Knowledge Discovery and Prediction Modeling of Protein-Drug Binding Kinetic by Integrating Machine Learning, Normal Mode Analysis and Molecular Dynamics Simulation

See Hong Chiu

Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu/gc_etds

Part of the Biochemistry Commons

Recommended Citation

https://academicworks.cuny.edu/gc_etds/886
KNOWLEDGE DISCOVERY AND PREDICTION MODELING OF PROTEIN-DRUG BINDING KINETICS BY INTEGRATING MACHINE LEARNING, NORMAL MODE ANALYSIS AND MOLECULAR DYNAMICS SIMULATION

By

SEE HONG CHIU

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2015
This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfying of the dissertation requirement for the degree of Doctor of Philosophy.

Lei Xie

Date

Chair of Examining Committee

Robert Haralick

Date

Executive Officer

Yinghao Wu

Saad Mneimneh

Susan L. Epstein

Tom Kurtzman

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

KNOWLEDGE DISCOVERY AND PREDICTION MODELING OF PROTEIN-DRUG BINDING KINETICS BY INTEGRATING MACHINE LEARNING, NORMAL MODE ANALYSIS AND MOLECULAR DYNAMICS SIMULATION

By

See Hong Chiu

Adviser: Doctor Lei Xie

One of the unaddressed challenges in drug discovery is that drug potency measured by protein-ligand binding affinity, such as IC\textsubscript{50} and K\textsubscript{d} in vitro, is not correlated with drug activity in vivo. Computational modeling is playing an increasing role in designing efficient therapeutics. However, existing computational methods for the high-throughput study of protein-ligand interactions (PLI) mainly focus on the prediction of the binding affinity. This is the combined effect of association (k\textsubscript{on}) and dissociation (k\textsubscript{off}) rate constants. Few works have been produced to predict k\textsubscript{off} or its reciprocal, residence time, which is a key measuring function of drug efficacy in vivo. This study addresses the unmet need of the accurate and scalable prediction of k\textsubscript{on} and k\textsubscript{off} simultaneously.

The fundamental strategy of our method is to develop a machine learning model using PLI kinetic features computed by normal mode analysis (NMA). To test our method, HIV-1 protease complex was used as a model system. There are three major findings of this study. First, kinetic properties are more important than thermal dynamic characteristics in determining protein-ligand binding kinetics. We propose that coherent conformational dynamics coupling between protein and ligand were proven to be more significant than pairwise residue binding energy in the prediction of kinetic rate.
constants. Second, NMA is an efficient method to capture conformational dynamics features for the large scale modeling of protein-ligand binding. Third, multi-target classification as well as multi-target regression, is a potentially valuable tool for modeling PLI kinetics. With the rapid increase of PLI kinetics data, the further improvement of proposed computational methodology may provide a powerful tool for large-scale modeling of PLI kinetics, thereby accelerating drug discovery process.
Acknowledgements

I want to take this opportunity to acknowledge those who helped me complete this thesis.

I want to thank my mentor, Dr. Lei Xie, for introducing me to the scientific area of integrating machine learning, protein-ligand interaction mechanism, molecular dynamics simulations, and normal mode analysis. Thank you Dr. Lei Xie for your patience in reviewing my thesis and for the many stimulating scientific discussions.

I want to thank my committee members, Dr. Susan L. Epstein, Dr. Saad Mneimneh, Dr. Tom Kurtzman, and Dr. Yinghao Wu, for evaluating my thesis and for their suggestions.

I want to thank Dr. Candido Cabo, Dr. Paula Whitlock, and Dr. Theodore Brown, for their guidance at the early stage of my thesis.

I want to thank executive officer Dr. Robert M. Haralick, and assistant program officer Dilvania M. Rodriguez, for all the administrative support over the years.

I want to thank Dr. Pablo Chacon, and Dr. Jose Ramon Lopez-Blanco at Chaconlab for their suggestions on the operation of iMOD system.

