-generation INSTIs Dolutegravir (DTG) and Bictegravir (BIC) display improved activity against RAL-resistant strains [101], [102], the advanced compounds are not immune to resistance [102]–[106]. In particular, Q148H/G140S changes in HIV-1 IN are associated with complete or partial loss of efficacy across the entire drug class. The mode of INSTI binding to the IN active site was first visualized in the context of the prototype foamy virus (PFV) intasome [107]. However, the limited ~15% amino acid sequence identity between PFV and HIV-1 INs greatly restricts the utility of PFV for studies of INSTI resistance and precludes its use as a template for structure-based lead optimization. Conversely, unfavourable biochemical properties of the HIV-1 intasome have impeded structural refinements to atomic resolution [108].

In order to establish a robust experimental system suitable for informing INSTI development, our collaborators evaluated IN proteins from primate lentiviruses that are highly related to circulating strains of HIV-1. The simian immunodeficiency virus from red-capped mangabeys (SIVrcm) is a direct ancestor of chimpanzee SIV [109], [110]. Because the HIV-1 pol gene is originally derived from SIVrcm, the viruses share as much as 75% IN amino acid sequence identity. SIVrcm IN displayed robust strand transfer activity in vitro, which was stimulated by the lentiviral IN host factor LEDGF/p75 [111], [112]. Reaction conditions were conducive for the formation of stable nucleoprotein complexes, which were competent for strand transfer activity and sensitive to INSTI inhibition. Examining the material by negative stain electron microscopy

(EM) revealed a heterogeneous population with the prominent presence of long linear polymers (hereafter referred to as stacks, Figure 4.1A).

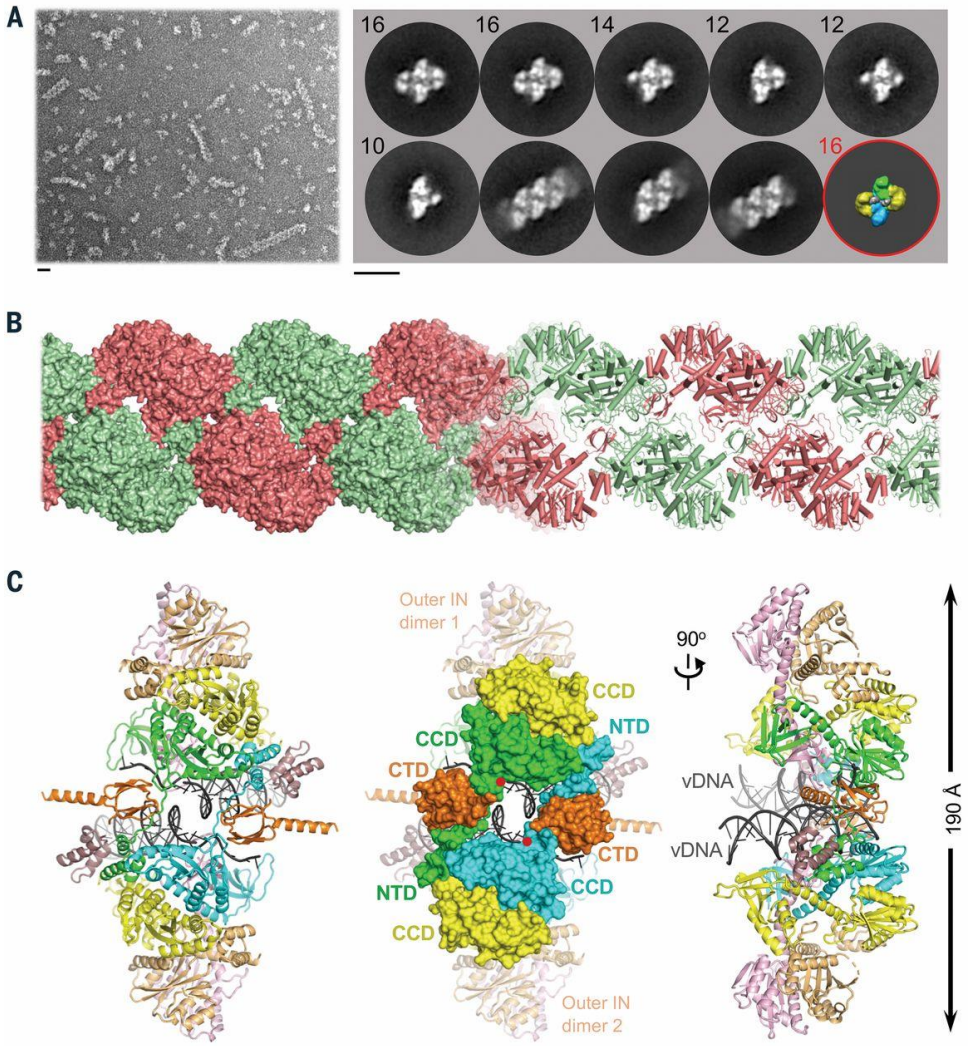


Figure 4.1. Reconstruction of the SIVrcm intasome core. (A) Raw image (left) and 2D class averages (right) of negatively stained SIVrcm intasome particles; apparent numbers of IN subunits are indicated for non-stacked classes. The envelope of the hexadecameric maedi-visna virus intasome (red circle; central and flanking IN tetramers in blue/green and yellow, respectively) is shown for comparison; scale bars are 0.2 nm. (B) Atomistic reconstruction of the SIVrcm intasome stack shown as space fill (left) and cartoons (right); separate repeat units are shown in alternating red and green colours. (C) Detailed view of a single intasomal repeat representing a pair of viral DNA ends (vDNA, grey cartoons) synapsed between a pair of IN tetramers (composed of yellow, orange, pink, and either green or cyan IN protomers; the active sites of the green and cyan molecules (red dots) catalyze DNA recombination). The repeat unit is completed by pairs of C-terminal (orange) and N-terminal (dark

magenta) domains donated by IN chains belonging to neighbouring repeats. These CTDs are critical to form the conserved intasome core (CIC), which is shown in space fill mode in the middle panel. CCD, catalytic core domain.

Reference-free classification revealed 2D averages that were strikingly similar to those observed in maedi-visna virus (MVV) intasome preparations (Figure 4.2A) [113]. However, while the latter behaved as a near-monodispersed population with a predominance of hexadecamers (tetramer-of-tetramers) of IN, the flanking IN tetramers of SIVrcm intasomes were notably disordered, often nucleating stack formation. Although HIV-1 IN assembly was much less efficient, it yielded particles visually indistinguishable from SIVrcm intasomes. These observations are consistent with polydispersity previously reported in HIV-1 intasomes assembled with a hyperactive IN mutant [108]. 2D class averages apparently corresponding to the dodecameric assembly from that study were readily identified in our wild type HIV-1 and SIVrcm intasome images (Figure 4.2A).

Our collaborators recorded micrograph movies of unstained SIVrcm intasome stacks in vitreous ice using a direct electron detector and refined the cryo-EM structure of an averaged intasome repeat unit. To prevent DNA binding to the target binding groove, which would occlude INSTI occupancy [114], the intasomes were prepared using A119D IN that precludes target DNA capture without affecting IN active site function [115]–[117]. The overall resolution of the reconstruction throughout the conserved intasome core (CIC) was 3.3 Å, while the local resolution of the active site region approached 2.8. In agreement with the resolution metrics, the cryo-EM density map was sufficiently detailed to build and refine an atomic model. The resulting model encompassed two IN tetramers with associated viral DNA ends, as well as two pairs of C- and N-terminal domains (CTDs and NTDs) donated by flanking stack units (Figure 4.1 B and C). Exchange of NTDs and CTDs between neighbouring intasomes forms the structural basis for stack formation.

Using available nucleotide sequence data [109], our collaborators engineered recombinant SIVrcm and evaluated its sensitivity to INSTIs. First- (RAL) and second- (DTG, BIC) generation INSTIs inhibited HIV-1 and SIVrcm at similar 50% effective concentrations (EC50) (Figure 4.2A). Q148H/G140S changes in IN rendered both HIV-1 and SIVrcm >2,000-fold resistant to RAL, while EC50 values of the second-generation INSTIs BIC and DTG increased ~5 to 8-fold against HIV-1 and 40 to 73-fold against SIVrcm. Importantly, the majority of residues that when altered confer INSTI resistance are conserved between HIV-1 and SIVrcm. An exception is Thr138: in HIV-1, E138T potentiates resistance of Q148H-containing viruses [19,20]. Concordantly, reverting Thr138 to Glu decreased DTG and BIC resistance of Q148H/G140S SIVrcm to the levels observed with HIV-1 Q148H/G140S. Moreover, T97A/L74M, which increase resistance of Q148H/G140S HIV-1 to second-generation INSTIs [106], exerted the same effect on SIVrcm.

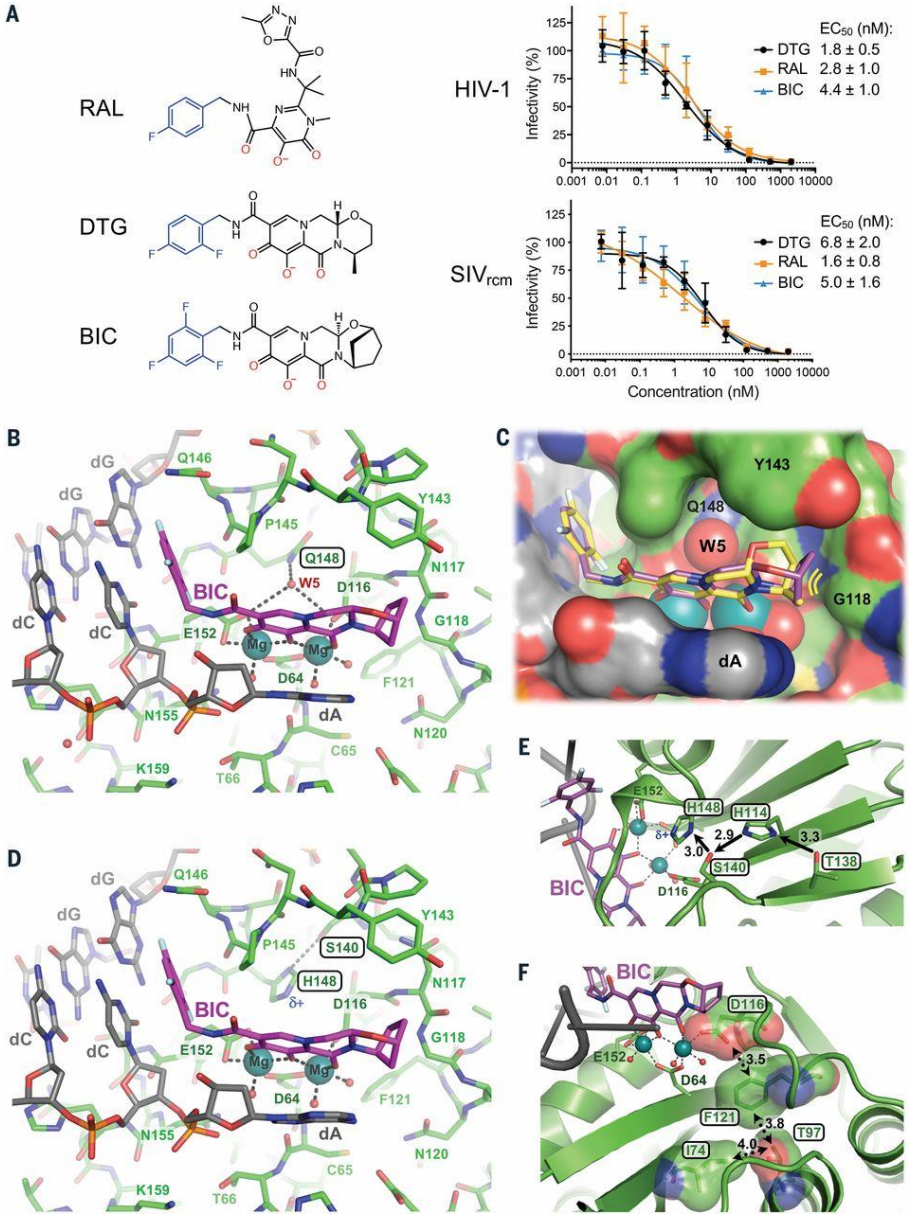


Figure 4.2. Binding modes of second-generation INSTIs in the IN active site. (A) Chemical structures of select first- (RAL) and second-generation (DTG, BIC) INSTIs (left; halo-benzyl groups in blue and metal-chelating oxygen atoms in red) and viral sensitivities (right). Results are averages and standard deviations of minimally n = 2 experiments, with each experiment conducted in triplicate; EC50 values are noted. (B) The active site of the SIVrcm intasome in complex with BIC; protein, DNA, and drug are shown as sticks. Blue spheres are Mg2+ ions, water molecules are shown as small red spheres. (C) Superposition of BIC (magenta) and DTG (yellow) bound structures with protein and DNA are shown in space-fill mode. Yellow lines accentuate proximity to IN β4-α2 connector. (D) Q148H/G140S active site bound to BIC. δ+ indicates increased electropositivity of the His148 Nε2 proton. (E) The extended hydrogen bond network that couples Thr138 to His148 in the Q148H/G140S SIVrcm intasome. Black arrows

indicate hydrogen bond donation; the corresponding interatomic distances are given in Ångstroms. (F) Long-range interactions of Ile74 and Thr97 with the chelating metal cluster via Phe121. Key amino acid residues are shown as sticks and semi-transparent van der Waals surfaces. Contacts between side-chain atoms are indicated by double-headed dotted arrows with distances given in Ångstroms.

Encouraged by these results, our collaborators acquired cryo-EM data on SIVrcm intasomes vitrified in the presence of INSTIs and Mg2+ ions. DTG- and BIC-bound structures were reconstructed to resolutions of 3.0 and 2.6 Å across the CIC, with local resolutions within active site regions of 2.8 and 2.4 Å. The inhibitors were defined remarkably well in density maps, allowing their refinements with bound Mg2+ ions and associated water molecules. The invariant IN active site carboxylates Asp64, Asp116, and Glu152 coordinate a pair of Mg2+ ions, which in turn interact with the metal chelating cores of the INSTIs (Figure 4.2B and C). As previously observed in PFV intasome crystals [107], the drugs displace the 3’ viral DNA nucleotide, which stacks against the central body of the INSTI. In agreement with low-level amino acid sequence identity, there are considerable differences in the environment of the small molecules in the SIVrcm and PFV structures.

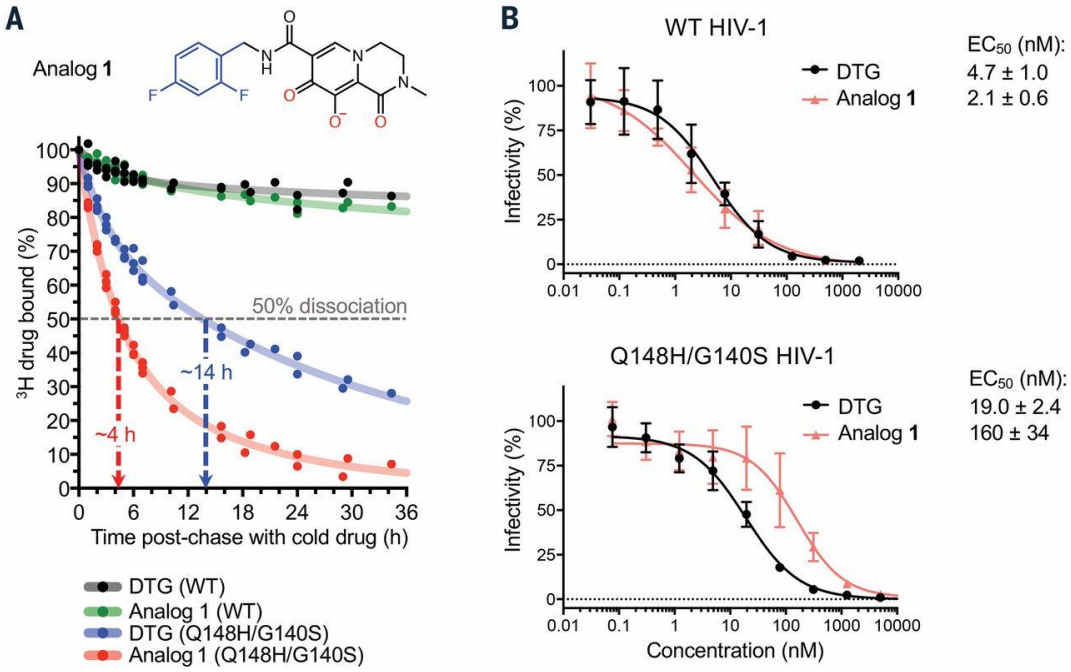


Figure 4.3. Effects of Q148H/G140S substitutions on DTG and analog 1 activities. (A) Structure of analog 1 (top; colours as in Figure 4.2A) and a time course of 3H-DTG and analog 1 dissociation from wild type and Q148H/G140S HIV-1 intasomes (bottom). Results from three independent experiments are plotted; each data point is an average of two measurements done in parallel; trendlines are for illustration purpose. Apparent INSTI dissociative half-times from the mutant intasome are indicated. (B) Activities of DTG and analog 1 against wild type (top) and Q148H/G140S (bottom) HIV-1. Results are averages and standard deviations of two independent experiments, with each experiment conducted in triplicate.

The map of the BIC complex revealed an interaction between the side-chain amide of Gln148 and the carboxylates of metal-chelating residues Glu152 and Asp116 via a water molecule (W5, Figure 4.2B). Molecular dynamics simulations confirmed the stability of this hydrogen bonding network (Figure 4.4). During the MD simulations, in the wt integrase the water W5 coordinating Glu152 and Asp116 is extremely stable, remaining the same molecule (represented with the same dot colour in Figure 4.4 A), while in the Q148H/G140S mutations the water molecule is continuously replaced by different water molecules (the different dot colours correspond to exchanges of the water molecule with bulk solvent in Figure 4.4 B). DTG and BIC intimately contact the backbone atoms of Asn117

and Gly118 from the IN β4-α2 connector, making respectively 8 and 12 contacts with interatomic distances ≤5 Å. Moreover, BIC makes three contacts with interatomic distances of 3.9-4.0 Å within this active site region. We obtained a truncated INSTI derivative lacking the heterocycle involved in these interactions to test their importance to drug potency (analog 1, Figure 4.3A). This modification was not expected to impact the metal chelating properties of the compound or its ability to stack with DNA bases, and indeed analog 1 and DTG similarly inhibited HIV-1 infection. However, in contrast to DTG, analog 1 was ~80-fold less effective against HIV-1 Q148H/G140S (Figure 4.3B). In agreement with published work [120], the amino acid substitutions increased the dissociative rate of DTG from HIV-1 intasomes, while their impact on the truncated derivative was much greater (Figure 4.3A). Collectively, these data implicate contacts with the β4-α2 connector as a crucial feature of the second-generation INSTIs.

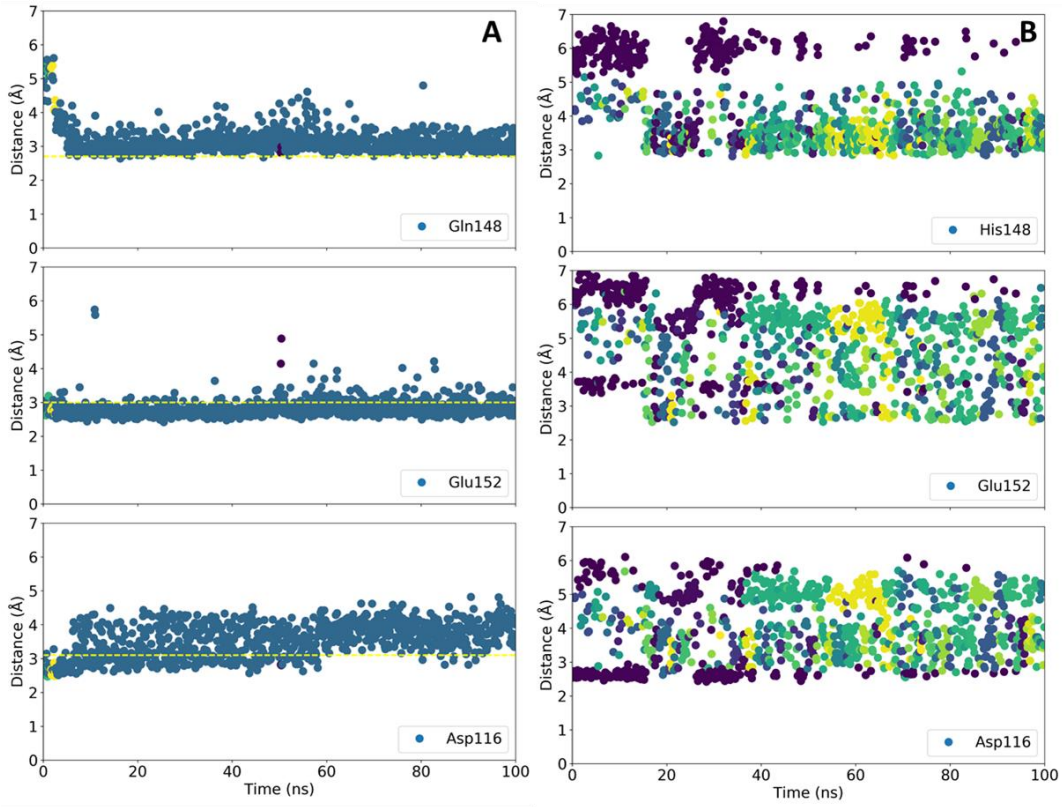


Figure 4.4. The behaviour of water molecules shared by Gln148 or His148 and active site carboxylates. (A) The structure of WT SIVrcm intasome bound to BIC was stripped of water molecules not directly coordinated to metal ions, embedded in the bulk solvent, and subjected to 100 ns of molecular dynamics. The bulk solvent-derived water molecule closest to Gln148 Ne2 and the carboxylates of Glu152 and Asp116 was identified in each frame of the simulation. The dot plots report the corresponding distances every 0.1 ns of the molecular dynamics. The position corresponding to that occupied by W5 in the structure becomes occupied by a stably bound water molecule during the initial 5 ns of the simulation. Alterations of dot colours correspond to exchanges of the water molecule with bulk solvent. (B) A similar analysis with the Q148H/G140S SIVrcm BIC structure. Here, a water molecule closest to His148 Ne2 and the carboxylates of Glu152 and Asp116 was identified in each frame of the simulation. Note frequent exchanges with bulk solvent; ~4 Å is considered to be the upper limit for hydrogen bonding.

To visualize the impact of the Q148H/G140S substitutions on drug binding, we imaged mutant SIVrcm intasomes in complex with BIC to a local resolution of 2.8. Ser140 and His148 side chains directly interact, and the latter positioned within 3.3 Å of the metal-chelating Glu152 carboxylate. In the refined model, steric clashes between the side chains are avoided by a 0.5-Å shift at the His148

Cα atom. Importantly, local crowding due to insertion of the mutant His148 side chain expelled water molecule W5 (Figure 4.4B), thus disturbing the secondary coordination shell of the Mg2+ ions. We also note that the amino acid changes caused a ~0.5-Å shift in the position of the bound drug; while arguably minor given the resolution of the cryo-EM map, the observed displacement agrees precisely with predictions by computational chemistry, illustrating the effect of the substitutions on drug binding. The Nε2 atom of His148, which intimately contacted the carboxylate of Glu152 (3.3 Å), is involved in bidentate coordination with one of the Mg2+ atoms. Importantly, the acidity of His148 Nε2 is increased due to hydrogen bonding of Nδ1 with Ser140 (Figure 4.2D). The Ser140-His148-Glu152 coupling is strikingly reminiscent of the non-catalytic Ser-His-Glu triad proposed as a stability determinant in α-amylases, representing a reversal of the charge relay system in hydrolase active sites [121], [122]. However, hydrogen bonding would require reorientation of IN Glu152 and His148 side chains, which would be incompatible with Mg2+ ion coordination and drug binding, suggesting an empirical interpretation of the INSTI resistance mechanism.

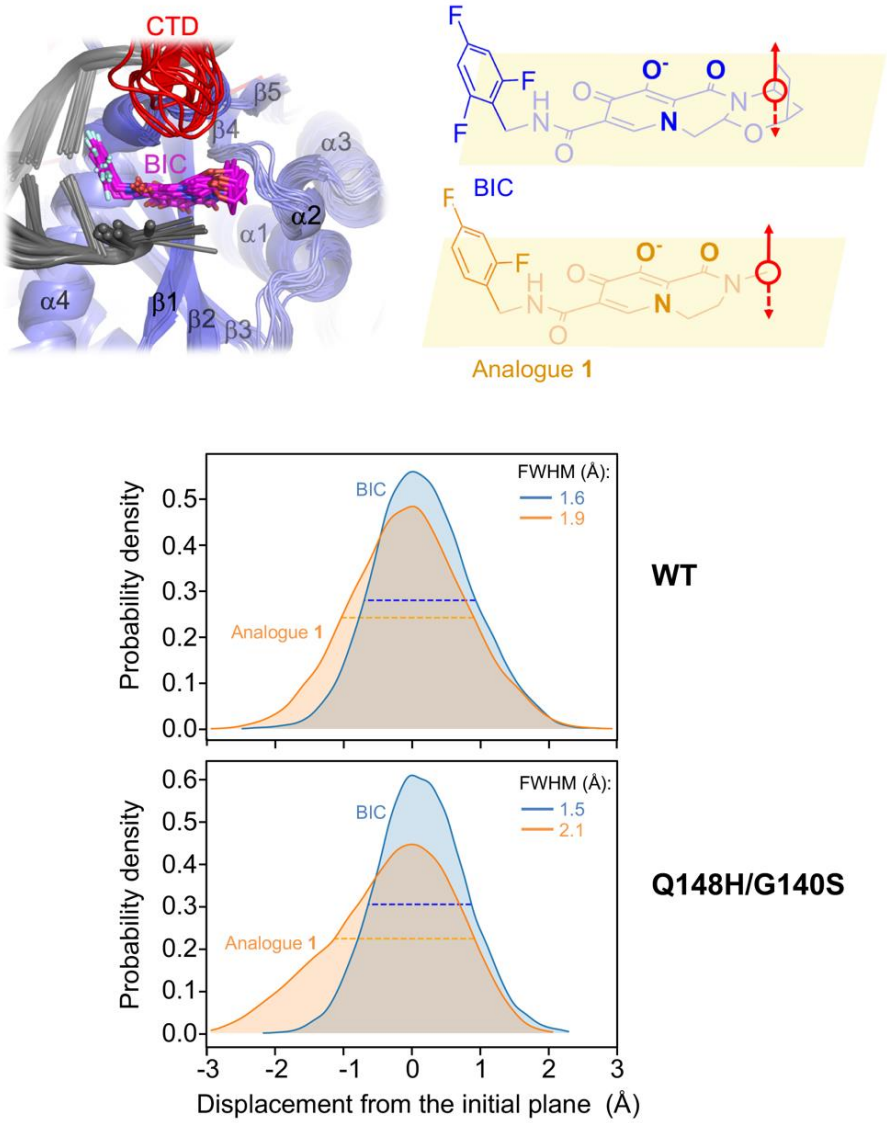


Figure 4.5. Dynamics of BIC and analogue-1 in the intasome active site. WT and Q148H/G140S SIVrcm intasomes bound to BIC or analogue 1 were subjected to MD simulation. The resulting frames (12,500 structures derived from a total of 250 ns simulation per condition) were aligned by Ca atoms of intasome active site residues. (A) A subset of 10 WT intasome-BIC complex frames separated by 10 ns of simulation. Protein and DNA are shown as cartoons and BIC as sticks. DNA is coloured grey and protein is coloured according to r.m.s. deviation from the initial position (blue, small displacement; red large displacement); the IN CTD and visible secondary structure elements of the CCD are indicated. (B) BIC and analogue 1 with the common carbon atoms closest to the b4-a2 connector when bound to the intasome active site indicated with red circles; arrowheads show direction of displacement chosen for the analysis. (C) Probability density for a given displacement of the chosen BIC (blue) or analogue-1 (orange) carbon atom from the initial plane defined by bolded atoms in panel B in

complex with WT (top) or Q148H/G140S (bottom) intasome. The full width at half maximum (FWHM) is listed for each distribution. Note a wider distribution of the atomic displacements in the case of analogue 1.

Our simulations show that analog 1 is considerably more dynamic in the active site compared to the full-sized molecule, the mobility of which is restricted through interactions with the β4-α2 connector (Figure 4.5 and Figure 4.6 **B,D**). By comparing the movement of a topologically identical atom in our multiple unbiased MD simulations we can see how the movement of that terminal segment of the ligand, presents a higher fluctuation in the case of analog 1 (Figure 4.6 **B, C** in the case of the wt and Q148H/G140S respectively), in contrast with BIC (Figure 4.6 **A, D** in the case of the wt and Q148H/G140S respectively) that shows a lower degree of freedom. The additional degree of freedom is expected to allow more extensive re-orientation of the truncated inhibitor, which may permit His148 to withdraw more electron density from the Mg2+-ligand cluster.

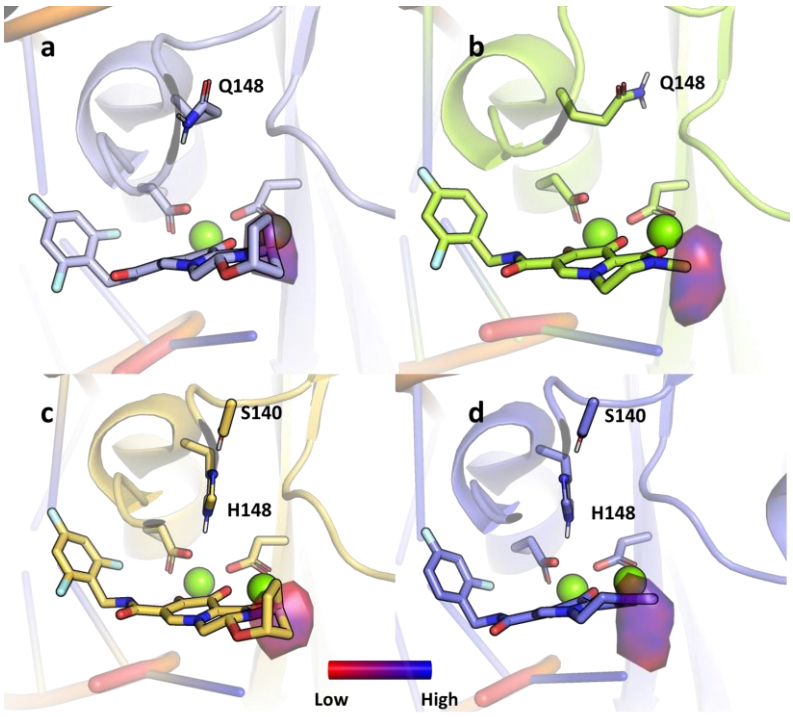


Figure 4.6. Spatial distribution of two topologically identical carbon atoms in ligands BIC (**a**) and analog 1 (**b**) in complex with wt integrase, and with

G140S/Q148H mutant integrase (respectively **c:** BIC and **d**: analog 1), the colouring of the surface represents the distribution of the C atom during the simulations.

