Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Doctoral dissertation

Algorithms for the analysis of molecular protein structures and drug-like ligands for modeling and simulation of residence time drug-molecular target

Author: mgr inż. Magdalena Ługowska
Supervisor: prof. dr hab. inż. Marek Kimmel
Co-supervisor: dr inż. Marcin Pacholczyk

Gliwice 2023

Contents

List of Figures

List of Tables

Journal publications

1. **Ługowska M.**, Pacholczyk M.; *Revealing ligand unbinding kinetics using advanced computational methods: a review*; Current Protocols in Molecular Biology; 2023 (under review); IF 2.74

2. **Ługowska M.**, Pacholczyk M. PDBrt: *A free database of complexes with measured drug-target residence time* [version 2; peer review: 1 approved, 1 approved with reservations]. F1000Research 2022, 10(Chem Inf Sci):1236 (https://doi.org/10.12688/f1000research.73420.2)

3. Magdziarz T., Mitusińska K., Gołdowska S., Płuciennik A., Stolarczyk M., **Ługowska M.**, Góra A.: *AQUA-DUCT: a ligands tracking tool*, Bioinformatics, vol. 33, nr 13, 2017, s. 2045-2046, DOI:10.1093/bioinformatics/btx125, IF=5.481

Conference appearances

1. **Ługowska M.**, Pacholczyk M.; 26th Annual International Conference on Research in Computational Molecular Biology; La Jolla, USA; May 22-25, 2022; *Investigation of τ-random acceleration molecular dynamics (τRAMD) method for determining drug-target residence time*

2. **Ługowska M.**, Pacholczyk M.; 9th International Work-Conference on Bioinformatics and Biomedical Engineering IWBBIO; Gran Canaria, Spain, June 27-30, 2022; *Relative drug-target residence time approximation by τ-random acceleration molecular dynamics (τRAMD)*

3. **Ługowska M.**, Pacholczyk M.; BIT21 - Bioinformatics in Torun, Poland; June 24, 2021; *Ligand dissociation pathway investigation by τ-random acceleration molecular dynamics (τRAMD)*

4. **Ługowska M.**, Pacholczyk M.; 23th Gliwice Scientific Meetings; Gliwice, Poland; November 22-23, 2019; *The PDBrt Database: a comprehensive collection of drug-target structures with residence time data*

5. **Ługowska M.**, Pacholczyk M.; BIT19 - Bioinformatics in Torun; Toruń, Poland; June 27-29, 2019;*The PDBrt: a free database of drug-target residence time*

6. **Ługowska M.**, Góra A.; 7th International Work-Conference on Bioinformatics and Biomedical Engineering IWBBIO; Grenada, Spain; May 8-10, 2019; *Determination of enantioselectivity of selected enzyme – in silico study*

Abstract

Drug development is a complex process that remains subject to risks and uncertainties. In its early days, much emphasis was placed on the equilibrium binding affinity of a drug to a given target, which is described by the equilibrium dissociation constant (K_d). However, there are a large number of drugs that exhibit non-equilibrium binding properties. For this reason, optimization of other kinetic parameters such as dissociation rate constants (k_{off}) and association rate constants (k_{on}), has become increasingly important to improve accuracy in measuring *in vivo* drug effects. To achieve this, the concept of drug-target residence time (τ) was developed to consider the continuous elimination of the drug, lack of equilibrium conditions and conformational dynamics of target molecules. Therefore, residence time has been shown to be a better estimate of lifetime potency than equilibrium binding affinity and is recognized as a key parameter in drug design. Nevertheless, being a single measure of drug potency, residence time provides a limited picture of binding kinetics and affinities.

Due to the complex, labor-intensive, and costly nature of experimental methods for determining binding kinetic parameters, and the rapid advancement of technology, the demand for high-throughput *in silico* methods to estimate binding kinetic parameters and determine their key factors is increasing.

This thesis describes basic assumptions and applications of computational methods available for understanding and analysis of ligand binding and unbinding kinetics as well as determination of drug residence time in a target molecule. This includes molecular

dynamics (MD), machine learning (ML) methods and their combination, as well as the use of Markov state models.

Because the existing bioinformatics databases do not contain complete information on the kinetic data of complexes with known crystallographic structure, the data were collected from published literature and transformed into a publicly available online database called PDBrt. Studies were performed for selected protein-ligand systems, the inhibitors of InhA (the enoyl acyl carrier protein reductase from Mycobacterium tuberculosis) - one of the key enzymes involved in the type II fatty acid biosynthetic pathway in M. tuberculosis. The heat shock protein inhibitor HSP90 and ligands of ENR, EGFR and HIV-1 proteins were also used to check the reproducibility of the τ Random Accelerated Molecular Dynamics (τRAMD) method.

The MD approach is used to analyze τRAMD to determine if the method is reproducible and applicable in calculating different relative residence times of drug-like compounds. When applied to a series of similar compounds, τRAMD was found to provide the most accurate prediction of residence times. The reproducibility of τRAMD was demonstrated - the results obtained are similar to those published. However, the study showed that the τRAMD method did not allow estimating relative residence times that correlate well with experimental values for structurally diverse compounds. This suggests that the method has limited application and is not applicable to a wide range of compounds.

A machine-learning algorithm was proposed to identify molecular features affecting protein-ligand binding kinetics for a set of similar compounds. Molecular dynamics simulations of τRAMD results were used as model input. The study confirmed that τRAMD provides information about the characteristics of the dissociation pathway since the obtained dissociation trajectories can be used to identify the interactions that occur and the conformational changes of the system at subsequent time points.

This information has been applied to further analyses, which led to the definition of key molecular properties for a series of InhA inhibitors. The proposed algorithm made it possible to obtain information on protein-ligand contacts that are specific to their residence times.

Chapter 1

Introduction

Drug efficacy depends not only on achieving the required concentration of the drug in the effector tissue, but also on its ability to bind to the receptor, i.e., the matched protein that regulates cell binding or assists in signal transduction (i.e., receptor affinity), and to activate or block the receptor. A common selection criterion in drug design is the equilibrium binding affinity of small chemical molecules (ligands) to a molecular target (receptor). The affinity describes the persistence of binding and the potency of the ligand's activation of the receptor. Potency is an important parameter that defines the potential of the ligand to efficiently activate the receptor to produce a strong response *in vivo*. Thus, affinity is a measure to quantify the efficacy of a drug and to determine the benefit of the interactions that occur between the drug molecule and its target. Therefore, drug design protocols are mainly based on molecules with high binding affinity. However, many drugs do not achieve equilibrium binding so that this approach does not always correspond to a higher efficacy of the drug under *in vivo* conditions. In recent years, it has been shown that it is possible to predict the efficacy of a new drug under *in vivo* conditions by measuring the binding kinetics, and that the rate of binding and dissociation of a drug with a receptor, along with the

pharmacokinetic properties, is the main feature that determines the biological activity profile of a drug (Copeland 2016, Vauquelin 2016). The concept of drug residence time in a molecular target that has been introduced takes into account the conformational dynamics of target molecules affecting drug binding and dissociation. An important observation is that in some cases this time correlates better with drug efficacy *in vivo* than binding affinity (Copeland et al. 2006). As a result, the residence time of the ligand at the target molecule (τ) is considered a reliable determinant of drug efficacy. Together with the kinetic parameters of the reaction, it plays a role in drug discovery programs. However, if residence time is used as the sole measure of drug efficacy, it provides a limited picture of reaction kinetics (Folmer 2018).

1.1 Ligand-receptor binding kinetics

It is assumed that series of sequential biochemical reactions within the cell begins with a substrate, also called a ligand (L), which is a small chemical molecule that interacts with an enzyme, also called a receptor (R). Cellular pathway originates outside the cell with the ligand (a molecule that is the initial stimulus) approaching a specific protein. The molecules form closely matched pairs, with the receptor usually recognizing a set of specific ligands, and the ligand binding one of a set of specific target receptors. Binding of the ligand to the receptor changes receptor's shape or activity, allowing it to transmit the signal or directly induce a change in the cell.

An important part of drug design is understanding and fully describing the kinetics of receptor-ligand (RL) binding and the molecular determinants of this fit. For example, the drug ibuprofen, one of the most commonly used analgesics, antipyretics, and anti-inflammatories, is a non-selective inhibitor of cyclooxygenase (COX), the enzyme responsible for converting fatty acids into prostaglandins, and belongs to the

group of NSAIDs (non-steroidal anti-inflammatory drugs). Prostaglandins are substances involved in the inflammatory process. Inhibition of COX leads to the blockade of prostaglandins. By reducing the production of prostaglandins, ibuprofen is expected to reduce the fever and pain associated with inflammation. However, it should be noted that blocking the COX enzyme has side effects. This enzyme has very important beneficial functions in the body, such as protecting the stomach lining. Therefore, long-term use of COX inhibitors is associated with adverse effects on the stomach and intestines.

Enzymatic reactions can be characterized by relevant kinetic (reaction rate constants) and thermodynamic (reaction equilibrium constants) parameters. Kinetic parameters include:

- k_{on}, the association rate constant ($M^{-1}s^{-1}$), indicating the rate of formation of the receptor-ligand complex
- k_{off}, the dissociation rate constant (s^{-1}), describing the rate of ligand release from the receptor-binding site, and
- τ, the residence time (s), a measure of the time the ligand spends in the receptor-binding site (the lifetime of the receptor-ligand complex) defined as the inverse of the dissociation rate constant.

The process of binding a ligand to a receptor is accompanied by a change in free energy described by the Gibbs potential as a function of enthalpic and entropic factors. The enthalpic factor may be related to the formation and/or breaking of hydrogen bonds, ionic interactions, or hydrophobic interactions, among others. The entropic factor refers to the changes in the number of degrees of freedom after the formation of a complex.

$$\Delta G \equiv G(R+L) - G(RL) = -k_b T ln(K_d) \Rightarrow K_d = e^{\frac{\Delta G}{k_b T}} \tag{1.1}$$

where ΔG - change in free energy of Gibbs bond, k_b - Boltzmann's constant, T - temperature. K_d, the equilibrium constant of the dissociation process, is a thermodynamic parameter that quantifies the strength of the interaction in the RL system:

$$K_d = \frac{[R][L]}{[RL]} \tag{1.2}$$

where [R] - receptor concentration, [L] - ligand concentration, [RL] - receptor-ligand complex concentration. K_d can be expressed in terms of kinetic parameters through the following relationship:

$$K_d = \frac{k_{off}}{k_{on}} \tag{1.3}$$

In the simplest case, the reaction of a ligand with a receptor is a one-step process:

$$R + L \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} RL \tag{1.4}$$

where $k_1 = k_{on}$, and $k_{-1} = k_{off}$ (Figure 1.1 a).

However, this model is not always sufficient to describe the interactions of drugs with their targets, which often involve multistep binding and dissociation. This led to the development of a two-step model of binding kinetics that accounts for conformational changes leading to increased complementarity between molecules.

$$R + L \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} RL^* \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} RL \tag{1.5}$$

where k_1, k_{-1}, k_2, k_{-2} kinetic constants, R - receptor, L - ligand, RL^* - transient receptor-ligand complex and RL - final protein-ligand complex (Figure 1.1 b).

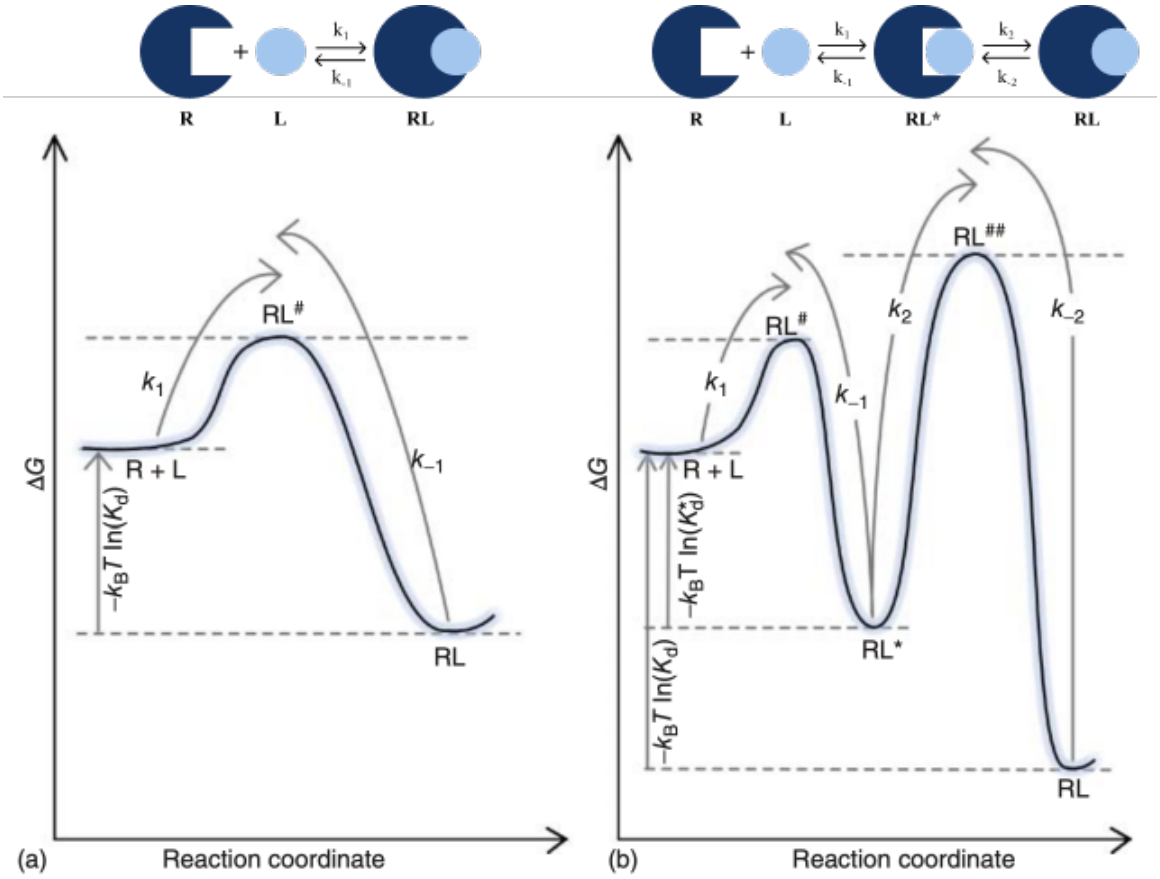


Figure 1.1: Schematic representation of the kinetics of the ligand-receptor binding reaction: (a) one-step and (b) two-step binding model. The diagrams show the energy picture with marked energy barriers (minima and maxima of free energy), whose height depends on the kinetic parameters. Graph adapted from (Romanowska et al. 2015).

In this model, a free drug encounters its target in a conformational state defined by a binding pocket that is suboptimal for the structure of the drug molecule. The initial phase of binding is an association process that forms an encounter complex (RL^*) defined by the association rate constant (k_1), the dissociation rate constant (k_{-1}), and the equilibrium dissociation constant (K_d) (Gabdoulline & Wade 1999, 2022). Initial binding is followed by another step in which the system must overcome the energy barriers created by conformational changes of the receptor and the ligand to form a

new stable state (RL) in which the binding pocket adopts a structure more similar to that of the drug molecule.

1.2 Residence time is an important factor in drug design

Residence time is a measure of how much time a ligand spends at a protein binding site. In other words, it is the residence time of a drug at a given target site. A drug is pharmacologically active as long as it remains bound to the receptor. Thus, the residence time is defined as the inverse of the dissociation rate constant $\tau = \frac{1}{k_{off}}$. This means that the concentration of a ligand does not affect its residence time in the target, and drugs with long residence times can remain bound even when their concentration falls below the equilibrium dissociation constant K_d. This is particularly important when the drug is cleared from the body, resulting in varying *in vivo* concentrations.

The residence time of many drugs has been shown to be more correlated with efficacy than binding affinity (Copeland 2016). Studies performed on a set of 12 different receptors in combination with 50 drugs have been presented where, in the vast majority of cases, higher efficacy was observed for drugs with long residence times than for short ones (Swinney 2004). In addition, the residence time may influence the interval between drug doses (Dowling & Charlton 2006, Vauquelin 2016).

However, it is important to note some uncertainties associated with the use of residence time in drug design programs. A simple model that accounts for the influence of both pharmacokinetics and binding kinetics on drug effects demonstrates that the binding time of a drug can be prolonged under the condition that pharmacokinetic clearance occurs faster than drug dissociation. However, the potential of using residence

time to determine drug action is reduced by known examples where drug dissociation is faster than elimination (Dahl & Akerud 2013).

1.3 Computational methods for residence time prediction

As interest in the residence time and importance of drug binding kinetics at the target binding site increases, *in silico* methods become more important. Primarily because the experimental methods commonly used are often very time consuming and costly. In addition, the use of computational methods to predict residence time and characterize reaction kinetics can support personalized medicine. Simulations tailored to the patient can accelerate the physician's decision to select the optimal drug from several potential drugs. Moreover, such calculations can be performed for compounds that have not yet been synthesized, which has significant implications for the cost and time of research. It should be noted that the developed *in silico* methods are based on experimental data that can confirm their reliability.

Computer-aided methods for estimating residence time and other kinetic parameters can be divided into two main groups. The first is a set of molecular dynamics methods with enhanced sampling. The second group are methods based on machine learning, often using molecular dynamics simulations.

1.3.1 Getting started with Molecular Dynamics

Molecular dynamics (MD) is an advanced computational tool for modeling systems of biological and chemical molecules. MD is used to describe the phenomena that occur at the atomic level, such as the interactions that occur in the model under study

during simulation (McDowell et al. 2007). The general protocol for classical molecular dynamics simulations is shown in Figure 1.2.

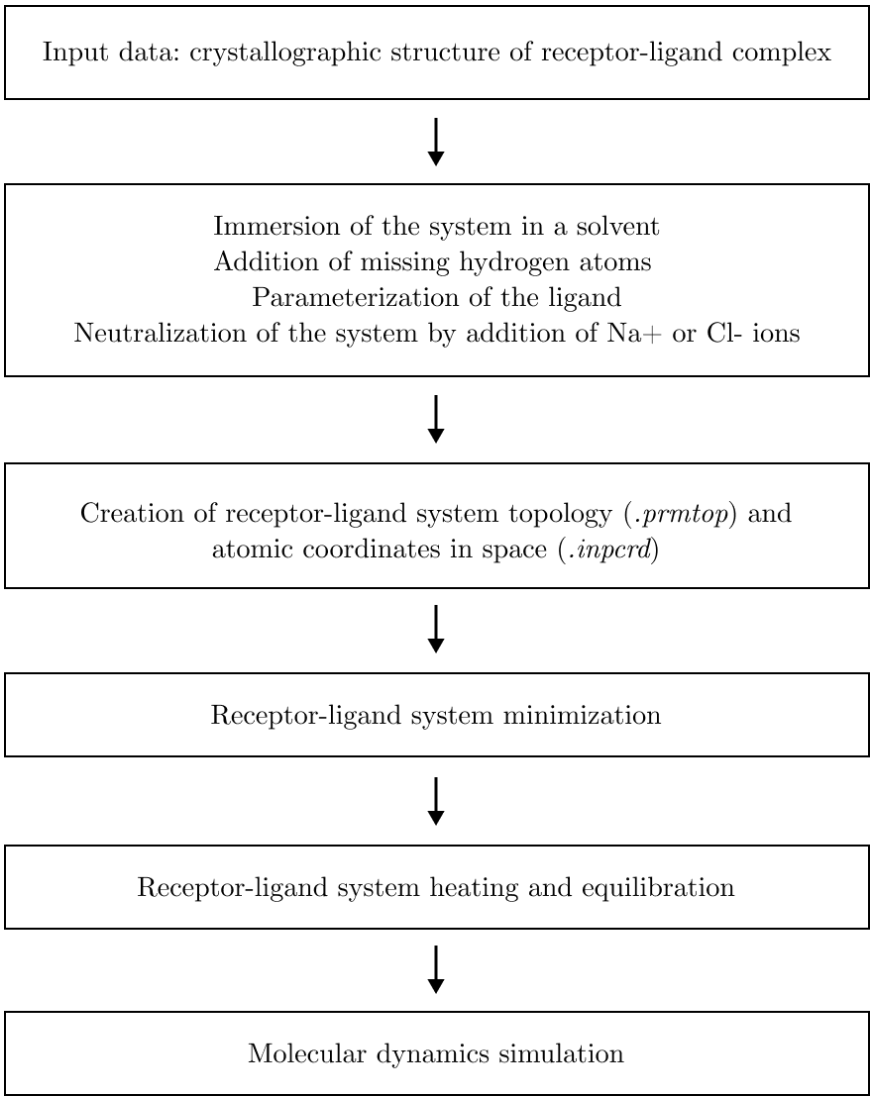


Figure 1.2: Protocol for preparing and performing molecular dynamics simulations of the receptor-ligand system.

Molecular dynamics provides information about the position of atoms at a given time during the simulation. The system is viewed as a collection of individual atoms, each with a charge and connected by covalent bonds. The motion takes place between the current time step and the next time step (the atom reaches a new position) ac-

cording to the principles of classical mechanics. The starting point for the simulation is a molecular model of the system under investigation. The structure of the system determines data about the spatial position of the atoms, their masses, and the position and value of the electric charges in the molecule. Then, random initial velocities are assigned to the atoms according to the Maxwell-Boltzmann distribution and Newton's equations of motion for the system are solved numerically. The force required to solve the Newtonian equation is determined from knowledge of the interaction potentials, thanks to the defined force field. Once the new position and velocity of the atoms is determined, a new force is again determined and the process is repeated. The whole process is repeated a certain number of times.

In biological systems, the macromolecules are immersed in a solvent consisting of water, ions, small organic molecules, and other macromolecules. Therefore, in order for the simulation to faithfully represent the behavior of the system, at least a small part of this environment should be included in the description of the system. When a solvent is included in the calculation, periodic boundary conditions are introduced into the simulation. The system is then constructed from an infinite number of identical elements, called periodic boxes, arranged so that there is no free space between their boundaries. The central box represents the actual system, and all the others are exact copies of it (they have the same speed and move in the same direction). Thus, when an atom leaves the simulation cell, its copy crosses the opposite boundary (replaces it) and the number of atoms in the cell does not change (Figure 1.3).