Finally, I want to thank my wife and daughter for their endless support and for encouraging me to do my best.
# Table of Contents

**Table of Contents** vii

**List of Tables** xiii

**List of Figures** xvi

**1 Introduction** 1

**2 Materials and Methods** 6

2.1 Phase 1 - 3D Molecular Structure of Ligand-Bound HIV-1 Complex 8

2.1.1 Detection Limits of SPR Biosensor 9

2.1.2 Simplified Molecular Input Line Entry Specification (SMILES) 9

2.1.3 Frog 13

2.1.4 Electronic High Throughput Screening 13

2.2 Phase 2 – Identify Residues close to Ligand 15

2.3 Phase 3 – Principal Dataset Construction 16

2.3.1 Training dataset DS-PIE 18

2.3.2 Training Datasets DS-RMLR and DS-RMRR 19

2.4 Phase 4 - Machine Learning 22

2.4.1 Cross-Validation 23

2.4.2 Performance Measurement 23

2.4.2.1 Quality Measure of Classification 24

2.4.2.2 Quality Measure of Regression 25

2.4.2.3 Quality Measure of Multi-Target Prediction 26

2.5 Feature Evaluation 26
2.5.1 Two-tailed Welch’s t-test
2.5.2 Feature Selection
2.6 Receiver Operating Characteristic Curve
2.7 Clustering

3 Molecular Dynamics Simulations
3.1 Ergodic Hypothesis
3.2 Potential Energy Function
3.3 Average Dielectric Constants of Different Types of Residues
3.4 Integration Algorithms
3.5 NAnoscale Molecular Dynamics – NAMD
3.5.1 Velocity Verlet Integration Algorithm
3.5.2 Implicit Solvation - \( GB^{OBC} \) Model
3.5.3 Scalability of Decomposition Strategy
3.5.4 Load Balancing
3.5.5 Stages in a Typical NAMD MD Simulation
3.6 The Role of Molecular Dynamic Simulation in this Study
3.7 Clustering
3.7.1 N-DS-PIE Clustering

4 Normal Mode Analysis
4.1 Standard Normal Mode Analysis
4.2 Elastic Network Model
4.3 Gaussian Network Model
4.4 Anisotropic Network Model
4.5 Normal Mode Analysis in Internal Coordinate
4.6 iMod
4.7 Conformational Change Covered by NMA
4.8 Utilities and Limitations
  4.8.1 Harmonic Deformation
  4.8.2 Timescales of Harmonic Motions
  4.8.3 Usefulness of Coarse-grained NMA
4.9 The Role of Normal Mode Analysis in this Study
  4.9.1 iMOD
4.10 Clustering
  4.10.1 N-DS-RMLR
  4.10.2 N-DS-RMRR
  4.10.2.1 Clustering Based on the Ligand Identity
  4.10.2.2 Clustering Based on the Residue Identity

5 Machine Learning

  5.1 Probabilistic Prediction
  5.2 Multi-Target Learning
  5.3 Neural Networks and Multilayer Perceptron
  5.4 Decision Trees
  5.5 Random Forest
    5.5.1 Multi-Target Random Forest Classification Algorithm of Clus
    5.5.2 Single-Target Random Forest Regression Algorithm of Scikit-Learn
  5.6 Linear Regression Algorithm
5.6.1 Single-Target Linear Regression Algorithm 101
5.6.2 Multi-Target Lasso Linear Regression Algorithm 102
5.7 K-Nearest Neighbors Algorithm 102
5.8 Results of Machine Learning Computation 103
5.8.1 Classification Results 103
5.8.2 Feature Selection 106
5.8.3 Regression Results 108
5.8.3.1 Single-Target Elastic Net Linear Regression Algorithm 108
5.8.3.2 Single-Target Lasso Linear Regression Algorithm 109
5.8.3.3 Binary-Target Lasso Linear Regression Algorithm 109
5.8.3.4 Single-Target Random Forest Algorithm 109
5.8.4 K-Nearest Neighbors 114
5.9 Impact of Individual Training Dataset 115
5.9.1 Binary-Target Random Forest Classification Algorithm 116
5.9.2 Single-Target Elastic net Linear Regression Algorithm 117
5.9.3 Single-Target Lasso Linear Regression Algorithm 119
5.9.4 Single-Target K-Nearest Neighbors Instance-Based Learning 121
5.9.5 Single-target Random Forest Regression 122
5.9.6 Binary-Target Lasso Regression Algorithm 124
5.9.7 Multi-Target Lasso versus Single-Target Lasso 126
5.10 Feature Evaluation 127
5.10.1 Protease Inhibitor Resistance Mutation 127
5.10.2 Receiver Operating Characteristic Curve 129
5.10.3 Welch’s T-Test