Our natural bond orbital analysis illustrates the changes of atomic charge distribution within the cluster in response to polarization by protonated His148 Nε2 and subtle conformational adaptations (Figure 4.7). It is easy to extend these observations to the other substitutions at position 148, Lys and Arg, both of which introduce electropositive functionalities to yield high-level INSTI resistance [103].

Further work will be required to unravel long-range interactions involved in boosting INSTI resistance by secondary changes such as E138T and L74M/T97A [118], [119]. As a start, we analysed respective side chains in our SIVrcm Q148H/G140S intasome structure. Thr138 is ideally positioned to hydrogen bond with Nδ1 of conserved residue His114, prompting it to donate its Nε2 proton to Ser140 (Figure 4.2E). This extended network, which may form a proton wire, is expected to reinforce Ser140 as a hydrogen bond donor for its interaction with His148 Nδ1, explaining why the E138T substitution can enhance the resistance of Q148H/G140S HIV-1[19,20]. SIVrcm IN residues Ile74 (the position occupied by Leu or Ile in HIV-1 strains) and Thr97 are in close proximity to the side chain of conserved Phe121, which is involved in van der Waals interactions with the metal-chelating carboxylate of Asp116 (Figure 4.2F). Readjustment of the Phe121 side chain in response to changes in its local packing environment serves as a likely conduit to perturb the structural integrity of the metal-chelating cluster.

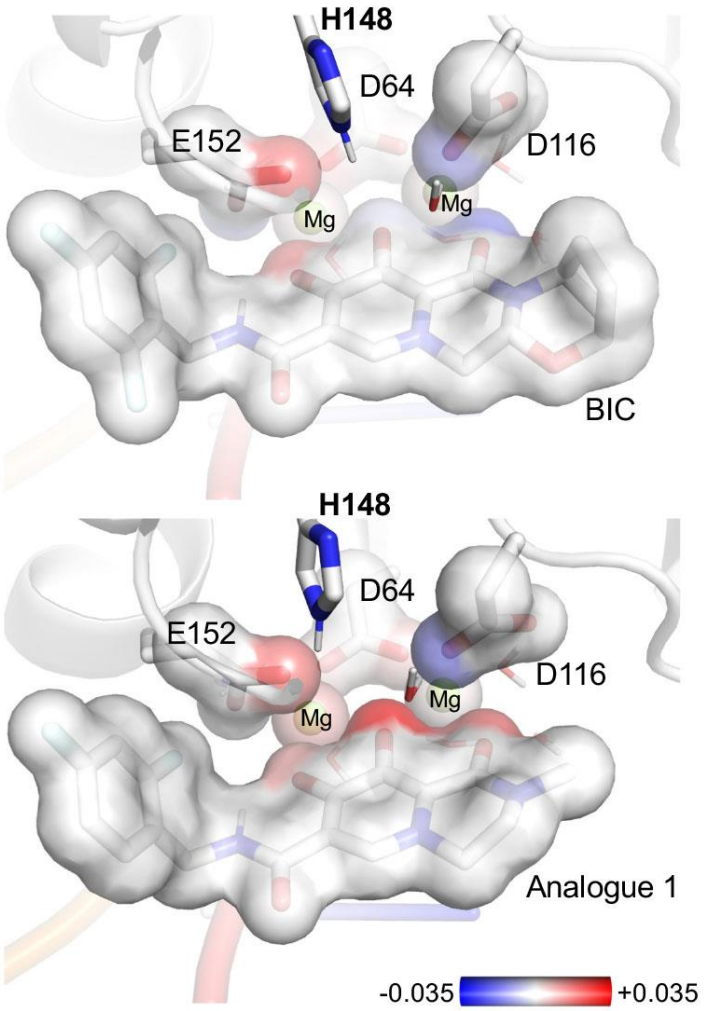


Figure 4.7. Active site polarizability changes due to Q148H/G140S substitutions for BIC (top) and analogue 1 (bottom). Natural bond orbital analysis results are shown for the active site Mg2+-ligand cluster. Protein residues, bound ligands, and metal ions from QM/MM minimized structures are represented as sticks and semi-transparent space-fill spheres. Colours indicate changes in charge distributions between the Q148H/G140S mutant and the wild type. Note the increased change in polarization of the metal chelating atoms of analogue 1 due to the amino acid substitutions.

The interactions with Mg2+ ions, which are nearly covalent in nature, are partly responsible for the extraordinary tight binding of INSTIs. Our results reveal that the chink in the armor of this drug class, exploited by the virus, is the extreme sensitivity of metal ions for the precise geometry and electronic properties of the ligand cluster [123], [124]. Each DNA-bound IN active site within the intasome catalyzes just one strand transfer event, allowing the virus to balance

INSTI resistance by detuning its active site while retaining sufficient replication capacity. However, extending the small molecules toward the IN backbone helps to stabilize optimal binding geometry and improve the resilience of the drug in the face of INSTI resistance mutations. We note that although DTG and BIC maximally extend to the β4-α2 connector, they leave substantial free space in the IN active site, which is occupied by solute molecules in our structures. Extension of the INSTI scaffolds to fill this space should be explored for the development of improved compounds.

4.4 Method

4.4.1 Molecular Dynamics

MD simulations were performed for four systems: the wild type and Q148H/G140S SIVrcm intasome bound to two types of ligands: BIC and analog 1. The initial models were assembled based on the experimental structures determined in this work, using N-terminal acetyl and C-terminal amide capping groups for missing residues. The ligands were parametrized using the general Amber force field (GAFF)[18]. The atomic partial charges were obtained from electrostatic potential calculations [125] at the level of density functional theory (DFT) ωB97X-D/def2TZVPP as implemented in Gaussian 09 Revision E [91], [126], [127]. The system was solvated by 100,000 -120,000 TIP3P water molecules depending on the ligand and mutations Na^+ and Cl^- ions were added to neutralize the system and set a salt concentration of 0.14 M.. From the original starting cryo-EM structure, water molecules directly involved in the metal coordination were retained. The system was minimized using a standard

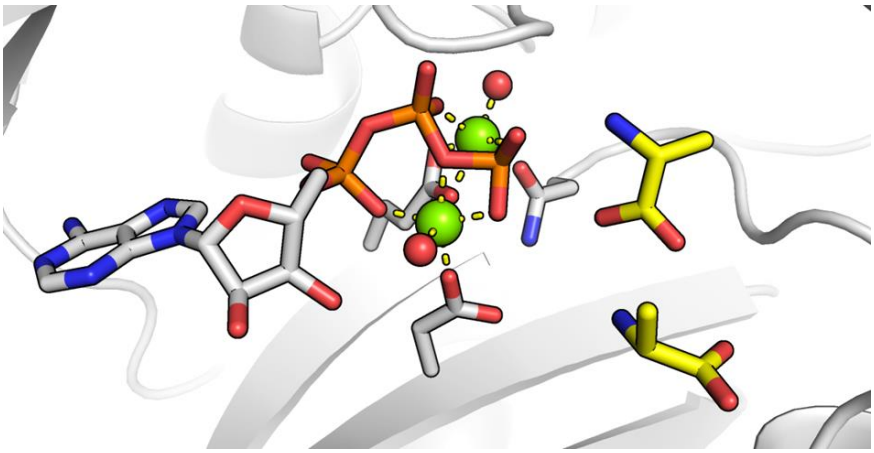
protocol via steepest descent algorithm for a total number of 150,000 steps, followed by 10ns equilibration with restrained heavy atoms (heavy atom of the backbone of the protein and the nucleic with an isotropic force of 1000 kJ mol-1nm-1) in constant number pressure and temperature (NPT) and constant number volume and temperature (NVT; up to 1 ns) at 300 K via standard MD procedure using NAMD 2.12 [92]. Pressure was maintained at 1 atm by a Nosé–Hoover Langevin piston [128]. Temperature was maintained at 298 K using Langevin dynamics with a damping coefficient γ of 0.5 ps^{-1} applied to all atoms. SHAKE[12] was applied to all bonds involving hydrogen and nonbonded interactions were calculated with a cutoff of 12 Å, and a switching distance of 10 Å. The particle mesh Ewald method was used for long-range electrostatic calculations with a grid density of >1 $Å^{-3}$ [129]. A series of unbiased MD simulations were performed for the four systems (wild type and G140S/Q148H IN bound with BIC or analog 1) to obtain multiple independent trajectories. A total of 10 replicas were run for all systems, and each simulation was 100 ns long.

4.4.2 Quantum Mechanics/Molecular Mechanics (QM/MM)

The experimentally resolved structures were subject to QM/MM minimizations to obtain geometry-optimized structures of the active site at the ab initio DFT level and assess the ligand-bound structures' quantitative response to the mutations. The MM region was described by CHARMM36 force field [130] while the hybrid functional B3LYP with the GTO basis set of 6-31G* [131] was utilized for the active site, as implemented in CHARMM [132] and Q-Chem 4.3 [10], respectively. QM/MM calculations were performed in a non-periodic fashion,

with atoms further than 25 Å from the ligand removed while those between 20 and 25Å were fixed. Population analysis was carried out for the QM region embedding the electrostatics of the MM atoms [133], using natural bond orbital (NBO) 3.1 scheme [134]–[136], as implemented in Gaussian 09 Revision E.

Chapter 5 Catalytic Mechanism and Druggability Study of D-ala-D-ala Ligase



5.1 Abstract

This work has been performed in collaboration with Dr Luiz Carvalho of the Francis Crick Institute. I present a combination of computational and experimental techniques, both done by me, where I analyse in detail the enzymatic mechanism of D-Ala-D-Ala Ligases in Mycobacterium tuberculosis (MtDdl). MtDdl is an essential enzyme for the cellular life of the bacteria, as it provides building block material for the cell wall. For this reason, it becomes essential to understand the mechanism of the enzyme to design new potent inhibitors, which can become a strong alternative drug against tuberculosis.

After modelling the protein in complex with its natural substrates, I performed QM/MM simulations to provide an atomistic insight behind the reaction mechanism that allows the formation of the peptide product through the ATP phosphorylation. The results suggest that the glutamate in position 239 is the most suitable base for the deprotonation of the second alanine. Those results have been confirmed through multiple "in vitro" experiments, including spectrophotometer and mass spectroscopy analysis. Thanks to the results obtained, I relied on the structure obtained from the QM/MM calculations to perform a virtual screening approach and suggest new molecules that inhibit the enzyme activity. The best candidates obtained from the virtual screening will be further analysed through molecular dynamic simulations and tested through "in vitro" spectrophotometric assay, to confirm the actual inhibition activity. It was discovered that four molecules present IC_{50} at the millimolar range, comparable with the known inhibitor D-cycloserine.

This work is an ongoing project. The next steps include more iteration for the QM/MM string simulations to obtain a better convergence, and more replica

and analysis from the MD simulations to obtain more statistical results from the virtual screening.

5.2 Introduction

Mycobacterium tuberculosis (Mt), the disease-causing agent of Tuberculosis (TB), remains a significant cause of mortality, particularly, childhood morbidity and mortality worldwide [137]. As for many bacterial pathogens, the cell wall of TB is an interesting and essential component of the organism, and, interfering with it can lead to the disruption and death of the TB cells [138].

D-Alanine-D-Alanine Ligase (Ddl) catalyses the ATP-driven reaction of ligation between two D-Alanine (D-Ala) molecules to form the D-Alanine-D-Alanine dipeptide [139]. This dipeptide is an essential building block of peptidoglycan that is the scaffold of the lipid-rich bacterial cell wall. Ddl belongs to the family of ATP-grasp enzymes, where two domains, α + β, *grasp* an ATP molecule, and through the phosphorylation of the nucleotide, it drives the energetically unfavourable formation of the peptide [140]. Because of its specificity, Ddl has always been considered as a target to stop bacterial infections, such as in *Mycobacterium tuberculosis*. D-cycloserine (DCS), was one of the first drug used for the treatment of Tuberculosis and included in the list of essential medicines for the World Health Organization (WHO) [141]. DCS is structurally similar to D-Ala; hence, its inhibition interferes with the natural substrate for Ddl [142], [143]. Allergic reactions, seizures, sleepiness are some of the important side effects of Cycloserine; addressing its usage only as second stage therapy or drug-resistant tuberculosis. Here the need to find a new and more efficient molecule that will

become an antibacterial agent. Several previous works have combined molecular modelling and biological study to find possibly inhibitors for Ddl, suggesting structures that recollect the enzyme's natural substrate or the intermediate product of the reaction [144]–[149].

Before addressing our work to develop a new inhibitor to this target; our attention is addressed to understand the enzymatic mechanism of Ddl.

Previous works suggested that the catalytic mechanism of Ddl consists of three stages [150]: An initial ATP-dependent phosphorylation of the first D-Ala to generate the acylphosphate intermediate forming a metastable enzyme intermediate complex, followed by the reaction between the second D-Ala and the acylphosphate forming a short-lived tetrahedral intermediate. Lastly, we have the release of the organic phosphate leading to the dipeptide product's formation (see Figure 5.1 for the reaction mechanism).

However, the full detailed mechanism of the peptide formation is still not fully characterised. An important question regarding the enzyme's catalytic reaction is which residue or residues acts as a base to deprotonate both the hydrogens of the second alanine and allows the formation of the tetrahedral intermediate. Up today, it is still unknown which base is involved in the deprotonation of the second D-Ala [139], [140]. In a work by Shi and co-worker [151], through mutagenesis work in *Escherichia Coli*, they excluded tyrosine 277 as the base.

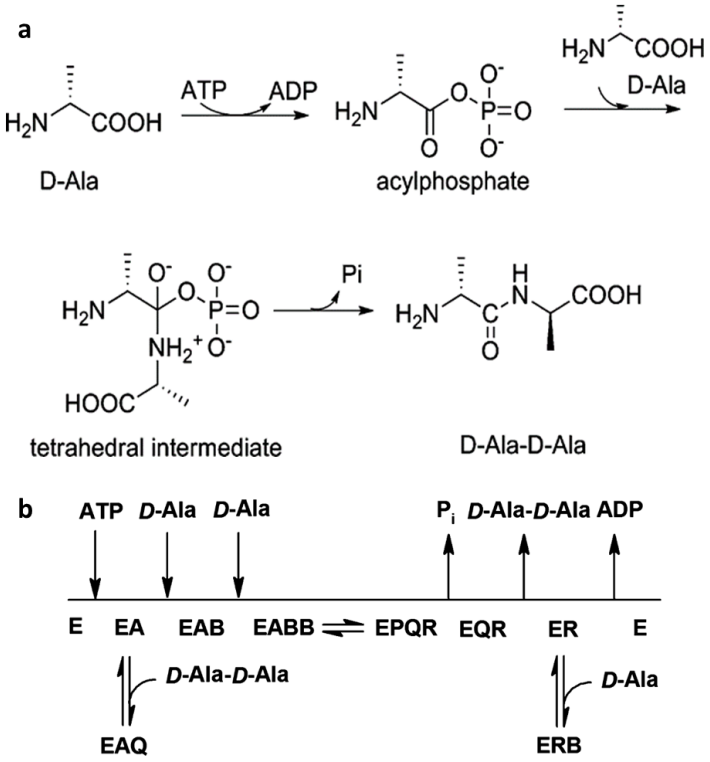


Figure 5.1. **a** catalytic and **b** kinetic mechanism of Ddl (From ref [143]).

Our homology modelling study found out that Glutamate 239 results in being very close to the second D-Ala and relatively conserved among DDL proteins. Part of my work is to provide a feasible mechanism of the catalytic activity of the enzyme. Our results suggest that glutamate 239 is the base required for the first deprotonation of the second alanine's amino group. Having a better idea on the mechanism of peptide formation, our next step was to provide experimental evidence of Glu239 being the base. This was achieved by comparing the wild type (wt) enzyme activity profile with three single-point mutants of MtDdl (E239Q, E239A, and Y277F). The enzymes' activity profile is tested through spectrophotometric and liquid chromatography-mass spectroscopy essays (LC-MS).

This result achieved by the QM/MM simulations and the experimental techniques, confirms our initial hypothesis of Glu239 being the base for the first deprotonation of the second D-Ala.

Having obtained a more exhaustive idea about the enzymatic mechanism of Ddl, we then presented a virtual screening approach to design novel inhibitors that target MtDdl. We performed a docking screening using the MolPort library, the results was manually checked and refined, and the molecules that present good docking results and relative stability along Molecular Dynamic (MD) simulations were tested enzymatically through spectrophotometer essay. To validate our method, we also tested enzymatically 40 random molecules from the Sigma-Aldrich library. From our work, we find that four molecules, two from the Molport and two from the Sigma library, presents an inhibition activity similar to DCS, making them interesting candidates for further drug development.

5.3 Methods

5.3.1 Initial Coordinates

Up to date, there are 46 Ddl crystal structures from 20 different organisms (enzyme commission number 6.3.2.4) deposited in the Protein Data Bank (PDB). Some of these structures contain the catalytic magnesium ions, ATP molecule or ADP + AlF_3, and inhibitors which mimicked the D-Alanine-D-Alanine dipeptide product. The only available crystal structure of *Mycobacterium*

tuberculosis DDL (PDB code: 3LWB) was chosen as a reference structure. It is an apo structure lacking 50 residues from chain A and 75 residues from chain B. By using multiple sequence alignments to identify enzymes belonging to the same enzyme class, structure 2ZDQ was chosen as the ATP and D-Alanine substrate template and 1EHI structure as the magnesium ions positional-template. Using molecular modelling of the structures mentioned before our model structure was built including the missing residues, magnesium ions, an ATP molecule and the two D-Alanine substrates, all necessary for enzyme activity. The M4T Server [152] was used for the standard 3D structure homology modelling.

5.3.2 QM/MM Simulation

Once we obtained the model, we use CHARMM-GUI [153] to parametrise the system, and we run 50 ns of equilibration and 100 ns of production trajectory using NAMD [92]. We used CHARMM36 [154] force field with periodic boundary condition and particle mesh Ewald [129] method for long-range electrostatic in combination with a 12 Å cutoff for the evaluation of the nonbonded interactions. The system was then trimmed to a sphere with 20 Å radius centred at the position of the first D-Ala. The QM/MM calculations were performed using Q-chem [10] coupled with the CHARMM [132] program. The QM region was treated with B3LYP [155] 6-31+G DFT level of theory, involving a total of 80 atoms, while for the MM region, we apply full electrostatic embedding [156]. Standard link atom treatment was used to connect the QM and MM regions. The QM/MM dynamics was performed using Langevin thermostat at 300K with 1 fs time step. For the QM region, we included: the ATP, only the atoms of the phosphate groups cut from C5' of the ribose, the two Mg^{2+}, the two alanine,

three water molecules and the terminal functional groups from the side chain of residues 239, 318, 330 and 332.

5.3.3 String Method Calculations

To determine the free energy path for the peptide formation, we implemented the finite-temperature string method [40]. As similarly described in Chapter 3, the method is defined by N windows that describe the reaction path, starting from ATP and the two alanine as a reactant, and arriving at the ADP + phosphate and the peptide as a product.

Distance	Atom 1	Atom 2
d1	ATP1 1501 PG	ATP1 1501 O3B
d2	ATP1 1501 PG	DAL1 401 OT1
d3	DAL2 402 N	DAL2 402 HT2
d4	DAL2 402 HT2	PROA 239 OE1
d5	DAL1 401 C	DAL1 401 OT1
d6	DAL1 401 C	DAL2 402 N
d7	DAL2 402 N	DAL2 402 HT3
d8	DAL2 402 HT3	WATA 75 OH2
d9	WATA 75 OH2	WATA 75 H2
d10	ATP1 1501 O3G	WATA 75 H2
d11	MG1 783 MG	ATP1 1501 O2G
d12	MG2 784 MG	ATP1 1501 O2A
d13	MG2 784 MG	ATP1 1501 O3B
d14	MG2 784 MG	ATP1 1501 O3G
d15	DAL2 402 HT3	DAL1 401 OT2
d16	DAL2 402 HT1	DAL1 401 OT2
d17	WATA 75 OH2	DAL2 402 HT1
d18	DAL2 402 N	WATA 75 OH2
d19	DAL2 402 N	DAL1 401 OT2
d20	DAL2 402 N	PROA 239 OE1

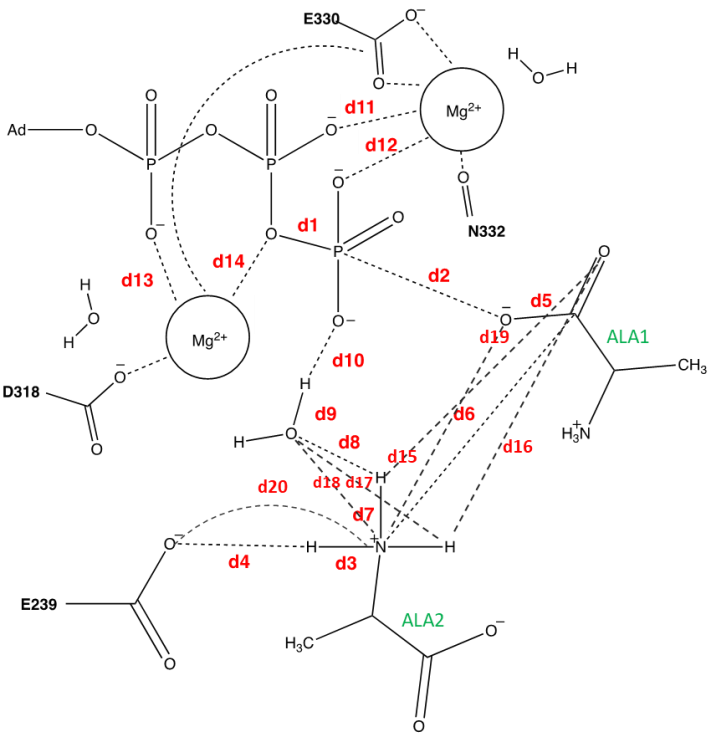


Figure 5.2. List and schematic representation of the reaction coordinate used to define the multidimensional space.

For each window, a set of collective variable, represented as interatomic distance, is defined, and for each of them, a harmonic potential is applied using

a force constant of 150 kcal/mol. This defines a multidimensional space where, for each window, each distance is restrained to a specific position. We then run 100 fs long QM/MM simulations for each of the 52 images. At the end of each iteration, a high-order polynomial (10) is applied to the average of each image's collective coordinates, and new positional restrain for the next iteration is generated. Between string simulations and updating the string values, the iteration is done till the variation of all collective coordinates fell below a given threshold. In this work, we used 20 distances to define the reaction path (see Figure 5.2). The results are then unbiased using WHAM with a convergence threshold of 0.001 kcal/mol to obtain the reaction's free energy profile.

5.3.4 Docking

The docking is carried out using Glide [157]. The last frame from our previous MD simulation, representing a minimised and dynamically stable system, is chosen as a starting structure for the docking calculations, including ATP and the metal ions. The docking grid area is defined as the region within 15 Å from the centre of mass of the first D-Ala. For the docking calculation, we included both the ATP and the two Mg^{2+} because we wanted to design novel drug molecules that directly coordinate the Mg^{2+} ions, analogously to the clinically used HIV IN drugs. The virtual screening was performed using the entire MolPort library (updated up to July 2018); first, we generate the 3D structures with their relative conformers of 7.5 millions of compounds using LigPrep obtaining a final library of 17 million compounds. For the docking, we used different precision levels integrated into the pipeline (Glide, with HTVS, SP and XP precision). We then parametrise and run long MD simulations of the

molecules showing the highest Glide score. The simulations aim to confirm that the selected molecules stay inside the pocket. MD simulations were performed using Desmond [158] software and OPLS3 [159] force field. Simulations have been analysed using the obtained trajectories, assessing both the ligand stability and the overall changes in the protein conformation, along the time. Four our validation, we performed the after-mentioned docking and MD simulation approach also with a randomly chosen molecule from the Sigma-Aldrich Cambridge library. We then choose the best 32 molecules from the MolPort library and the 40 molecules for Sigma-Aldrich for the enzymatic activity test.

5.3.5 Enzymatic Activity

WtDdl, E239Q, E239A, and Y277F Ddl were previously expressed and purified by a member of Carvalho laboratory. For each protein's enzymatic activity, the velocities are monitored continuously using UV-vis spectrometer (Shimadzu UV-2550 UV-Vis spectrophotometer, Milton Keynes, UK). The activity is measured by coupling our interested protein with pyruvate kinase and lactate dehydrogenase enzymes. Using this coupled enzyme, we can directly associate the peptide formation, from the ATP's cleavage to ADP, the last being one of the pyruvate kinases reaction substrates. Pyruvate, the product obtained from the pyruvate kinase, becomes the substrate of lactate dehydrogenase, oxidating NADH to NAD^+ (ε_{340} = 6220 M^{-1} cm^{-1}), (see Figure 5.3 for the full reaction mechanism).

All reactions are performed in 50 mM of HEPES buffer (pH 7.3), 10 mM MgCl2, 80 KCL, 0.2 mM NADH, 2mM of phosphoenolpyruvate at 37 °C at various

substrate concentrations of D-Ala (2-30 mM) and ATP (0.12-3mM). The reaction is initiated with the addition of enzyme at a range of 0.05-1 μM for wt and mutants respectively. The change in absorbance is recorded, and the initial linear rates are taken. These initial rates are then fitted against equation 1 to obtain the k_{cat} and the K_m for both substrates.

$$v = \frac{k_{cat}\,[ATP][Ala]}{K_{iATP}K_{Ala2} + K_{Ala2}[ATP] + K_{ATP}[Ala] + [ATP][Ala]} \quad (5.1)$$

Where K_{iATP} represents the inhibition constant of ATP, K_{Ala2} and K_{ATP} correspond to the Michaelis constant for D-Ala and ATP respectively.

For inhibitor screening, the same assay conditions were used as described before with some modifications. For this experiment, the same concentration of buffer was used, with saturating concentrations of substrates 3.6 mM of D-Ala, 3 mM of ATP in the presence of 1 mM of the inhibitor.

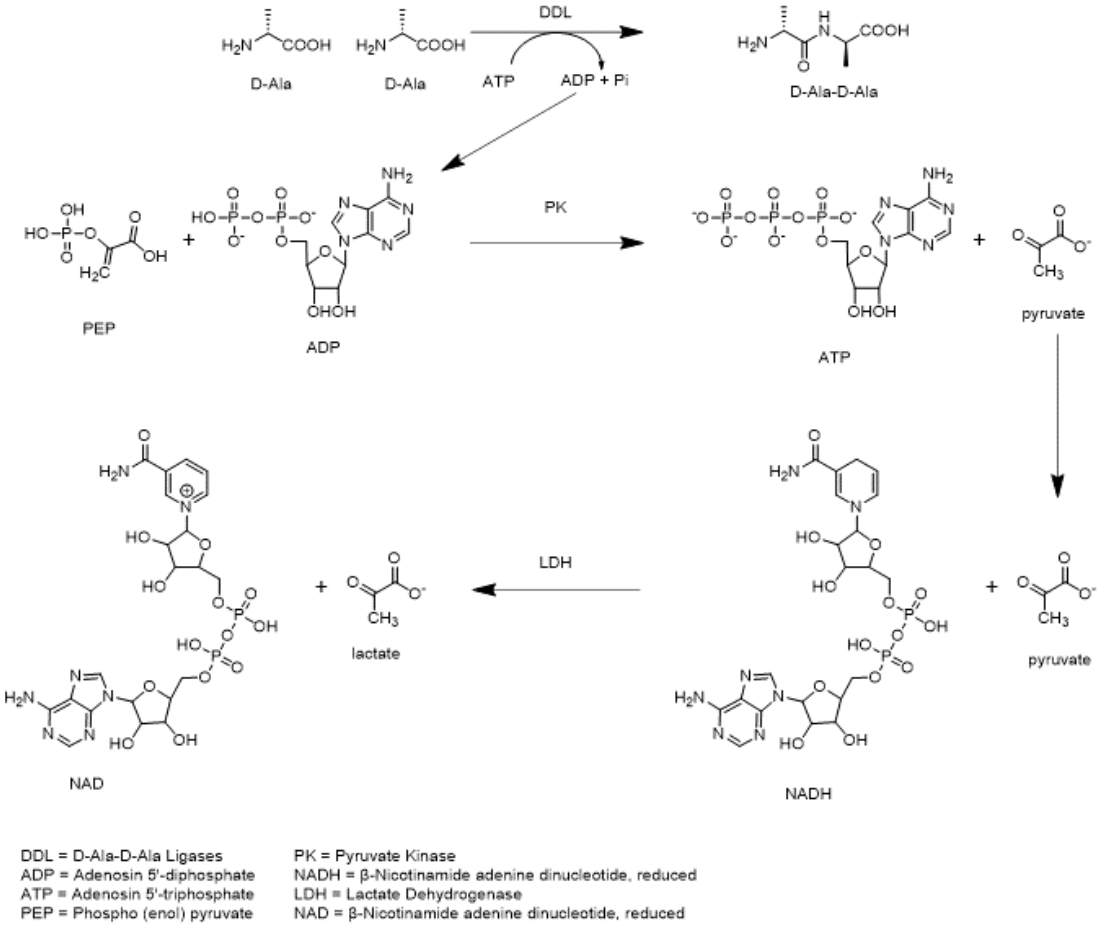


Figure 5.3. Reaction mechanism showing the coupled essay of PK/LDH involving PEP and NADH. The ADP produced from the initial reaction of DDL, becomes the substrate for the Pyruvate Kinase enzyme, converting Phospho (enol) pyruvate to produce ATP and Pyruvate. The conversion of pyruvate to lactate by lactate dehydrogenase (LDH). This step requires NADH which is oxidized to NAD+.

5.3.6 LC-MS

Chromatography is performed using a Cogent Diamond Hydride Type C silica column, 150 mm × 2.1 mm, at ambient temperature. Mobile phase A was 0.1% (v/v) formic acid in water. Mobile phase B was 0.1% (v/v) formic acid in acetonitrile. Analytes are eluted using a flow rate of 0.4 mL/min and the

following mobile phase gradient: 0 – 5 min, 85 – 70% B; 5 – 9 min, 70 – 5% B; 9 – 9.1 min, 5 – 85% B; The system is re-equilibrated to initial conditions for 4 minutes at the end of each run. The injection volume was 2 µL.

The ToF is operated in positive-ion mode with electrospray ionisation (ESI) using a dual AJS ESI source. Capillary, nozzle and fragmentor voltages were set at 3500 V, 2000 V and 110 V respectively. The nebuliser pressure is set at 35 psi, and the nitrogen drying gas flow rate is set at 13 L/min. The drying gas temperature is maintained at 250 °C. The sheath gas temperature and flow rate are at 350 °C and 12 L/min. Data are collected in the m/z range 50 – 1200 and saved in centroid mode. Dynamic mass axis calibration is achieved by continuous infusion of a reference mass solution, which enabled accurate mass spectral measurements with an error of less than five parts-per-million (ppm). The calibration curve is established over the D-Ala-D-Ala peptide concentration at 0, 0.4, 0.8, 1.6, 3.2, 4.8, 6.4, 8, 10, 12 mM.