Before the actual molecular dynamics simulation, the system under investigation must be prepared accordingly. First, the system is subjected to a minimization in which it assumes the most probable configuration compatible with the given conditions (temperature, pressure). In other words, the system reaches a local energy minimum and the stresses between the atoms that lead to destabilization of the system are

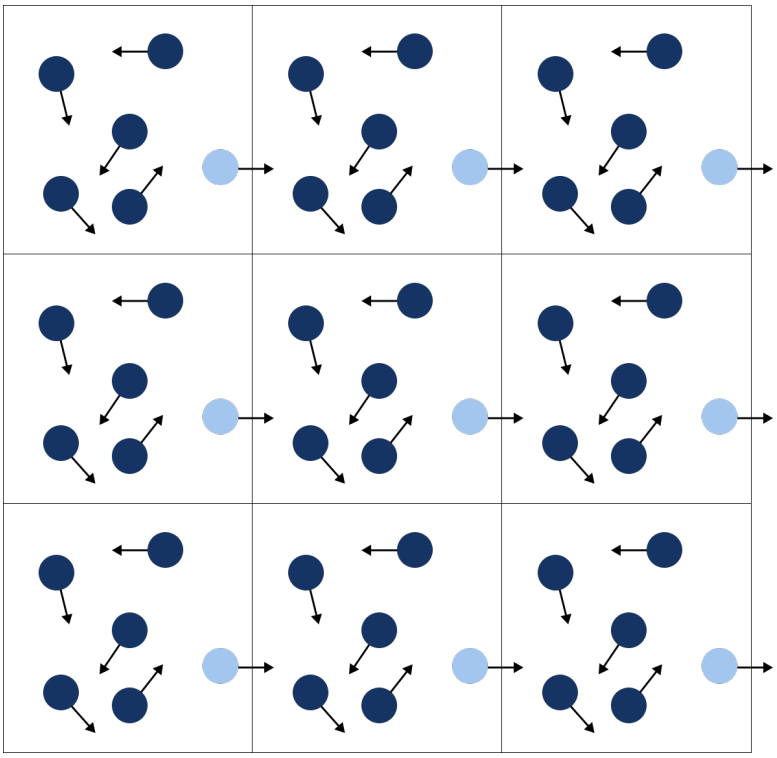


Figure 1.3: Periodic boundary conditions in two dimensions.

removed. Then the system is slowly heated, and the equilibrium of the whole system is reached. The system prepared in this way can then be subjected to a suitable molecular dynamics simulation.

Current computing power limits simulation times to the millisecond range. Since ligand dissociation usually occurs on a longer time scale than the possible simulation time, it can be concluded that its observation using molecular dynamics is impossible. Therefore, a series of methods based on molecular dynamics have been developed to observe the rare event of ligand dissociation from the receptor binding site.

1.3.2 Enhanced sampling methods

In the development of drugs using computational methods, two challenges can be observed that require special attention. One is the capture of the energy landscape, which includes kinetic and structural barriers. The other is the accurate determination of the kinetics of the receptor-ligand system under steady-state or nonequilibrium conditions. Sampling methods, which are modifications of classical molecular dynamics, increase the frequency of rare events that are the focus of research data. These methods can be divided into two groups in the context of predicting the residence time or dissociation rate constant (these parameters are interdependent) (see Table 1.1).

Table 1.1: Summary of sampling methods for receptor-ligand systems with the highest correlation of calculated (predicted) residence times with experimental times.

Method	Receptor	No. of compounds	Simulation time [ns]	Reference
Absolute residence time				
MSM	Trypsin/Benzamidine	1	58280	(Wu et al. 2018)
(Markov State Models)	Mdm2/PMI	1	500000	(Paul et al. 2017)
M-WEM (Markovian Weighted Ensemble Milestoning)	Trypsin/Benzamidine	1	480	(Ray et al. 2022)
AMS (Adaptive Multilevel Splitting)	Trypsin/Benzamidine	1	2300	(Teo et al. 2016)
Metadynamics	c-Src kinase-dasatinib	1	∼7000-8000	(Tiwary et al. 2017)
τRAMD + extrapolation	MtKatG-Isonazid	1	-	Maximova et al. (2021)
OPES (On-the-fly Probability Enhanced Sampling)	Trypsin/Benzamidine	1	3200	(Ansari et al. 2022)
Relative residence time				
τRAMD (τ Random Acceleration Molecular Dynamics)	HSP90-inhibitor	95	4-500	(Kokh et al. 2018, 2019)
Scaled MD	HSP90-inhibitor	7	200-7000	(Schuetz et al. 2019)
	GSK-3β	7	25-1000	(Gobbo et al. 2019)
	GK1	7	50-2000	(Mollica et al. 2015, 2016)
	HSP90-inhibitor	4		
	Grp78	4		
	A2A	4		
Steered MD	B-RAF	2	50	(Niu et al. 2016)
ABMD (Adiabatic-bias Molecular Dynamics)	GSK-3β	7	600	(Gobbo et al. 2019)
TMD (Targeted Molecular Dynamics)	HSP90-inhibitor	25	3-9	(Wolf et al. 2019)

The first group includes methods for determining the dissociation rate constant (k_{off}). The second group consists of methods for determining the relative residence time. That is, rather than estimating directly the experimental value (absolute value), they check how the calculated time, which corresponds to the length of the simulation, differs among compounds under study.

Methods for determining quantitative (absolute) values of the dissociation rate constant or the residence time require a large amount of simulation data as input.

Markov state models (MSMs) are constructed from simulation trajectories by sampling the conformational space of the receptor-ligand complex. The dynamics of the system are defined by a series of transitions between states, and the probabilities that the system starts in one state and transitions to the next on a given time scale are given by a transition matrix. An important assumption is that the probability of transition between states is independent of the trajectory history and conformational properties of the system, except for the current state. From this matrix, both the association and dissociation rates can be determined. In this approach, several independent MD simulations are performed, which can be started or stopped at any time, i.e., they can be of different lengths. Therefore, k_{off} values are predicted with lower accuracy than k_{on} because the release of ligands from the binding site is very rare and usually not observed spontaneously (Mondal et al. 2018). For this reason, several methods have been developed to build Markov state models for the calculation of (un)binding kinetics: Multi-ensemble Markov models (MEMMs) with transition-based reweight analysis (TRAM) (Wua et al. 2016) and their variant MBAR (TRAMMBAR) (Paul et al. 2017), Variational approach for Markov processes (VAMP) using deep neural networks (VAMPnets) (Mardt et al. 2018) and Deep Generative Markov State Model (DeepGenMSM) framework (Wu et al. 2018).

The approach that assumes that transitions between conformational states are

Markovian is Markovian Weighted Ensemble Milestoning (M- WEM) as a combination of Weighted Ensemble Milestoning and Markovian theory (Ray et al. 2022). Thus, for the ligand-receptor system, M- WEM can reproduce the experimental residence time, association and dissociation kinetics, and binding free energy in a short simulation time (nanoseconds).

One method that focuses on transitions between bound and unbound states to calculate the ligand dissociation constant is adaptive multilevel splitting (AMS) path sampling (Teo et al. 2016). The method is undoubtedly effective, and its results have been verified in the calculation of transition times for simple systems in both Monte Carlo and molecular dynamics simulations. Application of the method to MD simulations of protein-ligand dissociation has shown that the method can determine k_{off} with high accuracy. The metadynamics and its OPES variant, in which statistics are collected for multiple dissociation trajectories, allow determination of residence times that are close to experimental measurements. In addition, these improved sampling methods enable an understanding of ligand release pathways from the receptor binding site.

It should be noted that these methods are very computationally intensive, as evidenced by both the limited number of compounds for which the absolute value of the dissociation rate was calculated and the very short residence times of these molecules. In drug discovery, the main objective is to select the best compound with respect to a particular property (in this case, residence time). It can be concluded that the prediction of absolute values is not necessary, since relative methods can be used to compare two compounds at a much lower computational cost.

The method for determining relative residence time used for the largest number of compounds is τRAMD (Kokh et al. 2018). In this approach, an artificial, randomly directed force is used to remove a ligand from a binding site in a macromolecule, and

τRAMD correlates the average length of the dissociation pathway with an experimentally determined residence time. The movement of the ligand is evaluated according to this fixed number of simulation steps by calculating the distance between the new and the old position of the center of mass. If the distance change is less than the threshold, a new random force direction is generated. If the distance is greater than the threshold, the simulation continues for the next number of simulation steps with the same force direction. The authors found that the method works well to determine the relative residence time of similar and dissimilar compounds.

An enhanced sampling method for predicting the relative residence time of compounds with a wide range of experimentally determined residence times is Scaled MD (Schuetz et al. 2019). Its computational efficiency is much lower than the other relative methods, with an average computation time of about 1 microsecond (see Table 1.1). This performance can be improved by reducing the scaling factor, but this has been shown to negatively affect the accuracy of the relative residence time prediction.

Target MD, on the other hand, has a very low computational cost (see Table 1.1). The efficiency of the method has been shown to be highest for compounds with high similarity, which means that it will perform best in the optimization phase of the drug discovery process (Wolf et al. 2019).

Steered MD is characterized by the attachment of a Newtonian spring to the ligand, which pulls the ligand with varying force to give it a specific velocity. This approach allows the determination of the ΔG_{off} value (the change in free energy of the dissociation process), which is used to rank the residence times of two inhibitors. However, its application is limited to two compounds, so it requires more verification to determine its correct operation (Niu et al. 2016).

For compounds with high structural similarity, the presented methods for determining the relative residence time of a ligand in the receptor binding site show high

efficiency. Their performance seems to be lower for large (¿ 15-25) sets of compounds, except for the τRAMD method, which was tested on a set of 95 HSP90 protein inhibitors (see Table 1.1).

1.3.3 Getting started with Machine Learning

Machine learning (ML) is a subdomain of artificial intelligence that focuses on applying data and algorithms to reproduce the way humans learn and progressively improve their accuracy. In other words, it's the process of creating and using mathematical data models to make algorithms independent and decoupled from the need for direct human instruction. Machine learning-based algorithms make it possible to classify, predict, and describe data without requiring the user to explicitly solve the problem, but by finding trends in the data. Figure 1.4 illustrates how machine learning works.

Machine learning methods can be divided into four groups: supervised machine learning, unsupervised machine learning, semi-supervised machine learning, and enhanced machine learning.

Unsupervised learning uses algorithms to analyze and cluster unlabeled data sets. This group of algorithms includes cluster analysis algorithms: the K-Means method, hierarchical cluster analysis, DBSCAN (density-based spatial clustering of applications with noise), and algorithms for visualizing data and reducing its dimensionality: principal component analysis (PCA) or kernel principal component analysis. These methods have been used in the development of small molecules (Kadurin et al. 2017).

Supervised learning, on the other hand, uses labeled datasets to train algorithms. In this process, weights are adjusted until the model fits correctly. The input data is divided into training and testing sets in a cross-validation procedure to prevent the

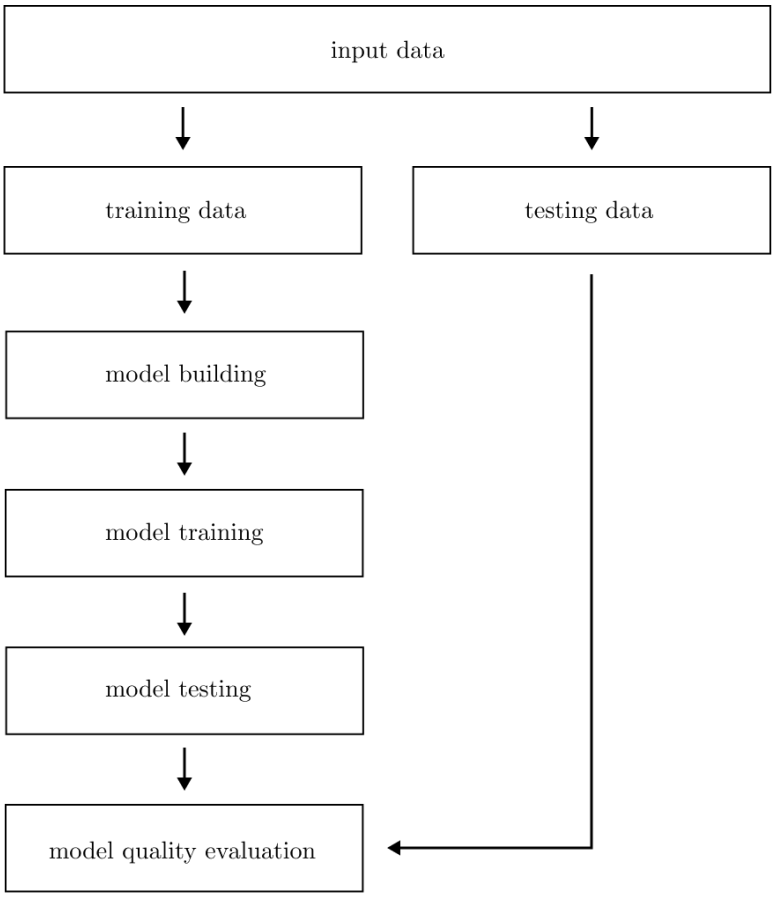


Figure 1.4: Workflow diagram for machine learning.

model from over- or under-fitting. Overfitting of the model occurs when the model makes good predictions in the training data, but the quality of the predictions in the test data set decreases significantly. This situation often occurs when the data set is too small relative to the complexity of the model used. Underfitting of the model, on the other hand, occurs when the model is too simple and cannot capture the dependencies in the training data set (Bashir et al. 2020).

There are two main applications of supervised learning: regression (value prediction) and classification (class prediction). Examples of algorithms for solving regression problems are: Linear Regression, Polynomial Regression, and algorithms for solving classification problems include: Decision Tree, k-NN, SVM, Random Forest, Logistic

Regression. Classification methods are most commonly used in the target identification phase of the drug discovery process. Regression methods are used to predict pharmacokinetics, activities, absorption, distribution, metabolism and excretion of a drug based on chemical structures and features (Bonaccorso 2017).

Evaluating the performance of a ML model, i.e., describing how well the model makes predictions, is an important part of developing a ML model. The metrics used vary depending on the problem. Mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R-squared) are typically used to evaluate prediction error rates and model performance in regression analysis. For classification models, numerical quality metrics such as TP (true positive), TN (true negative), FP (false positive), and FN (false negative) and their derivatives-such as overall classification performance, sensitivity, or specificity-or graphical metrics such as the confusion matrix and the ROC curve are used (Bonaccorso 2017).

1.3.4 Machine learning-based methods

Machine learning techniques have been used for molecular structural analysis, prediction of dynamic behavior, investigation of molecular dynamics trajectories, and molecular dynamics sampling (Wang et al. 2020). Rather than predicting the dissociation of a ligand or its residence time in a target, these methods are most commonly used to predict binding affinity. Examples of machine learning applications for affinity prediction include AtomNet (Wallach et al. 2015), COMBINE (Ortiz et al. 1995), KDEEP (Jiménez et al. 2018), OnionNet (Zheng et al. 2019), SIGN (Li et al. 2021), unsupervised deep learning for binding energy prediction (Yasuda et al. 2022), and others [for a detailed review, see elsewhere: (Dhakal et al. 2022)].

Most ML-based residence time prediction methods have been applied to HIV-1 and

HSP90 protease inhibitors. For the first time, the three-dimensional VolSurf lattice and partial least squares (PLS) modeling were used to develop predictive models of binding kinetics for HIV-1 and HSP90 inhibitors dataset (Qu et al. 2016).

Energetic and conformational features from molecular modeling were combined with a multitarget machine learning (MTML) approach to classify the HSP90 inhibitor system, with classes representing ligand binding kinetics.

A quantitative structure kinetic relationships (QSKR) for the dissociation rate constant (k_{off}) were determined using the COMBINE method, which was used as a tool to determine binding affinity (Ganotra & Wade 2018). In addition, the energy of intermolecular interactions obtained from molecular dynamics simulations can be imprinted into the molecular structure. It was shown that the dissociation rate is a function of the first half of the total dissociation process. A regression model was then developed based on a systematic analysis of protein-ligand binding interactions in dissociation trajectories. It cannot be excluded that the predictive power of these methods is overestimated because the test set did not contain structurally distinct ligands (they were closely structurally related to the training set).

The QSKR model for predicting the dissociation rate constant (k_{off}) of a ligand based on the structure of the receptor-ligand complex was applied to a larger and more diverse data set (406 ligands) (Su et al. 2020). The authors found that the model showed good predictive accuracy on external test sets that included multiple targets as well as a single target.

Random forest approach focused on structural descriptors from molecular dynamics simulations was used to demonstrate their importance in predicting kinetic rate constants for different receptor-ligand complexes (Amangeldiuly et al. 2020).

Similarly, to identify new receptor-ligand contacts that place the molecular system

in transient states, a method for transient state analysis (MLTSA) was developed (Badaoui et al. 2022). These novel interactions have been shown to contain key features that determine the kinetics of binding. However, the method has been applied to a small data set: cyclin-dependent kinase 2 (CDK2) and its two inhibitors.

The PCA-ML method for clustering dcTMD trajectories into clusters reflecting binding pathways and the RMSD-based clustering method for grouping pathways by their mean Euclidean distance are some of the newer methods for analyzing binding mechanisms in receptor-ligand system simulations. However, they are limited by the need for human intervention in the initial selection of pathways during model training or in deciding the boundaries of neighborhood networks (Bray et al. 2022).

1.4 Objectives and Motivation

An important element in the drug development process is characterising and understanding the reaction kinetics of ligand dissociation from the receptor target site. Molecular simulations are an important tool for describing the dissociation pathway, predicting kinetic parameters, including residence time, and determining structural features. In order to observe the occurrence of rare events during the simulation and to reduce the computational complexity, simplifications such as enhanced sampling are often used in these approaches. Short-lived events, such as the rearrangement of atoms in a molecule during the induced fitting step, are inaccurately described by these simplifications. This can be seen as a limitation of the simplifications. On the other hand, classical molecular dynamics allows us to understand these fast, important events. However, it cannot be applied to longer time scales, such as the residence time of a drug in a target, which can range from a few seconds to hours. In addition, the accuracy of the simulations is not stable, although they are often used as input to

machine learning-based algorithms. The reason for these studies is the need for more efficient and accurate methods to analyse receptor-ligand binding kinetics.

The aim of this work is to apply and test drug residence time solutions to investigate whether they can be used regardless of the size of the molecules or protein family, and to analyse structural interactions in the binding process of the InhA protein and its inhibitors in particular. The following questions will be answered by the research presented here:

- What are the most important receptor-ligand interactions that distinguish between long- and short-living ligands?
- Is the τRAMD method universal, and can it be applied to molecules of different sizes and with various structural similarities?
- How does relative residence time correlate with experimental measurements?

To answer the above questions, two new tools have been developed:

- PDBrt kinetic database publicly available on https://pdbrt.polsl.pl/ and
- tool for the automatic identification and analysis of molecular properties such as ligand-receptor interactions during molecular dynamic simulations.

Chapter 2

Methods

This chapter provides a theoretical overview of the various methods of advanced computational chemistry and a brief description of the software used. The order in which the methods are described follows the order in which they are used in subsequent chapters. Figure 2.1 shows the research protocol presented in this thesis.

2.1 τRAMD ligand dissociation pathways

τRAMD is an enhanced sampling method for molecular dynamics simulations. It was developed to calculate the relative residence time of pharmacological compounds in their molecular targets and to study the dissociation pathway of ligands from receptor binding sites (Kokh et al. 2018).

τRAMD simulations, in which a small randomly oriented force is applied to the center of mass of the ligand to accelerate its exit from the receptor active site, are performed on receptor-ligand systems immersed in a solvent. After a given time, the ligand movement is checked. A random change in force direction occurs if the change in position was less than a predefined threshold distance. When the ligand leaves the receptor binding site, the simulation ends. This condition is defined by specifying the

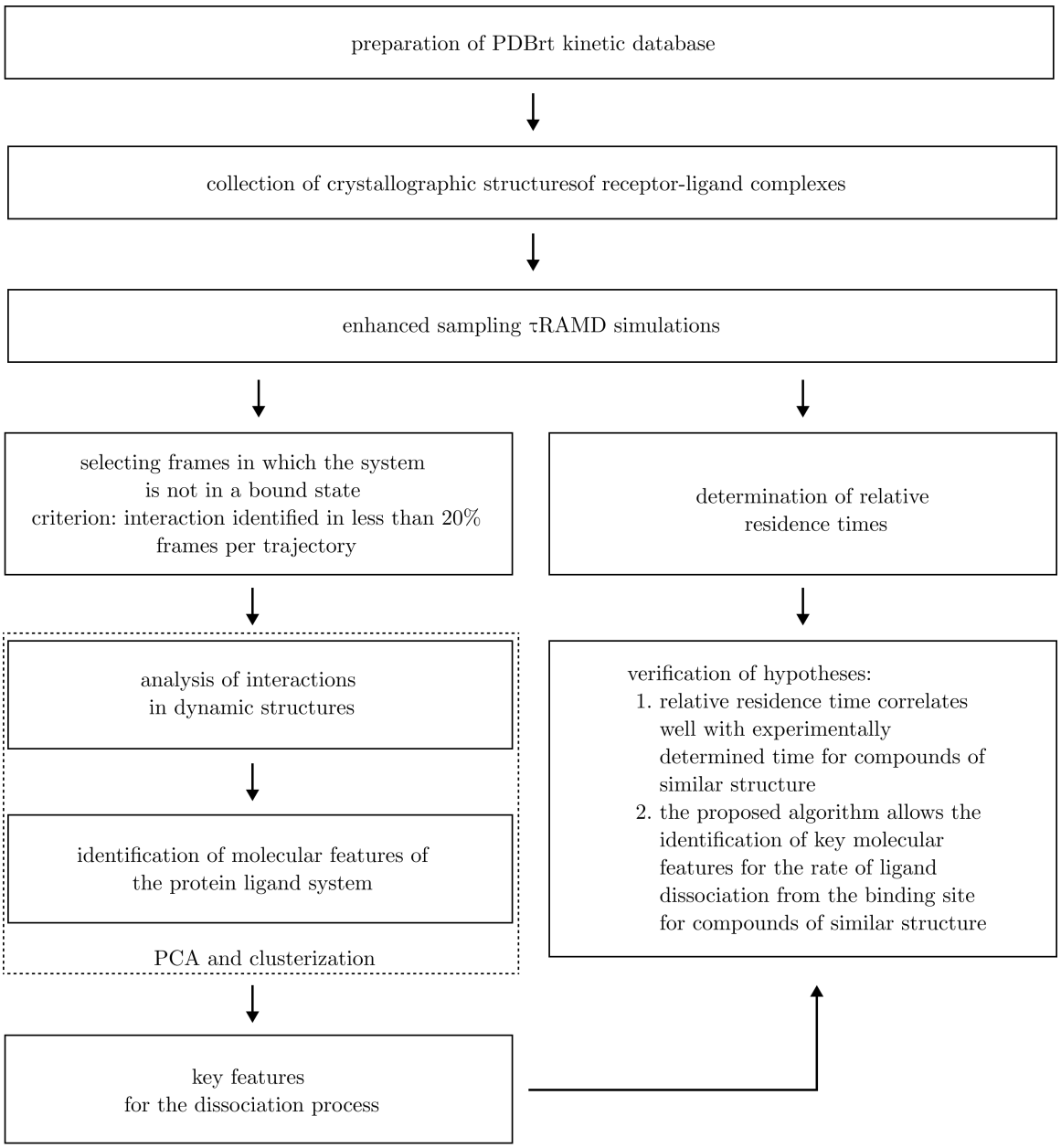


Figure 2.1: The research protocol presented in this paper.

ligand distance from the binding site corresponding to release in the configuration file. Simulation time is dependent on the residence time of the ligand in the target. Ligands with a longer residence time will require more time to leave the target (simulated time) or more force to leave the target in a given simulated time. No prior knowledge of the dissociation path or extensive parameterization is required for this method. The only parameter that needs careful definition is the magnitude of the applied force, which should not interfere with the calculated relative residence times. The only parameter that is carefully defined is the magnitude of the force, which must not interfere with the calculated relative residence time, i.e. the force must not force the ligand out of the binding site, so that the relative residence time estimates for each ligand are approximations, regardless of the actual dissociation rates. Due to the above mentioned characteristics, τRAMD is an efficient and relatively simple tool to estimate the relative residence time of drugs for molecular studies.