6 Structural Determinants of Protein-Ligand Binding Kinetics

6.1 Elementary Aspects of Ligand Protein Binding Kinetics

6.2 Ligand-Protein Binding Models

6.2.1 Kinetic Rate Model of Induced Fit Mechanism

6.2.2 Kinetic Rate Model of Selected Fit Mechanism

6.2.3 Kinetic Rate Model of Three-Step Mechanism

6.3 Residence Time

6.4 Metadynamics

6.5 Analysis of the log_{10}k_{on} and log_{10}k_{off} Results

6.5.1 Impact of Sample Distribution on log_{10}k_{on} Prediction

6.5.2 Impact of the Excited State of Ligand Bound Complex on log_{10}k_{off} Prediction

6.5.3 Analysis

6.6 Results of Feature Selection

6.7 Molecular Determinants of Ligand Binding and Unbinding Kinetics

7 Summary, Conclusion and Future Work

7.1 Summary

7.1.1 Machine Learning Prediction

7.1.2 Feature Selection, Welch’s t-test and PIRM

7.2 Conclusion

7.3 Future Work

7.3.1 Metadynamics Simulation for the Kinetic Role of Water
List of Tables

2.1 Association and dissociation rate constants (k_{on}, k_{off}) 10
2.2 Structural classification of thirty-three ligand-bound HIV-1 complexes 11
2.3 eHiTS docking results 13
2.4 The 23 Surface Point Types as Defined in eHiTS 16
2.5 Results of splitting 27
3.1 Ligand and residue names and their corresponding identities 50
3.2 Data analysis of the inertia results of N-DS-PIE 51
3.3 Results of the N-DS-PIE clustering based on the ligand identity 51
4.1 Data analysis of the inertia results of N-DS-RMLR clustering based on the ligand identity 72
4.2 Results of the N-DS-RMLR clustering based on the ligand identity 72
4.3 Data analysis of the inertia results of N-DS-RMRR clustering based on the ligand identity 76
4.4 Results of the N-DS-RMRR clustering based on the ligand identity 80
4.5 Data analysis of the inertia results of N-DS-RMRR-R clustering based on the residue identity 80
4.6 Results of the N-DS-RMRR-R clustering based on the residue identity 84
4.7 Residues in the descending order of the average normalized NTAVs 85
5.1 Results of the discretization 100
5.2 Results of the LOO cross-validation experiments of the binary-target classifiers for log_{10}k_{off} prediction 104
5.3 Results of the LOO cross-validation experiments of the binary-target classifiers for log_{10}k_{on} prediction

5.4 Quality measure of binary-target classification prediction

5.5 Attribute selection

5.6 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target elastic net regressors

5.7 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target lasso regressors

5.8 %deviation and Pearson’s correlation coefficient of the predicted results given by the binary-target lasso regressors

5.9 Results of log_{10}k_{on}

5.10 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target random forest regressors