The assay directly measures peptide formation at different time step for the wt and the three mutants. We kept the same concentration of the buffer as for the enzymatic assay (50 mM of HEPES, 80 mM of KCl, 10 mM of $MgCl_2$), and adding 3.6 mM of D-Ala and 3.1 mM of ATP. The reaction is carried at 37 °C; it will start with the addition of the enzyme, and it stops by quenching the sample with a solution of acetonitrile and formic acid (2%), 1:4 ratio. For each of the four protein (wt and the three mutants), we stop the reaction and detected the peptide concentration at the following times: 10-30-60-90-120-150-180-240 seconds.

5.4 Results and Discussion

5.4.1 The Catalytic Mechanism of The Peptide Formation

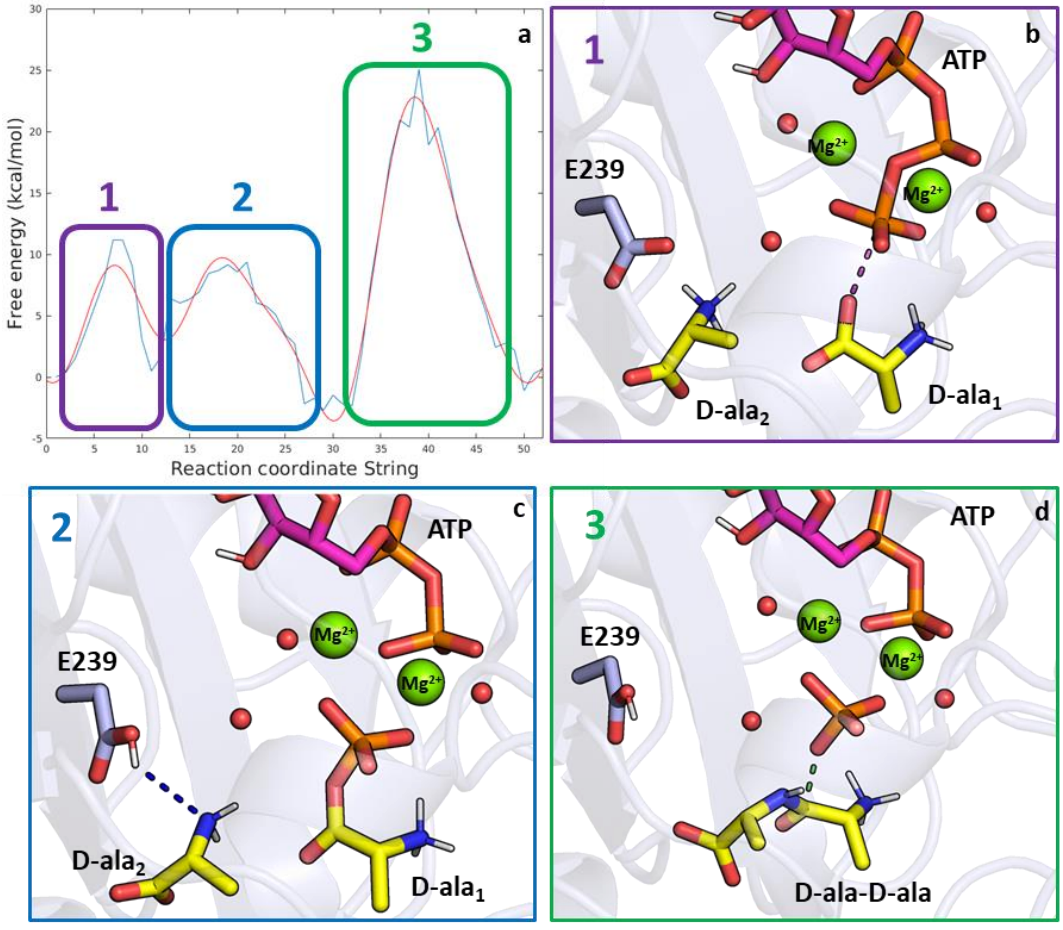


Figure 5.4. **a** Free energy profile projected onto the coordinate reaction windows, highlighted the three transition states observed from the QM/MM calculations, and the representative coordinate representation in **b**, **c**, and **d**. The first transition state (in purple) corresponds to the phosphorylation of the first D-Ala, the second transition state (blue) correspond to the deprotonation

of the second D-Ala and the third transition state is related to the peptide formation and release of the phosphate group.

A sampling of the conformational states through constrained minimisation scanning has been achieved through our QM/MM simulations. This scan provides us with the coordinate at a different stage of our reaction path. We then built the reaction path for the finite-temperature string method by dividing the reaction path into 52 windows. We then run 1ps long QM/MM simulations for each window for a total of 40 iterations. A histogram-free implementation, using the multidimensional WHAM method, was used from the combined data of the replica-exchange string simulations. The free energy profile, shown in Figure 5.4a, define the peptide formation process. According to the number of transition state barriers, the reaction mechanism can be divided into three steps:

1. Formation of the acylphosphate: The initial step is associated with the cleavage and transfer of the P_γ from the ATP to the first alanine (D-Ala_1), this is facilitated by the presence of the Mg^{2+} and water molecules (see Figure 5.4a). Here the carboxylate of D-Ala_1 attacks the γ-phosphate of ATP, forming the acylphosphate intermediate.
2. Deprotonation of the second D-Ala_2: The primary amine of the second alanine, needs to be in its deprotonated form to be able to attack the first alanine. In this work, we suggest that is glutamate 239 (Glu239) that steal the hydrogen from the amino group of D-Ala_2 (see Figure 5.4b). Our QM/MM calculations provide a good explanation for choosing this residue as the base for the deprotonation; this will also be confirmed by our mutagenesis study (see Chapter 5.4.2).
3. Peptide formation: The second deprotonation of the amino group of the D-Ala_2 allows the nitrogen to attack the phosphorylated carbonyl carbon of D-Ala_1, resulting first in the formation of a tetrahedral intermediate,

and then to the formation of the D-Ala-D-Ala peptide by cleavage of inorganic phosphate (see Figure 5.4d). This work suggests that a water molecule coordinated by the phosphate group is the proton acceptor. This step results in being the rate-limiting step of our free energy profile, showing an energy barrier of ~24 kcal/mol (Figure 5.4a).

Alternatively, for the second deprotonation, an alternative mechanism involves the same Glu239 acting as a base. However, the already protonated glutamate needs first to be deprotonated by a basic neighbour, to be able to attack the second hydrogen of the amino group. This mechanism, also known as a proton shuttle, has already been seen in different proteins such as serine proteases [160]. This part of the work is partially completed; as we are still running more iteration of the QM/MM string simulations in order to obtain a better convergence of the free energy profile.

5.4.2 Mutagenesis Study

Here I report the kinetic studies of Ddl with its native substrates. The results are obtained from the wt protein and the three single-point mutants E239Q, E239A and Y277F. For mutants, E239A and Y277F, no activity was observed at saturating substrate concentration. The E239Q did show some activity.

Because of the lack of saturation by D-Ala, as shown in Figure 5, it was not possible to measure V_{max} and K_m for E239A and Y277F. As expected, the wt protein shows the highest activity rate, compared with the other three mutations. This data demonstrates that all the mutations affect the enzyme

activity consistently; however, while E239A and Y277F show almost no activity at this condition, E239Q presents a slow enzymatic event.

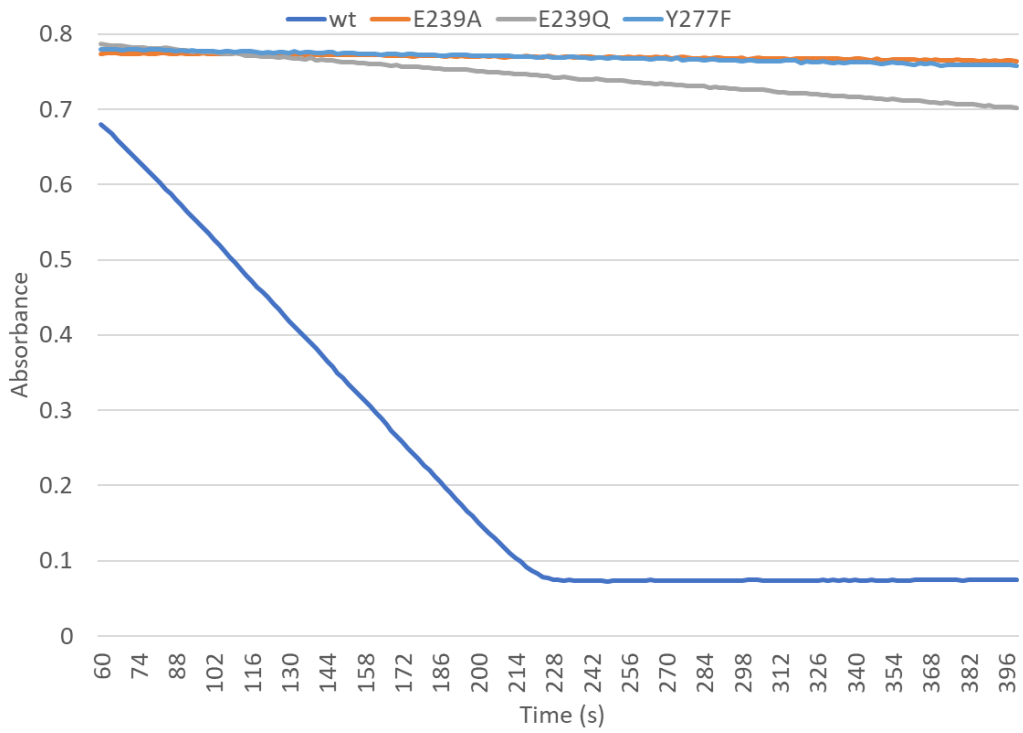


Figure 5.5. NADH's real-time absorption profile emission associated with the activity of the enzyme; wt: blue, E239A: orange, E239Q: grey, and Y277F: light blue.

The kinetic analysis was carried out, and K_m and k_{cat} (V_{max}/[Enzime]) were calculated as described previously in the method session 5.3.5 by testing the enzymatic activity at varying substrate concentrations of D-Ala and saturating ATP. Steady-state kinetic parameters for the wt enzyme are in good agreement with a previous work done by Prosser et al. [143], with the K_m for D-Ala at 8.8 mM & k_{cat} of 4.13 s^{-1}. We were also able to calculate the steady-state kinetic parameters for E239Q mutant, obtaining a K_m of 0.79 mM for D-Ala and a k_{cat} of 0.06 s^{-1}.

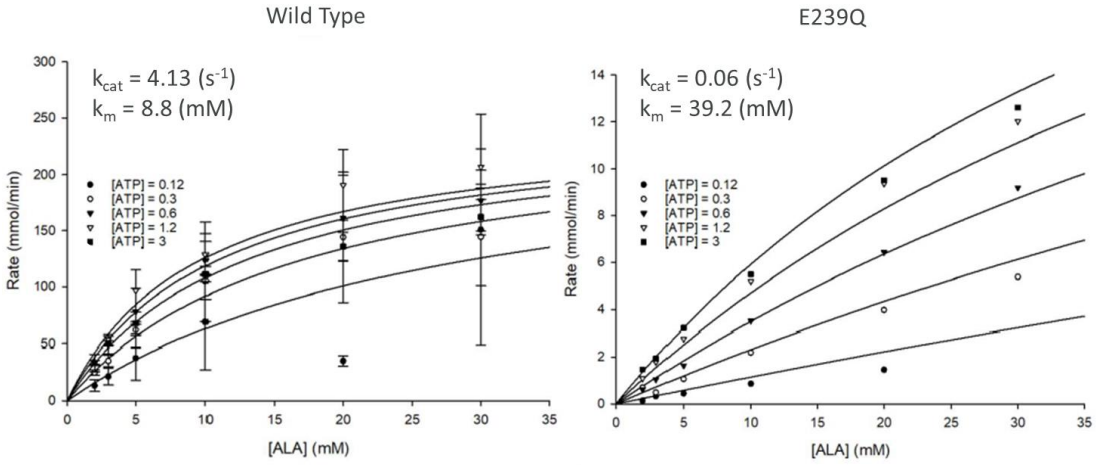


Figure 5.6 Steady-state activity of wt-MtDdl (left) and E239Q-MtDdl (right). Initial rates at varying concentrations of D-Ala (≤ 30 mM) and several fixed saturating concentrations of ATP: 0.12 mM (closed square), 1.2 mM (open triangle), 0.6 mM (closed triangle), 0.3 mM (open circle) and 0.12 mM (closed circle). For the wild type essay, the full experiment is performed in triplicate, and the standard error is plotted.

These results provide strong evidence of Glu239 being the base, as its substitution to glutamine results in a drastic reduction of activity (about ~100 fold), and a complete loss of activity when mutated to alanine. This can be explained as when the basic glutamate is replaced by glutamine, the glutamine's oxygen can still act as a weak proton acceptor, allowing the reaction to proceed, while a substitution to alanine, stop the activity of the enzyme. Furthermore, the data suggest that the mutation of E239 does not affect the D-Ala's affinity to the binding site, as the difference between the two K_m is not significant; but the k_{cat}, representing the velocity of the reaction is two orders of magnitude smaller in the mutant. Tyrosine 277, is another residue highly conserved in Ddl, but this residue was already demonstrated not to be the base [151]; however, the mutation of this protein, stop the activity of the enzyme, suggesting an important effect on the activity of the enzyme.

Using the LC-MS, we also checked the enzyme activity by detecting the product's concentration (D-Ala-D-Ala) at different reaction times. This experiment allows us to directly correlate the enzyme activity and the product formation, while by using the spectrophotometer, this was done indirectly by measuring the ADP formation and subsequently monitoring the reduction of NADH.

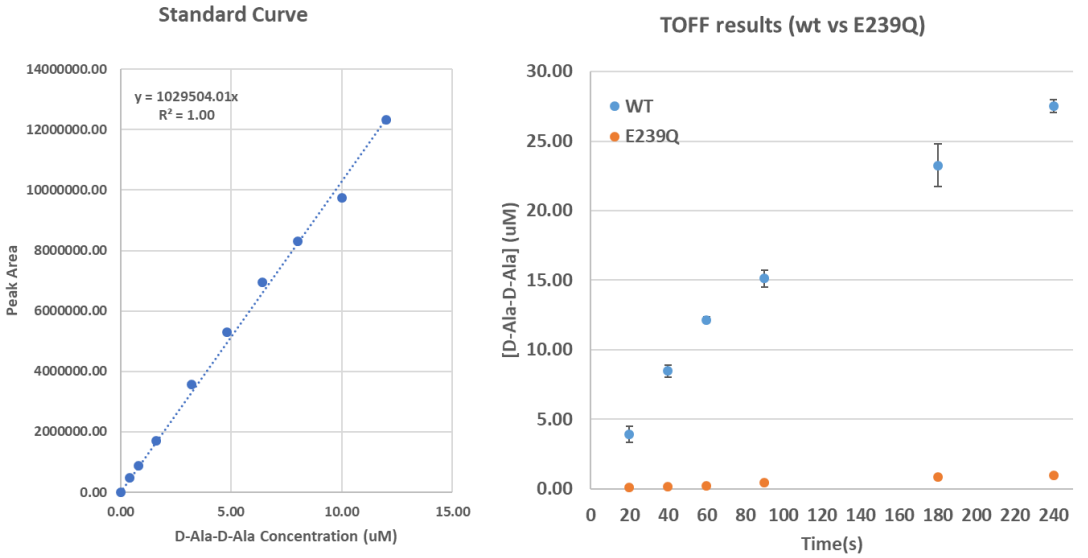


Figure 5.7. A calibration curve (left plot) for the LCMS essay. Stoichiometry of D-Ala-D-Ala along with the reaction time performed using the LCMS for the wild type (blue dots) and E239Q (orange dots).

Figure 5.7 confirms our previous results, showing a significant peptide formation with the wt while for E239Q, the formation of D-Ala-D-Ala is minimal.

5.4.3 Virtual Screening

For the docking calculations, I used the energy minimised initial structure of Ddl, in the bound ternary complex, where the enzyme is bound to the ATP and

the two D-Alas. The docking grid map is built with the ATP and the two Mg^{2+} included in our virtual screening, and we define the region for the docking sample in the same area of the two D-Ala. The pocket definition in this way is because the ATP, and therefore the two Mg2+, are the first substrate to bind and the last to leave (in the form of ADP + P_i) (see Figure 5.1) [161]. Docking results were obtained and ranked accordingly to the Glide scoring function. From the entire MolPort library, we obtained 1,627 compounds with good binding affinity, which we clustered using interaction fingerprint and classified using both the glide docking-score and ligand-target geometries and interactions (including solvent effects). Out of these 1,626 compounds, we selected the best 40 molecules and performed 150 ns long MD simulations for each ligand-protein complex to assess the binding pose's stability. We see that eight compounds leave the pocket from these MD simulations, suggesting that they might not be good candidates. We then purchased the remaining stable 32 compounds, and we perform the enzymatic inhibition assay (Figure 5.8).

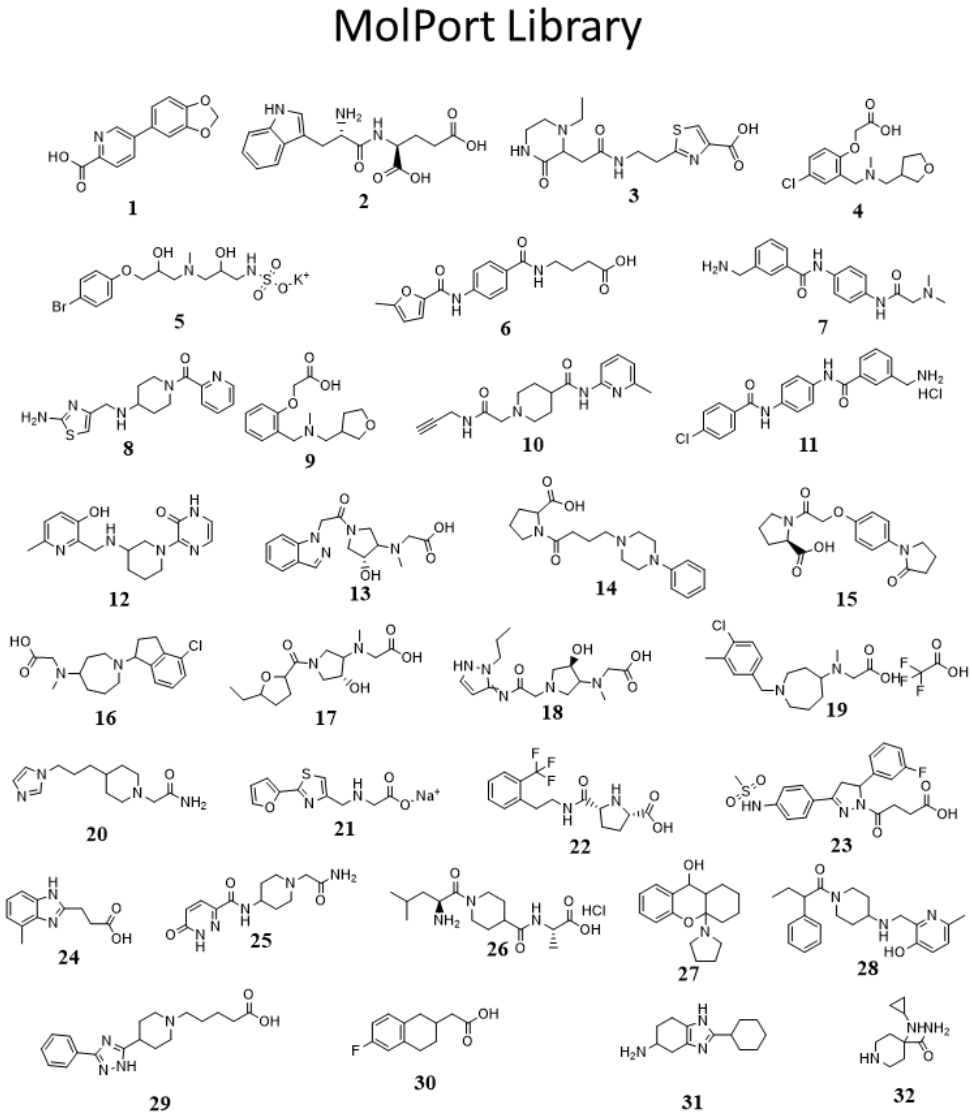


Figure 5.8. Chemical representation of the compound used from the MolPort library

From a preliminary analysis, we can see that a good number of molecules with a high docking score share structural similarity to the natural substrate of the enzyme or intermediate of the enzymatic reaction. From these compounds, we can see that 19 out of the 32 molecules present a carboxylic group, recollecting the carboxylic group of the D-Ala's natural reactant substrate (Figure 5.8). Furthermore, compound **2, 3, 6, 7, 10, 11, 15, 17, 18, 25, 26,** and **28** present an amide group, similar to product the D-Ala-D-Ala. Additionally, the presence of

negatively charged functional groups, such as the carboxylic group, is also explained by the presence of the Mg^{2+} ions; hence most of the docking pose involves a direct interaction of the molecule to either one or both the magnesium cations.

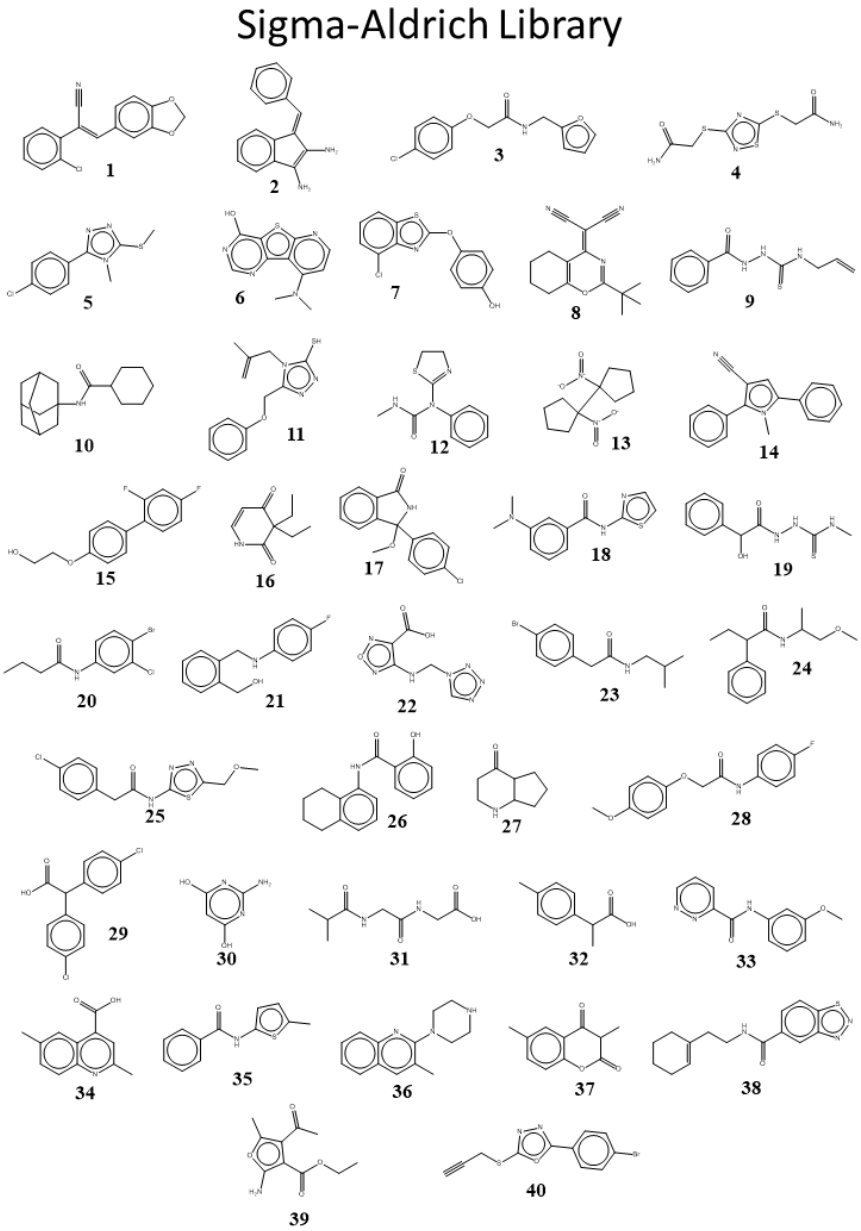


Figure 5.9. Chemical representation of the compound used from the Sigma-Aldrich

Additionally, to validate our method; we purchase 40 compounds from the Sigma-Aldrich Cambridge library. The molecules were randomly selected from the entire library and tested in vitro to validate our docking methodology. To these 40 compounds (Figure 5.9) we also performed a docking calculation to predict the binding poses, and we assessed the stability of these compound by

running 150 ns long MD simulations, using the same procedure as for the MolPort library. In this case, only four compounds (Sigma Aldrich: **1, 22, 25, 26**) remain inside of the active site.

5.4.4 Enzymatic Inhibition

The top-ranked compounds found from our docking calculations using the MolPort library and the 40 randomly selected molecules from the Sigma-Aldrich library, were selected for an in vitro evaluation. We use the product inhibition essay to assess the order of MtDdl binding and the product release following the same procedure as for the enzymatic assay. To assess the inhibition activity for each compound, first, we compare the activity rate of the enzyme with its natural substrates, from our previous experiments, with the enzyme in complex with its natural substrates and DCS, a known inhibitor for MtDdl. Out of the 32 molecules from the MolPort library, only 11 shows an inhibition similar to DCS, and similarly, from the 40 molecules of the Sigma Aldrich library, 12 shows inhibition.

MolPort		
Compound	Activity Rate	IC_{50} (mM)
Mol_1	0.129 ±0.024	
Mol_6	0.12 ±0.012	3.41
Mol_7	0.145 ±0.012	
Mol_8	0.133 ±0.009	
Mol_11	0.087 ±0.027	1.28
Mol_12	0.125 ±0.013	
Mol_13	0.127 ±0.024	
Mol_17	0.141 ±0.004	
Mol_23	0.133 ±0.008	
Mol_28	0.159 ±0.011	
Mol_30	0.147 ±0.016	

Sigma-Aldrich		
Compound	Activity Rate	IC_{50} (mM)
Sig_1	0.127 ±0.008	
Sig_2	0.077 ±0.02	1.99
Sig_6	0.117 ±0.02	
Sig_14	0.12 ±0.002	5.16
Sig_18	0.146 ±0.018	
Sig_20	0.148 ±0.005	
Sig_22	0.144 ±0.017	
Sig_25	0.121 ±0.011	
Sig_26	0.132 ±0.017	
Sig_35	0.142 ±0.009	
Sig_36	0.146 ±0.007	
Sig_40	0.145 ±0.015	

Compound	Activity Rate	IC_{50} (mM)
DCS	0.098 ±0.1	0.77
No inhibit.	0.167 ±0.012	

Figure 5.10. Ddl inhibition activity of the best docking results obtained for the MolPort and Sigma-Aldrich library. For the molecules presenting high activity (activity rate ≤ 0.12), we provide IC_{50} concentrations through initial velocity pattern analysis, using multiple concentration of the inhibitor (0.1 mM, 0.5 mM, 1 mM, 1.5 mM, 2 mM and 5 mM).

For the four compounds with the highest inhibition profile, we were also able to perform initial velocity pattern analysis, using multiple concentration of the inhibitor and calculate the IC_{50}. These four compounds were found to inhibit MtDdl with IC_{50} value at the millimolar range, with the same order of magnitude of DCS (Figure 5.11). Compound **6** and **11** from the MolPort library share a similar structure, with two amide group in both the molecules and by looking at the docking pose, they also show similar orientation in the active site. However, the two compounds from the Sigma-Aldrich library differs significantly from the natural substrates or product for this enzyme. Additionally, the MD simulations show that both the molecules results to be unstable in the binding pocket, suggesting that either the binding pose predicted by the docking calculations is not the optimal one, either that the

inhibition is not done at the orthosteric site but interacting with the protein with an important allosteric site (Figure 5.11).

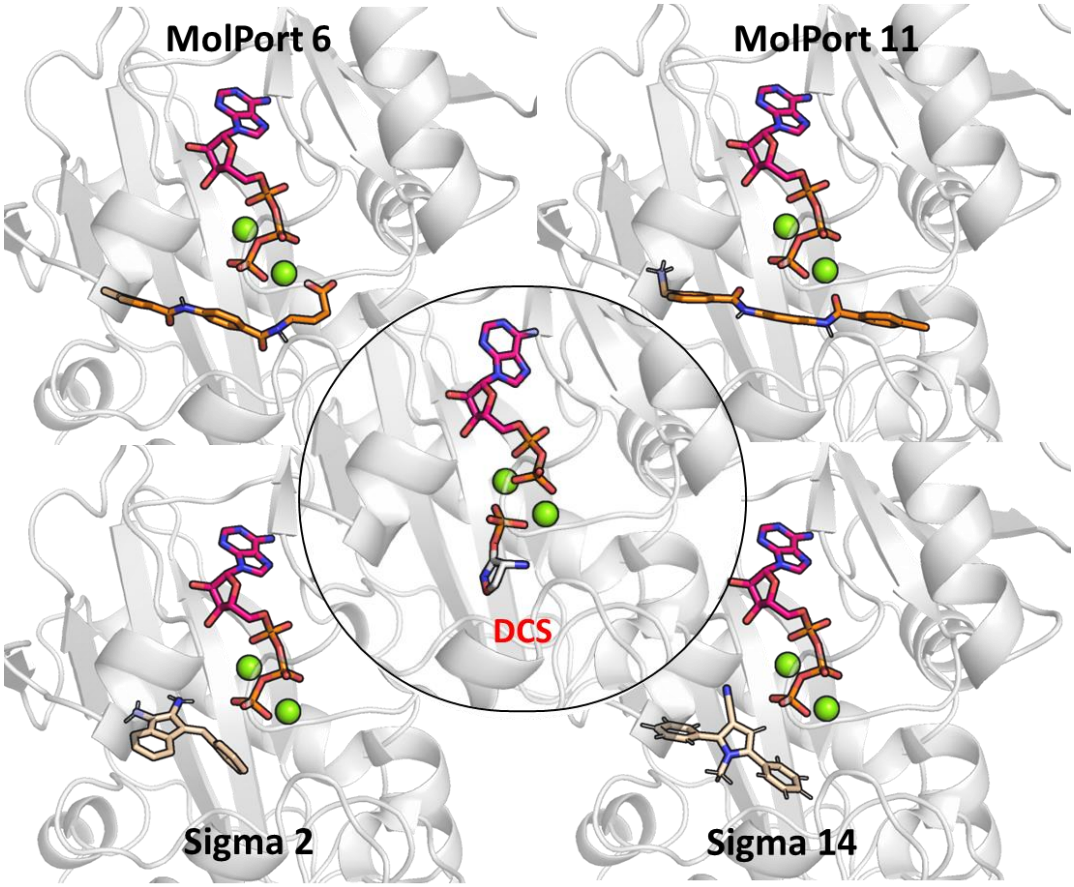


Figure 5.11: Graphical representation of the docking results for the four compound with inhibition activity, MolPort 6: top-left, MolPort 11: top-right, Sigma 2: lower-left, and Sigma 14: lower-right. The carbon atoms of the inhibitors are coloured in orange for the Molport molecules and beige for the Sigma molecules, while the carbon atoms of the ATP is coloured in magenta. In

the centre of the figure a representation if the phosphorylated DCS bound to Ddl obtained from PDB: 4C5A.