The MD simulation procedure with an additional force to accelerate the ligand output (τRAMD) was performed according to the published protocol (Kokh et al. 2018) and is shown in Figure 2.2.

The crystallographic structures of the HSP90 inhibitor, HSP90-gelanamycin, InhA inhibitor, ENR-triclosan, and EGFR-lapatinib complexes in the bound state were used as initial structures for τRAMD simulations (see Chapter 3). The crystallographic models are built based on an electron density map. This allows the localization of individual atoms based on spatial distribution analysis. In such structures, there is often a lack of information to accurately determine the position of all atoms, especially hydrogen atoms, which make up 50% of the atoms in the structure and have low electron density (Gilski 2014). The input systems were protonated with the PyMOL tool (Schrödinger & DeLano 2020) and parameterized with AmberTools (Case et al. 2022).

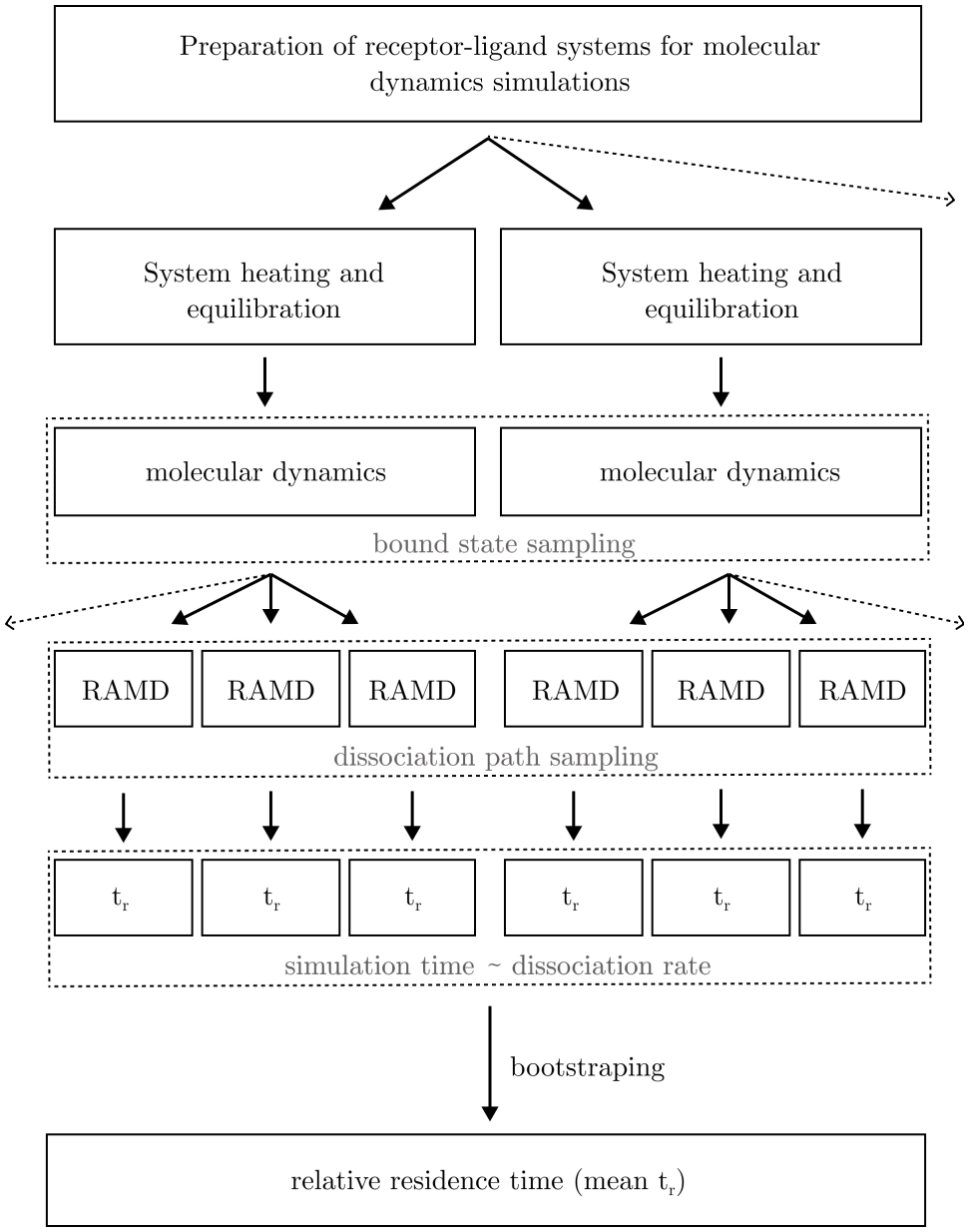


Figure 2.2: The τRAMD simulation protocol that takes into account obtaining of estimated relative residence times.

The tools antechamber, tleap and pmemd were selected from the package. Antechamber assigns partial charges to the ligand atoms and allows the atom types to be changed to those recognized by Amber14. The AM1-BCC method was used to calculate the partial atomic charges of the ligands (Jakalian et al. 2000, 2002). System topology

was built with tleap. This tool allows for charge neutralization (addition of Na^+ or Cl^- ions). The system is then immersed in a solvent using a force field of choice. The Amber ff14SB forcefield (Maier et al. 2015) was used for proteins, the General Amber Forcefield (Wang et al. 2004) (GAFF) for ligands, and the TIP3P model (Jorgensen et al. 1983) for water molecules. The water molecules were placed between the complex and the edge of the box, with the minimum distance being 10 Å. Crystallographic water molecules (water molecules obtained with the enzyme structure after crystallization) were added prior to immersing the complex in the solvent. Energy minimization, heating and equilibration calculations of the systems were then performed using the pmemd tool (Maier et al. 2015). The minimization was first carried out with a gradually decreasing harmonic force constant: 500, 50, and 5 kcal/Å^{2}mol, and then without any constraints at all. Each system was heated to 300 K with a heavy atom constraint of 50 kcal/Å^{2}mol. The systems were equilibrated: with the constraint decreasing from 50 to 10 kcal/Å^{2}mol, to 2 kcal/Å^{2}mol, and finally with no constraint. These atomic coordinates served as input for molecular dynamics simulations using NAMD software (Phillips et al. 2020). In the 2 ns simulations, Langevin dynamics were considered for a fixed temperature (300 K) and pressure (1 atm). τRAMD simulations were then performed using the resulting atomic coordinates and velocities.

In the τRAMD simulations, the motion of the molecule was evaluated after every 50 simulation steps with a randomly oriented force of 14 kcal/Å^{2}mol applied to the center of mass of the ligand. A length of 2 fs corresponded to one simulation step. If the position change was less than 0.025 Å, a new random force direction was generated. Otherwise, the same force continued to be applied. The end of the simulation, and thus the release of the ligand, was observed when the distance between the center of mass of the ligand and the receptor exceeded 40 Å. If no release was observed within 2 ns, the simulation was terminated. Every 100 fs, the coordinates of the trajectory

were saved.

For each system, the molecular dynamics simulation steps were repeated 5 times using NAMD software, which was treated as start files for τRAMD. From each starter file, a set of 10 dissociation trajectories was generated, resulting in a total of 50 dissociation simulations for each system. The residence time was defined as the simulation time required to dissociate the ligand in at least 50% of the trajectories, following the published procedure. The transient residence time was calculated as the mean of the (t_r) distribution for each simulated replicate using a bootstrap procedure. The average of all simulated repetitions for a given system was then used to estimate the correct relative residence time. To verify the method, a linear correlation was used between calculated and measured residence times on a logarithmic scale:

$$Y = aX + b \tag{2.1}$$

where $Y = log(\tau_{comp})$, $X = log(\tau_{exp})$.

The τRAMD dissociation time distributions were then subjected to statistical analysis. Given a sufficiently large sampling, the generated relative residence time distributions are expected to be asymptotic with respect to the Poisson statistic (ligand dissociation events are independent). The Poisson cumulative distribution function (*cdf*) was then calculated from the following formula:

$$P = 1 - e^{-\frac{t}{\tau}} \tag{2.2}$$

where τ=t_r, and compared to the empirical cumulative density function (*ecdf*) obtained from the dissociation probability distribution observed in τRAMD simulations. The Kolmogorov-Smirnov test was then applied. The maximum distance D between *cdf*

and *ecdf* was defined as the result:

$$D = sup|F_1(t) - F_2(t)| \tag{2.3}$$

where $F_1(t)$ is the cdf of the Poisson distribution at time tr (i.e. theoretical) and $F_2(t)$ is the empirical distribution function of the observed data (i.e. from the dissociation probability distribution).

2.2 Molecular features that determine residence time

Analyzing the interactions between ligands and receptors is an important factor in understanding the kinetics of ligand binding to receptors. Figure 2.3 illustrates the protocol used. From the τRAMD dissociation trajectory, receptor-ligand interactions were extracted as follows: (i) the position of the ligand center of mass and the coordinates of the atoms that make up the entire system were obtained from each frame (timepoint) of the trajectory and stored in separate *.pdb* files using a tcl script written for the VMD tool and executed from a Python script, (ii) the obtained coordinates of the position of the atoms in space were used as input for the identification of ligand-receptor interactions using the RDKit and ProLif libraries of Python (interaction classes: Hydrophobic, π-stacking, π-cation and cation-π, anionic and cationic, and H-bond donor and acceptor); (iii) each interaction was marked as "1" if an interaction was observed or "0", i.e., (v) on the basis of the frequency of occurrence of the interactions, a threshold was set which allowed the separation of the bound state from the transient and fully released states - for the purpose of the subsequent evaluation, those states of the system in which an interaction was detected in at least 20% of a single dissociation pathway have been removed.

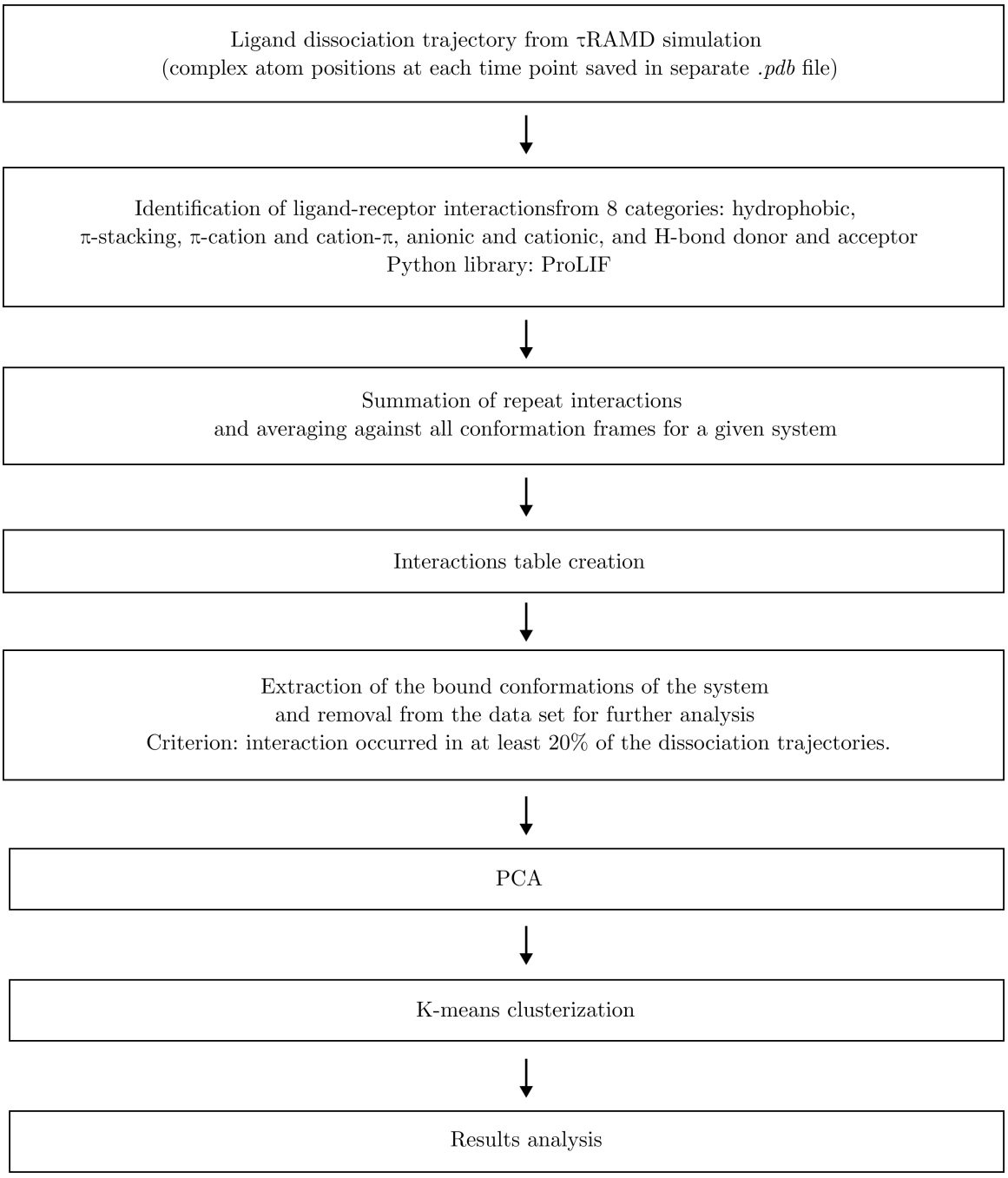


Figure 2.3: Flowchart of receptor-ligand interaction analysis during a ligand dissociation event, obtained from τRAMD simulations.

2.3 Software and Tools

The study presented in this thesis used a number of computational methods and visualization tools. Some of these are described above (see Sections 2.1 and 2.2). Scripts written in Python, bash, and tcl were used to analyze the data and to automate some of the tasks performed. In this section, the main tools and how they were used are described.

2.3.1 Receptor-ligand model preparation

Section 2.1 describes the **AmberTools** toolkit. These programs simulate and analyze biological and chemical molecules. They can be used as stand-alone tools or in combination with the Amber package. The tools antechamber, tleap, and pmemd from the AmberTools package were used in the work presented here. Antechamber is used for the creation of force fields for the molecules. Tleap is the most important tool for the preparation of the system for the simulation. And finally, pmemd allows molecular dynamics simulations to be performed. pmemd is similar to the sander code that also performs MD simulations, but it offers higher performance and a significant speed-up on the GPU.

2.3.2 Molecular dynamics simulation

Nanoscale Molecular Dynamics (NAMD) is a molecular dynamics simulation software (available at http://www.ks.uiuc.edu/Research/namd/). It has been developed using the Charm++ parallel programming model. The functions, parameters, and file formats of AMBER and CHARMM are fully compatible. It is often used for systems with up to millions of atoms due to the high performance achieved by the high parallelization of the code. In addition, by combining NAMD with the VMD visualization code, it is possible to perform interactive molecular dynamics. Since the τRAMD

simulation protocol is implemented in NAMD as a tcl script, the NAMD software was used to perform τRAMD simulations in this work.

2.3.3 Molecular structure visualization

PyMOL Molecular Graphics System (https://www.pymol.org/) is an application to visualize molecules in space. The tool allows the 3D visualization of models of biological and chemical molecules in a variety of representations and the performance of geometric manipulations. It also provides functions for sequence and structure analysis and for generating high-quality 3D images. In this work, PyMOL was used for the visual inspection of receptor-ligand systems as well as for the addition of missing hydrogen atoms in the molecular structures of both molecules.

Visual Molecular Dynamics (VMD) is a tool that has been developed for the visualization and analysis of biological systems, such as proteins or nucleic acids (Humphrey et al. 1996). In addition to the capabilities of PyMOL, VMD can be used to visualize and analyze trajectories obtained from molecular dynamics simulations. This tool was used for the visualization and analysis of dissociation trajectories generated by τRAMD simulations and is available at http://www.ks.uiuc.edu/Research/vmd/.

2.3.4 Identification of molecular features

A number of methods for learning and understanding the molecular structure of compounds are available in Python's **RDKit library** (https://www.rdkit.org/). In the work presented here, it was used for the visualization of 2D ligand molecules, the generation of molecular fingerprints of the compounds of interest, the calculation of the Tanimoto similarity coefficient.

ProLIF (https://github.com/chemosim-lab/ProLIF) is a Python library for fingerprinting intermolecular interactions extracted from molecular dynamics simulations.

Hydrophobic, π-stacking, π-cation and cation-π, anion and cation, and H-bond donor and acceptor interactions are identified by default. The library allows re-parameterization of the interaction types, as well as their extension to user-specified types. The algorithm expects the Molecule class of the RDKit library as input data describing the conformation of the ligand and receptor molecules. Then, it performs interaction recognition. Each interaction is marked as True if it occurred or as False if not identified at a given moment. The indices of the atoms responsible for the interaction can also be obtained. To facilitate further data manipulation, the library allows to return the identified interactions as a pandas DataFrame library object (Bouysset & Fiorucci 2021). To identify receptor-ligand interactions, this library was used in this work. The input files were extracted from the dissociation trajectories. These files contain information about the position of the system atoms at each time point of the simulation. The identified intermolecular contacts were converted into a Pandas DataFrame object and then stored as a table of intermolecular contacts.

2.3.5 Statistical data analysis

Principal Component Analysis (PCA) is an unsupervised machine learning technique to reduce the dimension of data containing n observations of m features. That is, the set has m dimensions or is described by m attributes. While minimizing the loss of information, PCA increases the interpretability of the data. The components are defined as a linear combination of the variables examined under the assumption that subsequent components are not correlated with each other. The analysis is based on the determination of an axis that preserves the maximum variance of the learning set. The method assumes that the first principal components will contain most of the variance of the original data. A correlation matrix between the input data is constructed in the first step of the algorithm. This matrix describes the correlation of

each attribute and is based on variables that have been normalized so that each input variable has the same variance. The next step is to determine the eigenvalues of the matrix. The eigenvalues determine how much of the variability is described by a given principal component. The total variance is the sum of the eigenvalues. The impact of the principal variables on the data of each component can be visualized. Each variable is represented as a load vector, the length, and direction of which determine how much influence the variable has on the principal components. Two variables that are close to each other on the graph have a positive correlation; if they are on opposite sides, their correlation is negative; and if they are perpendicular to each other, they have no correlation.

The **k-means method** is an iterative unsupervised machine learning algorithm for clustering data. It is based on the Euclidean distance of each point from a center, called the centroid. The data is divided into k clusters. The clusters contain similar objects such that the sum of the distances between them and the center of a given cluster is minimal. In the work presented here, the objects are individual receptor-ligand systems. The algorithm first creates k clusters using the elbow method. This method indicates the optimal number of clusters for a given data set. It then computes the centroids for each cluster and the distances of each point from the centroids. A given element is assigned to the nearest cluster. Again, the position of the centroid of each cluster is updated by setting it as the average of all the points assigned to it. The steps of the algorithm are repeated until convergence is achieved, i.e. the shift in the position of the cluster centroid is less than a certain threshold. In the study, this method was used to cluster data describing systems during ligand dissociation from the receptor binding site.

2.4 Automation of the data preparation process

Molecular dynamics simulations were run for 28 ligand-receptor systems. In order to streamline the process of data preparation, as well as the creation of the necessary configuration files, a protocol for the automation of these individual tasks was prepared. All scripts were written in Python (version 3.9). The code is available in the GitHub repository at https://github.com/mlugowska/residence-time-prediction.

Files containing atomic coordinate information for each subject were downloaded from the PDBrt database in *.pdb* format. A single file contained receptor or ligand information separately. These files are the input for the following application.

The first step is to add the missing hydrogen atoms to each of the molecular models of both the protein molecule and the ligand. This is done using the PyMOL tool, which adds hydrogen atoms based on the valence of each atom making up the molecule. The tool is also used to select and protonate water molecules, which are saved as a separate *.pdb* file. Because of the different naming conventions for hydrogen atoms in Amber and PyMOL, the name is changed to match the protein name. In the file describing the ligand, CONECT lines are automatically removed.

The antechamber tool from the AmberTools package is then used to convert the file containing the simplified ligand information into the .mol2 format. This format contains partial charges. These are calculated using the semi-empirical AM1-BCC method. The net charge of the ligand is equal to 0. The program then determines the bond types and atoms of the ligand using the parmchk package, and the ligand is parameterized using the tleap tool. The program creates topology files (*.prmtop*), coordinate files (*.inpcrd*) and *.pdb*.

Finally, a single .pdb file describing the entire system is created from the files containing the prepared protein, ligand, and water structures. This system is then immersed in a solvent, such as water, and is neutralized by the addition of Na^+ and

Cl^- ions, respectively. The program returns files that can be used as input for energy minimization simulations, system heating, and equilibration with Amber, containing the topology and coordinates of the entire system.

The automatic generation of files containing the configuration of the above amber simulation is the last step of the program. Any local computer can run the finished set. The Amber output files are used as input for NAMD heating and equilibration simulations, which in turn are used as input for RAMD simulations. The program automatically creates all necessary configuration files for each complex.

