

DISSERTATION
submitted
to the
Combined Faculty for the Natural Sciences and Mathematics
of
Heidelberg University, Germany
for the degree of
Doctor of Natural Sciences

Presented by:
M.Sc. Gaurav Kumar Ganotra
Born in Bhiwani, India
Oral examination:

COMPUTATIONAL STUDIES OF
DRUG-BINDING KINETICS

Referees:

Prof. Dr. Rebecca C. Wade

Prof. Dr. Frauke Gräter

Acknowledgements

First and foremost, I would like to express my immense gratitude to my supervisor Prof. Rebecca C. Wade for being a kind and patient mentor to me. I am highly indebted to Prof. Rebecca C. Wade for her unwavering mentorship, steadfast support and constant encouragement to make this journey intellectually stimulating. I thank her for providing a laboratory environment conducive to free thinking and her genuine support, both at professional and personal level that helped me accomplish my career goals. I am also very thankful to my second supervisor Prof. Frauke Gräter for co-supervising my PhD thesis and taking out her valuable time to attend and provide helpful suggestions in several of my thesis advisory committee meetings.

I am also very grateful to my mentors: Dr. Daria B. Kokh, Dr. Neil J. Bruce and Dr. Kashif S. Sadiq for their continuous support and excellent guidance throughout my PhD work. I am highly indebted to them for always extending their help and investing their time into my scientific training, especially developing my understanding in the field of molecular modeling and simulations. I would like to express my sincere thanks to all the current and past members of Molecular and Cellular Modeling (MCM) group at Heidelberg Institute for Theoretical Studies (HITS gGmbH), for providing a friendly and positive environment in the group for work. The interesting scientific and non-scientific discussions with them over the lunch and coffee breaks helped me enhance my scientific knowledge and to improve my social and interpersonal skills. I enjoyed and will always cherish the memories of our several group outings and retreats. And, I would like to thank especially Dr. Stefan Richter, the IT Mastermind in the group, for helping with the intricacies of software and hardware related issues. I would also like to specially thank Dr. Goutam Mukherjee for his great support and guidance during the last stage of my PhD.

I acknowledge the generous financial support from the Klaus Tschira Stiftung (KTS) gGmbH for supporting my research and helping me finish my PhD studies in Germany. I am also thankful for the fantastic institutional support and the core facilities at Heidelberg Institute for Theoretical Studies (HITS gGmbH) for providing state of the art computational resources, without which it would have been impossible to do good science. I also appreciate the educational training and the

strong administrative support that I received from the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp), Heidelberg University.

Finally, and most importantly, I would like to express my heartfelt gratitude to my family, especially my parents and my wife Namrata. They have been a source of great joy for me and their unconditional love and trust have always been the driving force for achieving my objectives. Without their blessings and affection, I would not have finished graduate school successfully. I owe them all my achievements and I am ever so grateful to them.

Publications arising from this thesis

1. **Gaurav K. Ganotra** and Rebecca C. Wade. Prediction of Drug–Target Binding Kinetics by Comparative Binding Energy Analysis. *ACS Medicinal Chemistry Letters* (2018), 9 (11): 1134-1139.

Chapter 3 is based on the material I wrote from this publication.

2. Christina Heroven, Victoria Georgi, **Gaurav K. Ganotra**, Paul E Brennan, Finn Wolfreys, Rebecca C. Wade, Amaury E. Fernández-Montalván, Apirat Chaikuad, Stefan Knapp. Halogen-aromatic π -interactions modulate inhibitor residence time. *Angewandte Chemie International Edition* (2018), 57(24):7220-7224.

The material I wrote from this publication is incorporated into Chapter 4.

3. Christina Heroven, Victoria Georgi, **Gaurav K. Ganotra**, Paul E Brennan, Finn Wolfreys, Rebecca C. Wade, Amaury E. Fernández-Montalván, Apirat Chaikuad, Stefan Knapp. Halogenaromatische π -Wechselwirkungen Modulieren Die Verweilzeit von Inhibitoren. *Angewandte Chemie* (2018), 130(24):7338–7343.

The material I wrote from this publication is incorporated into Chapter 4.

4. Neil J Bruce, **Gaurav K. Ganotra**, Daria B Kokh, S Kashif Sadiq, Rebecca C Wade. New approaches for computing ligand–receptor binding kinetics. *Current Opinion in Structural Biology* (2018), 49:1-10.

The material I wrote from this publication is incorporated into Chapter 1.

All reprints were made with permission from the publishers.

Abstract

The drug-receptor binding kinetics are defined by the rate at which a given drug associates with and dissociates from its binding site on its macromolecular receptor. The lead optimization stage of drug discovery programs usually emphasizes optimizing the affinity (as described by the equilibrium dissociation constant, K_d) of a drug which depends on the strength of its binding to a specific target. Since affinity is optimized under equilibrium conditions, it does not always ensure higher potency *in vivo*. There has been a growing consensus that, in addition to K_d , kinetic parameters (k_{on} and k_{off}) should be optimized to improve the chances of a good clinical outcome. However, current understanding of the physicochemical features that contribute to differences in binding kinetics is limited. Experimental methods that are used to determine kinetic parameters for drug binding and unbinding are often time consuming and labor-intensive. Therefore, robust, high-throughput *in silico* methods are needed to predict binding kinetic parameters and to explore the mechanistic determinants of drug-protein binding. As the experimental data on drug-binding kinetics is continuously growing and the number of crystallographic structures of ligand-receptor complexes is also increasing, methods to compute three dimensional (3D) Quantitative-Structure-Kinetics relationships (QSKRs) offer great potential for predicting kinetic rate constants for new compounds. COMparative BINDing Energy (COMBINE) analysis is one example of such approach that was developed to derive target-specific scoring functions based on molecular mechanics calculations. It has been used extensively to predict properties such as binding affinity, target selectivity, and substrate specificity. In this thesis, I made the first application of COMBINE analysis to derive Quantitative Structure-Kinetics Relationships (QSKRs) for the dissociation rates. I obtained models for k_{off} of inhibitors of HIV-1 protease and heat shock protein 90 (HSP90) with very good predictive power and identified the key ligand-receptor interactions that contribute to the variance in binding kinetics.

With technological and methodological advances, the use of all-atom unbiased Molecular Dynamics (MD) simulations can allow sampling upto the millisecond timescale and investigation of the kinetic profile of drug binding and unbinding to a receptor. However, the residence times of drug-receptor complexes are usually

longer than the timescales that are feasible to simulate using conventional molecular dynamics techniques. Enhanced sampling methods can allow faster sampling of protein and ligand dynamics, thereby resulting in application of MD techniques to study longer timescale processes. I have evaluated the application of τ -Random Acceleration Molecular Dynamics (τ RAMD), an enhanced sampling method based on MD, to compute the relative residence times of a series of compounds binding to Haspin kinase. A good correlation ($R^2 = 0.86$) was observed between the computed residence times and the experimental residence times of these compounds. I also performed interaction energy calculations, both at the quantum chemical level and at the molecular mechanics level, to explain the experimental observation that the residence times of kinase inhibitors can be prolonged by introducing halogen-aromatic π interactions between halogen atoms of inhibitors and aromatic residues at the binding site of kinases. I determined different energetic contributions to this highly polar and directional halogen-bonding interaction by partitioning the total interaction energy calculated at the quantum-chemical level into its constituent energy components. It was observed that the major contribution to this interaction energy comes from the correlation energy which describes second-order intermolecular dispersion interactions and the correlation corrections to the Hartree-Fock energy.

In addition, a protocol to determine diffusional k_{on} rates of low molecular weight compounds from Brownian Dynamics (BD) simulations of protein-ligand association was established using SDA 7 software. The widely studied test case of benzamidine binding to trypsin was used to evaluate a set of parameters and a robust set of optimal parameters was determined that should be generally applicable for computing the diffusional association rate constants of a wide range of protein-ligand binding pairs. I validated this protocol on inhibitors of several targets with varying complexity such as Human Coagulation Factor Xa, Haspin kinase and N1 Neuraminidase, and the computed diffusional association rate constants were compared with the experiments. I contributed to the development of a toolbox of computational methods: KBbox (<http://kbbox.h-its.org/toolbox/>), which provides information about various computational methods to study molecular binding kinetics, and different computational tools that employ them. It was developed to guide researchers on the use of the different computational and simulation approaches available to compute the kinetic parameters of drug-protein binding.

Zusammenfassung

Die Kinetik der Rezeptorbindung wird durch die Geschwindigkeit definiert, mit der ein bestimmtes Medikament mit einem makromolekularen Rezeptor assoziiert oder dissoziiert. Die Lead-Optimierungsphase von Drug-Discovery-Programmen konzentriert sich in der Regel auf die Optimierung der Affinität (beschrieben durch die Gleichgewichtsdissoziationskonstante, K_d) eines Medikaments, die von der Stärke seiner Bindung an ein bestimmtes Ziel abhängt. Da die Affinität unter Gleichgewichtsbedingungen optimiert wird, muss sie nicht unbedingt *in vivo* zu einer höheren Wirksamkeit führen. Es besteht ein wachsender Konsens darüber, dass neben K_d auch die kinetischen Parameter (k_{on} und k_{off}) optimiert werden sollten, um die Chancen auf ein gutes klinisches Ergebnis zu verbessern. Allerdings ist das Verständnis der physikalisch-chemischen Eigenschaften, die zu Unterschieden in der Bindungskinetik beitragen, derzeit begrenzt. Experimentelle Methoden, die zur Bestimmung kinetischer Parameter für die Bindung und Entbindung von Medikamenten eingesetzt werden, sind oft zeitaufwendig und arbeitsintensiv. Daher werden robuste, hochdurchsatzfähige *In-silico*-Methoden benötigt, um kinetische Bindungsparameter vorherzusagen und die mechanistischen Determinanten der Wirkstoff-Protein-Bindung zu untersuchen. Mit den kontinuierlich wachsenden experimentellen Daten zur Arzneimittelbindungskinetik und der zunehmenden Anzahl kristallographischer Strukturen von Liganden-Rezeptorkomplexen bieten Verfahren zur Berechnung dreidimensionaler (3D) Quantitativ-Struktur-Kinetik-Beziehungen (QSKRs) ein großes Potenzial zur Vorhersage kinetischer Geschwindigkeitskonstanten für neue Verbindungen. Die Comparative BINDing Energy (COMBINE)-Analyse ist ein solcher Ansatz, der entwickelt wurde, um zielgerichtete Scoring-Funktionen auf der Grundlage molekularmechanischer Berechnungen abzuleiten. Es wurde umfassend genutzt, um Eigenschaften wie Bindungsaffinität, Zielselektivität und Substratspezifität vorherzusagen. In dieser Arbeit habe ich die erste Anwendung der COMBINE-Analyse zur Ableitung von Quantitative Structure-Kinetics Relationships (QSKRs) für die Dissoziationsraten durchgeführt. Ich erhielt Modelle für k_{off} von Inhibitoren der HIV-1-Protease

und des Hitzeschockproteins 90 (HSP90) mit sehr guter Vorhersagekraft und identifizierte die wichtigsten Liganden-Rezeptoren-Interaktionen, die zur Varianz der Bindungskinetik beitragen.

Mit technologischen und methodischen Fortschritten kann der Einsatz von All Atomaren unbiased Molecular Dynamics (MD)-Simulationen bis in den Millisekundenbereich und die Untersuchung des kinetischen Profils der Medikamentenbindung und -entbindung an einen Rezeptor ermöglicht werden. Die Lebensdauer von Wirkstoff-Rezeptorkomplexen sind jedoch in der Regel länger als die Zeiten, die mit herkömmlichen molekularen Dynamikverfahren simuliert werden können. Verbesserte Verfahren können eine schnellere Probenahme von Protein- und Ligandendynamik ermöglichen, was zur Anwendung von MD-Techniken zur Untersuchung länger andauernder Prozesse führt. Ich habe die Anwendung von τ -Random Acceleration Molecular Dynamics (τ RAMD), einer verbesserten Probenahmemethode auf MD-Basis, zur Berechnung der relativen Verweilzeiten einer Reihe von Verbindungen, die an die Haspin-Kinase binden, ausgewertet. Es wurde eine gute Korrelation ($R^2=0,86$) zwischen den berechneten Verweilzeiten und den experimentellen Verweilzeiten dieser Verbindungen bestimmt. Ich habe auch Wechselwirkungsenergieberechnungen sowohl auf quantenchemischer als auch auf molekularmechanischer Ebene durchgeführt, um die experimentelle Beobachtung zu erklären, dass die Verweilzeiten von Kinase-Inhibitoren durch die Einführung von halogenaromatischen π Wechselwirkungen zwischen Halogenatomen von Inhibitoren und aromatischen Resten an der Bindestelle von Kinasen verlängert werden können. Ich bestimmte verschiedene energetische Beiträge zu dieser hochpolaren und gerichteten Halogen-Bindungswechselwirkung, indem ich die auf quantenchemischer Ebene berechnete gesamte Wechselwirkungsenergie in ihre konstituierenden Energiekomponenten aufteilte. Es wurde beobachtet, dass der Hauptbeitrag zu dieser Interaktionsenergie aus der Korrelationsenergie stammt, die intermolekulare Dispersionswechselwirkungen zweiter Ordnung und die Korrelationskorrekturen zur Hartree-Fock-Energie beschreibt.

Darüber hinaus wurde mit Hilfe der SDA 7-Software ein Protokoll zur Bestimmung der Diffusions k_{on} raten von niedermolekularen Verbindungen aus Brownian Dynamics (BD)-Simulationen von Protein-Liganden-Assoziationen erstellt. Der ausführlich untersuchte Testfall der Benzamidinbindung an Trypsin wurde verwendet, um eine Reihe von Parametern zu bewerten, und es wurde ein Satz optimaler Parameter bestimmt, der allgemein anwendbar sein sollten, um die Diffusionsratenkon-

stanten für Assoziation einer breiten Palette von Protein-Liganden-Bindungspaaren zu berechnen. Ich habe dieses Protokoll über Inhibitoren von mehreren Zielproteine mit unterschiedlicher Komplexität wie Human Coagulation Factor Xa, Haspin Kinase und N1 Neuraminidase validiert, und die berechneten Diffusionsratenkonstanten für Assoziation mit Experimenten verglichen. Ebenso habe ich an der Entwicklung einer Toolbox von Berechnungsmethoden mitgewirkt: KBbox (<http://kbbox.h-its.org/toolbox/>), die Informationen über verschiedene Berechnungsmethoden zur Untersuchung der molekularen Bindungskinetik und verschiedene Berechnungswerkzeuge, die diese verwenden, liefert. Dieses Tool wurde entwickelt, um Forscher bei der Suche nach neuen Simulationsmethoden unter der Berechnung von kinetischen Parameter zu unterstützen.

Contents

Abbreviations	xv
1 Introduction	1
1.1 Drug-binding kinetics and its importance	2
1.2 Current state-of-the-art in computing drug-binding kinetics	4
1.2.1 Fast binding of small, rather rigid ligands: the trypsin–benzamidine complex and fragment binding	5
1.2.2 Unbinding with conformational changes: kinase and heat shock protein 90 (HSP90) inhibitors	8
1.2.3 Binding to membrane proteins: G protein coupled receptor (GPCR) ligands	10
1.2.4 Binding of flexible ligands to flexible proteins: peptide binding to MDM2 protein and HIV-1 protease	11
1.3 Objectives and Motivation of the work	13
1.4 Organization of the thesis	15
2 Theoretical Methods and Software	17
2.1 COMparative BINding Energy (COMBINE) analysis	17
2.1.1 Partial least squares (PLS) regression	19
2.2 Brownian Dynamics	21
2.2.1 The concept of Brownian motion	21
2.2.2 Simulation of Diffusional Association (SDA) software	22
2.2.3 Calculation of protein-ligand association rates in SDA	26
2.3 Molecular Dynamics (MD) technique	27
2.3.1 τ Random Acceleration Molecular Dynamics (τ RAMD)	32
2.3.2 MM/GBSA free-energy calculations	32
2.3.3 Møller–Plesset energy calculations	33

2.4	Software and Tools	34
2.4.1	Simulation software	35
2.4.2	Structure preparation and general molecular modeling tools	36
2.4.3	Structure Visualization tools	38
3	Quantitative structure-kinetics relationships (QSKRs) for k_{off} values of HSP90 and HIV-1 protease inhibitors	39
3.1	Systems studied	41
3.1.1	Heat-shock protein 90 (HSP90)	41
3.1.2	HIV-1 protease	42
3.2	Dataset used for the COMBINE analysis	43
3.2.1	Heat-shock protein 90	43
3.2.2	HIV-1 protease	48
3.3	Methods	52
3.3.1	Preparation of protein and ligand structures	52
3.3.2	Generation of force field parameters and energy minimization	52
3.3.3	Selection of the training and the test datasets	54
3.3.4	Calculation of the interaction energy terms and generation of energy matrix for PLS analysis	54
3.3.5	PLS analysis	55
3.3.6	Model validation	55
3.4	Results	56
3.4.1	COMBINE analysis model for HSP90 inhibitors	56
3.4.2	Results: COMBINE analysis model for HIV-1 protease inhibitors	64
3.5	Concluding Discussions	69
4	Halogen-aromatic π interactions modulate inhibitor residence time	75
4.1	Background	75
4.2	Aim of the work	80
4.3	Methods	80
4.3.1	Quantum mechanical interaction energy calculations	80
4.3.2	Binding free energy calculations using MM/GBSA	81
4.3.3	τ -Random Acceleration Molecular Dynamics (τ -RAMD) simulations	83

4.4	Results and Discussions	85
4.4.1	The second order Møller-Plesset interaction energies (E_{MP2}) between the inhibitor and the gatekeeper residue correlate well with dissociation rate constants and equilibrium dissociation constants determined experimentally	85
4.4.2	Binding free energies calculated from MM/GBSA approach correlate with experimental parameters for the halogen-gatekeeper interaction.	89
4.4.3	Relative residence times from τ -RAMD simulations correlate with the experimentally measured residence times	90
5	Protocol for calculation of diffusional association rates for small molecules using Brownian dynamics	93
5.1	Overview	93
5.1.1	Trypsin	94
5.1.2	Human Coagulation Factor Xa	95
5.1.3	Haspin kinase	96
5.1.4	Neuraminidase	97
5.2	Methods	98
5.2.1	Preparation of protein and ligand structures	98
5.2.2	Preparation of PQR files	100
5.2.3	Grids preparation	100
5.2.4	Effective charges for protein and ligands	103
5.2.5	Calculation of diffusion coefficients	112
5.2.6	Generation of Reaction Criteria	113
5.2.7	Association rate calculation with SDA	116
5.3	Results	120
5.3.1	Diffusional association rate constants (k_{on}) computed for the trypsin-benzamidine association	120
5.3.2	Diffusional k_{on} rate constants computed for the inhibitors of Human Coagulation Factor Xa	122
5.3.3	Diffusional k_{on} rate constants computed for the inhibitors of Haspin kinase	124
5.3.4	Diffusional k_{on} rate constants computed for the inhibitors of Neuraminidase	126

5.4	Concluding Discussions	128
6	KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding	131
6.1	Technical Implementation	133
6.2	Database Structure	133
6.2.1	Query Interface to choose the methods	135
6.3	Group of Methods available in KBbox	136
6.4	List of Examples	137
6.5	Group of Computational tools in KBbox	137
6.6	List of Tutorials	138
6.7	Example Use Cases	139
6.7.1	What method should I use for a given project?	139
6.7.2	Where can I find information on previous applications of a method for studying kinetics?	141
7	Summary and Outlook	143
	Bibliography	149

Abbreviations

K_d Equilibrium dissociation constant [M].

β 2AR β 2 adrenergic receptor.

τ RAMD τ Random Acceleration Molecular Dynamics.

k_{off} Dissociation rate constant [s^{-1}].

k_{on} Association rate constant [$M^{-1}s^{-1}$].

3D Three dimensional.

aMetaD Adiabatic bias MD with metadynamics.

AMS Adaptive multistate splitting.

AMSM Adaptive Markov state modelling.

APBS Adaptive Poisson–Boltzmann Solver; a program for calculating electrostatic potentials for interacting protein and ligand molecules.

BD Brownian dynamics.

BEMD Bias-exchange metadynamics approach.

CDK8 Cyclin-dependent kinase 8.

CMD Conventional molecular dynamics.

CRF₁R Corticotropin-releasing factor type 1 receptor.

CV Collective variable.

ECM Effective Charges for Macromolecules in solvent.

ff14SB Amber force field ff14 Stony Brook.

fs femtosecond (10^{-15} s).

GAFF General Amber Force Field.

GB Generalised Born implicit model.

GPCR G protein coupled receptor.

Grp78 Glucose-regulated protein 78.

HIV Human immunodeficiency virus.

HSP90 Heat shock protein 90.

ITC Isothermal titration calorimetry.

MD Molecular dynamics.

MetaD) Metadynamics approach.

MFPT Mean First Passage Time.

MOE Molecular Operating Environment.

MOL2 Tripos molecular 3D structure file format.

MSM Markov state modelling.

NAMD Nanoscale Molecular Dynamics; a molecular dynamics simulation software.

NMA Normal mode analysis.

ns nanosecond (10^{-9} s).

OpenMP Open Multi-processing.

PBC Periodic Boundary Conditions.

PBE Poisson-Boltzmann equation.

PCA Principal component analysis.

PDB Protein Data Bank; also a molecular 3D structure file format.

PKPD Pharmacokinetic/Pharmacodynamic.

PLS Partial Least Squares Regression.

PME Particle Mesh Ewald summation method.

PMF Potential of Mean Force.

ps picosecond (10^{-12} s).

QM Quantum mechanics.

RAMD Random Acceleration Molecular Dynamics.

RMSD Root mean squared deviation.

rxna file format suffix for reaction criteria file, an input file required by SDA.

SDA Simulation of Diffusional Association Software.

SMD Steered molecular dynamics.

SPR Surface plasmon resonance.

tau Residence time [s].

TM Transition matrix.

UHBD University of Houston Brownian Dynamics software; here referred to as a format for electrostatic potential files accepted by the SDA software.

vdW van der Waals.

VMD Visual Molecular Dynamics software for molecular graphics.

WE Weighted ensemble path sampling approach.

Amino Acid Abbreviations

Amino acid	3 letter	1 letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Chapter 1

Introduction

Drug binding can be considered as a bimolecular reaction where a drug binds to its target receptor. Understanding the process of drug-receptor binding is crucial for structure-based drug design and is of fundamental importance for pharmaceutical research. The binding affinity, which determines the strength of drug-receptor binding, is usually considered as the most important quantitative metric for estimating the drug's efficacy on the basis of strength of target binding. Therefore, drug discovery programs mainly focus on the design of drug molecules with high receptor affinity and selectivity. For this reason, several computational methods based on molecular dynamics (MD) simulations have been developed to compute receptor-ligand binding affinities[1]. However, designing drug compounds to have high binding affinity does not always result into higher potency *in vivo*. Over the past few years, it has been becoming evident that drug binding kinetics may play a major role in efficacy. Recently, it has been realized that the efficacy of a drug is sometimes more correlated with its residence time at the receptor than the affinity[2]. This has led to widespread efforts in both industry and academia to consider the role of drug binding kinetics in their drug discovery programs[3]. Therefore, the demand for both computational and experimental methods for studying the drug-target binding kinetics is expected to rise. With advances in the computational power and availability of specialized architectures, now it is possible to apply simulation methodologies for relatively longer time scales (upto a few milliseconds) and this has enabled studies of the dynamics of ligand-receptor binding/unbinding using molecular simulations. In addition, progress in machine learning technologies and the availability of datasets of measured kinetic parameters has enabled the understanding of key ligand-receptor

features[3]. Since the experimental approaches that are commonly used to determine binding kinetic parameters are often time-consuming, labour intensive and expensive, there is a need to develop robust and improved *in silico* methods that can be used to compute and predict kinetic parameters for drug-receptor binding and can be used during the drug discovery and design process. With the growing level of interest in drug-binding kinetics, it can be expected that there will be an increased application of computational approaches to study drug-binding kinetics and that new methods will be developed for this purpose.

1.1 Drug-binding kinetics and its importance

The binding of a ligand (L) to its target receptor (R) can be considered as a bimolecular reaction which can be characterized by standard kinetic parameters: k_{on} , the association rate constant ($M^{-1}s^{-1}$) that defines the rate of formation of receptor-ligand complex (RL); k_{off} , the dissociation rate constant (s^{-1}), which measures the rate of dissociation of a receptor-ligand complex (RL); τ , the residence time (s), which describes the lifetime of a receptor-ligand complex and is given by the inverse of the dissociation rate constant ($\tau = 1/k_{off}$); and by a thermodynamic parameter, also known as the equilibrium dissociation constant, $K_d \equiv [L][R]/[LR]$, (units: $[M]$). K_d is related to the binding free energy ΔG as:

$$K_d = e^{\frac{+\Delta G}{k_B T}} \quad (1.1)$$

The simple one-step binding model with the transition state $RL^\#$ (see Figure 1.1 A) can be represented as:



where k_1 and k_{-1} are the rate constants for the association of receptor and ligand to form the receptor-ligand complex ($k_1 = k_{on}$) and for the complex to dissociate ($k_{-1} = k_{off}$), respectively (see Figure 1.1 A).

Under steady-state conditions:

$$K_d = \frac{k_{off}}{k_{on}} = \frac{k_{-1}}{k_1} \quad (1.3)$$

In general, the process of ligand-receptor binding can be represented as a two-step process with an intermediate state RL^* (see Figure 1.1 B):

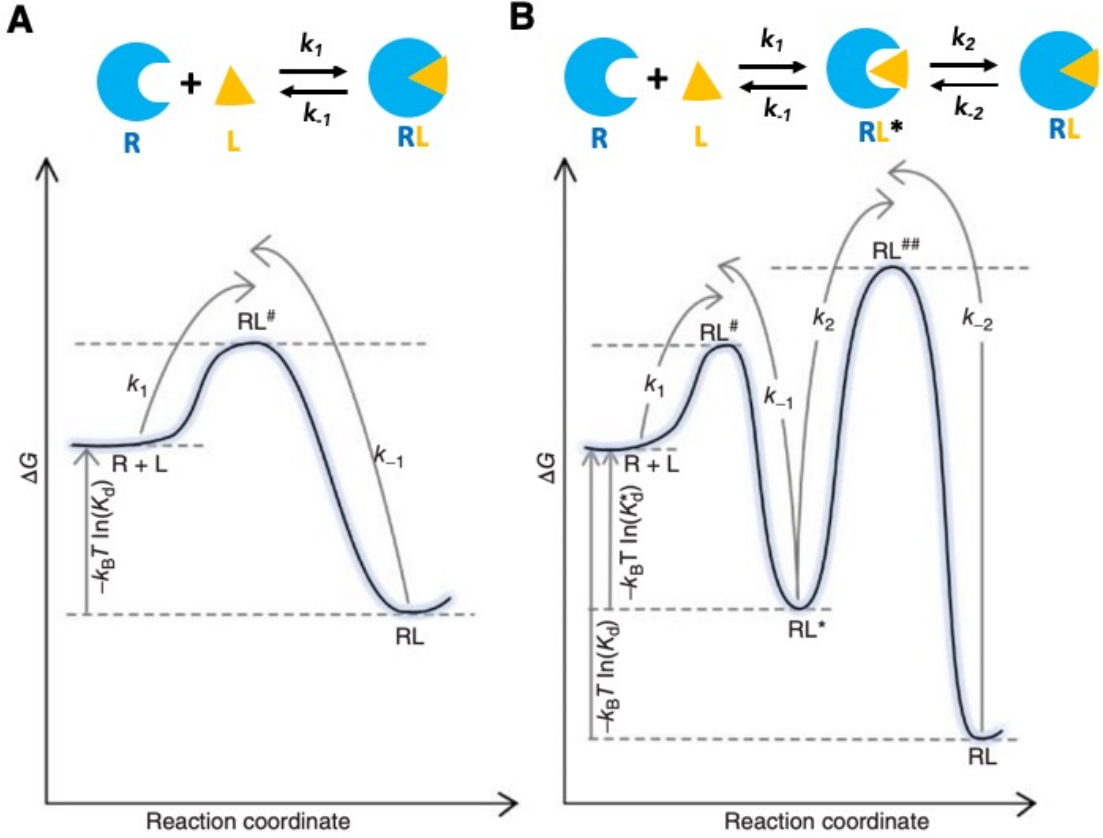


Figure 1.1: Two models for describing bimolecular receptor-ligand (R-L) binding, (A) one-step binding model and (B) two-step binding model. The plots show one-dimensional schemes of simplified energy landscapes where important free energy minima and maxima are marked; the heights of these barriers are related to the kinetic parameters (see text). The figure is adapted from Romanowska *et al.*[4]



Here, the first step of the binding process describes the diffusion-controlled approach of the ligand (L) towards the receptor (R) to form a diffusional encounter complex (RL^*) which is often characterized as a relatively stable, but not fully desolvated and ordered, arrangement of the ligand and receptor molecules[5, 6, 7]. The equilibrium dissociation constant (K_d^*) for this step can be given as $K_d^* = k_1/k_{-1}$ [M]. The second step of the binding process involves overcoming an energetic barrier (see $RL^{##}$ in Figure 1.1 B) to achieve a stable, low-energy bound state RL . This step is often referred to as an induced-fit step as it is associated with the formation of short-range interactions (such as hydrogen bonds) between the interacting molecules, the displacement of water molecules occupying the binding pocket, and the conformational changes thereby allowing the ligand and receptor molecules to

adapt to each other. This two-step model of binding leads to a more complicated relationship between the rates of forward and reverse reactions and k_{on}/k_{off} . If the second step is faster than the first step, the binding is considered to be diffusion-limited, and in this case k_{on} is mostly determined by the forward rate of the first binding step (k_1) and the ligand concentration, while k_{off} is mostly determined from the reverse rate of the second step (k_{-2}) with $\tau \sim 1/k_{-2}$ [2, 8].

1.2 Current state-of-the-art in computing drug-binding kinetics

In recent years, several promising computational methods have been developed for the computation of rate constants for ligand-receptor binding and understanding the mechanistic determinants of ligand-receptor binding processes[9] (see Figure 1.2); for reviews see [1, 2, 3, 10, 11, 12, 13, 14]. These methods include different types of enhanced sampling molecular dynamics simulations and the combination of energy-based models with chemo-metric analysis. Some of these approaches are developed for computing absolute association (k_{on}) and/or dissociation (k_{off}) rate constants, while others are developed for computing relative rate constants for a series of compounds. While some of these methods provide detailed information on pathways and binding/unbinding mechanisms, others just provide hints about the key determinants of the rate constants. The choice of an appropriate computational method depends on the level of complexity of the ligand-receptor binding process and the specific challenges posed by the system of interest. The magnitude of the association/dissociation rate constants, size and flexibility of the system are some of the key factors that must be taken into account while making a choice of the appropriate method[9]. Some of the computational methods make approximations that allow increasingly challenging systems to be studied, while others are more computationally rigorous and hence limited to certain classes of system (see following sections). Recently Bruce *et al.*[9] assessed the key computational approaches to compute rates for ligand-receptor binding processes, in terms of the classes of protein-ligand systems studied and the varying levels of complexity of the protein-ligand binding processes (see Figure 1.3). The following sections discuss the examples of the application of these computational methods to specific protein-ligand systems which are categorized into 4 main categories (similar to categories described in Bruce *et*

al.[9]) depending upon the level of flexibility of the proteins and ligands studied and the complexity of the binding process simulated. While a few of the systems are rather easier to simulate due to their low flexibility and faster binding, for others simulating them might be very challenging as the binding might involve large conformational changes and therefore, the high flexibility of the interacting protein and ligand molecules should be properly addressed.

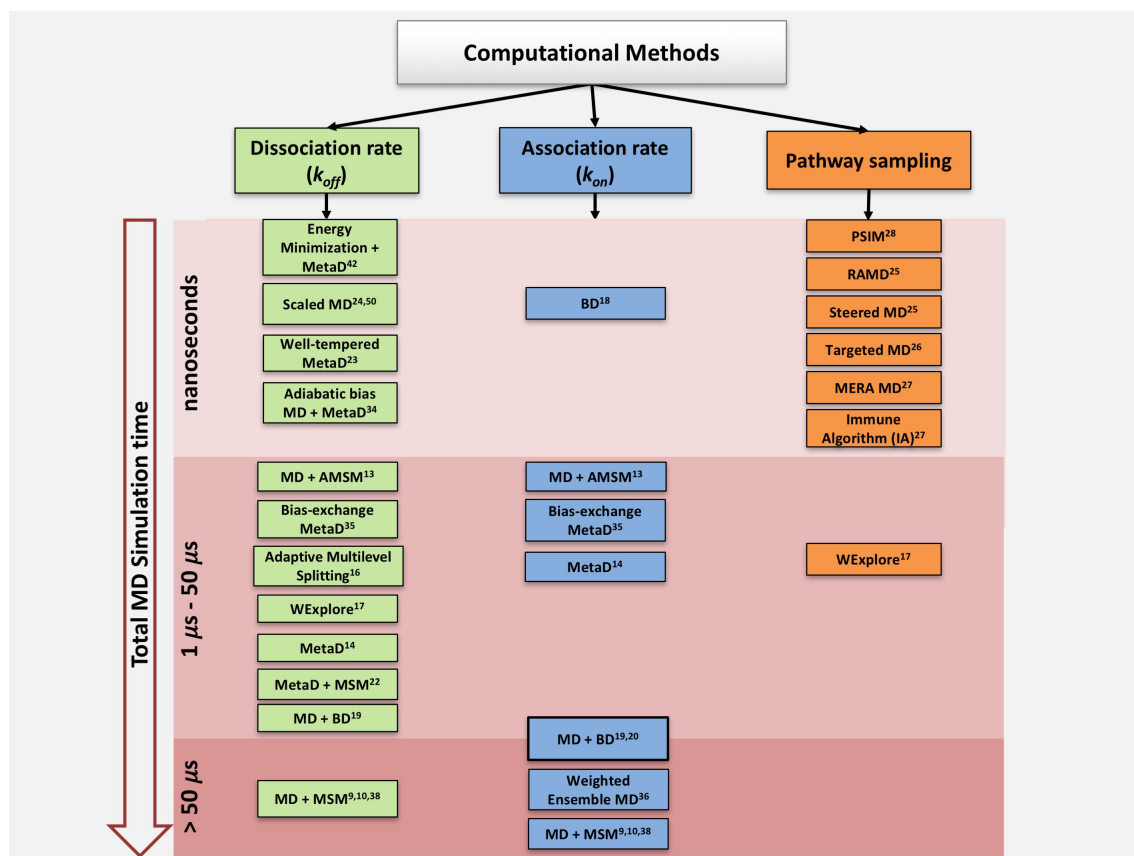


Figure 1.2: Methods for computing ligand-receptor kinetics. The figure is taken from Bruce *et al.*[9] and reproduced with permissions (Citations listed for recently published applications correspond to cited articles in Bruce *et al.*[9]). The simulation time ranges correspond to those in these applications. The simulation time depends on the properties of the system studied as well as the methods used.

1.2.1 Fast binding of small, rather rigid ligands: the trypsin–benzamidine complex and fragment binding

The relatively small size of both trypsin and benzamidine, their comparative rigidity, as well as their relatively fast binding, have made the trypsin–benzamidine complex a popular model system for developing and testing methods for computing

protein–ligand binding kinetics (see Figure 1.3 A). In 2011, Buch *et al.*[15] applied conventional molecular dynamics (CMD) simulations combined with Markov state modelling (MSM) to identify multiple intermediate states for trypsin-benzamidine binding and to compute the kinetics of an overall two-state binding model. In the MSM approach, only partial binding or unbinding transitions are observed from a simulated trajectory. These trajectories are first geometrically clustered in a predefined conformational subspace from which a transition matrix (TM) of discretized microstates is derived. Then, metastable states can be identified from the kinetic clustering of the microstates by using eigenvectors of the TM. The k_{on} and k_{off} values calculated by Buch *et al.*[15] using the MSM approach were 5-fold and 150-fold greater than the respective experimental values. Recently, Plattner and Noe [16] managed to compute kinetic rate constants closer to experimental values by applying a more rigorous multiscale model for MSMs to trypsin-benzamidine binding. However, they needed three times (150 μ s) more CMD sampling compared to the CMD sampling (50 μ s) required by Buch *et al.*[15]. Doerr and Fabritius[17] attempted to reduce this computational demand by demonstrating the application of the Adaptive Markov state model (AMSM) method to trypsin-benzamidine binding and they managed to sample ligand binding one order of magnitude faster than the classical sampling. AMSM methods iteratively perform multiple short trajectories of ensemble simulations. After each iteration, an MSM is constructed to learn a simplified model of the simulations and provide information on the locations of rarely sampled states. This information is then used to perform the next round of simulations to facilitate the crossing of the transition barrier.

The trypsin-benzamidine system has also been studied using the metadynamics (MetaD) approach[18]. In MetaD, a time-dependent biasing potential, represented by a sum of Gaussian functions, is added along a particular geometric coordinate (the so-called collective variable, CV) during a simulation, and this helps to sample the regions that are separated by notable energy barriers. The CVs, to represent the transition pathway or dissociation process (in the case of k_{off} computation), must be carefully chosen in MetaD. The method makes the critical assumption that the CVs chosen, represent the dissociation process as a single rate-limiting transition between two metastable states, and therefore the quality of the model can be evaluated by a statistical analysis. The k_{on} and k_{off} rate constants calculated with MetaD were 2-fold and 70-fold lower, respectively, than the experiment[18].

Teo *et al.*[19] applied the adaptive multistate splitting (AMS) method to compute k_{off} rates for the trypsin-benzamidine system and the computed k_{off} values were 2-fold slower than the experimental value. In the AMS method, an ensemble of simulation trajectories is started from the bound state and they are periodically pruned and restarted from the coordinates chosen so that the system progresses towards the unbound state. Prior knowledge of the transition paths is however not required in AMS. To enhance sampling of unbinding pathways for benzamidine from trypsin, Dickson and Lotz[20] applied WExplore, a method based on Weighted ensemble (WE) path sampling, and the computed k_{off} values were 10 times higher than experiment. The Weighted ensemble (WE) path sampling approach requires defining a reaction coordinate which divides the progression from an initial (unbound) state to a target (bound) state into several bins. A number of trajectories is started from the initial state having each trajectory assigned an equal weight or probability. After a short time interval, the current bin of each trajectory is recorded and the trajectories entering the new bins are either split or combined in order to have a predetermined number of trajectories for each bin. In this way, trajectories are periodically reweighted in a rigorous statistical manner, and the whole process is iterated several times to generate a weighted trajectory ensemble. This weighted ensemble provides information on transition probabilities and the evolving configurational distribution. The Brownian dynamics (BD) is another computationally inexpensive method that has recently been applied to compute diffusional association rate constants for inhibitors, such as oseltamivir binding to the neuraminidase[21]. BD is a stochastic method which uses an implicit continuum solvent model and the diffusional motion of solutes is propagated by integration of the overdamped Langevin equation. To account for solvent friction effects accurately, the random collisions are modelled with the Langevin equation. Due to the use of rigid structures, simplified force fields and an implicit solvent model in BD, it poses significant limitations for small but flexible molecules, as well as conformation dependent protein-binding. However, such limitations can be overcome by using multiscale methods that combine BD with MD. In such multiscale approaches, BD simulations can be used to model the initial diffusional association of two molecules, following which an MD-based regime can be used to simulate the formation of the final bound complex, thereby accounting for the flexibility of the molecules and the conformational changes. Votapka *et al.*[22] demonstrated the application of such a multiscale

approach to the trypsin-benzamidine binding using their software SEEKR[22], which uses a milestoning approach to combine BD and MD-based regimes. The k_{on} rates calculated from SEEKR were in good agreement with experiment whereas, the k_{off} value was within a factor of 10 of experimental values[22].

1.2.2 Unbinding with conformational changes: kinase and heat shock protein 90 (HSP90) inhibitors

Both protein kinases and HSP90 are considered as challenging targets for computing drug-binding kinetics because of their high binding site flexibility (see Figure 1.3 B). Metadynamics (MetaD)[18], an approach that uses a time-dependent energy function to enhance sampling of particular regions of configurational space, has been applied to a number of protein kinases for computation of k_{off} values. Tiwari *et al.*[23] applied MetaD to compute the dissociation rates of dasatinib from c-Src kinase. CVs used in this study were the distance between the ligand and the binding pocket, and a term describing the solvation state of the binding pocket. Casasnovas *et al.*[24] used MetaD to study the unbinding kinetics of a urea-based allosteric inhibitor from p38 MAP kinase. They have used two pathway-based CVs (along and perpendicular to the pathway) which were identified from 8 snapshots from steered MD (SMD) dissociation trajectories. The dissociation rates computed in both of the studies were in good agreement with experiment. Callegari *et al.*[25] attempted to estimate the relative k_{off} values of a set of cyclin-dependent kinase 8 (CDK8) inhibitors by proposing an alternative MetaD-based method. Seven CVs that encode both roto-translational and conformational motions of the ligand, were used for driving the ligands to the point of dissociation. The authors managed to rank a set of CDK8 inhibitors by their residence time, in good agreement with experiment. Mollica *et al.*[26] proposed scaled or smoothed-potential MD, another simpler approach for ranking k_{off} values, which does not involve the definition of CVs. This approach involves smoothing the system’s potential energy with a constant scaling parameter, resulting in the increased sampling of the conformational space. However, in applications to protein–ligand dissociation, a set of restraints is applied on all protein heavy atoms outside the binding site to prevent protein unfolding and keep the protein in its native conformation. The method was validated on several ligands of HSP90, Glucose-Regulated Protein (Grp78), and adenosine A2A receptor (A2A), and in all cases, method was able to rank these ligands correctly.

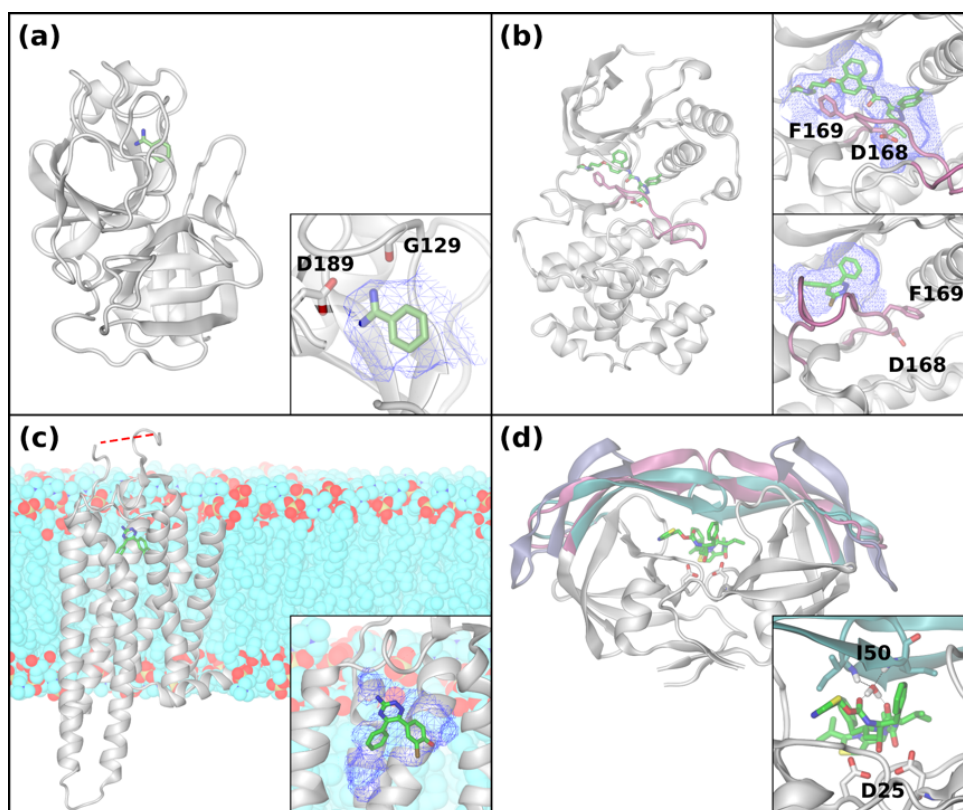


Figure 1.3: Examples of protein–ligand systems for which binding kinetics have been computed, illustrating some of the challenges posed for these calculations. The proteins are shown in cartoon representation along with their ligands and selected residues in stick representation. The insets show the molecular solvent accessible surface of the binding pockets as blue wireframes. (a) The trypsin–benzamidine complex is a classic model system for studying ligand binding due to its fast binding kinetics (PDB ID: 3PTB). Benzamidine binds in a surface exposed cleft, with the amidine forming hydrogen bonds with the sidechain of D189 and the backbone carbonyl of G129 (inset). (b) p38 MAP kinase with allosteric (upper inset, PDB ID: 1KV2) and orthosteric (lower insert, PDB ID: 4ZTH) inhibitors and the kinase activation loop and the DFG motif highlighted (pink). Computation of the binding kinetics of allosteric inhibitor binding to this and other kinases requires consideration of a switch of the D and F positions in the DFG loop, which opens a concealed cavity in the binding pocket that is otherwise blocked by the phenylalanine residue of the DFG loop. (c) The A2A GPCR with a triazine derivative (PDB ID: 3UZC) bound in the orthosteric binding pocket (inset). The dashed red line shows the location of a missing extracellular loop. The heterogeneous membrane-bound environment of GPCRs poses additional challenges both to the determination of accurate experimental structures and to simulation. (d) The HIV-1 protease homodimer with ritonavir bound (PDB ID: 1HXW). HIV-protease has two β -hairpin loops (flaps) that exist in a closed state (cyan) on ligand binding, but can also exist in semi-open (pink) or open (purple) conformations in the unbound form. This protein flexibility, as well as the often high flexibility of the ligands and the presence of a bridging water molecule (with H-bonds to the ligand and the backbone amide nitrogen of I50), need to be treated in computations of binding kinetics (inset). The figure is reproduced from Bruce *et al.*[9], with permissions.

Niu *et al.*[27] proposed a computationally inexpensive approach by using random acceleration molecular dynamics (RAMD) simulations and SMD to rank inhibitors of B-RAF serine/threonine kinase by their residence time. RAMD[28, 29] simulations were first used to obtain dissociation pathways of two inhibitors of B-RAF, followed by application of SMD to generate potentials of mean force, which provided a structural rationale for the difference in transition state barrier in qualitative agreement with the measured difference in binding kinetics. Recently Kokh *et al.*[29] developed the τ -Random Acceleration Molecular Dynamics (τ RAMD) procedure for estimation of the relative residence times and demonstrated its application for sets of diverse ligands of the N-terminal domain of HSP90. In τ -RAMD, an ensemble of MD simulations is run starting from the bound protein-ligand complex, and an artificial randomly oriented force is applied to the centre of mass of the bound ligand. By applying this artificial force, the unbinding of the ligands from the binding site can be observed in short simulations of a few nanoseconds. The authors obtained a good correlation between relative residence times computed from τ RAMD and the experimental residence times.

1.2.3 Binding to membrane proteins: G protein coupled receptor (GPCR) ligands

GPCRs are considered challenging targets for computational studies due to their heterogeneous membrane-bound environments, and since GPCR ligands may partition between the bilayer and solvent, it poses an extra challenge for their computer simulation (see Figure 1.3 C). However, due to the increasing availability of experimental structures of GPCR-ligand complexes over the past 10 years[30] and the importance of engineering the residence times of GPCR-targeting compounds[31], there has been increased interest in studying GPCRs computationally. Dror *et al.*[32] reported the first computational study of GPCR binding kinetics using conventional molecular dynamics (CMD), where they simulated the binding of antagonists and agonists to the β_2 adrenergic receptor (β_2 AR). The authors observed a total of 12 binding events from 50 CMD simulations, each of up to 19 μ s, of the membrane-bound receptors with alprenolol or dihydroalprenolol antagonists. A total of 10 replicas of the alprenolol or dihydroalprenolol antagonists was used to improve the sampling. By estimating the total time in which the ligand was available for binding, that is, in aqueous solution, not penetrating the membrane, and modelling binding as a first-

order Poisson process, the k_{on} rate for alprenolol binding to β_2 AR was calculated to be $3.1 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$, which was in good agreement with the experimental value of $1.0 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$. As mentioned in the previous section, Mollica *et al.*[26] applied scaled MD to correctly rank a congeneric series of four A2A antagonists based on their residence times.

Bortolato *et al.*[33] in 2015, reported a method that combines adiabatic bias MD with metadynamics (aMetaD) to distinguish between short (residence time, $\tau < 20$ min) and long ($\tau > 50$ min) residence time compounds of 3 GPCRs, including 12 ligands of the corticotropin-releasing factor type 1 receptor (CRF₁R). In adiabatic-bias MD[34], a time-dependent harmonic energy barrier is used to drive a system from an initial to a final state along a predefined reaction coordinate. The adiabatic-bias metadynamics (aMetaD) approach combines a time-dependent harmonic energy barrier to the ligand’s movements when it is not moving towards an unbound state with MetaD using two CVs, the distance along an unbinding pathway and the distance perpendicular to the pathway. From each simulation, a score is computed that describes the height of the traversed unbinding transition state barrier.

1.2.4 Binding of flexible ligands to flexible proteins: peptide binding to MDM2 protein and HIV-1 protease

Several studies have been published recently that have successfully applied enhanced sampling techniques to the calculation of rates of protein–peptide binding[35, 36]. Zwier *et al.*[36] computed the k_{on} rate for the binding of the N-terminal peptide fragment of p53 tumor suppressor to the MDM2 protein using a WE path sampling method. In this study, the k_{on} rate was calculated from 182 independent and continuous binding pathways obtained from a total of $\approx 120 \mu\text{s}$ of WE MD simulations, from unbound to encounter complex and from encounter complex to bound state. The computed k_{on} value ($7 \pm 4 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$) was within an order of magnitude of the experimental value of $9.2 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$ [37]. Recently, the mechanism of p53-MDM2 binding was studied in detail by Zhou *et al.*[38] using many unbiased CMD simulations. The authors constructed MSMs from 831 μs of CMD simulations for predicting p53 binding pathways, revealing both conformational selection and induced-fit. The k_{on} value ($2.5 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$) computed in this study was in good agreement with the experimental value. However, compared to experiment[37] (2.1 s^{-1}), the computed k_{off} value ($1.9 \times 10^5 \text{ s}^{-1}$) was strongly overestimated from tran-

sition path theory analysis due to insufficient sampling of binding and unbinding events.