5.5 Conclusion

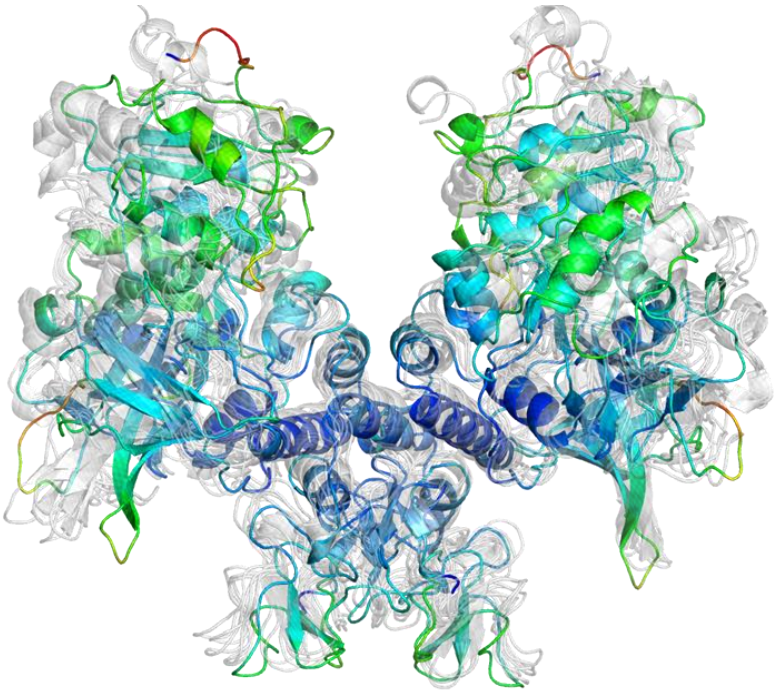
In this work, I presented a novel mechanism on the enzymatic activity of Ddl. This enzyme is an important protein in different bacteria, as its activity in forming the peptide D-Ala-D-Ala provides a critical scaffold for multiple bacterial cell-walls. My attention here is focused on the Ddl of *Mycobacterium tuberculosis*.

Despite the number of mechanistic studies regarding the catalytic activity of Ddl, not all the steps have been comprehensively addressed in terms of the specific details of their catalytic mechanism. An important question about this mechanism is which residue can act as a base for the deprotonation of the second alanine; to allows the peptide formation. Part of this work addressed this question. We demonstrated that the glutamate 239 is the first base that accepts the proton from the alanine and that a water molecule coordinated by the phosphate attached to the first alanine act as the base to extract the second hydrogen. We constructed a reaction path from the QM/MM calculations and calculated the free energy profile along the reaction path. Additionally, I performed single point mutation studies, to assess the protein's activity, when the interesting residues are mutated. Our conclusion, confirms the hypothesis of glutamate 239 being the base of the reaction.

Additionally, thanks to the results obtained from our QM/MM calculation, we assessed the important features that allow the peptide formation. We have conducted a virtual screening of two different libraries to identify novel

inhibitors for Ddl. Consequently, 46 molecules out of the 7.5 million compounds from the MolPort library and 40 randomly selected molecules from the Sigma-Aldrich library were tested for in vitro inhibition activity against MtDdl. For four molecules that show high inhibition activity, we calculated their relative IC_{50}. The results obtained showed for the last four molecules inhibition activity at the same order of magnitude to the known inhibitor D-cycloserine, providing an initial scaffold for the design of potent inhibitors for Tuberculosis treatment. Two of these inhibitors are structurally similar to the natural peptide product, constituting a promising starting point for further investigation.

Chapter 6 Modelling the Active SARS-Cov-2 Helicase Complex as a Basis for Structure-Based Inhibitor Design



6.1 Preface

The project started as a response to the biggest biological challenge of 2020, the Covid19 pandemic. The project aims to present a reliable structure of the SARS-Cov19 helicase protein in its holo form to provide key structural information of the catalytic site. Despite the helicase's fundamental activity in the viral transcription, we have very few information about the interaction between the protein and the natural substrates, RNA and ATP + Mg^{2+}. By performing long unbiased Molecular Dynamic simulations, we provided structural information of the helicase, providing key information to develop specific inhibitors to this target. In this work, my contributions were to set-up, perform and analyse all the simulations, interpret the results and write help draft the manuscript that is now published in Chemical Science:

Badaoui M., Berta D., Buigues P. J., Martino S. A., Pisliakov A. V., Elghobashi-Meinhardt N., Wells G., Harris S. A., Frezza E., Rosta E. (2020). Modelling the active SARS-CoV-2 helicase complex as a basis for structure-based inhibitor design.

6.2 Abstract

Having claimed over 1.7 million lives worldwide to date, the ongoing COVID-19 pandemic has created one of the biggest challenges to develop an effective drug to treat infected patients. Among all the proteins expressed by the virus, RNA helicase is an essential protein for viral replication, and it is highly conserved among the coronaviridae family. To date, there is no high-resolution structure of helicase bound with ATP and RNA. We present here structural insights and molecular dynamics (MD) simulation results of the SARS-CoV-2 RNA helicase both in its apo form and in complex with its natural substrates. Our structural information of the catalytically competent helicase complex provides valuable insights for the mechanism and function of this enzyme at the atomic level, a key to develop specific inhibitors for this potential COVID-19 drug target.

6.3 Introduction

Only a few approved drugs currently repurposed for treating COVID-19, a disease caused by the human coronavirus SARS-CoV-2, despite its close relatives, SARS-CoV and MERS-CoV are responsible for multiple outbreaks earlier this century. As of September 2020, Remdesivir and Dexamethasone are used in clinical practice [162], [163]. Therefore, drugs need to be developed that can be used against the viral replication to help patients overcome the disease in the most severe cases.

Here we focus on determining the catalytically active complex structures of the SARS-CoV-2 RNA helicase, labeled as Non-structural Protein (NSP) 13 (Figure 6.1). This protein is part of the Orf1ab polyprotein, which gets spliced to produce the enzymes required for viral replication. The RNA helicase performs two essential functions for viral replication making it an ideal drug target. It is thought to perform the first step in the 5’-capping of the viral RNA by its triphosphatase function hydrolyzing the 5'-triphosphate group to form diphosphate-RNA [164], [165]. Furthermore, its main helicase function is to enable RNA translocation and unwinding in an ATP-dependent mechanism during viral replication.

Accordingly, numerous studies have already demonstrated that it is possible to develop potent inhibitors of viral helicases as antiviral agents[166].

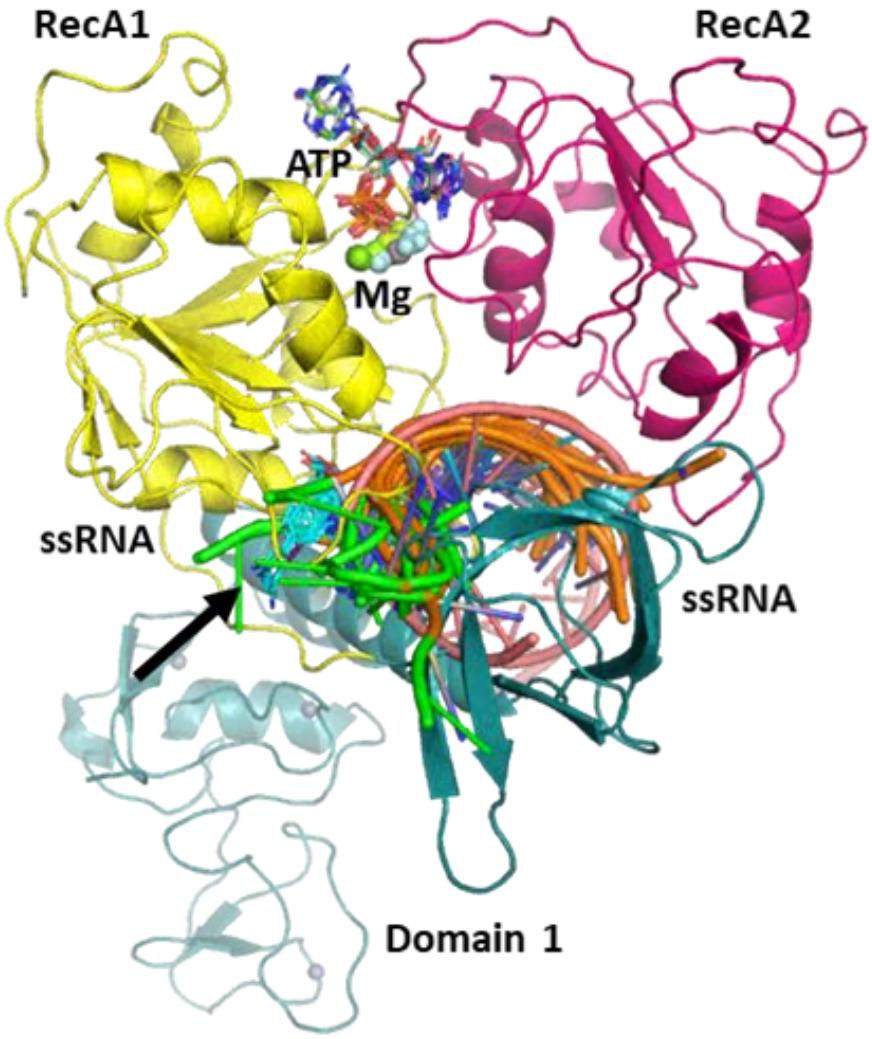


Figure 6.1. Cartoon representation of the RNA helicase NSP13 of SARS-CoV-2 monomer model composed of three domains: RecA1 (yellow), RecA2 (magenta), and Domain 1 (aquamarine). ATP analogues (sticks) along with Mg (green sphere) and single-stranded nucleic acids are depicted from aligned homologous structures. 3' ends of the nucleic acids present the same orientation in all chains (highlighted in green). Specific helicase inhibitor binding region with allosteric inhibitors displayed in cyan (black arrow).

Coronaviral RNA helicases share a high similarity. 600 out of the 601 residues of the SARS-CoV2 RNA helicase are identical to those of the SARS-CoV virus, and 70% match that of the MERS-CoV NSP13, demonstrating that these proteins are highly conserved within the coronaviridae family. Despite the importance of this target protein, currently, only the apo structure is available crystallographically. There are no structures currently known of the RNA

helicase bound with either inhibitors or nucleic acids, severely limiting the structure-based mechanistic understanding and the design of more potent drugs. Currently, the only experimentally available information on SARS-CoV-2 helicase comes from a recent study by Shi et al., that offer a low-resolution cryo-EM structure with ATP bound in the active site [167]. They fit the APO helicase's crystal structure from 6jyt to their cryo-EM density maps and refined using several software [168]. Recent works mainly focusing on the RNA-dependent RNA polymerase (RdRp) NSP12 [168], [169]. which is expressed in the polyprotein sequence just before the helicase, also yielded structures of the replication machinery, including low-resolution cryo-EM images of the helicase. Unfortunately, the level of resolution is too low in this structure to model the ATP pocket in a catalytically competent conformation.

Helicase Structures and Models

The recent July 2020 SARS-CoV-2 helicase structure (PDB ID 6zsl) and the almost identical SARS-CoV helicase structure from 2019 (PDB ID 6jyt)[170] were both resolved as crystallographic dimers (Figure 6.2a-b). Interestingly, the dimerization interface is different in the two cases, leading to structurally dissimilar complexes. In the cryo-EM structure of the RdRp complexed with the RNA helicase (and co-factors NSP7 and NSP8), the two helicase protomers are non-interacting (Figure 6.2c). Therefore, the catalytically active form of a monomer is of interest, and a dimer may not be the biologically functioning unit.

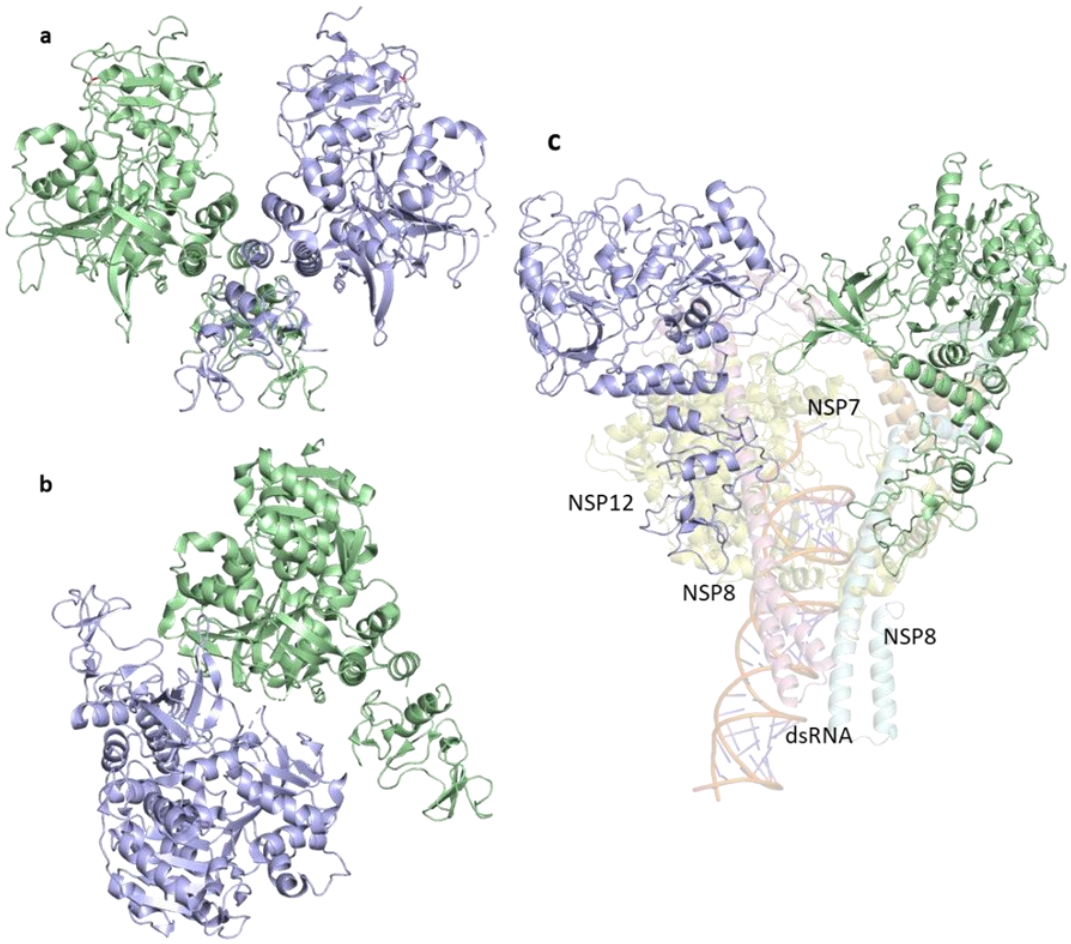


Figure 6.2. Structural comparison of the deposited PDB structures of the helicase dimer in SARS-CoV-1 (PDBID: 6jyt), SARS-CoV-2 (PDBID: 6zsl), and SARS-CoV-2 in complex with NSP7 NSP8 and NSP12 (PDBID: 6xez). The interaction between the two helicase monomers differs depending on the experimental method used to resolve the structures.

Here we present the first computational models of the SARS-CoV-2 RNA helicase with ATP and RNA substrates bound. We have performed sequence similarity searches to identify key domains and homologous sequences, suggesting structurally important conserved motifs. We also performed structural alignments of available homologous helicase crystal structures to help position the bound RNA and ATP substrates (Figure 6.1). A previous computational model by Hoffmann et al.[171] described the proposed helicase ATP interactions. However, there is no RNA incorporated in these modelled structures. Here, we also present long timescale MD simulations of both the apo

and ligand-bound states to address the flexibility and stability of our catalytically competent structures. Our results will help guide ongoing drug development with the identification of novel pockets.

6.4 Methods

6.4.1 Homologous Sequence Analysis

Sequence alignments were done using BLAST with default settings and BLOSUM62 distance matrix, requesting the most similar 1000 hits from UniProtKB [172]. Pairwise alignments for the obtained sequences were used to count identities and similarity for each residue in the SARS-CoV-2 helicase.

6.4.2 Homology Modelling

Proteins with crystal structures were aligned with mustang for a combined structural-sequence alignment[173]. The apo structure was based on PBD ID 6jyt [170]. Missing residues and the I570V mutagenesis were constructed in pymol. The position of the ATP was determined using the coordinates of PDB ID 2xzo [174], as a template, modifying residues around the ATP pocket, except for Arg443 which was modelled based on PBD ID 6jim [175]. The single-stranded RNA (ssRNA) was placed based on 2xzl [174].

6.4.3 Molecular Dynamics

The helicase model was used as a starting point for MD simulations. The system consists of the helicase, three Zn^{2+} ions, ATP, $Mg^{2+,}$ and ssRNA with eight uracil bases. The MD simulations were performed using NAMD 2.13 [176], using the CHARMM36 force field [177].

The system was solvated by 50,000 – 70,000 TIP3P water molecules resulting in a box of 120 Å per side. To neutralize the system and account for a 0.15 M KCl solution, we added 171 K^+ and 189 Cl^- ions [178]. Periodic boundary conditions (PBC) were used in all the simulations, and the particle mesh Ewald (PME) method was used for long-range electrostatic interactions. SHAKE algorithm was deployed to constraint the covalent bonds involving hydrogen atoms. A cutoff of 12 Å was used to treat non-bonding interactions.

The energy of the system was minimized using a standard protocol via steepest descent algorithm for a total number of 10,000 steps, followed by 50 ns equilibration with restrained heavy atoms (heavy atom of the backbone of the protein and the nucleic acid with an isotropic force of 1000 $kJmol^{-1}nm^{-1}$) in constant pressure and temperature (NPT) and constant volume and temperature (NVT; up to 1ns) at 303.15 K via standard MD procedure with a time step of 2 fs.

To help equilibrate the complexes, we used a harmonic constraint on selected contacts with a force constant of 10 kcal/mol for 15 ns in our preliminary MD simulations to maintain relevant contacts. These constraints were subsequently progressively reduced and removed during the next 20 ns, using the colvar function implemented in NAMD. For both the apo and the holo form, we performed five independent unbiased MD simulations, each 1 μs long, for a total of 5 μs simulations.

To compare simulation results obtained with MD, we also carried out MD simulations using GROMACS 2018 [179]–[182] with the Amber ff99+ parmbsc0+chioL3 force field [183], [184] for ssRNA and Amber14SB [185] for the helicase. To maintain the coordination of the Zn^{2+} ions, the ZAFF model was used[186]. The molecular systems were placed in a cubic box and solvated with TIP3P water molecules [178]. The distance between the solute and the box was set to at least 14 Å. The solute was neutralized with potassium cations and then K^+Cl^- ion pairs were added to reach the salt concentration of 0.15 M. We used the ion corrections of Joung et al.[187] as this force field has been shown to produce stable RNA structures [188]. The parameters for Mg^{2+} are taken from Ref. [189]. Long-range electrostatic interactions were treated using the particle mesh Ewald method[20], [129] with a real-space cut-off of 10 Å. The hydrogen bond lengths were restrained using P-LINCS [180], [190], allowing a time step of 2 fs [191]. Translational movement of the solute was removed every 1000 steps to avoid any kinetic energy build-up [192]. After energy minimization of the solvent and equilibration of the solvated system for 10 ns using a Berendsen thermostat (τ_T = 1 ps) and Berendsen pressure coupling (τ_P = 1 ps) [191], simulations were carried out in an NTP ensemble at a temperature of 300 K and a pressure of 1 bar using a Bussi velocity-rescaling thermostat[22] (τ_P = 1 ps) and a Parrinello-Rahman barostat (τ_P = 1 ps) [29]. During minimization and heating, all the heavy atoms of the solute were kept fixed using positional restraints. The restraints on the RNA and the protein backbone were relaxed slowly during the equilibration from 1000 $kJmol^{-1}$ nm^2 to 10 $kJmol^{-1}{\cdot}nm^2$.

Additional MD simulations, constructed using the AmberTools20 building package, were performed with the GPU version of Amber18 using the ff14SB force field to represent the protein [193], the ff99OL3 force field for the RNA [194], [195], ATP parameters from Meagher et al.[196] and parameters for Mg^{2+} are taken from Ref. [197]. The tetrahedral coordination state of the zinc was

maintained using the ZAFF bonded force field [197]. Note that additional parameters were required for the HIS-33 that interacted with the zinc via its epsilon nitrogen by reference to comparable parameters in the ZAFF using a hybrid of the centre ID 4 and 6 models [198]. For structures where ATP is bound, the octahedral coordination of the Mg^{2+} (which involves bonds to the ATP β and γ phosphate oxygen atoms, one with the oxygen of the Ser289 hydroxyl group and three structural water molecules) was constructed using the Chimera metal centre builder [199]. The solute was neutralized with potassium cations, then the protein was immersed in a box of TIP3P water molecules extending a minimum of 10 Å from the protein surface, and K^+Cl^- ion pairs were added to achieve a salt concentration of 0.14 M. MD simulations were performed in the NTP ensemble, with Berendsen temperature and pressure coupling. SHAKE was applied to all bonds involving hydrogen, allowing an MD integration timestep of 2 fs. Long-range electrostatic interactions were treated using the particle mesh Ewald method[20], [129] with a real-space cut-off of 12 Å. To equilibrate the protein and nucleo-protein complexes, the systems was initially energy minimized with positional restraints placed upon the solute, followed by minimization of both solvent and solute. The system was then heated to 300 K in the presence of positional restraints upon the solute, which were gradually reduced from 50 kcal/mol $Å^2$ to 1.0 kcal/mol $Å^2$ over a timescale of 100 ps. For the apo-helicase structure, all restraints were then removed. For the ATP-RNA helicase complex which included the coordinated Mg^{2+} ion, an additional 50 ns of equilibration was performed with harmonic distance restraints (set at 2.1 Å with a spring constant of 20 kcal/mol $Å^2$) to maintain the positions of coordinated atoms, and angle restraints imposing the octahedral geometry around the Mg^{2+} ion. An additional restraint was imposed to maintain the orientation of Asp374 and Glu375 to the adjacent coordinated water molecule, as observed in the MutS-ATP complex (PDB ID 1w7a,[200]). Three 1 μs

simulations of the apo-structure at a salt concentration of 140 mM, and one 1.5 μs simulation in neutralizing salt were performed. We have also obtained 1 μs simulations of the ATP-helicase (two replicas), the RNA-helicase and the ATP RNA-helicase complex. For all coordinated ATP Mg^{2+} metal centers, these equilibration protocols provide stable octahedral geometries, including the complexed water molecules, during unrestrained MD over 1 μs timescales.

Pocket analysis

For the analysis of the ATP pocket size, we used the open-source cavity detection software, Fpocket [201]. From our MD simulations of the apo and ATP-RNA helicase, we extracted 100 equally time-spaced protein structural snapshots from the last 200 ns simulation of both systems. We performed the pocket size analysis using several parameter sets using Fpocket and compared the results for our system. We then calculated the average volume of the ATP pocket for the most optimal parameter sets and compared these between the apo and the ATP-RNA complex simulations. The parameters we modified were: −m: minimum α-sphere size; -M: maximum α-sphere size -D: first clustering α-sphere size.

6.5 Results

6.5.1 Helicase Domains and their Sequence Homology

Following previous studies, the single-chain SARS-CoV-2 helicase can be divided into three domains, as depicted in Fig. 1. The sequence starts with

Domain 1 (residues 1-260), which features: a Zinc-binding domain (ZBD, residues 1-100), known to facilitate nucleic acid recognition; a Stalk region shaped by 2 adjacent alpha helices (residues 100-150) which functions as an interface connecting the ZBD with the rest of the Domain 1 (residues 150-260, also known as Domain 1B) that interacts with the RNA. The rest of the chain splits into RecA1 and RecA2 domains, which are well characterized in the superfamily 1B type helicases and bind ATP at their interface [202].

We have obtained the most homologous 1000 non-redundant sequences and their alignments from the UniProtKB library. About 10% of these sequences show similarities across the whole helicase sequence and the best 95 has 400 or more positives or similarities in the sequence alignment (Figure 6.3). These are all coronaviruses mainly derived from bat virome (beta and alphacoronaviruses), and affecting various hosts in the animal kingdom, including humans. Intriguingly, the next best sequence alignment only covers 235 amino acids as similar residues; all of these and subsequent aligned regions are specific to the RecA domains and range through all types of organisms.

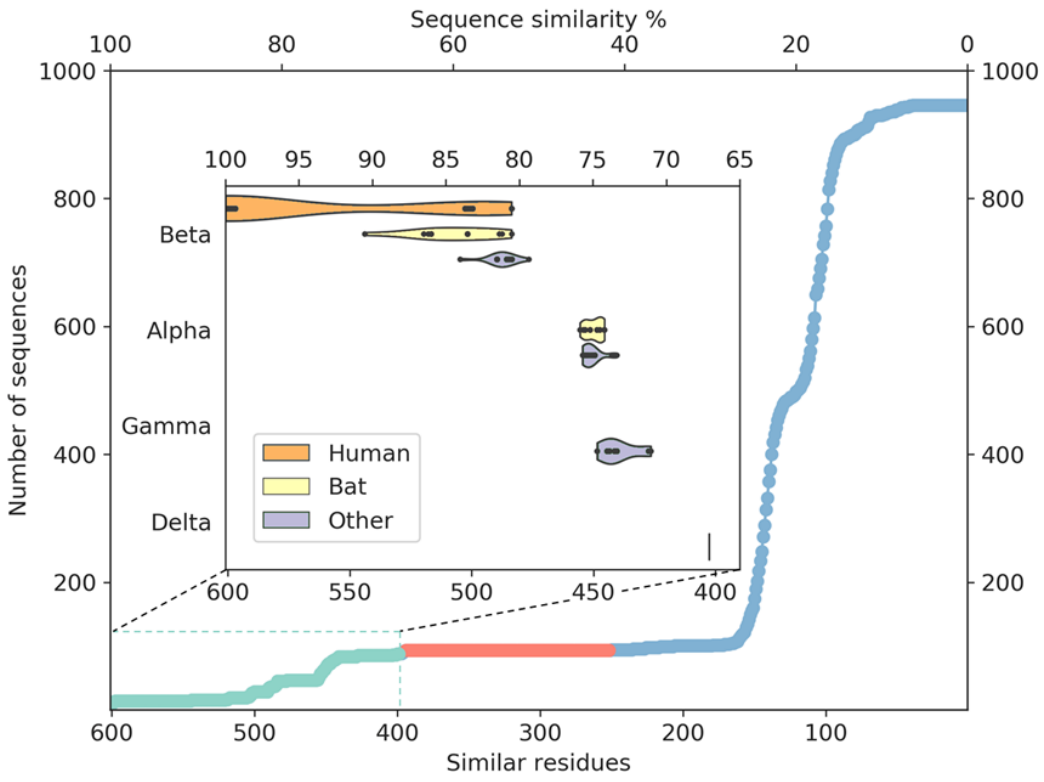


Figure 6.3. Distribution of the pairwise sequence alignments to the SARS-CoV-2 helicase. There are only members of coronaviridae above 398 matching residues (66%, lime circles, 95 entries). There are no sequences with medium similarity (235-394 similar residues, red circles). The closest relatives (95 sequences highlighted in lime dashed frame) are grouped in coronavirus subfamilies with principal hosts highlighted in the inset.

To evaluate any similarities to Domain 1 only, we also performed a search using only the first 230 residues. This search for sequences that match at least 70 residues resulted in the exact same 95 sequences as before, exclusively belonging to *coronaviridae*. An additional only 21 sequences match shorter segments of this domain between 1-230 residues, corresponding to a 22% sequence identity or below.

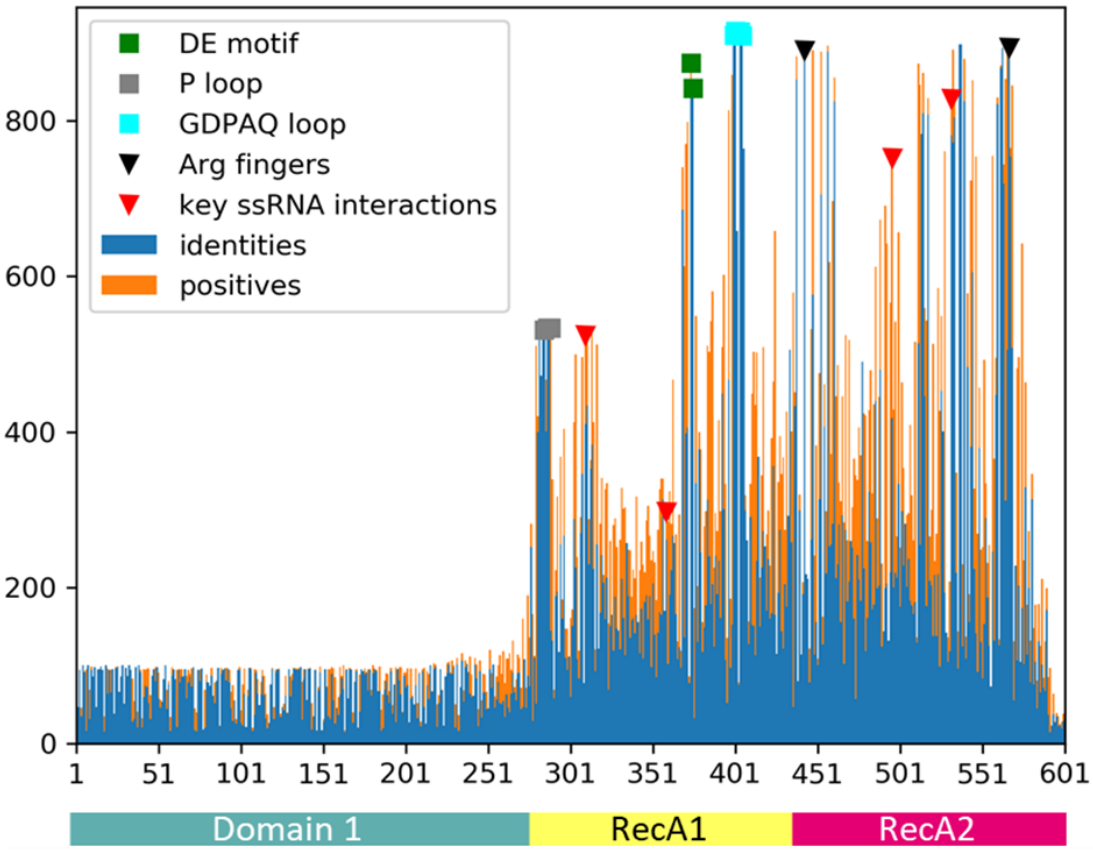


Figure 6.4. Sequence similarity (orange) and identity (blue) of the closest 946 sequences from UniProtKB using BLAST pairwise alignments to the 601-residue long SARS-CoV-2 RNA helicase. Domain 1 shows similarity only to the close relatives (95 sequences), while the RecA1 and RecA2 domains are more common across ATPase sequences. Key structural motifs are highlighted using symbols (P-loop: grey square, DE motif: green square, arginine fingers: black triangle, ssRNA interactions: red triangles).