It is also possible to generate scripts that execute a High Performance Computing (HPC) task in a slurm queue on a user-specified supercomputer to which the user has access. The ICM UW computing cluster (https://icm.edu.pl/en/) and PLGrid (https://www.plgrid.pl/en) were used for the research in the present thesis.

Chapter 3

Research data

3.1 PDBrt kinetic database creation

Many receptor-ligand complexes are well characterized experimentally and have been the subject of extensive structure-based drug discovery efforts. However, no information on the residence time of a drug in its molecular target is saved in the available bioinformatics databases. A database of experimental kinetic data, including residence times, deposited in the Protein Data Bank (PDB) has been developed with the goal of facilitating further research and development of residence time modeling. There are 59 complexes with experimentally measured residence times, including seven protein families and 56 small molecules, in the current version of the PDBrt database. Figure 3.1 and Table 3.1 provide summary statistics.

The acquisition of protein-ligand binding kinetics data involved a review of the available scientific literature to obtain the experimentally measured ligand-target residence time or dissociation rate (*koff*) and to verify that the complex in question matched the crystallographic structure available in the PDB database. Two methods were used to extract data from the literature. First, a combination of the following search terms was used to search the articles contained in the PubMed database

(https://pubmed.ncbi.nlm.nih.gov/): k_{off}, *residence time*, *kinetic binding*, and *drug-target kinetics*. Second, the primary reference of each PDB file was viewed in the RCSB PDB database (https://www.rcsb.org/). Each returned publication has been reviewed to determine kinetics, and each complex to determine crystallographic availability. A receptor-ligand complex was added to the PDBrt database if it met both criteria.

The database was initially created as a Microsoft Excel table. Table 2 shows the original data obtained from the literature. Missing data were left blank. A simple unit conversion was also performed to convert values from the manuscript to the units given in Table 3.2.

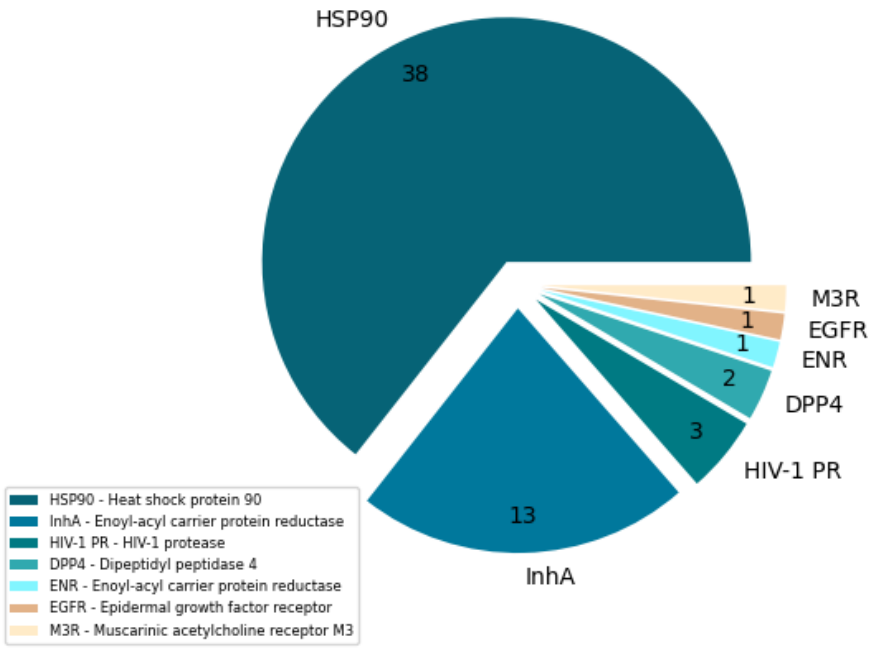


Figure 3.1: The number of protein families available in PDBrt. Figure adapted from (Ługowska & Pacholczyk 2021).

Table 3.1: List of chemical compounds in pairs with a given protein family.

Protein Name	No. of compounds	Receptor-ligand complex PDB ID	Ligand Code
DPP4	2	3BJM, 6B1E	BJM, LF7
EGFR	1	1XKK	FMM
ENR	1	1QSG	TCL
HIV-1 PR	3	6DJ1, 3EKX, 4DQB	AB1, 1UN, 017
HSP90	37	1YET, 2BSM, 2UWD, 2VCI, 2VCJ, 2YKI, 2YKJ, 5J20, 5J27, 5J2X, 5J64, 5J86, 5J9X, 5LNY, 5LNZ, 5LO5, 5LO6, 5LQ9, 5LR1, 5LR7, 5LRZ, 5LS1, 5NYH, 5NYI, 5OCI, 5OD7, 5ODX, 5T21, 6EI5, 6EL5, 6ELN, 6ELO, 6ELP, 6EY8, 6EY9, 6EYA, 6EYB, 6F1N	GDM, BSM, 2GG, 2GJ, 2EQ, YKI, YKJ, 6FJ, 6FF, 6DL, 6G7, 6GW, 6GC, 70K, 70Z, 70M, 70O, 72K, 72Y, 73J, 73Y, 73Z, 9EK, 2EQ, 9R8, H0T, 9RZ, 74E, B5Q, PU1, P4A, BAW, BA8, C4T, C4N, C4K, C3Z, C8W
INHA	13	2X22, 2X23, 4OHU, 4OIM, 4OXY, 4OYR, 5COQ, 5MTP, 5MTQ, 5MTR, 5UGS, 5UGT, 5UGU	TCU, TCU, 2TK, JUS, 1TN, 1US, TCU, 53K, XT3, XT0, XT5, XTW, XTV
M3R	1	4DAJ	0HK

From the available primary data, several properties were calculated. Equation 3.1 was used to calculate the residence time (minutes) and associated error if not defined in the reference literature.

$$\tau = \frac{1}{k_{off}} \tag{3.1}$$

This data was then entered into PDBrt. For each complex, files describing the three-dimensional structures of the molecules are automatically downloaded from the PDB database after the data file is uploaded. Receptor molecules, along with other components like water molecules and metal ions found in original structural description files, were saved in *.pdb*, while ligand molecules were saved in *.sdf*. After downloading from the RCSB PDB, neither the receptor nor the ligand underwent any structural optimization or modification.

Each complex deposited in PDBrt contains links to external databases such as RCSB PDB, PDBj, PDBe and PDBsum in addition to the information described above.

Table 3.2: Details of primary data submitted to PDBrt. The values are taken from the published manuscripts.

Fieldname	**Format**	**Unit**
PDB ID	String	
Ligand Name	String	
Ligand Inchi	String	
Ligand Smiles	String	
Ligand Formula	String	
Ligand Code	String	
Protein Name	String	
Protein Organism	String	
Complex Name	String	
Release Year	Integer	
Primary Reference	String	
Residence Time	Float	min
Residence Time Error		min
K_i	Float	nM
K_i Error	Float	nM
k_{on}	Float	$min^{-1}M^{-1}$
k_{on} Error	Float	$min^{-1}M^{-1}$
k_{off}	Float	min^{-1}
k_{off} Error	Float	min^{-1}

The PDBrt database is publicly available at https://pdbrt.polsl.pl/. Its source code is available in the GitHub repository: https://github.com/mlugowska/residence_time.

3.2 Mycobacterium enoyl acyl carrier protein reductase (InhA)

The global fight against the spread of multidrug-resistant tuberculosis (MDR-TB) has created an urgent need for new chemotherapeutic agents. The enoyl acyl carrier protein (ACP) reductase (InhA) is clinically one of the few targets for TB drug discovery. It is involved in fatty acid synthesis, mainly mycolic acid biosynthesis in *M. tuberculosis* (Prasad et al. 2021).

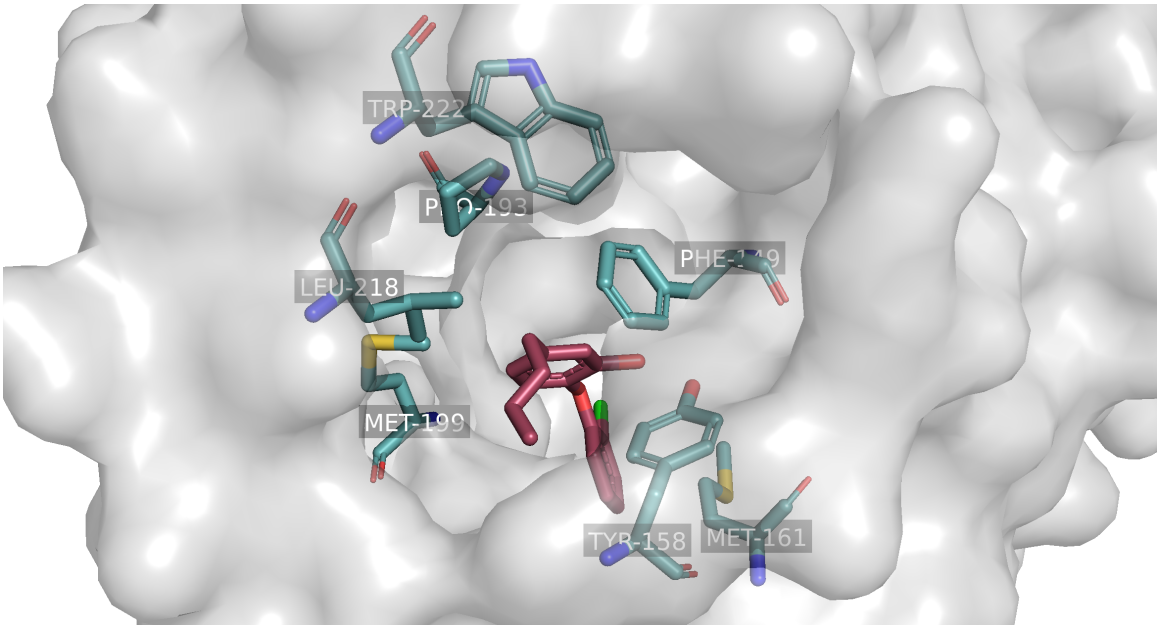


Figure 3.2: Crystallographic structure of InhA enzyme (example: PDB complex ID: 4OYR) visualized with PyMOL. The structure of the enzyme is shown in a "surface view" oriented towards the ligand binding pocket. Amino acids shown in the "sticks" representation that make up the active site of the enzyme: Phe149, Tyr158, Met161, Met199, Pro193, Leu218, and Trp222. In the binding pocket is the ligand (1US).

The enzymes are structurally unique, with deeper drug binding pockets in the active site. This makes the enzyme a unique target for anticancer drug development. InhA is

also the primary molecular target of the most potent anti-tuberculosis drug, isoniazid. Since this drug has no human orthologue, resistance problems can be avoided. This example shows that one of the most effective ways to combat tuberculosis is to inhibit Mycobacterium enoyl acyl carrier protein reductase (InhA).

The study used 10 InhA inhibitors for which both experimental kinetic measurements and crystallographic structures were available (see Figure 3.3a).

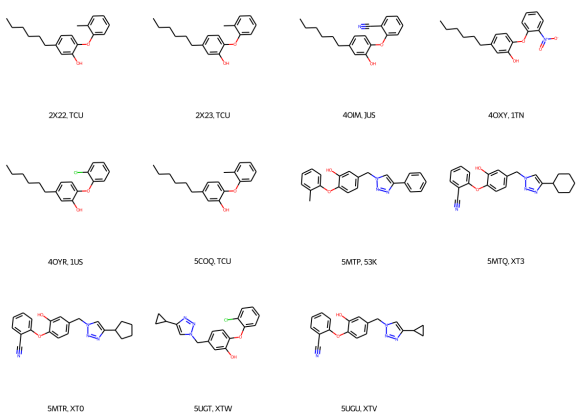
3.3 Heat-shock protein 90 (HSP90)

The most commonly studied systems for predicting kinetic parameters, including residence time, are heat shock proteins (HSP90) in complex with their inhibitors. They belong to the family of chaperone proteins. At the molecular level, Hsp90 proteins are involved in the folding of other proteins. Thus, they play a protective role in stabilizing proteins during heat shock. Because of their known role in stabilizing many proteins essential for tumor growth, they are also an anti-cancer target (Neckers et al. 1999). Hypoxia, low pH, and poor nutrition in cancer promote protein destabilization, which further increases dependence on Hsp90 activity (Solit & Chiosis 2008).

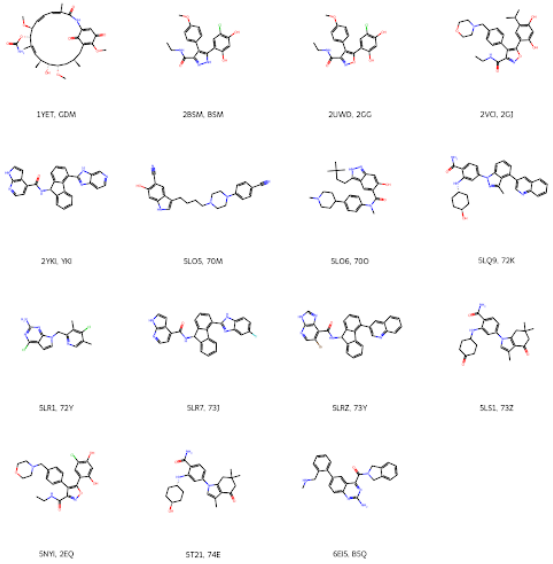
A total of 15 HSP90 inhibitors were used in the study, for which both experimental kinetic measurements and crystallographic structures were available (see Figure 3.3b).

3.4 Other study receptor-ligand systems

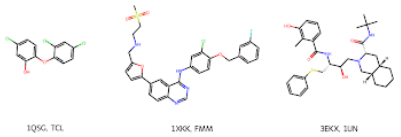
In order to validate the methods for the structural diversity of the ligands, the receptor-ligand systems shown in Figure 3.3c were used for the analysis. These are compounds complexed with (left to right) ENR, EGFR and HIV-1 PR.



(a)



(b)



(c)

Figure 3.3: 2D chemical structures of the (a) InhA inhibitors (b) HSP90 inhibitors and (c) other compounds used in the study. The RDKit Python library was used for the 2D visualizations.

Chapter 4

Relative residence time estimation using τRAMD

A dataset of HSP90 inhibitors was used as the first published example of using RAMD to predict drug residence time at a target site (Kokh et al. 2018). The τRAMD method was presented that extends RAMD with an ensemble method to obtain more accurate results for predicting average residence time. The authors correlate the results with experimental times. In this chapter, the accuracy of using τRAMD to predict drug residence time for molecular targets is evaluated. For this purpose, the published results of Kokh et al. were analyzed and extended to include the Geldanamycin molecule complexed with HSP90. This protocol was then repeated for 10 InhA ligands and one each of ENR, EGFR, and HIV-1 ligands. The goal of the analysis was to test the versatility and reproducibility of τRAMD.

4.1 Ligand similarity

Molecular structure comparison is one of the basic techniques used in various molecular modeling methods. Structurally similar molecules are expected to have similar

pharmacological profiles (Flower 1998). Using similarity measures allows to numerically determine the similarity of a set of ligands. A common measure is Tanimoto's similarity (Bajusz et al. 2015). Binary fingerprints are usually calculated for compared molecular structures. The fingerprint consists of a list of pre-defined structural fragments, or features, that are either present or absent in the structure. Each feature present is represented by a set bit (1).

$$S_{A,B} = \frac{N_{A,B}}{N_A + N_B - N_{A,B}} \tag{4.1}$$

where $S_{A,B}$ - the similarity of molecules A and B, N_A - the number of features "contained" in the structure of A [1, 0], that is, the situation when a certain structural characteristic expressed by the fingerprint is present in molecule A and not in B, N_B - the number of characteristics "contained" in the structure of B [0, 1], $N_{A,B}$ - the number of features (bits) "contained" in both fingerprints A and B [1, 1].

The Tanimoto similarity measure ranges between 0 and 1. $S_{A,B} = 1$ means that A and B are similar, not equal. It is assumed that a value of 0.85 ($S_{A,B} >= 0.85$) is the threshold for determining the high similarity of two chemical structures that they will have similar biological activity.

The ligand similarity analysis was performed using the RDKit library in Python. The following steps were included in the analysis protocol: (i) create a list of SMILES strings from which 2D molecular structures of the ligands were produced; (ii) drop ligands with the same SMILES string; (iii) generate molecular fingerprints for each molecule using the RDKit generator; (iv) compute the Tanimoto similarity measure for each ligand pair; (v) create sililarity matrix that presents data only once without repetition; (vi) find a suitable similarity score threshold above which a given percentage of the compound pairs are considered similar, (vii) cluster the obtained results using the Linking Hierarchy Cluster (LHC). The above analysis was performed for both

separate sets of ligands complexed with InhA (hereafter referred to as Set 1) and HSP90 receptors (hereafter referred to as Set 2), as well as for a set containing ligands complexed with ENR, EGFR, and HIV-1 (Set 3), and for all tested systems (Set 4). Table 4.1 shows details of the results obtained. Figure 4.1 shows the distribution of similarity scores for each set.

Table 4.1: Summary of ligand similarity analysis using Tanimoto coefficient.

	Set 1	**Set 2**	**Set 3**	**Set 4**
Number of fingerprints	9	15	3	27
Total compound pairs	81	225	9	729
Mean $S_{A,B}$	0.34	0.27	0.41	0.2
$S_{A,B}$ >= 0.85	13 (16.05%)	19 (8.44%)	3 (33.33%)	35 (4.8%)
$S_{A,B}$ >= 0.75	14 (17.28%)	20 (8.89%)	3 (33.33%)	37 (5.08%)
$S_{A,B}$ >= 0.65	17 (21%)	22 (9.78%)	3 (33.33%)	42 (5.76%)
$S_{A,B}$ >= 0.55	21 (25.92%)	26 (11.56%)	3 (33.33%)	51 (7%)
$S_{A,B}$ >= 0.45	28 (34.57%)	50 (22.22%)	3 (33.33%)	84 (11.5%)
$S_{A,B}$ >= 0.35	34 (41.98%)	105 (46.67%)	4 (44.44%)	170 (23.32%)
$S_{A,B}$ >= 0.2	45 (55.56%)	120 (53.33%)	4 (44.44%)	329 (45.13%)
The similarity score threshold above which a given percentage of compound pairs are considered similar				
95% of compound pairs	78	218	8	707
score at 95% percentile	0.79	0.35	1.0	0.43

Analysis of set 1 revealed that if two randomly selected compounds have a similarity score of 0.34 on average. If two ligands have a Tanimoto measure score of 0.35, which is near the average, the random compounds have a 42% chance of having a similarity score above 0.35. Therefore, it can be concluded that the two compounds are not similar to each other. The 95% percentile value is 0.79. This means that this value can be considered as a similarity score threshold, above which a certain percentage of compound pairs are determined to be similar. This is approximately 18% of the compound pairs.

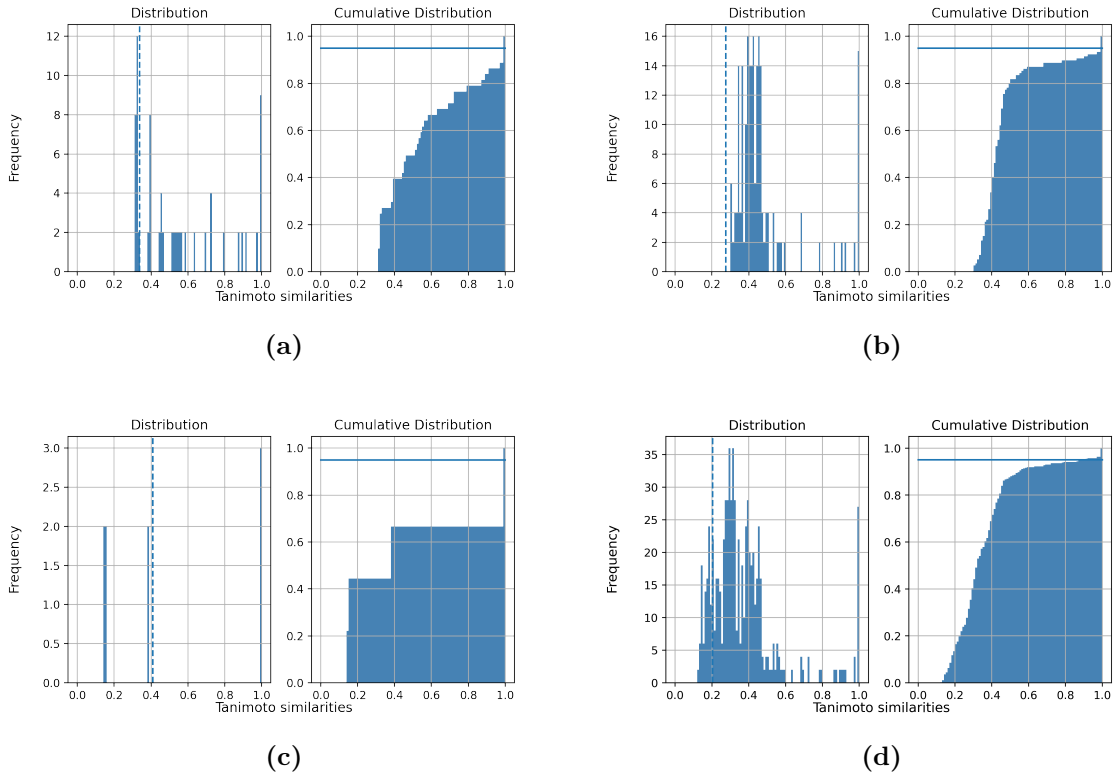


Figure 4.1: Histogram showing the distribution of scores between pairs of compounds for A) set 1, B) set 2, C) set 3, and D) set 4. Dotted line shows average.

In set 2, the average similarity score is 0.27, according to the similarity distribution between pairs of compounds. Approximately 12% of the randomly selected pairs of compounds have a similarity value that is greater than or equal to 0.55, and 8% of the pairs have a similarity value of at least 0.85. The value of the 95% percentile, i.e. the similarity threshold of the pairings, was found to be 0.35. This means that about 46% of the randomly selected pairs of compounds in the set could be described as similar.

For set 3, one can observe from the similarity distribution of compound pairs that the average similarity score of two random compounds is 0.41. This is the highest value among the studied sets. A similarity score greater than or equal to 0.55 or even at least 0.85 is observed for about 33% of the randomly selected compound pairs. If

two ligands have a Tanimoto measure score of .45, which is near the average, then the random compounds have a 33% chance of having a similarity score of .45 and above. Thus, it is reasonable to assume that the two compounds are not similar. The 95% percentile is 1, meaning this is the similarity threshold that none of the studied pairs exceeded.