5.11 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target k-NN regressors

5.12 Training datasets used for ML models for the predictions of residues of PIRM log_{10}k_{on} and log_{10}k_{off}

5.13 Results of splitting

5.14 Number of occurrence of attributes with two-tailed p-value < 0.05

6.1 Distribution of log_{10}k_{on} predicted results

6.2 Distribution of log_{10}k_{off} prediction

6.3 Results of assignment for the residues in F-RML

6.4 Common residues shared by F-RMLR, F-RMRR, and F-PIE
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5 Residue molecular properties</td>
<td>164</td>
</tr>
<tr>
<td>7.1 Summary of classification prediction</td>
<td>167</td>
</tr>
<tr>
<td>7.2 Summary of regression prediction</td>
<td>168</td>
</tr>
<tr>
<td>7.3 Summaries of the results of feature selection</td>
<td>169</td>
</tr>
<tr>
<td>7.4 Probability of having p-value &lt; 0.05 in the Welch’s T-Test</td>
<td>171</td>
</tr>
</tbody>
</table>
# List of Figures

2.1 Methodological process of this study 7
2.2 2D molecular structure of B268 11
2.3 Structure of interaction between aspartyl protease and peptide substrate 11
2.4 Transformation from the 2D molecular structure of B268 into its SMILES string 12
2.5 Conversion of B268 SMILES string into its 3D molecular structure 12
2.6 SASA TCL script with probe radius of 1.4Å for the A045-1AJV complex 17
2.7 Average dielectric constants of different types of amino acids 20
2.8 NAMDEnergy graphical user interface 20
2.9 A confusion matrix for a binary classifier 24
3.1 Molecular dynamics simulation algorithm 31
3.2 Stages in a NAMD MD simulation 45
3.3 Flowchart indicating the role of files as used by NAMD and VMD 45
3.4 Configuration file of MD simulation for A045-bound HIV-1 protein complex 46
3.5 Molecular model of A045-bound HIV-1 complex 48
3.6 Equilibrium of A045-bound HIV-1 complex model 48
3.7 Inertia value versus k-value 49
3.8 A to F Results of the N-DS-PIE clustering based on the ligand identity 54
4.1 Hookean pairwise potential 57
4.2 Description of GNM 59
4.3 Motional patterns for the fourth mode of lysozyme (PDB code:3lzt) 61
4.4 Internal coordinate system of protein 62
4.5 Cumulative overlap

4.6 Changes of energy landscape upon ligand binding

4.7 Superposition of ligand free HIV-1 protease (red) and DMP (green) bound HIV-1 protease (blue)

4.8 Superposition of the 44 residues of the aligned 3IXO (red) and the DMP bound 1QBS (blue) of 1st normal mode

4.9 Relative displacements of the 44 DMP (red) –residue (green/blue) pairs in the DMP bound HIV-1 complex (1st normal mode)

4.10 A to F Results of the N-DS-RMLR clustering based on the ligand identity

4.11 A to F Results of the N-DS-RMRR clustering based on the ligand identity

4.12 A to F Results of the N-DS-RMRR-R clustering based on the residue identity

4.13 Average normalized NTAVs (Å^2) of the 39 ligand bound complexes at the ground state (E2L)

5.1 Multilayer Perceptron Structure

5.2 Discretization

5.3 Iteration number versus accuracy

5.4 Results of log_{10}k_{on} prediction given by random forest regressors

5.5 Results of log_{10}k_{off} prediction given by random forest regressors

5.6 Results of binary-target random forest classification

5.7 Results of log_{10}k_{on} prediction of single-target elastic net regression

5.8 Results of log_{10}k_{off} prediction of single-target elastic net regression

5.9 Results of log_{10}k_{on} prediction of single-target lasso linear regression
5.10 Results of $\log_{10} k_{\text{off}}$ prediction of single-target lasso linear regression

5.11 Results of $\log_{10} k_{\text{on}}$ prediction of single-target

K-nearest neighbors instance-based learning

5.12 Results of $\log_{10} k_{\text{off}}$ prediction of single-target

K-nearest neighbors instance-based learning

5.13 Results of $\log_{10} k_{\text{on}}$ prediction of single-target random forest regression

5.14 Results of $\log_{10} k_{\text{off}}$ prediction of single-target random forest regression

5.15 Results of $\log_{10} k_{\text{on}}$ prediction of binary-target lasso linear regression

5.16 Results of $\log_{10} k_{\text{off}}$ prediction of binary-target lasso linear regression

5.17 MM-%deviation. Binary-target lasso linear regression versus

Single-target lasso linear regression

5.18 Active site PIRM and non-active site PIRM in Ritonavir-bound

HIV-1 complex (PDB code 1HXW)