Amongst crystal structures containing ATP analogues, most helicases have very low sequence similarity to NSP13. The closest homologues are 2xzo, 5mzn, and 6jim with 11.00%, 10.20%, and 8.40% sequence identity, respectively. Despite the low sequence identity, most residues in the ATP binding pocket are conserved. At the same time, the closest human sequence homologue based on our homology search, ZGRF1, a putative RNA helicase, shares only 22% sequence similarity, restricted to the RecA1 and RecA2 domains. This relatively narrow bandwidth of sequence similarity can be harnessed to design specific

inhibitors against the coronavirus RNA helicases that do not inhibit human proteins.

6.5.2 Structural Model of the ATP Binding Site

We modelled the ATP-bound active site using the 2xzo structure as a template (Figure 6.7). The essential Mg^{2+} ion cofactor coordinates both the β and γ-phosphates and a conserved Ser288. The active site contains a DE of the DEAD-motif of RNA helicases. The conserved Asp374 H-bonds with the Ser288 and one of the Mg-coordinating water molecules, whereas the Glu375 is positioned as the proton acceptor [171], [203], [204]. The γP is stabilized via H-bonds with Arg567 Lys289 and Gln404 through a water molecule, that are found in respectively 95, 56 and 57% of the homologous sequences analysed, while βP forms an H-Bond with Arg443. The sugar region of the ATP likely interacts also with Glu540 and Lys320 as seen in ten and four PDB structures, respectively. Unlike the highly conserved residues recognizing the triphosphate pocket, the environment of the sugar and purine moieties (Figure 6.7, right) shows a greater diversity. The purine ring is stabilized through multiple π stack interactions, from one side with Arg442, while in some helicases with tyrosine, and from the other side with His290 and Phe261. Additionally, there is an H-Bond between the amino group of the purine ring with Asn265, a residue that is more typically served by glutamine in similar sequences.

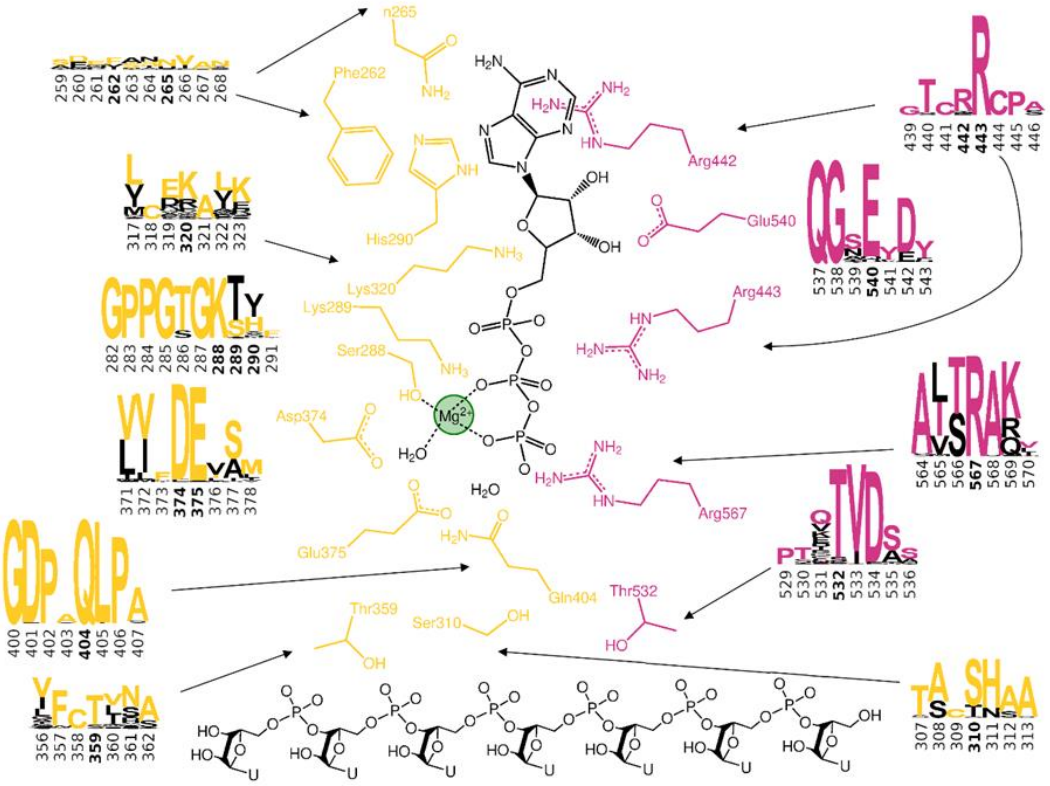


Figure 6.5. Conserved residues coordinating the ATP and RNA substrates of the SARS-CoV-2 helicase. Sequence conservation for RecA1 (orange) and RecA2 (magenta) domains are depicted in logos for each residue and its neighbours (data from Figure 6.4). The coloured letter represents the residues in the SARS-CoV-2 helicase sequence; depicted residue indices are bold in the logos.

A lack of specificity towards the purine group is likely due to the dual function of the SARS-CoV-2 helicase to aid the 5'-capping of the RNA by the triphosphate hydrolysis of most NTP substrates [164]. Due to these major differences, this area of the nucleotide-binding pocket can serve in the design of SARS-CoV-2-specific antiviral drugs.

6.5.3 Structural Model of the RNA Binding Site

To identify the main contact points with the ssRNA, it is first important to understand the unwinding function of the helicase. Filtering the related crystal structures to the ones containing RNA, we noticed that the RNA directionality relative to the ATP pocket is well defined (Figure 6.1). The unwinding is driven by a domain-wise translocation process, which moves the RNA one base in two steps [174], [175], [205]. Upon ATP hydrolysis, domain RecA2 translocate one base towards the 5' end of the RNA which is then followed by the RecA1 when a new molecule of ATP binds to the enzyme.

Domain 1, being in contact with the sidechain of the RNA, does not feature specific motifs, thus allowing different RNA bases to translocate. A long loop transition into the RecA1 and two domains are sandwiching the ATP pocket on the side of the RNA backbone. This region, equipped with the necessary functionalities to perform the ATP hydrolysis, has a higher degree of conservation along with the helicases. Both RecA domains have specific residues reliable for contact with the RNA phosphates, depicted in Figure 6.8. Thr359 in RecA1 and Thr532 are identified as the main anchoring points of the two domains. The base between these two threonine residues is coordinated by a backbone NH of His311, an interaction that is kept in RNA containing crystal structures. Ser310 is also reasonably conserved, although not directly featured in RNA coordination in this state of the enzyme.

Interestingly, the most conserved motif across the sequences is a GDP(A)Q loop interfacing between the RecA1 and RecA2 domains. This motif features Gln404, a residue that we consider to be important in the coordination of the nucleophilic water; moreover, it bridges the γP and the SH motif discussed earlier. We speculate these moieties play a role in the translocation of RecA2 upon ATP hydrolysis.

6.5.4 MD Simulations

6.5.4.1 Apo Structures

All replicas of the apo structure show low flexibility and no major changes in the backbone structure of the dimer. We analyze the overall flexibility of the dimers and compare our results with the experimental b-factor obtained in 6jyt and 6zsl (Figure 6.2). Our model, in common with the two crystal structures, shows higher flexibility on the external shell of the RecA2 domain, while the ATP and the RNA pockets appear to be more conserved. The ZBD shows low flexibility, in agreement with the b-value of 6jyt, but not with 6zsl (especially chain A), in which the temperature factor is higher, due to the different dimerization of the crystal structures.

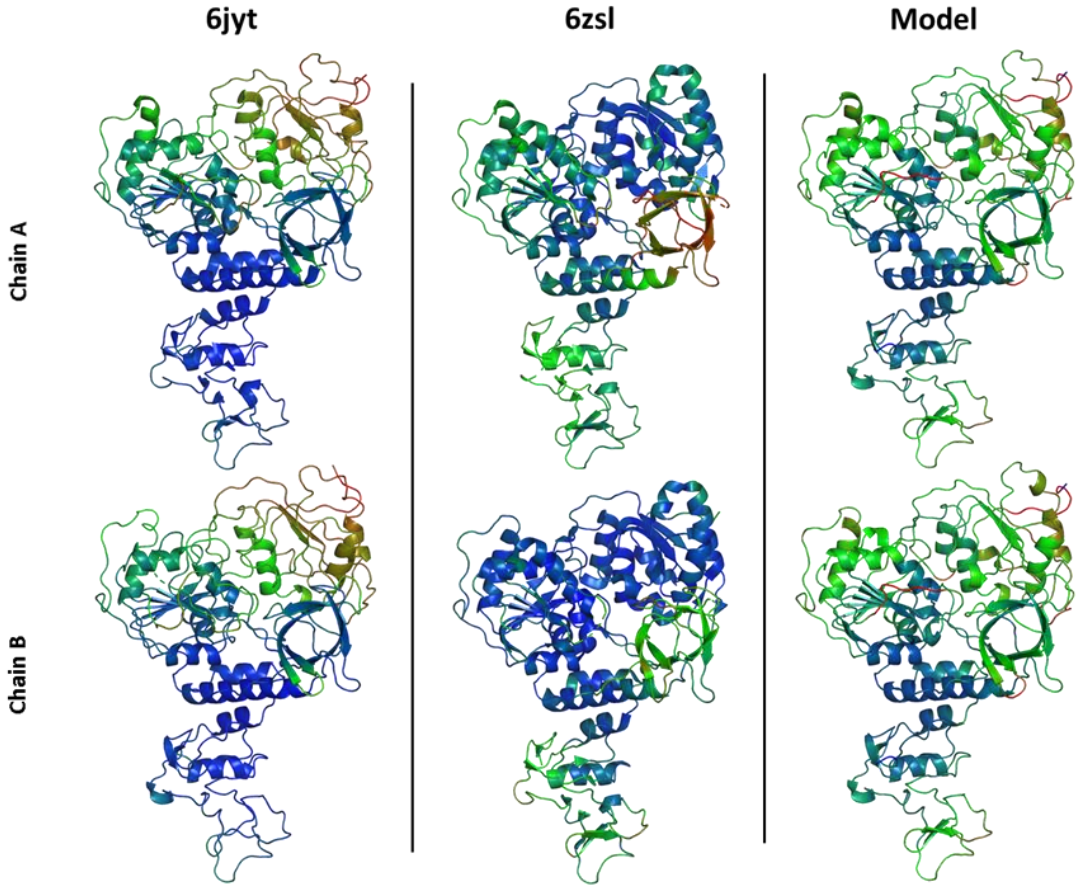


Figure 6.6. Conformational flexibility of the APO helicase protomers from 6jyt, 6zsl, and our model if the apo dimer from the MD simulations. The residues are coloured according to the deposited PDB B-factors (6jyt and 6zsl; from blue: low B-factor to red: high B-factor), and by the residue RMSD from the MD trajectory.

6.5.4.2 Pocket Analysis

We identify the cavity size of the ATP pocket using Fpocket. We tested different combinations of parameters to find a combination that yields more robust results. A comparison of the average pocket volume from the apo trajectory (394.6±157.5 Å3) with the average pocket volume from the ATP bound trajectory (367±146 $Å^3$) shows no substantial differences within the error of the analysis. This finding suggests that the ATP cavity structure is not altered significantly by the presence of ATP.

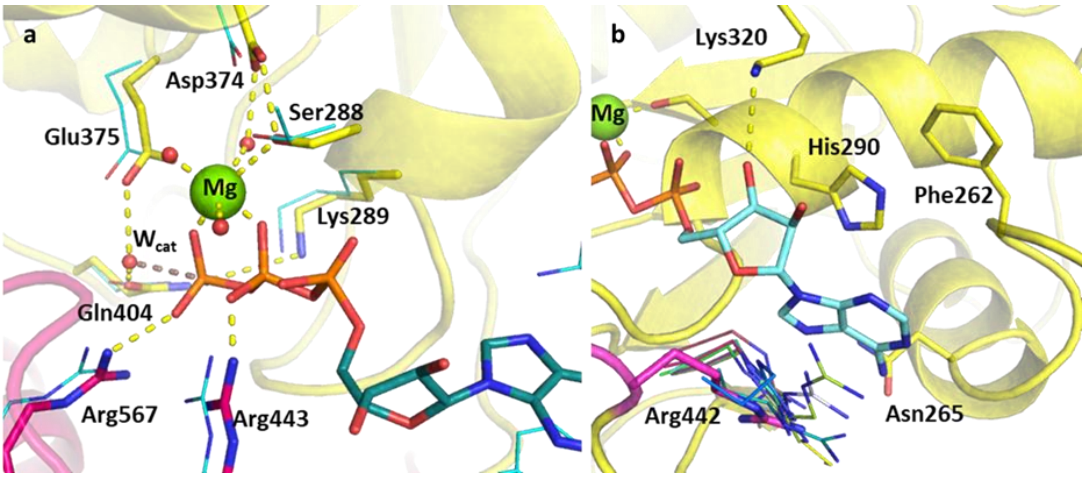


Figure 6.7. Structure of the ATP pocket aligned with homologous ATP-helicase complexes. RecA1 and RecA2 are shown in yellow and magenta, respectively. a) Main protein-substrate interactions of the triphosphate and magnesium ions are compared with alignment for PDB template 2xzo (cyan lines). b) Nucleotide-binding region focusing on Arg442 (magenta sticks) is aligned with homologous arginine residues (lines, PDB structures 5k8u, 5vhc, 5xdr, 5y4z, 5y6m, 5y6n, 6adx, 6ady, 6c90, and 6jim).

6.5.4.3 RNA Binding Site

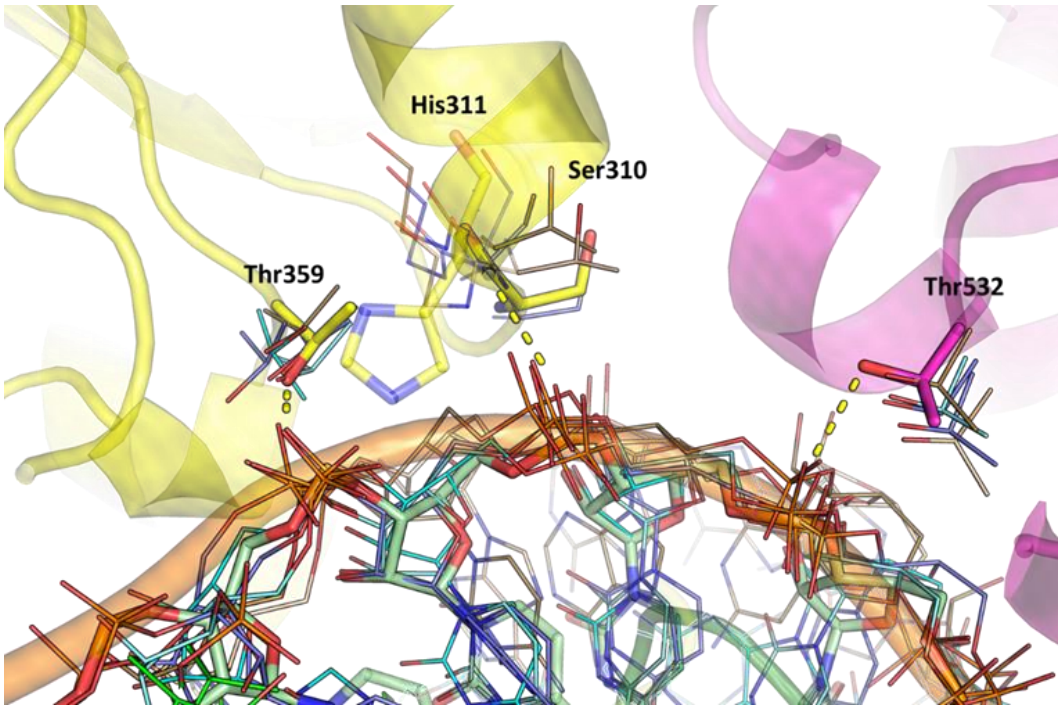


Figure 6.8 Structures of the RNA binding region aligned with existing RNA-helicase crystal structures complexed with RNA (depicted in lines). RecA1 and RecA2 domains are shown in yellow and magenta, respectively. Key residues (sticks) are labelled, and H-bonds are depicted in yellow dashes.

From the structural analysis and the homology modelling, we denote two important and well conserved interactions between the RNA and the helicase. Both interactions involve an H-bond between threonine and the O of the phosphate residue on the backbone of the RNA. Specifically, from Thr359 from RecA1 and Thr532 from RecA2. Additionally, another H-bond is made between the central RNA and the N of residue 311, this residue is not highly conserved, but the interaction between the backbone of this residue and the OP of the RNA is present in several PDB structures. A key residue close to the RNA pocket is Ser310; this residue is conserved (often present as a threonine) and appears to be important for the communication between the ATP pocket and the RNA pocket.

6.6 Summary

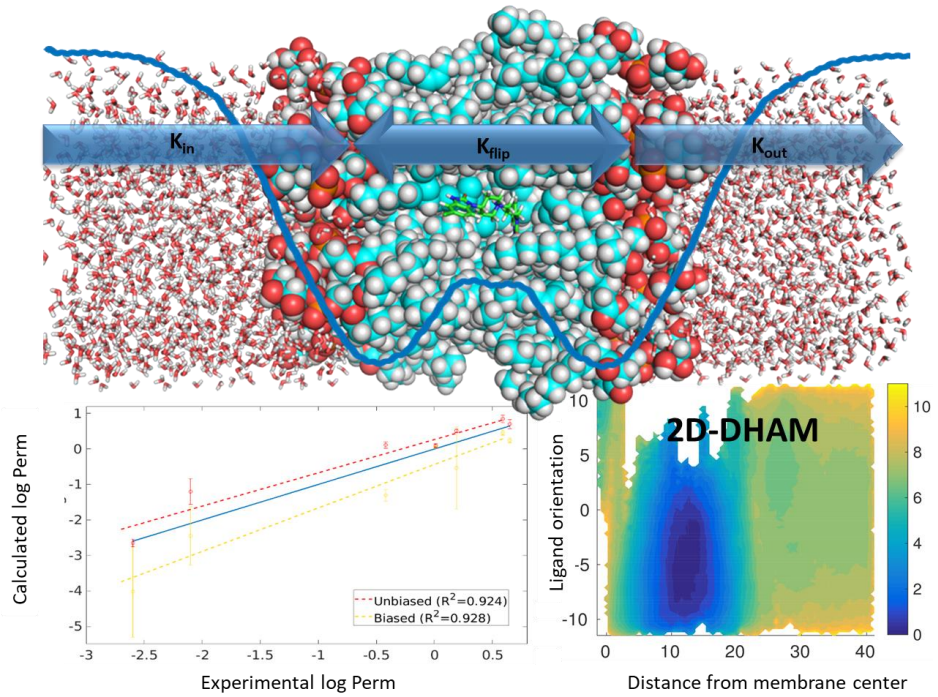
We present the catalytically competent computational model of the SARS-CoV-2 NSP13 ATP dependent RNA helicase. Our structure is the first to host ATP and a single strand RNA with detail to the binding modes of the substrates.

The analysis of homologous sequences shed light upon the specificity of the domain structure of the viral helicase yielding no match over 40% except close relatives from the coronaviridae family. Our model features two major improvements compared to the apo protein. The ATP pocket was reconstructed based on the structural conservation of the most similar crystal structures including signature motifs from phosphate-binding proteins such as the DE(AD) of helicases, the P loop, or the arginine fingers. Furthermore, we identified the main anchoring points of the ssRNA through the helicase, which are essential to understand the translocation driving the unwinding activity of NSP13.

With molecular dynamics, we have verified the stability of conserved interactions in our model as well as improved our initial model to host the nucleic acid. We assess the flexibility of the ATP pocket with and without the nucleotide bound detecting a well-maintained cavity.

Our work provides insight into one of the viral replication machinery elements, ideal for targeting by drug developers. Our structure can be the basis of structure-based compound design and screening. Moreover, elaborating the RNA translocations driven by the identified interactions can reveal other targetable states of the helicase.

Chapter 7 Calculating Kinetic Rates and Membrane Permeability from Biased Simulations



7.1 Preface

This work represents an early project I was working during the early years of my PhD. The project aimed to understand different types of molecules' behaviour while crossing a lipid membrane through molecular dynamic simulations. Using the Dynamic Histogram Analysis methods, I predicted the kinetic rates of seven know drugs crossing a membrane lipid bilayer and described the molecules' orientations using a 2D Markov model approach. The kinetic rates obtained from the model agrees with the experimental data.

My contribution to this work was to analyse the results from umbrella sampling simulations using the Dynamic Histogram Analysis method and define their preferred orientations while crossing the lipid membrane.

This work was successfully published in The Journal of Physical Chemistry B:

Badaoui, M., Kells, A., Molteni, C., Dickson, C. J., Hornak, V., & Rosta, E. (2018). Calculating Kinetic Rates and Membrane Permeability from Biased Simulations. *The Journal of Physical Chemistry B*, *122*(49), 11571-11578.

7.2 Abstract

We present a simple approach to calculate the kinetic properties of lipid membrane crossing processes from biased molecular dynamics simulations. We demonstrate that by using biased simulations, one can obtain highly accurate kinetic information with significantly reduced computational time. We describe how to conveniently calculate the transition rates to enter, cross and exit the membrane in terms of mean first passage times. By constructing a Markov model from the biased data using the Dynamic Histogram Analysis Method, the spectral properties of the resultant model allow for the easy calculation of free energy barriers and relaxation times. The permeability coefficients that are calculated from the relaxation times are found to correlate highly with experimentally calculated values. We show that more generally, certain calculated kinetic properties linked to the crossing of the membrane layer (e.g., barrier height, barrier crossing rates etc.) are good indicators of ordering drugs by permeability. Extending the analysis to a 2-D Markov model allows for a physical description of the membrane crossing mechanism.

7.3 Introduction

For a drug to be effective, not only has to bind strongly to its target, but it is also required to have good ADME (Absorption, Distribution, Metabolism, Excretion) profile [206]. An important factor for the absorption and the distribution is the drug's ability to cross the cell membrane to reach its target [207]–[209]. This became particularly important for drugs that act in the central nervous system and have to cross the blood-brain barrier [206]. This property is traditionally estimated by the lipophilicity of the drug. However, taking into

account only the lipophilicity of the molecule does not allow to fully understand the mechanism of membrane permeation (Figure 7.1), and for this reason, subsequent more refined models take into account additional physical parameters, such as depth-dependent partitioning and the resistance coefficient of the membrane.

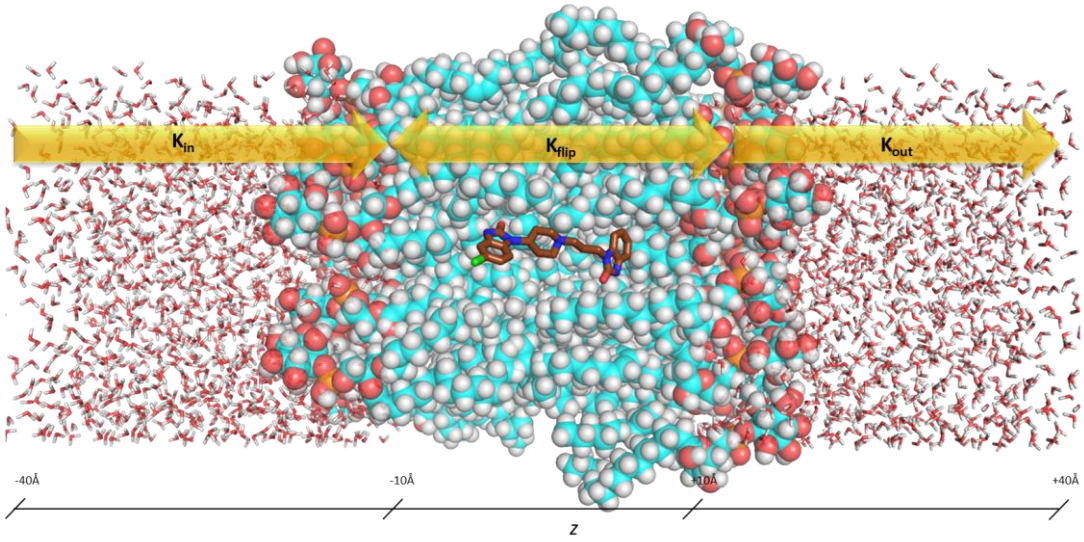


Figure 7.1. Representation of the system used in the molecular dynamics simulations: a drug molecule (in brown at the center of the image) interacts with and passes through a lipid membrane which is surrounded by water.

For a fully quantitative description, it has become fundamental to predict the kinetic behaviour of drugs addressing membrane interaction and permeation [206], [210]–[213]. Studies related to the transport of small ligands crossing various phospholipid membranes are the subject of increased interest in recent years [210], [214]–[217]. There are also significant challenges to investigating this behaviour experimentally. Eyer et al. [218] proposed a liposomal fluorescence assay method by which the permeation of weak basic drug-like solutes across the lipid membrane can be determined. However, details of membrane crossing mechanisms at an atomistic level are still missing experimentally [218]. Thanks to the dramatic recent development of computer

technology, molecular dynamics (MD) simulations are now capable of reaching biologically significant time scales and are becoming widely used in the pharmaceutical industry [213], [219]–[227]. In tandem with the improvement in simulation hardware and software, an important role has been played by the construction of mathematical models which allow the vast volumes of MD data to be processed in a statistically optimal manner. Markov state models (MSMs) have emerged as a useful tool for analyzing and understanding the results of these simulations. In fact, MSMs allow for the convenient combination of multiple MD trajectories into a single kinetic network model from which experimental observables and kinetic rates can be computed [211], [217], [228]–[230]. Using experimentally obtained permeabilities by Eyer et al.[218] across a lipid membrane for seven structurally unrelated drugs (Figure 7.2), Dickson et al.[231] recently demonstrated that accurate results for the permeability rates can be obtained by running long unbiased MD simulations [231].

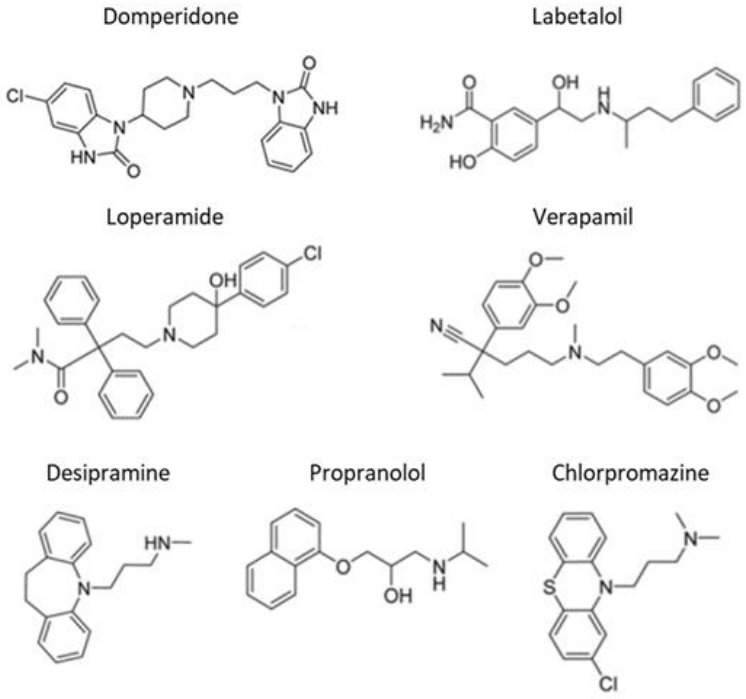


Figure 7.2. Chemical structure of the seven drugs analysed by Eyer et al.[218]

By using an MSM formalism, kinetic rates of the key steps in membrane crossing can then be estimated. However, very large computational resources are required for a sufficiently converged set of unbiased simulation trajectories to be analysed by MSMs. With the use of enhanced sampling biasing procedures, such as umbrella sampling (US), this computational time can be significantly reduced. The construction of MSMs from biased simulation data has not been traditionally possible. Biased simulations require the potential energy function of the system of interest to be modified such that the system is, for example, harmonically restrained to a given region of the energy landscape. This method is advantageous as it allows sampling of regions which might otherwise not be adequately visited during the simulation time. However, the kinetic behaviour observed is no longer representative of the true system, and as such, this needs to be accounted for when constructing the MSM. A recently derived unbiasing method, the dynamic histogram analysis method (DHAM), by Rosta and Hummer[232] uses a maximum likelihood estimate of the MSM transition probabilities given the observed transition counts during each biased trajectory and is found to produce often more accurate results than those of the more commonly used weighted histogram analysis method (WHAM) [42]. This unbiasing method is the first to use only biased US simulation data to obtain kinetic information directly by constructing the unbiased MSM. Here we determined the free energy profiles and kinetic rates of crossing a lipid membrane for the seven drugs represented in Figure 2 by using US biased simulations. All experimental kinetic permeation data used for comparison for these seven drug molecules was previously obtained by Eyer et al. [218]. Using this kinetic information, we aim to order the drugs according to their permeability coefficients (log Perm values). We analysed US simulation data to calculate kinetic rates for the entry into the membrane, flipping, and exit from

the membrane, and we compared it with that obtained from long unbiased simulations. All MD simulation data (unbiased and US biased) was previously obtained by Dickson et al. [231]. Here, we reanalysed the US biased data to obtain molecular kinetic rates for the membrane permeation using DHAM [232]. We found an excellent agreement between the kinetic properties of the drugs from US biased simulations compared with those from the combined biased and unbiased MD simulations, which are also in agreement with experimental permeation measurements, demonstrating that these calculations provide accurate in silico kinetic rates for these important dynamical processes. Additionally, we analysed the free energy surfaces corresponding to the orientations of the seven drug molecules while crossing the membrane by determining MSMs on a two-dimensional (2D) surface using DHAM, describing in detail the orientation for three of them. This method provides key insights into the drug permeation pathways and offers guidance for the design of molecules with required kinetic permeation properties.