HIV-1 protease is another challenging target known for its high flexibility and has been studied extensively using computer simulations. It has two extended β -hairpin loops, also known as flaps, that are known to exist in semi-open or open conformations in the unbound form and these flaps close when the ligand is bound (see Figure 1.3 D). For HIV-1 protease, the flap dynamics[39, 40] and the water-mediated H-bonds between the flaps and the ligand[40, 41] have been shown to be important for the binding and unbinding of its peptidomimetic inhibitors. Pietrucci *et al.*[35] investigated the binding mechanism of a peptide substrate to HIV-1 protease using the bias-exchange MetaD (BEMD) approach. BEMD involves running several replicas of metadynamics simulations for the same system at the same temperature, with each replica biased by a time-dependent potential acting on a different set of CVs. These replicas are then periodically allowed to exchange their configurations, thereby allowing the biasing of a virtually unlimited number of CVs simultaneously. The multidimensional nature of the bias makes it possible to explore a complex free energy landscape with high efficiency. In this study, the authors sampled a total of 7 CVs which accounted for features such as flap opening, bridging water molecules, and important physical interactions between the ligand and the protease. This conformational space was then used to construct a thermodynamic and kinetic model of the binding process based on the weighted-histogram approach. The computed k_{on} value ($1.26 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$) was roughly 10 times the the experimental value[42] ($\approx 0.16 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$) whereas the computed k_{off} value (57.1 s^{-1}) was overestimated compared to the experiment[42] ($\approx 0.2\text{-}0.4 \text{ s}^{-1}$). Since choosing correct CVs for BEMD is a very difficult task, such methods cannot be applied on a high-throughput scale. However, Sun *et al.*[43] recently demonstrated that for HIV-1 protease and several other targets, it is possible to get a good agreement of k_{off} with experiment for more slowly dissociating drug-like compounds ($k_{off} \approx 10^{-4} \text{ s}^{-1}$) using standard MetaD simulations even if the chosen set of CVs does not fully represent the slowest motion with a single bottle-neck transition. The two CVs chosen by authors in this study were the distance between the ligand and the binding pocket and the RMSD change of the binding pocket. Authors were however not able to accurately calculate the k_{off} value for the kinase studied. But for the other 5 systems that they studied, they managed to obtain the k_{off} values within about an order of magnitude when using

a single short (ns) MetaD simulation of each complex, starting from a minimized crystal structure.

Several chemometric approaches have also been used to derive quantitative structure–kinetics relationships (QSKRs) for HIV-1 protease inhibitors. Qu *et al.*[44] attempted to model the kinetic and thermodynamic properties of a series of HIV-1 protease inhibitors using Volsurf descriptors that were derived from Grid water and hydrophobic probes. The three-fold cross-validation (Q^2) coefficients for their optimal k_{off} and k_{on} models were 0.695 and 0.549, respectively. Such models however only include static structural characteristics and may not sufficiently capture the dynamic features of binding processes. Chiu and Xie[45] tried to address this problem by constructing multi-target machine learning classification models integrating energetic features with conformational variability features, derived from coarse-grained normal mode analysis, to classify HIV-1 protease inhibitors into binding kinetic classes.

1.3 Objectives and Motivation of the work

The computational studies to investigate protein-ligand interactions and to estimate kinetics of protein-ligand binding are often addressed using biomolecular simulation-based approaches. To reduce the computational complexity and to extend their application to larger systems and longer timescales, these approaches often make use of simplifications such as coarse-graining, use of implicit solvent and rigid body models, and enhanced sampling. But still, each of these methods has specific limitations and they usually do not address some of the key aspects of protein-ligand binding. For example, BD simulations can effectively model the diffusional encounter of protein and ligand molecules based on long-range electrostatic interactions but they do not explain short-range effects such as desolvation effects, side-chain rearrangements and conformational changes during the induced-fit step. The more detailed MD simulations on the other hand, can address these important issues but due to their high computational requirements, they cannot be applied to longer timescales corresponding to drug residence times which usually range from a few seconds to a few hours. Therefore, there is a need to develop and apply methods that allow the more accurate prediction and effective investigation of protein-ligand binding kinetics in a systematic manner and to address specific challenges in investigating

protein-ligand binding kinetics. This current work aims to apply different physics-based and bio- and chemoinformatics approaches to predict k_{on} or k_{off} rates for protein-ligand binding and to investigate the mechanistic determinants of protein-ligand binding kinetics. In such direction, the present thesis addresses the following fundamental questions:

- *Can high-throughput regression-based quantitative structure-kinetics relationship (QSKR) models be derived for a series of compounds using only the structural information from their complexes with a specific protein or receptor?*
- *Can these QSKR models accurately predict kinetic parameters for novel compounds?*
- *What are the key protein-ligand interactions that distinguish ligands with slow and fast binding kinetics?*
- *How can residence times for drug molecules be prolonged by introducing specific interactions such as halogen-aromatic interactions between halogenated drugs and aromatic residues of kinases?*
- *How can continuum solvent and rigid-body based BD simulations be used to assist high-throughput prediction of diffusional association rate constants for binding of drug-like compounds to their receptors?*
- *How accurately can the τ RAMD enhanced sampling procedure based on molecular dynamics simulations predict relative residence times for a series of compounds?*

In addition to answering the above questions, I contributed to the development of a toolbox of computational methods: KBbox (<http://kbbox.h-its.org/toolbox/>) to help researchers to guide them to use different computational methods available to study molecular binding kinetics. This toolbox consists of a collection of tutorials, example cases and a theoretical overview of the current state-of-the-art methods and tools used for computing the kinetic parameters of protein-ligand binding.

1.4 Organization of the thesis

This thesis consists of 7 chapters. Chapter 2 provides the theoretical basis for the different bio- and chemoinformatics and simulation-based approaches used in this work. An overview and theoretical background of methods, such as COMBINE analysis, Brownian dynamics and molecular dynamics is discussed. An introduction to molecular mechanics force-fields and basic information about the continuum solvent models used in this work is given. Also, the techniques commonly employed in molecular simulations are briefly introduced and an introduction to MM/GBSA free-energy calculations, and Møller–Plesset energy calculations is given. Further, chapter 2 provides a short overview on different software and tools used in this thesis.

Chapter 3 describes the application of the COMBINE analysis approach to derive Quantitative Structure-Kinetics Relationships (QSKRs) for k_{off} rates of Heat shock protein 90 (HSP90) and HIV-1 protease inhibitors. Results on predictive QSKR models derived for inhibitors of these two therapeutically important targets and the important protein-ligand interactions that distinguish slow and fast off-rate compounds are presented and discussed.

Chapter 4 discusses the modulation of the residence times of inhibitors by targeting the interactions between halogen atoms, commonly found in drugs, and the aromatic residues typically found in the drug binding sites on proteins. Using haspin, a serine/threonine kinase as a model system and halogen substituted tubercidin inhibitors (close analogues of ATP) as model inhibitors, it has been suggested that residence times of inhibitors can be increased by introducing halogen-aromatic π interactions between the halogen atom of the inhibitors and the aromatic gatekeeper residues of kinases. Results from quantum chemical interaction energy calculations, MM/GBSA free-energy calculations and τ RAMD are compared to the experimental findings and are discussed.

In Chapter 5, a protocol to compute diffusional associational rates for small molecule binding to proteins, using BD simulations with SDA, is presented. Simulation parameters that were optimized for setting up BD simulation runs for diffusional association of protein and small molecules, are presented and a standard workflow to analyse and compute diffusional k_{on} rates from SDA is described. Results for validation of the protocol on inhibitors of different protein systems of varying complexity are also discussed.

Chapter 6 describes the implementation and content of the KBbox (<http://kbbox.h-its.org/toolbox/>), a toolbox of computational methods for studying the kinetics of molecular binding. The software architecture and implementation of KBbox is briefly introduced, followed by discussion on the organization of different content in KBbox. Some of the use-cases are also described.

This thesis concludes with chapter 7 that consists of a brief conclusion and future directions on the application of computational approaches to investigate kinetic parameters of protein-drug binding.

Chapter 2

Theoretical Methods and Software

This chapter gives a theoretical overview of the different chemoinformatics and chemometric methods, as well as the methods based on bimolecular simulations, that were employed in this work. The sequence of the methods described follows the order in which they were employed in the subsequent chapters. In the end, a short summary of each of the different software and tools that were employed is presented along with the scope of their application for this study.

2.1 COMparative BINding Energy (COMBINE) analysis

COMBINE analysis[46] is an approach for deriving quantitative structure-activity relationships (QSAR) by exploiting the information contained in the 3D structures of receptor-ligand complexes. In COMBINE analysis, the binding free energy, ΔG , or a related property (such as K_d , k_{off} , k_{on} , pK_i , pIC_{50}) is correlated with a subset of weighted interaction energy components determined from the structures of energy-minimized receptor-ligand complexes. These interaction energy components are typically Lennard-Jones (LJ) and Coulombic interaction energies decomposed on a per amino acid residue basis. The binding free energies are calculated from energy-minimized ligand-receptor complexes using a standard molecular mechanics force field.

$$\Delta G = \sum_{i=1}^{n^r} w_i^{LJ} u_i^{LJ} + \sum_{i=1}^{n^r} w_i^C u_i^C + C \quad (2.1)$$

Where ΔG is the binding free energy, u_i^{LJ} and u_i^C are intermolecular Lennard-Jones and Coulombic interaction energies calculated between each ligand and n^r amino acid residues of the protein, w_i^{LJ} and w_i^C are weights or coefficients of these LJ and Coulombic interaction energy terms. If, a sufficiently large number of molecules with known activities and 3D structures of ligand-receptor structures for these molecules is used in the training set, the weights of these interaction energy terms can be estimated by linear fitting. Due to the fairly large number of residual interaction terms used in the linear fitting, the use of standard multiple regression techniques is avoided, and instead partial least squares (PLS) analysis is applied to perform statistical analysis to determine the weights and constant C . If required, the ligands can also be further divided into n^l fragments, and thus the equation 2.1 can be rewritten as:

$$\Delta G = \sum_{i=1}^{n^r} \sum_{j=1}^{n^l} w_{ij}^{LJ} u_{ij}^{LJ} + \sum_{i=1}^{n^r} \sum_{j=1}^{n^l} w_{ij}^C u_{ij}^C + C \quad (2.2)$$

To perform COMBINE analysis, an energy matrix is generated where the columns represent each of these interaction energy terms (independent variables) and the rows correspond to each ligand in the training set. The inhibitory activities or binding kinetics (dependent variable) of these ligands are added to the final column in the matrix. Then, the PLS method [47][48] is used to maximize the linear correlation between the independent and the dependent variables by performing rotations of this matrix in the latent variables (LV) or the Principal Components (PC) space. In order to exclude energy terms that do not contribute to binding from the QSAR, a variable selection procedure is carried out. The variable selection procedure involves evaluation of the effects of each independent variable on the model predictivity and is carried out iteratively using a combination of D-optimal and fractional factorial designs[49]. In this thesis, we did not perform variable selection procedure. Rather, we have only used a pre-screening procedure where only those interaction energy terms which have standard deviation higher than a specified threshold value were selected for PLS regression and the rest of the energy terms showing little or no variance across the training dataset, were eliminated from statistical analysis.

The main advantages of correlating inhibitory activity or kinetics (ΔG , K_d , k_{off} , k_{on} , pK_i , pIC_{50} etc.) with residue-based interaction energy components over simply correlating with total computed binding energy is that the resultant COMBINE analysis model can help to highlight key interactions that are important to explain

the observed variances in the biological activity. This information could help in providing insights for predicting the effects of point mutations in the protein and for designing compounds with improved binding properties. In addition, during PLS analysis, errors either resulting from the modeled 3D structures or from force-field parameterization can be at-least partly filtered out. The calculation of weights also allow an implicit description of terms contributing to the dependent variable which are not explicitly included in the model.

2.1.1 Partial least squares (PLS) regression

PLS regression [47, 48] is a statistical approach used to determine a linear regression model by projecting both the dependent and independent variables to a new space. PLS combines features from both principal component analysis (PCA) and multiple linear regression (MLR). As in PCA, orthogonal Principal Components (PCs) are extracted and a fitting procedure similar to MLR is performed to describe the response variable (biological activities of compounds). PLS is used to model the fundamental relations between two matrices, X and Y , by finding the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. In COMBINE analysis, the X matrix consists of independent variables which are interaction energy terms and, optionally, additional variables, such as desolvation energy terms.

$$X = \begin{pmatrix} C_1^1 & C_2^1 & \dots & C_M^1 & L_1^1 & L_2^1 & \dots & L_M^1 \\ C_1^2 & C_2^2 & \dots & C_M^2 & L_1^2 & L_2^2 & \dots & L_M^2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ C_1^N & C_2^N & \dots & C_M^N & L_1^N & L_2^N & \dots & L_M^N \end{pmatrix}$$

where C_j^i and L_j^i are the Coulombic and Lennard-Jones variables, respectively. N is the number of compounds and M is the number of residues in the protein.

The Y matrix consists of dependent variables i.e. the activities of the compounds.

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

where y_i is the individual activity of compound i . In PLS, the X and Y matrices are decomposed into one score matrix, T , and two different loading matrices, P and Q (see equation 2.3)

$$X = TP^T \quad \text{and} \quad Y = TQ^T \quad (2.3)$$

The score matrix T contains information about the projections of compounds onto the PCs. The PC space is normalized and has a mean of zero. The compounds which behave as outliers usually have higher scores. The loading matrices P and Q contain information about the variables in the so-called latent variables (LV) or PC space. Latent variables are orthogonal vectors obtained as linear combinations of the original variables in the X matrix. The coefficients in a given PC provide information on the relative weights of the different terms and it can be useful to deduce the importance of each individual ligand–residue interaction to explain the variance in activity. The quality of the fit for the training set of compounds can be evaluated using the regression coefficient (R^2), cross-validation correlation coefficients (Q^2), average absolute errors (AAE) and root-mean squared errors (RME) (see equation 2.4 to equation 2.9).

$$R^2 = \frac{\left[\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \langle \hat{y}_i \rangle) \right]^2}{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (\hat{y}_i - \langle \hat{y}_i \rangle)^2} \quad (2.4)$$

$$Q^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (2.5)$$

$$AAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (2.6)$$

$$RME = \frac{1}{N} \sum_{i=1}^N \sqrt{(\hat{y}_i - y_i)^2} \quad (2.7)$$

Where, \bar{y} is the average value of the experimental activities (y_1, y_2, \dots, y_N)

$$\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i \quad (2.8)$$

And, $\langle \hat{y} \rangle$ is the average value of the calculated activities ($\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N$)

$$\langle \hat{y} \rangle = \frac{1}{N} \sum_{i=1}^N \hat{y}_i \quad (2.9)$$

2.2 Brownian Dynamics

Brownian dynamics (BD) simulations are used to simulate the diffusional processes and dynamics of particles, such as proteins, that undergo Brownian motion. In BD, rigid body representations of interacting particles are used. The effect of solvent is modelled using a continuum, implicit solvent model and the stochastic and friction effects of surrounding water molecules and ions are introduced by additional terms in the motion equation. The flexibility of the interacting molecules can be simulated using a coarse grained force field or by switching between conformations on the fly. Since, the internal flexibility of the proteins is generally ignored, the conformational changes upon binding of the ligand cannot be captured in BD. However, having fewer degrees of freedom due to the rigid-body representation in implicit solvent, allows longer time scales (μs - ms range) of diffusional processes to be simulated and large systems (several μm) can also be studied using BD. With BD simulations, it is also possible to compute kinetic parameters of binding processes which is otherwise not possible with MD simulations. But, the computation of binding kinetics is generally limited to diffusional encounter complexes rather than bound complexes, as short-range effects are not modelled in BD.

2.2.1 The concept of Brownian motion

Brownian dynamics is named after Robert Brown who observed the random motions of pollen particles in water and he suggested that the stochastic collisions of particles with the solvent (water) resulted in their random motions. The diffusive displacement (Δr) of a particle in 3D in BD for a given time-step (Δt) can be determined from the following equation given by Einstein[50] and Smoluchowski[51].

$$\Delta r^2 = 6D\Delta t \tag{2.10}$$

Here, D is the translational diffusion coefficient of the particle and for spherical objects, it can be calculated using the following formula:

$$D = \frac{k_B T}{6\pi\eta a} \tag{2.11}$$

Here, k_B is the Boltzmann constant, T is the temperature in Kelvin, η is the viscosity of the solvent and a is the hydrodynamic radius of the particle.

Simulating diffusional encounter

Ermak and McCammon developed an algorithm to model Brownian motion for diffusional encounter of proteins by considering both translational (Δr) and rotational (Δw) motions[52].

The translational (Δr) displacement is given by:

$$\Delta r = (k_B T)^{-1} D_T F \Delta t + R \quad (2.12)$$

And, the rotational (Δw) displacement is given by:

$$\Delta w = (k_B T)^{-1} D_R \mathcal{T} \Delta t + \Theta \quad (2.13)$$

Here, D_T and D_R are translational and rotational diffusion coefficients of the molecule respectively, F and \mathcal{T} are the position-dependent interaction force and torque acting on the molecule and they are computed prior to taking the step Δt . R is the random displacement and it should satisfy the following conditions.

$$\langle R \rangle = 0 \quad (2.14)$$

$$\langle R^2 \rangle = 6D\Delta t \quad (2.15)$$

Similarly, Θ is the random rotational angle and it should also satisfy:

$$\langle \Theta \rangle = 0 \quad (2.16)$$

$$\langle \Theta^2 \rangle = 6D\Delta t \quad (2.17)$$

For simulating diffusional protein-ligand association in any BD software such as SDA, one of the molecules (usually a protein) is kept fixed and the movement of the ligand is simulated. Therefore, to account for the diffusive motion of both of the interacting molecules, the relative translational diffusion coefficient is used in the calculations.

2.2.2 Simulation of Diffusional Association (SDA) software

SDA (<https://mcm.h-its.org/sda/>) is a software package that can be used to run Brownian dynamics simulations of the diffusional association of solute molecules

in a continuum implicit solvent[53, 54]. Using rigid-body structures and a suitable force field, SDA can be used to apply BD simulations to calculate binding kinetics for protein-protein or protein-ligand association. It can also be used to perform rigid-body docking to record diffusional encounter complexes and to calculate bimolecular electron transfer rate constants. In addition, protein flexibility and hydrodynamic interactions can be introduced in SDA to account for protein’s internal motion and induced solvent effects. The theoretical background on how different interaction forces are computed in SDA, is outlined in the following sections.

Calculation of interaction forces

The diffusional association of interacting molecules is modelled as mainly driven by the Poisson-Boltzmann (PB) equation-derived electrostatic interaction and limited by the presence of exclusion forces. In SDA, electrostatic interaction, electrostatic desolvation and hydrophobic (non-polar) desolvation energies are used to simulate diffusional association of two solute molecules[54]. The total interaction energy (ΔG^{1-2}) between these solutes is calculated using the following equation in SDA:

$$\Delta G^{1-2} = \Delta G_{el}^{1-2} + \Delta G_{edesolv}^{1-2} + \Delta G_{DH}^{1-2}(r) + \Delta G_{np}^{1-2} + \Delta G_{rep}^{1-2} \quad (2.18)$$

Here, the first 3 terms approximate the Poisson-Boltzmann equation derived electrostatic energy between a pair of solutes and the last 2 terms account for the non-polar interactions between solutes and exclusion forces. ΔG_{el}^{1-2} is the long-range electrostatic interaction energy, $\Delta G_{edesolv}^{1-2}$ is the short-range electrostatic desolvation energy and ΔG_{DH}^{1-2} is the distance-dependent Debye-Hückel correction to the long-range energy term, accounting for the use of finite-sized grids in the calculations. ΔG_{np}^{1-2} is the non-polar desolvation energy which account for the change in the total solute-solvent interface area upon binding of solutes[55]. ΔG_{rep}^{1-2} is soft-core repulsion term that describes exclusion forces by applying a continuous, repulsive potential which prevents solutes from overlapping[56]. The exclusion forces can also be modelled as hard-core repulsion by defining an exclusion grid.

Electrostatic interactions and effective charges

The electrostatic potential (Φ) of a solute can be computed by solving the non-linear second-order PB equation which describes the distribution of electric potential in a

non-uniform dielectric. This equation (2.19) can be linearized and solved numerically for discrete grid points in the system r .

$$-\nabla \cdot (\vec{\epsilon}(r) \nabla \Phi(r)) = \rho(r) + \sum_i c_i q_i e^{-\frac{q_i \Phi}{k_B T}} \quad (2.19)$$

Here $\vec{\epsilon}(r)$ is position-dependent dielectric constant, $\Phi(r)$ is the electrostatic potential, $\rho(r)$ is molecular charge density, c_i is the concentration and q_i is the charge of ions in the solvent. k_B is the Boltzmann constant and T is the temperature in Kelvin. The Adaptive Poisson-Boltzmann Solver (APBS)[57] and University of Houston Brownian Dynamics (UHBD)[58] are two of the most commonly used programs to calculate electrostatic potentials of biomolecules by numerically solving the Poisson-Boltzmann equation. We have used APBS for generating electrostatic potentials of different protein and ligand systems used in this thesis.

Since it is computationally very expensive to calculate electrostatic interaction free energy between a pair of solutes at each time-step of the BD simulation, SDA uses the Effective Charge Model (ECM)[59] to approximate the PB theory derived electrostatic interaction. In this model, the electrostatic interaction energy between a solute pair (ΔG_{el}^{1-2}) is calculated as the interaction between PB derived electrostatic potential (Φ_{el}) of one solute and a set of effective charges (q_i) on the other solute, and vice-versa. The total interaction energy is multiplied by a factor of 1/2 to prevent double counting of the interaction. These effective charges are fitted in such a way that, they reproduce the electrostatic potential derived with PB in a heterogeneous dielectric medium, when they are placed in an uniform dielectric medium. For proteins, these effective charges are assigned on Lys, Arg, Glu, Asp residues, and C and N-termini of proteins. For small molecules, these charges are assigned on hydrogen bond donor-acceptor atoms (N,O,F,S), halogen atoms (Cl,Br,I) and on P and Fe atoms in case of co-factors.

$$\Delta G_{el}^{1-2} = \frac{1}{2} \sum_{i_1} q_{i_1} \Phi_{el_2}(r_{i_1}) + \frac{1}{2} \sum_{i_2} q_{i_2} \Phi_{el_1}(r_{i_2}) \quad (2.20)$$

Here, q_{i_n} is an effective charge on solute n and $\Phi_{el_m}(r_{i_n})$ is the electrostatic potential of solute m , at the position of effective charge q_{i_n} on solute n .

Electrostatic desolvation interaction

Binding of two solutes results in the exclusion of high dielectric solvent from the binding interface, which results in desolvation of the surface-lying charges. This creates an unfavorable contribution to the binding interaction between two solutes. This unfavorable penalty is corrected in SDA by an extended ECM model which includes an electrostatic desolvation correction term ($\Delta G_{edesolv}^{1-2}$), which is calculated as:

$$\Delta G_{edesolv}^{1-2} = \frac{1}{2} \sum_{i_1} q_{i_1}^2 \Phi_{edesolv_2}(r_{i_1}) + \frac{1}{2} \sum_{i_2} q_{i_2}^2 \Phi_{edesolv_1}(r_{i_2}) \quad (2.21)$$

Here, $\Phi_{edesolv_m}$ is the the electrostatic desolvation potential of solute m and it accounts for the effect of reduction in the dielectric constant at the interface. The electrostatic desolvation potential of a solute at point r is calculated by using the following formula.

$$\Phi_{edesolv}(r) = \alpha \frac{\epsilon_s - \epsilon_p}{\epsilon_s(2\epsilon_s + \epsilon_p)} \sum_j a_j^3 \frac{(1 + \kappa r_j)^2}{r_j^4} e^{-2\kappa r_j} \quad (2.22)$$

Here, α is an empirical scaling parameter, ϵ_s and ϵ_p are the dielectric constants of solvent and solute respectively, a_j is the radius of atom j and κ is the inverse of Debye length and it depends on ionic strength of the solvent.

Nonpolar desolvation interaction

The binding of two solutes at the interface leads to the reduction of total solute-solvent interface area which results in increased binding affinity. This interaction is modelled in SDA by nonpolar desolvation interaction energy[55] which is proportional to the solvent accessible surface area (SASA) of a solute that is obstructed by the interacting solute.

$$\Delta G_{np}^{1-2} = \sum_{i_1} SASA_{i_1} \Phi_{np_2}(r_{i_1}) + \sum_{i_2} SASA_{i_2} \Phi_{np_1}(r_{i_2}) \quad (2.23)$$

Here, $\Phi_{np_m}(r_{i_n})$ is the non-polar burial potential of solute m at the position of atom i_n of interacting solute n and $SASA_{i_n}$ is solvent-accessible surface area of surface atom i_n .

2.2.3 Calculation of protein-ligand association rates in SDA

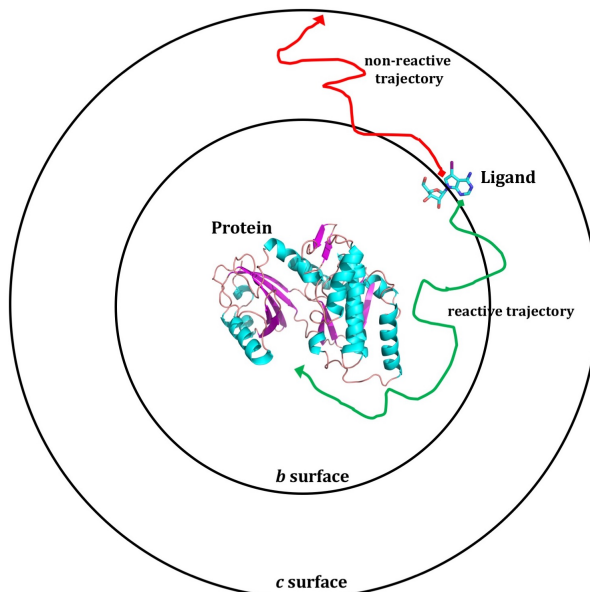


Figure 2.1: Schematic representation of the geometric setup for the diffusional association of protein and ligand molecules in BD simulations.

Association rate constants of two interacting molecules can be computed from BD simulations when the encounter is diffusion-controlled. The steady state rate constant $k_D(b)$ of two spherical molecules approaching each other at a separation distance $r = b$, can be given by the analytical Smoluchowski expression[60].

$$k_D(b) = \frac{4\pi D}{\int_b^\infty \frac{e^{\frac{E(r)}{k_B T}}}{r^2} dr} \quad (2.24)$$

Here D is the relative diffusion constant, $E(r)$ is the interaction potential acting between two spherical molecules. To obtain the association rate constant k_{on} of protein-ligand association in BD simulations, the steady-state rate constant from the previous equation is multiplied with the probability of formation of protein-ligand encounter complex β^∞ . BD trajectories are run starting at a relatively large protein-ligand separation b , where the centrosymmetric forces between protein and ligand are negligible. Each trajectory is stopped when either reaction criteria are satisfied or the molecules reach a much larger separation c (Figure 5.11). In case, the reaction conditions are satisfied, the trajectory is considered as reactive. In BD simulations, thousands of trajectories are run and β , the fraction of reactive

trajectories is calculated. To account for the possibility that in case of non-reactive trajectories, ligand may come back from c surface and form an encounter complex with protein, β is corrected by a multiplying factor Ω to get β^∞ [61].

$$k_{on} = k_D(b)\beta^\infty \quad (2.25)$$

$$\beta^\infty = \frac{\beta}{1 - (1 - \beta)\Omega} \quad (2.26)$$

Here, Ω describes the probability that the ligand at separation distance $c > b$ returns to b . It is given by:

$$\Omega = \frac{k_D(b)}{k_D(c)} \quad (2.27)$$

By substituting the values of β^∞ and Ω , the equation 2.28 can be written as:

$$k_{on} = k_D(b) \frac{\beta}{1 - (1 - \beta) \frac{k_D(b)}{k_D(c)}} \quad (2.28)$$

2.3 Molecular Dynamics (MD) technique

Molecular Dynamics (MD) simulation is one of the most common computer simulation techniques to monitor time-dependent processes of biological molecules. MD is widely used to gain insights into the molecular mechanisms of dynamic processes such as protein-folding, membrane transport, self-assembly and for studying thermodynamic properties and kinetics of bimolecular association. In MD simulation, movements for set of atoms and molecules are computed by numerically solving Newton's equations of motion as a function of time:

$$m_i \frac{\partial^2}{\partial t^2} \vec{r}_i = -\nabla E(\vec{r}_i) \quad (2.29)$$

where m_i and \vec{r}_i are the mass and position of the particle i , E is the total potential energy which depends on the positions of all particles in the system.

In order to run MD simulation, initial position \vec{r}_i and velocity \vec{v}_i of particles is determined for $t = 0$ and a short time step Δt is chosen. The force \vec{F} acting on each particle can be calculated from the total potential energy E as:

$$\vec{F}_i = -\frac{\partial E}{\partial \vec{r}_i} \quad (2.30)$$

And, the acceleration \vec{a}_i of each particle can be calculated from the force \vec{F}_i as:

$$\vec{a}_i = \frac{\vec{F}_i}{m_i} \quad (2.31)$$

The particles are moved for time Δt and a new set of positions is computed for the next time step ($t + \Delta t$):

$$\vec{r}_{i(t+\Delta t)} = \vec{r}_i + \vec{v}_i \Delta t + \frac{1}{2} \vec{a}_i \Delta t^2 \quad (2.32)$$

Therefore, for each MD time step, forces and velocities of particles are calculated and integrated for next time step. The iteration of the above process for subsequent time-steps gives the spatio-temporal evolution of the system. Since, small scale motions such as vibration of bonds or hydrogen atoms occur very fast, the time step of the integration chosen is usually very small (1 *fs*). For study of large scale motions, dynamics of bonds with hydrogen atoms becomes no longer important. Therefore, algorithms such as SHAKE[62] are used to constrain bonds with hydrogens which allow a bigger time step (2 *fs*) to be used for MD simulation, thereby resulting in increased computational efficiency.

Integration of Newton's equation of motion

To integrate Newton's equation of motion at finite time steps, a numerical integrator is required. Verlet[63] is one of the most commonly used integrators in MD simulations. The Verlet algorithm uses the Taylor expansion to approximate the particle's position and dynamic properties, where new positions $\vec{r}_{t+\Delta t}$ and acceleration at time $t + \Delta t$ are determined from the positions and acceleration at time t and from the positions of the previous step $\vec{r}_{t-\Delta t}$:

$$\vec{r}_{t+\Delta t} = \vec{r}_t + \vec{v}_t \Delta t + \frac{1}{2} \vec{a}_t \Delta t^2 + \dots \quad (2.33)$$

$$\vec{r}_{t-\Delta t} = \vec{r}_t - \vec{v}_t \Delta t + \frac{1}{2} \vec{a}_t \Delta t^2 - \dots \quad (2.34)$$

Summing both equations:

$$\vec{r}_{t+\Delta t} = 2\vec{r}_t - \vec{r}_{t-\Delta t} + \vec{a}_t \Delta t^2 \quad (2.35)$$

And, the velocity is calculated as:

$$\vec{v}_{t+\Delta t} = \frac{[\vec{r}_{t+\Delta t} - \vec{r}_{t-\Delta t}]}{2\Delta t} \quad (2.36)$$

The Verlet algorithm is time-reversible which means that if the direction of velocities of particles is reverse, the simulation will run in the reverse direction.

Molecular Mechanics force fields

A molecular mechanics force field is a set of energy functions or equations and the associated constants to define the potential energy of a molecular system as a function of its three-dimensional structure in molecular mechanics or molecular dynamics simulations. A force field is required in MD simulations to describe the time evolution of both bonding terms (such as bond lengths, bond angles, torsions) and the non-bonding electrostatic and van der Waals interactions between atoms. During MD simulations, pairwise potentials between atoms are calculated from this common potential energy function that describes both intramolecular and intermolecular interactions. Force fields are always optimized for specific classes of molecules, for example, AMBER *ff14SB*[\[64\]](#) was optimized for proteins and nucleic acids and *MMFF94*[\[65\]](#) was optimized for small organic molecules.

A force field typically includes set of following energy functions:

$$E_{\text{total}} = \sum_{\text{bonds}} K_r(r - r_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \\ + \sum_{i < j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}} \right] + \sum_{\text{H-bonds}} \left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right] \quad (2.37)$$

Temperature and pressure control in MD simulations

Like experiments, it is also important to describe the thermodynamic state of the system in MD simulations. Macroscopic properties that represent the thermodynamic state of a system include number of particles (N), pressure (P), temperature (T) and volume (V). An ensemble that describes a macroscopic or thermodynamic state is a collection of different microscopic properties. MD simulations can be run using different ensembles such as canonical (NVT), isothermal-isobaric (NPT), and microcanonical (NVE) ensembles. In MD simulations, integration of Newton's equation of motion results in a microcanonical NVE ensemble with constant number of particles, constant volume and constant energy. However, biological experiments

are usually performed at constant temperature and constant volume (NVT) and/or constant temperature and constant pressure (NPT). MD simulations therefore need to be performed with any of these ensembles to compare the simulation results with experiments.

For NVT ensemble, temperature need to be maintained constant using external thermal bath and for this there are several thermostat algorithms available that can be used to keep the temperature constant to any desired value. Berendsen weak-coupling method[66], Nosé–Hoover and Langevin piston are some of the commonly used thermostats in MD simulations. In the case of NPT ensemble, the constant pressure is maintained by allowing the volume of the simulation box to change using external barostats. The Nosé–Hoover Langewin piston[67] method, Berendsen weak-coupling[66] and Parinello-Rahman[68] methods are some of the common barostats used to maintain NPT conditions in MD simulations.

Implicit and Explicit Solvent Models

In order to model a realistic system, it is crucial to define the solvent environment of a biological system appropriately. For MD simulations of biomolecules, solvent effects can be treated using either explicit water model or implicit water model. In explicit solvent model, large number of water molecules are explicitly included in the simulation. This leads to enormous increase in the computational requirement, especially if longer time scales need to be simulated. Transferable intermolecular potential n point models (TIPnP)[69] from Jorgensen group and extended simple point charge model (SPC/E)[70] from Berendsen group are the most commonly used explicit water models. These models are based on standard non-polarizable force fields and they have been parameterized to reproduce the characteristic properties of real liquid water, such as density, diffusivity, energy and dielectricity. In this work, the TIP3P water model has been used to model the explicit water in MD.

Implicit solvent models tend to greatly reduce the number of degrees of freedom by using the continuum approximation of the discrete solvent, where the effect of solvent on solute molecules is described by a set of polar and non-polar terms included in the equation of energy. Using implicit solvent model has several advantages over explicit model such as reduced computational cost, instantaneous thermodynamic equilibrium with the solute, and enhanced conformational sampling. Poisson-Boltzmann (PB) and Generalized Born (GB)[71] are the most widely used implicit

solvent models in MD simulations of biomolecules. The GB model is an empirical approximation to the linear Poisson- Boltzmann equation. The GB model also includes the charge screening effects caused by ions and salt and can describe the solvent effects in MD fairly well but with very low computational cost. In order to choose the optimal water model for the simulations, one must consider several factors such as the molecular structure of biomolecules, available computational resources and the specific questions related to the calculation, that are being addressed.

Periodic boundary conditions (PBC)

MD simulations are used to simulate finite systems with a finite number of molecules. However, to predict properties at the bulk level, a relatively smaller system is used and periodic boundary conditions (PBC) are applied in MD to make the system look like an infinite one. By applying PBCs, a simulation box of unit size is replicated in all directions and it is assumed that all the molecular properties are identical in each of the unit cells. Existence of PBCs means that, if a particle exits from one side of the simulation box, then it is replaced by an image particle entering from the opposite side of box, thereby keeping the total number of particles constant in the simulation. Therefore, PBC helps to avoid any boundary effects caused by the box-edges and provides a homogeneous system to simulate bulk effects. The most common box shapes used in MD with PBCs are cubic, hexagonal, octahedron and rhombic dodecahedron. Using PBCs also has some limitations, for example, any fluctuations in the system with wavelength greater than the unit cell are not possible to observe. Moreover, the size of the simulation box chosen should be large enough to avoid any periodic artifacts caused by the artificial long-range interactions with the image molecule.

2.3.1 τ Random Acceleration Molecular Dynamics (τ RAMD)

τ RAMD is an enhanced sampling procedure based on MD simulations developed to compute relative residence times of drug-like compounds and to explore ligand exit pathways from the buried binding sites in proteins[29]. RAMD[28] simulations are performed in an explicit solvent with parameters similar to the standard MD simulations. In RAMD, during MD simulations of the bound protein-ligand complex, a small additional randomly oriented force is applied to the centre of mass of the ligand to accelerate its unbinding from the binding site. The movement of the ligand is assessed at regular time intervals and the direction of the force is reassigned randomly if the ligand’s movement is smaller than the specified threshold distance. In RAMD simulations, ligands that have higher residence times take longer to egress from the binding pocket or require application of a stronger force to exit within a specified simulation time. By applying this artificial force, the unbinding of the ligands from the binding site can be observed in short simulations of few nanoseconds. Therefore, RAMD may be very useful and computationally efficient approach to obtain relative estimates of residence times. The main advantage of RAMD over other enhanced sampling methods such as Smoothed potential MD[26] or Metadynamics[25], is that it does not require extensive parametrization or any prior knowledge of the dissociation pathway. The magnitude of the random force is the only parameter that needs to be set by the user in RAMD simulations and it should be carefully chosen within a reasonable range, so that it does not affect the computed relative residence times.

2.3.2 MM/GBSA free-energy calculations

In the molecular mechanics generalized Born surface area (MM/GBSA) method[72, 73], the binding free energy of a ligand to a protein to form a complex is obtained as the difference:

$$\Delta G^{bind} = G^{complex} - G^{receptor} - G^{ligand} \quad (2.38)$$

The free energy of each of the molecular systems is given by the expression:

$$G = E_{bnd} + E_{el} + E_{vdW} + G_{pol} + G_{np} - TS \quad (2.39)$$

where E_{bnd} , E_{el} and E_{vdW} are the standard molecular mechanics energy terms accounting for bonded, electrostatic and van der Waals interactions, respectively, in the gas phase. G_{pol} and G_{np} are polar and non-polar contributions to the solvation free energies, and S is the entropy contribution arising from changes in the dynamics upon ligand binding and it is calculated by a normal-mode analysis of the vibrational frequencies. In MM/GBSA, the generalized Born (GB) model is used to estimate G_{pol} , whereas G_{np} is obtained from a linear relation to the solvent accessible surface area (SASA).

2.3.3 Møller–Plesset energy calculations

In quantum chemistry, Moeller-Plesset perturbation theory (MP)[74] is one of the post-Hartree-Fock *ab initio* methods commonly implemented in many computational chemistry software packages. The Moeller-Plesset perturbation theory[74] improves on the Hartree-Fock method by adding electron-correlation effects using Rayleigh-Schrödinger perturbation theory to different orders (MP2, MP3, MP4 etc.) However, MP theory is not variational which means that the energy calculated by MP theory may be lower than the true ground state energy. *Ab initio* interaction energies using Møller-Plesset perturbation theory to second order (MP2)[75] were calculated using the GAMESS software[76], and partitioned into their constituent interaction energy terms using the many body interaction energy decomposition scheme (EDS) described by Góra *et al.*[77, 78]. In this scheme, the total interaction energy is calculated in a super-molecular approach as the difference between the total energy of a complex and the sum of the energies of its isolated constituents. In all calculations, the complex centered basis set (CCBS) was used consistently and the results are therefore basis set superposition error (BSSE) free due to the full counterpoise correction.

The total MP2 interaction energy (E_{MP2}) includes the components of the Hartree-Fock interaction energy (E_{SCF}) and the second order Coulomb correlation correction term (E_{CORR}). This correlation energy term (E_{CORR}) includes the second order intermolecular dispersion energy and the correlation corrections to the SCF components.

$$E_{MP2} = E_{SCF} + E_{CORR} \quad (2.40)$$

The Hartree-Fock interaction energy (E_{SCF}) was partitioned into a first order

Heitler-London component (E_{HL}) and a higher order Hartree-Fock delocalization interaction energy component (E_{SDEL}), which encompasses the induction and the associated exchange effects. Because their separation could lead to a non-physical charge transfer, this component was not partitioned any further.

$$E_{SCF} = E_{HL} + E_{DEL} \quad (2.41)$$

The Heitler-London interaction energy component (E_{HL}) can be separated into the first-order electrostatic interactions (E_{EL}) of monomers and the associated Heitler-London exchange repulsion energy (E_{EX}) due to the Fermi electron correlation effects. The electrostatic interaction energy (E_{EL}) was obtained as a first-order term in the polarization perturbation theory and the exchange repulsion term (E_{EX}) was calculated by subtracting the electrostatic interaction energy from the Heitler-London energy ($E_{EX} = E_{HL} - E_{EL}$).

$$E_{HL} = E_{EL} + E_{EX} \quad (2.42)$$

$E_{EL,MTP}$ refers to the electrostatic multipole component estimated from an atomic multipole expansion, $E_{EL,PEN}$ is the electrostatic penetration energy calculated by subtracting the electrostatic multipole component from the electrostatic interaction energy ($E_{EL,PEN} = E_{EL} - E_{EL,MTP}$)

$$E_{EL} = E_{EL,MTP} + E_{EL,PEN} \quad (2.43)$$

2.4 Software and Tools

In this project, a number of software packages and visualization and modelling tools was employed for a range of different applications such as running MD and BD simulations, modelling of ligands and protein-ligand complexes, visualization and analysis of 3D structures and trajectories from MD and BD simulations. For doing data analysis and to automate some of the tasks, several scripts were written using the Python programming language and bash scripting. Some tcl scripts, previously developed in the group, were also used for the running and analysis of RAMD simulations. All the plots used in this thesis were generated either using the Gnuplot program or with Microsoft Excel 2016. All the major tools/software employed and their application areas are briefly discussed below.

2.4.1 Simulation software

AMBER (Assisted Model Building with Energy Refinement)

AMBER refers to a suite of programs (<http://ambermd.org/>) used for running molecular dynamics simulations of proteins and nucleic acids[79]. It has a number of tools and program for the preparation of necessary input files, to setup and perform molecular dynamics simulations and to analyze the simulation results. AMBER has an efficient parallel scaling implementation making it one of the most widely used programs for biomolecular studies. The name Amber also refers to a set of molecular mechanics force fields used for the simulation of biomolecules. In this thesis, Amber was used for the preparation of topology files and the energy minimization of protein-ligand complexes for COMBINE analysis, for running MM/GBSA simulations, energy-minimization and equilibration of protein-ligand complexes before running RAMD simulations.

Version used: AMBER 14

NAMD (NAnoscale Molecular Dynamics program)

NAMD is a molecular dynamics simulation package (<http://www.ks.uiuc.edu/Research/namd/>) designed for high-performance simulation of large biomolecular systems[80]. NAMD includes a rich set of MD features such as multiple time stepping, constraints, and dissipative dynamics and can be used with the AMBER and CHARMM potential functions, parameters, and file formats. The code is highly parallelized as it can scale to thousands of processors on high-end parallel platforms. It can also be run on individual desktops and laptops. In addition, NAMD can be connected to the molecular graphics software VMD in order to provide an interactive simulation tool for modifying and viewing the running MD simulations. The τ RAMD[29] procedure has been implemented in the NAMD software using *tcl* scripts and, in this thesis, NAMD was used to run RAMD simulations of protein-ligand dissociation for inhibitors of haspin kinase.

Version used: NAMD 2.9

SDA (Simulation of Diffusional Association)

SDA[54, 53] is a software package (<http://mcm.h-its.org/sda7/>) to carry out Brownian dynamics simulations of the diffusional association of solute molecules (e.g.

proteins) in a continuum aqueous solvent. It can also be used to perform rigid-body docking to record Brownian dynamics trajectories or encounter complexes and to calculate bimolecular rate constants. In SDA, the interaction between the solutes is given by an approximation to the Poisson-Boltzmann equation-derived electrostatic interaction[59]. In addition, short-ranged hydrophobic desolvation and electrostatic desolvation forces can also be considered. In SDA, simulation of the diffusion of multiple proteins, in dilute or concentrated solutions can also be performed to study macromolecular crowding effects. In this work, SDA was used to perform Brownian dynamics simulations of protein and ligand association for calculation of diffusional association rate constants.

Version used: SDA7.1

2.4.2 Structure preparation and general molecular modeling tools

Schrödinger suite

Schrödinger (<https://www.schrodinger.com/>) is a suite of tools used for molecular modeling, drug-discovery and materials science research. In this thesis, Schrödinger was used to pre-process the structures of the protein-ligand complexes, to add missing side chains, to add disulphide bonds, and for optimizing the H-bond network to assign hydrogen atom positions.

Version used: Release 2015-4

MOE (Molecular Operating Environment)

MOE is an interactive integrated suite of applications (<https://www.chemcomp.com/>) that provides a wide range of functionality to support Molecular Modelling, Chemoinformatics, Structure- based Design, Virtual Screening and a broad range of life science applications. In this thesis, MOE was used to aid in preparation of protein structures, modeling of ligands and deciding correct protonation states for ligands and titratable residues.

Version used: MOE 11

AmberTools

AmberTools consists of several tools/packages that are either used independently or with the Amber program[79]. AmberTools include programs/tools to generate force fields for general organic molecules and metal centers, preparation programs for Amber simulations, programs for semi-empirical and DFTB quantum chemistry calculations, tools to compute numerical solutions to Poisson-Boltzmann models, programs for structure and dynamics analysis of trajectories. In this thesis, several programs from AmberTools, such as *antechamber*, *sqm*, *parmchk*, *RESP* and *LEaP*, were used to generate partial atomic charges for small molecules and to generate force field parameters for protein-ligand complexes, *ambpdb* was used for interconversion between different file formats, *MMPBSA.py* was used for free-energy calculations of haspin-inhibitor complexes, *cpptraj* was used for analysis of MD and RAMD trajectories from Amber and NAMD simulations.

Version used: AmberTools14

APBS (The Adaptive Poisson-Boltzmann Solver)

APBS (<http://www.poissonboltzmann.org/>) is a macromolecular electrostatics calculation program used for solving the equations of continuum electrostatics for large biomolecular systems[81]. The results of APBS calculations can be displayed as an electrostatic potential molecular surface using PyMOL. Most of the APBS functionality is available through the online *PDB2PQR* web server (http://nber-222.ucsd.edu/pdb2pqr_2.1.1/). In this thesis, APBS was used to generate electrostatic potential grids of protein and ligand molecules for BD simulations and the *PDB2PQR* webserver was used to generate PQR files for protein molecules.

Version used: APBS 1.4.1

HYDROPRO

HYDROPRO (<http://leonardo.inf.um.es/macromol/programs/hydropro/hydropro.htm>) is a computer program used to compute the hydrodynamic properties of rigid molecules (proteins, small nucleic acids, macromolecular complexes, etc.) from their atomic-level structure[82]. The HYDROPRO output comprises the basic hydrodynamic properties: translational diffusion coefficient, sedimentation coefficient, intrinsic viscosity, and relaxation times, along with the radius of gyration. In this thesis,

HYDROPRO was used to compute translational and rotational diffusion coefficients of protein and ligand molecules required for simulating diffusional association with SDA.

Version used: HYDROPRO10

2.4.3 Structure Visualization tools

PyMOL

PyMOL (<https://www.pymol.org/>) is a molecular visualization software used for the manipulation of structures and generating high quality 3D images of biological macromolecules, such as proteins. It also provides some basic functions that can be used to analyze molecular and chemical properties of biomolecules. In this thesis, we have used PyMOL for visual inspection of the protein-ligand complexes and PDB structures, and the creation and labelling of the crystallographic images.

Version used: PyMOL 1.7

VMD (Visual molecular dynamics)

VMD (<http://www.ks.uiuc.edu/Research/vmd/>) is another very commonly used visualization program designed for modelling, visualization and analysis of biological systems, such as proteins, nucleic acids, lipid bilayer assemblies[83]. Most importantly, VMD can be used to view and analyze the results of MD simulations and to visualize potential grids of the molecules. In this thesis, VMD was used to analyze the trajectories from RAMD and MD simulations and to visualize potential grids of proteins and ligands generated for running BD simulations in SDA.

Version used: VMD 1.9.3

Chapter 3

Quantitative structure-kinetics relationships (QSKRs) for k_{off} values of HSP90 and HIV-1 protease inhibitors

This Chapter is based on the following publication:

Prediction of Drug–Target Binding Kinetics by Comparative Binding Energy Analysis.

Gaurav K. Ganotra and Rebecca C. Wade, *ACS Medicinal Chemistry Letters* **2018** 9 (11), 1134-1139

DOI: 10.1021/acsmchemlett.8b00397

Quantitative structure activity relationships (QSARs) allow correlation of the physio-chemical and structural descriptors/properties of a class of molecules with their biological activities by applying regression-based machine learning techniques. Over the time, a number of classical regression techniques have been developed and successfully applied to derive QSARs for series of molecules[84, 85]. As more number of three-dimensional (3D) structures of ligand-protein complexes is becoming available, these QSAR approaches have been extended in three dimensions to derive 3D-QSARs by incorporating information on ligand and protein interactions into the models[86, 87, 88, 89]. COMparative BINDing Energy (COMBINE) analysis is one of such medium-throughput approaches that has been successfully ap-

plied to a number of protein targets to derive target specific scoring functions for the prediction of binding affinity and target selectivity[46, 90, 91, 92, 93, 94, 95]. In COMBINE analysis, ligand-receptor interaction energies are computed using a molecular mechanics model. These energies are then partitioned and subjected to regression-based methods, such as Partial Least Squares (PLS) regression, to derive a statistical model which relates the property of interest to weighted selected components of the ligand-receptor interaction energy. COMBINE analysis seeks to make complete and systematic use of the available information from 3D structures of receptor–ligand complexes and the measured bioactivities of compounds, by explicitly including information about the receptor–ligand interaction energies. This is in contrast to other 3D-QSAR approaches such as Comparative molecular field analysis (CoMFA)[86] or Molecular Similarity Indices in a Comparative Analysis (CoMSIA)[96] that only include information about the interaction properties of the ligands based on their 3D structures.