5.19 ROC curves of F-PIE, F-RMLR, and F-RMRR

6.1 Structure of HIV-1 protease

6.2 Catalytic mechanism for aspartic protease proposed by Suguna et al

6.3 (A) log (residence time, $\tau$) versus functional efficacy

(B) log(binding affinity, $K_d$) versus functional efficacy

6.4 Free energy profile of a drug (D) binding to a receptor (R)

to form a drug-receptor complex (DR)

6.5 Four-state model of protein-ligand binding

6.6 Kinetic rate model of the induced fit mechanism proposed by Weikl and von Deuster
6.7 Kinetic rate model of the selected fit mechanism proposed by Weikl and von Deuster

6.8 Three-step process of diffusion, free conformer selection, and refolding

6.9 Superposition of ligand-bound and ligand-free HIV-1 protease complexes

6.10 Reaction coordinate diagram of the escape trajectory for ligand dissociation following a retrograde induced-fit mechanism

6.11 Free-energy diagram of the escape trajectory for ligand dissociation following a retrograde induced fit mechanism

6.12 Example of metadynamics simulation in a one-dimensional model potential

6.13 %deviation of log10k_on prediction versus number of records in bins

6.14 %deviation of log10k_off prediction versus log10k_off

6.15 Residues of F-RMLR. Beads represent the 15 residues of chain A in F-RMLR

6.16 L76 and P81 in the HIV-1 active site

6.17 Ligand molecular weight versus log10k_on/log10k_off

6.18 Number of rotatable bond of ligand versus log10k_on/log10k_off

6.19 Ten active site residues (K45, I47, G48, G49, I50, G52, F53, I54, T80, P81) conduct significant conformational displacements upon ligand binding

6.20 Ligand-receptor interaction

6.21 Structure of cyclic urea inhibitor

7.1 Mechanism for the hydration of the keto amide inhibitor by HIV-1 protease
Chapter 1

Introduction

Discovering a new drug now costs more than two billion dollars and takes more than ten years. Only about one third of drugs in phase III clinical trials reach the market. The failure rate of a new drug is staggering. Phenotypic screening and target-based screening are the two essential approaches in drug discovering. Phenotypic screening is a strategy to identify molecules that cause a desirable change in phenotype in cellular or animal disease models without prior understanding of the molecular mechanism of action (MMOA). Only after the molecules have been discovered, an effort is made to determine the biological target of the molecules. A drawback of phenotypic screening is the challenge of optimizing the molecular characteristics of candidate drugs without the design parameters provided by prior MMOA knowledge [1]. Target-based screening is hypothesis-driven. Only drug candidate binding to a pre-designated protein target is assessed. A target-centric strategy provides a specific biological hypothesis to be tested and a beginning point for drug identification to do this with. A drawback of target-based screening is that the answer to the hypothesis may not be relevant to the disease pathogenesis or provide a sufficient therapeutic ratio [2]. These drawbacks account for the current high attrition rate in drug discovery.

One of the unaddressed fundamental challenges in drug discovery is that drug potency measured in vitro is not correlated with drug activity in vivo as in the human body. In pharmacological research, half maximal inhibitory concentration (IC\textsubscript{50}) and dissociation constant (K\textsubscript{d}) have been used as the measures of drug efficacy for years. As molecules in the human body are in a non-equilibrium condition, the efficacy of a drug
depends, not only on how strong it interacts with the protein, but also how long it resides in the protein. In 2011, Copeland revealed that drug efficacy in vivo is not defined by equilibrium conditions measured in vitro, but rather depends on the residence time \((\tau = 1/k_{off})\) of the receptor-ligand complex in vivo [3]. For example, geldenamycin has low affinity for Heat shock protein (Hsp90) in vitro with IC\(_{50}\) ~ 1 \(\mu\)M, in comparison to its nanomolar effects in vivo. In 2013, Pan et al. reported that residence time is highly correlated with functional efficacy of a series of agonists of the A\(_{2A}\) adenosine receptor \((r^2 = 0.95)\), but there is little correlation with binding affinity \((r^2 = 0.15)\) [4].