7.4 Method

7.4.1 Markov State Modelling

An MSM consists of a set of memoryless conditional probabilities between user-defined discrete states (in our implementation along with a finely discretized chosen reaction coordinate z), such that the value of $P(j,t|i,0)$ is the probability that the system is in state j at time t given that it was in state i initially. These conditional probabilities are typically calculated by determining the transition count matrix C_{ji}, which contains the count numbers of the observed transitions

from state *i* to *j*. The time parameter *t* is called the lagtime and must be chosen sufficiently large such that the Chapman−Kolmogorov test [233] is satisfied (i.e., that the relaxation time scales, τ of the system are insensitive to changes in the lagtime). To produce an MSM from enhanced sampling simulations in practice, we use a reaction coordinate of interest that was also employed to bias the MD simulation data. In the context of membrane permeation, it is desired to compute the kinetic rates with which the drug undergoes three important processes (see Figure 7.1): the rate at which it enters into (k_{in}), crosses (k_{flip}), and exits (k_{out}) from the membrane. The corresponding reaction coordinate is the distance between the center of mass (COM) of the ligand and the center of the lipid membrane (*z* coordinate as shown in Figure 7.1) was used. Unlike in typical MSM models consisting of only metastable states, here we discretized this coordinate into bins, where the number of bins is chosen sufficiently large to give a finely discretized coordinate but not so large as to give an under-sampling of transitions between bins. Once the bins have been determined, we count the number of observed transitions ($C_{ji}^k(t)$) between each pair of bins *i* and *j* in simulation *k* at the chosen lagtime *t*, as well as the number of times each bin is occupied ($n_i^k = \sum C_{ji}^k(t)$) during each simulation *k*. These values then provide the necessary conditional probabilities *Mji(t) = P(j,t|i,0)*. In the simplest unbiased case not enforcing detailed balance strictly, the maximum likelihood estimates are given by

$$M_{ji}(t) = \frac{\sum_j C_{ji}^k(t)}{\sum_k n_i^k} \tag{7.1}$$

For biased simulations where a biasing energy of u_i^k is applied to state i during simulation k, we employed the DHAM [232] to compute the unbiased MSM from the biased data as given by

$$M_{ji}(t) = \frac{\sum_j C_{ji}^k(t)}{\sum_k n_i^k \ \exp\left(-\frac{u_j^k - u_i^k}{2k_bT}\right)} \tag{7.2}$$

Equation 7.2 reduces to the unbiased equation when the biasing potentials are set to zero. Once an MSM has been constructed from simulation data, one is typically interested in determining the free energy profile as well as the kinetic information (relaxation times and mean first passage times). These quantities can be computed directly from the eigenvalues λ_n and eigenfunctions ψ_n of the transition matrix. All the eigenvalues of the transition matrix with detailed balance fall between 1 and 0 and can be arranged in decreasing order

$$1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n > 0 \tag{7.3}$$

The largest eigenvalue (equal to 1) gives the equilibrium populations of the states of the system (useful to find the free energy), while the second largest eigenvalue can be used to determine the time scale of the slowest relaxation process in the system via

$$\tau_2 = \frac{-t}{\ln(\lambda_2)} \tag{7.4}$$

The kinetic rates are computed by coarse graining our discretized states and the corresponding free energy profile into four regions, the outer water, outer membrane, inner membrane, and inner water regions (Figure 7.1), using the robust Perron cluster analysis (PCCA+) method [234] following Dickson et al. [231]. Once the clusters have been specified, we calculate the rates (k_{in}, k_{flip}, and k_{out}) from the Markov matrix as the inverse of the mean first passage times

(MFPT) [235] between the regions. The log Perm permeability values are typically calculated using Equation 7.5

$$\mathrm{Perm} = \frac{k_{slow} r}{3} \tag{7.5}$$

where k_{slow} is defined to be the rate of the slowest process in the system, i.e., the process which is most significant in describing the decay of the populations to equilibrium, and r is the radius of the liposome (100 nm). It should also be noted, that in this context, we use a base ten logarithm as is typically used when analyzing membrane permeability values. Typically, for membrane crossings, it is estimated by setting up a system of equations using the three rates (k_{in}, k_{flip}, and k_{out}) as inputs, solving these equations in a kinetic network model and fitting the time dependent populations to a biexponential curve. Here we propose a simpler and more direct approach to calculate the overall slowest relaxation time directly from the original Markov model and obtain a corresponding rate for k_{slow}. We compared and demonstrated that this simple approach is highly accurate and results in a similar k_{slow} estimate as the traditional approach. Recently, a number of advances of DHAM have been proposed where detailed balance is included [236]–[238]. We found that enforcing detailed balance did not lead to any observable changes in many US test cases we studied (data not shown); therefore, here we used the simplest original DHAM approach via Equation 2 [232], [236].

7.4.2 Simulation Details

Compounds were modelled with the parm@Frosst force field, a small molecule force field that extends AMBER ff99SB [239] and uses conformationally averaged AM1-BCC charges; lipids were modelled using the AMBER Lipid14

force field and water using the TIP3P model. MD simulations were run with AMBER16 and PMEMD CUDA on GPU cards [240]. The starting structures for the US simulations were obtained by placing each ligand at the centre of a POPC bilayer surrounded by water molecules (72 POPC and 60 water molecules per lipid) [241]. Three-dimensional periodic boundary conditions with the usual minimum image convention were employed. Energy minimization was performed by using the steepest descent method for 5000 steps and using the conjugate gradient method for a further 5000 steps. The system was then heated from 0 to 100 K using Langevin dynamics within a 5 ps constant volume run, with restraints on the drug molecule and lipids using a force constant of 10 kcal mol−1Å−2. Subsequently, the volume was allowed to change freely, increasing the temperature to 303 K. The Langevin collision frequency was γ = 1 ps−1, and anisotropic Berendsen control of the pressure around 1 atm was applied by coupling the periodic box with a time constant of 2 ps for 100 ps. The equilibration was completed after an additional 5 ns with the pressure relaxation time reduced to 1 ps in NPT, removing the restrain on the lipids. The SHAKE algorithm[242] was used to constrain the bonds involving hydrogen, and a time step of 2 fs was used. Using a pulling rate of 1 Å/ns, the drugs were then pulled out from the center of the system to outside the membrane, for a total of 40 Å (force constant of 1.1 kcal mol−1Å−2), in the NPT ensemble with semi-isotropic pressure scaling. During the simulations, a snapshot was saved every 1 Å, from the center z = 0 Å to z = 40 Å generating 40 windows. The results were calculated for one bilayer leaflet, and it was assumed that the second half behaves in the same way. This was achieved by reflecting the data along the z axis and adding 39 or 40 windows, depending on whether or not the window at z = 0 Å was reflected as well. Each US window was run for 20 ns to allow equilibration, followed by additional 80 ns of the production run in NVT

condition using an US force constant of 2.5 kcal $mol^{-1}Å^{-2}$. Configurations were recorded every 10 ps.

7.4.3 2D-DHAM

Analogously to the 1D case, we constructed a finely discretized 2D grid to determine the MSMs along with two reaction coordinates for the seven drugs. Specifically, for domperidone, loperamide, and labetalol, we analysed the drug's rotational movement during its passage across the membrane.

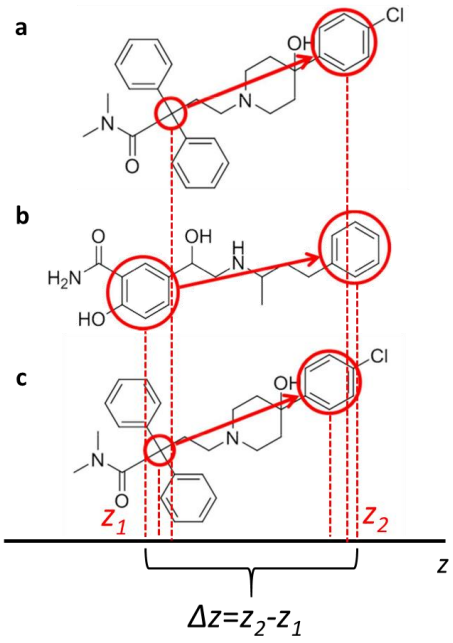


Figure 7.3. Definition of the Δz coordinate used in our 2D-DHAM analysis. The values are obtained by projecting the vector along the drug molecules' length, as shown by the red arrows, onto the z-axis. The vector describing the molecular length joins the COM of the circled atoms, as shown for Domperidone (a), Labetalol (b), and Loperamide (c).

As our first reaction coordinate, we used the same z coordinate as previously determined (distance from ligand COM to membrane center). For our second coordinate, we used the projection of the molecular orientation vector onto the z coordinate Δz, as a measurement of the orientation of the ligand with respect

to the membrane for two selected regions of the drug molecule along its length. Δz is equivalent to the molecular length scaled by the cosine of the angle between the z axes of the membrane and the molecular vector defined by the two ends of the ligand (Figure 7.3). With the projection along the z axes, we can predict the orientations of the molecules, for example when the ligand is oriented parallel to the membrane, Δz will be around 0 Å, as both ends are equidistant from the membrane, whereas when it is oriented perpendicular, then Δz will be equal the end-to-end length of the ligand (around −10 Å). The extremities of the ligand can be the COM of distal functional groups (e.g., benzene) or single atoms, as shown in Figure 7.3 for the molecules considered here. This 2D-DHAM analysis and the 2D free energy surfaces were used to find correlations between the rotation of the ligand and its position across the membrane, showing how the orientations of the ligand affected the free energies while crossing the membrane.

7.5 Results and Discussion

7.5.1 MSM Analysis of US Simulations

Using the Markov modelling methods and US simulation trajectories, the relaxation time, τ_2, was calculated by constructing MSMs at a range of lagtimes up to 300 ps with 1000 bins, as shown in Figure 7.4. Using a recently derived method for calculating the limiting relaxation time of an MSM [243], we determined the long lagtime limit of the relaxation time for each drug, as shown by the dashed lines in Figure 7.4. The relaxation times can be seen to level off in the region of lagtimes greater than 100 ps.

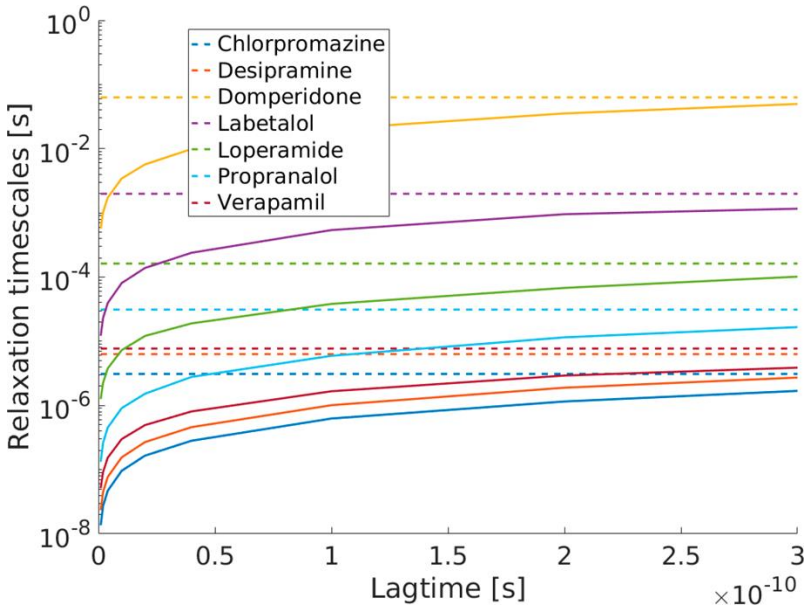


Figure 7.4. Relaxation time vs lagtime of the seven drugs (Figure 7.2). The dashed lines represent the long lagtime limit of the relaxation time obtained by a least-squares fitting to the relaxation times in the range of 1−300 ps.

In the analysis that follows, we chose to use a lagtime of 200 ps, as it is sufficiently large for τ_2 to be insensitive to the precise choice of the lagtime.

When calculating the MSMs with bin numbers of 600, 800, and 1000, at our chosen lagtime of 200 ps, there is almost no change in the obtained free energy profiles. We used 1000 bins for all subsequent analysis. Following this initial choice of parameters, seven Markov models were constructed with 1000 bins and a lagtime of 200 ps (100 000 simulation steps). This allows us to compute the free energy profiles for each drug and draw a comparison with the profiles obtained in the unbiased simulations using WHAM (Figure 7.5). Error bars were determined by dividing the data into two equal sections, determining the profiles independently, and calculating the variance. All of our free energy profiles show the same trend as the one calculated by Dickson et al. [231] (dotted lines in Figure 7.5) for the combined unbiased and biased MD data, and indeed, all of the WHAM predictions fall within the margin of error for the

DHAM results. While the PMF changes depending on whether the US window at z = 0 Å was reflected or not (Figure 7.5), the log Perm data are essentially unchanged. The asymmetry observed in the not fully reflected PMF profiles also suggests that longer simulations might be needed to reduce the error at this transition region. At the same time, we used a fraction of the data required for the unbiased simulations.

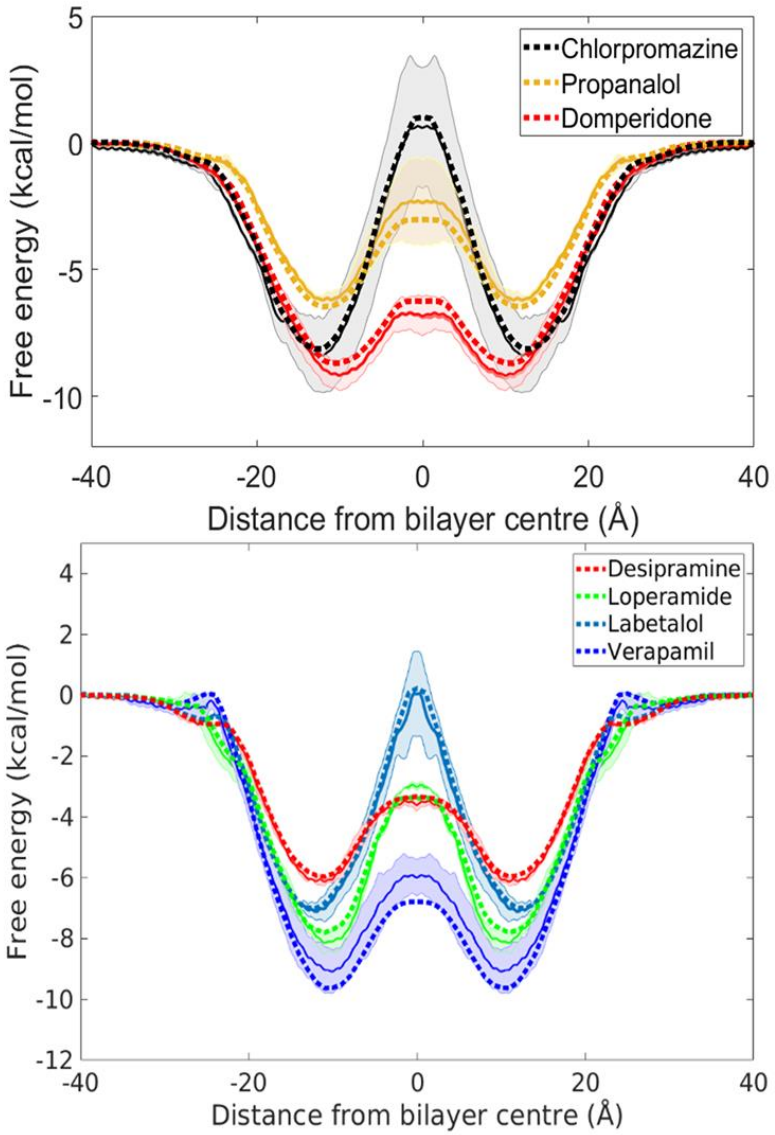


Figure 7.5. Free energy profiles calculated with DHAM from US simulations (solid lines) and WHAM using unbiased MD data (dashed lines). Errors are represented by the shaded area for US data.

We obtained the kinetics profile using the US data by Dickson et al. [231] with a total simulation time of 3.2 μs for each drug, whereas in the work done by Dickson et al. [231], the calculation of the kinetic profile required multiple unbiased simulations, with a total simulation time of 12.5 μs per drug. By analyzing the US data with DHAM, we are able to reduce the total time by at least 75% over using unbiased data.

7.5.2 Ordering Drugs According To Their Permeability

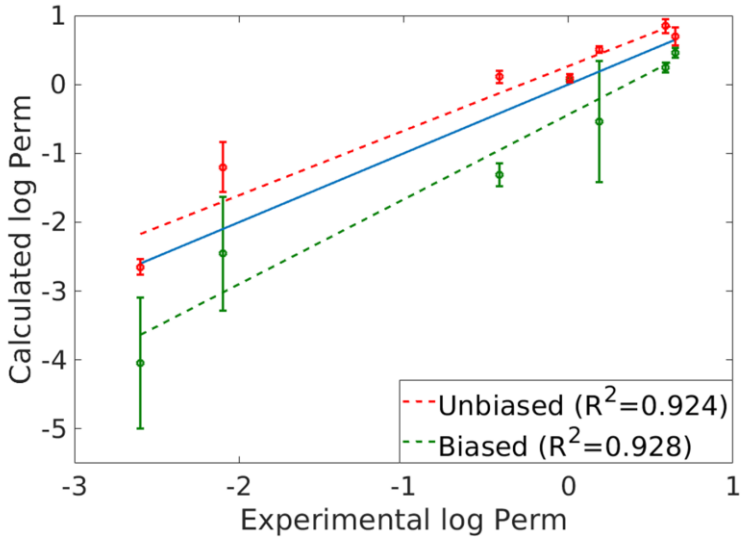


Figure 7.6. Log Perm values determined by the biased and unbiased simulations are compared with the experimental values [218]. The correlation in between the data sets is comparably high for both the biased and unbiased simulations (both have p-values of well below the 5% required to be statistically significant).

To determine the relative permeability, it is required to compare the rate of the slowest occurring process, k_{slow} corresponding to the crossing of the free energy barrier at the center of the membrane among the different drug molecules. Here, we considered several ways to estimate the relative ordering. First, we can use the overall relaxation time corresponding to the second eigenvalue of the MSM constructed for each drug using Equation 4. Second, we can make use of the free energy profile alone and compare the height of the free energy barrier across the different drugs, using an Arrhenius relationship

$$k_{slow} = Ae^{-\Delta G^{\ddagger}/k_B T} \tag{7.6}$$

Using the relaxation time obtained from the MSM in conjunction with the ΔG‡ calculated from the populations, we can determine the Arrhenius prefactor. We obtained similar prefactors for all of the drugs, with an average value of 9.72 ± 5.76 e+07 s^{-1}, 4 orders of magnitude bigger than the typical value of $A = \frac{k_B T}{H}$, considering a transmission coefficient close to 1. Third, as the barrier corresponds to the flipping process, we can use the rate constants determined by MFPT, assuming $k_{slow} \approx k_{flip}$. These three methods are each computationally simple to implement compared with alternative methods in the field of simulating a kinetic system from the calculated rates and performing a biexponential fit to the resultant time-dependent probabilities. The biased and unbiased calculated log Perm values correlate very well with the experimental data (Figure 7.6). The US simulation data displays similar R^2 values from the linear fit as the original kinetic data. The log Perm values from the combination of unbiased and biased potential of mean force (PMF) data with the discrete transition-based reweighting analysis method (dTRAM) of Dickson et al. [231] mostly lie above the experimental values predicting slightly faster permeation, while the biased values are almost all below the line. This slow time scale might be because our model was calculated at a larger lagtime. Increasing the lagtime will increase the relaxation time and, in turn, decrease the value of the rate of the slowest process, resulting in a smaller permeability value. We expect that the most accurate simulation-based rate estimates are calculated from all data (biased and unbiased) using longer lagtimes. Importantly, the process of ordering drugs according to their permeability is insensitive to the precise choice of lagtime. This can also be seen from Figure 7.4, where the ordering of the lines does not change as a function of the lagtime, predicting the same ordering in a lagtime independent manner. This demonstrates that equivalently high correlations can be found between the experimental and biased data as with the unbiased data. Furthermore, using the simple approach of the

relaxation time of the full Markov state model is an appropriate way to order the permeability of the drugs. By analyzing various kinetic quantities as predictors of the ordering of the drugs by permeability, we found that, in general, any sensible choice of the kinetic quantity which is closely related to the barrier crossing process will serve as an accurate indicator of drug ordering. The MSM relaxation times correlate very well with the calculated free energy barriers. The corresponding permeation obtained from the free energy barrier heights using an Arrhenius rate expression with a constant prefactor of k_BT/ h does not match the experimental log Perm values as closely as the MSM relaxation times. However, because the R^2 calculations are invariant under linear transformations, the free energy barrier can also be used to calculate log Perm values accurately. If the permeation is investigated using different membrane compositions, the Arrhenius prefactor may vary, and a kinetic comparison using MSMs might become necessary.

7.5.3 2D-DHAM

Using the 2D-DHAM analysis, we calculated the 2D free energy surface of all seven drugs. Here we illustrate the results on three of them, domperidone, loperamide, and labetalol, focusing on the rotation of the molecules while crossing the membrane. We also verified that the free energy barriers from the 2D-DHAM analysis agree well with 1D-DHAM results. Domperidone (Figure 7.7a), due to its polar characteristic, has a specific orientation inside the membrane. In the surrounding aqueous region, the molecule is free to rotate its z position between −40 and −25 Å. Once near the membrane, domperidone has a preferential orientation parallel to the surface of the membrane. Between the inter lipid region and the polar head (0 Å < z < 20 Å) of the membrane, it orientates perpendicular to the z coordinate showing a particular preference

where the structure is parallel to the membrane surface. Due to its dipole moment, in between the two phospholipidic layers, domperidone switches position, preferring a parallel orientation with its more polar end pointing toward the water along the z coordinate. This phenomenon is known as solute hopping [244]. The second compound, labetalol (Figure 7.7b), has an even stronger polar side, due to the presence of both hydroxyl groups and an amide group. On the other end, the molecule has a hydrophobic side, showing an overall "lipid-like" structure. When the drug is near the polar head of the membrane, it keeps its polar region close to the polar side of the membrane. Once at the intermembrane layer, it has a rapid interchange of orientation, keeping always its polar region close to the polar region of the membrane closest to bulk water. Loperamide (see Figure 7.7c) is the most hydrophobic of the three drugs, it prefers a specific orientation only when entering the membrane, with its hydroxylic group facing the membrane headgroups.

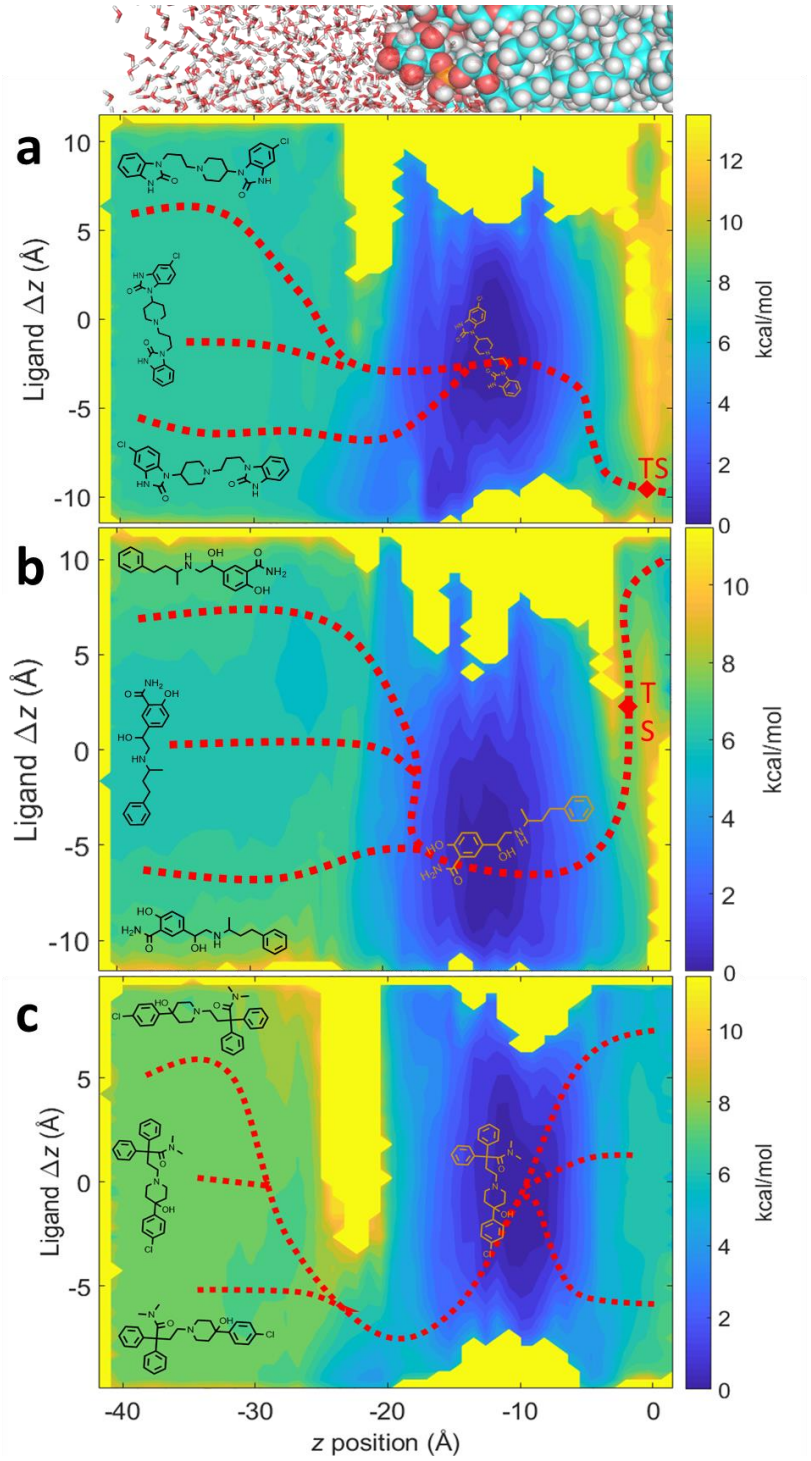


Figure 7.7. 2D free energy surfaces of (a) domperidone, (b) labetalol, and (c) loperamide along with the absolute z position of the ligand, and the Δz coordinate for each molecule (schematic representation of the molecule

orientation is also shown). The preferred paths for membrane crossing are shown as a function of the molecule orientation (red dotted lines).

Once entered, it tends to have relatively high rotational freedom. As quantitatively assessed by 2D free energy surfaces as a function of the z and Δz coordinates, depending on the polarity and symmetry of the molecule, once inside the membrane, molecules have specific preferential orientations during the passage across the membrane. Several works have already been done to analyze the orientation of the ligands while crossing the membrane [245]–[248], and our results and general trends from the 2D-DHAM analysis agree with these previous works. Polar molecules keep their polar region facing toward the polar heads of the membrane, while more lipophilic compounds have a higher rotational freedom. Furthermore, polarity and charge distribution also determine the orientation of entry and the corresponding free energy pathways into the lipid membrane.

7.6 Conclusion

We demonstrate that by performing a series of biased simulations of a drug molecule near a lipid membrane, highly accurate equilibrium and kinetic information can be determined by constructing an MSM using DHAM. This gives results which agree closely with experiment and achieve similar levels of accuracy as those attained by much longer unbiased MD simulations. Furthermore, we present a simpler method for calculating permeability coefficients from MD simulation data by calculating the relaxation time directly from the MSMs. We also find that if the goal is to order the drugs according to permeability, then most kinetic quantities correlate with the free energy barrier to cross the membrane, indicating that linear transformations would give an excellent approximation to the experimental log Perm value. While this is very

promising to order drugs in the same membrane environment, possibly such correlation with the barrier height no longer holds across different membrane/aqueous environments. We found that the prefactor in the Eyring equation differed by about 4 orders of magnitude from k_BT/h. This could potentially be due to the fact that the diffusion coefficients are very different inside the membrane that has a very different dielectric constant than water, or it could be due to other factors, including the choice of the reaction coordinates affecting the transmission coefficient. Finally, we constructed 2D free energy surfaces and corresponding MSMs for three of our drug molecules and interpreted the crossing mechanisms in terms of the physical processes occurring during the simulations. The molecular properties, i.e., charge distribution and lipophilicity, of the solute determine specific rotational preferences and pathways during the membrane entrance and crossing processes. Our results demonstrate that DHAM is capable of providing accurate molecular kinetic information from purely biased simulations. As the range of systems with biased simulations is very flexible, we plan to apply this method in multiple applications. We can determine unbinding rates in molecular systems, such as in host-guest complexes, e.g., the competitive binding of ethanol and methanol with cucurbiturils in nanoaggregates of Au nanoparticles in an aqueous environment [249] or for catalytic rates of enzyme-catalysed chemical reactions, such as the reaction mechanism of lipoxygenases [250]. Future work will be addressed to larger ligand permeability data sets, the kinetic prediction of ligand-protein unbinding, and other important relevant kinetic processes.

Chapter 8 Conclusion and Perspectives

New drugs are continually required to fight diseases, the most evident example is the recent spread of Covid19 pandemic. One of the major bottlenecks in the drug discovery process is the lack of

efficacy. The thermodynamic and kinetic understanding of the interactions between a ligand and its target is fundamental to improve the development of new drugs. MD simulations can be used as complementary solution to speed up the drug discovery process. In this thesis, I have presented several works in which the core subject is the study of ligand-protein interactions at atomic levels using MD simulations. The aim of the thesis is to provide insights of MD methods that can be used in drug discovery, and the effectiveness of these methods by applying it to relevant biological systems.

The first work presented in this thesis represents a state of the art method that predict unbinding kinetics of protein ligand complexes. The method consists of using enhanced MD simulations to first sample the unbinding reaction path and collect valuable information on the mechanism of unbinding. The novelty of the method lies in using an automatic way to iteratively add and remove collective variables during the unbinding trajectory. The method allows us to discover novel interactions not available when collective variables (CVs) are selected *a priori*, using methods where CVs are usually defined from the initial bound structure. From the unbinding trajectory we obtain a set of CVs and an unbinding path that is then used in conjunction with the well establish finite temperature string method to calculate the absolute free energy barrier of the unbinding process. The results obtained agree with the experimental data, showing how the method can be used as an alternative to other already

established methods. Additionally, we focused our attention on the transition state ensemble, where using a combination of unbiased MD simulations and a novel machine learning analysis, we can identify the most important features involved at the TS of the unbinding process. Because the method is fully automated, it can be easily applied to any biological system of interest as well as to other type of processes.

The second work presented in this thesis shows how using multiple unbiased MD simulations and QM calculations we are able to obtain insight into the mechanism of inhibition of two molecules that target the HIV integrase protein. Thanks to the collaboration with experimentalists at the Francis Crick Institute, we used the coordinates obtained from high-resolution cryo-EM images as the starting point for our long unbiased MD simulations. From the analysis of these simulations, we were able to understand the reason behind the higher potency of the second-generation HIV integrase inhibitors compared to the first-generation drugs. Additionally, we showed how the presence of two known mutations that cause resistance to the first-generation HIV integrase inhibitors affects the affinity of the simulated molecules by disrupting the coordination of the drugs with the two magnesium ions present in the active site. The results obtained from the simulations allow us to better understand the interaction of HIV integrase inhibitors and will allow us to design novel more potent drugs to treat HIV.

The next work presented, shows how using a combination of computational and experimental works, both carried out by me, we are able to decipher the catalytic mechanism of DDL, an important target against *Mycobacterium Tuberculosis*, and to design a set of new molecules that inhibit the activity of the protein. By using QM/MM calculations and kinetic experiments on wt and single point mutants, we defined which residues are important in the active site of the protein. Using this knowledge, we then performed docking calculations

and MD simulations using as a library a set of purchasable compounds, testing in vitro our best candidates. Our results showed that four tested molecules present similar inhibition activity to D-cycloserine, a known FDA-approved drug in *tuberculosis*. The molecules with high inhibition activity will be further tested and used as scaffold for the design of novel inhibitors targeting DDL.