Over the past few years, the interest in the evaluation of drug-binding kinetics (k_{on} and k_{off}) during lead optimization has increased considerably due to their influence on the time course of a drug’s effect. Since experimental assays used to determine kinetic parameters for drug binding/unbinding are usually time consuming and labor-intensive, robust, efficient and high-throughput *in silico* methods are much in demand to predict kinetic parameters accurately and provide insights into the mechanistic determinants of drug-protein binding. These insights can help in the rational modulation of the binding kinetics during lead optimization. In this work, we have applied COMBINE analysis to derive quantitative structure-kinetics relationships (QSKRs) for the dissociation rate constants (k_{off}) by studying two large and chemically diverse sets of inhibitors of the well-characterized drug targets, heat-shock protein 90 (HSP90) and HIV-1 protease. COMBINE analysis was originally developed to derive QSARs for binding affinity; here, we provide its first application to derive QSKRs for binding kinetic parameters. By performing COMBINE analysis, we obtained QSKRs for dissociation rate constants (k_{off}) of HSP90 and HIV-1 protease inhibitors with very good predictive ability. 70 structurally diverse inhibitors of HSP90 and 36 inhibitors of HIV-1 protease with available experimental kinetics data and co-crystallized or modelled protein-inhibitor complexes were used to derive target-specific predictive models for k_{off} rates. We have also identified key protein-inhibitor interactions that distinguish inhibitors with slow and fast off-rates.

3.1 Systems studied

For demonstrating the first application of COMBINE analysis for deriving QSKRs, we studied two established and well-studied drug targets: HSP90 and HIV-1 protease. Both proteins have been well characterized experimentally and have been the subject of extensive structure-based drug discovery efforts. However, these two targets present different challenges for the prediction of drug-binding kinetics as they show high binding site flexibility and inhibitors with both slow and fast binding kinetics are known[97, 40].

3.1.1 Heat-shock protein 90 (HSP90)

HSP90 is one of the common chaperone proteins that assists in the proper folding of other proteins and stabilizes proteins against elevated temperatures. It is known for its role in stabilizing a number of proteins essential for tumor growth, and it is therefore an anti-cancer target[98]. The name HSP90 comes from the fact that it weighs approximately 90 kiloDaltons (kDa). The N-terminal domain (NTD) of HSP90 is a highly conserved domain and with a mass of approximately 25 kDa. The binding pocket for ATP is situated in the NTD and therefore the ATPase function of the NTD can be blocked by designing small molecule inhibitors that bind to the ATP binding pocket. Blocking of the ATPase function disrupts the chaperone activity of HSP90 which leads to degradation of client proteins and hence suppressed tumor growth[99].

The structures of the NTD of HSP90 (N-HSP90) in complex with inhibitors are known to have high plasticity and exist in "loop-in", "helical" or "loop-out" conformations which differ at the side of the ATP-binding site where α -helix3 is located (see Figure 3.1). Both loop conformations (loop-in and loop-out) have been observed in the crystallographic structures of unbound apo-protein as well as in the holo-structures with different small inhibitors bound to the ATP-binding pocket[97]. In contrast, the helix conformation, with a complete α -helix3, has been observed only in holo-structures and only when the bound inhibitor occupies a transient hydrophobic subpocket between α -helix3 and the beta-strands, in addition to the ATP binding site (see Figure 3.1 A).

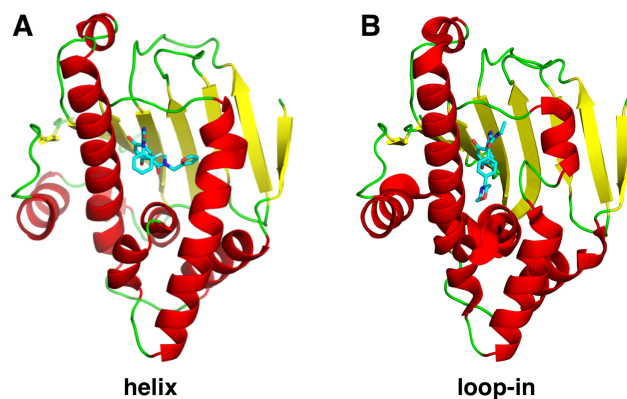


Figure 3.1: A) Helix conformation of the N-HSP90 in complex with the inhibitor bound to the ATP binding-site (PDB ID: 5J20[97]). B) Loop-in conformation of the N-HSP90 in complex with the inhibitor bound to the ATP binding-site (PDB ID: 5NYI[97]). The structures of N-HSP90 are shown in ribbon representation where helices, β -sheets and loops are colored in red, yellow and green, respectively, and inhibitors are shown with cyan stick representation.

3.1.2 HIV-1 protease

HIV-1 protease is a homodimeric aspartyl protease and it specifically cleaves the precursor Gag and Gag-Pol polyproteins into various viral capsid and other structural proteins. As HIV-1 protease plays a critical role in viral maturation for producing infectious virus particles, it is an attractive target for AIDS therapy[100].

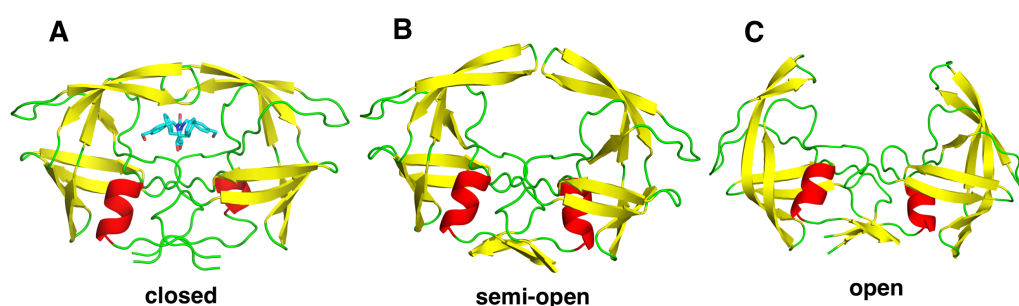


Figure 3.2: A) The HIV-1 protease homodimer with DMP323 bound (PDB ID: 1QBS[101]). HIV-protease has two β -hairpin loops (flaps) that exist in a closed state on ligand binding. B) semi-open and C) open conformations of the flaps in unbound HIV-1 protease (conformations shown were obtained from the MD simulation run of unliganded HIV-1 protease with PDB ID: 1HXW[102]). The structures of HIV-1 protease are shown in ribbon representation where helices, β -sheets and loops are colored in red, yellow and green, respectively and the DMP323 is shown with cyan stick representation.

Each subunit of HIV-1 protease is a polypeptide chain consisting of 99 residues. The active site is located at the interface between both subunits and is composed of the two conserved catalytic triplets (Asp25-Thr26-Gly27)[103]. Specific inhibitors of HIV-1 protease, including clinically approved drug molecules, bind to the substrate binding pocket[41] which is mainly formed by the side chains of Arg8, Leu23, Asp25, Gly27, Ala28, Asp29, Asp30, Val32, Ile47, Gly48, Gly49, Ile50, Phe53, Leu76, Thr80, Pro81, Val82, and Ile84 residues of both subunits. Each monomer contains an extended β -sheet region (a glycine-rich loop), also known as the flap. These flaps exist in a closed state in the liganded form, and can exist in either semi-open or open conformations in the unbound form (see Figure 3.2).

3.2 Dataset used for the COMBINE analysis

3.2.1 Heat-shock protein 90

For generating the COMBINE analysis model of HSP90, 70 inhibitors with available experimental kinetics measurements, were used (see Table 3.1 for SMILES strings and Figure 3.3 for their chemical structures). These inhibitors are structurally very diverse and they belong to 11 different chemical classes: resorcinol, indazole, hydroxylindazole, aminoquinazoline, benzamide, aminopyrrolopyrimidine, 7-imidazopyridine, 7-azaindole, aminothienopyridine, 6-hydroxyindole, adenine and 2-aminopyridine (see Figure 3.3). All of these inhibitors block the ATPase function of HSP90 by binding to its ATP binding pocket located in the N-terminal domain of HSP90 (N-HSP90). 57 of these inhibitors bind to the helical conformation of N-HSP90 and hence will be referred to as "helix-binders" in the following sections. The remaining 13 inhibitors bind to the N-HSP90 in the loop conformation and will be termed "loop-binders". The experimental k_{off} , k_{on} and K_d values for these inhibitors were available from Kokh *et al.*[29]. The range of experimental k_{off} rate constants for these inhibitors spans over 4 orders of magnitude with the fastest and slowest dissociating inhibitors having k_{off} values of 0.83 s^{-1} and 0.0001 s^{-1} , respectively, and therefore ideal for deriving QSKRs using COMBINE analysis. In addition, 3D crystallographic structures of protein-inhibitor complexes for 37 of these inhibitors are available in the PDB database. For the remaining 33 inhibitors, it was possible to model their bound complex with the protein by introducing small substitutions

into similar compounds complexed with N-HSP90.

Compound Id	SMILES
1	<chem>CCNC(=O)c1noc(-c2cc(C(C)C)c(O)cc2O)c1-c1ccc(C[NH+]2CCOCC2)cc1</chem>
2	<chem>CCNC(=O)c1noc(-c2cc(Cl)c(O)cc2O)c1-c1ccc(OC)cc1</chem>
3	<chem>CCNC(=O)c1[nH]nc(-c2cc(Cl)c(O)cc2O)c1-c1ccc(OC)cc1</chem>
4	<chem>CCNC(=O)c1noc(c2cc(Cl)c(O)cc2O)c1c3ccc(C[NH+]4CCOCC4)cc3</chem>
5	<chem>O=c1[nH]nc(-c2cc(Br)c(O)cc2O)n1-c1cccc1F</chem>
6	<chem>COc1ccc(-c2c(C#N)c(N)nc3sc(C(N)=O)c(N)c23)cc1OCCCC(=O)O</chem>
7	<chem>CCc1cc(-c2n[nH]c(C)c2-c2cccc2F)c(O)cc1O</chem>
8	<chem>O=c1[nH]nc(-c2ccc(O)cc2O)n1-c1cccc1F</chem>
9	<chem>Cc1n[nH]c2cc(O)c(-c3ccnn3-c3cccc3)cc12</chem>
10	<chem>COc1ccc(-c2c(-c3ccc(O)cc3O)n[nH]c2C)cc1</chem>
11	<chem>CN(Cc1cccc1)C(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2F)c(O)cc1O</chem>
12	<chem>Cc1cccc1-n1c(-c2cc(C(=O)N(C)Cc3cccc3)c(O)cc2O)n[nH]c1=O</chem>
13	<chem>CCCN(C)C(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2F)c(O)cc1O</chem>
14	<chem>Oc1cc(O)c(-c2ccnn2-c2cccc2Cl)cc1CCc1cccc1</chem>
15	<chem>CCCN(C)S(=O)(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2F)c(O)cc1O</chem>
16	<chem>CC(C)N(C)S(=O)(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2F)c(O)cc1O</chem>
17	<chem>Br1cnc2[nH]cnc2c1C(=O)NC1c2cccc2-c2c(-c3cnc4cccc4c3)cccc21</chem>
19	<chem>O=C(NC1c2cccc2-c2c(-c3nc4ccncc4[nH]3)cccc21)c1ccnc2[nH]ccc12</chem>
20	<chem>Cc1nn(-c2ccc(C(N)=O)c(N[C@H]3CC[C@H](O)CC3)c2)c2cccc(-c3cnc4cccc4c3)c12</chem>
21	<chem>Cc1en(-c2ccc(C(N)=O)c(N[C@H]3CC[C@H](O)CC3)c2)c2c1C(=O)CC(C)(C)C2</chem>
22	<chem>Cc1en(-c2ccc(C(N)=O)c(NC3CCC(=O)CC3)c2)c2c1C(=O)CC(C)(C)C2</chem>
23	<chem>CC(C)N(C)S(=O)(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2Cl)c(O)cc1O</chem>
24	<chem>CCCN(C)C(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2C)c(O)cc1O</chem>
25	<chem>Cc1cccc1-n1c(-c2cc(C(=O)N(C)Cc3cccc3)c(O)cc2O)n[nH]c1=O</chem>
26	<chem>CCCCCN(C)C(=O)c1cc(-c2ccnn2-c2cccc2C)c(O)cc1O</chem>
27	<chem>Cc1cccc(CN(C)C(=O)c2cc(-c3n[nH]c(=O)n3-c3cccc3C)c(O)cc2O)c1</chem>
28	<chem>Cc1cccc1-n1c(-c2cc(C(=O)N(C)CC3CCCO3)c(O)cc2O)n[nH]c1=O</chem>
29	<chem>CCCN(C)C(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2)c(O)cc1O</chem>
30	<chem>Cc1cccc1-n1cccc1-c1cc(C(=O)N(C)Cc2ccco2)c(O)cc1O</chem>
31	<chem>Cc1cccc1-n1c(-c2ccc(O)cc2O)n[nH]c1=O</chem>
32	<chem>O=c1[nH]nc(-c2ccc(O)cc2O)n1-c1cccc1Cl</chem>
33	<chem>CCc1cccc1-n1c(-c2ccc(O)cc2O)n[nH]c1=O</chem>

Table 3.1 continued from previous page

Compound Id	SMILES
34	<chem>CCCN(C)C(=O)c1cc(-c2n[nH]c(=O)n2-c2cccc2F)c(O)cc1O</chem>
35	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc5c(c4)OCO5)cc23)c1</chem>
36	<chem>C[NH+]1CCC(c2ccc(N(C)C(=O)c3cc4c(CCC(C)(C)C)n[nH]c4cc3O)cc2)CC1</chem>
37	<chem>CCCCN(C)C(=O)c1n[nH]c2cc(O)c(C(=O)N(C)c3ccc(N4CCOCC4)cc3)cc12</chem>
38	<chem>CN(Cc1ccc(Cl)cc1)C(=O)c2cc3c(Cc4ccccc4)n[nH]c3cc2O</chem>
39	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)Cc4ccccc4)cc23)c1</chem>
40	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)Cc4ccc(Cl)cc4)cc23)c1</chem>
41	<chem>Oc1cc2[nH]nc(Cc3ccccc3)c2cc1-c1ccnn1-c1ccccc1</chem>
42	<chem>Cc1ccc(N(C)C(=O)c2cc3c(Cc4cccc(C)c4)n[nH]c3cc2O)cc1</chem>
43	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CCOCC5)cc4)cc23)c1</chem>
44	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccccc4)cc23)c1</chem>
45	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CCCCC5)cc4)cc23)c1</chem>
46	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N(C)C)cc4)cc23)c1</chem>
47	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CC[NH2+]CC5)cc4)cc23)c1</chem>
48	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CCN(C)CC5)cc4)cc23)c1</chem>
49	<chem>CO[C@H]1CCN(C(=O)c2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CCOCC5)cc4)cc23)C1</chem>
50	<chem>CO[C@H]1CCCN(C(=O)c2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CCOCC5)cc4)cc23)C1</chem>
51	<chem>CN(C(=O)c1cc2c(C(=O)N3CCCC3)n[nH]c2cc1O)c1ccc(N2CCOCC2)cc1</chem>
52	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(N5CCOCC5=O)cc4)cc23)c1</chem>
53	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4ccc(F)cc4)cc23)c1</chem>
54	<chem>COc1cccc(N(C)C(=O)c2cc3c(Cc4cccc(C)c4)n[nH]c3cc2O)c1</chem>
55	<chem>Cc1cccc(Cc2n[nH]c3cc(O)c(C(=O)N(C)c4cccc(C)c4)cc23)c1</chem>
56	<chem>CN(C(=O)c1cc2c(cc1O)[nH]nc2C(=O)N1CCOCC1)c1ccc(N2CCOCC2)cc1</chem>
57	<chem>CN(C(=O)c1cc2c(C(=O)N3CCCC3)n[nH]c2cc1O)c1ccc(N2CCOCC2)cc1</chem>
58	<chem>Nc1nc(C(=O)N2Cc3ccc(O)cc3C2)c2ccccc2n1</chem>
59	<chem>Nc1nc(C(=O)N2Cc3ccccc3C2)c2cc(O)ccc2n1</chem>
60	<chem>C[NH+]1CCN(S(=O)(=O)c2ccccc2-c2ccc3nc(N)nc(C(=O)N4Cc5ccccc5C4)c3c2)CC1</chem>
61	<chem>Cc1ccc2nc(N)nc(C(=O)N3Cc4ccccc4C3)c2c1</chem>
62	<chem>CNCc1ccccc1-c1ccc2nc(N)nc(C(=O)N3Cc4ccccc4C3)c2c1</chem>
63	<chem>Nc1nc(C(=O)N2Cc3ccccc3C2)c2cc(-c3cc(F)c(F)cc3CCc3nnn[nH]3)ccc2n1</chem>
64	<chem>Nc1nc(C(=O)N2Cc3ccccc3C2)c2ccccc2n1</chem>
65	<chem>Nc1nc(C(=O)N2Cc3ccccc3C2)c2cc(-c3ccccc3O)ccc2n1</chem>
66	<chem>COc1c(C)enc(Cn2cc(C#CCC(C)(C)O)c3c(Cl)nc(N)nc32)c1C</chem>

Table 3.1 continued from previous page

Compound Id	SMILES
67	<chem>Cc1cnc(Cn2ccc3c(Cl)nc(N)nc32)c(C)c1Cl</chem>
68	<chem>C#CCCCn1c(Cc2cc(OC)c(OC)c(OC)c2Cl)nc2c(N)nc(F)nc21</chem>
69	<chem>N#Cc1ccc(N2CCN(CCCc3c[nH]c4cc(O)c(C#N)cc34)CC2)cc1</chem>
70	<chem>Nc1cc(C(=O)NC2c3ccccc3-c3c(-c4nc5ccncc5[nH]4)cccc32)ccn1</chem>

Table 3.1: List of the SMILES strings for the 70 HSP90 inhibitors used for the COMBINE analysis.

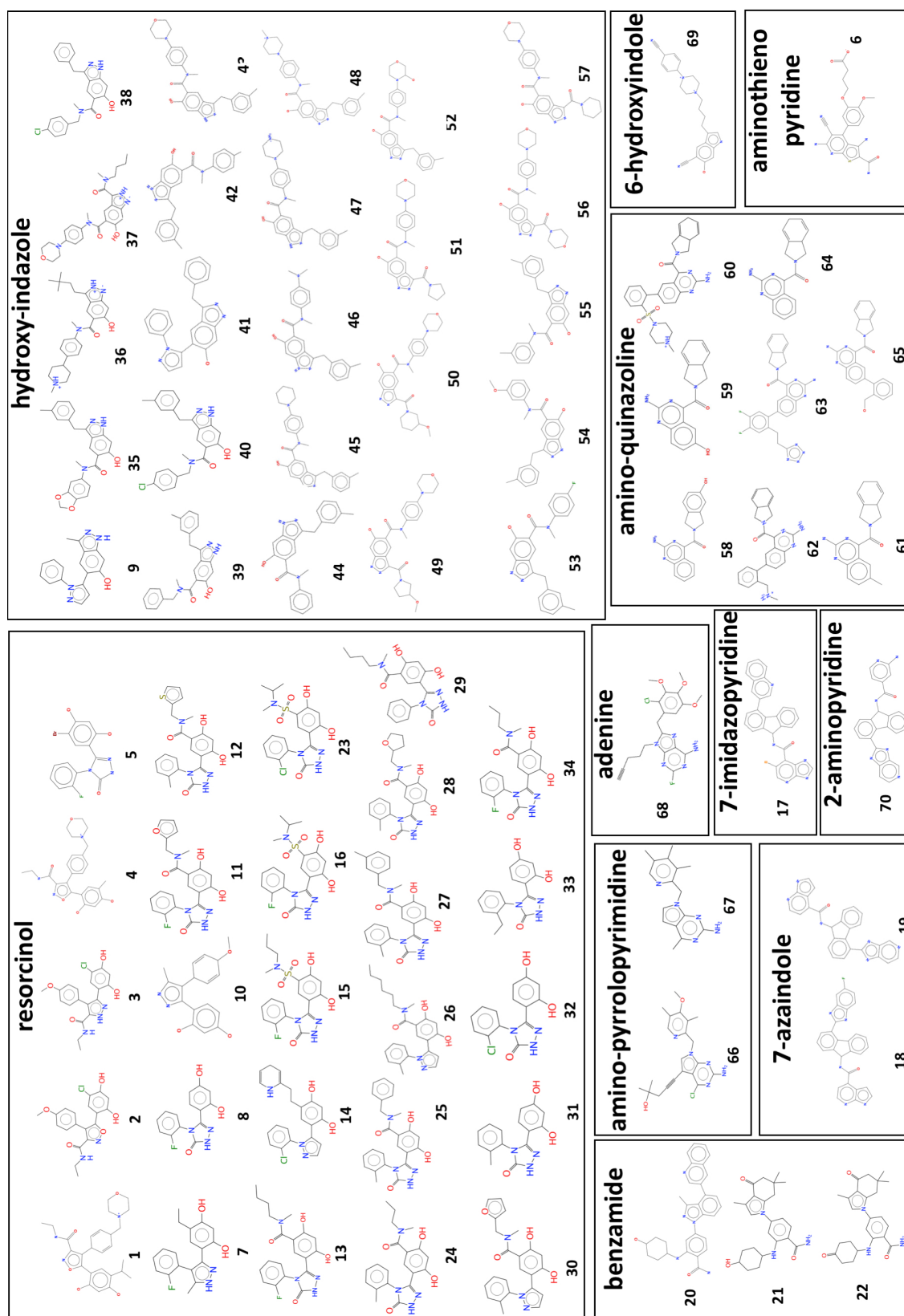


Figure 3.3: 2D chemical structures of inhibitors of HSP90 used for the COMBINE analysis. These 70 inhibitors belong to 11 different chemical classes: resorcinol, hydroxyl-indazole, aminoquinazoline, benzamide, aminopyrrolopyrimidine, 7-imidazopyridine, 7-azaindole, aminothienopyridine, 6-hydroxyindole, adenine and 2-aminopyridine.

3.2.2 HIV-1 protease

For the COMBINE analysis of HIV-1 protease, we decided to consider 36 inhibitors in our analysis as it was possible to either model their bound structure with HIV-1 protease based on analogy or their co-crystallized structure was available in the PDB database (see Table 3.2 for SMILES strings and Figure 6.1 for their chemical structures). For 12 of these inhibitors, their co-crystallized structures with HIV-1 protease were available in the PDB database and the remaining 24 protease-inhibitor complexes were modelled by introducing small substitutions into co-crystallized structures of similar compounds complexed with HIV-1 protease. The experimental measurements of k_{off} rates for these compounds were available from Markgren *et al.*[?]. The k_{off} values of these inhibitors span over 5 orders of magnitude with the fastest and slowest dissociating inhibitors having k_{off} rate constants of 0.00022 s^{-1} and 83.3 s^{-1} , respectively. These inhibitors are also structurally very diverse with their scaffolds belonging to different chemical classes such as cyclic ureas, cyclic sulfamides, linear analogues of compound B268, and non-analogues of B268 (see Figure 6.1).

Compound Id	SMILES
B435	<chem>O[C@@H]([C@@H](O)[C@@H](OCc1cccc1)C(=O)NC2[C@H](O)Cc3cccc23)[C@@H](OCc4cccc4)C(=O)NC5cccc5</chem>
A047	<chem>CNC(=O)c1ccc(CN2[C@H](COc3cccc3)[C@H](O)[C@@H](O)[C@@H](COc4cccc4)N(Cc5cccc5)S2(=O)=O)c1</chem>
A023	<chem>OCCc1ccc(CN2[C@H](COc3cccc3)[C@H](O)[C@@H](O)[C@@H](COc4cccc4)N(Cc5ccc(CCO)c5)S2(=O)=O)c1</chem>
A024	<chem>COC(=O)c1ccc(CN2[C@H](COc3cccc3)[C@H](O)[C@@H](O)[C@@H](COc4cccc4)N(Cc5ccc(CO)cc5)S2(=O)=O)cc1</chem>
B429	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(cc1)c2cccn2)[C@H](O)[C@@H](O)[C@@H](OCc3ccc(cc3)c4cccn4)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
B409	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(cc1)c2ccsc2)[C@H](O)[C@@H](O)[C@@H](OCc3ccc(cc3)c4ccsc4)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
B268	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1cccc1)[C@H](O)[C@@H](O)[C@@H](OCc2cccc2)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
A045	<chem>CNC(=O)c1ccc(CN2[C@H](COc3cccc3)[C@H](O)[C@@H](O)[C@@H](COc4cccc4)N(Cc5ccc(c5)C(=O)NC)S2(=O)=O)c1</chem>
B425	<chem>O[C@H](C[C@@H](OCc1cccc1)C(=O)NC2[C@H](O)Cc3cccc23)[C@@H](O</chem>

Compound Id	SMILES
	<chem>Cc4ccccc4)C(=O)NC5[C@@H](O)Cc6ccccc56</chem>
A021	<chem>OCc1ccc(CN2[C@H](COc3ccccc3)[C@H](O)[C@@H](O)[C@@H](COc4ccccc4)N(Cc5ccc(CO)cc5)S2(=O)=O)cc1</chem>
saquinavir	<chem>CC(C)(C)NC(=O)[C@@H]1C[C@@H]2CCCC[C@@H]2CN1C[C@@H](O)[C@H](Cc3ccccc3)NC(=O)[C@H](CC(=O)N)NC(=O)c4ccc5ccccc5n4</chem>
indinavir	<chem>CC(C)(C)NC(=O)[C@@H]1CN(Cc2cccnc2)CCN1C[C@@H](O)C[C@@H](Cc3ccccc3)C(=O)N[C@@H]4[C@H](O)Cc5ccccc45</chem>
ritonavir	<chem>CC(C)[C@H](NC(=O)N(C)Cc1csc(n1)C(C)C)C(=O)N[C@H](C[C@H](O)[C@H](Cc2ccccc2)NC(=O)OCc3cnsc3)Cc4ccccc4</chem>
DMP323	<chem>OCc1ccc(CN2[C@H](Cc3ccccc3)[C@H](O)[C@@H](O)[C@@H](Cc3ccccc3)N(Cc3ccc(CO)cc3)C2=O)cc1</chem>
nelfinavir	<chem>Cc1c(O)cccc1C(=O)N[C@H](CSc1ccccc1)[C@H](O)CN1C[C@H]2CCCC[C@H]2C[C@H]1C(=O)NC(C)(C)C</chem>
B369	<chem>O[C@H]([C@@H](O)[C@@H](OCc1ccccc1)C(=O)NC2[C@@H](O)Cc3ccccc23)[C@@H](OCc4ccccc4)C(=O)NC5[C@H](O)Cc6ccccc56</chem>
B388	<chem>CC(C)C(N(C)C(=O)[C@H](OCc1ccccc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccccc2)C(=O)NC3[C@@H](O)Cc4ccccc34)C(=O)O</chem>
A038	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(cc1)C2=C(O)C(=O)CCC2)[C@H](O)[C@@H](O)[C@@H](OCc3ccc(cc3)C4=C(O)C(=O)CCC4)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
A037	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(\C=C\C(=O)OC)cc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccc(\C=C\C(=O)OC)cc2)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
B440	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(cc1)e2nccs2)[C@H](O)[C@@H](O)[C@@H](OCc3ccc(cc3)c4nccs4)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
B439	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(CCc2ccccc2)cc1)[C@H](O)[C@@H](O)[C@@H](OCc3ccc(CCc4ccccc4)cc3)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
B408	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(Br)cc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccc(Br)cc2)C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
B412	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccc(cc1)e2cccc(e2)[N+](=O)[O-])[C@H](O)[C@@H](O)[C@@H](OCc3ccc(cc3)c4ccccc4)[N+](=O)[O-])C(=O)N[C@@H](C(C)C)C(=O)NC)C(C)C</chem>
U75875	<chem>CC[C@@H](C)[C@H](NC(=O)[C@H](C(C)C)[C@@H](O)[C@H](O)[C@H](CC1CCCCC1)NC(=O)[C@H](Cc2c[nH+]c[nH]2)NC(=O)COc3ccccc4ccccc34)C(=O)NCc5ccccc5</chem>

Compound Id	SMILES
A008	<chem>OCc1ccc(CN2C(COc3ccccc3)[C@H](O)[C@@H](O)C(COc4ccccc4)N(Cc5ccc(CO)cc5)C2=O)cc1</chem>
B277	<chem>CCCO[C@H]([C@H](O)[C@@H](O)[C@@H](OCCC)C(=O)N[C@@H](C(C)C)C(=O)NC)C(=O)N[C@@H](C(C)C)C(=O)NC</chem>
A030	<chem>C\C(=N/O)\c1cccc(CN2[C@H](COc3ccccc3)[C@H](O)[C@@H](O)[C@@H](COc4ccccc4)N(Cc5cccc(c5)\C(=N\O)\C)S2(=O)=O)c1</chem>
A015	<chem>CN(C(Cc1cccc1)C(=O)O)C(=O)[C@H](OCc2ccccc2)[C@H](O)[C@@H](O)[C@@H](OCc3ccccc3)C(=O)N(C)C(Cc4ccccc4)C(=O)O</chem>
A016	<chem>CN(C(Cc1ccc(O)cc1)C(=O)O)C(=O)[C@H](OCc2ccccc2)[C@H](O)[C@@H](O)[C@@H](OCc3ccccc3)C(=O)N(C)C(Cc4ccc(O)cc4)C(=O)O</chem>
A017	<chem>CC(O)C(N(C)C(=O)[C@H](OCc1ccccc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccccc2)C(=O)N(C)C(C(C)O)C(=O)O)C(=O)O</chem>
B322	<chem>CCC(C)C(N(C)C(=O)[C@H](OCc1ccccc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccccc2)C(=O)N(C)C(C(C)CC)C(=O)O)C(=O)O</chem>
B365	<chem>CC(C)C(N(C)C(=O)[C@@H](C[C@@H](O)[C@@H](OCc1ccccc1)C(=O)N(C)C(C(C)C)C(=O)O)OCc2ccccc2)C(=O)O</chem>
B347	<chem>CC(C)C(N(C)C(=O)[C@H](OCc1ccccc1)[C@@H](O)[C@H](O)[C@@H](OCc2ccccc2)C(=O)N(C)C(C(C)C)C(=O)O)C(=O)O</chem>
A018	<chem>CSCCC(N(C)C(=O)[C@H](OCc1ccccc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccccc2)C(=O)N(C)C(CCSC)C(=O)O)C(=O)O</chem>
B249	<chem>COC(=O)C(NC(=O)[C@H](OCc1ccccc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccccc2)C(=O)NC(C(C)C)C(=O)OC)C(C)C</chem>
B376	<chem>CNC(=O)[C@@H](NC(=O)[C@H](OCc1ccccc1)[C@H](O)[C@@H](O)[C@@H](OCc2ccccc2)C(=O)N[C@H](C(=O)NC)c3ccccc3)c4ccccc4</chem>

Table 3.2: List of the SMILES strings for the 36 HIV-1 protease inhibitors used for COMBINE analysis.

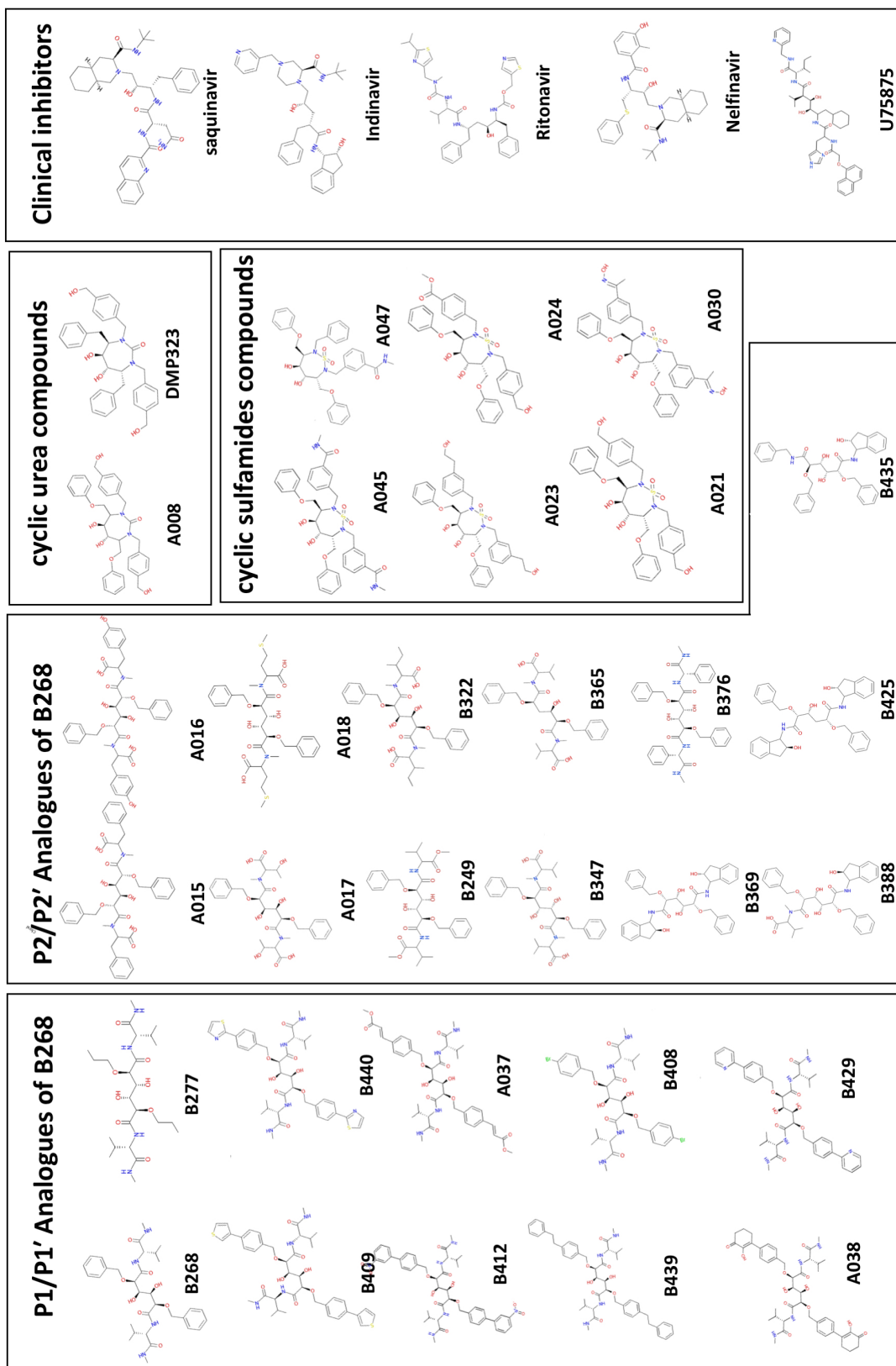


Figure 3.4: 2D chemical structures of inhibitors of HIV-1 protease used for COMBINE analysis.

3.3 Methods

3.3.1 Preparation of protein and ligand structures

Coordinates of the 37 crystallographic structures of protein-inhibitor complexes for N-HSP90 and 12 structures of HIV-1 protease-inhibitor complexes were downloaded from the PDB database (<https://www.rcsb.org/>). The remaining 33 protein-inhibitor complexes for N-HSP90 and 24 complexes for HIV-1 protease were modeled based on analogy by introducing small substitutions into similar compounds complexed with the proteins, using the Schrödinger software (release 2015-4, Schrödinger, LLC, New York). The *Protein Preparation Wizard*[104] of the Schrodinger suite was used to prepare and pre-process the structures of the bound complexes. The preparation of complexes involved addition of missing side chains and disulphide bonds, deletion of crystallographic waters present, and the optimization of the hydrogen bonding network to assign hydrogen atom positions. The protonation states of titratable residues were assigned at pH 7.0 using the PROPKA[105] program available through the *Protein Preparation Wizard* of Schrodinger. To get rid of bad contacts and steric clashes, all of the prepared complex structures were subjected to initial energy minimization using the *Impref* module [104] of the Schrodinger suite with default parameters and the *OPLS3* force field. The *Impref* minimization is a two-step relaxation procedure in which first the rotatable hydrogen atoms are minimized with all the torsional potentials removed, and then an all-atom minimization is performed that is terminated either when the system is fully converged or when it reaches a heavy-atom RMSD from the initial structure of 0.30 Å.

3.3.2 Generation of force field parameters and energy minimization

The partial atomic charges of the inhibitors were calculated using the RESP approach, where the *RESP*[106] program was used to fit the atom-centered charges to the molecular electrostatic potential (MEP) grid computed by the GAMESS program[76]. The *LEap* program of the Amber14 software[79] was used to prepare the force field parameters and topology files for all the protein-inhibitor complexes. The *ff14SB*[107] and the General Amber Force Field (*GAFF*) were used for the proteins and inhibitors, respectively. For energy minimization, the *PME MD* module of

Amber14 software was used. The Amber minimization protocol involved 4 different minimization procedures with gradually decreasing restraints on heavy atoms (100 kcal/mol.², 100 kcal/mol.² and 5 kcal/mol.²) to no positional restraints in the final minimization procedure. For each minimization procedure, 500 steps of steepest-descent minimization followed by 500 steps of conjugate-gradient minimization were applied. Minimization was performed using implicit solvent and a distance dependent dielectric constant ($4r$) was used.

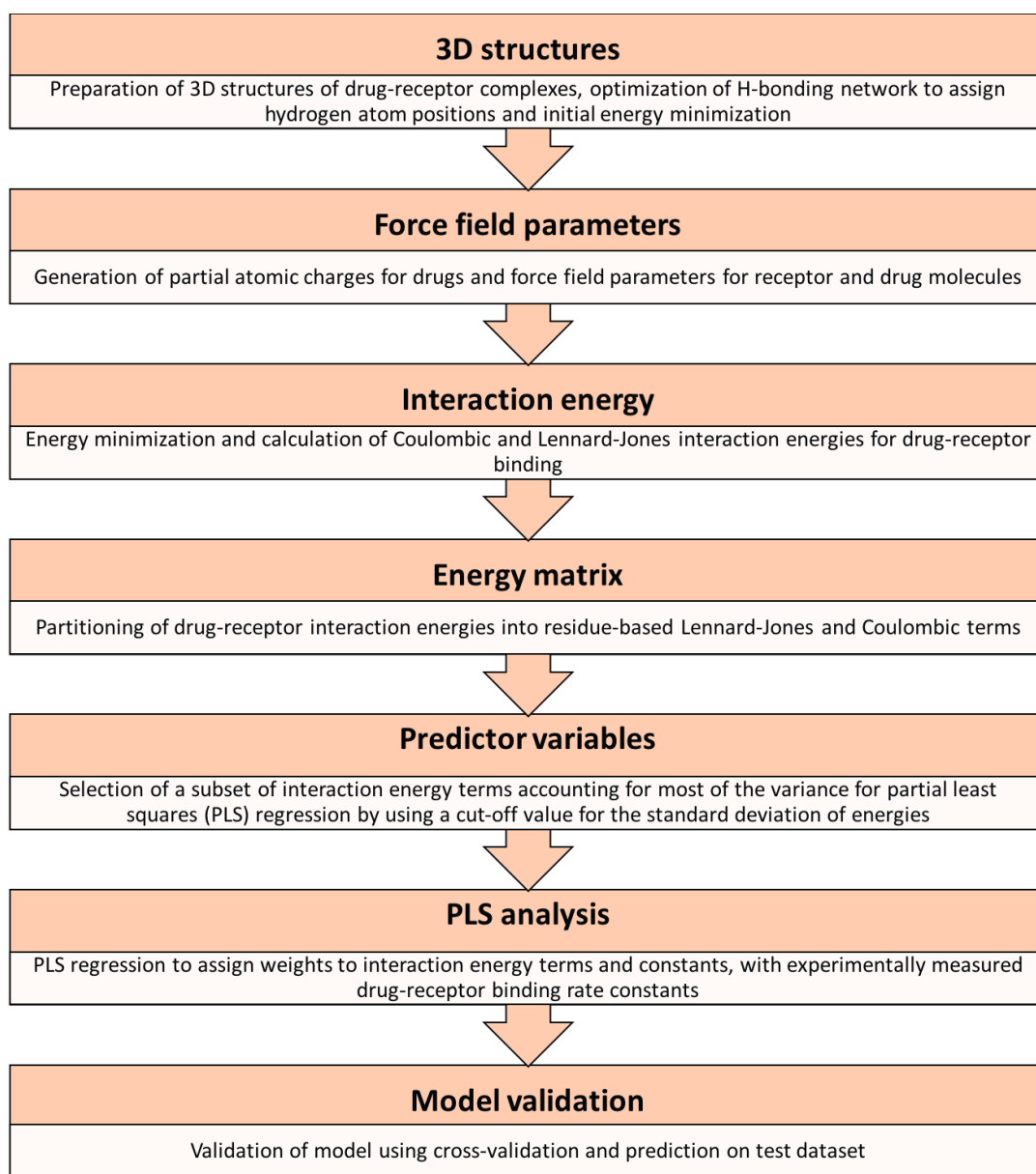


Figure 3.5: Schematic outline of the different steps involved in applying COMBINE analysis to derive a QSKR to predict drug-binding kinetics.

3.3.3 Selection of the training and the test datasets

Four (compounds **6**, **30**, **65** and **69**) of the 70 inhibitors of HSP90 were detected as outliers during the chemometric analysis as they diminished the quality of the model significantly. Interestingly, three of these compounds (**30**, **65** and **69**) were also identified as outliers in the recent work by Kokh *et al.*[29] where the authors used τ RAMD, an enhanced sampling procedure based on molecular dynamics simulations, to calculate the relative residence times of HSP90 inhibitors. Therefore, we decided to exclude these 4 outliers from our dataset and the remaining 66 inhibitors were considered for further analysis. For generating training and test datasets for COMBINE analysis for HSP90, all the inhibitors were ranked from high to low k_{off} values and every fifth inhibitor ($\approx 20\%$) in the ranked list was selected for the test set, while the remaining ($\approx 80\%$) inhibitors were selected for the training set. Therefore the training set and test set consisted of 53 and 13 inhibitors, respectively.

Out of the 36 inhibitors in the HIV-1 protease dataset, 3 inhibitors (**U75875**, **B249** and **B376**) were identified as outliers and hence not considered further. Two of these outliers: **B249** and **B376**, which are dihydroxy analogues of compound **B268**, have a variety of substituents at the valine side chains of **B268** and no crystal structures were available for them. These small substitutions in **B268** resulted in large increases of k_{off} rates by almost 1000-fold, and this effect was not captured by the COMBINE analysis as the modeled complexes of these compounds were very similar to the reference structure. Due to the smaller size of the dataset, we decided to train our COMBINE analysis model with all of the 33 inhibitors and therefore no separate test-set was chosen. The model was only validated with different cross-validation methods such as leave-one-out, leave-two-out and leave-three-out cross-validation.

3.3.4 Calculation of the interaction energy terms and generation of energy matrix for PLS analysis

The *gCOMBINE* program[108] was used for the calculations of LJ and Coulombic interaction energies between protein and inhibitors using Amber force field parameters generated by the *LEap* program of the Amber14 software. *gCOMBINE* decomposes the total LJ and Coulombic interaction energy between protein and the bound inhibitor on a per-residue basis, thereby resulting in a matrix of interaction energy

terms where each energy term corresponds to Coulombic or LJ interaction energy between one of the amino acid residues of the protein and the bound inhibitor. Since there are 207 amino acid residues in the N-HSP90, gCOMBINE generated 207 Coulombic and 207 LJ energy terms for all HSP90 inhibitors. Similarly, 198 Coulombic and 198 LJ energy terms were calculated for all HIV-1 protease inhibitors corresponding to 198 amino acid residues in the HIV-1 protease dimer (each monomer of protease has 99 amino acids).

3.3.5 PLS analysis

PLS analysis was also performed using the gCOMBINE program. Only those interaction energy terms that showed variance across the entire training dataset, and have a standard deviation greater than the specified cutoff value, were selected for PLS analysis. Different cutoff values in the range of 0.2–1.0 kcal/mol were tested for both datasets. For HSP90, choosing a standard deviation cut-off of 0.25 kcal/mol resulted in the most robust model with the least sensitivity and best predictive performance (Q^2) observed in different cross-validation methods used. The best model for HIV-1 protease was obtained when a cutoff of 0.65 kcal/mol was chosen. Then the weights (or the contributions) of these interaction energy terms and their projection over different numbers of latent variables were determined from PLS regression by correlating the interaction energies with the experimental $\log_{10}(k_{off})$ values. Projections were made for up to 10 latent variables for both datasets. Regression coefficients (R^2), average absolute errors (AAE) and root-mean squared errors (RME) for different models obtained with projection over different numbers of latent variables were calculated by gCOMBINE.

3.3.6 Model validation

In order to access the sensitivity of the different models obtained, the models obtained from PLS regression were subjected to different validation techniques such as: leave-one-out (LOO), leave-two-out (L2O), leave-three-out (L3O) and random groups of 7. The model with the best predictive power and least sensitivity was selected as the best model.

3.4 Results

3.4.1 COMBINE analysis model for HSP90 inhibitors

Lennard-Jones and Coulombic interaction energies were computed between inhibitors and 207 residues in the N-terminal domain of HSP90 for all 66 N-HSP90-inhibitor complexes in the training and test datasets (see Figure 3.6). As seen in Figure 3.6, the interaction energies between bound inhibitors and amino acid residues close to the active site (labelled residues) show high variation across the entire dataset and should therefore be considered further for PLS regression. On the other hand, the interaction energies between inhibitors and residues located far from the active site are almost negligible.

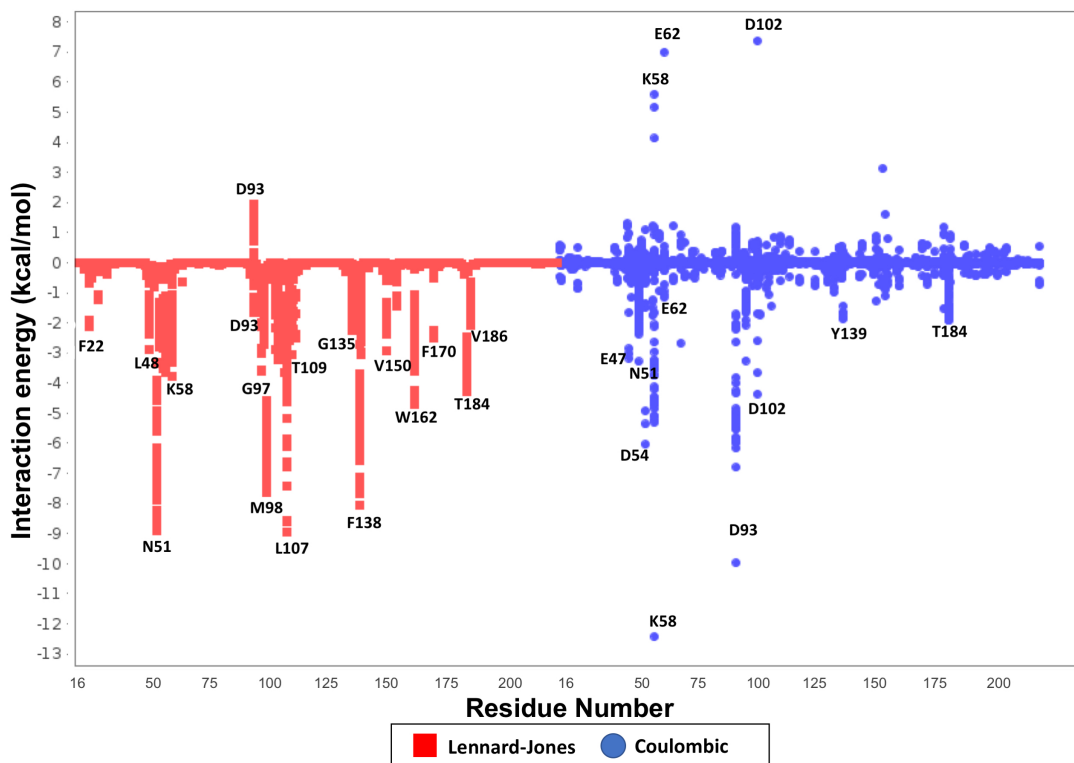


Figure 3.6: Interaction energies between N-HSP90 amino acid residues and inhibitors. The first 207 columns on the x-axis correspond to Lennard-Jones energies (kcal/mol) for each residue and the last 207 columns correspond to Coulombic energies (kcal/mol) between each inhibitor and different residues. Each column has 66 data points corresponding to the 66 inhibitors used for the COMBINE analysis.

The best QSKR model having least sensitivity and the best predictive power (Q^2) for k_{off} rates constants of HSP90 inhibitors was obtained when a standard deviation

cutoff of 0.25 kcal/mol was used to select a subset of interaction energy terms for PLS analysis. A total of 42 inhibitor–residue interaction energy terms (12 coulombic and 30 LJ terms) that have standard deviation higher than 0.25 kcal/mol in the training dataset were used in the PLS regression (see Figure 3.7). Nine amino acid residues: N51, D54, K58, D93, G97, D102, L103, Y139, and T184, make contributions of both coulombic and LJ interaction energies to the QSKR model.