For set 4, from the distribution of similarity between pairs of compounds, it can be observed that two randomly selected compounds have an average similarity score of 0.2. About 7% of randomly selected pairs of compounds have similarity scores of 0.55 and above, and about 6% have similarity scores of 0.65 and above. A similarity score greater than or equal to 0.85 is found in only 4.8% of the compounds studied. There is a 45% chance that the randomly selected compounds will have a similarity score greater than or equal to 0.2 if two ligands have a Tanimoto measure score of 0.2, which is close to the mean.

Therefore, the conclusion that there is no similarity between the two compounds would be a reasonable one. Tanimoto scores can range from 0 (no resemblance) to 1 (similar molecules), with a mid-range of 0.5. Therefore, to consider two compounds as similar, a Tanimoto score of 0.55 is not sufficient. Based on the generated score distribution graph, only about 7% of the selected compound pairs had a score above that. However, calculating the 95% percentile, which is the number greater than 95% of the values in the dataset, yielded a value of 0.43. This means that this value can be considered as a similarity score threshold, above which a certain percentage of compound pairs are determined to be similar. This is approximately 11% of the compound pairs.

The similarity scores were plotted on a triangular correlation heatmap (see Figure 4.3), where the data were represented in a single, nonrepetitive manner, i.e., the categories were correlating with each other in a single instance. The idea of obtaining a

triangular correlation map is to remove the data above it so that it is represented only once, since the data is symmetrical on the diagonal from the top left to the bottom right. The items on the diagonal are where categories of the same type correlate.

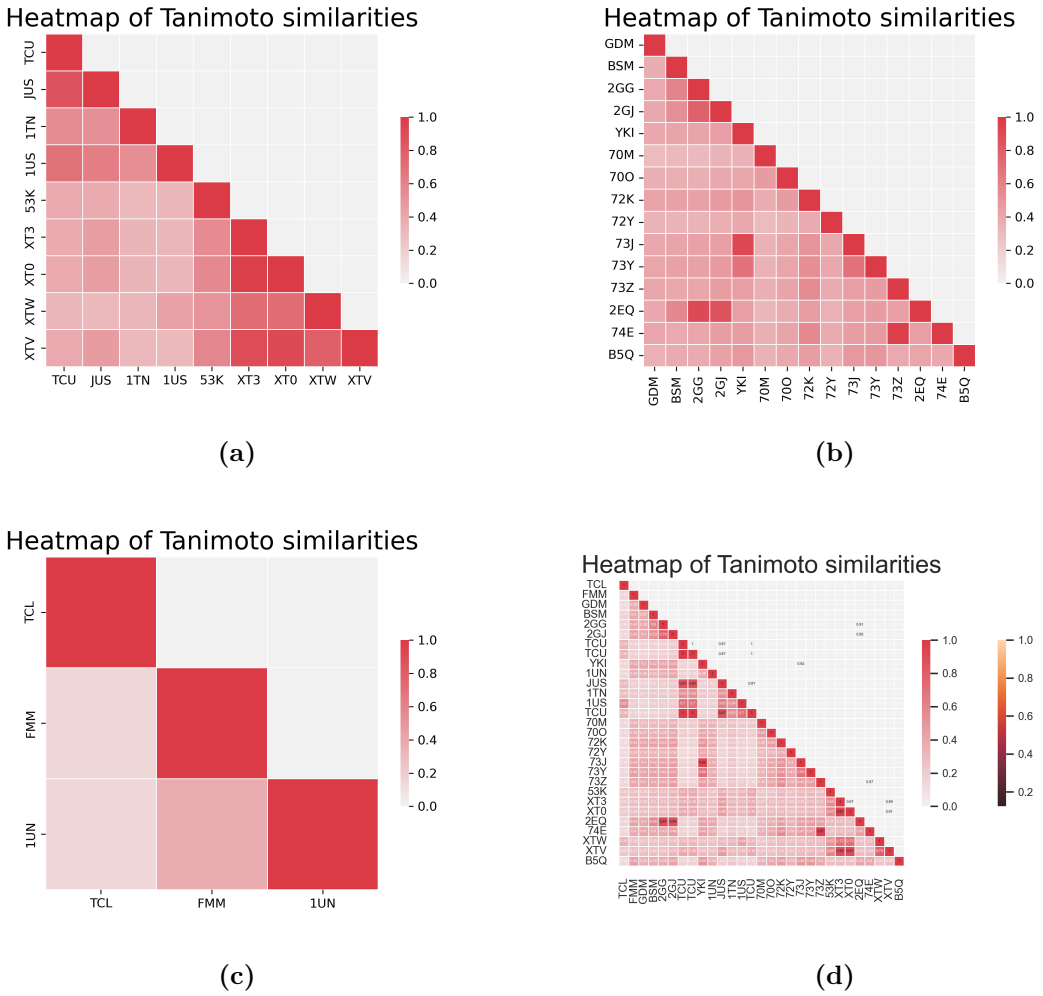
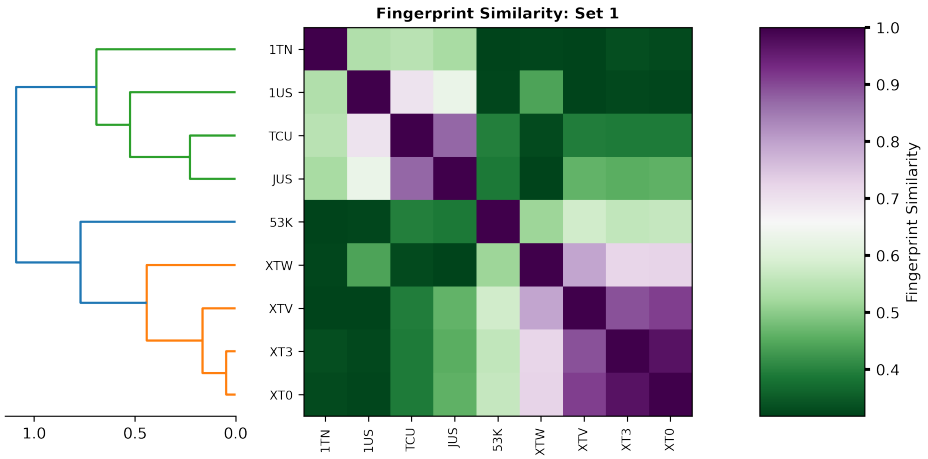


Figure 4.2: Triangular correlation heatmap for a) set 1, b) set 2, c) set 3, and d) set 4.

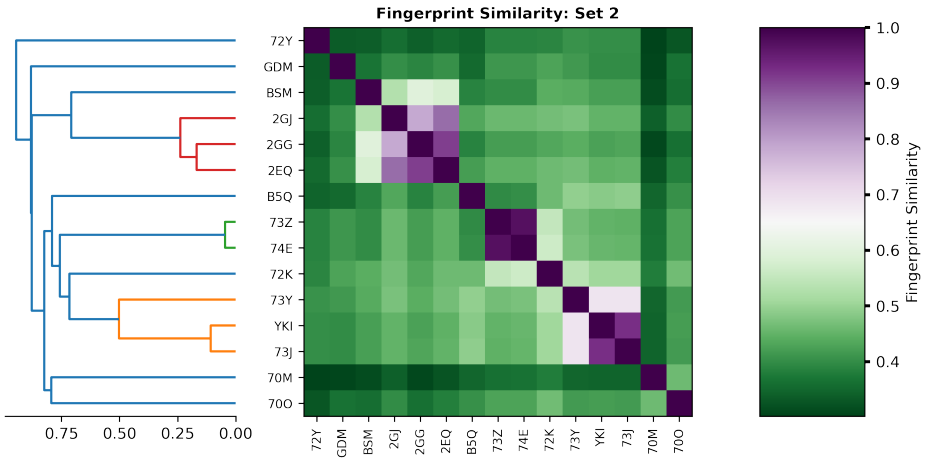
The entire set of compounds studied is diverse, as indicated by the average value of the Tanimoto similarity measure for set 4 (0.2). Ligands complexed with one receptor show less structural diversity. Their average Tanimoto values are as follows 0.34 (set 1), 0.27 (set 2) and 0.41 (set 3).

Based on the similarity of the fingerprints, the presented approach was able to identify several clusters. These are shown in Figure 4.3 along with a heat map of the molecular similarities identified.

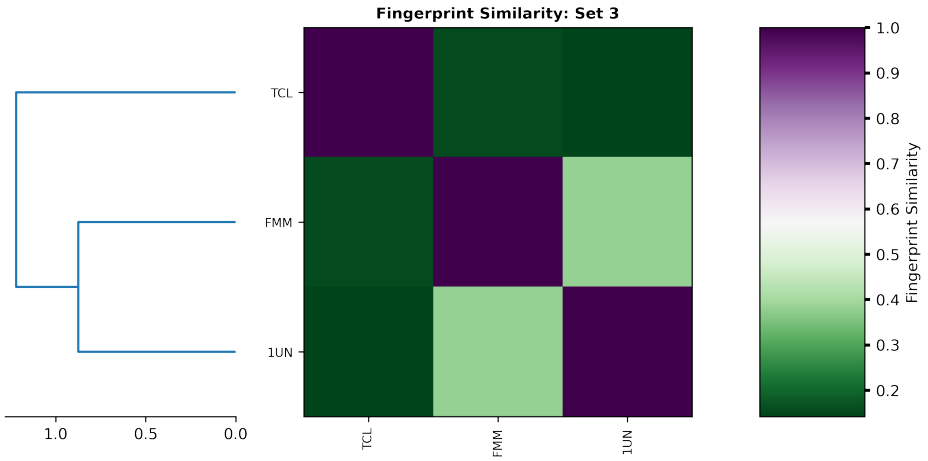
Four groups of ligands were identified based on their similarity in set 4. Group 1 is highlighted in orange. It consists of 5 ligands with the following codes: TCU (InhA), TCL (ENR), JUS (InhA), 1TN (InhA), and 1US (InhA). Group 2, shown in green, contains 5 compounds: XT0, XTV, XT3, XTW, and 53K, which form complexes with InhA protein.



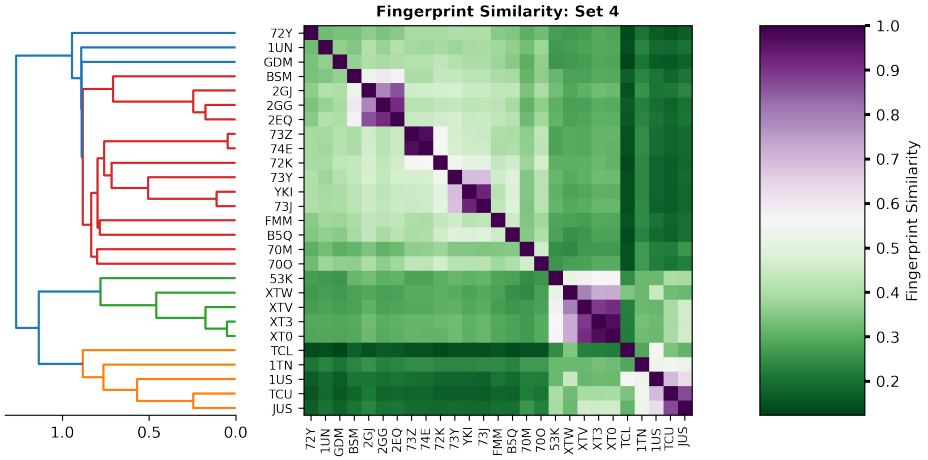
(a)



(b)



(c)



(d)

Figure 4.3: Heatmaps of molecular similarity by Tanimoto similarity index for (a) set 1 (b) set 2 (c) set 3 and (d) set 4. On the left are the results of the clustering of compounds in the form of dendrograms, which show the relationships between the objects of the particular set.

Group 3, shown in red, contains 14 ligands that form complexes with HSP90 protein, as well as one from group 3: FMM (EGFR). The last group, marked in blue, contains 3 compounds: 72Y (HSP90), GDM (HSP90) and 1UN (HIV-1).

The similarity within the groups is confirmed by the identified ligand groups. Sep-

arate groups were identified for the compounds in set 1 that form a complex with the InhA protein. Similarly, the compounds in group 2 that form a complex with HSP90 are similar, with the exception of 72Y and GDM. Two of the compounds in set 3 are structurally similar to HSP90 inhibitors and one is similar to InhA inhibitors.

4.2 Application for HSP90 inhibitors

A τRAMD method for predicting the residence time of a drug in a target was presented by (Kokh et al. 2018). Section 2.1 describes the protocol of the method. The authors applied τRAMD to a set of 70 HSP90 ligands with different chemical compositions and showed that for 55 of them there is a strong correlation between the average length of the dissociation trajectory (τ_{comp}) and the experimentally measured residence time. Furthermore, for congeneric, i.e. structurally similar, sets of ligands, a correlation between τ_{comp} and the experimentally determined residence time was observed. The authors concluded that τRAMD is an efficient method with broad applicability for optimizing the residence time of a drug target (Kokh et al. 2018).

15 HSP90 inhibitors were analyzed using the τRAMD protocol, 14 of which are included in the published data set and one, with PDB ID: 1YET, which was not previously analyzed, added in order to verify the reproducibility of the results and thus the reliability of the method. The molecular models are publicly available and all compounds were crystallized in the inhibitor-HSP90 complex. Thus, the same initial structure can be used for testing. Table 4.2 summarizes the results.

Table 4.2: Summarized information for 15 HSP90 inhibitors. The experimentally determined residence time, τ_{exp} (min), was taken from PDBrt. τ_{comp} (Kokh et al. 2018) (ns) is the calculated relative residence time from the manuscript. τ_{comp} repeat (ns) is the repeated calculated relative residence time averaged over 5 sets of τRAMD simulations performed for each system.

PDB	τ_{exp} (min)	τ_{comp} (Kokh et al. 2018) (ns)	SD_{comp} (Kokh et al. 2018) (ns)	τ_{comp} repeat (ns)	SD_{comp} repeat (ns)
1YET	400	n.a.	n.a.	1.67	1.17
2BSM	1.7	2.2	0.65	3.04	1.81
2UWD	7.9	2.6	0.98	0.09	0.02
2VCI	167	13.0	3.6	0.29	0.21
2YKI	58.5	4.0	1.28	3.86	1.53
5LO5	0.04	1.0	0.29	1.25	0.82
5LO6	12	2.5	0.76	0.05	0.04
5LQ9	122.5	4.8	1.33	4.29	1.15
5LR1	0.7	0.1	0.03	0.15	0.06
5LR7	88.2	4.6	1.58	3.99	1.59
5LRZ	60	1.5	0.46	1.82	0.94
5LS1	34.4	3.6	1.64	2.0	0.56
5NYI	166.7	8.8	2.11	4.39	1.27
5T21	21.8	2.6	0.75	2.18	0.46
6EI5	3.7	0.16	0.04	0.15	0.05

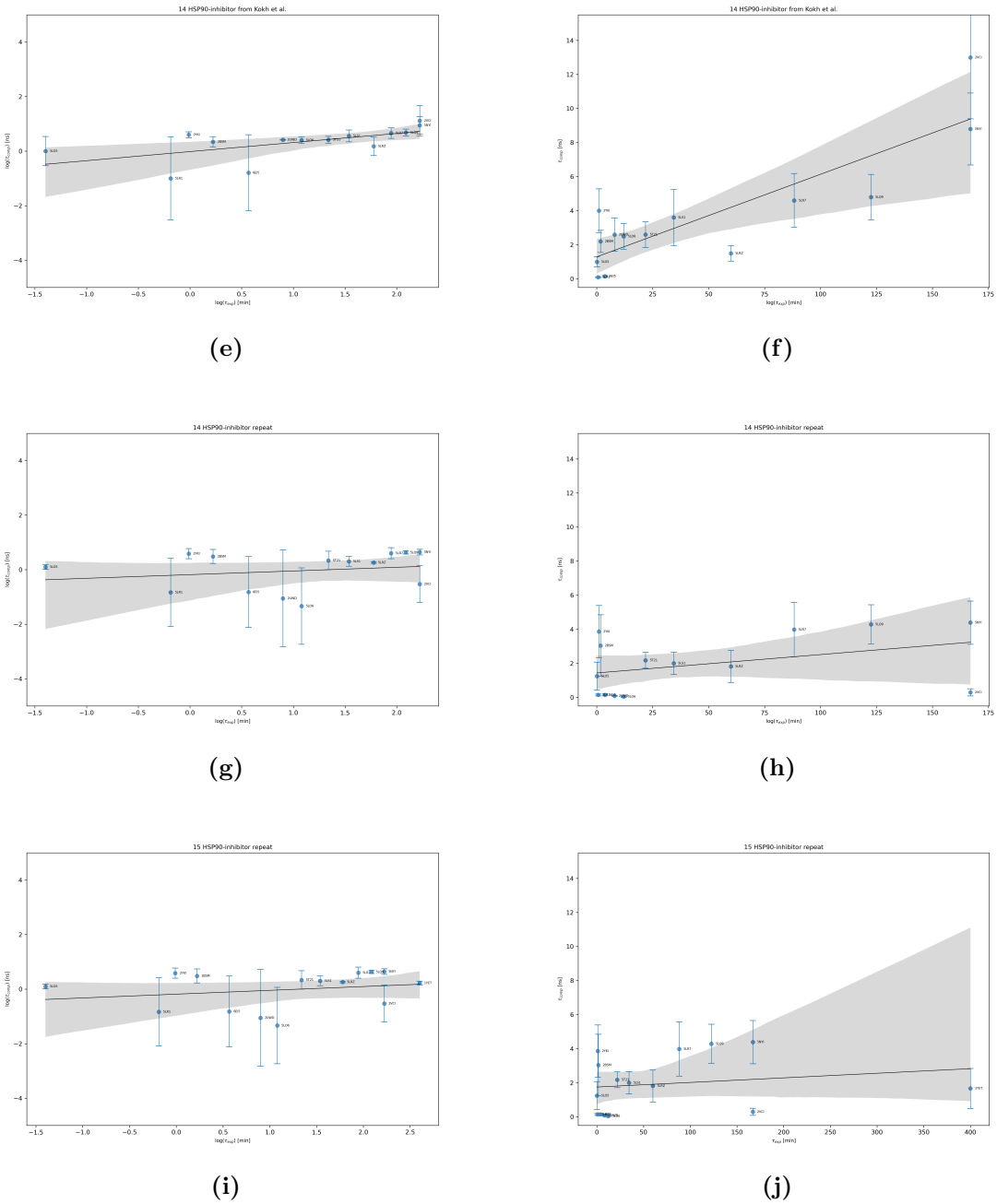
Since the relative residence time calculated in present work is the average of 5 sets of τRAMD simulations for each system, while in (Kokh et al. 2018) the average is of 4 to 8 sets, the results are found to be in good agreement, except for the 2VCI system. The results from the table are plotted as log and ordinal (Figure 4.4) scales using linear fitting. The black line shows the linear regression of all points except the

gray area, which indicates the area within the residual standard deviation of the linear fit calculated at 0.95 confidence. The error bars show the standard deviations of the calculated residence times.

Pearson's coefficient was used to determine the correlation of the data. For the results presented in (Kokh et al. 2018) (Figure 4.4a and 4.4b), the correlation coefficient is $R^2 = 0.86$, which indicates a very strong correlation between the calculated and the experimental data. For the repeated simulations of exactly the same systems (Figure 4.4c and 4.4d), except for the 2VCI complex, the correlation coefficient is R^2=0.4, which indicates a moderately positive relationship. The data with the 1YET-complex (Figure 4.4e and 4.4f) show a weak correlation with R^2=0.2.

Note that for the 5LR1, 6EI5, 2UWD, 5LO6 and 2VCI systems, the calculated standard deviation is relatively high, which indicates high variability in individual simulation time. The insufficient number of τRAMD simulation sets is a possible reason for this result. For the remaining systems, the standard deviation assumes low levels, suggesting low dispersion around the mean.

For each system, a Gaussian distribution of the residence times obtained from different output replicas was generated (see Figure 4.5). The top row of the single graph shows the variation of calculated effective residence times from τRAMD simulations for different initial replicates for the 15 studied HSP90 complexes - Gaussian distributions of residence times were generated using a bootstrapping procedure applied to a set of dissociation times (simulations) of τRAMD; red lines indicate the obtained residence times, t_r, the black line indicates the probability density function (*pdf*). The bottom row of the graph shows a comparison of the Poisson cumulative distribution function (*cdf*), indicated by the black line, and the empirical cumulative density function (*ecdf*) obtained from the dissociation probability distribution (observed data), represented by blue dots. The red line shows the mean value of the residence time, τ_{comp}. Each graph



(e) **(f)**

(g) **(h)**

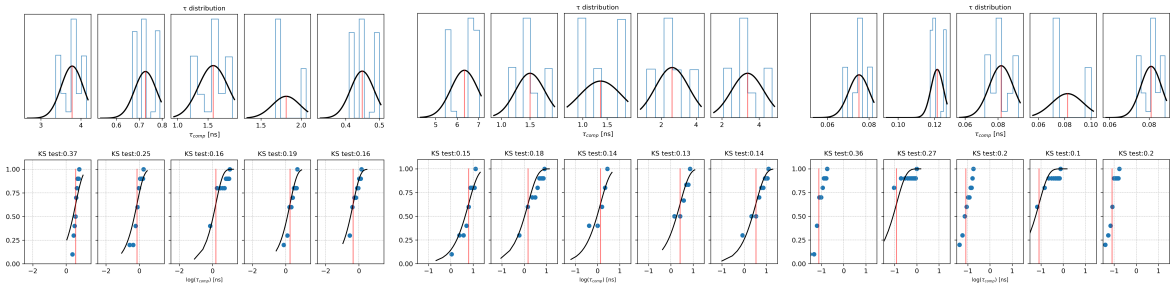
(i) **(j)**

Figure 4.4: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic scale (left) and ordinal (right) for (a, b) 14 HSP90 inhibitors and published results (c, d) 14 HSP90 inhibitors and replicate results (e, f) 15 HSP90 inhibitors.