The fourth project presented in this thesis, started as a response to the ongoing pandemic of Covid19. The project aimed to present a reliable holo structure of the SARS-Covid-19 helicase, aiming to understand the catalytic mechanism of the protein and to provide an accurate initial structure for virtual screening. To gain key insights into the structure and dynamics of the complete holoenzyme in addition to the experimentally available apo protein, we modelled a fully assembled complex with both the ATP and ssRNA substrates. We identified highly conserved anchoring points in the core of the helicase for polynucleotide binding, which are essential to understand the translocation driving the unwinding activity of NSP13. We confirmed the stability of the conserved interactions through multiple unbiased MD simulations and compared the MD results while using multiple force fields. Furthermore, from the simulations generated, we were able to discover allosteric pockets that can be used for virtual screening allowing the design of possible inhibitors targeting this protein.

Lastly, in Chapter 7, I presented the efficacy of the DHAM method in the case of passive membrane permeation of seven known drugs. Here we showed an efficient method for calculating permeability coefficients using the data obtained from biased and unbiased MD simulations by calculating the relaxation time directly from the MSMs. Additionally, by introducing a 2D-DHAM analysis, we are able to describe the orientation of the ligands while crossing the membrane, showing how, according to the functional groups present in each molecule, the orientation of the drug changes along the

membrane crossing path. The method has been validated with experimental results, suggesting that it can be used with larger datasets, or different compositions of the lipid bilayer.

References

[1] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics, Vol. III: The new millennium edition: Quantum Mechanics*. Hachette UK, 2015.

[2] J. A. McCammon, B. R. Gelin, and M. Karplus, "Dynamics of folded proteins," *Nature*, vol. 267, no. 5612, pp. 585–590, 1977, doi: 10.1038/267585a0.

[3] J. Harvey, *Computational Chemistry*. Oxford University Press, 2018.

[4] B. J. Alder and T. E. Wainwright, "Phase transition for a hard sphere system," *The Journal of Chemical Physics*, vol. 27, no. 5. American Institute of PhysicsAIP, pp. 1208–1209, Nov. 13, 1957, doi: 10.1063/1.1743957.

[5] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," *Nature Structural Biology*, vol. 9, no. 9. pp. 646–652, 2002, doi: 10.1038/nsb0902-646.

[6] S. A. Hollingsworth and R. O. Dror, "Molecular Dynamics Simulation for All," *Neuron*, vol. 99, no. 6. Cell Press, pp. 1129–1143, Sep. 19, 2018, doi: 10.1016/j.neuron.2018.08.011.

[7] S. A. Adcock and J. A. McCammon, "Molecular dynamics: Survey of methods for simulating the activity of proteins," *Chemical Reviews*, vol. 106, no. 5. American Chemical Society , pp. 1589–1615, May 2006, doi: 10.1021/cr040426m.

[8] L. Verlet, "Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules," *Phys. Rev.*, vol. 159, no. 1, pp. 98–103, Jul. 1967, doi: 10.1103/PhysRev.159.98.

[9] W. F. Van Gunsteren and H. J. C. Berendsen, "A Leap-Frog Algorithm for Stochastic Dynamics," *Mol. Simul.*, vol. 1, no. 3, pp. 173–185, 1988, doi: 10.1080/08927028808080941.

[10] Y. Shao *et al.*, "Advances in molecular quantum chemistry contained in the Q-Chem 4 program package," *Mol. Phys.*, vol. 113, no. 2, pp. 184–215, Jan. 2015, doi: 10.1080/00268976.2014.952696.

[11] A. R. Leach and A. R. Leach, *Molecular modelling: principles and applications*. Pearson education, 2001.

[12] H. C. Andersen, "Rattle: A 'velocity' version of the shake algorithm for molecular dynamics calculations," *J. Comput. Phys.*, vol. 52, no. 1, pp. 24–34, Oct. 1983, doi:

10.1016/0021-9991(83)90014-1.

[13] W. D. Cornell *et al.*, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *J. Am. Chem. Soc.*, vol. 117, no. 19, pp. 5179–5197, May 1995, doi: 10.1021/ja00124a002.

[14] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1656–1676, Oct. 2004, doi: 10.1002/jcc.20090.

[15] W. L. Jorgensen and J. Tirado-Rives, "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin," *J. Am. Chem. Soc.*, vol. 110, no. 6, pp. 1657–1666, Mar. 1988, doi: 10.1021/ja00214a001.

[16] A. D. MacKerell *et al.*, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, Apr. 1998, doi: 10.1021/jp973084f.

[17] K. Vanommeslaeghe *et al.*, "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields," *J. Comput. Chem.*, vol. 31, no. 4, pp. 671–690, Mar. 2010, doi: 10.1002/jcc.21367.

[18] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general Amber force field," *J. Comput. Chem.*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004, doi: 10.1002/jcc.20035.

[19] P. P. Ewald, "Die Berechnung optischer und elektrostatischer Gitterpotentiale," *Ann. Phys.*, vol. 369, no. 3, pp. 253–287, Jan. 1921, doi: 10.1002/andp.19213690304.

[20] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," *J. Chem. Phys.*, vol. 103, no. 19, pp. 8577–8593, 1995.

[21] D. Frenkel and B. Smit, *Understanding molecular simulation: From algorithms to applications*. 1996.

[22] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, no. 1, p. 14101, 2007.

[23] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *J. Chem. Phys.*, vol. 72, no. 4, pp. 2384–2393, Feb. 1980, doi: 10.1063/1.439486.

[24] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," *J. Chem. Phys.*, vol. 81, no. 1, pp. 511–519, Aug. 1984, doi: 10.1063/1.447334.

[25] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, no. 3, pp. 1695–1697, Mar. 1985, doi: 10.1103/PhysRevA.31.1695.

[26] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. Dinola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Aug. 1984, doi: 10.1063/1.448118.

[27] D. S. Lemons and A. Gythiel, "Paul Langevin's 1908 paper 'On the Theory of Brownian Motion' ['Sur la théorie du mouvement brownien,' C. R. Acad. Sci. (Paris) 146 , 530–533 (1908)]," *Am. J. Phys.*, vol. 65, no. 11, pp. 1079–1081, Jun. 1997, doi: 10.1119/1.18725.

[28] H. J. C. C. Berendsen, J. P. M. M. van Postma, W. F. Van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, doi: 10.1063/1.448118.

[29] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–7190, 1981.

[30] G. J. Martyna, D. J. Tobias, and M. L. Klein, "Constant pressure molecular dynamics algorithms," *J. Chem. Phys.*, vol. 101, no. 5, pp. 4177–4189, Aug. 1994, doi: 10.1063/1.467468.

[31] Y. Song and E. A. Mason, "Statistical-mechanical basis for accurate analytical equations of state for fluids," *Fluid Phase Equilib.*, vol. 75, no. C, pp. 105–115, Aug. 1992, doi: 10.1016/0378-3812(92)87010-K.

[32] Y. Duan and P. A. Kollman, "Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution," *Science (80-.).*, vol. 282, no. 5389, pp. 740–744, Oct. 1998, doi: 10.1126/science.282.5389.740.

[33] D. E. Shaw *et al.*, "Atomic-level characterization of the structural dynamics of proteins," *Science (80-.).*, vol. 330, no. 6002, pp. 341–346, Oct. 2010, doi: 10.1126/science.1187409.

[34] C. Abrams and G. Bussi, "Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration," *Entropy*, vol. 16, no. 1, pp. 163–199, Dec. 2013, doi: 10.3390/e16010163.

[35] J. C. Phillips *et al.*, "Scalable molecular dynamics on CPU and GPU architectures with NAMD," *J. Chem. Phys.*, vol. 153, no. 4, p. 044130, Jul. 2020, doi: 10.1063/5.0014475.

[36] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, "Role of Molecular Dynamics and Related Methods in Drug Discovery," *Journal of Medicinal Chemistry*, vol. 59, no. 9. American Chemical Society, pp. 4035–4061, May 12, 2016, doi: 10.1021/acs.jmedchem.5b01684.

[37] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," *J. Comput. Phys.*, vol. 23, no. 2, pp. 187–199, Feb. 1977, doi: 10.1016/0021-9991(77)90121-8.

[38] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 20, pp. 12562–12566, Oct. 2002, doi: 10.1073/pnas.202427399.

[39] H. Fukunishi, O. Watanabe, and S. Takada, "On the Hamiltonian replica exchange

method for efficient sampling of biomolecular systems: Application to protein structure prediction," *J. Chem. Phys.*, vol. 116, no. 20, pp. 9058–9067, May 2002, doi: 10.1063/1.1472510.

[40] W. E, W. Ren, and E. Vanden-Eijnden, "Finite temperature string method for the study of rare events," *J. Phys. Chem. B*, vol. 109, no. 14, pp. 6688–6693, Apr. 2005, doi: 10.1021/jp0455430.

[41] J. D. Chodera and F. Noé, "Markov state models of biomolecular conformational dynamics," *Curr. Opin. Struct. Biol.*, vol. 25, pp. 135–144, 2014, doi: 10.1016/j.sbi.2014.04.002.

[42] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," *J. Comput. Chem.*, vol. 13, no. 8, pp. 1011–1021, Oct. 1992, doi: 10.1002/jcc.540130812.

[43] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased Markov state model," *J. Chem. Theory Comput.*, vol. 11, no. 1, pp. 276–285, Jan. 2015, doi: 10.1021/ct500719p.

[44] A. M. Ferrenberg and R. H. Swendsen, "New Monte Carlo technique for studying phase transitions," *Phys. Rev. Lett.*, vol. 61, no. 23, pp. 2635–2638, Dec. 1988, doi: 10.1103/PhysRevLett.61.2635.

[45] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Phys. Rev.*, vol. 136, no. 3B, p. B864, Nov. 1964, doi: 10.1103/PhysRev.136.B864.

[46] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," *Phys. Rev.*, vol. 140, no. 4A, p. A1133, Nov. 1965, doi: 10.1103/PhysRev.140.A1133.

[47] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *J. Chem. Phys.*, vol. 98, no. 7, pp. 5648–5652, Aug. 1993, doi: 10.1063/1.464913.

[48] A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme," *J. Mol. Biol.*, vol. 103, no. 2, pp. 227–249, May 1976, doi: 10.1016/0022-2836(76)90311-9.

[49] O. Barabás *et al.*, "Catalytic mechanism of α-phosphate attack in dUTPase is revealed by X-ray crystallographic snapshots of distinct intermediates, 31P-NMR spectroscopy and reaction path modelling," *Nucleic Acids Res.*, vol. 41, no. 22, pp. 10542–10555, Dec. 2013, doi: 10.1093/nar/gkt756.

[50] R. A. Copeland, *Evaluation of Enzyme Inhibitors in Drug Discovery: A Guide for Medicinal Chemists and Pharmacologists: Second Edition*. John Wiley and Sons, 2013.

[51] R. A. Copeland, "The drug-target residence time model: A 10-year retrospective," *Nature Reviews Drug Discovery*, vol. 15, no. 2. Nature Publishing Group, pp. 87–95, Feb. 03, 2016, doi: 10.1038/nrd.2015.18.

[52] M. Bernetti, M. Masetti, W. Rocchia, and A. Cavalli, "Kinetics of Drug Binding and Residence Time," *Annu. Rev. Phys. Chem. Annu. Rev. Phys. Chem. 2019*, vol. 70, pp. 143–171, Jun. 2019, doi: 10.1146/annurev-physchem-042018.

[53] R. A. Copeland, D. L. Pompliano, and T. D. Meek, "Drug–target residence time and its implications for lead optimization," *Nat. Rev. Drug Discov.*, vol. 5, no. 9, pp. 730–739, Sep. 2006, doi: 10.1038/nrd2082.

[54] H. Lu and P. J. Tonge, "Drug-target residence time: Critical information for lead optimization," *Current Opinion in Chemical Biology*, vol. 14, no. 4. NIH Public Access, pp. 467–474, Aug. 2010, doi: 10.1016/j.cbpa.2010.06.176.

[55] M. Bernetti, A. Cavalli, and L. Mollica, "Protein-ligand (un)binding kinetics as a new paradigm for drug discovery at the crossroad between experiments and modelling," *MedChemComm*, vol. 8, no. 3. Royal Society of Chemistry, pp. 534–550, Mar. 23, 2017, doi: 10.1039/c6md00581k.

[56] A. Ruiz-Garcia, M. Bermejo, A. Moss, and V. G. Casabo, "Pharmacokinetics in Drug Discovery," *J. Pharm. Sci.*, vol. 97, no. 2, pp. 654–690, Feb. 2008, doi: 10.1002/jps.21009.

[57] D. A. Schuetz *et al.*, "Kinetics for Drug Discovery: an industry-driven effort to target drug residence time," *Drug Discovery Today*, vol. 22, no. 6. Elsevier Ltd, pp. 896–911, Jun. 01, 2017, doi: 10.1016/j.drudis.2017.02.002.

[58] R. J. Darling and P. A. Brault, "Kinetic exclusion assay technology: Characterization of molecular interactions," *Assay and Drug Development Technologies*, vol. 2, no. 6. Mary Ann Liebert, Inc. 2 Madison Avenue Larchmont, NY 10538 USA , pp. 647–657, Dec. 27, 2004, doi: 10.1089/adt.2004.2.647.

[59] R. H. Rose, S. J. Briddon, and S. J. Hill, "A novel fluorescent histamine H 1 receptor antagonist demonstrates the advantage of using fluorescence correlation spectroscopy to study the binding of lipophilic ligands," *Br. J. Pharmacol.*, vol. 165, no. 6, pp. 1789–1800, Mar. 2012, doi: 10.1111/j.1476-5381.2011.01640.x.

[60] K. Herrick-Davis, E. Grinde, A. Cowan, and J. E. Mazurkiewicz, "Fluorescence correlation spectroscopy analysis of serotonin, adrenergic, muscarinic, and dopamine receptor dimerization: The oligomer number puzzle," *Mol. Pharmacol.*, vol. 84, no. 4, pp. 630–642, Oct. 2013, doi: 10.1124/mol.113.087072.

[61] N. J. Bruce, G. K. Ganotra, D. B. Kokh, S. K. Sadiq, and R. C. Wade, "New approaches for computing ligand–receptor binding kinetics," *Curr. Opin. Struct. Biol.*, vol. 49, pp. 1–10, Apr. 2018, doi: 10.1016/j.sbi.2017.10.001.

[62] D. J. Huggins *et al.*, "Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity," *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 9, no. 3, p. e1393, May 2019, doi: 10.1002/wcms.1393.

[63] D. Huang and A. Caflisch, "The Free Energy Landscape of Small Molecule Unbinding," *PLoS Comput. Biol.*, vol. 7, no. 2, p. e1002002, Feb. 2011, doi: 10.1371/journal.pcbi.1002002.

[64] S. Wolf, B. Lickert, S. Bray, and G. Stock, "Multisecond ligand dissociation

dynamics from atomistic simulations," *Nat. Commun.*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-020-16655-1.

[65] R. K. Mamidala, V. Ramana, S. G, M. Lingam, R. Gannu, and M. R. Yamsani, "Factors Influencing the Design and Performance of Oral Sustained/Controlled Release Dosage Forms," *Int. J. Pharm. Sci. Nanotechnol.*, vol. 2, no. 3, pp. 583–594, Nov. 2009, doi: 10.37285/ijpsn.2009.2.3.1.

[66] S. D. Lotz and A. Dickson, "Unbiased Molecular Dynamics of 11 min Timescale Drug Unbinding Reveals Transition State Stabilizing Interactions," *J. Am. Chem. Soc.*, vol. 140, no. 2, pp. 618–628, Jan. 2018, doi: 10.1021/jacs.7b08572.

[67] S. Haldar *et al.*, "A Multiscale Simulation Approach to Modeling Drug-Protein Binding Kinetics," *J. Chem. Theory Comput.*, vol. 14, no. 11, pp. 6093–6101, Nov. 2018, doi: 10.1021/acs.jctc.8b00687.

[68] L. W. Votapka, B. R. Jagger, A. L. Heyneman, and R. E. Amaro, "SEEKR: Simulation Enabled Estimation of Kinetic Rates, A Computational Tool to Estimate Molecular Kinetics and Its Application to Trypsin-Benzamidine Binding," *J. Phys. Chem. B*, vol. 121, no. 15, pp. 3597–3606, Apr. 2017, doi: 10.1021/acs.jpcb.6b09388.

[69] P. Tiwary, V. Limongelli, M. Salvalaglio, and M. Parrinello, "Kinetics of protein-ligand unbinding: Predicting pathways, rates, and rate-limiting steps," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 5, pp. E386–E391, Feb. 2015, doi: 10.1073/pnas.1424461112.

[70] A. Cavalli, A. Spitaleri, G. Saladino, and F. L. Gervasio, "Investigating drug-target association and dissociation mechanisms using metadynamics-based algorithms," *Accounts of Chemical Research*, vol. 48, no. 2. American Chemical Society, pp. 277–285, Feb. 17, 2015, doi: 10.1021/ar500356n.

[71] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," *J. Chem. Phys.*, vol. 126, no. 5, p. 054103, Feb. 2007, doi: 10.1063/1.2432340.

[72] J. Fidelak, J. Juraszek, D. Branduardi, M. Bianciotto, and F. L. Gervasio, "Free-energy-based methods for binding profile determination in a congeneric series of CDK2 inhibitors," *J. Phys. Chem. B*, vol. 114, no. 29, pp. 9516–9524, Jul. 2010, doi: 10.1021/jp911689r.

[73] D. A. Schuetz *et al.*, "Predicting Residence Time and Drug Unbinding Pathway through Scaled Molecular Dynamics," *J. Chem. Inf. Model.*, vol. 59, no. 1, pp. 535–549, Jan. 2019, doi: 10.1021/acs.jcim.8b00614.

[74] A. M. Capelli and G. Costantino, "Unbinding pathways of VEGFR2 inhibitors revealed by steered molecular dynamics," *J. Chem. Inf. Model.*, vol. 54, no. 11, pp. 3124–3136, Nov. 2014, doi: 10.1021/ci500527j.

[75] Y. Niu, S. Li, D. Pan, H. Liu, and X. Yao, "Computational study on the unbinding pathways of B-RAF inhibitors and its implication for the difference of residence time: insight from random acceleration and steered molecular dynamics simulations.," *Phys. Chem. Chem. Phys.*, vol. 18, no. 7, pp. 5622–9, Feb. 2016, doi: 10.1039/c5cp06257h.

[76] R. Casasnovas, V. Limongelli, P. Tiwary, P. Carloni, and M. Parrinello, "Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations," *J. Am. Chem. Soc.*, vol. 139, no. 13, pp. 4780–4788, Apr. 2017, doi: 10.1021/jacs.6b12950.

[77] S. K. Lüdemann, V. Lounnas, and R. C. Wade, "How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms," *J. Mol. Biol.*, vol. 303, no. 5, pp. 797–811, Nov. 2000, doi: 10.1006/jmbi.2000.4154.

[78] D. B. Kokh *et al.*, "Estimation of Drug-Target Residence Times by τ-Random Acceleration Molecular Dynamics Simulations," *J. Chem. Theory Comput.*, vol. 14, no. 7, pp. 3859–3869, Jul. 2018, doi: 10.1021/acs.jctc.8b00230.

[79] J. Juraszek, G. Saladino, T. S. Van Erp, and F. L. Gervasio, "Efficient numerical reconstruction of protein folding kinetics with partial path sampling and pathlike variables," *Phys. Rev. Lett.*, vol. 110, no. 10, p. 108106, Mar. 2013, doi: 10.1103/PhysRevLett.110.108106.

[80] M. A. Morando *et al.*, "Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase," *Sci. Rep.*, vol. 6, Apr. 2016, doi: 10.1038/srep24439.

[81] R. Evans, L. Hovan, G. A. Tribello, B. P. Cossins, C. Estarellas, and F. L. Gervasio, "Combining Machine Learning and Enhanced Sampling Techniques for Efficient and Accurate Calculation of Absolute Binding Free Energies," *J. Chem. Theory Comput.*, vol. 16, no. 7, pp. 4641–4654, Jul. 2020, doi: 10.1021/acs.jctc.0c00075.

[82] J. Rydzewski and O. Valsson, "Finding multiple reaction pathways of ligand unbinding," *J. Chem. Phys.*, vol. 150, no. 22, p. 221101, Jun. 2019, doi: 10.1063/1.5108638.

[83] E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, "Constrained reaction coordinate dynamics for the simulation of rare events," *Chem. Phys. Lett.*, vol. 156, no. 5, pp. 472–477, Apr. 1989, doi: 10.1016/S0009-2614(89)87314-2.

[84] L. Hovan, F. Comitani, and F. L. Gervasio, "Defining an Optimal Metric for the Path Collective Variables," *J. Chem. Theory Comput.*, vol. 15, no. 1, pp. 25–32, Jan. 2019, doi: 10.1021/acs.jctc.8b00563.

[85] E. Rosta, M. Nowotny, W. Yang, and G. Hummer, "Catalytic mechanism of RNA backbone cleavage by ribonuclease H from quantum mechanics/molecular mechanics simulations," *J. Am. Chem. Soc.*, vol. 133, no. 23, pp. 8934–8941, Jun. 2011, doi: 10.1021/ja200173a.

[86] H. Jung, R. Covino, and G. Hummer, "Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations," arXiv, Jan. 2019. Accessed: Feb. 25, 2021. [Online]. Available: http://arxiv.org/abs/1901.04595.

[87] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," *Annu. Rev. Phys. Chem.*, vol. 71, pp. 361–390, Nov. 2019,

doi: 10.1146/annurev-physchem-042018-052331.

[88] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, "Unsupervised Learning Methods for Molecular Simulation Data," *Chemical Reviews*. American Chemical Society, 2021, doi: 10.1021/acs.chemrev.0c01195.

[89] J. B. Dunbar *et al.*, "CSAR data set release 2012: Ligands, affinities, complexes, and docking decoys," *J. Chem. Inf. Model.*, vol. 53, no. 8, pp. 1842–1852, Aug. 2013, doi: 10.1021/ci4000486.

[90] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, Jul. 2015, doi: 10.1021/acs.jctc.5b00255.

[91] M. J. Frisch *et al.*, "Gaussian 09 Revision E." Gaussian, Inc., Wallingford CT, 2016, [Online]. Available: http://gaussian.com/.

[92] J. C. Phillips *et al.*, *Scalable molecular dynamics with NAMD*, vol. 26, no. 16. John Wiley and Sons Inc., 2005, pp. 1781–1802.

[93] I. Lans *et al.*, "Theoretical study of the mechanism of the hydride transfer between ferredoxin-NADP+ reductase and NADP+: The role of Tyr303," *J. Am. Chem. Soc.*, vol. 134, no. 50, pp. 20544–20553, Dec. 2012, doi: 10.1021/ja310331v.

[94] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[95] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure-activity relationships," *J. Chem. Inf. Model.*, vol. 55, no. 2, pp. 263–274, Feb. 2015, doi: 10.1021/ci500747n.

[96] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines."

[97] H. Wang *et al.*, "Exploring the Reaction Mechanism of HIV Reverse Transcriptase with a Nucleotide Substrate," *J. Phys. Chem. B*, vol. 124, no. 21, pp. 4270–4283, May 2020, doi: 10.1021/acs.jpcb.0c02632.

[98] V. Ovchinnikov, M. Karplus, and E. Vanden-Eijnden, "Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI," *J. Chem. Phys.*, vol. 134, no. 8, p. 85103, Feb. 2011, doi: 10.1063/1.3544209.

[99] Y. Li *et al.*, "Insights on structural characteristics and ligand binding mechanisms of CDK2," *International Journal of Molecular Sciences*, vol. 16, no. 5. MDPI AG, pp. 9314–9340, Apr. 23, 2015, doi: 10.3390/ijms16059314.

[100] K. Anstett, B. Brenner, T. Mesplede, and M. A. Wainberg, "HIV drug resistance against strand transfer integrase inhibitors," *Retrovirology*, vol. 14, no. 1. BioMed Central Ltd., p. 36, Jun. 05, 2017, doi: 10.1186/s12977-017-0360-7.

[101] B. A. Johns *et al.*, "Carbamoyl pyridone HIV-1 integrase inhibitors 3. A

diastereomeric approach to chiral nonracemic tricyclic ring systems and the discovery of dolutegravir (S/GSK1349572) and (S/GSK1265744)," *J. Med. Chem.*, vol. 56, no. 14, pp. 5901–5916, Jul. 2013, doi: 10.1021/jm400645w.

[102] M. Oliveira *et al.*, "Selective resistance profiles emerging in patient-derived clinical isolates with cabotegravir, bictegravir, dolutegravir, and elvitegravir," *Retrovirology*, vol. 15, no. 1, p. 56, Aug. 2018, doi: 10.1186/s12977-018-0440-3.

[103] S. J. Smith, X. Z. Zhao, T. R. Burke, and S. H. Hughes, "Efficacies of Cabotegravir and Bictegravir against drug-resistant HIV-1 integrase mutants," *Retrovirology*, vol. 15, no. 1, p. 37, May 2018, doi: 10.1186/s12977-018-0420-7.

[104] H. T. Pham *et al.*, "The s230r integrase substitution associated with virus load rebound during dolutegravir monotherapy confers low-level resistance to integrase strand-transfer inhibitors," *J. Infect. Dis.*, vol. 218, no. 5, pp. 698–706, Jul. 2018, doi: 10.1093/infdis/jiy175.

[105] I. E. A. Wijting *et al.*, "HIV-1 resistance dynamics in patients with virologic failure to dolutegravir maintenance monotherapy," *J. Infect. Dis.*, vol. 218, no. 5, pp. 688–697, Jul. 2018, doi: 10.1093/infdis/jiy176.

[106] W. W. Zhang, P. K. Cheung, N. Oliveira, M. A. Robbins, P. Richard Harrigan, and A. Shahid, "Accumulation of multiple mutations in vivo confers cross-resistance to new and existing integrase inhibitors," *J. Infect. Dis.*, vol. 218, no. 11, pp. 1773–1776, Oct. 2018, doi: 10.1093/infdis/jiy428.

[107] S. Hare, S. S. Gupta, E. Valkov, A. Engelman, and P. Cherepanov, "Retroviral intasome assembly and inhibition of DNA strand transfer," *Nature*, vol. 464, no. 7286, pp. 232–236, Mar. 2010, doi: 10.1038/nature08784.

[108] D. O. Passos *et al.*, "Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome," *Science (80-.).*, vol. 355, no. 6320, pp. 89–92, Jan. 2017, doi: 10.1126/science.aah5163.

[109] S. Ahuka-Mundeke *et al.*, "Full-length genome sequence of a simian immunodeficiency virus (SIV) infecting a captive agile mangabey (Cercocebus agilis) is closely related to SIVrcm infecting wild red-capped mangabeys (Cercocebus torquatus) in Cameroon," *J. Gen. Virol.*, vol. 91, no. 12, pp. 2959–2964, 2010, doi: 10.1099/vir.0.025767-0.

[110] P. M. Sharp, G. M. Shaw, and B. H. Hahn, "Simian Immunodeficiency Virus Infection of Chimpanzees," *J. Virol.*, vol. 79, no. 7, pp. 3891–3902, Apr. 2005, doi: 10.1128/jvi.79.7.3891-3902.2005.

[111] P. Cherepanov, "LEDGF/p75 interacts with divergent lentiviral integrases and modulates their enzymatic activity in vitro," *Nucleic Acids Res.*, vol. 35, no. 1, pp. 113–124, Jan. 2007, doi: 10.1093/nar/gkl885.

[112] S. Hare, M. C. Shun, S. S. Gupta, E. Valkov, A. Engelman, and P. Cherepanov, "A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75," *PLoS Pathog.*, vol. 5, no. 1, pp. e1000259–e1000259, Jan. 2009, doi: 10.1371/journal.ppat.1000259.

[113] A. Ballandras-Colas *et al.*, "A supramolecular assembly mediates lentiviral DNA integration," *Science (80-.).*, vol. 355, no. 6320, pp. 93–95, Jan. 2017, doi: 10.1126/science.aah7002.

[114] A. S. Espeseth *et al.*, "HIV-1 integrase inhibitors that compete with the target DNA substrate define a unique strand transfer conformation for integrase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 21, pp. 11244–11249, Oct. 2000, doi: 10.1073/pnas.200139397.

[115] W. M. Konsavage, S. Burkholder, M. Sudol, A. L. Harper, and M. Katzman, "A Substitution in Rous Sarcoma Virus Integrase That Separates Its Two Biologically Relevant Enzymatic Activities," *J. Virol.*, vol. 79, no. 8, pp. 4691–4699, Apr. 2005, doi: 10.1128/jvi.79.8.4691-4699.2005.

[116] M. G. Nowak, M. Sudol, N. E. Lee, W. M. Konsavage, and M. Katzman, "Identifying amino acid residues that contribute to the cellular-DNA binding site on retroviral integrase," *Virology*, vol. 389, no. 1–2, pp. 141–148, 2009, doi: 10.1016/j.virol.2009.04.014.

[117] G. N. Maertens, S. Hare, and P. Cherepanov, "The mechanism of retroviral integration from X-ray structures of its key intermediates," *Nature*, vol. 468, no. 7321, pp. 326–329, Nov. 2010, doi: 10.1038/nature09517.

[118] R. W. Shafer, "Rationale and uses of a public HIV drug-resistance database," in *Journal of Infectious Diseases*, Sep. 2006, vol. 194, no. SUPPL. 1, doi: 10.1086/505356.

[119] J. M. George *et al.*, "Rapid Development of High-Level Resistance to Dolutegravir with Emergence of T97A Mutation in 2 Treatment-Experienced Individuals with Baseline Partial Sensitivity to Dolutegravir," *Open Forum Infect. Dis.*, vol. 5, no. 10, Oct. 2018, doi: 10.1093/ofid/ofy221.

[120] K. E. Hightower *et al.*, "Dolutegravir (S/GSK1349572) exhibits significantly slower dissociation than raltegravir and elvitegravir from wild-type and integrase inhibitor-resistant HIV-1 integrase-DNA complexes," *Antimicrob. Agents Chemother.*, vol. 55, no. 10, pp. 4552–4559, Oct. 2011, doi: 10.1128/AAC.00157-11.

[121] D. Blow, "More of the catalytic triad," *Nature*, vol. 343, no. 6260. pp. 694–695, 1990, doi: 10.1038/343694a0.

[122] J. C. Marx, J. Poncin, J. P. Simorre, P. W. Ramteke, and G. Feller, "The noncatalytic triad of α-amylases: A novel structural motif involved in conformational stability," *Proteins Struct. Funct. Genet.*, vol. 70, no. 2, pp. 320–328, Feb. 2008, doi: 10.1002/prot.21594.

[123] M. E. Maguire and J. A. Cowan, "Magnesium chemistry and biochemistry," *BioMetals*, vol. 15, no. 3. pp. 203–210, 2002, doi: 10.1023/A:1016058229972.

[124] M. M. Harding, "Geometry of metal-ligand interactions in proteins," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 57, no. 3, pp. 401–411, 2001, doi: 10.1107/S0907444900019168.

[125] C. M. Breneman and K. B. Wiberg, "Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis," *J. Comput. Chem.*, vol. 11, no. 3, pp. 361–373, Apr. 1990, doi: 10.1002/jcc.540110311.

[126] J. Da Chai and M. Head-Gordon, "Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections," *Phys. Chem. Chem. Phys.*, vol. 10, no. 44, pp. 6615–6620, Nov. 2008, doi: 10.1039/b810189b.