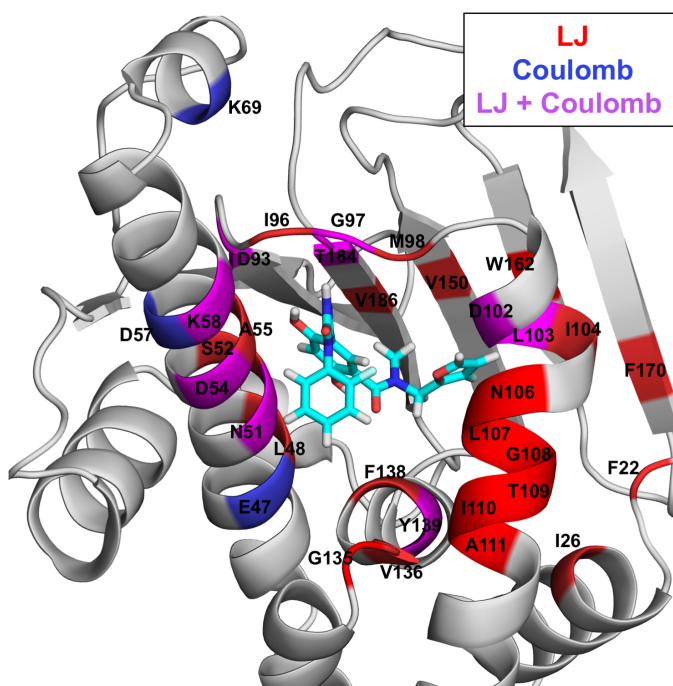


Figure 3.7: Key protein-inhibitor interactions identified from the COMBINE analysis of HSP90 inhibitors. 30 LJ and 12 coulombic protein residue–inhibitor interaction energy terms were selected based on variance over the inhibitors for deriving the PLS model. On the crystal structure (PDB ID: 5J20) of compound 11 (cyan sticks) complexed with N-HSP90 (ribbon representation), the residues are colored according to whether their coulombic (blue), LJ (red), or both coulombic and LJ (magenta) interaction energies with the bound inhibitor contribute to the model. The figure is taken from **Ganotra and Wade, 2018**[109].

Then the predictor variables (interaction energies) and the response variable (experimental $\log_{10}(k_{off})$ values) were projected over different numbers of latent variables in PLS analysis and the weights (or the contributions) of the 42 interaction energy terms were determined from PLS regression (see Figure 3.9). The regression coefficients and standard mean errors were determined for different numbers of latent variables (see Table 3.3). The models were then subjected to different cross-validation techniques to access their sensitivity and predictive ability (see Table 3.5). The model with three latent variables was found to have the best predictive power

and least sensitivity with a R^2 of 0.80 and a leave-one-out (LOO) cross-validated correlation coefficient (Q^2) of 0.69 (see Table 3.4 and Figure 3.8). The average absolute error (AAE_T) and the root mean squared error (RME_T) for the training set were calculated to be 0.37 ($\log_{10}(s^{-1})$ units) and 0.46 ($\log_{10}(s^{-1})$ units), respectively (Table 3.3). The model obtained has good predictive power as the correlation coefficient for the test-set (R^2_{PRED}) with 13 compounds was calculated to be 0.86 with an AAE_{TP} of 0.33 and RME_P of 0.37 (see Figure 3.8 and Table 3.6). The values of the average absolute error (AAE_V) and the root mean squared error (RME_V) for different cross-validation sets were found to be consistent for the different cross-validation methods used (Table 3.5).

LV	R^2	Q^2_{LOO}	AAE_T	AAE_V	RME_T	RME_V	R^2_{PRED}	AAE_P	RME_P
1	0.43	0.31	0.61	0.67	0.78	0.85	0.33	0.60	0.79
2	0.75	0.66	0.39	0.45	0.51	0.60	0.71	0.44	0.52
3	0.80	0.69	0.37	0.45	0.46	0.57	0.86	0.33	0.37
4	0.82	0.72	0.35	0.44	0.44	0.55	0.87	0.32	0.35
5	0.85	0.71	0.32	0.45	0.40	0.55	0.86	0.31	0.36
6	0.87	0.71	0.30	0.45	0.38	0.56	0.89	0.26	0.33
7	0.88	0.69	0.27	0.45	0.36	0.58	0.89	0.27	0.33
8	0.89	0.66	0.26	0.47	0.34	0.60	0.86	0.32	0.36
9	0.90	0.64	0.25	0.46	0.33	0.62	0.85	0.33	0.37
10	0.91	0.55	0.23	0.51	0.31	0.70	0.69	0.48	0.54

Table 3.3: Summary of the models derived for different numbers of latent variables (LVs) for the COMBINE analysis for k_{off} rate constants of N-HSP90 inhibitors. The models were derived using the $\log_{10}(k_{off})$ value (unit of k_{off} rates in s^{-1}) as the response variable in the PLS analysis. The table lists the regression coefficient (R^2) for the training set, the correlation coefficient for leave-one-out cross validation sets (Q^2_{LOO}), average absolute errors (AAE_T and AAE_V) and root mean squared errors (RME_T and RME_V) for the training set and leave-one-out validation sets, respectively, the correlation coefficient (R^2_{PRED}) for the test-set (prediction set), average absolute error (AAE_P) and root mean squared error (RME_P) for the test-set.

Compound Id	N-HSP90 Binding site conformation	PDB Id	Experimental $\log_{10}(k_{off}(s^{-1}))$	Fitted $\log_{10}(k_{off}(s^{-1}))$ (PLS regression)	Predicted $\log_{10}(k_{off}(s^{-1}))$ (LOO validation)
1	loop	2VCI	-4.00 ± 0.00	-3.45	-2.57
3	loop	2BSM	-2.00 ± 0.04	-2.54	-2.47
5	loop	5J2X	-1.85 ± 0.07	-1.21	-1.16
7	loop	6ELO	-1.20 ± 0.02	-0.77	-0.65
8	loop	5J64	-0.68 ± 0.07	-0.60	-0.65
9	loop	n.a.	-0.08 ± 0.03	-0.86	-1.13
10	loop	6ELN	-0.60 ± 0.03	-1.12	-1.25
11	helix	5J20	-3.48 ± 0.03	-2.47	-2.37
12	helix	5J86	-2.75 ± 0.09	-2.80	-2.79
13	helix	5J9X	-2.77 ± 0.12	-2.43	-2.41
14	helix	6ELP	-0.76 ± 0.06	-1.91	-2.16
15	helix	5J27	-2.19 ± 0.03	-2.02	-2.00
16	helix	5J86	-1.85 ± 0.05	-2.09	-2.13
17	helix	5LRZ	-3.56 ± 0.01	-3.99	-3.89
18	helix	5LR7	-3.72 ± 0.16	-3.18	-2.59
20	helix	5LQ9	-3.87 ± 0.01	-4.18	-4.11
21	helix	5LS1	-3.31 ± 0.12	-2.89	-2.80
22	helix	5T21	-3.12 ± 0.03	-2.61	-2.52
23	helix	n.a.	-2.02 ± 0.02	-1.80	-1.76
24	helix	n.a.	-2.33 ± 0.07	-2.06	-2.04
27	helix	n.a.	-2.92 ± 0.04	-2.53	-2.52
28	helix	n.a.	-2.34 ± 0.08	-2.59	-2.66
29	helix	n.a.	-2.52 ± 0.04	-2.43	-2.44
31	loop	n.a.	-0.96 ± 0.17	-1.26	-1.34
33	loop	n.a.	-1.15 ± 0.10	-0.64	-0.58
36	helix	5LO6	-2.86 ± 0.12	-3.21	-3.32
37	helix	5LNZ	-2.70 ± 0.04	-3.13	-3.14
38	helix	6EY8	-1.54 ± 0.02	-2.09	-2.08
39	helix	6EFU	-1.65 ± 0.02	-1.95	-1.93
40	helix	6EY9	-1.76 ± 0.01	-2.16	-2.13
41	helix	6EY8	-0.63 ± 0.04	-1.72	-1.87
42	helix	n.a.	-2.30 ± 0.08	-2.19	-2.17
43	helix	5OCI	-3.17 ± 0.00	-2.81	-2.77

Compound Id	N-HSP90 Binding site conformation	PDB Id	Experimental $\log_{10}(k_{off}(s^{-1}))$	Fitted $\log_{10}(k_{off}(s^{-1}))$ (PLS regression)	Predicted $\log_{10}(k_{off}(s^{-1}))$ (LOO validation)
44	helix	n.a.	-2.04 ± 0.05	-1.93	-1.91
46	helix	n.a.	-2.63 ± 0.07	-2.76	-2.78
47	helix	n.a.	-2.91 ± 0.03	-2.76	-2.75
48	helix	n.a.	-3.12 ± 0.07	-2.98	-2.94
50	helix	5ODX	-3.53 ± 0.02	-3.34	-3.28
51	helix	5NYH	-2.62 ± 0.01	-2.60	-2.56
52	helix	n.a.	-2.86 ± 0.14	-2.78	-2.73
55	helix	n.a.	-2.11 ± 0.27	-2.10	-2.08
56	helix	n.a.	-1.88 ± 0.02	-2.62	-2.67
57	helix	n.a.	-3.04 ± 0.12	-2.78	-2.72
58	helix	n.a.	-0.26 ± 0.12	-0.45	-0.63
59	helix	n.a.	-0.24 ± 0.02	-0.45	-0.57
60	helix	5OD7	-3.62 ± 0.11	-3.85	-3.64
62	helix	6EI5	-2.34 ± 0.04	-2.11	-2.06
63	helix	n.a.	-2.82 ± 0.05	-2.83	-3.25
64	helix	n.a.	-0.26 ± 0.04	-0.21	-0.29
66	helix	n.a.	-2.90 ± 0.08	-2.26	-2.17
67	helix	5LR1	-1.59 ± 0.02	-0.73	-0.68
68	helix	6EL5	-1.48 ± 0.02	-2.04	-2.19
70	helix	2YKJ	-3.00 ± 0.06	-2.65	-2.27

Table 3.4: Comparison of $\log_{10}(k_{off})$ values calculated by the COMBINE analysis model in PLS regression (column 5) and the experimental $\log_{10}(k_{off})$ values from Ref.[29] (column 4) for different N-HSP90 inhibitors (53 compounds) used for training the COMBINE analysis model. The $\log_{10}(k_{off})$ values predicted from leave-one-out (LOO) cross-validation for the training set of inhibitors are given in the last column.

Validation	Q^2	AAE _V	RME _V
Leave-one-out (LOO)	0.69	0.45	0.57
Leave-two-out (L2O)	0.69	0.45	0.58
Leave-three-out (L3O)	0.68	0.46	0.59
Random groups of 7 (10 iterations)	0.68	0.46	0.59

Table 3.5: Statistical measures of correlation for the COMBINE Analysis Models Derived for $\log(k_{off})$ of HSP90 inhibitors. Cross-validated correlation coefficient (Q^2), average absolute errors (AAE_V) and root mean squared errors (RME_V) for different validation methods for the PLS model derived with 3 latent variables for HSP90 inhibitors.

Compound Id	N-HSP90 Binding site conformation	PDB Id	Experimental $\log_{10}(k_{off}(\text{s}^{-1}))$	Predicted $\log_{10}(k_{off}(\text{s}^{-1}))$
2	loop	2UWD	-2.70 ± 0.03	-2.45
4	loop	5NYI	-4.00 ± 0.00	-3.77
19	helix	2YKI	-3.55 ± 0.07	-3.23
25	helix	n.a.	-2.96 ± 0.21	-2.45
26	helix	n.a.	-2.00 ± 0.07	-2.44
32	loop	n.a.	-0.92 ± 0.07	-1.32
34	helix	n.a.	-2.38 ± 0.05	-2.90
35	helix	6EYA	-2.27 ± 0.03	-2.31
45	helix	n.a.	-3.13 ± 0.05	-2.67
49	helix	n.a.	-2.86 ± 0.06	-2.91
53	helix	n.a.	-1.50 ± 0.22	-1.96
54	helix	n.a.	-1.79 ± 0.10	-2.24
61	helix	n.a.	-0.58 ± 0.12	-0.42

Table 3.6: Comparison of $\log_{10}(k_{off})$ values predicted by the COMBINE analysis model and the experimental $\log_{10}(k_{off})$ values from Ref.[29] (column 4) for different N-HSP90 inhibitors used in the test set (13 compounds) for validation of the COMBINE analysis model.

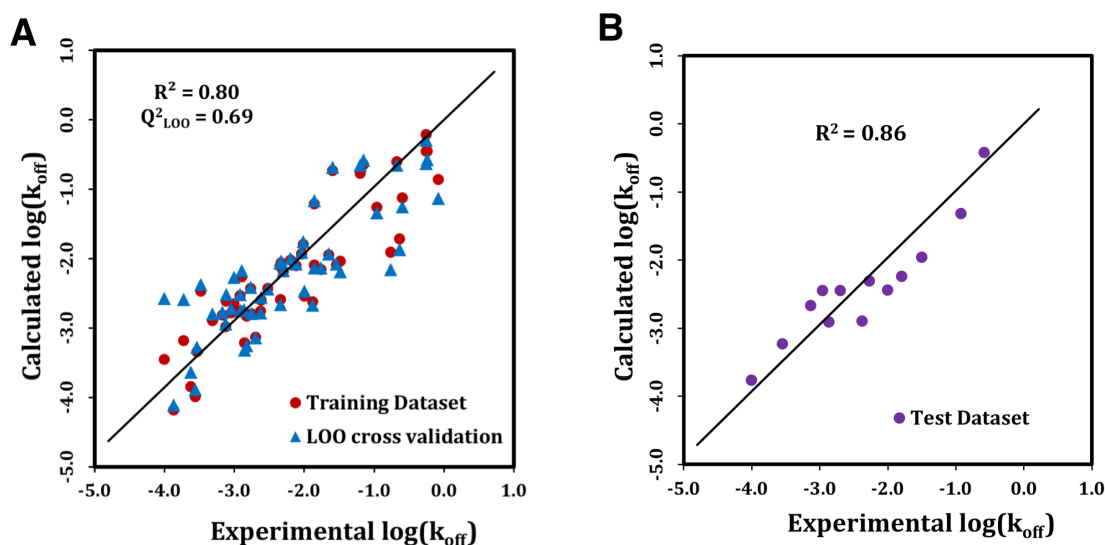


Figure 3.8: A) Plot of calculated vs experimental $\log(k_{off})$ values for the training data set ($R^2 = 0.80$) and LOO cross-validation ($Q^2 = 0.69$). B) Plot of calculated vs experimental $\log(k_{off})$ values for the test data set with 13 compounds ($R^2_{PRED} = 0.86$). The diagonal straight lines in both plots corresponds to $y = x$ (ideal case).

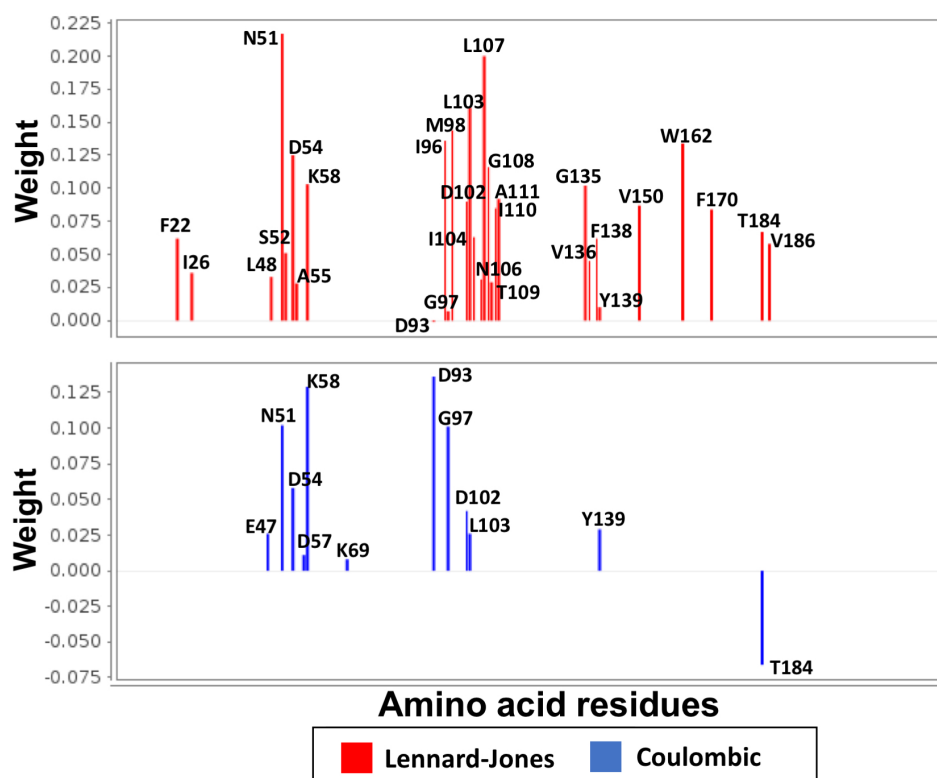


Figure 3.9: Weights for different LJ and coulombic interaction energy contributions derived from the PLS analysis (projection to 3 latent variables). The figure is taken from Ganotra and Wade, 2018^[109].

The major contribution to the k_{off} rate constants comes from the LJ energies of the hydrophobic residues lining the binding pocket (see Figure 3.9). Therefore, compounds with slow k_{off} rates tend to have bulky hydrophobic groups mediating strong LJ interactions with the nonpolar binding site residues. Most of the helix-binders are relatively bulkier in size and have lower k_{off} rate constants, as they have additional hydrophobic moieties which occupy a transient hydrophobic cavity formed between α -helix3 and the β -strands and mediate strong van der Waals interactions with hydrophobic residues L103, I104, N106, L107, G108, T109, I110, and A111 (see Figure 3.10). On the other hand, loop-binders are usually smaller in size and have relatively higher k_{off} rates. Loop-binders that have lower k_{off} rates have additional polar moieties mediating coulombic interactions with amino acid residues such as N51, E47, and G97, thereby stabilizing the bound-state (see Figure 3.10).

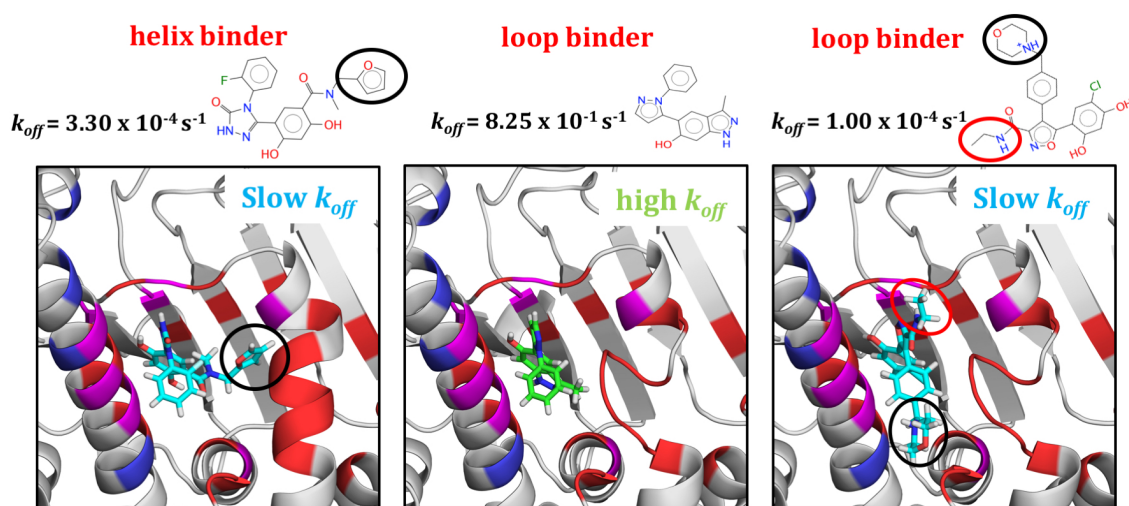


Figure 3.10: Comparison of the binding modes and the key interactions for a helix-binder (compound 11, crystal structure PDB ID: 5J20), a faster dissociating loop-binder (compound 9, model based on PDB ID: 5OCI), and a slower dissociating loop-binder (compound 4, crystal structure PDB ID: 5NYI), respectively. Hydrophobic moieties (shown with a black circle in the left panel) of helix-binders occupy a transient hydrophobic cavity formed by the helix conformation of N-HSP90 and mediate strong LJ interactions with hydrophobic residues. Most of the loop binders are smaller in size and dissociate faster (middle panel). Some of the slower dissociating loop-binders have additional polar moieties (marked with red and black circles in the right panel) that mediate additional electrostatic interactions with the binding-site residues. The figure is taken from **Ganotra and Wade, 2018**[109].

3.4.2 Results: COMBINE analysis model for HIV-1 protease inhibitors

HIV-1 protease is a homodimer with each monomer consisting of 99 amino acid residues. Therefore, for each of the 33 protease inhibitors in the training dataset, 198 coulombic and 198 LJ energies were calculated using gCOMBINE (see Figure 3.11).

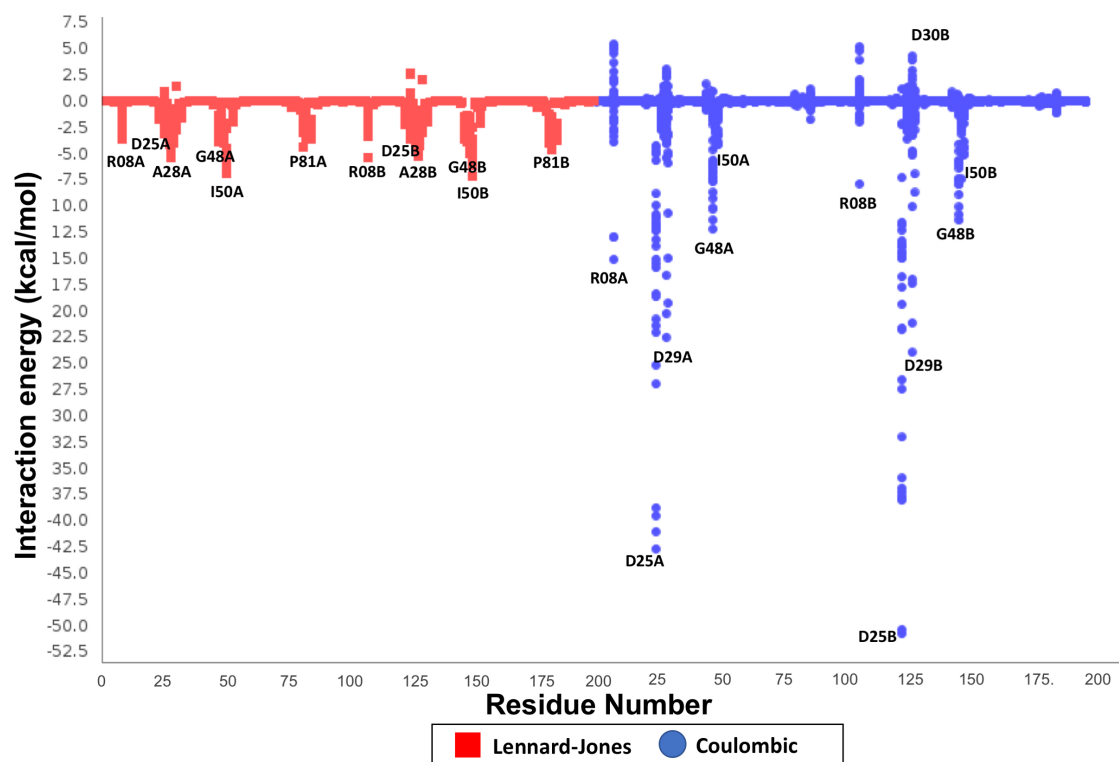


Figure 3.11: Interaction energies between HIV-1 protease residues and inhibitors. Lennard-Jones and Coulombic interaction energies were computed between the inhibitors and 198 amino acid residues of the protease dimer using the gCOMBINE program. The first 198 columns on the x-axis correspond to Lennard-Jones energies (kcal/mol) for each residue and the last 198 columns correspond to Coulombic energies (kcal/mol) between each inhibitor and different residues. Each column has 33 data points corresponding to the 33 inhibitors used for the COMBINE analysis.

To select a subset of interaction energy terms for PLS, a standard deviation cutoff range (from 0.2 to 1.0 kcal/mol) was tested and the choice of a cutoff of 0.65 kcal/mol resulted in the best model. Seventeen coulombic and 17 LJ terms that have standard deviations higher than the cutoff value were used for PLS analysis (see Figure 3.12). The models were derived for different numbers of latent variables and validated using several validation methods (see Tables 3.7 and 3.8). The model

with the best predictive ability and least sensitivity was obtained when projection was made to six latent variables. The R^2 , AAE_T , and RME_T for the training set are 0.94, 0.26 ($\log_{10}(s^{-1})$ units), and 0.34 ($\log_{10}(s^{-1})$ units), respectively (see Table 3.9 and Figure 3.13). The Q^2 value for different validation methods ranged from 0.51 to 0.70 (Table 3.8).

LV	R^2	Q_{LOO}^2	AAE_T	AAE_V	RME_T	RME_V
1	0.20	0.01	1.03	1.14	1.23	1.36
2	0.38	-0.01	0.86	1.09	1.09	1.38
3	0.56	0.06	0.70	1.02	0.91	1.33
4	0.74	0.34	0.55	0.89	0.70	1.11
5	0.83	0.38	0.46	0.83	0.57	1.08
6	0.94	0.70	0.26	0.58	0.34	0.75
7	0.96	0.77	0.23	0.52	0.27	0.66
8	0.97	0.83	0.20	0.47	0.23	0.57
9	0.98	0.83	0.18	0.47	0.20	0.56
10	0.98	0.83	0.16	0.48	0.19	0.57

Table 3.7: Summary of the models derived for different numbers of latent variables (LVs) for the COMBINE analysis for k_{off} rate constants of HIV-1 protease inhibitors. The models were derived using the $\log_{10}(k_{off})$ value (unit of k_{off} rates in s^{-1}) as the response variable in the PLS analysis. The table lists the regression coefficient (R^2) for the training set, the correlation coefficient for leave-one-out cross validation sets (Q_{LOO}^2), average absolute errors (AAE_T and AAE_V) and root mean squared errors (RME_T and RME_V) for the training set and leave-one-out validation sets, respectively. The model with 6 LVs displayed the best predictive performance and least sensitivity in different cross-validation methods used.

Validation	Q^2	AAE_V	RME_V
Leave-one-out (LOO)	0.70	0.58	0.75
Leave-two-out (L2O)	0.51	0.68	0.96
Leave-three-out (L3O)	0.52	0.68	0.95
Random groups of 7 (10 iterations)	0.60	0.63	0.86

Table 3.8: Cross-validated correlation coefficient (Q^2), average absolute errors (AAE_V) and root mean squared errors (RME_V) for different validation methods for the PLS model derived with 6 latent variables for HIV-1 protease inhibitors.

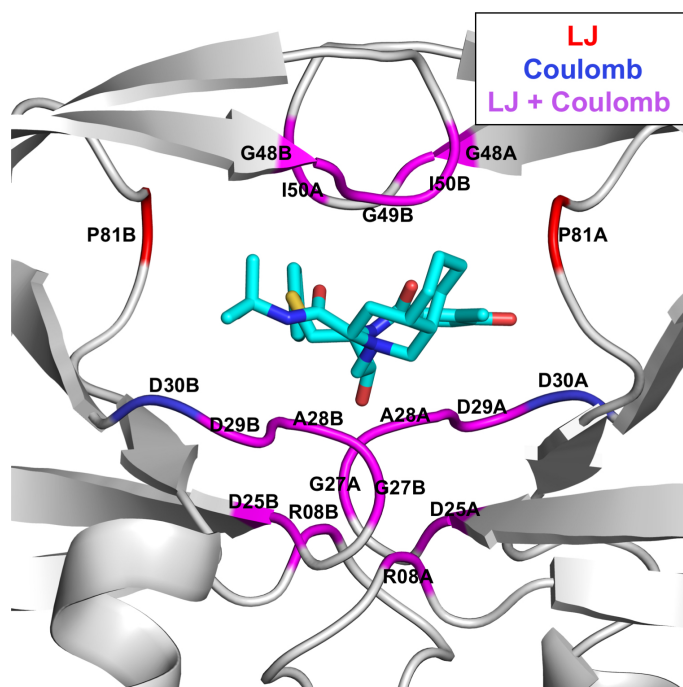


Figure 3.12: Key protein-inhibitor interactions identified from the COMBINE analysis of HIV-1 protease inhibitors. 17 LJ and 17 coulombic protein residue-inhibitor interactions were selected based on variance over the inhibitors. Residues are shown on the crystal structure (PDB ID: 1OHR) of nelfinavir (cyan sticks) bound to HIV-1 protease (ribbon representation) colored according to whether their LJ (red), coulombic (blue), or both LJ and coulombic (magenta) interaction energy terms, contribute to the PLS model. The figure is taken from **Ganotra and Wade, 2018**[109].

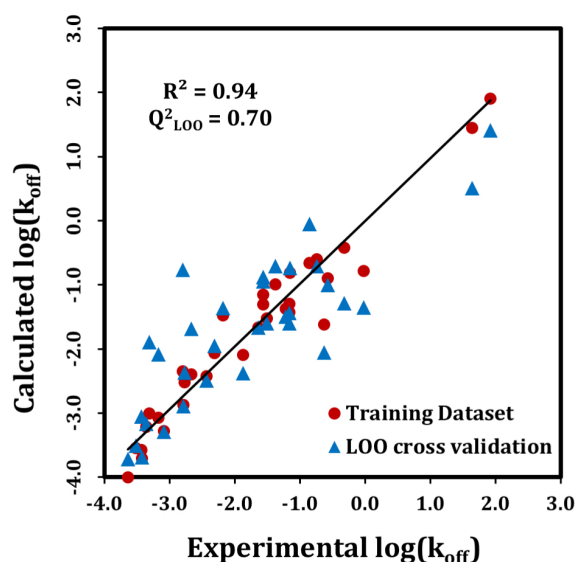


Figure 3.13: Correlation plot for experimental $\log(k_{off})$ values vs $\log(k_{off})$ values calculated by COMBINE analysis of HIV-1 protease inhibitors for the training dataset ($R^2=0.94$) and leave-one-out (LOO) cross-validation ($Q^2=0.78$). The straight line corresponds to $y=x$ (ideal case).

Compound Id	PDB Id	Experimental $\log_{10}(k_{off}(s^{-1}))$	Fitted $\log_{10}(k_{off}(s^{-1}))$ (PLS regression)	Predicted $\log_{10}(k_{off}(s^{-1}))$ (LOO validation)
B435	1D4H	-2.19 ± 0.09	-1.48	-1.36
A047	1G2K	-1.16 ± 0.10	-0.81	-0.74
A023	n.a.	-0.86 ± 0.11	-0.66	-0.05
A024	1G35	-1.16 ± 0.10	-1.30	-1.44
B429	n.a.	-3.43 ± 0.09	-3.71	-3.69
B409	1EC1	-3.37 ± 0.13	-3.21	-3.16
B268	n.a.	-2.44 ± 0.05	-2.43	-2.49
A045	n.a.	-0.58 ± 0.09	-0.90	-1.00
B425	1D4I	-0.63 ± 0.00	-1.62	-2.06
A021	n.a.	-1.56 ± 0.04	-1.16	-0.94
saquinavir	3OXC	-3.64 ± 0.06	-4.00	-3.72
indinavir	2BPX	-2.80 ± 0.04	-2.35	-0.77
ritonavir	1HXW	-2.67 ± 0.06	-2.39	-1.69
DMP323	1QBS	1.92 ± 0.12	1.91	1.41
nelfinavir	1OHR	-3.18 ± 0.04	-3.07	-2.08
B369	1EBY	-1.88 ± 0.19	-2.09	-2.38
B388	n.a.	-1.64 ± 0.15	-1.66	-1.67
A038	n.a.	-3.31 ± 0.02	-3.01	-1.89
A037	n.a.	-3.44 ± 0.04	-3.57	-3.05
B440	n.a.	-3.52 ± 0.02	-3.54	-3.51
B439	n.a.	-2.79 ± 0.06	-2.87	-2.89
B408	n.a.	-2.77 ± 0.02	-2.52	-2.37
B412	n.a.	-3.09 ± 0.19	-3.28	-3.29
A008	n.a.	1.64 ± 0.15	1.45	0.51
B277	n.a.	-2.31 ± 0.17	-2.07	-1.96
A030	n.a.	-1.38 ± 0.13	-1.00	-0.71
A015	n.a.	-0.03 ± 0.37	-0.78	-1.35
A016	n.a.	-1.22 ± 0.22	-1.37	-1.50
A017	n.a.	-0.75 ± 0.09	-0.61	-0.71
B322	n.a.	-1.17 ± 0.29	-1.43	-1.60
B365	n.a.	-1.51 ± 0.06	-1.52	-1.60
B347	n.a.	-1.57 ± 0.05	-1.31	-0.88
A018	n.a.	-0.32 ± 0.20	-0.43	-1.28

Table 3.9 continued from previous page

Compound	PDB	Experimental	Fitted	Predicted
Id	Id	$\log_{10}(k_{off}(\text{s}^{-1}))$	$\log_{10}(k_{off}(\text{s}^{-1}))$ (PLS regression)	$\log_{10}(k_{off}(\text{s}^{-1}))$ (LOO validation)

Table 3.9: Comparison of $\log_{10}(k_{off})$ values calculated by the COMBINE analysis model in PLS regression (column 4) and the experimental $\log_{10}(k_{off})$ values from Ref.[?] (column 3) for different HIV-1 protease inhibitors used for training the COMBINE analysis model. The $\log_{10}(k_{off})$ values predicted from leave-one-out (LOO) cross-validation for the training set of inhibitors are given in the last column.

Of the 17 coulombic and 17 LJ interactions considered in the PLS analysis, many make an unfavorable contribution to the dissociation kinetics (see Figure 3.14). It was observed that some of the interactions of the inhibitors, specifically with the residues in the flap region of HIV-1 protease, favor fast unbinding. For example, the cyclic urea and cyclic sulfamide inhibitors have direct polar contacts with the I50 residues located in the flap regions of the HIV-1 protease dimer and have fast dissociation rates. The flaps are very dynamic in nature and are known to exist in different conformations ranging from open to semiclosed to closed. Their fast movements could lead to these small cyclic compounds being driven out of the binding pocket. The cyclic urea inhibitors A008 and DMP323 have the highest k_{off} rate constants, and they have hydroxyl groups that make hydrogen bonds with the amide backbone atoms of both D30 residues in the bound complexes. The interaction with D30B was identified as unfavorable by the COMBINE analysis (Figure 3.14, bottom inset). The acyclic inhibitors, on the other hand, are peptidomimetic and have relatively slow dissociation rates. They do not form direct contacts with the flap residues and their aromatic groups mediate favorable LJ interactions with residues such as P81 and R08 (Figure 3.14, top inset). In some of the crystal structures of acyclic inhibitors complexed with HIV-1 protease, bridging waters mediate the interaction between the inhibitors and binding site residues such as D30 and I50. While interfacial water molecules can be considered explicitly in COMBINE analysis[95], we omitted them in this study. Thus, the effect of the water-mediated interactions that tend to correlate with slow dissociation rates appears to be represented implicitly by direct hydrogen-bonding to the corresponding residues in the complexes of fast dissociating inhibitors having negative weights in the PLS model.

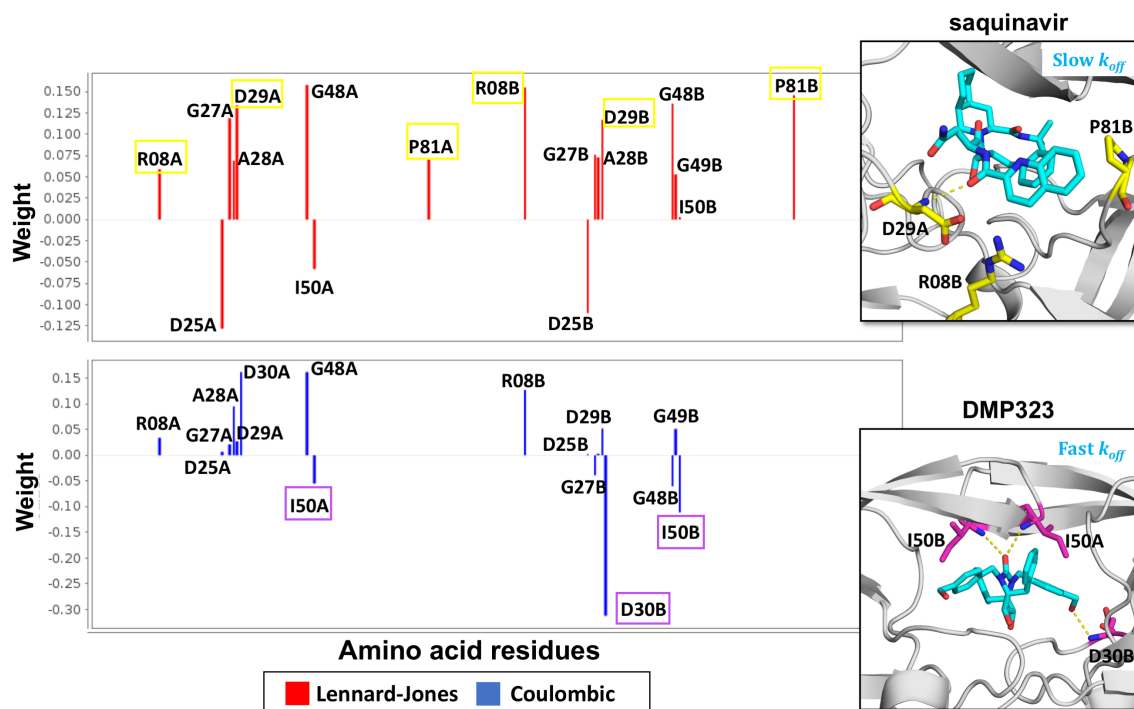


Figure 3.14: Weights for different LJ and coulombic interaction energy terms derived from the PLS analysis (projection to six latent variables, the value of constant C was 0.134). A negative weight means that an energetically favorable (negative) interaction energy term tends to shorten the residence time. The labels of some of the interaction energy terms that characterize slow and fast dissociating inhibitors are boxed, and the corresponding residues are also shown in the inset figures. The top inset shows a few of the interactions (yellow) contributing to the long residence time of the slowly dissociating inhibitor saquinavir ($k_{off} = 0.00023 \text{ s}^{-1}$) and the bottom inset shows the interactions (magenta) contributing to the short residence time of a very fast dissociating cyclic urea inhibitor DMP323 ($k_{off} = 83.3 \text{ s}^{-1}$) in the crystal structures with PDB IDs 3OXC and 1QBS, respectively. The figure is taken from **Ganotra and Wade, 2018**[109].

3.5 Concluding Discussions

Using COMBINE analysis, we have obtained QSKRs for k_{off} rates with very good predictive power ($Q_{LOO}^2 = 0.69$, $R_{PRED}^2 = 0.86$ for N-HSP90 and $Q_{LOO}^2 = 0.70$ for HIV-1 protease) and identified the key ligand–receptor interactions that contribute to the variance in binding kinetics. These specific interaction energy components provide insights into the mechanisms of specific slow and fast dissociating classes of compounds. Additionally, COMBINE analysis could be used to predict the effect of specific mutations in the protein on the dissociation kinetics of its inhibitors. COMBINE analysis was originally developed to derive QSARs for binding affinity (or K_d , the equilibrium dissociation constant) for a congeneric series of compounds

with a similar binding mode to a protein target. Here, we have not used congeneric series, but rather diverse sets of compounds with very different scaffolds and binding modes. We find that our COMBINE analysis models for K_d are not as predictive as the COMBINE models for k_{off} for these diverse sets of compounds (see Tables 3.10, 3.11, 3.12, 3.13). For deriving the model for K_d of the N-HSP90 inhibitors, the full dataset of 66 compounds was initially used for training. 3 outliers (compounds 17, 50 and 67) were later removed from the PLS analysis to improve the quality of the model. Therefore, the final model for K_d was trained with 63 compounds. In spite of training the model with the whole dataset, the model for K_d had only weak predictive ability with $R^2 = 0.59$ and $Q_{LOO}^2 = 0.41$ for 3 latent variables (see Tables 3.10 and 3.11). Similarly, for deriving the model for K_d of HIV-1 protease inhibitors, the full dataset of 36 compounds was initially used for training. 3 outliers (compounds B435, A037 and B249) were later removed from the PLS analysis to improve the quality of the model. Therefore, the final model for K_d of protease inhibitors was trained with 33 compounds.

LV	R^2	Q_{LOO}^2	AAE_T	AAE_V	RME_T	RME_V
1	0.33	0.13	0.66	0.74	0.90	1.02
2	0.55	0.34	0.58	0.68	0.73	0.89
3	0.59	0.41	0.57	0.67	0.70	0.84
4	0.64	0.39	0.52	0.67	0.66	0.85
5	0.69	0.40	0.49	0.67	0.61	0.85
6	0.70	0.32	0.48	0.68	0.60	0.90
7	0.72	0.16	0.47	0.74	0.58	1.00
8	0.74	0.09	0.45	0.76	0.56	1.04
9	0.74	-0.17	0.44	0.82	0.55	1.18
10	0.75	-0.75	0.44	0.91	0.54	1.44

Table 3.10: Summary of the models derived for different numbers of latent variables (LVs) for the COMBINE analysis for K_d of the N-HSP90 inhibitors. The models were derived using the $\log_{10}(K_d)$ value (unit of K_d is M) as the response variable in the PLS analysis. The table lists the regression coefficient (R^2) for the training set, the correlation coefficient for leave-one-out cross validation sets (Q_{LOO}^2), average absolute errors (AAE_T and AAE_V) and root mean squared errors (RME_T and RME_V) for the training set and leave-one-out validation sets, respectively.

Validation	Q^2	AAE_V	RME_V
Leave-one-out (LOO)	0.41	0.67	0.84
Leave-two-out (L2O)	0.41	0.68	0.84
Leave-three-out (L3O)	0.38	0.70	0.86
Random groups of 7 (10 iterations)	0.37	0.69	0.87

Table 3.11: Statistical measures of correlation for the COMBINE Analysis models Derived for $\log(K_d)$ of N-HSP90 inhibitors. The table lists the cross-validated correlation coefficient (Q^2), average absolute errors (AAE_V) and root mean squared errors (RME_V) for different validation methods for the PLS model derived with 3 latent variables for N-HSP90 inhibitors.

LV	R^2	Q^2_{LOO}	AAE_T	AAE_V	RME_T	RME_V
1	0.30	0.08	1.04	1.22	1.23	1.42
2	0.39	0.16	0.95	1.13	1.15	1.35
3	0.55	0.27	0.78	1.05	0.99	1.26
4	0.64	0.37	0.68	0.96	0.88	1.17
5	0.69	0.45	0.63	0.85	0.82	1.09
6	0.78	0.53	0.50	0.79	0.69	1.02
7	0.80	0.53	0.47	0.78	0.65	1.01
8	0.83	0.53	0.45	0.76	0.60	1.01
9	0.86	0.43	0.41	0.89	0.54	1.11
10	0.89	0.24	0.39	1.01	0.50	1.28

Table 3.12: Summary of the models derived for different numbers of latent variables (LVs) for the COMBINE analysis for K_d of the HIV-1 protease inhibitors. The models were derived using the $\log_{10}(K_d)$ value (unit of K_d is M) as the response variable in the PLS analysis. The table lists the regression coefficient (R^2) for the training set, the correlation coefficient for leave-one-out cross validation sets (Q^2_{LOO}), average absolute errors (AAE_T and AAE_V) and root mean squared errors (RME_T and RME_V) for the training set and leave-one-out validation sets, respectively.

The R^2 and Q^2_{LOO} for the COMBINE analysis model derived with 6 latent variables are 0.78 and 0.53 respectively (see Tables 3.12 and 3.13). We do however, obtain better statistics for a COMBINE model for K_d generated with a smaller data set of resorcinol compounds that inhibit HSP90 and have a similar scaffold (Table 3.14). A possible explanation for the better predictions for k_{off} than K_d may be that dissociation rates are independent of the unbound state, and therefore differences in ligand and protein desolvation and conformational free energies are not so

important.

Validation	Q^2	AAE_V	RME_V
Leave-one-out (LOO)	0.53	0.79	1.02
Leave-two-out (L2O)	0.46	0.81	1.08
Leave-three-out (L3O)	0.44	0.81	1.10
Random groups of 5 (10 iterations)	0.48	0.83	1.06

Table 3.13: Statistical measures of correlation for the COMBINE Analysis models Derived for $\log(K_d)$ of HIV-1 protease inhibitors. The table lists the cross-validated correlation coefficient (Q^2), average absolute errors (AAE_V) and root mean squared errors (RME_V) for different validation methods for the PLS model derived with 6 latent variables for HIV-1 protease inhibitors.

Validation	Q^2	AAE_V	RME_V
Leave-one-out (LOO)	0.49	0.47	0.57
Leave-two-out (L2O)	0.45	0.47	0.59
Leave-three-out (L3O)	0.47	0.46	0.58
Random groups of 5 (10 iterations)	0.43	0.49	0.61

Table 3.14: Statistical measures of correlation for the COMBINE Analysis models Derived for the $\log(K_d)$ of the resorcinol series of inhibitors of N-HSP90. For deriving the model for K_d , a smaller dataset of 25 inhibitors belonging to the resorcinol series was used for training. 3 outliers (compounds 23, 28, 30) were later removed from the PLS analysis to improve the quality of the model. Therefore, the final model for K_d was trained with 22 compounds. The table lists the cross-validated correlation coefficient (Q^2), average absolute errors (AAE_V) and root mean squared errors (RME_V) for different validation methods used. These statistical measures correspond to a model derived with 4 latent variables in PLS analysis.

The current applications to HSP90 and HIV-1 protease data sets with very diverse sets of inhibitors, using both crystal structures and modeled protein-inhibitor complexes, demonstrates the potential of COMBINE analysis as a robust QSKR approach with increasing scope for application as more data sets of measured kinetic parameters become available. COMBINE analysis complements a growing number of methods based on biomolecular simulation and machine learning to predict drug-target binding kinetics[9]. Indeed, a possible extension of the COMBINE analysis approach would be to the analysis of structures from molecular dynamics

simulations, including intermediates along drug binding or unbinding pathways.

Chapter 4

Halogen-aromatic π interactions modulate inhibitor residence time

This Chapter is based on the following publication:

Halogen–Aromatic π Interactions Modulate Inhibitor Residence Times.

Christina Heroven, Victoria Georgi, Gaurav K. Ganotra, Paul E Brennan, Finn Wolfreys, Rebecca C. Wade, Amaury E. Fernández-Montalván, Apirat Chaikuad, Stefan Knapp, *Angew. Chemie Int. Ed.* 2018, 57 (24), 7220–7224.

4.1 Background

Designing drug molecules with longer residence times may result in increased drug efficacy and prolonged inhibition after the free drug concentration has dropped owing to *in vivo* clearance. Having slow off-rates specific for the target may also result in kinetic-selectivity over off-targets with high dissociation rates despite similar binding constants[110]. Kinases are particularly dynamic proteins and after G-protein-coupled receptors (GPCRs), kinases are the second most important group of drug targets[111] under study with more than 150 kinase inhibitors in preclinical trials waiting FDA approval[112]. There has been several reasons attributed to the slow off-rates of several kinase inhibitors already approved as clinical drugs. In most cases, an induced fit binding mechanism results in slow dissociation rates, where the dissociation of the drug will require the structural rearrangement of the target. For example, slow binding kinetics of a type II inhibitor of p38 MAP kinase, BIRB-796, was suggested to be the result of its binding to an inactive conformation in which

the DFG motif is displaced in a so-called “DFG-out” conformation[113]. However, not all type II inhibitors that bind to “DFG-out” conformation show slow binding kinetics[114]. Indeed, recent study by Schneider *et al.*[115] suggested that the slow off-rates were the result of efficient hydrophobic contacts rather than the kinetic dissociation barrier introduced by the DFG-out transition. In the case of the type I CDK inhibitor roniciclib, the long residence time was considered to be the result of changes in the arrangement of water molecules coupled to conformational adaptation of the DFG motif[116]. In some cases, the presence of water-shielded hydrogen bonds can also lead to slow off-rates[117].

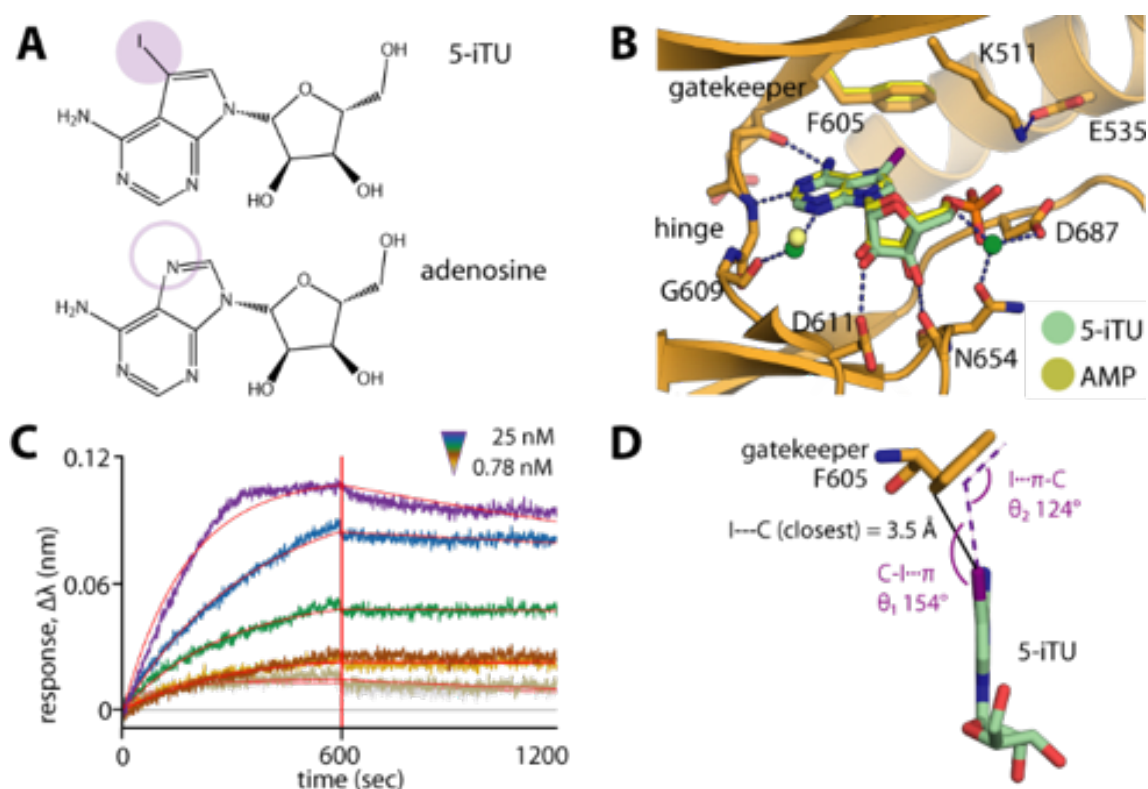


Figure 4.1: The 5-iodotubercidin inhibitor (5-iTU) exhibits tight binding with slow dissociation kinetics from haspin. A) Chemical structures of 5-iTU and adenosine. B) Superimposition of haspin-5iTU and AMP (pdb id: 4ouc) reveals similar binding modes of the two compounds. C) BLI sensorgram suggests slow kinetic behavior of the 5-iTU-haspin interaction. D) The iodide and the benzene moieties of 5-iTU and F605, respectively, are located in close proximity with a favorable geometry for a halogen- π bond. The figure is taken from **Heroven *et al.***[118] with permissions.