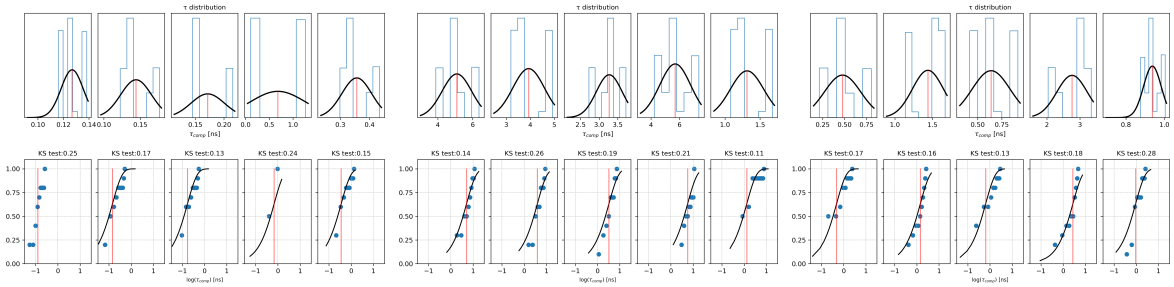
corresponds to a specific compound, for which the results of the Kolmogorov-Smirnov test, KS, are also shown.

The Kolmogorov-Smirnov test statistic took values below 0.37 for each tested system as the maximum vertical distance between *cdf* and *ecdf*. For a sample size of n=10 (i.e., 10 dissociation trajectories for each set of τRAMD simulations), the assumed confidence level α=0.05 is 0.41. The null hypothesis that the distributions are similar is true (can be accepted) when comparing the two values. That is, the distribution of measured dissociation times is similar to a Poisson distribution with time t_r.

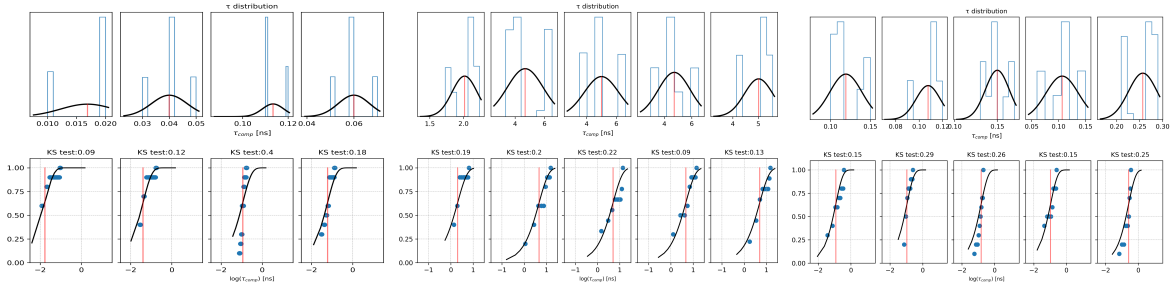
Boxplots were also generated to provide information about the location, dispersion, and shape of the data distribution (see Figure 4.6). A single box represents data from a series of τRAMD simulations (i.e., 10 dissociation trajectories). The red line represents the mean and the orange line represents the median. The open circle indicates outliers.



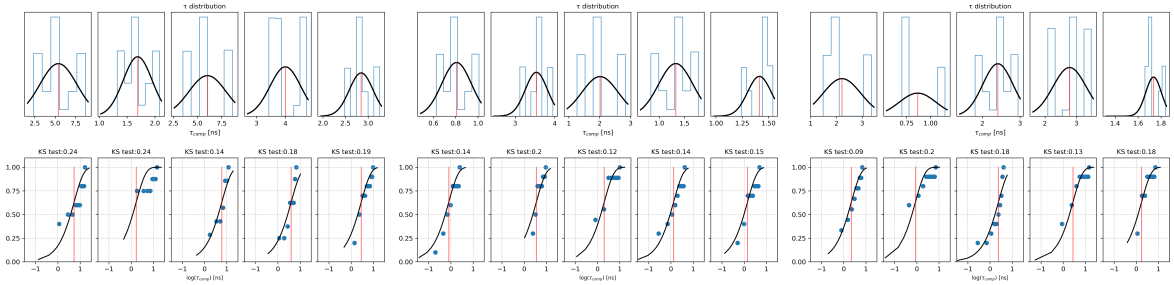
(a) 1YET **(b)** 2BSM **(c)** 2UWD



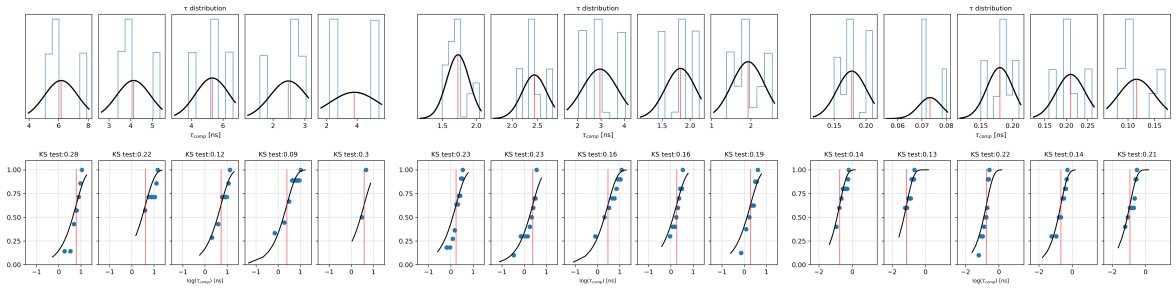
(d) 2VCI **(e)** 2YKI **(f)** 5LO5



(g) 5LO6 **(h)** 5LQ9 **(i)** 5LR1

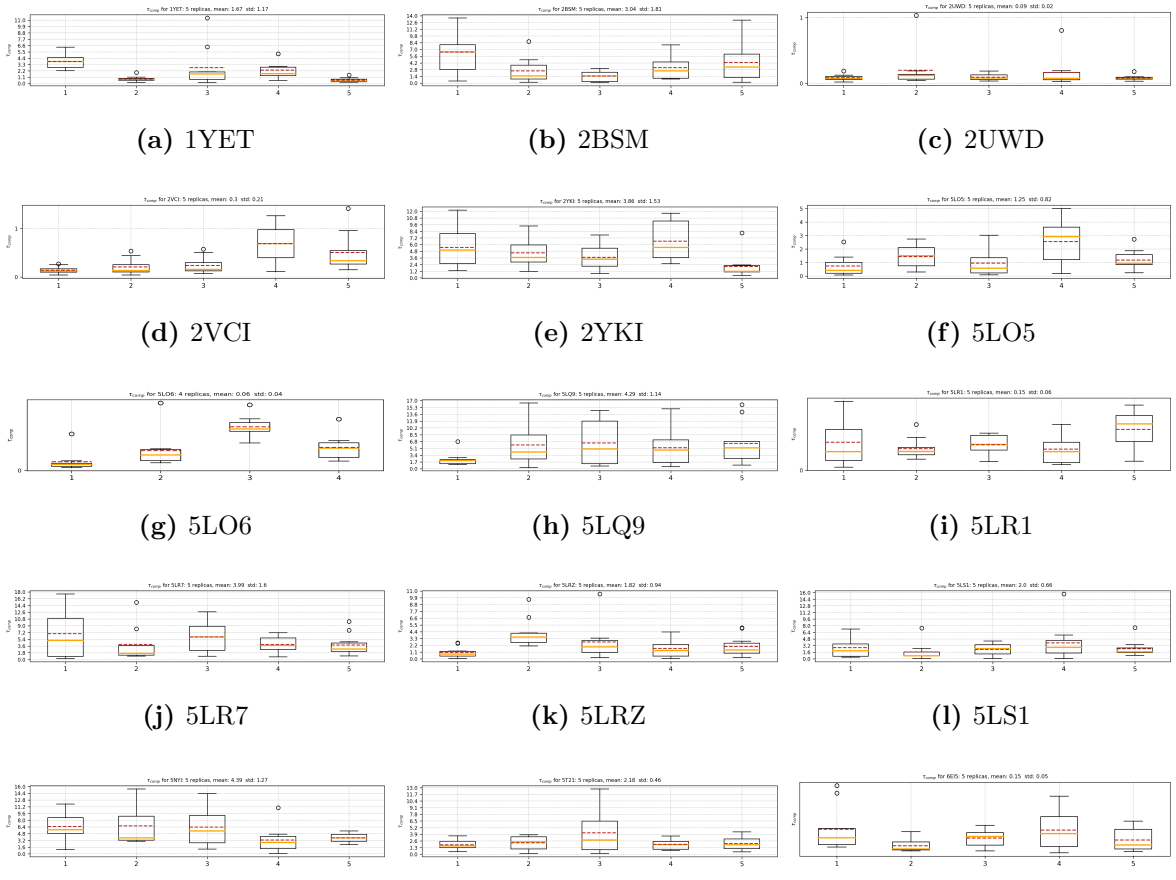


(j) 5LR7 **(k)** 5LRZ **(l)** 5LS1



(m) 5NYI **(n)** 5T21 **(o)** 6EI5

Figure 4.5: Distribution analysis of the residence times obtained from the τRAMD dissociation trajectories for 15 HSP90 inhibitors.



(a) 1YET **(b)** 2BSM **(c)** 2UWD

(d) 2VCI **(e)** 2YKI **(f)** 5LO5

(g) 5LO6 **(h)** 5LQ9 **(i)** 5LR1

(j) 5LR7 **(k)** 5LRZ **(l)** 5LS1

(m) 5NYI **(n)** 5T21 **(o)** 6EI5

Figure 4.6: Box plots of τRAMD dissociation trajectory residence times for 15 HSP90 inhibitors.

4.3 Application for InhA inhibitors

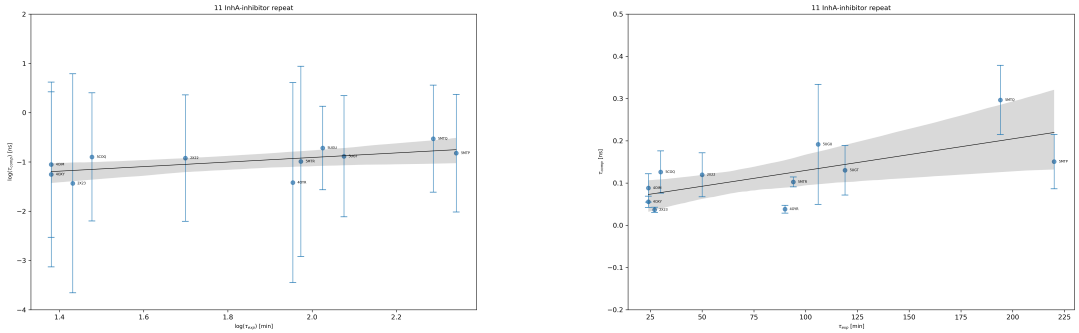
A set of 10 ligands of the InhA protein was subjected to τRAMD analysis (11 complexes). A summary of the results is given in Table 4.3.

Table 4.3: Summarized data for 11 InhA inhibitors. Experimental residence time, τ_{exp} (min), was obtained from PDBrt. τ_{comp} (ns) is the calculated relative residence time averaged over 5 τRAMD simulations for each system.

PDB	τ_{exp} (min)	τ_{comp} (ns)	SD_{comp} (ns)
2X22	24	0,09	0,03
2X23	24	0,06	0,01
4OIM	50	0,12	0,05
4OXY	27	0,04	0,005
4OYR	90	0,04	0,009
5COQ	30	0,13	0,05
5MTP	94	0,1	0,01
5MTQ	119	0,13	0,06
5MTR	106	0,19	0,14
5UGT	220	0,15	0,06
5UGU	194	0,3	0,08

The results were plotted on a logarithmic scale (Figure 4.7a) and an ordinal scale (Figure 4.7b) with a linear fit.

The correlation of the data was determined by Pearson's coefficient, the value of which is R^2=*0.68*, which indicates a relatively strong correlation between calculated and experimental data. It can be noted that for the 5UGU, 5MTQ, 5MTP, 5UGT systems, the calculated standard deviation takes on relatively high values, indicating a large variation in individual simulation times. In this case, too, a possible reason for this result is the insufficient number of τRAMD simulation sets. The standard

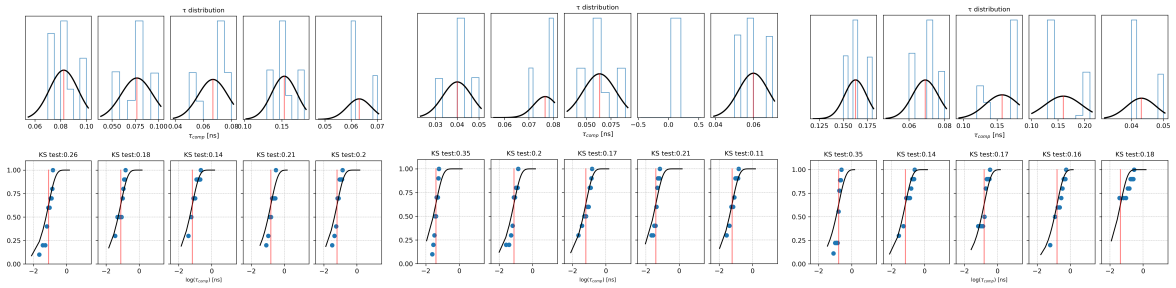


(a) **(b)**

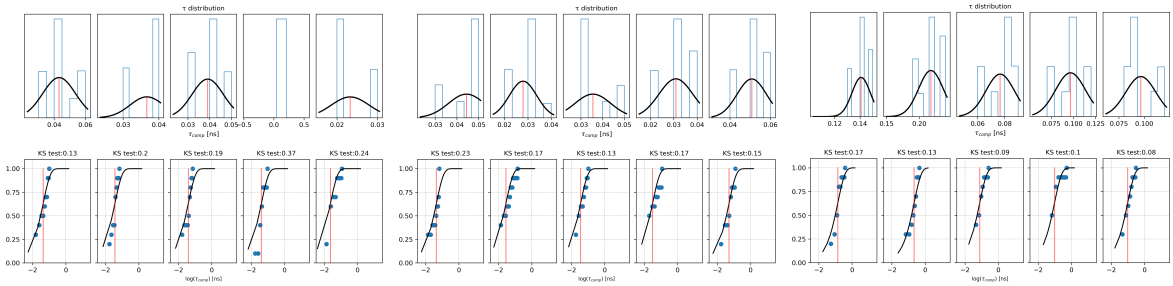
Figure 4.7: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic scale (left) and ordinal (right) for 10 InhA inhibitors (11 complexes).

deviation for the remaining layouts takes on low values, indicating a small dispersion of values around the mean.

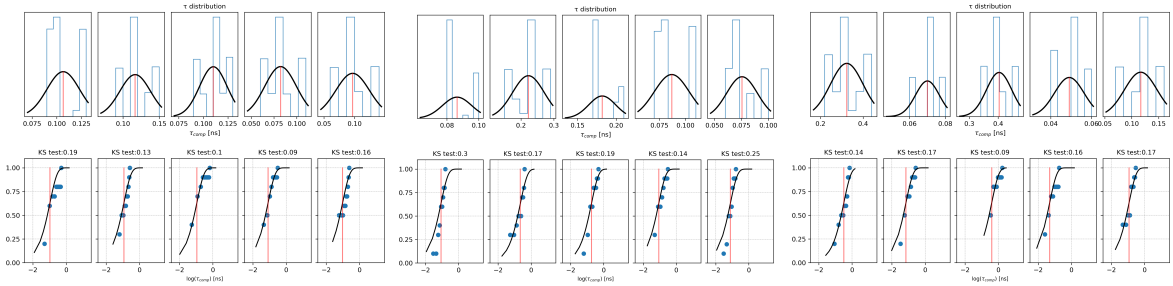
For each system, a Gaussian distribution of the obtained residence times from different output replicas was generated (Figure 4.8). The low D-values of the Kolmogorov-Smirnov test (the highest calculated statistic is less than 0.37), with an assumed confidence level of α=0.05 for a sample size of n = 10 (i.e., 10 dissociation trajectories for each set of τRAMD simulations) of 0.41, indicate that the null hypothesis that the distributions are similar can be accepted. This implies that the measured dissociation time distribution resembles a Poisson distribution with time t_r.



(a) 2X22 **(b)** 2X23 **(c)** 4OIM



(d) 4OXY **(e)** 4OYR **(f)** 5COQ

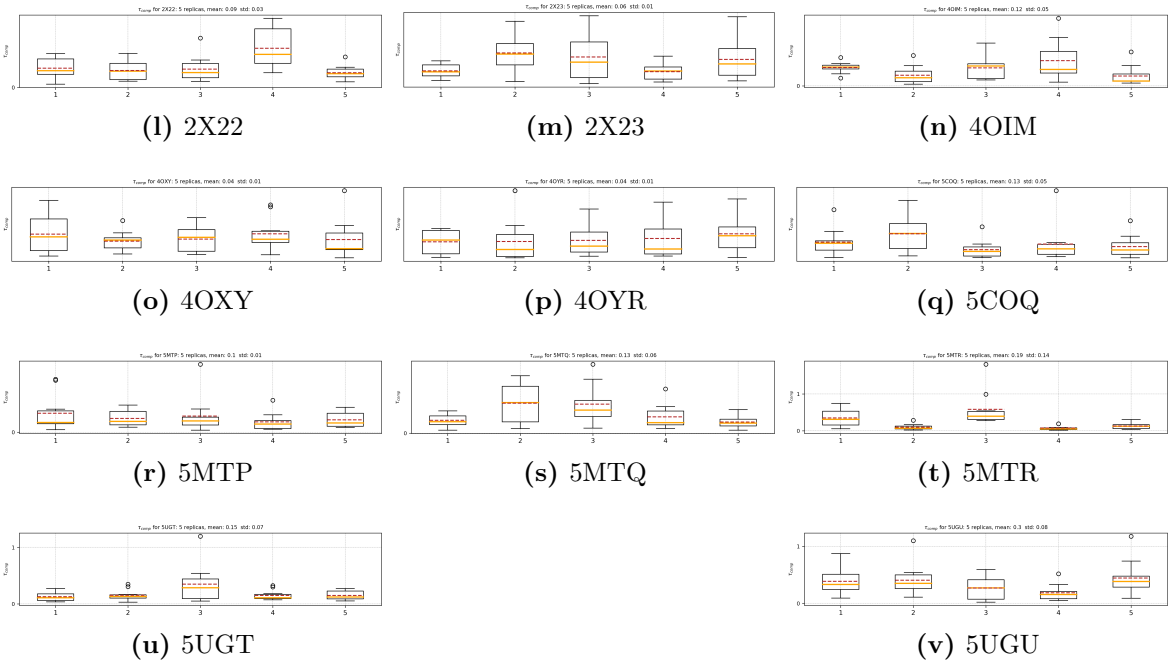


(g) 5MTP **(h)** 5MTQ **(i)** 5MTR

(j) 5UGT **(k)** 5UGU

Figure 4.8: Distribution analysis of the residence times obtained from the τRAMD dissociation trajectories for 11 InhA inhibitors.

Boxplots containing information about the location, dispersion and shape of the data distribution were also generated (Figure 4.9).



(l) 2X22 **(m)** 2X23 **(n)** 4OIM

(o) 4OXY **(p)** 4OYR **(q)** 5COQ

(r) 5MTP **(s)** 5MTQ **(t)** 5MTR

(u) 5UGT **(v)** 5UGU

Figure 4.9: Box plots of τRAMD dissociation trajectory residence times for 10 InhA inhibitors (11 complexes).

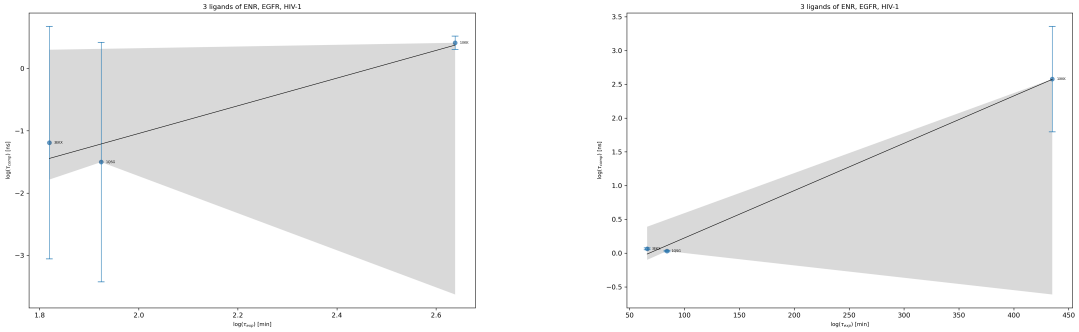
4.4 Application for ENR, EGFR and HIV-1 ligands

τRAMD analysis was then performed for ENR, EGFR and HIV-1 ligands. Table 4.3 summarizes the results of the calculated relative dissociation times together with the original data.

Table 4.4: Summarized data for for 3 ligands of ENR, EGFR and HIV-1. Experimental residence time, τ_{exp} (min), was obtained from PDBrt. τ_{comp} (ns) is the calculated relative residence time averaged over 5 τRAMD simulations for each system.

PDB	τ_{exp} (min)	τ_{comp} (ns)	SD_{comp} (ns)
1QSG	84	0.03	0.01
1XKK	435	2.58	0.78
3EKX	66	0.06	0.014

The results from the table were visualized in a scatter plot on a logarithmic scale (Figure 4.10a) and an ordinal scale (Figure 4.10b) with a linear fit performed.



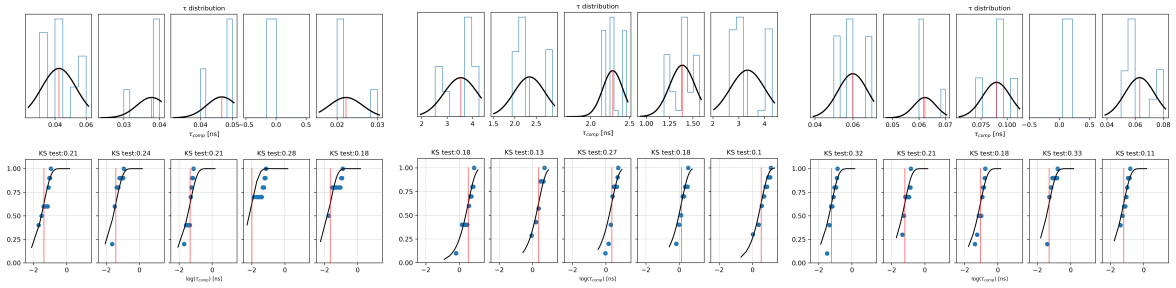
(a) **(b)**

Figure 4.10: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic scale (left) and ordinal (right) for 3 ligands of ENR, EGFR and HIV-1.