[127] F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy," *Phys. Chem. Chem. Phys.*, vol. 7, no. 18, pp. 3297–3305, Sep. 2005, doi: 10.1039/b508541a.

[128] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, "Constant pressure molecular dynamics simulation: The Langevin piston method," *J. Chem. Phys.*, vol. 103, no. 11, pp. 4613–4621, Jun. 1995, doi: 10.1063/1.470648.

[129] T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log (N) method for Ewald sums in large systems," *J. Chem. Phys.*, vol. 98, no. 12, pp. 10089–10092, 1993.

[130] R. B. Best *et al.*, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 Dihedral Angles," *J. Chem. Theory Comput.*, vol. 8, no. 9, pp. 3257–3273, Sep. 2012, doi: 10.1021/ct300400x.

[131] M. M. Francl *et al.*, "Self-consistent molecular orbital methods. XXIII. A polarization-type basis set for second-row elements," *J. Chem. Phys.*, vol. 77, no. 7, pp. 3654–3665, 1982, doi: 10.1063/1.444267.

[132] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, Jul. 2009, doi: 10.1002/jcc.21287.

[133] C. M. Smith and G. G. Hall, "The approximation of electron densities," *Theor. Chim. Acta*, vol. 69, no. 1, pp. 63–69, Jan. 1986, doi: 10.1007/BF00526293.

[134] A. E. Reed, R. B. Weinstock, and F. Weinhold, "Natural population analysis," *J. Chem. Phys.*, vol. 83, no. 2, pp. 735–746, 1985, doi: 10.1063/1.449486.

[135] A. E. Reed, L. A. Curtiss, and F. Weinhold, "Intermolecular Interactions from a Natural Bond Orbital, Donor—Acceptor Viewpoint," *Chem. Rev.*, vol. 88, no. 6, pp. 899–926, Sep. 1988, doi: 10.1021/cr00088a005.

[136] S. R. Kimura, H. P. Hu, A. M. Ruvinsky, W. Sherman, and A. D. Favia, "Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics," *J. Chem. Inf. Model.*, vol. 57, no. 6, pp. 1388–1401, Jun. 2017, doi: 10.1021/acs.jcim.6b00623.

[137] M. Pai *et al.*, "Tuberculosis," *Nature Reviews Disease Primers*, vol. 2, no. 1. Nature Publishing Group, pp. 1–23, Oct. 27, 2016, doi: 10.1038/nrdp.2016.76.

[138] G. D. Wright and C. T. Walsh, "D-Alanyl-D-alanine Ligases and the Molecular Mechanism of Vancomycin Resistance," 1992. Accessed: Nov. 27, 2020. [Online].

Available: https://pubs.acs.org/sharingguidelines.

[139] M. V. Fawaz, M. E. Topper, and S. M. Firestine, "The ATP-grasp enzymes," *Bioorg. Chem.*, vol. 39, no. 5–6, pp. 185–191, 2011, doi: 10.1016/j.bioorg.2011.08.004.

[140] J. L. Pederick, A. P. Thompson, S. G. Bell, and J. B. Bruning, "D-alanine–D-alanine ligase as a model for the activation of ATP-grasp enzymes by monovalent cations," *J. Biol. Chem.*, vol. 295, no. 23, pp. 7894–7904, Apr. 2020, doi: 10.1074/JBC.RA120.012936.

[141] Y. Li *et al.*, "Cycloserine for treatment of multidrug-resistant tuberculosis: A retrospective cohort study in China," *Infect. Drug Resist.*, vol. 12, pp. 721–731, 2019, doi: 10.2147/IDR.S195555.

[142] S. Halouska, R. J. Fenton, D. K. Zinniel, D. D. Marshall, R. G. Barletta, and R. Powers, "Metabolomics analysis identifies d-alanine-d-alanine ligase as the primary lethal target of d-cycloserine in mycobacteria," *J. Proteome Res.*, vol. 13, no. 2, pp. 1065–1076, Feb. 2014, doi: 10.1021/pr4010579.

[143] G. A. Prosser and L. P. S. Carvalho, "Kinetic mechanism and inhibition of *M ycobacterium tuberculosis* d-alanine: d-alanine ligase by the antibiotic d-cycloserine," *FEBS J.*, vol. 280, no. 4, pp. 1150–1166, Feb. 2013, doi: 10.1111/febs.12108.

[144] M. Hrast, B. Vehar, S. Turk, J. Konc, S. Gobec, and D. Janežič, "Function of the D -alanine: D -alanine ligase lid loop: A molecular modeling and bioactivity study," *J. Med. Chem.*, vol. 55, no. 15, pp. 6849–6856, Aug. 2012, doi: 10.1021/jm3006965.

[145] V. Škedelj *et al.*, "6-Arylpyrido[2,3-d]pyrimidines as novel ATP-competitive inhibitors of bacterial D-Alanine:D-Alanine ligase," *PLoS One*, vol. 7, no. 8, p. e39922, Aug. 2012, doi: 10.1371/journal.pone.0039922.

[146] S. Liu *et al.*, "Allosteric inhibition of Staphylococcus aureus D-alanine:D-alanine ligase revealed by crystallographic studies," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 41, pp. 15178–15183, Oct. 2006, doi: 10.1073/pnas.0604905103.

[147] W. H. Parsons *et al.*, "Phosphinic Acid Inhibitors of D-Alanyl-D-Alanine Ligase," *J. Med. Chem.*, vol. 31, no. 9, pp. 1772–1778, Sep. 1988, doi: 10.1021/jm00117a017.

[148] A. Kovač *et al.*, "Discovery of new inhibitors of D-alanine:D-alanine ligase by structure-based virtual screening," *J. Med. Chem.*, vol. 51, no. 23, pp. 7442–7448, Dec. 2008, doi: 10.1021/jm800726b.

[149] A. Kovač *et al.*, "Diazenedicarboxamides as inhibitors of d-alanine-d-alanine ligase (Ddl)," *Bioorganic Med. Chem. Lett.*, vol. 17, no. 7, pp. 2047–2054, Apr. 2007, doi: 10.1016/j.bmcl.2007.01.015.

[150] L. S. Mullins, L. E. Zawadzke, C. T. Walsh, and F. M. Raushel, "Kinetic evidence for the formation of D-alanyl phosphate in the mechanism of D-alanyl-D-alanine ligase," *J. Biol. Chem.*, vol. 265, no. 16, pp. 8993–8998, 1990, Accessed: Jul. 08, 2020. [Online]. Available: https://www.jbc.org/content/265/16/8993.short.

[151] Y. Shi and C. T. Walsh, "Active Site Mapping of Escherichia coli D-Ala-D-Ala Ligase by Structure-Based Mutagenesis," *Biochemistry*, vol. 34, no. 9, pp. 2768–2776, 1995, doi: 10.1021/bi00009a005.

[152] N. Fernandez-Fuentes, C. J. Madrid-Aliste, B. K. Rai, J. E. Fajardo, and A. Fiser, "M4T: A comparative protein structure modeling server," *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, p. W363, Jul. 2007, doi: 10.1093/nar/gkm341.

[153] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," *J. Comput. Chem.*, vol. 29, no. 11, pp. 1859–1865, Aug. 2008, doi: 10.1002/jcc.20945.

[154] J. Huang and A. D. Mackerell, "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data," *J. Comput. Chem.*, vol. 34, no. 25, pp. 2135–2145, Sep. 2013, doi: 10.1002/jcc.23354.

[155] A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange," *J. Chem. Phys.*, vol. 98, no. 7, pp. 5648–5652, 1993, doi: 10.1063/1.464913.

[156] H. L. Woodcock, M. Hodošček, A. T. B. Gilbert, P. M. W. Gill, H. F. Schaefer, and B. R. Brooks, "Interfacing Q-Chem and CHARMM to perform QM/MM reaction path calculations," *J. Comput. Chem.*, vol. 28, no. 9, pp. 1485–1502, Jul. 2007, doi: 10.1002/jcc.20587.

[157] R. A. Friesner *et al.*, "Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes," *J. Med. Chem.*, vol. 49, no. 21, pp. 6177–6196, Oct. 2006, doi: 10.1021/jm051256o.

[158] K. J. Bowers *et al.*, "Scalable algorithms for molecular dynamics simulations on commodity clusters," *SC '06 Proc. 2006 ACM/IEEE Conf. Supercomput.*, 2006, Accessed: Dec. 06, 2020. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.2121.

[159] W. L. Jorgensen and J. Tirado-Rives, "The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin," *J. Am. Chem. Soc.*, vol. 110, no. 6, pp. 1657–1666, 1988, doi: 10.1021/ja00214a001.

[160] A. Warshel, G. Naray-Szabo, F. Sussman, and J. K. Hwang, "How do serine proteases really work?," *Biochemistry*, vol. 28, no. 9, pp. 3629–3637, May 1989, doi: 10.1021/bi00435a001.

[161] G. A. Prosser and L. P. S. De Carvalho, "Kinetic mechanism and inhibition of Mycobacterium tuberculosis d-alanine: D-alanine ligase by the antibiotic d-cycloserine," *FEBS J.*, vol. 280, no. 4, pp. 1150–1166, Feb. 2013, doi: 10.1111/febs.12108.

[162] "U.S. Food and Drug Administration: Coronavirus Disease 2019 (COVID-19) EUA Information," *U.S. Food and Drug Administration*, 2021. https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization#coviddrugs.

[163] EMA, "First COVID-19 treatment recommended for EU authorisation," 2020.

https://www.ema.europa.eu/en/news/first-covid-19-treatment-recommended-eu-authorisation.

[164] K. A. Ivanov, V. Thiel, J. C. Dobbe, Y. van der Meer, E. J. Snijder, and J. Ziebuhr, "Multiple Enzymatic Activities Associated with Severe Acute Respiratory Syndrome Coronavirus Helicase," *J. Virol.*, vol. 78, no. 11, pp. 5619–5632, Jun. 2004, doi: 10.1128/JVI.78.11.5619-5632.2004.

[165] J. A. Tanner *et al.*, "The severe acute respiratory syndrome (SARS) coronavirus NTPase/helicase belongs to a distinct class of 5′ to 3′ viral helicases," *J. Biol. Chem.*, vol. 278, no. 41, pp. 39578–39582, Oct. 2003, doi: 10.1074/jbc.C300328200.

[166] A. D. Kwong, B. G. Rao, and K. T. Jeang, "Viral and cellular RNA helicases as antiviral targets," *Nat. Rev. Drug Discov.*, vol. 4, no. 10, pp. 845–853, 2005, doi: 10.1038/nrd1853.

[167] J. Chen *et al.*, "Structural basis for helicase-polymerase coupling in the SARS-CoV-2 replication-transcription complex," *Cell*, vol. 182, no. 6, pp. 1560–1573, Jul. 2020, doi: 10.1016/j.cell.2020.07.033.

[168] Q. Peng *et al.*, "Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2," *Cell Rep.*, vol. 31, no. 11, p. 107774, Jun. 2020, doi: 10.1016/j.celrep.2020.107774.

[169] W. Yin *et al.*, "Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir," *Science (80-.).*, vol. 368, no. 6498, pp. 1499–1504, Jun. 2020, doi: 10.1126/science.abc1560.

[170] Z. Jia *et al.*, "Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis," *Nucleic Acids Res.*, vol. 47, no. 12, pp. 6538–6550, Jul. 2019, doi: 10.1093/nar/gkz409.

[171] M. Hoffmann *et al.*, "Three dimensional model of severe acute respiratory syndrome coronavirus helicase ATPase catalytic domain and molecular design of severe acute respiratory syndrome coronavirus helicase inhibitors," *J. Comput. Aided. Mol. Des.*, vol. 20, no. 5, pp. 305–319, May 2006, doi: 10.1007/s10822-006-9057-z.

[172] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, Oct. 1990, doi: 10.1016/S0022-2836(05)80360-2.

[173] A. S. Konagurthu, J. C. Whisstock, P. J. Stuckey, and A. M. Lesk, "MUSTANG: A Multiple Structural Alignment Algorithm."

[174] S. Chakrabarti *et al.*, "Molecular Mechanisms for the RNA-Dependent ATPase Activity of Upf1 and Its Regulation by Upf2," *Mol. Cell*, vol. 41, no. 6, pp. 693–703, Mar. 2011, doi: 10.1016/j.molcel.2011.02.010.

[175] Y.-S. Law *et al.*, "Structural insights into RNA recognition by the Chikungunya virus nsP2 helicase," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 19, pp. 9558–9567, May 2019, doi: 10.1073/pnas.1900656116.

[176] J. C. Phillips *et al.*, "Scalable molecular dynamics on CPU and GPU architectures with NAMD," *J. Chem. Phys.*, vol. 153, no. 4, p. 044130, Jul. 2020, doi: 10.1063/5.0014475.

[177] A. D. MacKerell *et al.*, "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins," *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3586–3616, Apr. 1998, doi: 10.1021/jp973084f.

[178] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *J. Chem. Phys.*, vol. 79, no. 2, pp. 926–935, Jul. 1983, doi: 10.1063/1.445869.

[179] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: fast, flexible, and free," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1701–1718, 2005.

[180] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, 2008.

[181] S. Pronk *et al.*, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.," *Bioinformatics*, vol. 29, no. 7, pp. 845–854, 2013, doi: 10.1093/bioinformatics/btt055.

[182] M. J. Abraham *et al.*, "Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, 2015, doi: 10.1016/j.softx.2015.06.001.

[183] A. Pérez *et al.*, "Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers," *Biophys. J.*, 2007, doi: 10.1529/biophysj.106.097782.

[184] M. Zgarbová *et al.*, "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles," *J. Chem. Theory Comput.*, vol. 7, no. 9, pp. 2886–2902, Aug. 2011, doi: 10.1021/ct200162x.

[185] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.," *J Chem Theory Comput*, vol. 11, no. 8, pp. 3696–3713, 2015, doi: 10.1021/acs.jctc.5b00255.

[186] M. B. Peters, Y. Yang, B. Wang, L. Füsti-Molnár, M. N. Weaver, and K. M. Merz, "Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF)," *J. Chem. Theory Comput.*, vol. 6, no. 9, pp. 2935–2947, Aug. 2010, doi: 10.1021/ct1002626.

[187] I. S. Joung and T. E. Cheatham, "Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations," *J. Phys. Chem. B*, vol. 112, no. 30, pp. 9020–9041, Jul. 2008, doi: 10.1021/jp8001614.

[188] I. Beššeová, M. Otyepka, K. Réblová, and J. Šponer, "Dependence of A-RNA simulations on the choice of the force field and salt strength," *Phys. Chem.*

Chem. Phys., vol. 11, no. 45, p. 10701, Nov. 2009, doi: 10.1039/b911169g.

[189] O. Allnér, L. Nilsson, and A. Villa, "Magnesium ion--water coordination and exchange in biomolecular simulations," *J. Chem. Theory Comput.*, vol. 8, no. 4, pp. 1493–1502, 2012.

[190] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: a linear constraint solver for molecular simulations," *J. Comput. Chem.*, vol. 18, no. 12, pp. 1463–1472, 1997.

[191] H. J. C. Berendsen, J. P. M. van Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *J. Chem. Phys.*, vol. 81, no. 8, pp. 3684–3690, 1984.

[192] S. C. Harvey, R. K.-Z.-. Z. Tan, and T. E. Cheatham, "The flying ice cube: velocity rescaling in molecular dynamics leads to violation of energy equipartition," *J. Comput. Chem.*, vol. 19, no. 7, pp. 726–740, 1998.

[193] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB," *J. Chem. Theory Comput.*, vol. 11, no. 8, pp. 3696–3713, Aug. 2015, doi: 10.1021/acs.jctc.5b00255.

[194] A. Pérez *et al.*, "Refinement of the AMBER force field for nucleic acids: Improving the description of α/γ conformers," *Biophys. J.*, vol. 92, no. 11, pp. 3817–3829, 2007, doi: 10.1529/biophysj.106.097782.

[195] M. Zgarbová *et al.*, "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles," *J. Chem. Theory Comput.*, vol. 7, no. 9, pp. 2886–2902, Sep. 2011, doi: 10.1021/ct200162x.

[196] K. L. Meagher, L. T. Redman, and H. A. Carlson, "Development of polyphosphate parameters for use with the AMBER force field," *J. Comput. Chem.*, vol. 24, no. 9, pp. 1016–1025, Jul. 2003, doi: 10.1002/jcc.10262.

[197] O. Allnér, L. Nilsson, and A. Villa, "Magnesium Ion–Water Coordination and Exchange in Biomolecular Simulations," *J. Chem. Theory Comput.*, vol. 8, no. 4, pp. 1493–1502, Apr. 2012, doi: 10.1021/ct3000734.

[198] P. Li and K. M. Merz Jr., "ZAFF Modeling Tutorial," 2015. https://ambermd.org/tutorials/advanced/tutorial20/ZAFF.htm (accessed Oct. 29, 2020).

[199] E. F. Pettersen *et al.*, "UCSF Chimera - A visualization system for exploratory research and analysis," *J. Comput. Chem.*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.

[200] J. H. G. Lebbink, A. Fish, A. Reumer, G. Natrajan, H. H. K. Winterwerp, and T. K. Sixma, "Magnesium coordination controls the molecular switch function of DNA mismatch repair protein MutS," *J. Biol. Chem.*, vol. 285, no. 17, pp. 13131–13141, Apr. 2010, doi: 10.1074/jbc.M109.066001.

[201] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: An open source platform

for ligand pocket detection," *BMC Bioinformatics*, vol. 10, no. 1, p. 168, May 2009, doi: 10.1186/1471-2105-10-168.

[202] B. Leonaité *et al.*, "Sen1 has unique structural features grafted on the architecture of the Upf1-like helicase family," *EMBO J.*, vol. 36, pp. 1590–1604, 2017, doi: 10.15252/embj.201696174.

[203] I. Briguglio, S. Piras, P. Corona, and A. Carta, "Inhibition of RNA Helicases of ssRNA + Virus Belonging to Flaviviridae, Coronaviridae and Picornaviridae Families," *Int. J. Med. Chem.*, vol. 2011, p. 22, 2011, doi: 10.1155/2011/213135.

[204] X. Yang *et al.*, "Mechanism of ATP hydrolysis by the Zika virus helicase," *FASEB J.*, vol. 32, no. 10, pp. 5250–5257, 2018, doi: 10.1096/fj.201701140R.

[205] T. C. Appleby *et al.*, "Visualizing ATP-dependent RNA translocation by the NS3 helicase from HCV," *J. Mol. Biol.*, vol. 405, no. 5, pp. 1139–1153, Feb. 2011, doi: 10.1016/j.jmb.2010.11.034.

[206] D. Schuster, C. Laggner, and T. Langer, "Why Drugs Fail - A Study on Side Effects in New Chemical Entities," *Curr. Pharm. Des.*, vol. 11, no. 27, pp. 3545–3559, Oct. 2005, doi: 10.2174/138161205774414510.

[207] D. Smith *et al.*, "Passive lipoidal diffusion and carrier-mediated cell uptake are both important mechanisms of membrane permeation in drug disposition," *Molecular Pharmaceutics*, vol. 11, no. 6. American Chemical Society, pp. 1727–1738, Jun. 02, 2014, doi: 10.1021/mp400713v.

[208] A. Avdeef, *Transport Model*. 2012.

[209] R. V. Swift and R. E. Amaro, "Back to the Future: Can Physical Models of Passive Membrane Permeability Help Reduce Drug Candidate Attrition and Move Us Beyond QSPR?," *Chem. Biol. Drug Des.*, vol. 81, no. 1, pp. 61–71, Jan. 2013, doi: 10.1111/cbdd.12074.

[210] A. Finkelstein, "Water and nonelectrolyte permeability of lipid bilayer membranes," *J. Gen. Physiol.*, vol. 68, no. 2, pp. 127–135, Aug. 1976, doi: 10.1085/jgp.68.2.127.

[211] D. Sezer and T. Oruç, "Protonation Kinetics Compromise Liposomal Fluorescence Assay of Membrane Permeation," *J. Phys. Chem. B*, vol. 121, no. 20, pp. 5218–5227, May 2017, doi: 10.1021/acs.jpcb.7b01881.

[212] K. F. Hermann *et al.*, "Kinetics of lipid bilayer permeation of a series of ionisable drugs and their correlation with human transporter-independent intestinal permeability," *Eur. J. Pharm. Sci.*, vol. 104, pp. 150–161, Jun. 2017, doi: 10.1016/j.ejps.2017.03.040.

[213] R. E. Amaro and A. J. Mulholland, "Multiscale methods in drug design bridge chemical and biological complexity in the search for cures," *Nat. Rev. Chem.*, vol. 2, no. 4, pp. 1–12, Apr. 2018, doi: 10.1038/s41570-018-0148.

[214] S. D. Krämer, *Absorption prediction from physicochemical parameters*, vol. 2, no. 9. Elsevier Current Trends, 1999, pp. 373–380.

[215] D. Bemporad, J. W. Essex, and C. Luttmann, "Permeation of Small Molecules

through a Lipid Bilayer: A Computer Simulation Study," *J. Phys. Chem. B*, vol. 108, no. 15, pp. 4875–4884, Apr. 2004, doi: 10.1021/jp035260s.

[216] J. Witek *et al.*, "Interconversion Rates between Conformational States as Rationale for the Membrane Permeability of Cyclosporines," *ChemPhysChem*, vol. 18, no. 23, pp. 3309–3314, Dec. 2017, doi: 10.1002/cphc.201700995.

[217] N. Pokhrel and L. Maibaum, "Free Energy Calculations of Membrane Permeation: Challenges Due to Strong Headgroup-Solute Interactions," *J. Chem. Theory Comput.*, vol. 14, no. 3, pp. 1762–1771, Mar. 2018, doi: 10.1021/acs.jctc.7b01159.

[218] K. Eyer *et al.*, "A liposomal fluorescence assay to study permeation kinetics of drug-like weak bases across the lipid bilayer," *J. Control. Release*, vol. 173, no. 1, pp. 102–109, Jan. 2014, doi: 10.1016/j.jconrel.2013.10.037.

[219] S. J. Marrink and H. J. C. Berendsen, "Simulation of water transport through a lipid membrane," *J. Phys. Chem.*, vol. 98, no. 15, pp. 4155–4168, 1994, doi: 10.1021/j100066a040.

[220] T. X. Xiang and B. D. Anderson, "Liposomal drug transport: A molecular perspective from molecular dynamics simulations in lipid bilayers," *Advanced Drug Delivery Reviews*, vol. 58, no. 12–13. Adv Drug Deliv Rev, pp. 1357–1378, Nov. 30, 2006, doi: 10.1016/j.addr.2006.09.002.

[221] J. Ulander and A. D. J. Haymet, "Permeation Across Hydrated DPPC Lipid Bilayers: Simulation of the Titrable Amphiphilic Drug Valproic Acid," *Biophys. J.*, vol. 85, no. 6, pp. 3475–3484, 2003, doi: 10.1016/S0006-3495(03)74768-7.

[222] R. W. Tejwani, M. E. Davis, B. D. Anderson, and T. R. Stouch, "DRUG DISCOVERY INTERFACE: Functional Group Dependence of Solute Partitioning to Various Locations within a DOPC Bilayer: A Comparison of Molecular Dynamics Simulations with Experiment," *J. Pharm. Sci.*, vol. 100, no. 6, pp. 2136–2146, Jun. 2011, doi: 10.1002/jps.22441.

[223] M. Paloncýová, K. Berka, and M. Otyepka, "Convergence of free energy profile of coumarin in lipid bilayer," *J. Chem. Theory Comput.*, vol. 8, no. 4, pp. 1200–1211, Apr. 2012, doi: 10.1021/ct2009208.

[224] T. S. Carpenter, D. A. Kirshner, E. Y. Lau, S. E. Wong, J. P. Nilmeier, and F. C. Lightstone, "A Method to Predict Blood-Brain Barrier Permeability of Drug-Like Compounds Using Molecular Dynamics Simulations," *Biophys. J.*, vol. 107, no. 3, pp. 630–641, Aug. 2014, doi: 10.1016/j.bpj.2014.06.024.

[225] C. H. Tse, J. Comer, Y. Wang, and C. Chipot, "Link between Membrane Composition and Permeability to Drugs," *J. Chem. Theory Comput.*, vol. 14, no. 6, pp. 2895–2909, Jun. 2018, doi: 10.1021/acs.jctc.8b00272.

[226] Q. Zhu *et al.*, "Entropy and Polarity Control the Partition and Transportation of Drug-like Molecules in Biological Membrane," *Sci. Rep.*, vol. 7, no. 1, p. 17749, Dec. 2017, doi: 10.1038/s41598-017-18012-7.

[227] H. A. L. Filipe *et al.*, "Quantitative Assessment of Methods Used to Obtain Rate Constants from Molecular Dynamics Simulations - Translocation of Cholesterol

across Lipid Bilayers," *J. Chem. Theory Comput.*, vol. 14, no. 7, pp. 3840–3848, Jul. 2018, doi: 10.1021/acs.jctc.8b00150.

[228] C. T. Leahy, R. D. Murphy, G. Hummer, E. Rosta, and N. V. Buchete, "Coarse Master Equations for Binding Kinetics of Amyloid Peptide Dimers," *J. Phys. Chem. Lett.*, vol. 7, no. 14, pp. 2676–2682, Jul. 2016, doi: 10.1021/acs.jpclett.6b00518.

[229] V. S. Pande, K. Beauchamp, and G. R. Bowman, "Everything you wanted to know about Markov State Models but were afraid to ask," *Methods*, vol. 52, no. 1, pp. 99–105, Sep. 2011, doi: 10.1016/j.ymeth.2010.06.002.Everything.

[230] B. E. Husic and V. S. Pande, "Markov State Models: From an Art to a Science," *Journal of the American Chemical Society*, vol. 140, no. 7. American Chemical Society, pp. 2386–2396, Feb. 21, 2018, doi: 10.1021/jacs.7b12191.

[231] C. J. Dickson, V. Hornak, R. A. Pearlstein, and J. S. Duca, "Structure-kinetic relationships of passive membrane permeation from multiscale modeling," *J. Am. Chem. Soc.*, vol. 139, no. 1, p. jacs.6b11215, Jan. 2016, doi: 10.1021/jacs.6b11215.

[232] E. Rosta and G. Hummer, "Free energies from dynamic weighted histogram analysis using unbiased Markov state model," *J. Chem. Theory Comput.*, vol. 11, no. 1, pp. 276–285, Jan. 2015, doi: 10.1021/ct500719p.

[233] W. C. Swope *et al.*, "Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and a β-hairpin peptide," *J. Phys. Chem. B*, vol. 108, no. 21, pp. 6582–6594, May 2004, doi: 10.1021/jp037422q.

[234] S. Röblitz and M. Weber, "Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification," *Adv. Data Anal. Classif.*, vol. 7, no. 2, pp. 147–179, May 2013, doi: 10.1007/s11634-013-0134-6.

[235] C. D. Meyer, "An alternative expression for the mean first passage matrix," *Linear Algebra Appl.*, vol. 22, no. C, pp. 41–47, 1978, doi: 10.1016/0024-3795(78)90055-1.

[236] L. S. Stelzl, A. Kells, E. Rosta, and G. Hummer, "Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations," *J. Chem. Theory Comput.*, vol. 13, no. 12, pp. 6328–6342, Dec. 2017, doi: 10.1021/acs.jctc.7b00373.

[237] L. S. Stelzl and G. Hummer, "Kinetics from Replica Exchange Molecular Dynamics Simulations," *J. Chem. Theory Comput.*, vol. 13, no. 8, pp. 3927–3935, Aug. 2017, doi: 10.1021/acs.jctc.7b00372.

[238] H. Wu, F. Paul, C. Wehmeyer, and F. Noé, "Multiensemble Markov models of molecular thermodynamics and kinetics," *Proc. Natl. Acad. Sci.*, vol. 113, no. 23, pp. E3221–E3230, 2016, doi: 10.1073/pnas.1525092113.

[239] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins: Structure, Function and Genetics*, vol.

65, no. 3. Proteins, pp. 712–725, Nov. 15, 2006, doi: 10.1002/prot.21123.

[240] D. A. Case *et al.*, "The Amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16. NIH Public Access, pp. 1668–1688, Dec. 2005, doi: 10.1002/jcc.20290.

[241] C. J. Dickson *et al.*, "Lipid14: The amber lipid force field," *J. Chem. Theory Comput.*, vol. 10, no. 2, pp. 865–879, Feb. 2014, doi: 10.1021/ct4010307.

[242] J.-P. P. Ryckaert, G. Ciccotti+, H. J. C. C. Berendsen, G. Ciccotti, and H. J. C. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *J. Comput. Phys.*, vol. 23, no. 3, pp. 327–341, Mar. 1977, doi: 10.1016/0021-9991(77)90098-5.

[243] A. Kells, A. Annibale, and E. Rosta, "Limiting relaxation times from Markov state models," *J. Chem. Phys.*, vol. 149, no. 7, p. 072324, Aug. 2018, doi: 10.1063/1.5027203.

[244] H. Träuble, "The movement of molecules across lipid membranes: A molecular theory," *J. Membr. Biol.*, vol. 4, no. 1, pp. 193–208, Dec. 1971, doi: 10.1007/BF02431971.

[245] D. Bemporad, C. Luttmann, and J. W. Essex, "Behaviour of small solutes and large drugs in a lipid bilayer from computer simulations," *Biochim. Biophys. Acta - Biomembr.*, vol. 1718, no. 1–2, pp. 1–21, Dec. 2005, doi: 10.1016/j.bbamem.2005.07.009.

[246] M. Orsi, M. G. Noro, and J. W. Essex, "Dual-resolution molecular dynamics simulation of antimicrobials in biomembranes," *J. R. Soc. Interface*, vol. 8, no. 59, pp. 826–841, Jun. 2011, doi: 10.1098/rsif.2010.0541.

[247] R. Sun, J. F. Dama, J. S. Tan, J. P. Rose, and G. A. Voth, "Transition-Tempered Metadynamics Is a Promising Tool for Studying the Permeation of Drug-like Molecules through Membranes," *J. Chem. Theory Comput.*, vol. 12, no. 10, pp. 5157–5169, Oct. 2016, doi: 10.1021/acs.jctc.6b00206.

[248] M. Orsi and J. W. Essex, "Permeability of drugs and hormones through a lipid bilayer: Insights from dual-resolution molecular dynamics," *Soft Matter*, vol. 6, no. 16, pp. 3797–3808, Aug. 2010, doi: 10.1039/c0sm00136h.

[249] B. De Nijs *et al.*, "Smart supramolecular sensing with cucurbit[: N] urils: Probing hydrogen bonding with SERS," *Faraday Discuss.*, vol. 205, no. 0, pp. 505–515, Dec. 2017, doi: 10.1039/c7fd00147a.

[250] R. Suardíaz, P. G. Jambrina, L. Masgrau, À. González-Lafont, E. Rosta, and J. M. Lluch, "Understanding the Mechanism of the Hydrogen Abstraction from Arachidonic Acid Catalyzed by the Human Enzyme 15-Lipoxygenase-2. A Quantum Mechanics/Molecular Mechanics Free Energy Simulation," *J. Chem. Theory Comput.*, vol. 12, no. 4, pp. 2079–2090, Apr. 2016, doi: 10.1021/acs.jctc.5b01236.