Herein, we present data that suggest that targeting the interactions between halogen atoms, commonly found in drugs, and the aromatic residues, which are also typically found in drug binding sites on proteins, can be utilized to design inhibitors

with long residence times. We (Heroven *et al.*[118]) chose haspin, a serine/threonine kinase with a known three-dimensional structure, as a model system and the close analogue of ATP, 5-iodotubercidin (5-iTU) as a model inhibitor for an archetypal active state (type I) kinase-inhibitor binding mode. From the analysis of 3D crystallographic structures of haspin, it was observed that the binding modes of both 5-iTU and the nucleoside adenosine, are highly conserved (see Figure 4.1 A,B). Both of these molecules are very similar except for the presence of the iodide moiety in 5-iTU, in close proximity to the F605 gatekeeper, which forms a halogen–aromatic π interaction (see Figure 4.1 D). Compared to ATP or adenosine ($K_d = 180 \mu\text{M}$), 5-iTU showed a very high affinity ($K_d = 0.78 \text{ nM}$) for haspin and an unexpectedly long residence time. This tight binding of 5-iTU with slow binding kinetics was further confirmed by isothermal titration calorimetry (ITC), biolayer interferometry (BLI), and surface plasmon resonance (SPR) experiments (see Figure 4.1 C for BLI results). We (Heroven *et al.*[118]) therefore hypothesized that this halogen- π interaction between iodide of 5-iTU and aromatic gatekeeper residue (phenylalanine) of haspin contributes to most of the increase in the binding free energy (ΔG) and could be responsible for the long residence time of 5-iTU.

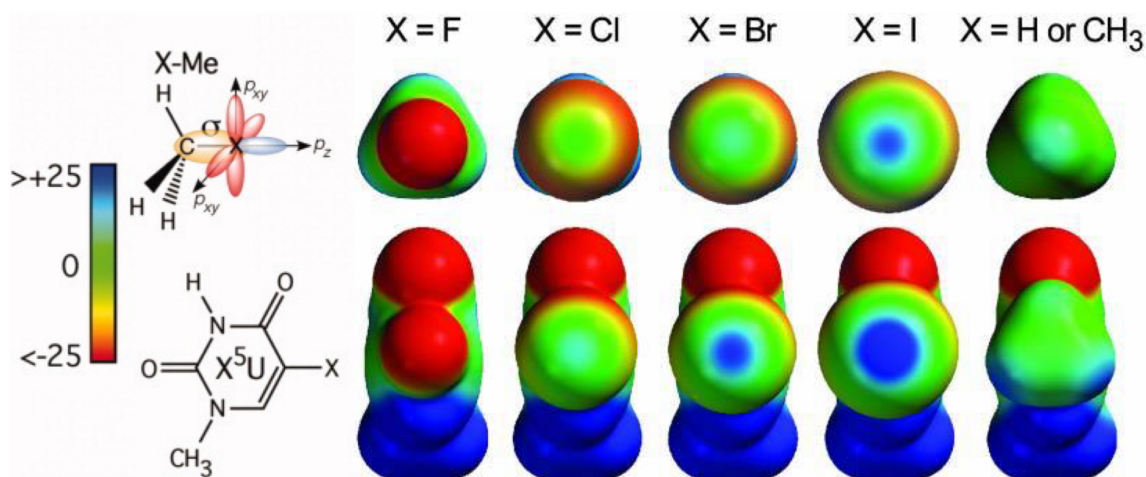


Figure 4.2: The σ -hole model and the polarization of the electrostatic surface potential. Electrostatic surface potentials of halogenated methane (top) and uridine nucleobase (bottom), adapted from Auffinger *et al.*[119] and Scholfield *et al.*[120]. Potential energies are shown from red (negative) to green (neutral) to blue (positive), viewing into the halogen atoms down the X-C bond. The resulting polarization increases with the atomic size of the halogen, following the series $F < Cl < Br < I$. For comparison, the potential surface of methane and methylated uridine is shown on the right.

Quantum mechanical calculations suggest that there is an anisotropic distribution of the electrostatic potential across the surface of the halogen[121]. According to σ -hole model[122], this direction polarization is the result of the covalent σ -bond (C-X) between a carbon atom and the halogen. Halogens have five electrons in the p -atomic orbitals of their valence shell. While both p_x and p_y orbitals have paired electrons, it is the single valence electron of the p_z orbital that is involved in formation of the covalent C-X σ -bond. This depopulation of the p_z orbital directly opposite the C-X σ -bond creates a hole, also known as a σ -hole, that partially exposes the positive nuclear charge of the halogen. The magnitude of the σ -hole depends on the the polarizability of the halogen, which follows the series I > Br > Cl > F (see Figure 4.2). This σ -hole magnitude also depends on the electron-withdrawing ability of the molecule that the halogen is covalently bound to. As the partial positive charge exposed along the C-X σ -bond diminishes with decreasing size of the halogen atom, the iodide in 5-iTU was substituted by smaller halogen atoms (Br, Cl and F) and the affinities and binding kinetics of these 5-iTU derivatives were characterized using experimental assays. The affinities of 5-iTU derivatives diminished with decreasing size of the substituted halogen atom, as confirmed by ITC experiments (see Figure 4.3 A). In comparison to 5-iTU, removal of the halogen atom in tubercidin (TU), resulted in a 42-fold decrease in the potency (see Figure 4.3 A). Similarly, substitution of iodide in 5-iTU with a smaller sized fluoride led to a eightfold decrease of potency in 5-fluorotubercidin (5-fTU). Binding kinetics of these five synthesized 5-tubercidin halogen derivatives with haspin were performed using three independent techniques: kinetic probe competition assays (kPCAs), BLI, and SPR. The K_d values determined by these three independent methods correlated well with each other and also with the binding constants determined in solution by ITC (see Figure 4.3 B). The off-rates (k_{off}) calculated from three different experimental showed the same behavior with the off-rates increasing with decreasing halogen size from the 5-iodo- to the 5-fluoro-substituted tubercidin, and the unsubstituted tubercidin showed the fastest off-rate (see Figure 4.3 C,D).

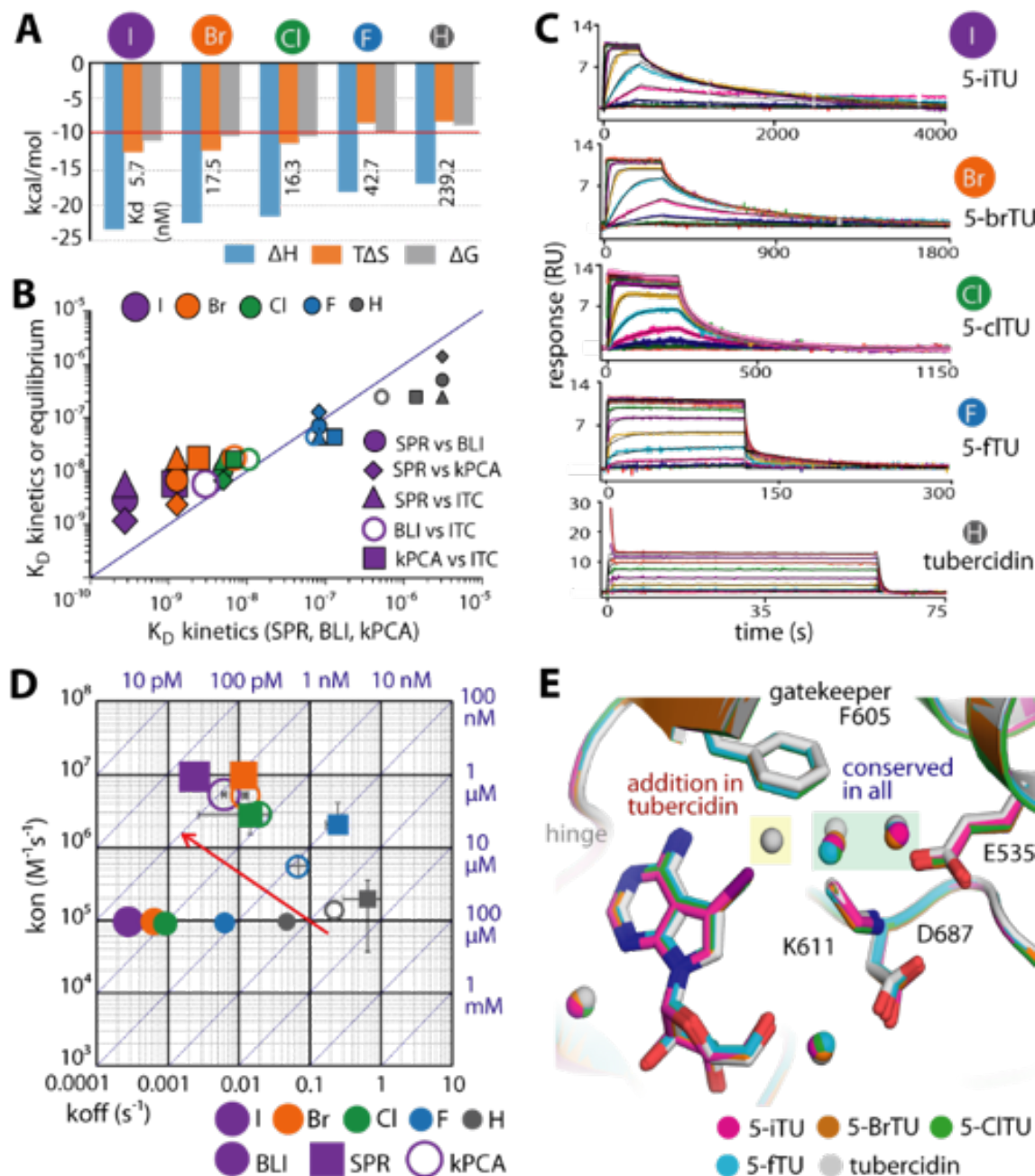


Figure 4.3: Binding kinetics of haspin with five tubercidin derivatives harboring halogen substituents at the 5-position. A) ITC thermodynamic binding parameters. B) Comparison of dissociation constants (K_D) measured by ITC, BLI, SPR and kPCA shows good correlation of the measured equilibrium data. C) SPR sensorgrams demonstrating increasingly slow dissociation rates with increasing size of the halogens. D) Rate plot with Isoaffinity Diagonal (RaPID) of k_{on} and k_{off} constants measured by BLI, SPR and kPCA. The red arrow indicates the trend to increasing k_{on} and decreasing k_{off} upon increasing the atomic radii of the halogens. E) Crystal structures reveal conserved binding modes of all five tubercidin derivatives, albeit with an additional water molecule adjacent to the inhibitor and F605 gatekeeper in tubercidin. The figure is taken from Heroven *et al.*[118] with permissions.

4.2 Aim of the work

Based on the observations from kinetic, thermodynamic, and structural measurements performed for characterization of the halogen-aromatic π interactions in the haspin-inhibitor complexes, we wanted to answer few important questions such as:

- 1) How is the polarization-mediated anisotropic charge distribution on halogen atoms involved in mediating such a strong electrostatic interaction in the binding pocket of kinase?,
- 2) What are the different energetic components that contribute to this electrostatic interaction?,
- 3) Can the mechanism of prolonging drug-target residence times by introducing aromatic-halogen interactions be applied to other kinases having aromatic residues at the gatekeeper position?

To answer these questions, we decided to perform quantum mechanical calculations as it was important to consider polarization and electronic effects for explaining such polarization mediated interactions. Therefore, to analyze the nature of the interactions of the core inhibitor scaffold with the gatekeeper aromatic residues, we calculated post-Hartree-Fock *ab initio* interaction energies using the Møller–Plesset perturbation theory (MP) to second order (MP2). Then, we evaluated how different contributions to the interaction energy, derived from Energy Decomposition Scheme based on Hybrid Variation-Perturbation Theory, perform in scoring the residence times and affinities. To account for the complete protein structure in the computation of the binding free energies of the haspin–ligand complexes, we used the classical MM/GBSA approach with an implicit solvent model. We also applied a new computational approach based on Molecular dynamics Simulations: τ -Random Acceleration Molecular Dynamics (τ -RAMD)[29] for prediction of relative dissociation rates.

4.3 Methods

4.3.1 Quantum mechanical interaction energy calculations

The energy contributions of the inhibitor-aromatic gatekeeper interaction in haspin-inhibitor complexes were calculated using *ab initio* Møller–Plesset perturbation the-

ory to second order (MP2). The *Protein Preparation Wizard*[104] of the Schrödinger suite (release 2015-4, Schrödinger, LLC, New York) was used to pre-process the X-ray crystallographic structures of the haspin-inhibitor complexes, to add missing side chains and to optimize the H-bond network. The *impref*[104] program of the Schrödinger suite was used for energy minimization using the *OPLS3*[123] force field. The default minimization protocol using *impref* involves optimization of the position of hydrogen atoms followed by all-atom minimization where non-hydrogen atoms are restrained with a harmonic potential using a force constant of 25 kcal/mol.Å². The coordinates of the inhibitor and the gatekeeper phenylalanine residue were extracted from energy-minimized structures of haspin-inhibitor complexes. The termini of the phenylalanine residue were capped with hydrogen atoms and their positions were optimized using the *OPLS3* force field in the *Maestro* program of the Schrödinger suite (release 2015-4, Schrödinger, LLC, New York). In the case of the gatekeeper mutants, the corresponding gatekeeper residues (tyrosine and threonine) were prepared in the same way. The *def2TZVP* basis set was used for all calculations and effective core potentials (ECPs) were used for the iodine atom. *Ab initio* interaction energies at the MP2 level were calculated using the GAMESS[76] software, and partitioned into their constituent interaction energy terms using the many body interaction energy decomposition scheme (EDS) described by Góra *et al.*[78, 77].

4.3.2 Binding free energy calculations using MM/GBSA

The molecular mechanics-generalized Born surface area (*MM/GBSA*) method was used to estimate the binding free energy of the inhibitors to haspin kinase. The initial coordinates of the haspin-inhibitor complexes were obtained from the co-crystallized structures. The *Protein Preparation wizard*[104] of the *Schrödinger* suite (release 2015-4, Schrödinger, LLC, New York) was used for pre-processing of the structures, formation of disulfide bonds, addition of hydrogen atoms and assigning protonation states at *pH* 7.0. The *pmemd* module of the *Amber14* software suite [79] was used to perform the Molecular dynamics (MD) simulations together with the *Amber ff14SB*[107] force field for protein. The *LEap* module of *AmberTools14* was used to construct the topologies of the haspin-inhibitor complexes. The ligand parameters were generated based on the generalized Amber force field[124] (*GAFF*). To improve the description of charge, dipole moment and geometry of halogenated compounds in molecular mechanical calculations, the positive region (σ -hole) centered on the

halogen atom was represented by an extra-point charge (*EP*). This inclusion of an *EP* results in improved modeling of halogen-bonding in MD simulations[125]. The force field parameters for this *EP* were taken from Ibrahim *et al.*[125]. For generation of the partial atomic charges for the ligands, the *RESP*[106] program was used to fit the atom-centered charges to the molecular electrostatic potential (MEP) grid computed by the *GAMESS*[76] program. The system was centred and aligned with the axes to minimize the volume. The system was then solvated using the *TIP3P*[126] water model by immersing the protein-ligand complex in a cubic box of water molecules, such that the shortest distance between the edge of the solvation box and the complex is 10 Å. The net charge ($-2e$) of the system was then neutralized by adding counter ions such as Na⁺ and Cl⁻ ions. For each system, energy minimization was performed in three 1500-cycle consecutive runs using the steepest descent minimization method followed by switching to the conjugate gradient method after 500 cycles. Gradually decreasing harmonic restraints with force constants of 500, 1 and 0 kcal/mol.Å² were used for non-hydrogen atoms in three consecutive runs. Energy minimization was followed by 1 ns of gradual heating from 10 K to 300 K with harmonic restraints with a force constant of 50 kcal/mol.Å² acting on non-hydrogen atoms. Then the system was equilibrated for 1 ns under NPT conditions at 300K, with heavy atoms (except solvent Na⁺ and Cl⁻ ions) harmonically restrained with a force constant of 50 kcal/mol.Å². This was followed by an NPT equilibration of 2 ns without any positional restraints. The potential energy function and atomic coordinates were calculated using a 2 fs time step. The *SHAKE*[62] algorithm was used to constrain all the bonds involving hydrogen atoms. The Particle Mesh Ewald (PME)[127] method was used to calculate the electrostatic interactions. A cut-off of 10 Å was set for generating the non-bonded pair list and this pair list was updated after every 100 steps. After equilibration, data were collected over 6 ns of a simulation run for binding free energy calculations and 3000 sets of atomic coordinates were saved every 2 ps.

Then, MMGBSA calculations of the binding free energy were performed using the *MMPBSA.py* module implemented in the Amber14 analysis tools. A single-trajectory approach was used in which receptor, ligand and complex geometries were extracted from a single MD trajectory. All the ions and water molecules were stripped from the trajectory snapshots. A salt concentration of 0.15 M and the Born implicit solvent model (*igb=2*) was used. Each binding free energy was com-

puted as the sum of a molecular mechanics term (ΔE_{gas}), a Gibbs solvation term ($\Delta\Delta G_{solvation}$) and an entropic contribution ($T\Delta S_{solute}$). For the entropic contribution to binding free energy, we computed translational and rotational entropies with a rigid rotor model using the *MMPBSA.py* module. The calculation of vibrational entropies using normal-mode analysis with *MMPBSA.py* failed due to the inclusion of the *EP* in the force field. The free energy of binding for some of the derivatives is positive since vibrational and conformational entropy terms are neglected.

4.3.3 τ -Random Acceleration Molecular Dynamics (τ -RAMD) simulations

τ -RAMD is a computationally efficient procedure that involves application of Random acceleration molecular dynamics simulations[28], to compute relative dissociation rates for ligand unbinding from the binding site of proteins[29]. To perform τ -RAMD simulations, X-ray crystallographic structures of haspin-inhibitor complexes were used as the starting structures. The *Protein Preparation wizard*[104] of the Schrodinger suite (release 2015-4, Schrödinger, LLC, New York) was used to pre-process the structures, to add missing side chains, and to optimizing the H-bond network. The topologies of the systems were constructed using the *LEap* program of the Amber14 software. The Amber *ff14SB*[107] force field was used for the proteins, the General Amber Force Field[124] (*GAFF*) for inhibitors, and the *TIP3P*[126] model for waters. The partial atomic charges of the ligands were calculated according to the *AM1-BCC*[128, 129] method using *Antechamber* module of Amber14. The systems were then solvated in a cubic box of water, with water molecules extending at least 14Å between the complex and the edge of the box. Na⁺ counter ions were added to neutralize the net charge of the system. The *pmemd* module of Amber14 was used for carrying out energy minimization in four 1500 cycle consecutive runs using the steepest descent minimization method followed by switching to the conjugate gradient method after 500 cycles. Gradually decreasing harmonic restraint force constants of 500, 1 and 0.005 kcal/mol.Å² were used for heavy atoms in the first three consecutive runs and no positional restraints were used in the final minimization run. Each system was then heated for 40 *ps* from 10 *K* to 300 *K* with a restraint with a force constant of 50 kcal/mol.Å² acting on the heavy atoms. Then, the systems were equilibrated in the NPT ensemble using a Langevin thermostat and Nosé–Hoover Langevin pressure control to maintain the system at 1 *atm* and 300

K. A two-stage equilibration for 400 *ps* (200 *ps* in each stage) was performed using the *pmemd* module of Amber14. In the first stage, heavy atoms except Na⁺ and Cl⁻ ions were restrained with a harmonic with a force constant of 50 kcal/mol.Å². In the second stage, NPT equilibration with no restraints was performed for each system. The *SHAKE* algorithm was used to constrain all bonds involving hydrogen atoms and a time step of 2 *fs* was used. Electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method. The atomic coordinates of the equilibrated system generated with AMBER were used as input for the production run with the NAMD software[80]. Then, 2 *ns* long standard MD simulations were run using the *NAMD* software with Langevin dynamics applied for constant temperature (300K) and pressure control (1 *atm*). Atom pairs that were less than 14 Å apart were included in the pair list, and non-bonded interactions were calculated at every step for atom pairs that were within 12 Å cut-off distance. The atomic coordinates and velocities obtained after 2 *ns* of this production run were used as the starting input for the Random Acceleration Molecular Dynamics (*RAMD*) simulations. The *RAMD* simulation procedure[28] was implemented as a Tcl wrapper around the NAMD software[80], and recently this Tcl wrapper was modified to take the force magnitude rather than acceleration as an input parameter and to use new functions available from version 2.10 of NAMD onward[29].

The *RAMD* simulations were performed in an explicit solvent with parameters identical to the standard MD simulations. A randomly oriented force (F) of magnitude 8.0 kcal/mol.Å was applied to the centre-of-mass of the ligand and the movement of the ligand was assessed after every 50 MD simulation steps (100 *fs*). If the change in distance r moved by the ligand in this time was less than a threshold distance of $r_{min} = 0.025$ Å, a new random direction of the force was generated. Otherwise, the simulation was continued for the next 50 simulation steps with the same direction of the force. The simulation was stopped when the distance between the ligand and protein centre-of-mass exceeded 30 Å or if the ligand exit was not observed within 3 *ns*. Coordinates were saved at 1 *ps* intervals. A set of 20 *RAMD* dissociation trajectories was generated for each ligand by varying the initial direction of the artificial force F . The ligand egress times of the *RAMD* trajectories were recorded for all the ligands employed in this method. As in the procedure described by Kokh *et al.*[29], the residence time, τ , was defined as the simulation time required for ligand dissociation in 50% of the trajectories. A bootstrapping procedure

(200 sets, each of the sets containing 80% of the trajectories chosen randomly) was employed to compute the final residence time, τ_{comp} , and its standard deviation.

4.4 Results and Discussions

4.4.1 The second order Møller-Plesset interaction energies (E_{MP2}) between the inhibitor and the gatekeeper residue correlate well with dissociation rate constants and equilibrium dissociation constants determined experimentally

The second-order Møller–Plesset interaction energies (E_{MP2}) between the 5-iTU derivatives and the gatekeeper phenylalanine (F605) residue were calculated at consecutive levels of quantum mechanical theory and are outlined in Tables 4.1. The contributions of isolated constituent energy terms to total interaction energy are given in Table 4.2. E_{MP2} energies were also calculated between 5-iTU and mutated gatekeeper residues (Tyrosine and Threonine) at the 605 position (F605Y and F605T) (see Tables 4.3 and 4.4). The coordinates for these tyrosine and threonine gatekeepers were extracted from the resolved crystallographic structures of mutant haspin in complex with 5-iTU. We obtained a good correlation between interaction energies calculated at the MP2 level of theory (E_{MP2}) and the dissociation rate constants (k_{off}) and affinities determined experimentally (see Figure 4.4 A,C). E_{MP2} energies correlate well with the size of halogens substituted on the inhibitors. The Pearson correlation coefficient (R^2) between calculated E_{MP2} energies (kcal/mol) for 5-iTU derivatives with the gatekeeper (F605) residue and their $\log(k_{off})$ values from SPR experiments is 0.93. Similarly, calculated E_{MP2} energies also correlated well with affinities (K_d) from ITC experiments with a Pearson correlation coefficient of 0.80. The E_{MP2} interaction energy calculated for 5-iTU with F605 was approximately 1.5 kcal/mol higher than the E_{MP2} energy calculated for tubercidin (Table 4.1).

Inhibitor	$\log k_{off}$ (SPR)	$\log k_{off}$ (BLI)	$\log K_d$ (SPR)	$\log K_d$ (ITC)	E_{EL}	E_{HL}	E_{SCF}	E_{MP2}
5-iTU	-2.62	-3.56	-9.10	-8.24	-1.85	2.79	2.15	-3.01
5-brTU	-1.92	-3.19	-8.59	-7.76	-1.09	2.22	1.78	-2.74
5-clTU	-1.84	-3.04	-8.18	-7.79	-0.52	1.91	1.58	-2.24
5-ftTU	-0.60	-2.20	-6.81	-7.37	0.17	1.15	0.94	-1.44
tubercidin	-0.18	-1.32	-5.59	-6.62	-0.09	0.65	0.47	-1.48

Table 4.1: Total interaction energy [kcal. mol⁻¹] between tubercidin derivatives and gatekeeper Phe 605 residue at consecutively increasing levels of quantum mechanical theory. E_{EL} is the electrostatic energy only, E_{HL} includes the Heitler-London energy, E_{SCF} includes the Hartree-Fock energy as well, and E_{MP2} is the full Moeller-Plesset second order energy. k_{off} values were measured by SPR and BLI, and K_d values were measured by SPR and ITC.

Inhibitor	$\log k_{off}$ (SPR)	$\log k_{off}$ (BLI)	$\log K_d$ (SPR)	$\log K_d$ (ITC)	$E_{EL,MTP}$	$E_{EL,PEN}$	E_{EX}	E_{DEL}	E_{CORR}
5-iTU	-2.62	-3.56	-9.10	-8.24	2.48	-4.33	4.63	-0.64	-5.16
5-brTU	-1.92	-3.19	-8.59	-7.76	1.03	-2.12	3.30	-0.42	-4.54
5-clTU	-1.84	-3.04	-8.18	-7.79	-0.21	-0.32	2.44	-0.34	-3.82
5-ftTU	-0.60	-2.20	-6.81	-7.37	0.13	0.04	0.97	-0.21	-2.37
tubercidin	-0.18	-1.32	-5.59	-6.62	-0.07	-0.01	0.74	-0.18	-1.94

Table 4.2: Contribution of the different interaction energy terms to the total interaction energy, E_{MP2} [kcal. mol⁻¹], between tubercidin derivatives and the gatekeeper Phe 605 residue. $E_{EL,MTP}$ is the electrostatic multipole term, $E_{EL,PEN}$ is the penetration electrostatic term, E_{EX} is the exchange term, E_{DEL} is the delocalization term, and E_{CORR} is the correlation energy term. k_{off} values were measured by SPR and BLI, and K_d values were measured by SPR and ITC.

Partitioning of E_{MP2} into its constituent energy components (see Table 4.2) using a many-body interaction energy decomposition scheme showed that the major contribution to E_{MP2} comes from the correlation energy (E_{CORR}). E_{CORR} describes second-order intermolecular dispersion interactions and the correlation corrections to the Hartree-Fock energy. Similar to E_{MP2} , E_{CORR} increases in magnitude with an increase in the size of the halogen. A very high correlation was observed between E_{CORR} and the dissociation rates measured experimentally with $R^2 = 0.97$

(see Figure 4.4 B). This indicates the importance of the halogen interaction with the aromatic gatekeeper for the prolongation of residence times as the halogen size increases. The computed *ab initio* energies also correlate for the interaction of 5-iTU with the F605Y mutant but the magnitude of the interaction energy of 5-iTU with the threonine mutant (F605T) was underestimated (see Table 4.3 and Figure 4.4).

System	log k_{off} (SPR)	log k_{off} (BLI)	log K_d (SPR)	log K_d (ITC)	E_{EL}	E_{HL}	E_{SCF}	E_{MP2}
Wild type	-2.62	-3.56	-9.10	-8.24	-1.85	2.79	2.15	-3.01
F605Y	ND	-3.37	ND	-8.34	-2.27	3.01	2.30	-3.13
F605T	-1.72	-2.91	-7.89	-8.38	-0.78	1.38	0.99	-1.14

Table 4.3: Total interaction energy [kcal. mol⁻¹] between 5-iTU and the gatekeeper residue for the wild type and the two mutants, at consecutively increasing levels of quantum mechanical theory. E_{EL} is the electrostatic energy only, E_{HL} includes the Heitler-London energy, E_{SCF} includes the Hartree-Fock energy as well, and E_{MP2} is the full Moeller-Plesset second order energy. k_{off} values were measured by SPR and BLI, and K_d values were measured by SPR and ITC.

System	log k_{off} (SPR)	log k_{off} (BLI)	log K_d (SPR)	log K_d (ITC)	$E_{EL,MTP}$	$E_{EL,PEN}$	E_{EX}	E_{DEL}	E_{CORR}
Wild type	-2.62	-3.56	-9.10	-8.24	2.48	-4.33	4.63	-0.64	-5.16
F605Y	ND	-3.37	ND	-8.34	0.41	-2.68	5.28	-0.71	-5.43
F605T	-1.72	-2.91	-7.89	-8.38	-6.97	6.19	2.17	-0.39	-2.13

Table 4.4: Contribution of the different interaction energy terms to the total interaction energy, E_{MP2} [kcal.mol⁻¹] between 5-iTU and the gatekeeper residue for the wild type and the two mutants. $E_{EL,MTP}$ is the electrostatic multipole term, $E_{EL,PEN}$ is the penetration electrostatic term, E_{EX} is the exchange term, E_{DEL} is the delocalization term, and E_{CORR} is the correlation energy term. k_{off} values were measured by SPR and BLI, and K_d values were measured by SPR and ITC.

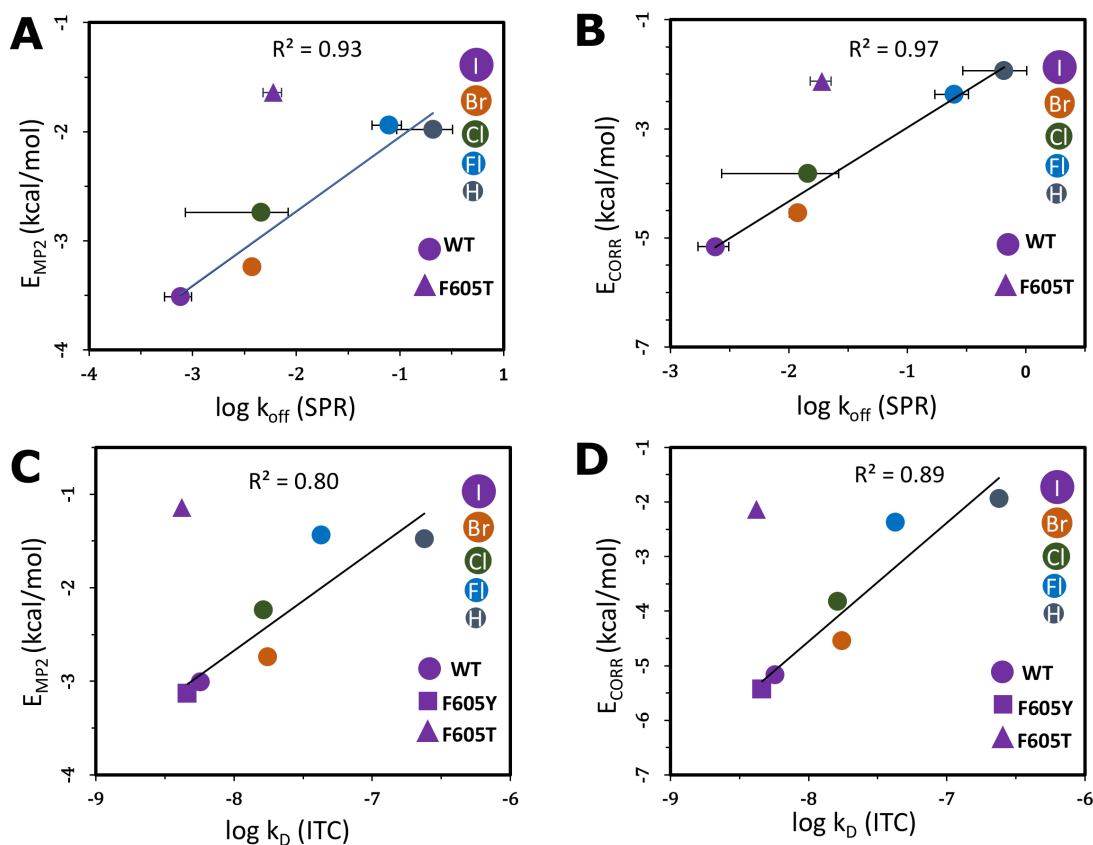


Figure 4.4: Correlation plots of computed quantum mechanical energies against experimental binding parameters. A) Second-order Møller-Plesset interaction energy (E_{MP2}) between the tubercidin derivatives and the gatekeeper residue versus the experimental (SPR) dissociation rate constants (k_{off}) of the tubercidin derivatives. B) The second-order correlation correction energy term (E_{CORR}) for the interaction between the tubercidin derivatives and the gatekeeper residue versus the experimental (SPR) dissociation rate constants (k_{off}) of the tubercidin derivatives. This correlation energy (E_{CORR}) includes second-order intermolecular dispersion interactions and the correlation corrections to the Hartree-Fock (HF) energy. C) Second-order Møller-Plesset interaction energy (E_{MP2}) between tubercidin derivatives and gatekeeper residue versus the experimental (ITC) binding affinities (K_d) of the tubercidin derivatives. D) Second-order correlation correction energy term (E_{CORR}) for the interaction between the tubercidin derivatives and the gatekeeper residue versus the experimental (ITC) binding affinities (K_d) of the tubercidin derivatives. The correlation coefficients (R^2) and the linear fits were computed omitting the outlier data points for the F605T mutant. The error bars for the K_d (ITC) values are smaller than the size of the data point symbols.

4.4.2 Binding free energies calculated from MM/GBSA approach correlate with experimental parameters for the halogen-gatekeeper interaction.

In order to account for the complete protein structure, we computed the binding free energies of the haspin-inhibitor complexes using the classical MM/GBSA approach with an implicit solvent model (see Table 4.5). Each binding free energy was computed as the sum of a molecular mechanics term (ΔE_{gas}), a Gibbs solvation term ($\Delta\Delta G_{solvation}$) and an entropic contribution ($T\Delta S_{solute}$). Some ΔG_{MMGBSA} values are positive as they only include translational and rotational entropic terms and do not include vibrational and conformational entropy contributions.

The binding free energies computed correlate well with the calorimetric data measured by ITC, with the Pearson correlation coefficient, $R^2 = 0.83$ between enthalpic energies computed by MM/GBSA ($\Delta E_{gas} + \Delta\Delta G_{solvation}$) and ITC enthalpies (ΔH_{ITC}). Also, a good correlation ($R^2 = 0.62$) was observed between the binding free energies computed from MM/GBSA (ΔG_{MMGBSA}) and the ITC binding free energies (ΔG_{ITC}) of the interactions between haspin and halogenated derivatives of 5-iTU. Therefore, the binding free energies computed from MM/GBSA, are consistent with the increasingly favorable enthalpic contribution to binding as the halogen size increases (see Figure 4.5 and Table 4.5).

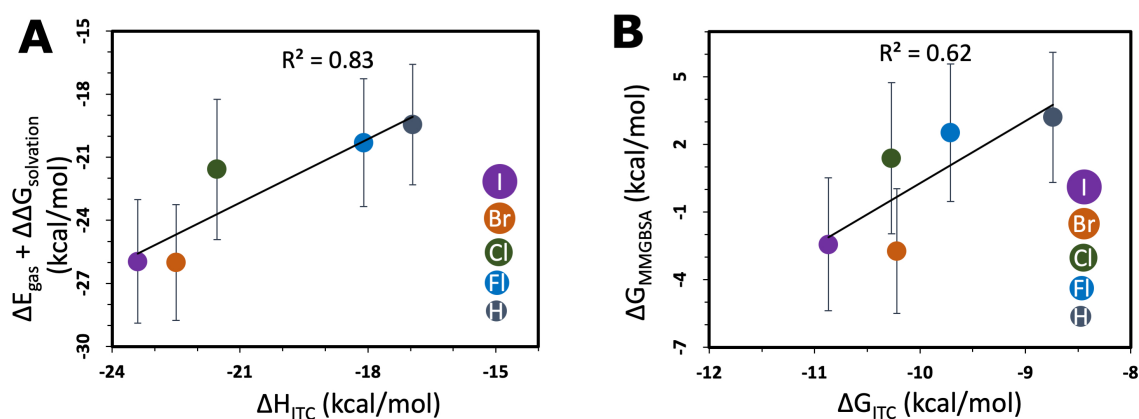


Figure 4.5: Correlation of calculated binding free energies with experimental parameters for the halogen-gatekeeper interaction. A) MMGBSA internal and solvation contributions ($\Delta E_{gas} + \Delta\Delta G_{solvation}$) vs the ITC enthalpies (ΔH_{ITC}) and (B) MMGBSA binding free energies (ΔG_{MMGBSA}) vs the ITC binding free energies (ΔG_{ITC}) of the interactions between haspin and TU derivatives.

Inhibitor	$\Delta E_{gas} + \Delta\Delta G_{solvation}$ (kcal.mol ⁻¹)	$T\Delta S_{MMGBSA}$ (kcal.mol ⁻¹)	ΔG_{MMGBSA} (kcal.mol ⁻¹)
5-iTU	-25.95 ± 2.93	-23.52 ± 0.02	-2.43 ± 2.95
5-brTU	-26.00 ± 2.75	-23.27 ± 0.02	-2.73 ± 2.77
5-clTU	-21.58 ± 3.33	-22.97 ± 0.02	1.39 ± 3.35
5-ftTU	-20.31 ± 3.04	-22.83 ± 0.01	2.52 ± 3.05
tubercidin	-19.45 ± 2.87	-22.65 ± 0.02	3.20 ± 2.89

Table 4.5: Binding free energies calculated using the MMGBSA approach for the binding of tubercidin derivatives with haspin.

4.4.3 Relative residence times from τ -RAMD simulations correlate with the experimentally measured residence times

The ligand egress times of 5-iTU derivatives from the haspin’s binding site were recorded for the set of 20 dissociation trajectories simulated for each inhibitor using τ -RAMD protocol. The RAMD residence time (τ_{comp}) and its standard deviation was computed for each of the 5-iTU derivative using a bootstrapping procedure where bootstrapping was performed for a total 200 sets, and in each of the sets 80% of the trajectories were chosen randomly (see Table 4.6).

Inhibitor	Experimental residence time (SPR) τ_{exp} (s)	Experimental residence time (BLI) τ_{exp} (s)	Computed residence time τ_{comp} (s)
5-iTU	416.67	3623.19	1.62 ± 0.37
5-brTU	84.03	1557.63	1.29 ± 0.45
5-clTU	68.97	1084.60	0.89 ± 0.17
5-ftTU	4.01	158.23	0.95 ± 0.15
tubercidin	1.52	21.10	0.64 ± 0.18

Table 4.6: Experimental residence times (τ_{exp}) and the computed residence times from τ -RAMD (τ_{comp}) for tubercidin derivatives.

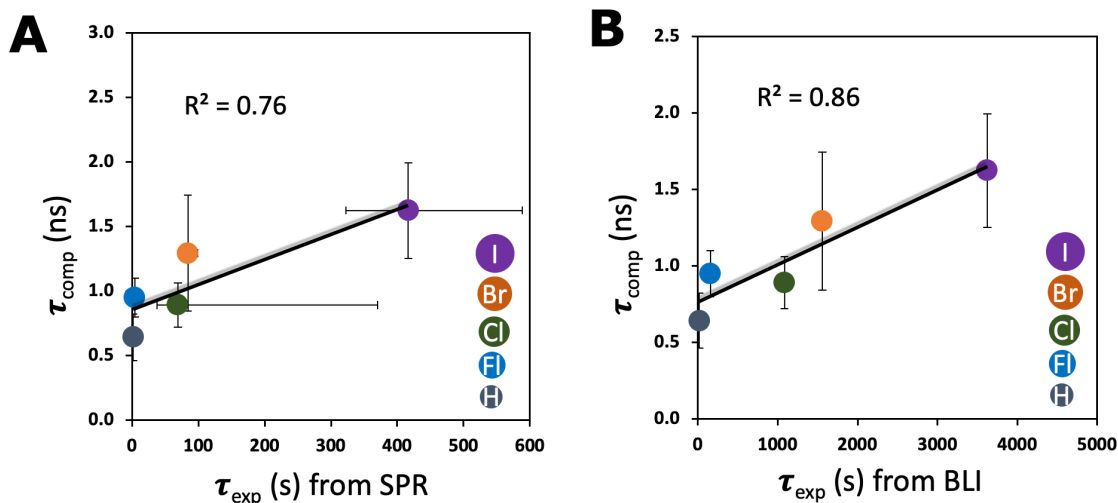


Figure 4.6: Correlation plot of experimental residence times (τ_{exp}) from (A) SPR experimental assay and (B) BLI measurements versus the computed residence time (τ_{comp}) from the τ RAMD procedure as described in Kokh *et al.*[29]. The solid line in both plots corresponds to a linear fit with R^2 values labelled on the plot.

A good correlation was observed between the computed RAMD residence time (τ_{comp}) and experimental residence times (τ_{exp}) of halogenated inhibitors of haspin from SPR and BLI assay (see Figure 4.6 A and B). The computed residence times (τ_{comp}) correlated better with τ_{exp} values from BLI experiments with Pearson correlation coefficient, R^2 of 0.86 compared to R^2 of 0.76 for τ_{exp} from SPR assay. However, difference in residence time between 5-clTU and 5-ftTU is not distinguished; suggesting some effects of halogens may not be captured by the classical MM force field.

In this chapter, the biophysical and structural data presented from Heroven *et al.*[118] on 5-halogen-substituted tubercidin derivatives along with our computational results using quantum chemical interaction energy calculations and MM/GBSA binding free energy calculations, suggest that it could be a viable strategy to increase the residence times of inhibitors by mediating their interactions with aromatic residues of proteins by incorporating heavier halogen atoms into inhibitors. Our results provide a good basis for further research on this topic and it would be interesting to explore the role of presence of halogen atoms to longer residence times of drug candidates especially because many approved drugs have halogen atoms[130] and the aromatic residues are also found frequently in the binding site of proteins, especially kinases.

Chapter 5

Protocol for calculation of diffusional association rates for small molecules using Brownian dynamics

5.1 Overview

The Brownian dynamics (BD) simulation technique is used to simulate the diffusive dynamics of particles, such as proteins, that undergo Brownian motion. It involves use of an implicit solvent model and the stochastic and friction effects of the surrounding solvent are introduced in a separate term in the equation of motion. BD is used to simulate protein-protein association or diffusion of multiple proteins to investigate biomolecular diffusion, binding kinetics, and the effects of macromolecular crowding. The procedure for simulating diffusional association of protein molecules with Brownian dynamics simulations using the SDA software[54, 53, 131] is well established and numerous applications have been reported in the past where SDA was used to compute diffusional association rate constants (k_{on}) for protein-protein association and the computed rates correlated well with the experiments[131, 53, 7, 132, 6, 5, 133, 134]. However, this procedure to compute diffusional k_{on} rates using SDA has not been well optimized for computation of association kinetics for binding of protein and small molecules. Considering the increasing interest in simulating association of drug-like molecules with their target proteins to compute kinetic parameters for binding, we decided to optimize the simu-

lation parameters in SDA and implement a generalized protocol using SDA software that allows calculation of diffusional k_{on} rates for small molecules by running BD simulations of diffusional association of protein and small molecules. We also implemented new algorithms for assigning effective charge sites for small molecules and for the systematic definition of reaction criteria as python scripts. These algorithms differ from the algorithms already implemented for protein-protein association. The implemented protocol was validated for several inhibitors of 4 different targets of varying levels of structural complexity (see the following sections). While for some of these protein-ligand systems, binding is mainly electrostatically driven, for others short-range hydrophobic interactions play a key role in the binding as the inhibitors are hydrophobic in nature.

5.1.1 Trypsin

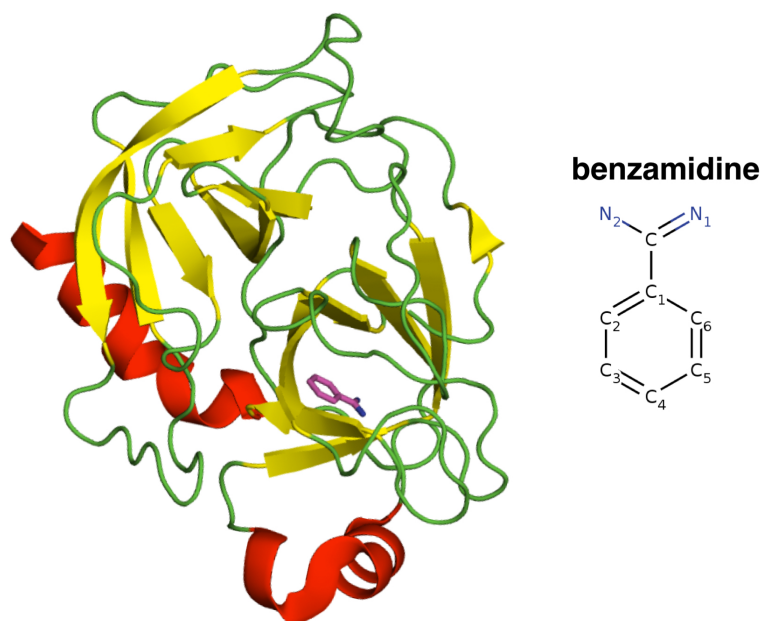


Figure 5.1: 3D crystal structure of trypsin in complex with benzamidine (PDB ID: 3PTB[135]). Trypsin is shown in ribbon representation with helices, β -sheets and loops shown in red, yellow and green, respectively, and the bound benzamidine is shown in magenta stick representation.

Trypsin is a pancreatic serine protease and it hydrolyses proteins by cleaving peptides on the C-terminal side of the amino acid residues, lysine and arginine. Trypsin is characterized by the catalytic triad His57, Asp102 and Ser195. Trypsin is a globular protein composed of 220 residues. The protein is composed of 13 beta-strands, 4

regions of alpha-helix, and six disulfide bridges. The Ca^{2+} -binding loop extends from Glu70 to Glu80. Due to its relatively small size (25 kDa) and monomeric structure, the trypsin–benzamidine complex has proven to be a popular model system for developing and testing methods for computing protein–ligand binding kinetics.

5.1.2 Human Coagulation Factor Xa

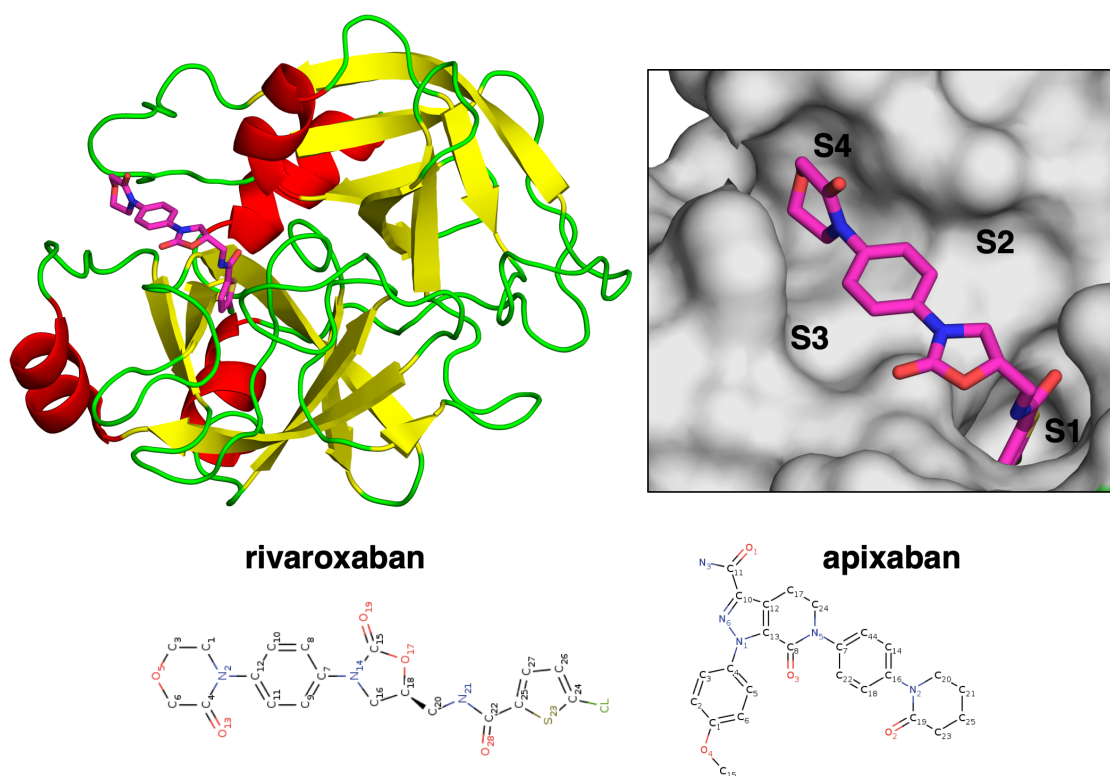


Figure 5.2: 3D crystal structure of the Human Coagulation Factor X activated (FXa) with rivaroxaban bound (PDB ID: 2W26[136]). FXa is shown in ribbon representation with helices, β -sheets and loops shown in red, yellow and green, respectively, and rivaroxaban is shown in magenta stick representation. The inset on the right shows the active site of FXa with rivaroxaban bound and the four subpockets: S1, S2, S3 and S4 are labelled. The 2D structures of the FXa inhibitors studied, rivaroxaban and apixaban, are also shown.