The Pearson correlation coefficient, R^2=*0.99*, indicates that the relationship between the calculated and experimental data is very strong.

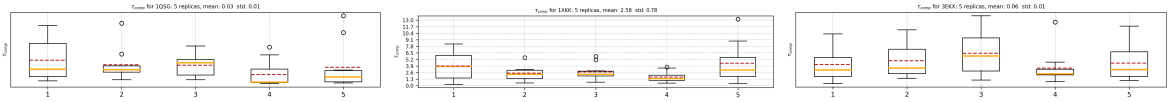
For the 1XKK system, the calculated standard deviation assumes high values, indicating a high variability of the individual simulation times. A possible reason for this result is also the insufficient number of τRAMD simulation sets. The standard deviation for the other layouts becomes low, indicating low scatter around the average. For each system, the Gaussian distribution of the residence times obtained from the different output replicas was plotted (Figure 4.11), along with boxplots (Figure 4.12).

The highest value of D in the data set examined is 0.33, indicating that the null hypothesis of similarity of distributions can be accepted, given a confidence level of α=0.05 and a sample size of n=10 (10 dissociation trajectories for each set of τRAMD simulations). The distribution of the dissociation times is comparable to the Poisson distribution with the time t_r.



(a) 1QSG **(b)** 1XKK **(c)** 3EKX

Figure 4.11: Distribution analysis of the residence times obtained from the τRAMD dissociation trajectories for for 3 ligands of ENR, EGFR and HIV-1.

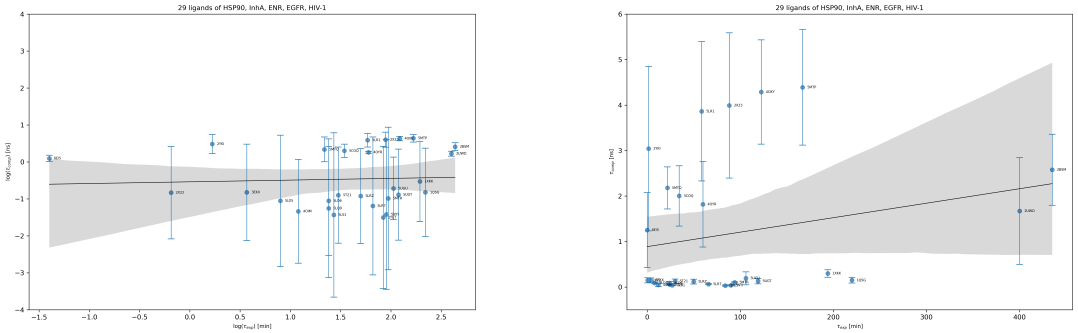


(a) 1QSG **(b)** 1XKK **(c)** 3EKX

Figure 4.12: Box plots of τRAMD dissociation trajectory residence times for 3 ligands of ENR, EGFR and HIV-1.

4.5 Application for all systems under study

In order to verify the universality of the method, the correlation of $\tau_{komp}$with the experimentally determined time was checked for all the systems under study. The Pearson's coefficient was 0.23 (R^2 =*0.23*). This is considered a negligible correlation. This suggests that the accuracy of the method may be lower for compounds with greater structural variation.



(a) **(b)**

Figure 4.13: Correlation plot of τ_{comp} with τ_{exp} on a logarithmic scale (left) and ordinal (right) for all studied receptor-ligand systems.

4.6 Summary and Conclusion

To estimate the relative residence time, the τRAMD method was applied to 5 different protein systems. To verify the reproducibility of the method, the analysis has been performed on a set of 15 HSP90 protein inhibitors, 14 of which have been previously analyzed (Kokh et al. 2018). In addition, to test the versatility of the method, 10 inhibitors of the InhA protein and 3 ligands of the ENR, EGFR, and HIV-1proteins were tested. Thus, 28 ligands with different structures were included in the total set.

The analysis performed led to the conclusion that τRAMD is reproducible, but in some cases more simulation sets should be performed to reduce data scatter around the mean. It was observed that for compounds of similar structure, i.e. differing by a small modification, e.g. a functional group shift, the τRAMD results show a good or strong correlation with the experimentally determined residence time. However, for a structurally diverse set of ligands, the method showed poor performance, suggesting a limited application of τRAMD within the drug discovery process.

Chapter 5

Ligand properties that affect residence time

Besides determining relative residence times, τRAMD simulations provide insight into dissociation mechanisms and are a good starting point for analyzing ligand dissociation pathways from the receptor binding pocket. This chapter describes an approach to identify the key interactions occurring in the τRAMD dissociation trajectories that affect drug residence time in its molecular target for the InhA enzyme inhibitors studied.

5.1 Feature generation

Section 2.2 describes the procedure for generating fingerprints of ligand-receptor interactions from τRAMD dissociation trajectories. First, the interactions defined as features for the machine learning algorithm were extracted. The number of simulation snapshots (a snapshot is the next observed conformational change in the system) and the number of identified contacts are shown in Table 5.1.

Table 5.1: Summary of the number of simulation snapshots and identified receptor-ligand interactions for each compounds in all dissociation trajectories.

Complex ID \ Replica	1		2		3		4		5		Total			
	Snapshots	No. of interactions	Snapshots	No. of interactions	Snapshots	No. of interactions	Snapshots	No. of interactions	Snapshots	No. of interactions	Snapshots	No. of interactions	Snapshots	No. of interactions
2X22	8946	98	7861	118	8475	154	13108	132	6807	120	45197	205		
2X23	4195	129	7920	125	7081	157	4032	123	6596	119	29824	228		
4OIM	14395	127	9365	99	15832	203	22082	105	8631	135	70296	276		
4OXY	4745	202	3559	138	3921	142	4815	124	3843	89	20883	275		
4OYR	4207	128	4278	133	4504	144	4075	174	5804	150	23668	257		
5COQ	15134	167	21378	192	9428	117	13632	124	11781	122	71356	273	983775	832
5MTP	19052	95	13925	124	16289	172	10375	112	12434	105	72075	222		
5MTQ	9516	120	22112	207	21371	147	12089	142	8233	106	73321	286		
5MTR	100258	133	9615	135	84757	150	7356	142	13088	133	215074	235		
5UGT	12849	150	15165	148	85774	153	15095	203	14869	198	143752	309		
5UGU	85154	170	23148	208	18892	124	16365	167	74770	113	218327	294		

In 983 775 snapshots, the total number of identified interaction fingerprints was 832. The number of features was reduced in the next step by: (i) removal of the bound state on the assumption that the interaction describing this state occurs in at least 20% of the snapshots of a single dissociation trajectory; (ii) removal of noise, which is defined as a very rare event that does not affect the dissociation rate (the set threshold is the occurrence of contact in less than 5% of all trajectories of a given complex). Dissociation transients were identified along with relevant events using this approach. The data was reduced to 24 features and 35986 snapshots for further analysis. Table 5.2 lists all identified interaction fingerprints.

A preliminary identification of amino acids likely to affect drug residence time in a molecular target can be made by analyzing Table 5.2. Only in the 5MTQ complex, characterized by one of the longer residence times (119 min), compound XT3 showed hydrophobic interactions with the amino acids Ala33, Cys242, Gln31, Ile9, Leu245, Leu4, Lys7, Phe96, and Trp248, and van der Waals interactions with the amino acids Cys242, Leu4, Ser246, and Trp248. A hydrophobic interaction with the amino acid Arg41 was only observed with the ligand 1US in the 4OYR complex with a residence time of 90 min. Hydrophobic and van der Waals interactions with the amino acid Arg42 are only found in compounds with a long residence time (5MTR, 5UGT, 5UGU). The exception is the TCU compound in the 2X23 complex (residence time 24 min), where this hydrophobic interaction was also identified. XTW ligand from the 5UGT complex contacted most frequently with the Arg42 amino acid and Ile15. A similar situation was observed for interacting with amino acid Asp41, especially hydrophobic. Only the XTW compound with the longest residence time (220 min) shows a pi-cation interaction with the amino acid Arg42. Only compounds of the 4OIM and 5COQ complexes, which are characterized by short residence times, interact with the amino acid Gln99.

Table 5.2: Frequencies of identified ligand-amino acid interactions for the tested InhA protein inhibitors.

Protein residue	ALA33	ARG41	ARG42			ASP41		CYS242		GLN31	GLN99	ILE15	ILE9	LEU196	LEU245	LEU4		LYS7	PHE40		PHE96	SER246	TRP248		Residence time (min)
Interaction type	Hydrophobic	Hydrophobic	Hydrophobic	PiCation	VdWContact	Hydrophobic	VdWContact	Hydrophobic	VdWContact	Hydrophobic	Hydrophobic	Hydrophobic	Hydrophobic	Hydrophobic	Hydrophobic	Hydrophobic	VdWContact	Hydrophobic	Hydrophobic	VdWContact	Hydrophobic	VdWContact	Hydrophobic	VdWContact		
2X22	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5,81%	0%	0%	0%	24
2X23	0%	0%	8,14%	0%	0%	7,10%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0,00%	0%	0%	0%	24
4OIM	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5,48%	5,58%	0%	0%	0%	0%	0%	0%	0%	0%	7,93%	0%	0%	0%	50
4OXY	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	27
4OYR	0%	5,14%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	90
5COQ	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	6,14%	0%	0%	0%	0%	0%	0%	0%	0%	5,96%	0%	10,21%	0%	0%	0%	30
5MTP	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	94
5MTQ	5,92%	0%	0%	0%	0%	0%	0%	7,30%	6,30%	5,68%	0%	0%	7,04%	0%	8,10%	8,92%	6,95%	5,03%	0%	0%	5,18%	7,89%	8,12%	5,98%	119
5MTR	0%	0%	9,08%	0%	6,04%	7,59%	0%	0%	0%	0%	0%	6,89%	0%	0%	0%	0%	0%	0%	0%	5,23%	0%	8,77%	0%	0%	0%	106
5UGT	0%	0%	6,29%	5,89%	13,20%	7,56%	7,05%	0%	0%	0%	0%	14,36%	0%	0%	0%	0%	0%	0%	0%	5,34%	5,32%	0%	0%	0%	0%	220
5UGU	0%	0%	0%	0%	6,18%	0%	0%	0%	0%	0%	0%	9,28%	0%	8,91%	0%	0%	0%	0%	0%	6,80%	0%	7,05%	0%	0%	0%	194

Only the compound with the longest residence time in the 5UGT complex had hydrophobic interactions with the amino acids Leu116 and Phe40.

Interactions with the amino acids Arg42 (pi and van der Waals), Asp41 (van der Waals), Ile115 (hydrophobic) and Phe40 (van der Waals) are thought to be particularly characteristic of a compound with long residence times (the longest in the dataset studied) in complex with the InhA enzyme.

5.2 Identification of key interaction fingerprints

Standardization of data sets is an important step in analysis because it removes bias from the original variables. Standardized variables have similar variance. Analysis of the table of interactions of the studied complexes using PCA allowed isolation of the main factors for all InhA inhibitor systems studied. Figure 5.1 shows how the primary variables correlate with the principal components.

The individual and cumulative variance percentages of the analyzed data are shown on Figure 5.2. The first two components model the most variance in the data, best describing its structure, and the plot below show that these components describe only about 50% of the total variance of the data. Nevertheless, PCA analysis provides valuable information.

The position of the samples in the coordinate system defined by the principal components after PCA analysis and k-means data clustering is shown in Figure 5.3. The elbow method was used to determine the number of clusters. This method performs k-means clustering on the data set for a range of k values (in the thesis, a range of 2 to 8). It then calculates the average score for all clusters for each value of k. The default calculation is the distortion score, which is the sum of the squared lengths of each point from their associated midpoint. The point of the elbow at which the rate of decrease changes is then detected.

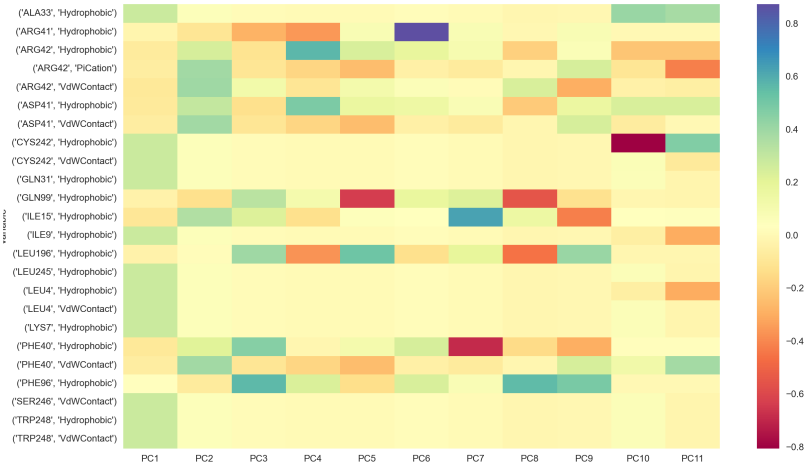


Figure 5.1: Features correlation with the principal components.

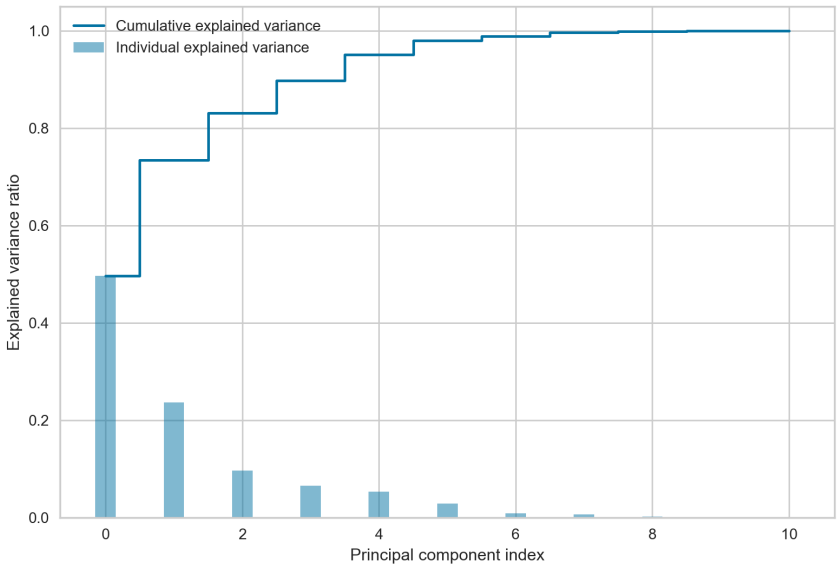


Figure 5.2: Individual and cumulative data variance percentages described by the 11 main factors for all studied complexes.

Figure 5.3a and 5.3b show groups of receptor-ligand contacts and Figure 5.3c and 5.3d - ligands. The expected result was separate clusters consisting of ligands with

similar residence times, or interaction fingerprints specific to the given residence time lengths.

As shown in Figure 5.3a, the optimal number of clusters for grouping the identified interaction fingerprints is 5. Table 5.3 provides a detailed summary of the contacts that form a given cluster, along with information about the receptor-ligand system in which they were observed.

The hydrophobic interaction with the amino acid Phe96 occurs in most of the complexes and therefore forms a separate group 2. This interaction does not affect the residence time of the ligand. The interactions of groups 0 and 3 occur only in systems with a ligand with a long residence time, and these are interactions that can be considered to be characteristic of compounds with a long residence time. Group 1 consists of interactions that, with the exception of two complexes with a relatively short residence time ligand: 4OIM and 2X23, have mostly been identified in complexes with long residence time compounds. Group 4 is characterized by contact identified only for compounds with short residence times, while additionally there is an interaction identified for both a compound with short and those with long residence times.

Figure 5.3c shows that the optimal number of clusters for ligand grouping by residence time is 4. Table 5.4 provides a detailed list of the ligands that make up a given cluster, along with their residence time information.

Group 0 includes compounds with a residence time of less than 100 minutes. Ligands with residence times greater than 100 minutes are grouped as 1, 2, and 3. A separate cluster was identified for the 5UGT complex containing the ligand with the longest residence time in the group.

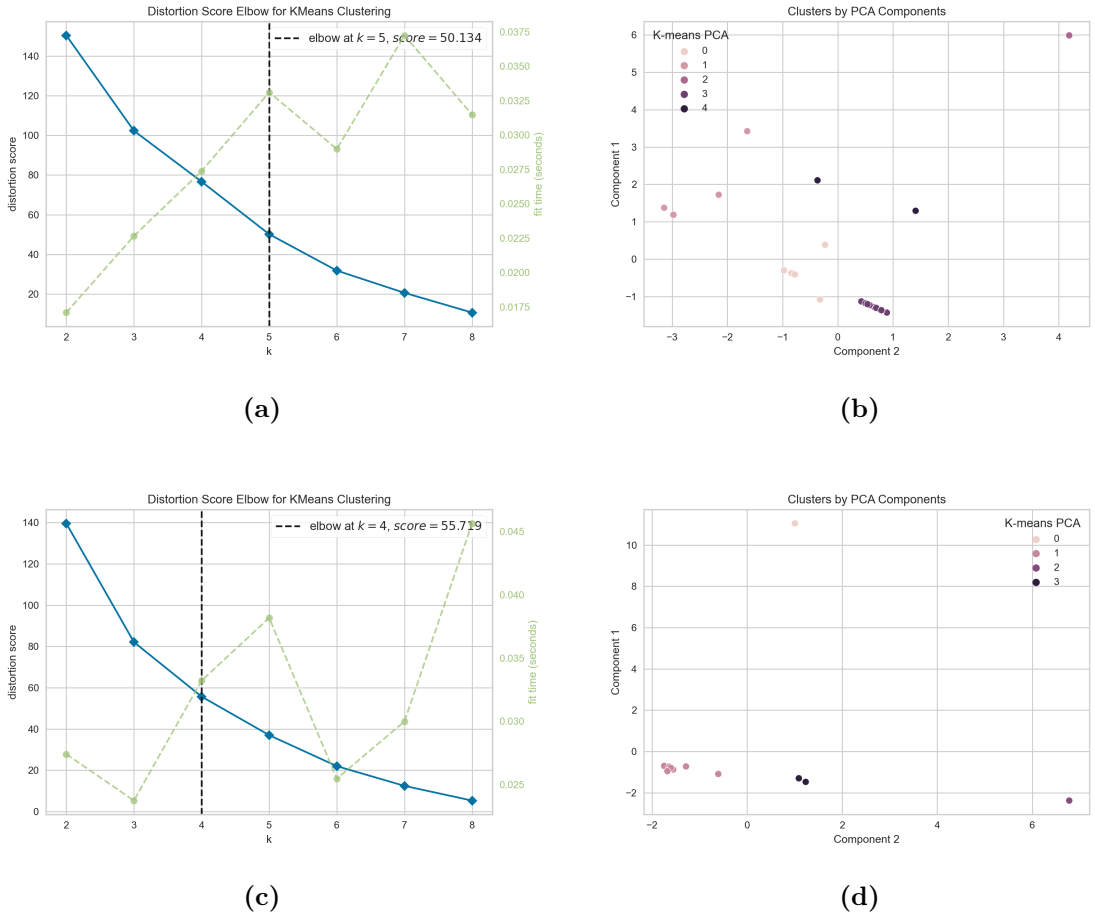


Figure 5.3: Sample projection onto the space defined by the first two principal factors.

Table 5.3: Quantitative summary of the size of each interaction fingerprint group.

Cluster ID	Cluster size	Interaction fingerprints that belong to cluster	receptor-ligand PDB IDs with residence times in which interaction occured
0	5	ARG41 Hydrophobic	4OYR (90 min)
		ARG42 PiCation	5UGT (220 min)
		ASP41 VdWContact	5UGT (220 min)
		PHE40 VdWContact	5UGT (220 min)
		LEU196 Hydrophobic	5UGU (194 min)
1	4	ILE15 Hydrophobic	4OIM (50 min), 5MTR (106 min), 5UGT (220 min), 5UGU (194 min)
		ARG42 Hydrophobic	2X23 (24 min), 5MTR (106 min), 5UGT (220 min)
		Arg42 VdWContact	5MTR (106 min), 5UGT (220 min), 5UGU (194 min)
2	1	ASP41 Hydrophobic	2X23 (24 min), 5MTR (106 min), 5UGT (220 min)
		Phe96 Hydrophobic	2X22 (24 min), 4OIM (50 min), 5COQ (30 min), 5MTQ (119 min), 5MTR (106 min), 5UGU (194 min)
3	12	SER246 VdWContact	5MTQ (119 min)
		LYS7 Hydrophobic	
		LEU4 VdWContact	
		Leu4 Hydrophobic	
		ALA33 Hydrophobic	
		ILE9 Hydrophobic	
		TRP248 Hydrophobic	
		GLN31 Hydrophobic	
		CYS242 VdWContact	
		Cys242 Hydrophobic	
		LEU245 Hydrophobic	
		TRP248 VdWContact	
4	2	GLN99 Hydrophobic	4OIM (50 min), 5COQ (30 min)
		PHE40 Hydrophobic	5COQ (30 min), 5MTR (106 min), 5UGT (220 min), 5UGU (194 min)

Table 5.4: Quantitative summary of the size of each ligand group.

Cluster ID	Cluster size	receptor-ligand PDB IDs with residence times that belong to cluster
0	7	2X22 (24 min), 2X23 (24 min), 4OIM (50min), 4OXY (27 min), 4OYR (90 min), 5COQ (30 min), 5MTP (94 min)
1	1	5MTQ (119 min)
2	1	5UGT (220 min)
3	2	5UGU (194 min), 5MTR (106 min)

A projection of the weights onto the space defined by the first two principal components is made in order to examine which factors are responsible for the differentiation of the samples.

Each variable (interaction) is represented in the form of a vector, which direction and length determine how much the variable influences each principal component. Thus, it may be concluded that the largest contribution to the formation of the first component (the values of the coefficients are the highest) is the van der Waals interaction with the amino acid Trp248. Hydrophobic contacts with Leu196, Arg41, Gln99 and Phe96 also contribute to the first principal component. The last three together with hydrophobic interactions with Phe40, Arg42, Asp41, Ile15 and van der Waals interactions with Arg42 and Phe40 form the second principal component.

From the weight projections, it can be concluded that hydrophobic interactions with Gln99 and Arg41 are highly correlated. The hydrophobic interactions with the amino acids Phe40, Arg42, Asp41 and Ile15 and the van der Waals interactions with Arg42 and Phe40 are also positively correlated with each other. This interaction set

is negatively correlated with hydrophobic interactions with Gln99, Arg41 and Phe96. The hydrophobic interaction between the ligand and Leu196 is negatively correlated with the van der Waals interaction with amino acid Trp248 which shows no correlation with other interactions. The interaction with Trp248 was only observed with the ligand in the 5MTQ complex (119 min residence time) and with Leu196 with the ligand in the 5UGU complex (194 min) (for the occurrence of each interaction see Table 5.4).