Coagulation Factor X activated (FXa) is a trypsin-like serine endo-peptidase and an important enzyme (EC 3.4.21.6) of the blood coagulation cascade. It is formed by the proteolysis of Factor X (FX). It is composed of two disulfide-linked subunits that catalyze prothrombin to form thrombin (Factor IIa). Thrombin is also a serine protease and it processes fibrinogen to form fibrin (Factor Ia). Fibrin, as the name suggests, is a fibrous, non-globular protein which polymerizes with platelets and results in the formation of a blood clot. Because of its important role in the blood

coagulation cascade, FXa is an important target for the treatment for thromboembolic disorders and a number of selective, direct and indirect fXa inhibitors have been approved for clinical use such as rivaroxaban[137], apixaban[138], betrixaban[139], edoxaban[140] and fondaparinux[141]. Optimizing the k_{on} of FXa inhibitors may be important for the *in vivo* activities as k_{on} for targeting free FXa has been identified to influence the clinical coagulation behavior[142].

The active site of FXa can be subdivided into four sub-pockets: S1, S2, S3 and S4[143](see Figure 5.2). Direct FXa inhibitors such as rivaroxaban and apixaban bind in an L-shaped conformation where one part of the inhibitor occupies the anionic S1 pocket and another part occupies the S4 pocket[136](see Figure 5.2, inset). The S1 sub-pocket is surrounded by residues His57, Asp189, Ser195 and Tyr228, and determines the major component of selectivity[137]. On the other hand, the S4 sub-pocket consists of a narrow hydrophobic channel formed by aromatic residues, such as Tyr99, Phe174 and Trp215. The first generation FXa inhibitors make direct electrostatic interaction of a basic arginine-mimic P1 group with Asp189 at the bottom of the S1 pocket[144, 145]. However, these basic groups are also generally critical for oral bioavailability. A new class of oral fXa inhibitors has been developed that does not require the presence of a basic P1 group and instead makes favourable non-basic interactions with residues in the S1 sub-pocket. Rivaroxaban is one such example which has a chlorothiophene moiety and its chlorine atom interacts with the aromatic ring of Tyr228 at the bottom of the S1 pocket[145]. This chlorine–Tyr228 interaction accounts for high potency and oral bioavailability for the rivaroxaban.

5.1.3 Haspin kinase

Haspin, or germ cell-specific gene 2 protein (GSG2), is an atypical serine/threonine kinase known for its role in cell cycle regulation[146, 147]. Haspin phosphorylates Thr3 on histone H3 during prometaphase, which provides a signal for Aurora B kinase to localize to the centromere of mitotic chromosomes[148]. Aurora B is part of the chromosome passenger complex (CPC) and regulates several steps in mitotic progression. The overall structure of the Haspin kinase corresponds to the conserved eukaryotic protein kinase fold, which consists of two subunits: an amino-terminal lobe (“N lobe”) and a larger carboxy-terminal lobe (“C lobe”)[149, 150]. The "N lobe" is composed of an α -helix (helix α C) and a five-stranded β -sheet, while the C lobe is predominantly helical[149, 150]. Both of these lobes are connected by a

“hinge region” and the ATP binding pocket resides deeply between these lobes.

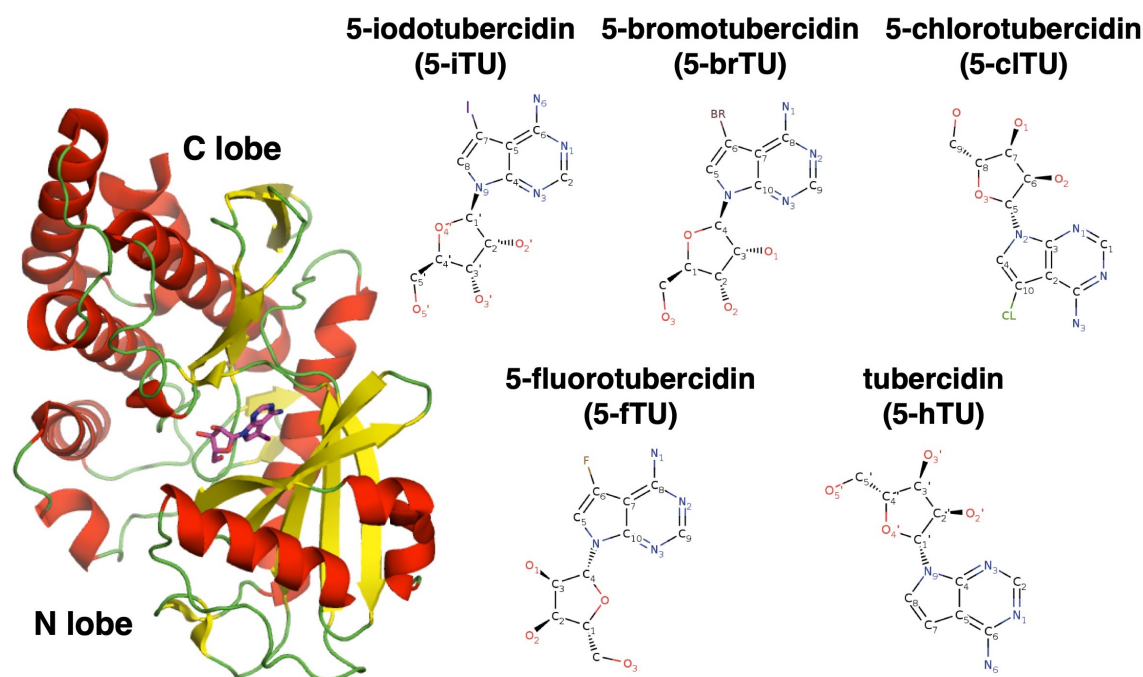


Figure 5.3: 3D crystal structure of the Haspin kinase in complex with 5-iodotubercidin (PDB ID: 6G34[118]). Haspin is shown in ribbon representation with helices, β -sheets and loops shown in red, yellow and green color respectively, and the 5-iodotubercidin (5-iTU) is shown in magenta stick representation. 2D structures of haspin inhibitors studied, 5-iTU, 5-brTU, 5-clTU, 5-ftTU and tubercidin are also shown.

5.1.4 Neuraminidase

Neuraminidases are key targets for drug development against influenza because of their significant role in the release of the virus from an infected cell. Neuraminidase is a homotetramer with circular symmetry and is composed of four identical subunits[151]. The four active sites (one in each subunit) are located in a deep depression on the upper surface. The binding of its natural substrate, sialic acid, to the active site results in the clipping of the glycosidic linkage between sialic acid receptor and sialic acid[152]. The structure of the active site depends upon a proximal four-fold coordinated Ca^{2+} ion and the flexibility of the 150-loop consisting of amino acid residues 148-151[151]. Several subtypes of Neuraminidase (N1, N4, N5, N8) are known to have an open 150-loop structure in the unbound form, that closes upon drug binding. The existence of Neuraminidase in an open 150-loop conformation leads to the presence of an additional cavity also known as the 150-cavity, which is absent in the bound closed form. Other Neuraminidase subtypes (N2, N3, N6, N7,

N9) have been shown to exist in a closed conformation in both bound and unbound forms, without a 150-cavity[153].

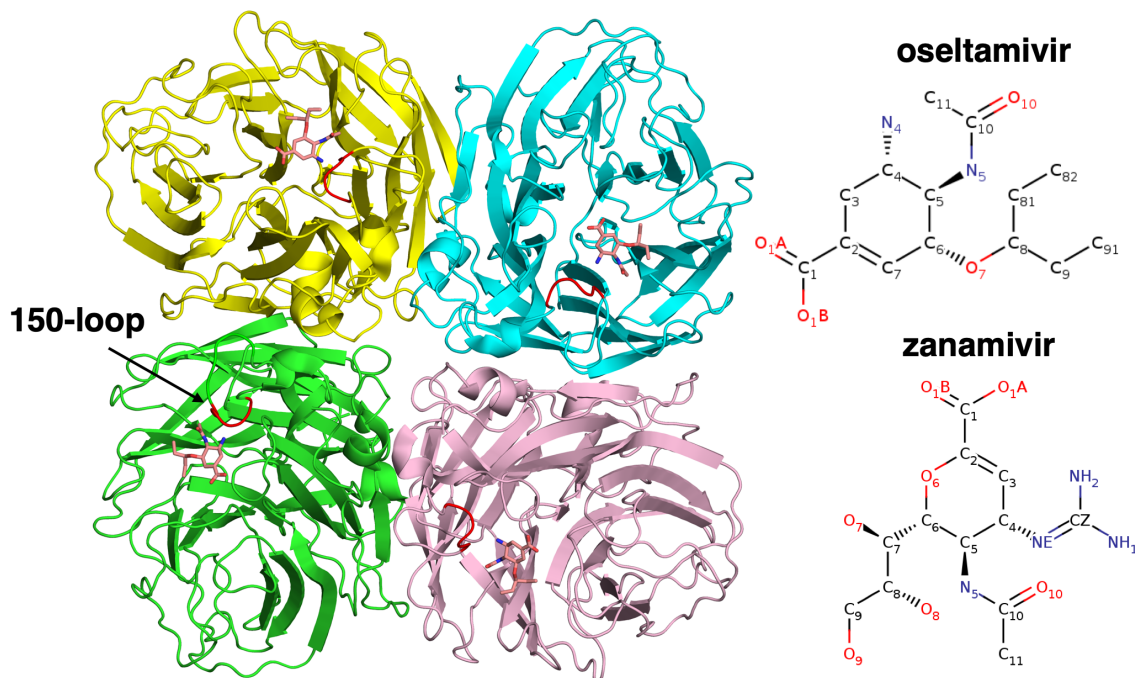


Figure 5.4: Tetrameric crystal structure of the N1 Neuraminidase (NA) complexed with oseltamivir (PDB ID: 2HU4[153]). NA is shown in ribbon representation with different monomeric units shown with different colors and oseltamivir (orange stick representation) bound to each monomer. The 150-loop is shown in red color. The 2D structures of NA inhibitors studied, oseltamivir and zanamivir, are also shown.

Currently, oseltamivir and zanamivir are the two FDA-approved drugs used to battle influenza. They are transition-state analogues of the natural substrate, sialic acid, and therefore, bind more tightly to the active site than the substrate. Both oseltamivir and zanamivir are zwitterionic in nature and they comprise a central ring structure with a carboxyl and an amino moiety.

5.2 Methods

5.2.1 Preparation of protein and ligand structures

The 3D coordinates of all the protein and ligand molecules studied were extracted from the co-crystallized protein-ligand complexes available in the PDB database. Table 5.1 lists the PDB IDs used for extracting the coordinates of the different

protein-ligand systems studied. N1 Neuraminidase (NA) has a tetrameric structure and each monomeric unit has a Ca^{2+} ion situated close to the active site and it is tetrahedrally coordinated by backbone as well as side chain atoms. This Ca^{2+} ion seemed to be important for structural stability around the active site of NA. Since the coordinates for Ca^{2+} ions were missing in the PDB structures of the NA-inhibitor complexes used in our study, they were taken from a high-resolution structure of NA in complex with oseltamivir (PDB ID: 3TI6[154]) and modelled into all NA-inhibitor crystal structures. The *Protein preparation wizard*[104] of the Schrödinger suite (release 2015-4, Schrödinger, LLC, New York) was used for pre-processing of the protein structures e.g. to add disulfide linkages and missing atom names, remove crystallographic waters and for capping polypeptide chain termini. Any missing residues or empty loops in the protein were modelled using the *Prime* program[155] of Schrödinger. Then, the protein and ligand structures were protonated at pH 7.0 according to residue pKa values calculated by *PROPKA*[105]. The hydrogen-bonding network of the protein-ligand complexes was optimized with the Schrödinger suite to avoid any steric clashes and ensure favorable atom orientations. Then the complexes were energy-minimized with a default energy minimization procedure using the *Impref*[104] program of Schrödinger and OPLS3 force field. The *Impref* minimization involves a two-step relaxation in which first the rotatable hydrogen atoms are minimized with all the torsional potentials removed, and then an all-atom minimization is performed that is terminated either when the system is fully converged or when it reaches a heavy-atom RMSD from the initial structure of 0.30 Å. The coordinates of the protein and ligand molecules were then extracted from the minimized complexes into separate PDB files.

Protein	Ligand	PDB ID	Resolution (Å)
Trypsin	benzamidine	3PTB	1.70
Coagulation Factor Xa	rivaroxaban	2W26	2.08
Coagulation Factor Xa	apixaban	2P16	2.30
Haspin kinase	5-iTU	6G34	1.76
Haspin kinase	5-brTU	6G35	1.55
Haspin kinase	5-clTU	6G36	1.46
Haspin kinase	5-ftTU	6G37	1.48

Protein	Ligand	PDB ID	Resolution (Å)
Haspin kinase	5-hTU	6G38	1.47
Neuraminidase	oseltamivir	2HTY (Neuraminidase)	2.50
		2HU4 (oseltamivir)	2.50
Neuraminidase	zanamivir	2HTY (Neuraminidase),	2.50
		2HTQ (zanamivir)	2.20

Table 5.1: List of PDB structures of different protein-ligand complexes used for computation of diffusional association rates using SDA.

5.2.2 Preparation of PQR files

The *PDB2PQR* web server[156] (http://nbc-222.ucsd.edu/pdb2pqr_2.1.1/) was used to generate PQR files for protein structures for being used later in continuum electrostatics calculations. PQR files are PDB files where the occupancy and B-factor columns have been replaced by per-atom charge and radius. The AMBER force field was used to generate PQR files with the PDB2PQR server and the protonation states determined from *PROPKA* in the previous step were retained. The *PDB2PQR webserver* allows the setup, execution, and analysis of Poisson-Boltzmann electrostatics calculations. However PDB2PQR cannot perform calculations for bound ligand molecules and Ca^{2+} ions. Therefore, AmberTools[79] was used to generate PQR file for ligands. Partial atomic charges for ligand molecules were generated using the *RESP*[106] program of AmberTools by fitting atom-centered charges to the electrostatic potential computed by the *GAMESS*[76] program with 6-31+G basis sets. The force field parameters for ligand molecules (inpcrd-and prmtop-files) were prepared with *LEaP* program of AmberTools with the GAFF[124] force field, and these parameters were then used by *ambpdb* program of AmberTools to generate PQR files. Coordinates, charges (+2e) and radii (1.713 Å) for Ca^{2+} ions were manually added into PQR files of corresponding proteins.

5.2.3 Grids preparation

To calculate interaction forces acting between protein and ligand molecules during the BD simulations of diffusional association, charges and radii from PQR files were

used to generate grids of the electrostatic potential. Electrostatic potentials were calculated by solving the linearized Poisson-Boltzmann equation using the *APBS*[57] software. *APBS* calculates the electrostatic potential of the solute with respect to its environment for every grid point of a 3D system. In addition to the PQR file, *APBS* requires an input file with information on grid dimensions and grid spacing. The grid sizes for different systems were chosen so as to ensure electrostatic potentials with isovalues of ± 0.01 kcal/mol/e fit into the boxes. A grid size of 129x129x129 points was used for Trypsin, Coagulation Factor Xa and Haspin kinase whereas a larger grid size of 161x161x161 points was used for Neuraminidase due to its relatively bigger size. For all the ligands, a grid size of 65x65x65 points was used. A grid spacing of 1.0 Å was consistently used for all protein and ligand molecules. The temperature and ionic strength used for the grid calculations were specific to the system studied, depending upon the experimental conditions in which rates were measured (see Table 5.2). The solvent dielectric constant, protein dielectric constant, and the ionic radius were set to 78, 2, and 1.5 Å, respectively. “Single Debye-Hückel” boundary condition (*befl sdh*) was used where the potential at the boundary is set to the values prescribed by a Debye-Hückel model for a single sphere with a point charge, dipole, and quadrupole. For mapping the point charges to the grid for a Poisson-Boltzmann calculation, traditional trilinear interpolation (*chgm spl0*) was used where each charge is mapped onto the nearest-neighbor grid points. The dielectric and ion-accessibility coefficients (*srfm* parameter) were defined using the *smol* flag where the dielectric coefficient is defined based on a molecular surface definition, and the ion-accessibility coefficient is defined by an inflated van der Waals model. The electrostatic potential grids calculated by *APBS* are in units of kT/e and cannot be directly used with SDA. Therefore, they were rescaled and converted into UHBD format (kcal/mol.e) using the *convert_grid* program of the SDA package. To take the desolvation forces into account, electrostatic and hydrophobic desolvation grids for the protein and ligand molecules were calculated using the *make-edhdlj-grid* tool of SDA. A grid size of 110x110x110 points with a 1.0 Å spacing was used and ionic strength, solvent dielectric constant and ion radius were set to 0 mM, 78 and 1.5 Å, respectively. The value of empirical scaling parameter, α , for electrostatic desolvation grids was set to 0.36 in the SDA input file. The parameterisation is based on the work by Gabdoulline and Wade[55] to reproduce the Poisson-Boltzmann interaction energy of plastocyanin and cytochrome f. When α is

set to 0.36, electrostatic desolvation potential is calculated assuming that no salt is present. Therefore, ionic strength was set to 0 in the input file used for calculation of electrostatic desolvation grids.

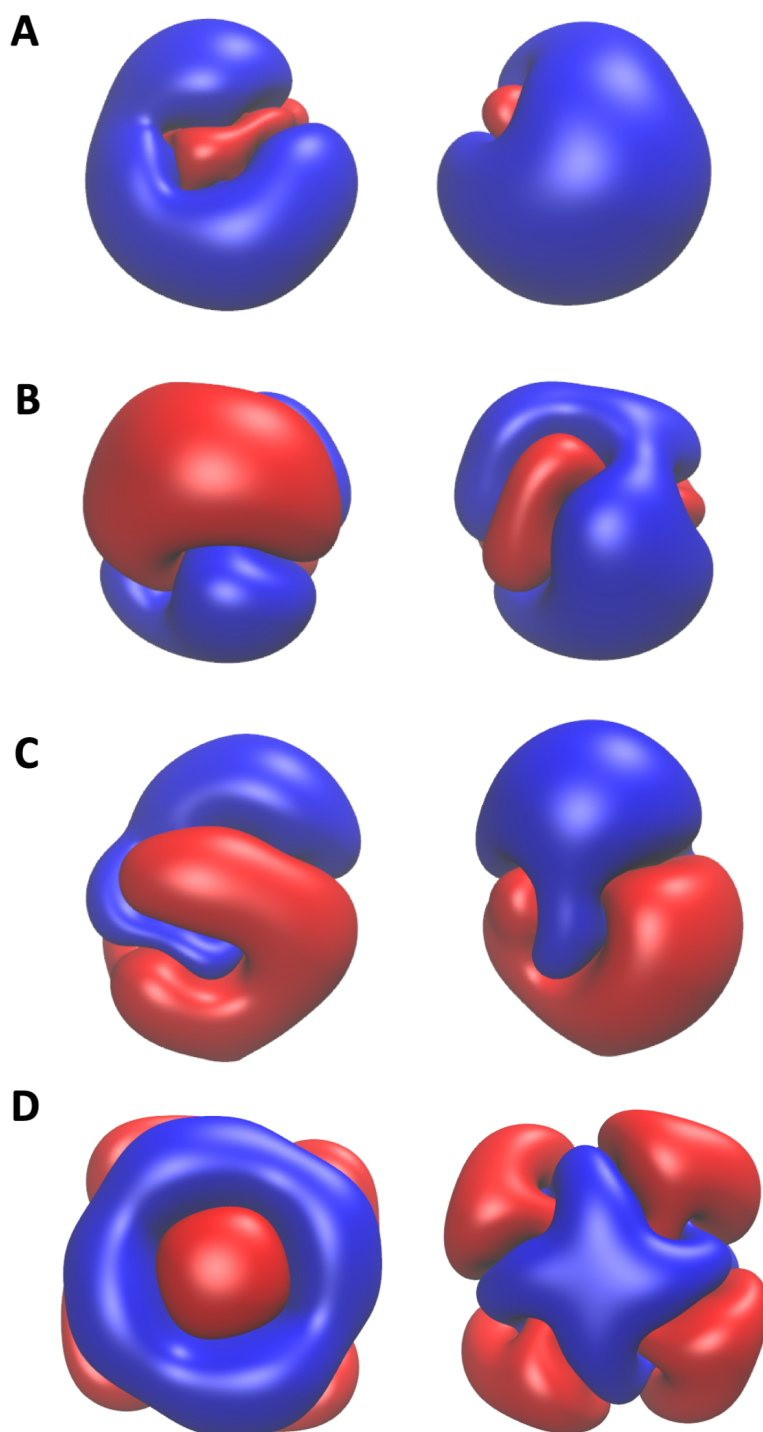


Figure 5.5: Electrostatic Potential Grids of A) Trypsin, B) Human Coagulation Factor Xa, C) Haspin and D) Neuraminidase. For each system studied, electrostatic potential from the front and the back are shown. Isosurfaces shown in the figure correspond to the isovalue of ± 0.01 kcal/mol/e units. The grid sizes chosen for different systems are given in Table 5.2.

Protein-ligand system simulated	Ionic Strength mM	Temperature °C	Electrostatics grid size for protein (\AA^3)	Reference for experimental conditions
Trypsin-benzamidine	100	25	$129 \times 129 \times 129$	[157]
FXa-rivaroxaban	150	37	$129 \times 129 \times 129$	[158]
FXa-apixaban	150	37	$129 \times 129 \times 129$	[158]
Haspin-5-iTU	150	25	$129 \times 129 \times 129$	[118]
Haspin-5-brTU	150	25	$129 \times 129 \times 129$	[118]
Haspin-5-clTU	150	25	$129 \times 129 \times 129$	[118]
Haspin-5-ftTU	150	25	$129 \times 129 \times 129$	[118]
Haspin-5-hTU	150	25	$129 \times 129 \times 129$	[118]
Neuraminidase-oseltamivir	100	25	$161 \times 161 \times 161$	[159]
Neuraminidase-zanamivir	100	25	$161 \times 161 \times 161$	[159]

Table 5.2: Different experimental conditions (Ionic strength and Temperature) used for preparation of electrostatic grids for different protein-ligand complexes simulated with SDA for computation of diffusional association rates.

5.2.4 Effective charges for protein and ligands

Effective Charges for Macromolecules in solvent (ECM) are fitted charges in a uniform dielectric that can reproduce the electrostatic potential of the molecule computed with the use of all partial atomic charges in a heterogeneous dielectric. The accurate evaluation of electrostatic forces and interaction free energies for protein-ligand association is computationally very demanding for realistic systems with thousands of atomic charges in an environment with a non-uniform dielectric permittivity and a solvent of non-zero ionic strength. Therefore, a small number of effective charges are calculated for each molecule that reproduce the intermolecular electrostatic interactions with high accuracy in a uniform dielectric[160]. Determining the effective charge sites is relatively simple for proteins where the test charges are placed on the carboxylate oxygens of Asp, Glu, and the C-terminus, and the amine

nitrogens of Lys, Arg, and the N-terminus. However, this approach does not work for chemical compounds, cofactors etc. Therefore, a python script was written to pick effective charge sites for small molecules and assign appropriate test charges to them (see Figure 5.6 for the protocol).

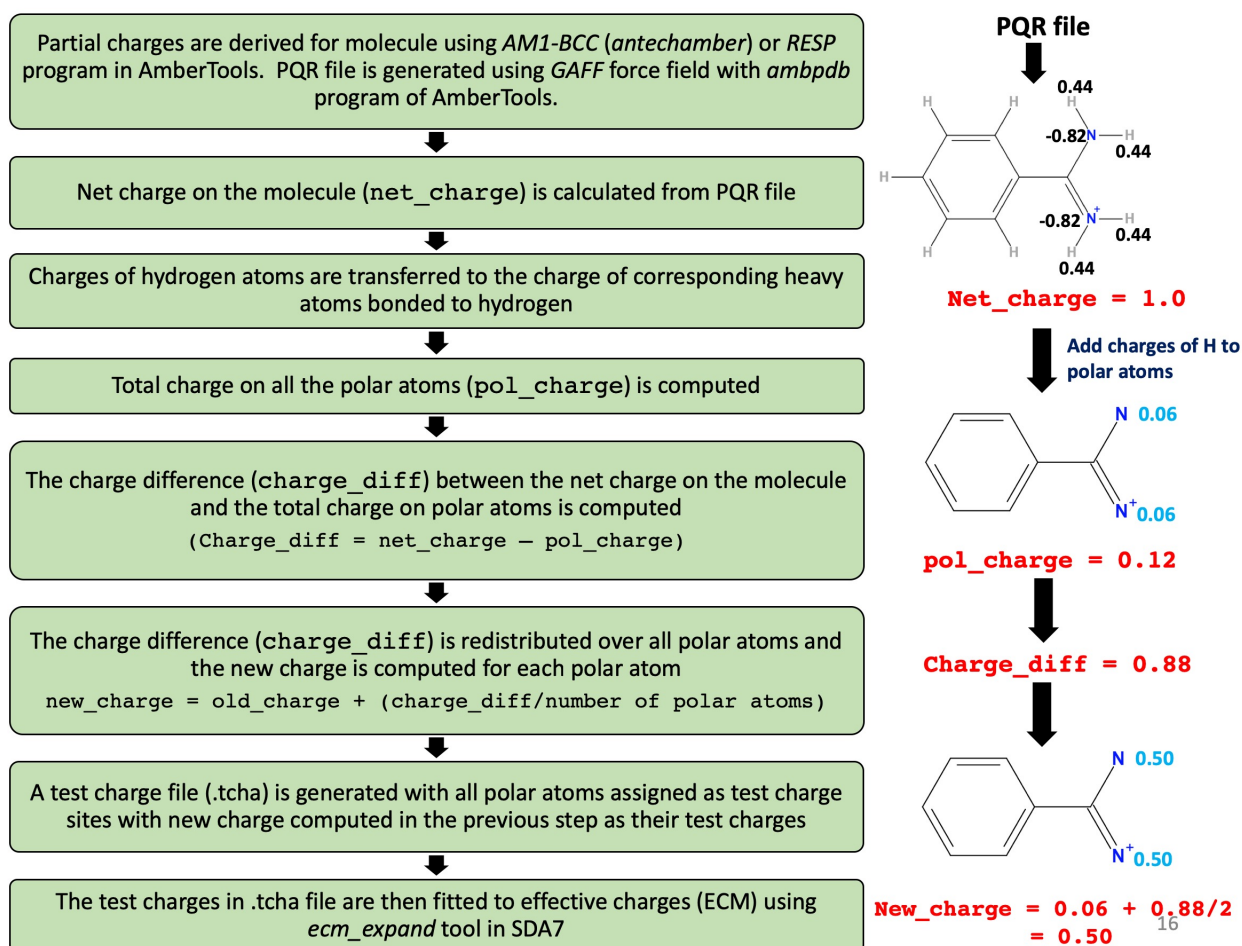


Figure 5.6: Algorithm to assign test charges for computing effective charges for small molecules.

The script reads in the PQR file and determines the net charge on the molecule. Then, the N, O, S, F, Cl, Br, I, P and Fe atoms present in drug like compounds and organic cofactors are marked as effective charge sites. The partial atomic charges of hydrogen atoms covalently bonded to these effective charge sites are added to the respective partial charges of the effective charge site atoms. Then, the charge difference between the net charge of the molecule and the cumulative sum of the partial charges of the effective charge sites is calculated. This charge difference is then redistributed equally to all the effective charge sites, so that the total test charge on all effective charge sites is equal to the net charge of molecule. These updated test charges are then written to an output file in a .tcha format, which is similar to pdb

format with occupancy column replaced by the test charge of atoms. This output .tcha file is then used by the *ecm_expand* tool of SDA to generate ECM charges. The molecular structures of inhibitors of Human Coagulation Factor Xa, Haspin kinase and N1 Neuraminidase are given in Figures 5.8, 5.9 and 5.10, respectively. Tables 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11 and 5.12 list the test charges and effective charges assigned to the polar atoms of benzamidine, rivaroxaban, apixaban, 5-iTU, 5-brTU, 5-clTU, 5-ftTU, 5-hTU, oseltamivir and zanamivir, respectively.

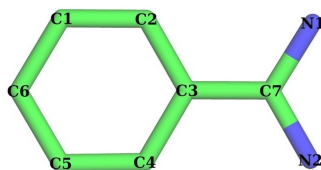


Figure 5.7: Atomic structure of benzamidine. Benzamidine has a net charge of +1e. Test and effective charges for benzamidine are given in Table 5.3.

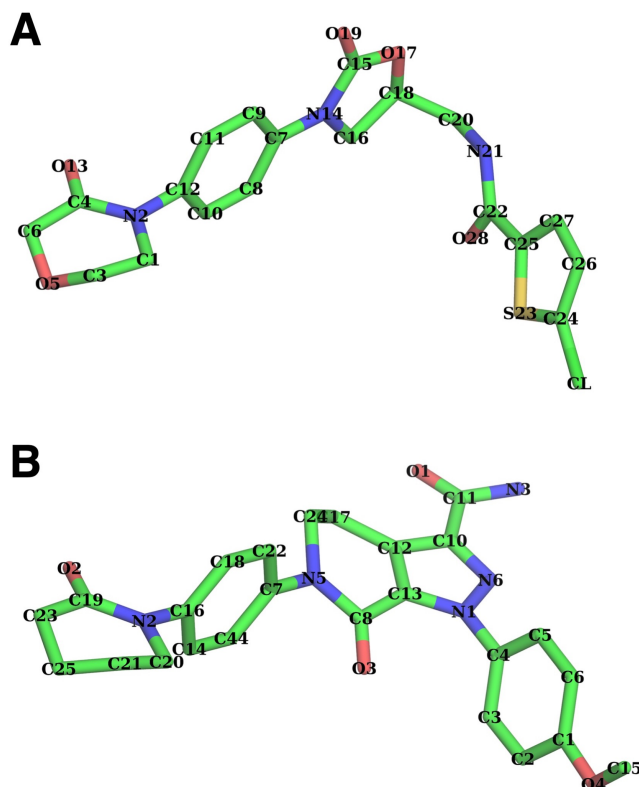


Figure 5.8: Atomic structures of A) rivaroxaban and B) apixaban. Both rivaroxaban and apixaban are neutral. Test and effective charges for rivaroxaban and apixaban are given in tables 5.4 and 5.5, respectively.

Atom Name	Test charge (e)	Effective charge (e)
N1	0.50	0.524
N2	0.50	0.534
Net effective charge (e)		1.058
Electric dipole moment (Debye)	0.02 D	0.05 D (4.84 D)
Solvent accessible surface area (Å²)	311.15	

Table 5.3: Test charges and effective charges assigned to polar atoms of benzamidine. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges. The net electric dipole moment of the molecule calculated using RESP atomic charges is given in parenthesis.

Atom Name	Test charge (e)	Effective charge (e)
CL	0.242	0.128
N2	0.217	0.865
O5	-0.130	-0.118
O13	-0.297	-0.606
N14	0.197	0.281
O17	-0.091	-0.196
O19	-0.279	-0.455
N21	0.033	0.901
S23	0.346	-0.401
O28	-0.239	-0.377
Net effective charge (e)		0.021
Electric dipole moment (Debye)	22.21 D	8.74 D (4.11 D)
Solvent accessible surface area (Å²)	997.46	

Table 5.4: Test charges and effective charges assigned to polar atoms of rivaroxaban. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of rivaroxaban shown in Figure 5.8 A. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges. The net electric dipole moment of the molecule calculated using RESP atomic charges is given in parenthesis.

Atom Name	Test charge (e)	Effective charge (e)
N1	0.494	0.445
O1	-0.329	-0.611
N2	0.209	0.827
O2	-0.291	-0.735
N3	0.186	0.414
O3	-0.245	-0.966
O4	-0.072	-0.009
N5	0.189	0.785
N6	-0.143	-0.143
Net effective charge (e)		0.006
Electric dipole moment (Debye)	5.41 D	9.98 D (7.19 D)
Solvent accessible surface area (\AA^2)	1171.93	

Table 5.5: Test charges and effective charges assigned to polar atoms of apixaban. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of apixaban shown in Figure 5.8 B. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges.

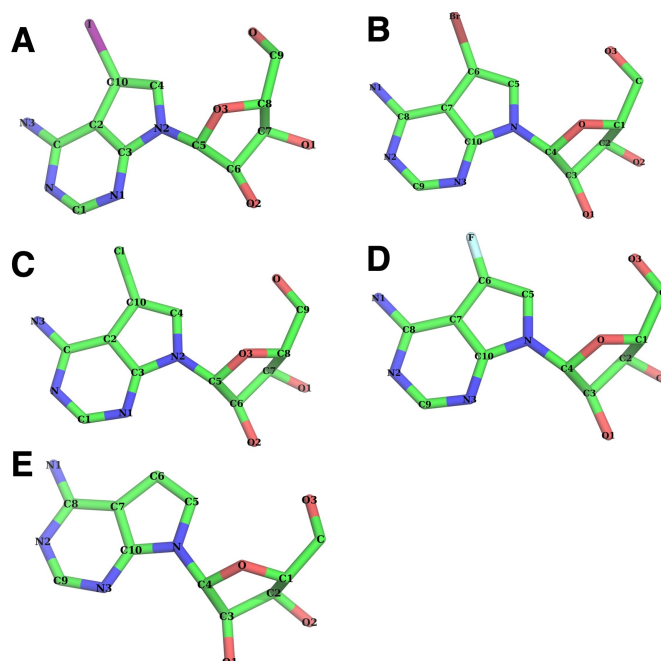


Figure 5.9: Atomic structures of A) 5-iTU, B) 5-brTU, C) 5-clTU, D) 5-ftTU, and E) 5-hTU. All 5-iTU derivatives are neutral. Test and effective charges for 5-iTU derivatives are given in tables 5.6, 5.7, 5.8, 5.9 and 5.10, respectively.

Atom Name	Test charge (e)	Effective charge (e)
N	-0.450	-0.171
O	0.092	0.481
N1	-0.399	-0.715
O1	0.041	0.008
N2	0.267	-0.059
O2	0.099	0.163
N3	0.164	0.608
O3	-0.100	0.131
I	0.285	-0.474
Net effective charge (e)		0.028
Electric dipole moment (Debye)	13.77 D	14.61 D (7.54 D)
Solvent accessible surface area (\AA^2)	679.64	

Table 5.6: Test charges and effective charges assigned to polar atoms of 5-iTU. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of 5-iTU shown in Figure 5.9 A. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges.

Atom Name	Test charge (e)	Effective charge (e)
N	0.254	0.060
O	-0.095	-0.237
N1	0.177	0.306
O1	0.105	-0.211
N2	-0.446	-0.040
O2	0.051	0.330
N3	-0.402	-0.451
O3	0.102	0.518
Br	0.254	-0.291
Net effective charge (e)		-0.016
Electric dipole moment (Debye)	13.30 D	12.48 D (7.83 D)
Solvent accessible surface area (\AA^2)	674.31	

Table 5.7: Test charges and effective charges assigned to polar atoms of 5-brTU. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of 5-brTU shown in Figure 5.9 B. The net electric dipole moment of the molecule calculated using RESP atomic charges is given in parenthesis.

Atom Name	Test charge (e)	Effective charge (e)
N	-0.449	0.305
O	0.103	0.442
N1	-0.397	-0.859
O1	0.056	0.145
N2	0.251	-0.345
O2	0.111	0.091
N3	0.187	0.236
O3	-0.088	0.207
Cl	0.225	-0.250
Net effective charge (e)		-0.027
Electric dipole moment (Debye)	13.19 D	15.14 D (7.77 D)
Solvent accessible surface area (Å²)	669.62	

Table 5.8: Test charges and effective charges assigned to polar atoms of 5-clTU. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of 5-clTU shown in Figure 5.9 C. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges. The net electric dipole moment of the molecule calculated using RESP atomic charges is given in parenthesis.

Atom Name	Test charge (e)	Effective charge (e)
N	0.238	-0.257
O	-0.078	0.185
N1	0.206	0.309
O1	0.105	0.042
N2	-0.428	0.226
O2	0.077	0.180
N3	-0.386	-0.825
O3	0.111	0.433
F	0.158	-0.316
Net effective charge (e)		-0.024
Electric dipole moment (Debye)	12.14 D	14.86 D (7.57 D)
Solvent accessible surface area (Å²)	658.29	

Table 5.9: Test charges and effective charges assigned to polar atoms of 5-ftTU. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of 5-ftTU shown in Figure 5.9 D.

Atom Name	Test charge (e)	Effective charge (e)
O3	0.173	0.305
O	-0.032	0.655
O1	0.180	0.579
O2	0.076	-0.432
N	0.327	-0.658
N3	-0.478	-0.980
N2	-0.503	0.268
N1	0.257	0.239
Net effective charge (e)		-0.024
Electric dipole moment (Debye)	15.58 D	12.01 D (7.12 D)
Solvent accessible surface area (Å ²)	631.48	

Table 5.10: Test charges and effective charges assigned to polar atoms of 5-hTU. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of 5-hTU shown in Figure 5.9 E. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges. The net electric dipole moment of the molecule calculated using RESP atomic charges is given in parenthesis.

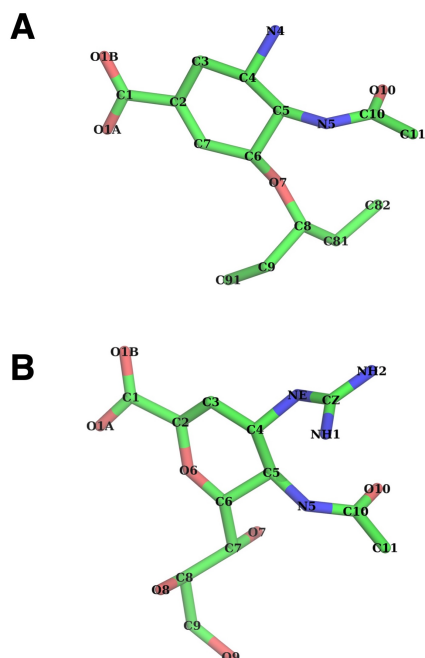


Figure 5.10: Atomic structures of A) oseltamivir and B) zanamivir. Both oseltamivir and zanamivir are neutral and zwitter-ionic. Test and effective charges for oseltamivir and zanamivir are given in tables 5.11 and 5.12, respectively.

Atom Name	Test charge (e)	Effective charge (e)
O1A	-0.408	-0.880
O1B	-0.425	-0.571
N4	0.858	1.331
N5	0.180	0.844
O7	0.041	0.378
O10	-0.246	-1.086
Net effective charge (e)		0.017
Electric dipole moment (Debye)	20.73 D	24.50 D (22.24 D)
Solvent accessible surface area (\AA^2)	675.42	

Table 5.11: Test charges and effective charges assigned to polar atoms of oseltamivir. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of oseltamivir shown in Figure 5.10 A. The table also lists the solvent accessible surface area of the molecule, and the electric dipole moment of the molecule calculated after assignment of test charges and effective charges. The net electric dipole moment of the molecule calculated using RESP atomic charges is given in parenthesis.

Atom Name	Test charge (e)	Effective charge (e)
NE	0.272	0.492
NH1	0.304	0.498
NH2	0.285	0.270
O1A	-0.464	-0.593
O1B	-0.477	-0.674
N5	0.207	0.303
O6	0.100	0.197
O7	0.078	-0.188
O8	0.005	0.427
O9	0.037	-0.136
O10	-0.348	-0.572
Net effective charge (e)		0.024
Electric dipole moment (Debye)	25.82 D	21.11 D (25.20 D)
Solvent accessible surface area (\AA^2)	741.86	

Table 5.12: Test charges and effective charges assigned to polar atoms of zanamivir. The positions of these atoms corresponding to effective charge sites can be visualized on the atomic structure of zanamivir shown in Figure 5.10 B.

5.2.5 Calculation of diffusion coefficients

During BD simulations using SDA software, diffusional motion of a mobile solute is modelled according to the translational and rotational Ermak-McCammon equations. Since one of the solutes (the protein) is kept fixed, SDA uses a relative translational diffusion constant D to account for diffusion of both protein and ligand. D is calculated as the sum of the translational diffusion constant of the protein (D_{prot}) and the translational diffusion constant of the ligand (D_{lig}). In our BD simulations, translational and rotational diffusion coefficients for the protein and the ligand were calculated using the HYDROPRO[82] software. A partial specific volume of $0.73 \text{ cm}^3\text{mol}^{-1}$, solvent density of 1.0 g/cm^3 , and solvent viscosity of 0.0091 poises was used in the HYDROPRO input. The calculation mode (*INDMODE* parameter) was set to 1 which corresponds to the atomic-level primary model and shell-based methodology with up to 2000 minibeads. For proteins, the hydrodynamic radius (*AER*) was set to 2.9 \AA which is the recommended value to be used for proteins when *INDMODE* is set to 1. The lowest (*SIGMIN*) and the highest (*SIGMAX*) value for sigma, the minibead radius, was set to 1.0 \AA and 2.0 \AA , respectively. The *NSIG* parameter was set to 6 which corresponds to the number of values of the radius of the minibead. However, for ligands, a smaller value of the hydrodynamic radius ($AER = 1.2 \text{ \AA}$) was used. This value was optimized by Dr. Ariane Nunes-Alves, based on control calculations on several small chemical compounds to reproduce their experimental diffusion coefficients (data unpublished). Also, for ligands, the *NSIG* parameter was set to -1 where the program estimates the two extreme values of sigma, and therefore, there is no need to define *SIGMIN* and *SIGMAX* values. The diffusion coefficients of the different protein and ligand molecules simulated in this study are given in table 5.13.

Molecule Name	Translational diffusion	Rotational diffusion
	coefficient $\text{\AA}^2/\text{ps}$	coefficient $\text{radian}^2/\text{ps}$
Trypsin	0.010550	0.000015
Coagulation Factor Xa	0.010460	0.000014
Haspin kinase	0.008663	0.000008
Neuraminidase	0.005707	0.000002
benzamidine	0.092250	0.009354

Molecule Name	Translational diffusion	Rotational diffusion
	coefficient	coefficient
	$\text{\AA}^2/\text{ps}$	radian ² /ps
rivaroxaban	0.052000	0.001482
apixaban	0.048800	0.001173
5-iTU	0.063310	0.002976
5-brTU	0.063830	0.003063
5-clTU	0.063230	0.003080
5-ftTU	0.064360	0.003165
5-hTU	0.065460	0.003408
oseltamivir	0.063450	0.002729
zanamivir	0.060030	0.002330

Table 5.13: Translational and rotational diffusion coefficients of different protein and ligand molecules calculated using HYDROPRO.

5.2.6 Generation of Reaction Criteria

For docking and association rate calculations with SDA, the user needs to specify a set of reaction criteria that define the formation of an encounter complex. Usually, donor-acceptor atom pairs in the bound protein-ligand complexes are considered as reaction criteria but other types of interacting atom pairs can also be included. To generate reaction criteria for two interacting solutes automatically, we wrote a python script *ReactionCriteria.py* that generates reaction criteria by taking into account the hydrogen bonding interactions, halogen- π interactions, and the π - π interactions in the bound protein-ligand complex. It requires a total of three mandatory parameters as input: 1) PDB file for protein, 2) PDB file for ligand, and 3) the reaction distance (in \AA). The MOL2 file for the ligand can also be provided as an optional fourth argument, if π - π interactions between the protein and ligand also need to be considered for generating reaction criteria. In our association rate calculations, we only considered possible donor-acceptor pairs (within 3.5 \AA) and halogen- π interactions (within 4.5 \AA) between protein and ligand molecules for generating reaction criteria. Tables 5.14, 5.15, 5.16 and 5.17 lists the set of pairs of

atoms used as reaction criteria for performing association rate calculations for the different protein-ligand systems simulated.

Trypsin	benzamidine
D171 OD1	N2
D171 OD1	N1
D171 OD2	N1
S172 OG	N1
S172 O	N2
S172 O	N1
G196 O	N2

Table 5.14: Reaction Criteria for SDA association rate constant calculations for benzamidine binding to Trypsin.

FXa-rivaroxaban		FXa-apixaban	
FXa	rivaroxaban	FXa	apixaban
T98 N	O5	R143 NH1	N3
Y99 N	O5	E146 O	N3
G219 N	O17	C191 SG	N3
G219 N	O19	Q192 N	N6
G219 O	N21	G216 N	O3
Y228 CD1	CL	C220 SG	N3

Table 5.15: Reaction Criteria for SDA association rate constant calculations for inhibitors of Human Coagulation Factor Xa (FXa).

Haspin-5-iTU		Haspin-5-brTU		Haspin-5-clTU		Haspin-5-ftTU		Haspin-5-hTU	
Haspin	5-iTU	Haspin	5-brTU	Haspin	5-clTU	Haspin	5-ftTU	Haspin	5-hTU
F605 CG	I	F605 CG	Br	F605 CG	CL	F605 CG	F	F605 CG	H10
E606 O	N3	E606 O	N1	E606 O	N3	E606 O	N1	E606 O	N1
G608 N	N	G608 N	N2	G608 N	N	G608 N	N2	G608 N	N2
D611 OD2	O1	D611 OD2	O1	D611 OD2	O1	D611 OD2	O1	D611 OD2	O1
D611 OD2	O2	D611 OD2	O2	D611 OD2	O2	D611 OD2	O2	D611 OD2	O2
G653 O	O1	G653 O	O2	G653 O	O1	G653 O	O2	G653 O	O2

Table 5.16: Reaction Criteria for SDA association rate constant calculations for inhibitors of Haspin kinase.

oseltamivir-Neuraminidase		zanamivir-Neuraminidase	
oseltamivir	Neuraminidase	zanamivir	Neuraminidase
N4	E119 OE2	O1B	R118 NH2
N4	D151 OD1	NH2	E119 OE1
O10	R152 NH2	NE	E119 OE2
O1A	R292 NH1	NH2	D151 O
O1A	R292 NH2	NE	D151 OD1
O1A	Y347 OH	O10	R152 NH2
O1A	R371 NH1	NH2	R156 NH1
O1B	R371 NH1	NH1	W178 O
		NH2	W178 O
		O9	R224 NE
		O9	E276 OE1
		O8	E276 OE2
		O1A	R292 NH1
		O8	R292 NH1
		O1A	R292 NH2
		O8	R292 NH2
		O1A	Y347 OH
		O1A	R371 NH1
		O1B	R371 NH2
		O1A	Y406 OH
		O1B	Y406 OH
		O6	Y406 OH

Table 5.17: Reaction Criteria for SDA association rate constant calculations for inhibitors of N1 Neuraminidase (NA).

5.2.7 Association rate calculation with SDA

Simulation setup

Using trypsin-benzamidine as a model system, some of the parameters were optimized to simulate protein-ligand association with SDA. For protein-ligand association, a smaller *probep* radius of 1.40 Å (representative of hydrogen bond distance) was used than the relatively larger radius of 1.7-2.0 Å used for protein-protein association in previous studies. The value of the radius used for protein-protein association

studies is chosen to represent the atomic radius of surface (non hydrogen) atoms and therefore the value of this radius depends on the type of force field used, for example, a smaller value (0.5 Å) should be used with the ProMetCS force field as it includes a Lennard-Jones term. In the simulation setup, the protein molecules were kept fixed at the center of a spherical system and the BD trajectories of ligand molecules were started from a large center-to-center distance (of mass), *b surface* of 100 Å at which the centrosymmetric forces acting between protein and ligand molecules are assumed to be negligible (iso-potential values of electrostatic grids were lower than ± 0.01 kcal/mol units). A trajectory was stopped when the ligand left the outer *c surface* = 300 Å. An encounter complex was considered to be formed when protein and ligand satisfied 1 to 5 independent reaction contacts (*nb-contacts* parameter in SDA input file), with the minimum distance between independent contacts (*dind*) being 3 Å. Due to smaller size of the ligands and to ensure presence of at least 2 independent contacts, a smaller *dind* value of 3 Å was chosen for ligands compared to 6 Å used for proteins. Due to the small size of the benzamidine, the distance criteria for independent contacts (*dind*) was set to 2 Å to have at least 2 independent contacts because having *dind*=3 Å resulted in only 1 independent contact.

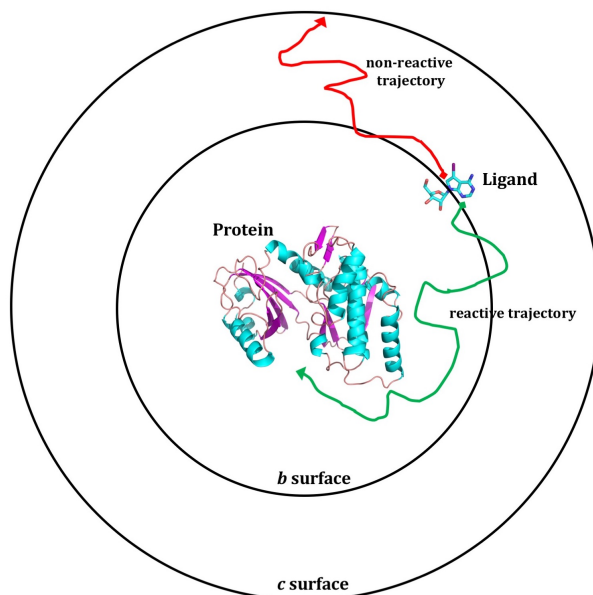


Figure 5.11: Schematic representation of the geometric setup for the diffusional association of protein and ligand molecules in BD simulations.

Analysis of SDA output

Diffusional association rate constants were computed for encounter complexes satisfying one to five reaction contacts at different reaction distances starting from 3.0 Å to 20.0 Å, with each reaction window separated by 0.5 Å. For each system, two sets of simulations were run: with and without taking into account hydrophobic desolvation (HD) potentials in simulating diffusional association (see Figure 5.12). Hydrophobic desolvation potential grid values were multiplied by a factor of -0.013 ($hdfct = -0.013$) to compute the short-range attractive nonpolar interaction forces. For each set of simulations, four replica simulation runs were run with different starting position of ligand at the *b surface* to gain statistically relevant results. 50,000 trajectories were simulated in each run (total $4 \times 50,000 = 200,000$) of simulations when hydrophobic desolvation (HD) potentials were considered. For simulations without hydrophobic desolvation taken into account, 500,000 trajectories were simulated for each run (total 2 million trajectories were run). To increase the statistical significance, rate constants from all four replica simulations of the same system were averaged and a standard error was calculated using the SDA integrated tool *nos2rates*.