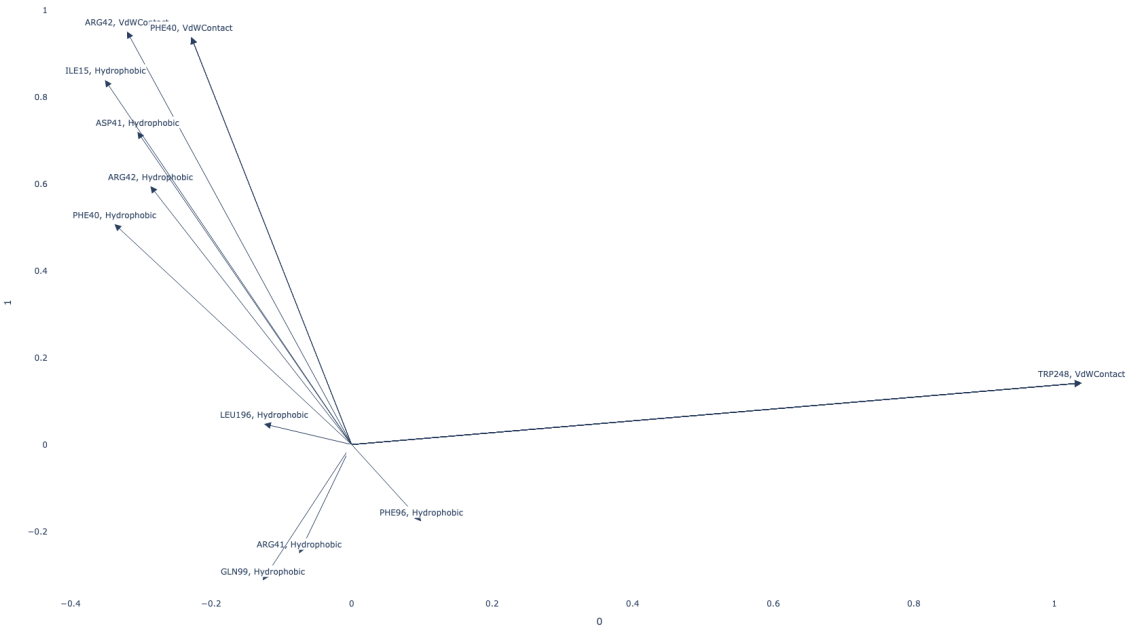


Figure 5.4: Projection of the weights on the space of the first two principal components.

5.3 Summary and Conclusion

To identify the molecular properties of the InhA protein and its inhibitors during the dissociation process characteristic of a given residence time, an approach that takes into account the transition states of the system and the interactions that take place between the protein molecule and the ligand molecule has been proposed. The

approach requires no prior knowledge of the binding mechanism and is based on PCA analysis and k-means clustering.

The identified groups form interaction fingerprints characteristic of ligands with a specific residence time, as well as groups that distinguish ligands based on the length of residence time in a molecular target.

An additional analysis of molecular descriptors, which allow a deeper insight into the molecular properties of the protein-ligand system, would probably be an interesting extension of the method. This will be an area for further development of the method.

Chapter 6

Discussion and Conclusions

The objective of this work was to apply and verify drug residence time prediction solutions to determine if they could be used regardless of the size of the molecules or the protein family, and to identify the molecular properties of the molecules involved in the dissociation process for InhA protein inhibitors. The most important questions that the research presented here has tried to answer are as follows: What are the major receptor-ligand interactions that distinguish ligands with long- and short-residence times? Is the τRAMD method universal and applicable to molecules of different size and structural similarity? What is the correlation between the relative residence time and experimental measurements?

Since no database of ligand binding kinetics was available, data were collected from literature and published in an online database (https://pdbrt.polsl.pl). The PDBrt database contains a total of 59 ligand entries for a total of 7 different families of proteins. The database will be continuously expanded with new available data.

τRAMD has been presented as a computationally efficient method for the prediction of relative residence time. Its versatility and reproducibility were investigated in this work. The analysis was performed on a set of 15 inhibitors of the HSP90 protein, 14 of which have been previously analysed (Kokh et al. 2018). 11 inhibitors of the InhA

protein and 3 ligands of the ENR, EGFR and HIV-1 proteins were also included. The total set included 29 ligands of various structural types. The resulting dissociation trajectories were then used as input for analysis of protein-ligand contacts, which are critical for the residence time. This analysis was carried out for a set of 11 inhibitors of the InhA protein.

Is the τRAMD method universal and applicable to molecules of different size and structural similarity? What is the correlation between the relative residence time and experimental measurements?

The analysis shows that τRAMD is reproducible. However, in some cases more sets (repetitions) of simulations should be performed to reduce the scatter of the data around the mean. The τRAMD results show a good or strong correlation with the experimentally determined residence time for compounds with a similar structure, i.e. differing by a small modification such as a shift of the functional group. However, for structurally diverse ligands, the method performed poorly, suggesting limited applicability of τRAMD in drug discovery.

What are the major receptor-ligand interactions that distinguish ligands with long- and short-residence times?

The identification of interactions critical for the residence time of InhA protein inhibitors was possible using the presented approach to analyze the molecular properties of receptor-ligand systems. A set of features was generated from the dissociation trajectories of the studied complexes. It was assumed that interactions occurring in at least 20% of a single trajectory describe the bound state, and that interactions occurring in less than 5% of all trajectories do not affect the residence time. Features meeting the above assumptions were removed from the feature set in the subsequent analysis. The proposed approach uses PCA and k-means cluster analysis, and does not require prior knowledge of binding mechanisms.

Analyzing the frequency of occurrence of a given interaction in the complexes allowed identifying key amino acids that are likely to have a significant impact on differentiating the tested compounds by residence time:

- the hydrophobic interaction with the amino acid Leu196, as well as the van der Waals with Phe40, Asp41, Arg42 and the π-cation with Arg42 promote longer residence times, as they were identified only in complexes with the ligands with the longest residence times in the studied data set (106, 194 and 220 min),
- for compounds with relatively short residence times (30 and 50 min), the hydrophobic interaction between the ligand and the amino acid Gln99 is characteristic.

Principal Component Analysis (PCA) identified the factors responsible for the differentiation of ligand residence times, and thus to verify the previously defined amino acids. These factors are as follows:

- van der Waals interaction with the amino acid Trp248,
- hydrophobic interactions with the amino acids Gln99 and Arg41,
- hydrophobic interactions with amino acids Phe40, Arg42, Asp41, Ile15 and van der Waals interactions with Arg42 and Phe40.

For compounds of similar structure, the research shows that the relative residence time correlates well with the experimentally determined time. The proposed algorithm can be used to identify key molecular features for the rate at which ligands dissociate from the binding site for structurally similar compounds.

Bibliography

Amangeldiuly, N., Karlov, D. & Fedorov, M. V. (2020), 'Baseline model for predicting protein-ligand unbinding kinetics through machine learning', *Journal of Chemical Information and Modeling* **60**, 5946–5956.

Ansari, N., Rizzi, V. & Parrinello, M. (2022), 'Water regulates the residence time of benzamidine in trypsin'.
URL: *http://arxiv.org/abs/2204.05572 http://dx.doi.org/10.1038/s41467-022-33104-3*

Badaoui, M., Buigues, P. J., Berta, D., Mandana, G. M., Gu, H., Földes, T., Dickson, C. J., Hornak, V., Kato, M., Molteni, C., Parsons, S. & Rosta, E. (2022), 'Combined free-energy calculation and machine learning methods for understanding ligand unbinding kinetics', *Journal of Chemical Theory and Computation* **18**, 2543–2555.

Bajusz, D., Rácz, A. & Héberger, K. (2015), 'Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?', *Journal of Cheminformatics* **7**.

Bashir, D., Montanez, G. D., Sehra, S., Segura, P. S. & Lauw, J. (2020), 'An information-theoretic perspective on overfitting and underfitting'.
URL: *https://arxiv.org/abs/2010.06076*

Bonaccorso, G. (2017), *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*, Packt Publishing.

Bouysset, C. & Fiorucci, S. (2021), 'Prolif: a library to encode molecular interactions as fingerprints', *Journal of Cheminformatics* **13**.

Bray, S., Tänzel, V. & Wolf, S. (2022), 'Ligand unbinding pathway and mechanism analysis assisted by machine learning and graph methods'.
URL: *http://arxiv.org/abs/2205.09894 http://dx.doi.org/10.1021/acs.jcim.2c00634*

Case, D., Aktulga, H., Belfon, K., Ben-Shalom, I., Berryman, J., Brozell, S., Cerutti, D., III, T. C., Cisneros, G., Cruzeiro, V., Darden, T., Duke, R., Giambasu, G., Gilson, M., Gohlke, H., Goetz, A., Harris, R., Izadi, S., Izmailov, S., Kasavajhala, K., Kaymak, M., King, E., Kovalenko, A., Kurtzman, T., Lee, T., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Machado, M., Man, V., Manathunga, M., Merz, K., Miao, Y., Mikhailovskii, O., Monard, G., Nguyen, H., O'Hearn, K., Onufriev, A., Pan, F., Pantano, S., Qi, R., Rahnamoun, A., Roe, D., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shajan, A., Shen, J., Simmerling, C., Skrynnikov, N., Smith, J., Swails, J., Walker, R., Wang, J., Wang, J., Wei, H., Wolf, R., Wu, X., Xiong, Y., Xue, Y., York, D., Zhao, S., & Kollman, P. (2022), 'Amber 22'.
URL: *https://ambermd.org/AmberTools.php*

Copeland, R. A. (2016), 'The drug-target residence time model: A 10-year retrospective', *Nature Reviews Drug Discovery* **15**, 87–95.
URL: *http://dx.doi.org/10.1038/nrd.2015.18*

Copeland, R. A., Pompliano, D. L. & Meek, T. D. (2006), 'Drug-target residence time and its implications for lead optimization', *Nature Reviews Drug Discovery* **5**, 730–739.

Dahl, G. & Akerud, T. (2013), 'Pharmacokinetics and the drug-target residence time concept'.

Dhakal, A., McKay, C., Tanner, J. J. & Cheng, J. (2022), 'Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions'.

Dowling, M. R. & Charlton, S. J. (2006), 'Quantifying the association and dissociation rates of unlabelled antagonists at the muscarinic m 3 receptor', *British Journal of Pharmacology* **148**, 927–937.

Flower, D. R. (1998), 'On the properties of bit string-based measures of chemical similarity', *Journal of Chemical Information and Computer Sciences* **38**(3), 379–386.
URL: *https://doi.org/10.1021/ci970437z*

Folmer, R. H. (2018), 'Drug target residence time: a misleading concept', *Drug Discovery Today* **23**, 12–16.
URL: *https://doi.org/10.1016/j.drudis.2017.07.016*

Gabdoulline, R. R. & Wade, R. C. (1999), 'On the protein-protein diffusional encounter complex'.

Gabdoulline, R. R. & Wade, R. C. (2022), 'Biomolecular diffusional association gabdoulline and wade 205'.

Ganotra, G. K. & Wade, R. C. (2018), 'Prediction of drug-target binding kinetics by comparative binding energy analysis', *ACS Medicinal Chemistry Letters* **9**, 1134–1139.

Gilski, M. (2014), 'Wysokorozdzielcza krystalografia makromolekuŁ high resolution crystallography of macromolecules'.

Gobbo, D., Piretti, V., Martino, R. M. C. D., Tripathi, S. K., Giabbai, B., Storici, P., Demitri, N., Girotto, S., Decherchi, S. & Cavalli, A. (2019), 'Investigating drug-

target residence time in kinases through enhanced sampling simulations', *Journal of Chemical Theory and Computation* **15**, 4646–4659.

Humphrey, W., Dalke, A. & Schulten, K. (1996), 'VMD – Visual Molecular Dynamics', *Journal of Molecular Graphics* **14**, 33–38.

Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. (2000), 'Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method', *Journal of Computational Chemistry* **21**, 132–146.

Jakalian, A., Jack, D. B. & Bayly, C. I. (2002), 'Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation', *Journal of Computational Chemistry* **23**, 1623–1641.

Jiménez, J., Škalič, M., Martínez-Rosell, G. & Fabritiis, G. D. (2018), 'Kdeep: Protein-ligand absolute binding affinity prediction via 3d-convolutional neural networks', *Journal of Chemical Information and Modeling* **58**, 287–296.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983), 'Comparison of simple potential functions for simulating liquid water', *The Journal of Chemical Physics* **79**, 926–935.

Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. (2017), 'druGAN: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico', *Molecular Pharmaceutics* **14**(9), 3098–3104. PMID: 28703000.
URL: *https://doi.org/10.1021/acs.molpharmaceut.7b00346*

Kokh, D. B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H. P., Dreyer, M. K., Frech, M., Lowinski, M., Vallee, F., Bianciotto, M., Rak, A. & Wade, R. C.

(2018), 'Estimation of drug-target residence times by τ-random acceleration molecular dynamics simulations', *Journal of Chemical Theory and Computation* **14**, 3859–3869.

Kokh, D. B., Kaufmann, T., Kister, B. & Wade, R. C. (2019), 'Machine learning analysis of τramd trajectories to decipher molecular determinants of drug-target residence times', *Frontiers in Molecular Biosciences* **6**.

Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D. & Xiong, H. (2021), Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity, Association for Computing Machinery, pp. 975–985.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. (2015), 'ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb', *Journal of Chemical Theory and Computation* **11**(8), 3696–3713. PMID: 26574453.
URL: *https://doi.org/10.1021/acs.jctc.5b00255*

Mardt, A., Pasquali, L., Wu, H. & Noé, F. (2018), 'Vampnets for deep learning of molecular kinetics', *Nature Communications* **9**.

Maximova, E., Postnikov, E. B., Lavrova, A. I., Farafonov, V. & Nerukh, D. (2021), 'Protein–ligand dissociation rate constant from all-atom simulation', *The Journal of Physical Chemistry Letters* **12**(43), 10631–10636. PMID: 34704768.
URL: *https://doi.org/10.1021/acs.jpclett.1c02952*

McDowell, S. E., Špačková, N., Šponer, J. & Walter, N. G. (2007), 'Molecular dynamics simulations of rna: An in silico single molecule approach'.

Mollica, L., Decherchi, S., Zia, S. R., Gaspari, R., Cavalli, A. & Rocchia, W. (2015),

'Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations', *Scientific Reports* **5**.

Mollica, L., Theret, I., Antoine, M., Perron-Sierra, F., Charton, Y., Fourquez, J. M., Wierzbicki, M., Boutin, J. A., Ferry, G., Decherchi, S., Bottegoni, G., Ducrot, P. & Cavalli, A. (2016), 'Molecular dynamics simulations and kinetic measurements to estimate and predict protein-ligand residence times', *Journal of Medicinal Chemistry* **59**, 7167–7176.

Mondal, J., Ahalawat, N., Pandit, S., Kay, L. E. & Vallurupalli, P. (2018), 'Atomic resolution mechanism of ligand binding to a solvent inaccessible cavity in t4 lysozyme', *PLoS Computational Biology* **14**.

Neckers, L., Mimnaugh, E. & Schulte, T. W. (1999), 'Hsp90 as an anti-cancer target', *Drug Resistance Updates* **2**(3), 165–172.
URL: *https://www.sciencedirect.com/science/article/pii/S1368764699900821*

Niu, Y., Li, S., Pan, D., Liu, H. & Yao, X. (2016), '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**, 5622–5629.

Ortiz, A. R., ttst M Teresa Pisabarro, Gago, J. F. & Wade, R. C. (1995), 'Prediction of drug binding affinities by comparative binding energy analysis'.

Paul, F., Wehmeyer, C., Abualrous, E. T., Wu, H., Crabtree, M. D., Schöneberg, J., Clarke, J., Freund, C., Weikl, T. R. & Noé, F. (2017), 'Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations', *Nature Communications* **8**.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. (2020), 'Scalable molecular dynamics with namd'.

URL: *www.ks.uiuc.edu.*

Prasad, M. S., Bhole, R. P., Khedekar, P. B. & Chikhale, R. V. (2021), 'Mycobacterium enoyl acyl carrier protein reductase (inha): A key target for antitubercular drug discovery', *Bioorganic Chemistry* **115**, 105242.

URL: *https://www.sciencedirect.com/science/article/pii/S0045206821006192*

Qu, S., Huang, S., Pan, X., Yang, L. & Mei, H. (2016), '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**, 2061–2068.

Ray, D., Stone, S. E. & Andricioaei, I. (2022), 'Markovian weighted ensemble milestoning (m-wem): Long-time kinetics from short trajectories', *Journal of Chemical Theory and Computation* **18**(1), 79–95. PMID: 34910499.

URL: *https://doi.org/10.1021/acs.jctc.1c00803*

Romanowska, J., Kokh, D. B., Fuller, J. C. & Wade, R. C. (2015), '1 computational approaches for studying drug binding kinetics'.

Schrödinger, L. & DeLano, W. (2020), 'Pymol'.

URL: *http://www.pymol.org/pymol*

Schuetz, D. A., Bernetti, M., Bertazzo, M., Musil, D., Eggenweiler, H. M., Recanatini, M., Masetti, M., Ecker, G. F. & Cavalli, A. (2019), 'Predicting residence time and drug unbinding pathway through scaled molecular dynamics', *Journal of Chemical Information and Modeling* **59**, 535–549.

Solit, D. B. & Chiosis, G. (2008), 'Development and application of hsp90 inhibitors', *Drug Discovery Today* **13**(1), 38–43.
URL: *https://www.sciencedirect.com/science/article/pii/S1359644607004217*

Su, M., Liu, H., Lin, H. & Wang, R. (2020), 'Machine-learning model for predicting the rate constant of protein-ligand dissociation', *Wuli Huaxue Xuebao/ Acta Physico - Chimica Sinica* **36**.

Swinney, D. C. (2004), 'Biochemical mechanisms of drug action: what does it take for success?'.
URL: *www.nature.com/reviews/drugdisc*

Teo, I., Mayne, C. G., Schulten, K. & Lelièvre, T. (2016), 'Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time', *Journal of Chemical Theory and Computation* **12**(6), 2983–2989. PMID: 27159059.
URL: *https://doi.org/10.1021/acs.jctc.6b00277*

Tiwary, P., Mondal, J. & Berne, B. J. (2017), 'How and when does an anticancer drug leave its binding site?', *Science Advances* **3**(5), e1700014.
URL: *https://www.science.org/doi/abs/10.1126/sciadv.1700014*

Vauquelin, G. (2016), 'Effects of target binding kinetics on in vivo drug efficacy: koff, konand rebinding', *British Journal of Pharmacology* pp. 2319–2334.

Wallach, I., Dzamba, M. & Heifets, A. (2015), 'Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery'.
URL: *http://arxiv.org/abs/1510.02855*

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. (2004), 'Develop-

ment and testing of a general amber force field', *Journal of Computational Chemistry* **25**, 1157–1174.

Wang, Y., Ribeiro, J. M. L. & Tiwary, P. (2020), 'Machine learning approaches for analyzing and enhancing molecular dynamics simulations', *Current Opinion in Structural Biology* **61**, 139–145.

Wolf, S., Amaral, M., Lowinski, M., Vallée, F., Musil, D., Güldenhaupt, J., Dreyer, M. K., Bomke, J., Frech, M., Schlitter, J. & Gerwert, K. (2019), 'Estimation of protein-ligand unbinding kinetics using non-equilibrium targeted molecular dynamics simulations', *Journal of Chemical Information and Modeling* **59**, 5135–5147.

Wu, H., Mardt, A., Pasquali, L. & Noe, F. (2018), 'Deep generative markov state models'.
URL: *http://arxiv.org/abs/1805.07601*

Wua, H., Paul, F., Wehmeyer, C. & Noéa, F. (2016), 'Multiensemble markov models of molecular thermodynamics and kinetics', *Proceedings of the National Academy of Sciences of the United States of America* **113**, E3221–E3230.

Yasuda, I., Endo, K., Yamamoto, E., Hirano, Y. & Yasuoka, K. (2022), 'Differences in ligand-induced protein dynamics extracted from an unsupervised deep learning approach correlate with protein–ligand binding affinities', *Communications Biology* **5**.

Zheng, L., Fan, J. & Mu, Y. (2019), 'Onionnet: A multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction', *ACS Omega* **4**, 15956–15965.

Ługowska, M. & Pacholczyk, M. (2021), 'Pdbrt: A free database of complexes with measured drug-target residence time', *F1000Research* **10**, 1236.

Appendix A

Abbreviations

3D Three dimensional

abMD Adiabatic bias molecular dynamics

AMS Adaptive multilevel splitting

cdf Poisson cumulative distribution function

CDK2 cyclin-dependent kinase 2

COM Centre of mass

COMBINE Comparative binding energy

DBSCAN Density-based spatial clustering of applications with noise

ecdf Empirical cumulative density function

EGFR Epidermal growth factor receptor

ENR Enoyl-acyl carrier protein reductase

ff14SB Amber force field ff14 Stony Brook

fs Femtosecond (10^{-15} s)

GAFF General Amber Force Field

HIV-1 PR HIV-1 protease

HSP]Heat shock protein

InhA Mycobacterium enoyl acyl carrier protein reductase

K_d Equilibrium dissociation constant

k_{off} Dissociation rate constant

k_{on} Association rate constant

KS Kolmogorov-Smirnov test

MAE Mean absolute error

MD Molecular dynamics

ML Machine learning

MSM Markov state models

NAMD Nanoscale Molecular Dynamics

ns Nanosecond (10^{-9} s)

PCA Principal component analysis

PDB Protein Data Bank

PDBrt Protein Data Bank residence time

PLS Partial least-squares

QSKR Quantitative structure-kinetic relationship

τ**RAMD** τRandom accerelation molecular dynamics

RMSE Root mean squared error

ROC Receiver operating characteristic

τ Residence time

Appendix B

Amino acids abbreviations

Ala Alanine

Arg Arginine

Asn Asparagine

Asp Aspartic acid

Cys Cysteine

Glu Glutamic acid

Gln Glutamine

Gly Glycine

His Histidine

Ile Isoleucine

Leu Leucine

Lys Lysine

Met Methionine

Phe Phenylalanine

Pro Proline

Ser Serine

Thr Threonine

Trp Tryptophan

Tyr Tyrosine

Val Valine