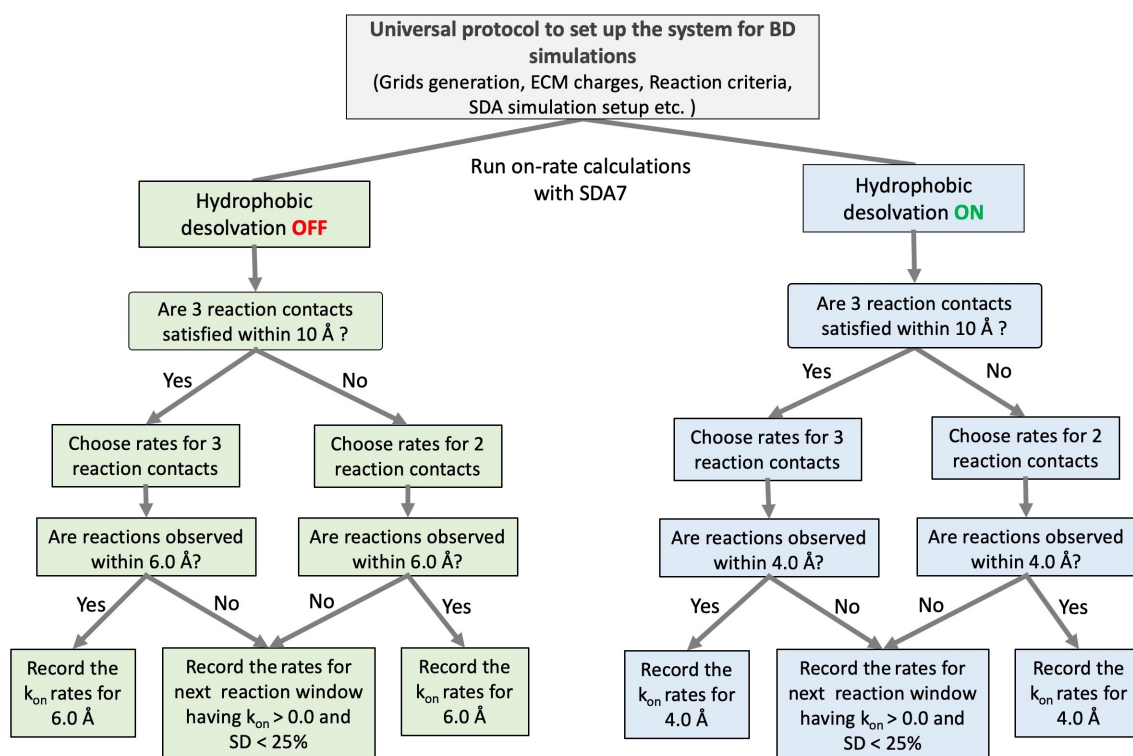


Figure 5.12: Schematic representation of BD simulations protocol to explore determinants of k_{on} rates for protein-ligand binding using SDA 7.

For computing diffusional association rates, encounter complexes satisfying 3 independent reaction contacts between protein and ligand were considered by default. In cases where reactions were not observed within 10 Å, rates for 2 reaction contacts were taken into account (refer to Figure 5.12). However, for certain systems, only encounter complexes with a maximum of 2 independent reaction contacts were formed due to limited accessibility of the binding pocket due to the specific binding mode. For such systems, encounter complexes satisfying 2 reaction contacts were considered for computing association rates. Because desolvation interactions are only relevant at distances less than 6 Å, for simulations run with only electrostatic forces and no hydrophobic desolvation potentials, the reaction window starting at 6 Å (having $k_{on} > 0$ and standard deviation $< 25\%$) was used to record the association rates.

Protein-ligand system simulated	Number of reaction contacts satisfied in encounter complexes	Distance for recording diffusional k_{on} (with Hydrophobic Desolvation)	Distance for recording diffusional k_{on} (without Hydrophobic Desolvation)
		(Å)	(Å)
Trypsin-benzamidine	2	4.0	6.0
FXa-rivaroxaban	2	5.0	6.5
FXa-apixaban	2	6.5	8.0
Haspin-5-iTU	3	11.0	NA
Haspin-5-brTU	3	10.0	NA
Haspin-5-clTU	3	10.0	NA
Haspin-5-ftTU	3	11.0	NA
Haspin-5-hTU	3	10.5	NA
Neuraminidase-oseltamivir	3	4.0	6.0
Neuraminidase-zanamivir	3	4.0	6.0

Table 5.18: Reaction criteria considered for defining successful encounter complexes and distances at which diffusional k_{on} rates were recorded in both sets of simulations (with and without hydrophobic desolvation forces) for different protein-ligand complexes studied.

In simulations run with hydrophobic desolvation potentials, a smaller reaction window of 4 Å was considered as cut-off window because short-range attractive hydrophobic interaction leads to closer contacts. For more details on the criteria used to record on-rates, please refer to protocol in Figure 5.12.

5.3 Results

5.3.1 Diffusional association rate constants (k_{on}) computed for the trypsin-benzamidine association

The Trypsin–benzamidine complex, due to its small size and monomeric structure, is a popular model system frequently used for developing and testing the methods for computing protein–ligand binding kinetics. Herein, we have also used the diffusional association of benzamidine to trypsin for optimizing some of the SDA input parameters used in our BD protocol to compute diffusional association (k_{on}) rate constants for protein–ligand association. Diffusional association rate constants were computed for the association of trypsin and benzamidine using SDA and 2 sets of simulations were run with hydrophobic desolvation (HD) interaction forces included in the first set and excluded in the second set of simulations (Figure 5.13 B). The encounter complexes between trypsin and benzamidine were formed starting at a distance of 3.0 Å irrespective of the presence or absence of HD forces in the simulation of diffusional encounter. The encounter complexes formed (in both simulations with and without HD forces included) at 3.0 Å almost reproduced the bound state with an RMSD of less than 1.0 Å between the orientation of benzamidine in the encounter complex and the crystallized structure of trypsin–benzamidine complex (Figure 5.13 A).

The computed diffusional association rate constants were consistently higher at all reaction distances when HD forces were taken into account. A sharp decline in the association rates was observed at distances below 8.0 Å when HD forces were not considered. However, when HD forces were considered in the BD simulations, the short-range hydrophobic forces become important as observed by considerably faster on-rates, even at shorter distances (Figure 5.13 B). The HD forces within 6 Å might result in some effective 2D surface diffusion of the ligand, thereby resulting in closer contacts and faster on-rates. However, we did not analyse the individual trajectories

to visualize any surface diffusion. The diffusional k_{on} rate constant computed for binding of benzamidine to trypsin was approximately 15 fold overestimated ($45.9 \pm 2.16 \times 10^7 \text{ M}^{-1}\text{s}^{-1}$) compared to the experimental value ($2.9 \times 10^7 \text{ M}^{-1}\text{s}^{-1}$), when HD forces were considered. When HD forces were excluded, the computed on-rate ($0.82 \pm 0.14 \times 10^7 \text{ M}^{-1}\text{s}^{-1}$) was about 3 fold lower than the experimental value (for which no information on experimental error was given).

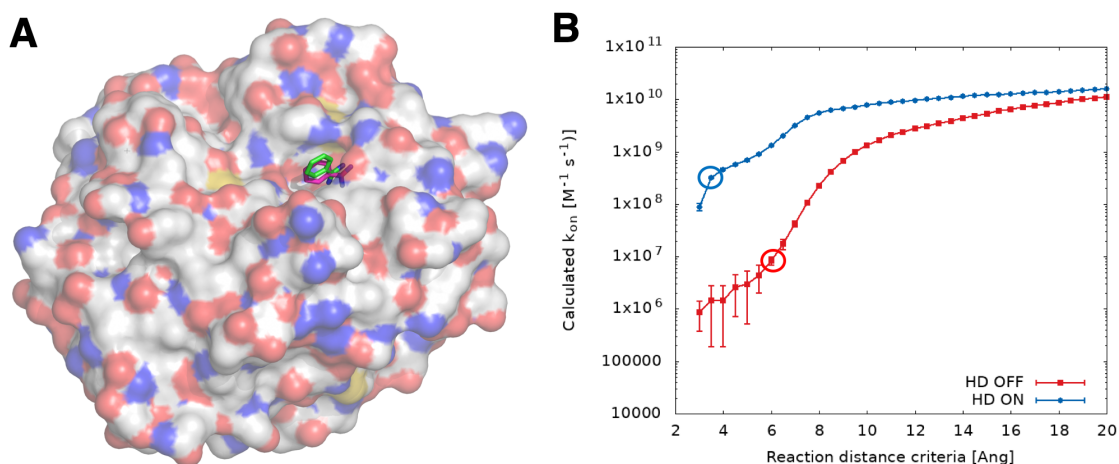


Figure 5.13: A) Comparison of the orientation of benzamidine (shown with stick representation in magenta) in the co-crystallized structure (PDB Id: 3PTB) with trypsin (shown with surface representation) and in the encounter complex formed during BD simulations with hydrophobic desolvation potentials using SDA (stick representation in green). B) Comparison of diffusional k_{on} rates calculated for the association of benzamidine with trypsin with and without inclusion of HD forces in the diffusional association. Based on the analysis protocol described in Figure 5.12, different reaction windows were selected specific to the protein-ligand system studied and the type of simulation, with their corresponding computed k_{on} values marked with circles (refer to Table 5.19 for exact numbers).

Protein-ligand system simulated	Experimental k_{on} value ($\times 10^6 \text{ M}^{-1}\text{s}^{-1}$)	Computed k_{on} value (with HD) ($\times 10^6 \text{ M}^{-1}\text{s}^{-1}$)	Computed k_{on} value (without HD) ($\times 10^6 \text{ M}^{-1}\text{s}^{-1}$)	Reference for experimental value
Trypsin-benzamidine	29.0	459.1 ± 21.56	8.17 ± 1.43	[157]
FXa-rivaroxaban	29.00 ± 6.0	169.6 ± 22.55	1.45 ± 0.34	[158]
FXa-apixaban	7.30 ± 1.60	88.55 ± 11.43	0.77 ± 0.17	[158]

Protein-ligand system simulated	Experimental k_{on} value ($\times 10^6 \text{ M}^{-1}\text{s}^{-1}$)	Computed k_{on} value (with HD) ($\times 10^6 \text{ M}^{-1}\text{s}^{-1}$)	Computed k_{on} value (without HD) ($\times 10^6 \text{ M}^{-1}\text{s}^{-1}$)	Reference for experimental value
Haspin-5-iTU	9.39 ± 2.03	39.14 ± 6.82	NA	[118]
Haspin-5-brTU	9.71 ± 2.67	31.26 ± 2.01	NA	[118]
Haspin-5-clTU	2.79 ± 1.26	11.83 ± 0.71	NA	[118]
Haspin-5-ftTU	2.06 ± 2.07	24.25 ± 4.12	NA	[118]
Haspin-5-hTU	0.20 ± 0.16	10.10 ± 2.50	NA	[118]
Neuraminidase- oseltamivir	2.52 ± 0.21	111.25 ± 6.80	3.21 ± 0.09	[159]
Neuraminidase- zanamivir	0.95 ± 0.08	552.50 ± 2.97	9.60 ± 0.40	[159]

Table 5.19: Comparison of experimental association rate constants and computed association rate constants from SDA for different protein-ligand complexes simulated (HD = hydrophobic desolvation potentials).

5.3.2 Diffusional k_{on} rate constants computed for the inhibitors of Human Coagulation Factor Xa

Association rate constants computed for association of rivaroxaban and apixaban to Human Coagulation Factor Xa using SDA are shown in Figure 5.14 C,D. Both rivaroxaban and apixaban bind to Factor Xa in an L-shaped conformation where one part of the ligand occupies the anionic S1 pocket and another part occupies the S4 pocket (see Figure 5.14 A,B). Since both protein and inhibitor molecules were modelled as rigid bodies and internal conformational flexibility was neglected, the encounter complexes of rivaroxaban and apixaban with Factor Xa satisfied maximum 2 reaction contacts within 10 Å reaction distance. Due to differences in the size of the inhibitors and the diversity of atoms contributing to the reaction criteria, the encounter complexes and hence, the rates were computed starting from different reaction windows. For rivaroxaban, association events were observed starting at 4.5 Å when hydrophobic desolvation forces were also taken into account, and at 5.5 Å when hydrophobic desolvation forces were excluded in the simulation of diffusional

encounter (see Figure 5.14 C and D). Apixaban, on the other hand, formed encounter complexes with Factor Xa starting from 6 Å in the simulations run with hydrophobic desolvation forces and from 7.5 Å when hydrophobic desolvation forces were ignored.

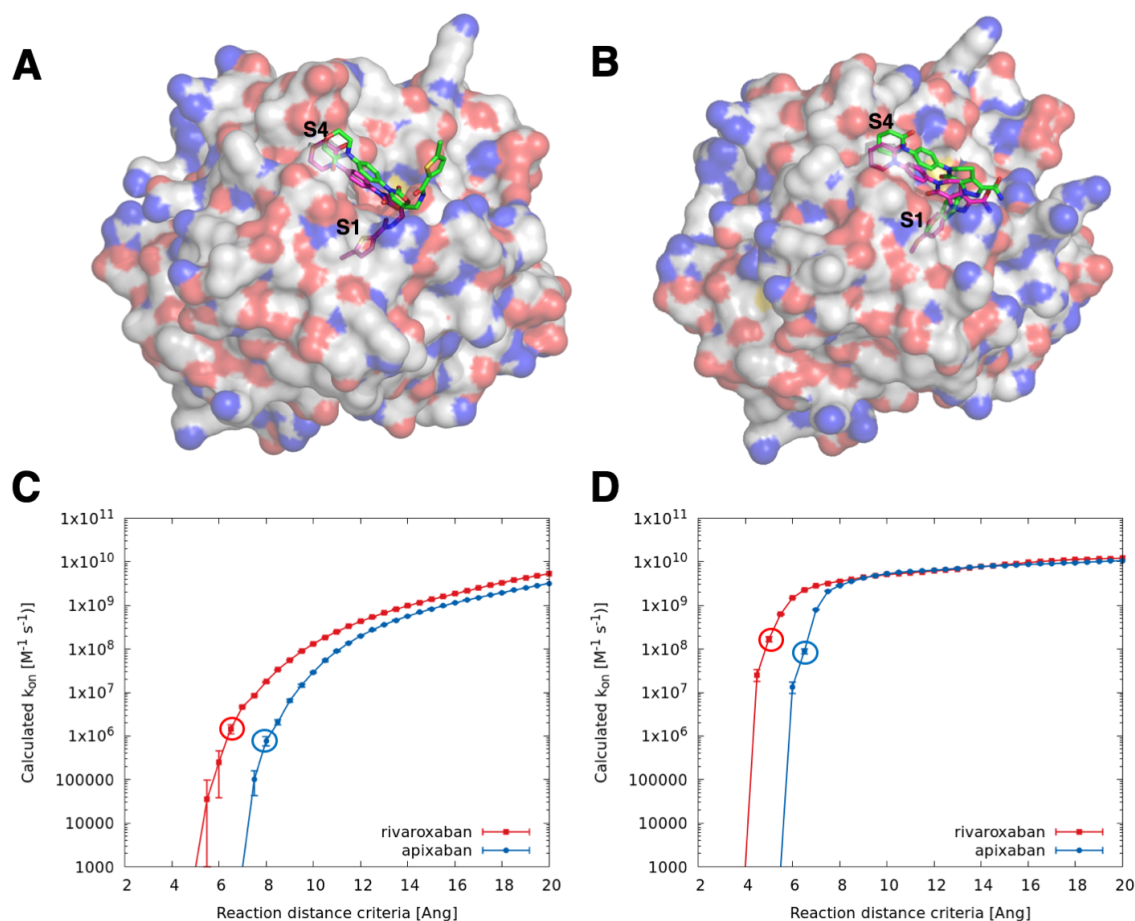


Figure 5.14: The orientations of A) rivaroxaban and B) apixaban (shown with stick representations in magenta) in the co-crystallized structures (PDB Ids: 2W26, 2P16 respectively) with Coagulation Factor Xa (shown with surface representation, colored by polarity of residues with positively charged, negatively charged and neutral residues shown in blue, red, and gray color respectively) and in the encounter complex formed during BD simulations with hydrophobic desolvation forces using SDA (shown with stick representations in green). In the encounter complexes of rivaroxaban with FXa, rivaroxaban did not occupy the S1 sub-pocket of FXa’s binding site and its chlorothiophene moiety was surface exposed. On the other hand, apixaban occupied the S1 sub-pocket in the majority of its encounter complexes with FXa formed at shorter distances. Comparison of diffusional k_{on} rate constants calculated for association of rivaroxaban and apixaban with Coagulation Factor Xa: C) without HD forces included, and D) with HD forces included in the simulation setup. The rates for shown for encounter complexes forming 2 reaction contacts. Based on the analysis protocol described in Figure 5.12, different reaction windows were selected specific to the protein-ligand system studied and the type of simulation, with their corresponding computed k_{on} values marked with circles (refer to Table 5.19 for exact numbers).

In calculations with HD forces included, the rate constants for reaction criteria distances between 15 Å and 20 Å remained stable at around $1.0 \times 10^{10} \text{ M}^{-1}\text{s}^{-1}$. For both rivaroxaban and apixaban, this value corresponds to the diffusional motion of ligands at a distance where they are not yet influenced by the distinct conformation and electrostatic distribution of the active site. At reaction criteria distances between 4 and 8 Å, the association rate constants dropped for both rivaroxaban and apixaban. The rate constant values are lower at shorter distances as encounter complex formation depends on the proximity and correct orientation of bound ligands. The latter requirement prevented multiple contacts being made at short distances. The rate constants computed from BD simulations with hydrophobic desolvation included, were overestimated for both rivaroxaban and apixaban, with computed k_{on} values almost 5 fold and 12 fold higher than the experimental values for rivaroxaban and apixaban, respectively (see Table 5.19). The computed k_{on} for rivaroxaban and apixaban from BD simulations without considering hydrophobic desolvation were underestimated by about 1 order of magnitude compared to the experiments. However, both of the simulation protocols estimated higher k_{on} values for rivaroxaban compared to apixaban, which is in agreement with the experimental observations (see Figure 5.14 C,D). The logP value of apixaban (2.22, source: DrugBank[161]) is higher than the logP of rivaroxaban (1.74, source: DrugBank[161]) suggesting it to be more hydrophobic than the rivaroxaban.

5.3.3 Diffusional k_{on} rate constants computed for the inhibitors of Haspin kinase

Association rate constants computed for the association of 5-iTU derivatives to Haspin at different reaction distances are shown in Figure 5.15 B. Lack of the conformational flexibility and rigid modelling of both haspin and inhibitor molecules restricted the access to the binding site by 5-iTU derivatives, with encounter complexes starting to form at only distances greater than 8 Å from the binding pocket (see Figure 5.15 A). In the BD simulations without HD forces taken into account, encounter complexes were observed only from distances higher than 15 Å. Therefore, for inhibitors of Haspin, we only computed rates from BD simulations with HD forces included in simulating diffusional encounter. Diffusional k_{on} rate constants observed for all 5-iTU derivatives showed huge standard deviations at smaller distances and therefore rate constants were recorded at distances starting from 10 Å

where the conditions specified in the SDA protocol (Figure 5.12) were met (refer to Figure 5.15 B and Tables 5.19 and 5.18 for the specific reaction windows considered for different 5-iTU derivatives and the corresponding rate constants computed for these reaction windows).

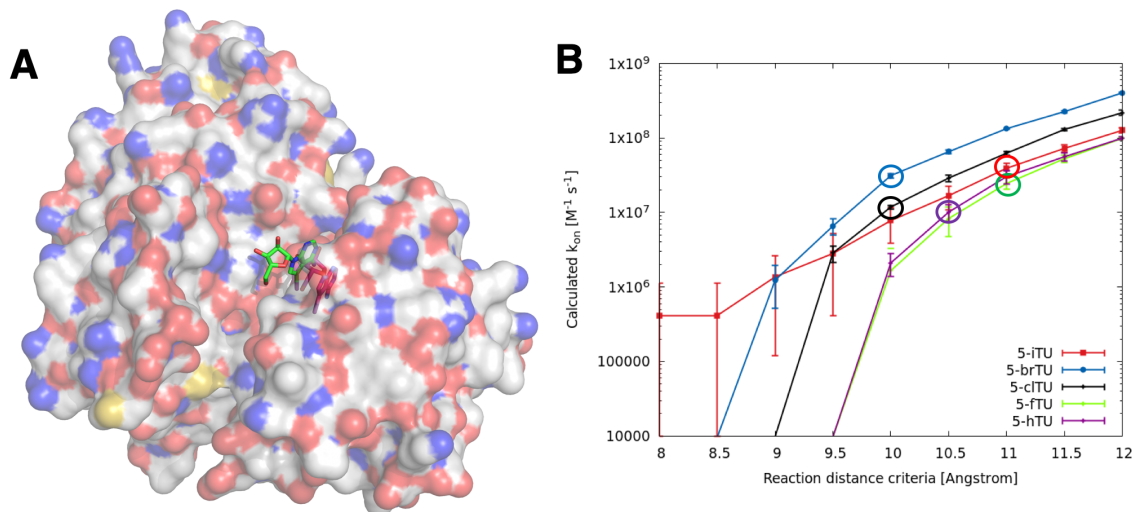


Figure 5.15: A) The orientations of 5-iTU (shown with stick representation in magenta) in the co-crystallized structure (PDB Id: 6G34) with Haspin (shown with surface representation, colored by polarity of residues with positively charged, negatively charged and neutral residues shown in blue, red, and gray color respectively) and in the encounter complex formed during BD simulations with hydrophobic desolvation forces (stick representation in green). B) Comparison of diffusional k_{on} rates calculated for association of 5-iTU derivatives with Haspin with HD forces included in the simulation setup. Based on the analysis protocol described in Figure 5.12, different reaction windows were selected specific to the protein-ligand system studied and the type of simulation, with their corresponding computed k_{on} values marked with circles (refer to Table 5.19 for exact numbers).

Diffusional k_{on} rate constants computed were higher for inhibitors containing bigger halogen atoms such as 5-iTU, 5-brTU and 5-clTU than the inhibitors with a smaller halogen (5-ftu) or no halogen substituted (5-hTU) (see Figure 5.15 B). This is consistent with the experimental observations that substitution of bigger halogens on tubercidin results in mediation of halogen- π interaction between halogen of inhibitors and aromatic ring of phenylalanine gatekeeper residue in the binding site of Haspin, thereby resulting in an increase of k_{on} rates and decrease of k_{off} rates. Since diffusional k_{on} rate constants of inhibitors with bigger halogens were consistently higher even at large reaction distances, changes in electrostatics induced by halogen substitutions might result in stronger long-range electrostatic interaction between the inhibitors and Haspin, leading to faster diffusion of halogenated in-

hibitors towards Haspin’s binding site. Although, SDA was not able to exactly rank these 5-iTU derivatives based on their k_{on} rate constants, it managed to capture the effects of changes in electrostatics and non-polar interactions of inhibitors due to substitution of bigger halogens. The diffusional k_{on} computed for all 5-iTU derivatives were overestimated compared to the experimental values (see Table 5.19) as we computed the rates at larger reaction distances, because at shorter distances, rates were statistically not reliable and showed large error bars (see Figure 5.15 B). These large errors at shorter distances could be attributed to the limited access to the binding site, with BD trajectories in only a subset of simulations satisfying the requirements for formation of an encounter complex.

5.3.4 Diffusional k_{on} rate constants computed for the inhibitors of Neuraminidase

Diffusional k_{on} rate constants computed by SDA for diffusional association of oseltamivir and zanamivir with Neuraminidase at different reaction distances using the 2 different simulation protocols (with and without HD forces) are shown in Figures 5.16 B and 5.16 C. The relatively exposed binding site of Neuraminidase, the presence of a number of charged amino acid residues in its binding site, and the zwitter ionic nature of both inhibitors contributed to the strong electrostatic interaction between the protein and the inhibitors with encounter complexes observed at smaller distances from 4 Å. Due to the differences in the size of the ligands and the diversity of atoms contributing to the reaction criteria, oseltamivir made up to three, zanamivir up to five reaction contacts in the encounter complexes. When only electrostatic forces were considered and no HD forces in the association, the on-rates computed were higher for zanamivir than for oseltamivir (Figure 5.16 B). The computed k_{on} rate constant ($3.21 \pm 0.09 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$) for oseltamivir was very close to the experimental values ($2.52 \pm 0.21 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$). In fact, this k_{on} value computed for oseltamivir by our protocol is very similar to the k_{on} value ($5.17 \pm 0.08 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$) computed by Sung *et al.*[21] with their BD simulation procedure (without including HD term in the simulations). The k_{on} rate constant computed ($9.60 \pm 0.40 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$) for zanamivir was almost 10-times higher than the experimental value ($0.95 \pm 0.08 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$). The electric dipole moment of zanamivir (25.20 Debye) calculated from RESP atomic charges was higher than the electric dipole moment of oseltamivir (22.24 Debye) suggesting it to be slightly more polar than

oseltamivir. This additional polarity might be because of the presence of additional polar moieties on zanamivir such as diaminomethyl and trihydroxypropyl groups which may result in stronger electrostatics interactions with the Neuraminidase and this could be the reason why computed on-rates for zanamivir were higher than for oseltamivir. Or, the higher on-rates for zanamivir could be due to the fact that it might be relatively easier for zanamivir to satisfy the criteria of formation of 3 reaction contacts in the encounter complex due to its bigger size and presence of more number of polar atoms on the list of contacts provided as the reaction criteria.

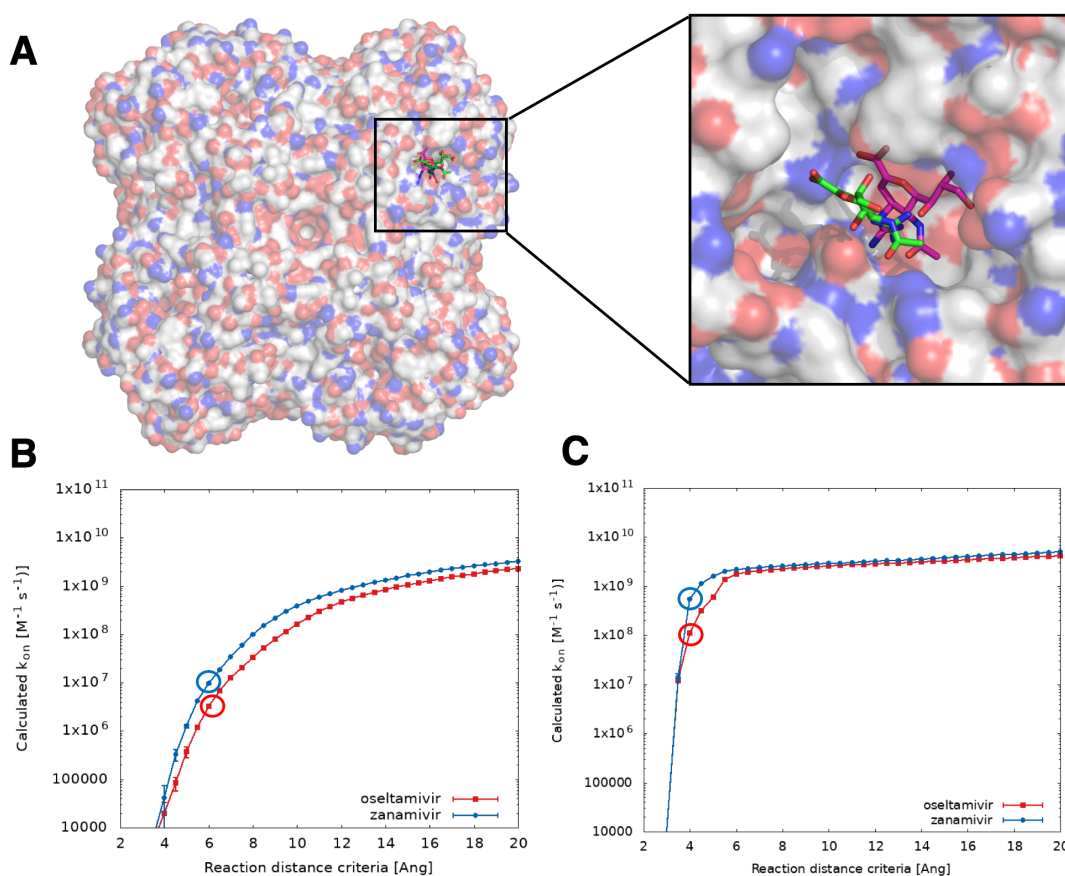


Figure 5.16: A) The orientations of the oseltamivir (shown with stick representation in magenta) in the co-crystallized structure with Neuraminidase (shown with surface representation, colored by polarity of residues with positively charged, negatively charged and neutral residues shown in blue, red, and gray color respectively) and in the encounter complex formed during BD simulations with hydrophobic desolvation forces (stick representation in green). Comparison of diffusional k_{on} rates calculated for association of oseltamivir and zamamivir with Neuraminidase: B) without HD forces included, and C) with HD forces included in the simulation setup. Rates are shown for encounter complexes satisfying the criteria for 3 reaction contacts. Based on the analysis protocol described in Figure 5.12, different reaction windows were selected specific to the protein-ligand system studied and the type of simulation, with their corresponding computed k_{on} values marked with circles(refer to Table 5.19 for exact numbers).

In the simulation protocol with HD forces included, computed on-rates for both oseltamivir and zanamivir were very high (higher than $10^9 \text{ M}^{-1}\text{s}^{-1}$) even at shorter distances (from 6 Å). After 7-8 Å, rates become stable at around $10^9 \text{ M}^{-1}\text{s}^{-1}$. This trend was observed for both oseltamivir and zanamivir and describes the diffusional motion of ligands in a distance, at which they are not yet influenced by the distinct conformation and electrostatic distribution of the active site. At reaction criteria distances between 4 and 6 Å, association rate constants dropped for both oseltamivir and zanamivir, but the drop was higher for oseltamivir. Still, the computed k_{on} rate constants ($111.25 \pm 6.80 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$) were approximately 40 times higher for oseltamivir than the experiments ($2.52 \pm 0.21 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$). Similarly, for zanamivir, computed k_{on} value ($552.50 \pm 2.97 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$) was highly overestimated by more than 2 orders of magnitude compared to the experimental value of $0.95 \pm 0.08 \times 10^6 \text{ M}^{-1}\text{s}^{-1}$. Also, Neuraminidase are known to have an open 150-loop structure in the unbound form, that closes upon drug binding, and the polar residues in the 150-loop (Asp151, Arg152) interact with polar side chains of the inhibitors. This suggests that the slower conformational changes of protein and inhibitor molecules are associated with the formation of a fully bound protein-ligand complex and these short-range conformational changes are not modelled in BD simulations, and we have only used the Neuraminidase structure with an open 150-loop conformation (PDB Id: 2HTY). Therefore, SDA results suggest that zanamivir, inspite of having higher diffusional on-rates, has the slower experimental k_{on} value than oseltamivir, meaning that conformational adaptation might be slower for binding of zanamivir to Neuraminidase than for oseltamivir.

5.4 Concluding Discussions

The current application of our protocol to different protein-ligand systems for the calculation of diffusional association rate constants of the ligands demonstrate that the protocol has some limitations in correctly ranking the ligands according to their experimental on-rates. In some cases, especially for inhibitors binding to Coagulation Factor Xa and Haspin kinase, the protocol managed to correctly distinguish the inhibitors with slow and fast on-rates. However, for inhibitors binding to Neuraminidase, the protocol failed to capture the effects of faster binding of oseltamivir compared to zanamivir. Since our protocol is solely based on BD simulations where

we model the interacting molecules as rigid bodies and neglect the internal conformational flexibility of molecules, our protocol may not completely capture the short-range effects of binding of small but flexible ligand molecules, especially when the protein-ligand binding is conformation dependent. Although we have optimized a set of parameters for the diffusional association of protein-ligand systems using SDA software, this list is not complete and there are several other parameters that need to be evaluated and optimized on a diverse set of systems. The short-range desolvation forces acting between protein and ligand molecules might be sensitive to the size of grid-spacing used for generating desolvation potentials grids. For the generation of desolvation grid potentials, we have used the grid spacing value of 1 Å, that was used for protein-protein association studies in the past. We believe that different values of grid spacing need to be evaluated for protein-ligand association, especially because the rates (hence the binding) are very sensitive to the presence or absence of short-range hydrophobic desolvation forces as observed in our calculations when using HD term in the simulations. Moreover, the current dataset on protein-ligand systems needs to be extended further and studied to have a well optimized and robust protocol. As we have already discussed, this protocol, solely based on BD simulations alone would not be sufficient to model the complete binding process but it can serve as a good starting point for multiscale modelling, where, for example, the less demanding BD simulations can be run using this protocol to model the initial diffusional encounter of protein and ligand molecules, following which a more computationally rigorous MD-based regime can be used to account for the flexibility and conformational changes to simulate the formation of the final bound complex. The association rates from these two different approaches can be combined using specialized techniques such as milestoning to compute the on-rates for the complete binding.

Chapter 6

KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding

I have contributed to the development of KBbox (<http://kbbox.h-its.org/toolbox/>), a webserver which provides information about various computational methods to study molecular binding kinetics, and different computational tools that employ them. It is developed as an effort to guide less experienced researchers in the use of different computational and simulation approaches available to compute the kinetics parameters of drug-protein binding. The toolbox lists and provides an overview of the current state-of-the-art computational approaches for studying molecular binding kinetics, with methods ranging from relatively high-throughput regression-based cheminformatics methods to computationally intensive atomic-level simulation approaches. KBbox provides a curated list of published applications of the methods, providing users with an easy-to-find reference list. For a number of methods, detailed tutorials are also provided that give the user an introduction into how to run the calculations and to reproduce some of the example cases. KBbox provides a query interface that asks a series of questions relating to structural and kinetics data available to them, and the data they wish to calculate, and provides them with a list of methods found in the toolbox that match their query, sorted approximately by the computational resources required for their application. The web server is easily extendable, allowing us to add new methods to the toolbox as they are developed and published.

List of methods
Grouped into 4 categories:

- QSKR Approaches
- Molecular Simulation
- PKPD Modelling
- Molecular Modelling

Total 19 methods at present

Example Cases

- 33 examples at present
- 31 published, 2 unpublished
- Each example linked to 1 or more computational methods

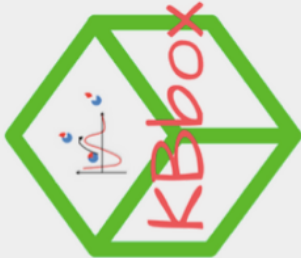
List of tutorials

- In-depth tutorials of computational methods
- 5 tutorials at present
- Each tutorial describes how to reproduce one of the example cases

List of tools
Grouped into 5 categories:

- Simulation Tools
- Preparation and General Modelling Tools
- PKPD Modelling Tools
- Data Analysis Tools
- Structure Visualization Tools

Total 18 tools at present



K4DD Project Contact Us

KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding

Welcome to KBbox!

Here you can find information about various **computational methods** for studying molecular binding kinetics, and the **computational tools** that employ them.

KBbox is developed to be of use to researchers interested in applying these methods in their work. To help with this, KBbox provides an updated list of **examples of published work**, along with **detailed tutorials** to guide less experienced researchers.

KBbox is still being developed, and new content will be added continually. If you have any examples of published computational work or computational tools that you think would be useful to add, please **contact us** so we can add them to the site.

Workflow to find a method that suits user's needs

Not sure which method to use for a particular problem? [Click here!](#)

Figure 6.1: A snapshot of the homepage of KBbox toolbox website (<http://kbbbox.h-its.org/toolbox/>). The key features/content of the toolbox are highlighted in different colors.

6.1 Technical Implementation

KBbox was developed using the high-level python web framework Django (Version 2.2, Retrieved from <https://djangoproject.com>) for serving dynamic HTML content created from information stored in its database (see Database Structure). The responsive web interface of the KBbox is provided through the use of JavaScript and CSS plugins from the Bootstrap 3[162] framework. Biopython's[163] Bio.Entrez module is used for programmatic access to Entrez[164], a data retrieval system that provides users access to NCBI's databases such as PubMed. The Bio.Entrez parser allows for example to search PubMed or download GenBank records from within a Python script. This functionality is used in KBbox to add new example cases directly by providing either a valid PubMed Id or a valid DOI identifier, and the data is parsed using the Beautiful Soup parser. KBbox uses the CKEditor rich text editor in the Django-CKEditor python package to enable writing the content directly in web pages e.g. to add and format rich text for new methods, tools or tutorials. In addition, KBbox also uses Pillow (version 4.0, retrieved from <https://pillow.readthedocs.io/en/stable/>), a Python Imaging Library (PIL) fork which provides image processing capabilities to the Python interpreter.

6.2 Database Structure

KBbox uses SQLite, the default database available in the Django framework. SQLite is included in Python and it is a fast, self-contained and highly-reliable SQL database engine. The data structure used in KBbox is shown in the Figure 6.2. The main data table contains the computational methods that comprise the toolbox (class *CompMethod*). Each entry in the table contains a short summary description of the method and a more detailed introduction to the method. It also contains a number of Boolean parameters that are used for querying the methods in the toolbox, to find which methods match the user's needs. These relate to the data that the user wishes to obtain (association rates— k_{on} ; dissociation rates— k_{off} ; pharmacokinetic/pharmacodynamics predictions—PKPD), whether training or atomic structural data is required by the method, and whether the method is able to provide absolute data, or only relative data. Finally, the table contains an integer (*comp_cost* field) describing the approximate computational cost of the method

from 1 (least expensive) to 5 (most expensive). Values of 1 or 2 relate to methods that can be run in short time on a desktop computer. Values in the range 3 to 5 correspond to longer simulation-based methods. Each method is sorted into one group (class *CompMethodGroup*), to allow KBbox to organize the methods into different classes allowing users to search for methods more easily. Currently, these groups are molecular modeling, molecular simulation, PKPD modelling and QSKR (qualitative structure – kinetics relationships) approaches (discussed in the following sections).

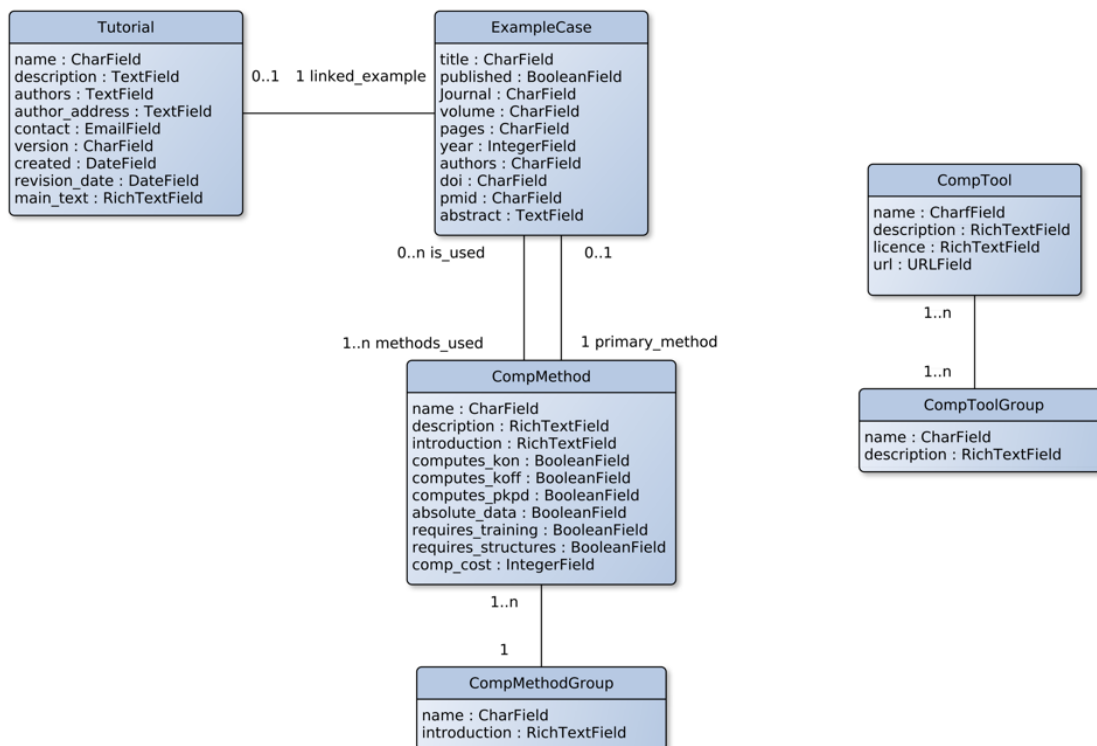


Figure 6.2: Entity relationship diagram showing the data representation used in KBbox. (<http://kbbox.h-its.org/toolbox/>). For each of the 6 classes, name of the fields and their data types are shown. This figure was prepared by Dr. Neil J. Bruce.

Each method is linked to one or more examples of previously published research, or a report of currently unpublished data, in which that method was either the primary method used, or one of a set of methods applied (class *ExampleCase*). Each row in the table describing these examples contains author information and a flag to say whether the work is published or unpublished. For published examples, citation data is also provided. For some of the examples, tutorials are also provided these are recorded in an additional table (class *Tutorial*).

A separate table is used to populate the list of computational tools described by KBbox (class *CompTool*). Each row of this table contains a description of the

tool, its license and a URL to its web page. In a similar manner to the methods, the tools are grouped into classes (class *CompToolGroup*). Currently, these are data analysis tools, PKPD modelling tools, preparation and general modelling tools, simulation tools and structure visualization tools (discussed in the following sections). A particular tool could be a member of more than one group.

6.2.1 Query Interface to choose the methods

The toolbox also provides a query interface that asks users about information on the amount of structural and kinetic data they have, and the data they want to calculate, and suggests them a list of appropriate methods that they could use. This list of methods is sorted based on the amount of computational resources required by the methods. The query is built based on the information provided by users on the data they want to estimate (k_{on} , k_{off} or PKPD modelling), amount of 3D information available on protein-ligand complexes, and whether experimental kinetic data is available for the ligands (see Figure 6.3). The methods that match this query are provided, with methods sorted by the computational resources required for their application (*comp_cost* field in *CompMethod* class).

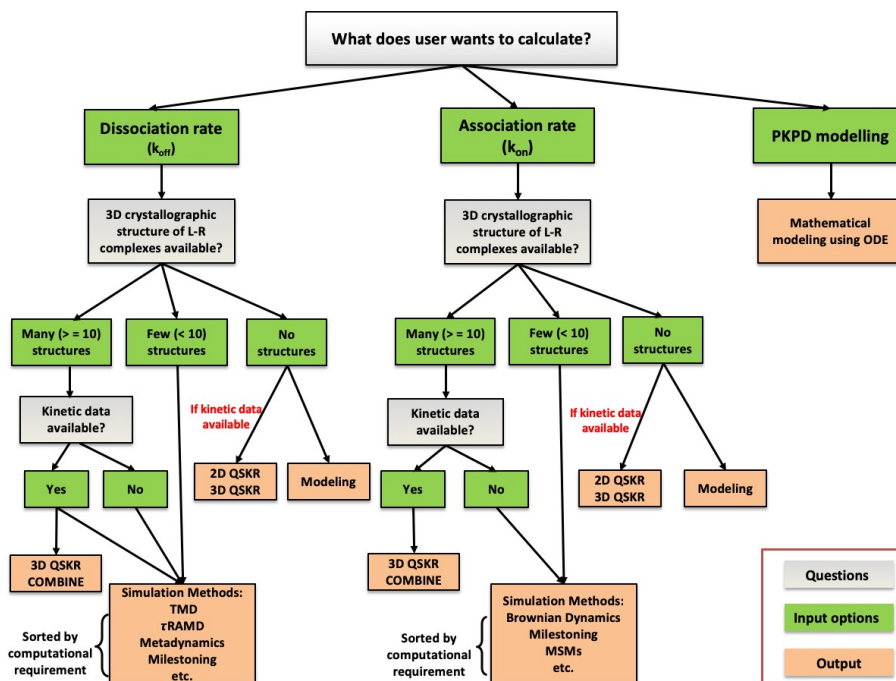


Figure 6.3: Schematic outline of the query building workflow of the Interface to suggest list of appropriate methods to the users. Note: This schema served as the starting point for the design of the interface, however the design evolved with time and therefore, the actual workflow implemented on the website might be different from the one shown here.

6.3 Group of Methods available in KBbox

The **Methods** webpage provides a description of a number of computational methods that can be used to investigate binding kinetics. At present, the toolbox provides information on 19 different computational approaches that have been broadly categorized into the following 4 groups:

- **QSKR Approaches**

QSKR (quantitative structure-kinetic relationship) are regression or classification models that relate a set of physio-chemical properties or molecular descriptors to the binding kinetics of compounds. QSKR regression models use regression techniques to relate a set of molecular descriptors (referred to as "predictor variables" in Machine Learning terminology) to the kinetic parameter ("response variable") and predicted parameter is a continuous value. QSKR classification models, on the other hand, relate the predictor variables to a categorical value (e.g. slow or fast, active or inactive) of the response variable.

- **Molecular Simulation**

Molecular simulation is a type of N-body simulation technique used for studying the physical movements of atoms and molecules. In molecular simulation, the dynamics of the system can be studied by allowing the atoms and molecules to interact for a fixed period of time. In the most common version, forces between the interacting particles and their potential energies are calculated using interatomic potentials or molecular mechanics force fields, and the trajectories of atoms and molecules are determined by numerically solving the Newton's equations of motion.

- **PKPD Modelling**

PKPD modeling (pharmacokinetic/pharmacodynamic modeling) combines dose-concentration relationships (pharmacokinetics) and concentration-effect relationships (pharmacodynamics) into one set of mathematical expressions, that allows to establish and evaluate dose-concentration-response relationships and subsequently describe and predict the effect-time courses resulting from a drug dose.

- **Molecular Modelling**

Molecular modelling tools are used to create and modify the 3D structures of

molecules. When experimental structural data is missing, they can be used to predict the structures of biomolecules, biomolecular complexes and of complexes formed between biomolecules and small molecules. They are also used to prepare existing structures for use in simulations.

6.4 List of Examples

The different computational methods described in KBbox are linked to one or several of the example cases (also available under **Examples** webpage link) where they were successfully applied. The toolbox currently lists total 33 example cases of which 31 are already published.

6.5 Group of Computational tools in KBbox

The **Tools** webpage lists a number of different computational tools that are employed in binding kinetics studies. At the moment, there are total 18 computational tools and software that are grouped under the following 5 categories:

- **Simulation Tools**

Currently, this section consists of information on 7 different simulation tools such as Amber, Amber Tools, CHARMM, Gromacs etc.

- **Preparation and General Modelling Tools**

There are total 12 different tools available in this category that are generally used for preparation and modeling of 3D structures and potential grids for use in simulation and docking based methods.

- **Pharmacokinetic and Pharmacodynamics (PKPD) Modelling Tools**

This category currently lists only one PKPD modelling tool: Berkeley Madonna, which is a mathematical modelling software package used for numerically solving ordinary differential equations, difference equations and multi-dimensional transcendental algebraic equation roots.

- **Data Analysis Tools**

This section lists tools and programming languages commonly used for analysing, plotting and visualizing output from molecular simulations and chemoinformatics methods.

- **Structure Visualisation Tools**

This category lists tools such as PyMOL, VMD or MOE commonly used to visualize 3D structures, trajectories from molecular simulations and potential grids of biomolecules.

6.6 List of Tutorials

KBbox has list of tutorials available under the **Tutorials** webpage to introduce newcomers to the different computational methods required to reproduce some of the example cases. These tutorials are also linked to their published example cases. These tutorials provide scripts, input dataset and input files along with detailed instructions on using specific computational method for computing rates for binding kinetics. Currently, KBbox has 5 different tutorials available on different methods:

- **Estimation of relative residence times of protein-ligand complexes using τ -Random Acceleration Molecular Dynamics (τ RAMD)**

This tutorial guides the user through the process of setting up and running RAMD simulations for estimation of the relative residence time (τ) of a protein-small molecule complex[29]. The procedure is demonstrated for a complex of a low molecular weight compound with the N-terminal domain of the heat shock protein, HSP90.

- **Data exploration and linear regression of a kinetic dataset using R**

This tutorial guides the user through the process of doing Multiple Linear regression and data exploration on 16 MAP38 kinase inhibitors within the software package R. Explorative data analysis is carried out on this dataset, containing precalculated physicochemical descriptors. Multiple linear regression and correlation analysis are utilized to identify descriptors influencing k_{off} .

- **Compartmental modelling and simulation in Berkeley Madonna**

This tutorial demonstrate the use of compartmental modelling and simulation in Berkeley Madonna in predicting the receptor occupancy time profile in a body tissue after intravenous administration of a receptor ligand. In this tutorial, the selective dopamine D2 antagonist raclopride is used as an example.

The pharmacokinetics (PK) and dopamine D2 receptor occupancy (RO) in brain after intravenous administration of raclopride to rat are simulated[165].

- **Prediction of the rate of formation of a protein-protein complex using SDA**

This tutorial describes the use of SDA to perform Brownian dynamics simulations to predict the bimolecular association rate constant for the formation of a protein-protein complex[131].

- **Generation of Quantitative structure-kinetics relationships (QSKRs) using Comparative Binding Energy (COMBINE) Analysis**

This tutorial guides the user through the process of setting up and running COMparative BINDing Energy (COMBINE) analysis to derive Quantitative structure-kinetics relationship (QSKR) for dissociation rate constants (k_{off}) of inhibitors of a drug target[109]. The procedure is demonstrated for a dataset of 70 inhibitors of heat shock protein 90 (HSP90) belonging to 11 different chemical classes.

6.7 Example Use Cases

KBbox is useful for a variety of users with different needs and experiences. Here we outline two potential use cases of the web server.

6.7.1 What method should I use for a given project?

A PhD student with some basic experience in computational modeling and simulation, is interested in starting a new project where he has data on residence times of a set of inhibitors for a given protein target. He also has crystallographic data for these inhibitors bound to the target. He wants to use computational modeling to predict the determinants of short and long residence time compounds and use this knowledge to predict residence times of compounds with no experimental data available. He arrives on the KBbox home page (Figure 6.4, black box), and click on the button “Not sure what method to use for a particular problem? Click here!”, and is then asked a set of questions relating to the data available to him, and the data he is interested in calculating. After these questions are answered, KBbox checks the database, and a list of methods, that match his query are presented to him

(Figure 6.4, red boxes, anticlockwise). These methods are sorted approximately by the computational resources required for the calculation. The student clicks on each entry in this list, and he is taken to pages that give an overview of the methods, along with a curated list of examples of previous applications of each method, with links to the relevant journal articles. The student selects COMBINE analysis[46] as the method he is interested in, and he then follows the link to the tutorial that describes how to perform COMBINE analysis on his data (Figure 6.4, red boxes, anticlockwise).

The image shows a screenshot of the KBbox website. The main page features a navigation bar with 'Methods', 'Examples', 'Tutorials', and 'Tools'. The main content area is titled 'KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding'. Below this, there is a 'Welcome to KBbox!' section and a list of methods. A red box highlights a link: 'Not sure which method to use for a particular problem? Click here!'. Below the main page, two example use cases are shown in red boxes. The first is 'Find a method to suit your needs', which includes a search form and a list of methods. The second is 'Comparative Binding Energy (COMBINE) Analysis', which includes an introduction and a list of example cases. A green box highlights the 'Example Cases' section of the 'τ-Random Acceleration Molecular Dynamics (τRAMD)' page, which includes a list of example cases and a tutorial link.

KBbox: A Toolbox of Computational Methods for Studying the Kinetics of Molecular Binding

Welcome to KBbox!

Here you can find information about various computational methods for studying molecular binding kinetics, and the computational tools that employ them.

KBbox is developed to be of use to researchers interested in applying these methods in their work. To help with this, KBbox provides an updated list of examples of published work, along with detailed tutorials to guide less experienced researchers.

KBbox is still being developed, and new content will be added continually. If you have any examples of published computational work or computational tools that you think would be useful to add, please contact us so we can add them to the site.

Not sure which method to use for a particular problem? Click here!

Find a method to suit your needs

What do you want to calculate? -

- Dissociation rate constants or residence times

Do you have existing kinetic data with for a number of similar complexes, with which you can train a model? -

- Yes, I have training data

Do you have structural data to define the complex formed by the binding partners? -

- Yes, I have structural data

Do you need to obtain absolute data values, or are relative values sufficient? -

- Absolute data is required

The following methods are identified that match your needs. They are listed in the approximate order of the computational resources that are required from low to high:

- Linear Regression of Kinetic Data with Chemical Descriptors
- Comparative Binding Energy (COMBINE) Analysis
- τ-Random Acceleration Molecular Dynamics (τRAMD)
- Steered molecular dynamics
- Targeted molecular dynamics
- Smoothed or Scaled Molecular Dynamics (Scaled MD)
- Adaptive Multilevel Spitting (AMS)
- Adiabatic-bias Molecular Dynamics (ABMD)
- Metadynamics
- Bias-exchange Metadynamics (BEMD)
- Unbiased molecular dynamics simulation
- Markov State Modeling

Comparative Binding Energy (COMBINE) Analysis

Introduction

COMBINE analysis uses partial least squares regression to derive a system-specific model for binding free energy or a related experimental observable based on weighting of force field energy terms describing contributions to the protein - ligand interaction energy.

COMBINE approach is used to derive a target-specific scoring function to compute binding free energy or a related property by exploiting the information contained in the 3D structures of receptor-ligand complexes. Energy-minimized structures of protein-ligand complexes are used to compute protein - ligand interaction energies using a molecular mechanics force field. These energies are then partitioned and subjected to regression based methods such as Partial Least Squares (PLS) regression, to derive a statistical model which relates the experimental observable to weighted selected components of the protein - ligand interaction energy. The interaction energy components are typically Lennard-Jones and Coulombic terms from a standard force field, decomposed on a per residue basis.

Example Cases

For examples of previously performed studies in which τ-Random Acceleration Molecular Dynamics (τRAMD) was the primary method used, see the following example cases:

- Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations.
- Machine Learning Analysis of τRAMD Trajectories to Decipher Molecular Determinants of Drug-Target Residence Times

Example Cases

For examples of previously performed studies in which Comparative Binding Energy (COMBINE) Analysis was the primary method used, see the following example cases:

- Prediction of Drug-Target Binding Kinetics by Comparative Binding Energy Analysis.

Tutorials

The following tutorial describes the use of Comparative Binding Energy (COMBINE) Analysis:

- Generation of Quantitative structure-kinetics relationships (QSKRs) using Comparative Binding Energy (COMBINE) Analysis

Figure 6.4: Representation of two of the example use cases of KBbox.

6.7.2 Where can I find information on previous applications of a method for studying kinetics?

A postdoctoral researcher is interested in using τ -Random Acceleration Molecular Dynamics[29] (τ RAMD) to study the unbinding of a set of compounds from a target protein and to rank them according to their relative residence times. While he has heard about the method, he does not have experience of running these simulations, and is looking for more information on the method and examples of previously published research using τ RAMD. He arrives on the KBbox homepage and then clicks on the “Methods” button at the top of the page. From here he selects “Molecular Simulation” from the menu on the left-hand side of the page and find the entry for τ -Random Acceleration Molecular Dynamics (Figure 6.4, green boxes, anticlockwise). After clicking on this entry, he is presented with an overview of the method, and a list of published examples of its application to binding kinetics studies. He/she follows the links for the examples and is taken to the relevant journal pages.

The computation of binding kinetics is an active area of interest, and the development and application of more robust and advanced computational approaches is expected to increase in the future. Therefore, the KBbox will be extended continuously to include data on new computational methods and their example cases. Also, more tutorials on different computational approaches will be added in the future. Hence, KBbox will be useful to the scientific community and will provide information on state-of-the-art of computational methods available to investigate and estimate kinetic parameters for molecular binding.

Chapter 7

Summary and Outlook

A number of recent studies suggesting the better correlation of efficacy of a drug with its residence time at its receptor than its affinity has led to widespread efforts in both industry and academia to consider the role of drug binding kinetics in drug discovery programs. This has resulted in a growing demand for *in-silico* and experimental methods that can estimate or predict kinetic parameters of drug-protein binding. In addition, understanding the mechanistic determinants of drug-target binding kinetics is important for aiding the design of lead molecules with optimized kinetic properties.

One of the important aims of this thesis was to use the available information from structures of protein-drug complexes and experimentally determined kinetic parameters of protein-drug binding to derive Quantitative Structure-Kinetics Relationships (QSKRs). For this purpose, I employed Comparative Binding Energy (COMBINE) analysis, to derive protein-specific scoring functions for the k_{off} rate constants of inhibitors of HIV-1 protease and HSP90. For both of these protein systems, I managed to derive statistical models which relate the k_{off} rate constants of their inhibitors to weighted selected components of the drug-receptor interaction energy. Unlike the congeneric series of compounds normally used for training such linear regression models, herein I have used diverse sets of inhibitors that have very different scaffolds and binding modes. These models were found to have good predictive ability as assessed using different cross-validation methods and a validation data set. These models can therefore be used to make predictions for off-rates of novel inhibitors of these proteins. Using COMBINE analysis, I was also able to identify key protein-inhibitor interactions that explain the variance in the binding kinetics of

the inhibitors. These specific components of interaction energy provide insights into the mechanisms of specific slow and fast dissociating classes of inhibitors. My results on two different targets with very diverse sets of inhibitors considered, suggests that the COMBINE analysis has potential as a robust and medium-throughput QSKR method and its scope of application is expected to grow as more data on measured kinetic parameters becomes available.

Recent work on Haspin kinase and its halogenated inhibitors has demonstrated that inhibitors with long residence times can be designed by introducing a halogen-aromatic π interaction between a halogen atom which is commonly found in the drugs, and an aromatic residue in the binding site. Substitution of an iodide moiety in tubercidin to form 5-iodotubercidin (5-iTU), a close analogue of ATP, resulted in the formation of a halogen-aromatic π interaction with the F605 gatekeeper residue present in the binding site of Haspin, as confirmed from the analysis of the 3D crystallographic structures of Haspin complexed with 5-iTU. 5-iTU shows a very high affinity for Haspin and also a very long residence time compared to ATP and unsubstituted tubercidin. Characterization of the affinities and binding kinetics of 5-iTU derivatives (substituted with smaller halogen atoms: Br, Cl and F) with different experimental assays showed that the affinities as well as the residence times of 5-iTU derivatives diminish with the decreasing size of the substituted halogen atom. I performed quantum mechanical interaction energy calculations to analyze the nature of the polarization mediated interactions of the core inhibitor scaffold with the gatekeeper aromatic residues. I calculated the second-order Møller–Plesset interaction energies (E_{MP2}) between the 5-iTU derivatives and the gatekeeper phenylalanine (F605) residue at consecutive levels of quantum mechanical theory and partitioned the E_{MP2} energy into its constituent energetic components using a many-body energy decomposition scheme. I found that the correlation energy (E_{CORR}) makes a major contribution to the total E_{MP2} interaction energy between 5-iTU derivatives and the gatekeeper residue. Also, I observed a very high correlation between E_{CORR} and the experimentally measured residence times of 5-iTU derivatives with a correlation coefficient (R^2) of 0.97. This correlation energy explains the second-order intermolecular dispersion interactions and the correlation corrections to the Hartree–Fock energy. I also computed binding free energies of the Haspin-inhibitor complexes using the classical MM/GBSA approach, to account for the complete protein structure. The computed binding free energies correlated well with the calorimetric data obtained

from experiments suggesting that the enthalpic contribution to binding increases with the increase in size of the halogen atom substituted on inhibitors. In addition, the residence times of the 5-iTU derivatives computed with the τ -RAMD procedure correlated well with the experimentally measured residence times.

I then went on to establish a protocol for the high-throughput calculation of diffusional association rate constants for protein-small molecule binding by running continuum solvent and rigid-body based Brownian dynamics simulations using the SDA 7 software. A number of simulation parameters were accessed using the association of trypsin and benzamidine as a test system and an optimized set of parameters was derived that should be generally applicable to simulating diffusional association of a wide-range of protein-ligand binding pairs. I also established standard guidelines for recording diffusional on-rates corresponding to specific reaction conditions. I validated this protocol on several inhibitors of different targets of varying complexities. I observed that the protocol had limitations in explaining the binding of small but flexible molecules, as well as conformation dependent protein-ligand binding. However, this protocol can serve as a good starting point for multi-scale approaches that combine BD with MD, where BD simulations can be run using this protocol to model the initial diffusional association of protein and ligand molecules, following which an MD-based regime can be used to account for the flexibility and conformational changes to simulate the formation of the final bound complex. The association rates from these two different approaches can be combined using specialized techniques such as milestoning, as demonstrated by Votapka *et al.*[22] using their software SEEKR.

In addition, I contributed to the development of KBbox, a toolbox of computational methods, which provides access to information on state-of-the-art computational methods to study molecular binding kinetics, and example cases and tutorials for these methods. KBbox also includes a collection of tutorials that provide the users with an introduction into how to use different computational approaches to compute the kinetic parameters of protein-ligand binding. Due to the growing interest in evaluation of drug-binding kinetics, a plethora of computational methods based on biomolecular simulations and chemoinformatics has emerged recently that are designed to compute either k_{on} or k_{off} or both. Some of these methods can provide absolute values of k_{on} and/or k_{off} whereas others can be used to get relative rates or they can rank or classify ligands according to their binding kinetics.

Also, depending upon the complexity of the protein-ligand system being studied, the assumptions of the methods, and the amount of data available on the system, the computational requirement and accuracy of these methods vary to a great extent. Therefore, the choice of an appropriate method for a specific problem is non-trivial. To help the users in choosing appropriate methods, KBbox provides a query interface that asks users a series of questions related to the amount of structural and kinetics data available, data that users wish to calculate, and suggests them a list of methods for their calculations, and these methods are sorted by the level of computational resources required for their application. Also, KBbox is designed to be easily extendable, so that the data on the newly developed methods and their published examples can be added. We therefore believe that it will be useful to continuously maintain, and regularly update this toolbox, and that will help the researchers with an interest in studying drug-binding kinetics to use different state-of-the-art computational methods for their system of interest.

Despite a lot of progress being made in the development and application of computational approaches to compute binding kinetic parameters, there are still many different challenges for computing k_{on} and k_{off} rates. For computing k_{on} rates accurately, meaningful encounter states must be found. On the other hand, for correct evaluation of k_{off} rates, a method should be able to effectively capture the factors determining the escape of a tightly-bound ligand and the associated transition barrier. In addition, computational methods make a critical assumption that the force fields used are able to fully represent binding and unbinding processes for computing drug-binding kinetics. However, the employed molecular force-fields are parameterized to reproduce the equilibrium populations of free-energy wells, rather than the transition barrier heights between them. Therefore, it is important to recognize the shortcomings in commonly-used force fields so that future force fields can provide improved representations of barrier heights. Similarly, water models used in simulation approaches also need to be significantly improved so that the computational methods are able to reproduce diffusionally-limited kinetic data. At present, there is no method available that can calculate absolute values of both k_{on} and k_{off} accurately in the same computational framework with only modest computational resources. Since the development of computational methods to compute drug-binding kinetics is a very active area of research, one can expect further testing, refinement and validation of these methods, as well as new approaches in the next few years. More-

over, the improvement in accuracy of computational methods will contribute to a thorough understanding of ligand–receptor structure–kinetics relationships.

Bibliography

- [1] Marco De Vivo, Matteo Masetti, Giovanni Bottegoni, and Andrea Cavalli. Role of molecular dynamics and related methods in drug discovery. *Journal of Medicinal Chemistry*, 59(9):4035–4061, 2016.
- [2] Robert A Copeland, David L Pompliano, and Thomas D Meek. Drug–target residence time and its implications for lead optimization. *Nature Reviews Drug discovery*, 5(9):730, 2006.
- [3] Doris A. Schuetz, Wilhelmus Egbertus Arnout de Witte, Yin Cheong Wong, Bernhard Knasmueller, Lars Richter, Daria B. Kokh, S. Kashif Sadiq, Reggie Bosma, Indira Nederpelt, Elena Segala, Marta Amaral, Dong Guo, Dorothee Andres, Victoria Georgi, Leigh A. Stoddart, Steve Hill, Robert M. Cooke, Chris De Graaf, Rob Leurs, Matthias Frech, Rebecca C. Wade, Elizabeth Cunera Maria de Lange, Adriaan P. IJzerman, Anke Müller-Fahrnow, and Gerhard F. Ecker. Kinetics for Drug Discovery: an industry-driven effort to target drug residence time. *Drug Discovery Today*, 6446(17), 2017.
- [4] Julia Romanowska, Daria B Kokh, Jonathan C Fuller, and Rebecca C Wade. *Computational Approaches for Studying Drug Binding Kinetics*, chapter 11, pages 211–235. John Wiley & Sons, Ltd, 2015.
- [5] Razif R. Gabdouliline and Rebecca C. Wade. On the protein–protein diffusional encounter complex. *Journal of Molecular Recognition*, 12(4):226–234, 1999.
- [6] Razif R. Gabdouliline and Rebecca C. Wade. Biomolecular diffusional association. *Current Opinion in Structural Biology*, 12(2):204 – 213, 2002.
- [7] Adrian H Elcock, Razif R Gabdouliline, Rebecca C Wade, and J.Andrew McCammon. Computer simulation of protein-protein association kinetics:

- acetylcholinesterase-fasciculin11 edited by b. honig. *Journal of Molecular Biology*, 291(1):149 – 162, 1999.
- [8] Hao Lu and Peter J Tonge. Drug–target residence time: critical information for lead optimization. *Current Opinion in Chemical Biology*, 14(4):467 – 474, 2010. Next Generation Therapeutics.
- [9] Neil J Bruce, Gaurav K Ganotra, Daria B Kokh, S Kashif Sadiq, and Rebecca C Wade. New approaches for computing ligand–receptor binding kinetics. *Current Opinion in Structural Biology*, 49(Supplement C):1–10, 2018.
- [10] Samia Aci-Sèche, Sonia Ziada, Abdennour Braka, Rohit Arora, and Pascal Bonnet. Advanced molecular dynamics simulation methods for kinase drug discovery. *Future Medicinal Chemistry*, 8(5):545–566, 2016.
- [11] Andrea Cavalli, Andrea Spitaleri, Giorgio Saladino, and Francesco L Gervasio. Investigating drug–target association and dissociation mechanisms using metadynamics-based algorithms. *Accounts of Chemical Research*, 48(2):277–285, 2014.
- [12] Lillian T Chong, Ali S Saglam, and Daniel M Zuckerman. Path-sampling strategies for simulating rare events in biomolecular systems. *Current Opinion in Structural Biology*, 43:88–94, 2017.
- [13] Noelia Ferruz and Gianni De Fabritiis. Binding kinetics in drug discovery. *Molecular Informatics*, 35(6-7):216–226, 2016.
- [14] Xiaodong Pang and Huan-Xiang Zhou. Rate constants and mechanisms of protein–ligand binding. *Annual Review of Biophysics*, 46:105–130, 2017.
- [15] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National Academy of Sciences*, 108(25):10184–10189, 2011.
- [16] Nuria Plattner and Frank Noé. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models. *Nature Communications*, 6:7653, 2015.

- [17] S Doerr and G De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *Journal of Chemical Theory and Computation*, 10(5):2064–2069, 2014.
- [18] Pratyush Tiwary, Vittorio Limongelli, Matteo Salvalaglio, and Michele Parrinello. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National Academy of Sciences*, 112(5):E386–E391, 2015.
- [19] Ivan Teo, Christopher G Mayne, Klaus Schulten, and Tony Lelièvre. Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine–trypsin dissociation time. *Journal of Chemical Theory and Computation*, 12(6):2983–2989, 2016.
- [20] Alex Dickson and Samuel D Lotz. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by wexplore. *Biophysical Journal*, 112(4):620–629, 2017.
- [21] Jeffrey C Sung, Adam W Van Wynsberghe, Rommie E Amaro, Wilfred W Li, and J Andrew McCammon. Role of secondary sialic acid binding sites in influenza n1 neuraminidase. *Journal of the American Chemical Society*, 132(9):2883–2885, 2010.
- [22] Lane W Votapka, Benjamin R Jagger, Alexandra L Heyneman, and Rommie E Amaro. Seekr: simulation enabled estimation of kinetic rates, a computational tool to estimate molecular kinetics and its application to trypsin–benzamidine binding. *The Journal of Physical Chemistry B*, 121(15):3597–3606, 2017.
- [23] Pratyush Tiwary, Jagannath Mondal, and Bruce J Berne. How and when does an anticancer drug leave its binding site? *Science Advances*, 3(5):e1700014, 2017.
- [24] Rodrigo Casasnovas, Vittorio Limongelli, Pratyush Tiwary, Paolo Carloni, and Michele Parrinello. Unbinding kinetics of a p38 map kinase type ii inhibitor from metadynamics simulations. *Journal of the American Chemical Society*, 139(13):4780–4788, 2017.
- [25] Donatella Callegari, Alessio Lodola, Daniele Pala, Silvia Rivara, Marco Mor, Andrea Rizzi, and Anna Maria Capelli. Metadynamics simulations distinguish

- short-and long-residence-time inhibitors of cyclin-dependent kinase 8. *Journal of Chemical Information and Modeling*, 57(2):159–169, 2017.
- [26] Luca Mollica, Sergio Decherchi, Syeda Rehana Zia, Roberto Gaspari, Andrea Cavalli, and Walter Rocchia. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Scientific Reports*, 5:11539, 2015.
- [27] Yuzhen Niu, Shuyan Li, Dabo Pan, Huanxiang Liu, and Xiaojun 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. *Physical Chemistry Chemical Physics*, 18(7):5622–5629, 2016.
- [28] Susanna K Lüdemann, Valère Lounnas, and Rebecca 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. *Journal of Molecular Biology*, 303(5):797–811, 2000.
- [29] Daria B Kokh, Marta Amaral, Joerg Bomke, Ulrich Gradler, Djordje Musil, Hans-Peter Buchstaller, Matthias K Dreyer, Matthias Frech, Maryse Lowinski, Francois Vallee, et al. Estimation of drug-target residence times by τ -random acceleration molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 14(7):3859–3869, 2018.
- [30] Robert M Cooke, Alastair JH Brown, Fiona H Marshall, and Jonathan S Mason. Structures of g protein-coupled receptors reveal new opportunities for drug discovery. *Drug Discovery Today*, 20(11):1355–1364, 2015.
- [31] Christofer S Tautermann. Impact, determination and prediction of drug-receptor residence times for gpcrs. *Current Opinion in Pharmacology*, 30:22–26, 2016.
- [32] Ron O Dror, Albert C Pan, Daniel H Arlow, David W Borhani, Paul Maragakis, Yibing Shan, Huafeng Xu, and David E Shaw. Pathway and mechanism of drug binding to g-protein-coupled receptors. *Proceedings of the National Academy of Sciences*, 108(32):13118–13123, 2011.

- [33] Andrea Bortolato, Francesca Deflorian, Dahlia R Weiss, and Jonathan S Mason. Decoding the role of water dynamics in ligand–protein unbinding: Crf1r as a test case. *Journal of Chemical Information and Modeling*, 55(9):1857–1866, 2015.
- [34] Massimo Marchi and Pietro Ballone. Adiabatic bias molecular dynamics: a method to navigate the conformational space of complex molecular systems. *The Journal of Chemical Physics*, 110(8):3697–3702, 1999.
- [35] Fabio Pietrucci, Fabrizio Marinelli, Paolo Carloni, and Alessandro Laio. Substrate binding mechanism of hiv-1 protease from explicit-solvent atomistic simulations. *Journal of the American Chemical Society*, 131(33):11811–11818, 2009.
- [36] Matthew C Zwier, Adam J Pratt, Joshua L Adelman, Joseph W Kaus, Daniel M Zuckerman, and Lillian T Chong. Efficient atomistic simulation of pathways and calculation of rate constants for a protein–peptide binding process: application to the mdm2 protein and an intrinsically disordered p53 peptide. *The Journal of Physical Chemistry Letters*, 7(17):3440–3445, 2016.
- [37] Oliver Schon, Assaf Friedler, Mark Bycroft, Stefan MV Freund, and Alan R Fersht. Molecular mechanism of the interaction between mdm2 and p53. *Journal of Molecular Biology*, 323(3):491–501, 2002.
- [38] Guangfeng Zhou, George A Pantelopulos, Sudipto Mukherjee, and Vincent A Voelz. Bridging microscopic and macroscopic mechanisms of p53-mdm2 binding with kinetic network models. *Biophysical Journal*, 113(4):785–793, 2017.
- [39] Suresh Karthik and Sanjib Senapati. Dynamic flaps in hiv-1 protease adopt unique ordering at different stages in the catalytic cycle. *Proteins: Structure, Function, and Bioinformatics*, 79(6):1830–1840, 2011.
- [40] Yu-ming M Huang, Mark Anthony V Raymundo, Wei Chen, and Chia-en A Chang. Mechanism of the association pathways for a pair of fast and slow binding ligands of hiv-1 protease. *Biochemistry*, 56(9):1311–1323, 2017.
- [41] Alexander Wlodawer and Jiri Vondrasek. Inhibitors of hiv-1 protease: a major success of structure-assisted drug design. *Annual Review of Biophysics and Biomolecular Structure*, 27(1):249–284, 1998.

- [42] Eric S Furfine, Eric D’Souza, Kenneth J Ingold, Johann J Leban, Thomas Spector, and David JT Porter. Two-step binding mechanism for hiv protease inhibitors. *Biochemistry*, 31(34):7886–7891, 1992.
- [43] Huiyong Sun, Youyong Li, Mingyun Shen, Dan Li, Yu Kang, and Tingjun Hou. Characterizing drug–target residence time with metadynamics: how to achieve dissociation rate efficiently without losing accuracy against time-consuming approaches. *Journal of Chemical Information and Modeling*, 57(8):1895–1906, 2017.
- [44] Sujun Qu, Shuheng Huang, Xianchao Pan, Li Yang, and Hu Mei. Constructing interconsistent, reasonable, and predictive models for both the kinetic and thermodynamic properties of hiv-1 protease inhibitors. *Journal of Chemical Information and Modeling*, 56(10):2061–2068, 2016.
- [45] See Hong Chiu and Lei Xie. Toward high-throughput predictive modeling of protein binding/unbinding kinetics. *Journal of Chemical Information and Modeling*, 56(6):1164–1174, 2016.
- [46] A R Ortiz, M T Pisabarro, F Gago, and R C Wade. Prediction of drug binding affinities by comparative binding energy analysis. *Journal of Medicinal Chemistry*, 38:2681–2691, 1995.
- [47] Svante Wold, Arnold Ruhe, Herman Wold, and WJ Dunn, III. The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743, 1984.
- [48] Svante Wold, Michael Sjöström, and Lennart Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.
- [49] George EP Box, J Stuart Hunter, and William Gordon Hunter. *Statistics for experimenters: design, innovation, and discovery*, volume 2. Wiley-Interscience New York, 2005.
- [50] Albert Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560, 1905.

- [51] Marian Von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der Physik*, 326(14):756–780, 1906.
- [52] Donald L Ermak and J A McCammon. Brownian dynamics with hydrodynamic interactions. *The Journal of Chemical Physics*, 69(4):1352–1360, 1978.
- [53] Razif R Gabdoulline and Rebecca C Wade. Brownian dynamics simulation of protein–protein diffusional encounter. *Methods*, 14(3):329–341, 1998.
- [54] Michael Martinez, Neil J Bruce, Julia Romanowska, Daria B Kokh, Musa Ozboyaci, Xiaofeng Yu, Mehmet Ali Öztürk, Stefan Richter, and Rebecca C Wade. SDA 7: A modular and parallel implementation of the simulation of diffusional association software. *Journal of Computational Chemistry*, 36(21):1631–1645, 2015.
- [55] Razif R Gabdoulline and Rebecca C Wade. On the contributions of diffusion and thermal activation to electron transfer between phormidium laminosum plastocyanin and cytochrome f: Brownian dynamics simulations with explicit modeling of nonpolar desolvation interactions and electron transfer events. *Journal of the American Chemical Society*, 131(26):9230–9238, 2009.
- [56] Paolo Mereghetti, Razif R Gabdoulline, and Rebecca C Wade. Brownian dynamics simulation of protein solutions: structural and dynamical properties. *Biophysical Journal*, 99(11):3782–3791, 2010.
- [57] Nathan A Baker, David Sept, Simpson Joseph, Michael J Holst, and J Andrew McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, 98(18):10037–10041, 2001.
- [58] Malcolm E Davis, Jeffrey D Madura, Brock A Luty, and J Andrew McCammon. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian dynamics program. *Computer Physics Communications*, 62(2-3):187–197, 1991.
- [59] Razif R Gabdoulline and Rebecca C Wade. Effective charges for macromolecules in solvent. *The Journal of Physical Chemistry*, 100(9):3868–3878, 1996.

- [60] M v Smoluchowski. Versuch einer mathematischen theorie der koagulationskinetik kolloider lösungen. *Zeitschrift für Physikalische Chemie*, 92(1):129–168, 1918.
- [61] Scott H Northrup, Stuart A Allison, and J Andrew McCammon. Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *The Journal of Chemical Physics*, 80(4):1517–1524, 1984.
- [62] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, 1977.
- [63] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159:98–103, 1967.
- [64] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.
- [65] Thomas A Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996.
- [66] 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. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.
- [67] William G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.
- [68] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981.
- [69] Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable

- potential functions. *The Journal of Chemical Physics*, 112(20):8910–8922, 2000.
- [70] H J C Berendsen, J R Grigera, and T P Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, 1987.
- [71] Vickie Tsui and David A Case. Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers: Original Research on Biomolecules*, 56(4):275–291, 2000.
- [72] Peter A Kollman, Irina Massova, Carolina Reyes, Bernd Kuhn, Shuanghong Huo, Lillian Chong, Matthew Lee, Taisung Lee, Yong Duan, Wei Wang, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–897, 2000.
- [73] Tingjun Hou, Junmei Wang, Youyong Li, and Wei Wang. Assessing the performance of the mm/pbsa and mm/gbsa methods. 1. the accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82, 2010.
- [74] Chr. Møller and M. S. Plesset. Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618–622, 1934.
- [75] Martin Head-Gordon, John A. Pople, and Michael J. Frisch. Mp2 energy evaluation by direct methods. *Chemical Physics Letters*, 153(6):503 – 506, 1988.
- [76] Michael W Schmidt, Kim K Baldridge, Jerry A Boatz, Steven T Elbert, Mark S Gordon, Jan H Jensen, Shiro Koseki, Nikita Matsunaga, Kiet A Nguyen, Shunjun Su, et al. General atomic and molecular electronic structure system. *Journal of Computational Chemistry*, 14(11):1347–1363, 1993.
- [77] Robert W Gora, Wojciech Bartkowiak, Szczepan Roszak, and Jerzy Leszczynski. A new theoretical insight into the nature of intermolecular interactions in the molecular crystal of urea. *The Journal of Chemical Physics*, 117(3):1031–1039, 2002.

- [78] Robert W Gora, W Andrzej Sokalski, Jerzy Leszczynski, and Virginia B Pett. The nature of interactions in the ionic crystal of 3-pentenenitrile, 2-nitro-5-oxo, ion (- 1), sodium. *The Journal of Physical Chemistry B*, 109(5):2027–2033, 2005.
- [79] X. Wu D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. and P.A. Kollman. “AMBER 14”, University of California, San Francisco, 2014.
- [80] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.
- [81] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E Felberg, David H Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W Gohara, Todd Dolinsky, Robert Konecny, David R Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J Holst, J Andrew McCammon, and Nathan A Baker. Improvements to the APBS biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018.
- [82] A Ortega, D Amorós, and J García de La Torre. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic-and residue-level models. *Biophysical Journal*, 101(4):892–898, 2011.
- [83] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.
- [84] Jitender Verma, Vijay M. Khedkar, and Evans C. Coutinho. 3D-QSAR in Drug Design - A Review. *Current Topics in Medicinal Chemistry*, 10(1):95–115, 2010.
- [85] Harun M. Patel, Malleshappa N. Noolvi, Poonam Sharma, Varun Jaiswal, Sumit Bansal, Sandeep Lohan, Suthar Sharad Kumar, Vikrant Abbot,

- Saurabh Dhiman, and Varun Bhardwaj. Quantitative structure-activity relationship (QSAR) studies as strategic approach in drug discovery. *Medicinal Chemistry Research*, 23(12):4991–5007, 2014.
- [86] Richard D. Cramer, David E. Patterson, and Jeffrey D. Bunce. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *Journal of the American Chemical Society*, 110(18):5959–5967, 1988.
- [87] Gabriele Cruciani and Kimberly A. Watson. Comparative Molecular Field Analysis Using GRID Force-Field and GOLPE Variable Selection Methods in a Study of Inhibitors of Glycogen Phosphorylase b. *Journal of Medicinal Chemistry*, 37(16):2589–2601, 1994.
- [88] Angel R Ortiz, Manuel Pastor, Albert Palomer, Gabriele Cruciani, Federico Gago, and Rebecca C Wade. Reliability of comparative molecular field analysis models: effects of data scaling and variable selection using a set of human synovial fluid phospholipase a2 inhibitors. *Journal of Medicinal Chemistry*, 40(7):1136–1148, 1997.
- [89] Comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) studies on α 1A-adrenergic receptor antagonists based on pharmacophore molecular alignment. *International Journal of Molecular Sciences*, 12(10):7022–7037, 2011.
- [90] Manuel Pastor, Carlos Pérez, and Federico Gago. Simulation of alternative binding modes in a structure-based QSAR study of HIV-1 protease inhibitors. *Journal of Molecular Graphics and Modelling*, 15(6):364–371, 1997.
- [91] Carlos Pérez, Manuel Pastor, Angel R. Ortiz, and Federico Gago. Comparative binding energy analysis of HIV-1 protease inhibitors: Incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *Journal of Medicinal Chemistry*, 41(6):836–852, 1998.
- [92] Juan José Lozano, Manuel Pastor, Gabriele Cruciani, Katrin Gaedt, Nuria B. Centeno, Federico Gago, and Ferran Sanz. 3D-QSAR methods on the basis of ligand-receptor complexes. Application of COMBINE and GRID/GOLPE

- methodologies to a series of CYP1A2 ligands. *Journal of Computer-Aided Molecular Design*, 14(4):341–353, 2000.
- [93] T. Wang and R. C. Wade. Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. *Journal of Medicinal Chemistry*, 44(6):961–971, 2001.
- [94] J. Kmuníček, S. Luengo, F. Gago, A. R. Ortiz, R. C. Wade, and J. Damborský. Comparative binding energy analysis of the substrate specificity of haloalkane dehalogenase from *Xanthobacter autotrophicus* GJ10. *Biochemistry*, 40(30):8905–8917, 2001.
- [95] Ting Wang and Rebecca C. Wade. Comparative Binding Energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. *Journal of Medicinal Chemistry*, 45(22):4828–4837, 2002.
- [96] Gerhard Klebe, Ute Abraham, and Thomas Mietzner. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *Journal of Medicinal Chemistry*, 37(24):4130–4146, 1994.
- [97] M Amaral, D B Kokh, J Bomke, A Wegener, H P Buchstaller, H M Eggenweiler, P Matias, C Sirrenberg, R C Wade, and M Frech. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. *Nature Communications*, 8(1):2276, 2017.
- [98] Len Neckers, Edward Mimnaugh, and Theodor W Schulte. Hsp90 as an anti-cancer target. *Drug Resistance Updates*, 2(3):165–172, 1999.
- [99] Bonnie Tillotson, Kelly Slocum, John Coco, Nigel Whitebread, Brian Thomas, Kip A West, John MacDougall, Jie Ge, Janid A Ali, Vito J Palombella, Emmanuel Normant, Julian Adams, and Christian C Fritz. Hsp90 (Heat Shock Protein 90) Inhibitor Occupancy Is a Direct Determinant of Client Protein Degradation and Tumor Growth Arrest in Vivo. *The Journal of Biological Chemistry*, 285(51):39835–39843, 2010.
- [100] CHRISTINE DEBOUCK. The HIV-1 Protease as a Therapeutic Target for AIDS. *AIDS Research and Human Retroviruses*, 8(2):153–164, 1992.

- [101] Patrick Y. S. Lam, Yu Ru, Prabhakar K. Jadhav, Paul E. Aldrich, George V. DeLuca, Charles J. Eyermann, Chong-Hwan Chang, George Emmett, Edward R. Holler, Wayne F. Daneker, Liangzhu Li, Pat N. Confalone, Robert J. McHugh, Qi Han, Renhua Li, Jay A. Markwalder, Steven P. Seitz, Thomas R. Sharpe, Lee T. Bacheler, Marlene M. Rayner, Ronald M. Klabe, Linyee Shum, Dean L. Winslow, David M. Kornhauser, David A. Jackson, Susan Erickson-Viitanen, and C. Nicholas Hodge. Cyclic hiv protease inhibitors: synthesis, conformational analysis, p2/p2' structureactivity relationship, and molecular recognition of cyclic ureas. *Journal of Medicinal Chemistry*, 39(18):3514–3525, 1996. PMID: 8784449.
- [102] D J Kempf, K C Marsh, J F Denissen, E McDonald, S Vasavanonda, C A Flentge, B E Green, L Fino, C H Park, and X P Kong. Abt-538 is a potent inhibitor of human immunodeficiency virus protease and has high oral bioavailability in humans. *Proceedings of the National Academy of Sciences*, 92(7):2484–2488, 1995.
- [103] Irene T Weber and Yuan-Fang Wang. HIV-1 Protease: Role in Viral Replication, Protein–Ligand X-Ray Crystal Structures and Inhibitor Design. In *Aspartic Acid Proteases as Therapeutic Targets*, pages 107–137. Wiley-VCH Verlag GmbH & Co. KGaA, 2010.
- [104] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimaju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3):221–234, 2013.
- [105] Chresten R Søndergaard, Mats H M Olsson, Michał Rostkowski, and Jan H Jensen. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *Journal of Chemical Theory and Computation*, 7(7):2284–2295, 2011.
- [106] François-Yves Dupradeau, Adrien Pigache, Thomas Zaffran, Corentin Savineau, Rodolphe Lelong, Nicolas Grivel, Dimitri Lelong, Wilfried Rosanski, and Piotr Cieplak. The R.E.D. Tools: Advances in RESP and ESP charge derivation and force field library building, 2010.

- [107] James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.
- [108] Rubén Gil-Redondo, Javier Klett, Federico Gago, and Antonio Morreale. gCOMBINE: A graphical user interface to perform structure-based comparative binding energy (COMBINE) analysis on a set of ligand-receptor complexes. *Proteins: Structure, Function, and Bioinformatics*, 78(1):162–172, 2010.
- [109] Gaurav K Ganotra and Rebecca C Wade. Prediction of drug–target binding kinetics by comparative binding energy analysis. *ACS Medicinal Chemistry Letters*, 9(11):1134–1139, 2018.
- [110] Robert A. Copeland, David L. Pompliano, and Thomas D. Meek. Drug–target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery*, 5(9):730–739, 2006.
- [111] Philip Cohen. Protein kinases—the major drug targets of the twenty-first century? *Nature Reviews Drug Discovery*, 1(4):309, 2002.
- [112] Tudor I Oprea, Cristian G Bologa, Søren Brunak, Allen Campbell, Gregory N Gan, Anna Gaulton, Shawn M Gomez, Rajarshi Guha, Anne Hersey, and Jayme Holmes. Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery*, 17(5):317, 2018.
- [113] Christopher Pargellis, Liang Tong, Laurie Churchill, Pier F. Cirillo, Thomas Gilmore, Anne G. Graham, Peter M. Grob, Eugene R. Hickey, Neil Moss, Susan Pav, and John Regan. Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nature Structural Biology*, 9(4):268–272, 2002.
- [114] Nicole Willemsen-Seegers, Joost C M Uitdehaag, Martine B W Prinsen, Judith R F de Vetter, Jos de Man, Masaaki Sawa, Yusuke Kawase, Rogier C Buijsman, and Guido J R Zaman. Compound selectivity and target residence time of kinase inhibitors studied with surface plasmon resonance. *Journal of Molecular Biology*, 429(4):574–586, 2017.

- [115] E. V. Schneider, J. Bottcher, R. Huber, K. Maskos, and L. Neumann. Structure-kinetic relationship study of CDK8/CycC specific compounds. *Proceedings of the National Academy of Sciences*, 110(20):8081–8086, 2013.
- [116] Pelin Ayaz, Dorothee Andres, Dennis A Kwiatkowski, Carl-Christian Kolbe, Philip Lienau, Gerhard Siemeister, Ulrich Lucking, and Christian M Stegmann. Conformational adaption may explain the slow dissociation kinetics of roniciclib (BAY 1000394), a type I CDK inhibitor with kinetic selectivity for CDK2 and CDK9. *ACS Chemical Biology*, 11(6):1710–1719, 2016.
- [117] Peter Schmidtke, F Javier Luque, James B Murray, and Xavier Barril. Shielded hydrogen bonds as structural determinants of binding kinetics: application in drug design. *Journal of the American Chemical Society*, 133(46):18903–18910, 2011.
- [118] Heroven Christina, Georgi Victoria, Ganotra Gaurav K., Brennan Paul, Wolfreys Finn, Wade Rebecca C., Fernández-Montalván Amaury E., Chaikuad Apirat, and Knapp Stefan. Halogen–Aromatic π Interactions Modulate Inhibitor Residence Times. *Angewandte Chemie International Edition*, 57(24):7220–7224, 2018.
- [119] Pascal Auffinger, Franklin A Hays, Eric Westhof, and P Shing Ho. Halogen bonds in biological molecules. *Proceedings of the National Academy of Sciences*, 101(48):16789–16794, 2004.
- [120] Matthew R Scholfield, Crystal M Vander Zanden, Megan Carter, and P Shing Ho. Halogen bonding (X-bonding): A biological perspective. *Protein Science*, 22(2):139–152, 2013.
- [121] Tore Brinck, Jane S Murray, and Peter Politzer. Surface electrostatic potentials of halogenated methanes as indicators of directional intermolecular interactions. *International Journal of Quantum Chemistry*, 44(S19):57–64, 1992.
- [122] Timothy Clark, Matthias Hennemann, Jane S Murray, and Peter Politzer. Halogen bonding: the σ -hole. *Journal of Molecular Modeling*, 13(2):291–296, 2007.

- [123] Edward Harder, Wolfgang Damm, Jon Maple, Chuanjie Wu, Mark Reboul, Jin Yu Xiang, Lingle Wang, Dmitry Lupyan, Markus K Dahlgren, Jennifer L Knight, et al. Opls3: a force field providing broad coverage of drug-like small molecules and proteins. *Journal of Chemical Theory and Computation*, 12(1):281–296, 2015.
- [124] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.
- [125] Mahmoud A A Ibrahim. Molecular mechanical study of halogen bonding in drug discovery. *Journal of Computational Chemistry*, 32(12):2564–2574, 2011.
- [126] William L Jorgensen, Jayaraman Chandrasekhar, Jeffrey D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.
- [127] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An $n \log(n)$ method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.
- [128] Araz Jakalian, Bruce L Bush, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *Journal of Computational Chemistry*, 21(2):132–146, 2000.
- [129] Araz Jakalian, David B Jack, and Christopher I Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: II. parameterization and validation. *Journal of Computational Chemistry*, 23(16):1623–1641, 2002.
- [130] Zhijian Xu, Zhuo Yang, Yingtao Liu, Yunxiang Lu, Kaixian Chen, and Weiliang Zhu. Halogen bond: its role beyond drug–target binding affinity for drug discovery and development. *Journal of chemical information and modeling*, 54(1):69–78, 2014.
- [131] R R Gabdouliline and R C Wade. Simulation of the diffusional association of barnase and barstar. *Biophysical Journal*, 72(5):1917–1929, 1997.

- [132] Razif R Gabdoulhine and Rebecca C Wade. Protein-protein association: investigation of factors influencing association rates by Brownian dynamics simulations¹¹Edited by B. Honig. *Journal of Molecular Biology*, 306(5):1139–1155, 2001.
- [133] F De Rienzo, R R Gabdoulhine, M C Menziani, P G De Benedetti, and R C Wade. Electrostatic analysis and Brownian dynamics simulation of the association of plastocyanin and cytochrome f. *Biophysical Journal*, 81(6):3090–3104, 2001.
- [134] Michal Harel, Alexander Spaar, and Gideon Schreiber. Fruitful and futile encounters along the association reaction between proteins. *Biophysical Journal*, 96(10):4237–4248, 2009.
- [135] MARKUS Marquart, JOCHEN Walter, JOHANN Deisenhofer, WOLFRAM Bode, and ROBERT Huber. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallographica Section B: Structural Science*, 39(4):480–490, 1983.
- [136] Susanne Roehrig, Alexander Straub, Jens Pohlmann, Thomas Lampe, Josef Pernerstorfer, Karl-Heinz Schlemmer, Peter Reinemer, and Elisabeth Perzborn. Discovery of the novel antithrombotic agent 5-chloro-N-({(5 S)-2-oxo-3-[4-(3-oxomorpholin-4-yl) phenyl]-1, 3-oxazolidin-5-yl} methyl) thiophene-2-carboxamide (BAY 59-7939): an oral, direct factor Xa inhibitor. *Journal of Medicinal Chemistry*, 48(19):5900–5908, 2005.
- [137] Elisabeth Perzborn, Susanne Roehrig, Alexander Straub, Dagmar Kubitza, and Frank Misselwitz. The discovery and development of rivaroxaban, an oral, direct factor Xa inhibitor. *Nature Reviews Drug Discovery*, 10(1):61, 2011.
- [138] Sangeeta Bhanwra and Kaza Ahluwalia. The new factor Xa inhibitor: apixaban. *Journal of Pharmacology & Pharmacotherapeutics*, 5(1):12, 2014.
- [139] Jona Lekura and James S Kalus. Overview of betrixaban and its role in clinical practice. *The Bulletin of the American Society of Hospital Pharmacists*, 75(15):1095–1102, 2018.

- [140] Henri Bounameaux and A John Camm. Edoxaban: an update on the new oral direct factor Xa inhibitor. *Drugs*, 74(11):1209–1231, 2014.
- [141] David Bergqvist. Review of fondaparinux sodium injection for the prevention of venous thromboembolism in patients undergoing surgery. *Vascular Health and Risk Management*, 2(4):365, 2006.
- [142] X Zhou, DRH Huntjens, and RAHJ Gilissen. A systems pharmacology model for predicting effects of factor xa inhibitors in healthy subjects: assessment of pharmacokinetics and binding kinetics. *CPT: pharmacometrics & systems pharmacology*, 4(11):650–659, 2015.
- [143] Herbert Nar. The role of structural information in the discovery of direct thrombin and factor Xa inhibitors. *Trends in Pharmacological Sciences*, 33(5):279–288, 2012.
- [144] Marc Nazaré, David W Will, Hans Matter, Herman Schreuder, Kurt Ritter, Matthias Urmann, Melanie Essrich, Armin Bauer, Michael Wagner, and Jörg Czech. Probing the subpockets of factor Xa reveals two binding modes for inhibitors based on a 2-carboxyindole scaffold: a study combining structure-activity relationship and X-ray crystallography. *Journal of Medicinal Chemistry*, 48(14):4511–4525, 2005.
- [145] Elisabeth Perzborn, Susanne Roehrig, Alexander Straub, Dagmar Kubitzka, Wolfgang Mueck, and Volker Laux. Rivaroxaban: a new oral factor xa inhibitor. *Arteriosclerosis, thrombosis, and vascular biology*, 30(3):376–381, 2010.
- [146] Jeyanthi Eswaran, Debasis Patnaik, Panagis Filippakopoulos, Fangwei Wang, Ross L Stein, James W Murray, Jonathan M G Higgins, and Stefan Knapp. Structure and functional characterization of the atypical human kinase haspin. *Proceedings of the National Academy of Sciences*, 106(48):20198–20203, 2009.
- [147] Fabrizio Villa, Paola Capasso, Marcello Tortorici, Federico Forneris, Ario de Marco, Andrea Mattevi, and Andrea Musacchio. Crystal structure of the catalytic domain of Haspin, an atypical kinase implicated in chromatin organization. *Proceedings of the National Academy of Sciences*, 106(48):20204–20209, 2009.

- [148] Jun Dai, Sammy Sultan, Stephen S Taylor, and Jonathan M G Higgins. The kinase haspin is required for mitotic histone H3 Thr 3 phosphorylation and normal metaphase chromosome alignment. *Genes & Development*, 19(4):472–488, 2005.
- [149] Morgan Huse and John Kuriyan. The conformational plasticity of protein kinases. *Cell*, 109(3):275–282, 2002.
- [150] Louise N Johnson, Martin E M Noble, and David J Owen. Active and inactive protein kinases: structural basis for regulation. *Cell*, 85(2):149–158, 1996.
- [151] J N Varghese, W G Laver, and Peter M Colman. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature*, 303(5912):35, 1983.
- [152] F G Blix, A Gottschalk, and E Klenk. Proposed nomenclature in the field of neuraminic and sialic acids. *Nature*, 179(4569):1088, 1957.
- [153] Rupert J Russell, Lesley F Haire, David J Stevens, Patrick J Collins, Yi Pu Lin, G Michael Blackburn, Alan J Hay, Steven J Gamblin, and John J Skehel. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature*, 443(7107):45, 2006.
- [154] Christopher J Vavricka, Qing Li, Yan Wu, Jianxun Qi, Mingyang Wang, Yue Liu, Feng Gao, Jun Liu, Enguang Feng, Jianhua He, et al. Structural and functional analysis of laninamivir and its octanoate prodrug reveals group specific mechanisms for influenza na inhibition. *PLoS pathogens*, 7(10):e1002249, 2011.
- [155] Matthew P Jacobson, David L Pincus, Chaya S Rapp, Tyler J F Day, Barry Honig, David E Shaw, and Richard A Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004.
- [156] Todd J Dolinsky, Paul Czodrowski, Hui Li, Jens E Nielsen, Jan H Jensen, Gerhard Klebe, and Nathan A Baker. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(suppl_2):W522–W525, 2007.

- [157] Florent Guillain and Darwin Thusius. The Use of Proflavin as an Indicator in Temperature-Jump Studies of the Binding of a Competitive Inhibitor to Trypsin. *Journal of American Chemical Society*, 92(9):5534–5536, 1970.
- [158] G Jourdi, V Siguret, A C Martin, J L Golmard, A Godier, C M Samama, P Gaussem, I Gouin-Thibault, and B Le Bonniec. Association rate constants rationalise the pharmacodynamics of apixaban and rivaroxaban. *Thrombosis and Haemostasis*, 114(1):78–86, 2015.
- [159] Patrick J. Collins, Lesley F. Haire, Yi Pu Lin, Junfeng Liu, Rupert J. Russell, Philip A. Walker, John J. Skehel, Stephen R. Martin, Alan J. Hay, and Steven J. Gamblin. Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. *Nature*, 453(7199):1258–1261, 2008.
- [160] R R Gabdouliline and R C Wade. Effective Charges for Macromolecules in Solvent. *The Journal of Physical Chemistry*, 100(9):3868–3878, 1996.
- [161] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2017.
- [162] Mark Otto. Bootstrap from twitter. *Developer Blog. Twitter.*, 19 August, 2011.
- [163] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 03 2009.
- [164] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 41(D1):D8–D20, 11 2012.
- [165] Yin Cheong Wong, Trayana Ilkova, Rob C. van Wijk, Robin Hartman, and Elizabeth C.M. de Lange. Development of a population pharmacokinetic model to predict brain distribution and dopamine d2 receptor occupancy of raclopride in non-anesthetized rat. *European Journal of Pharmaceutical Sciences*, 111:514 – 525, 2018.