Moreover, since IC\(_{50}\) and K\(_d\) depend on the measurement of the combined effect of k\(_{on}\) and k\(_{off}\), they are actually insufficient to explain the impact of conformational dynamics on PLI, as the same value of K\(_d\) can come from infinite number of combinations of k\(_{on}\) and k\(_{off}\). Additionally, since K\(_d\) is dependent on the free energy difference between the bound and unbound states but is independent on the transition state of PLI, it is inadequate to elucidate the binding kinetics of PLI [4,5].

Computational modeling plays an increasing role in designing efficient therapeutics. Most of the drugs act by binding to receptors, but the binding process has proven difficult to elucidate. MD simulations have been reported to be capable to capture the binding process, from beginning to end, in full atomic detail. Unfortunately, ligand binding and unbinding events are often on a long-time scale ranging from milliseconds to days, far beyond the current capability of MD simulations. For this reason, metadynamics has been developed not only to improve sampling in molecular dynamics simulations of a system where ergodicity is hindered by the form of the system’s energy landscape, but also adopted as a powerful technique for reconstructing the free-energy surface as a
function of few selected degrees of freedom. In 2011, Buch et al. presented a kinetic model for the binding process of serine protease β-trypsin inhibitor benzamidine obtained from MD simulations of free ligand binding. In addition to the kinetic pathway of the binding process, the binding free energy and the kinetic constants \(k_{on}\) and \(k_{off}\) of the process were also reported [6]. In 2004, Gervasio et al. applied a metadynamics method successfully to the docking of ligands on flexible receptors in water solution. The method is able not only to find the docked geometry and to predict the binding affinity \(\Delta G_{\text{binding}}\) but also to explore the entire docking process from the solution to the docking cavity, including barriers and intermediate minima. Four docking cases were examined in the study, including β-trypsin/benzamidine, β-trypsin/chlorobenzamidine, immunoglobulin McPC-603/phosphocholine, and cyclin-dependent kinase2/staurosporine [7].

Several predictive models for kinetic constants of protein-protein interaction (PPI) have been built. However, existing models all use static characteristics of PPI to predict the binding affinity. Researchers including Bai et al. in 2010 [8], and Moal and Bates in 2012 [9] have built predictive models for kinetic constants of PPI. But the molecular attributes in their models only covered static characteristics such as the percentage of residues in α-helix, the buried surface area of protein, and the interfacial electrostatic interaction energy between interfacial residues in the model built by Bai et al., and the proportion of charged residues at the interface, the hydrogen bonding potential, and the proportion of polar atoms at the interface in the model built by Moal and Bates. What lacks in their models are dynamics features of protein structure and of PPI. In addition, existing methods predict \(k_{on}\) and \(k_{off}\) independently. As a matter of fact, they are dependent in nature.
To tackle the above problems, we conducted a study to use machine learning to build models to predict $k_{on}$ and $k_{off}$ with kinetic features as well as thermal dynamic features in the training datasets. In this study, ligand-bound HIV-1 protease was used as an example to build machine learning models with three principal training datasets including DataSet-Pairwise Interaction Energy (DS-PIE), DataSet-Relative Movement Ligand-Residue (DS-RMLR), and DataSet-Relative Movement Residue-Residue (DS-RMRR) for kinetic rate constants: $k_{on}$ and $k_{off}$. NMA was used to build two of the training datasets, DS-RMLR and DS-RMRR, which cover the kinetic properties of the ligand-bound HIV-1 complex at ground state, where DS-RMLR concerns the relative movement of a ligand-residue pair and DS-RMRR pertains to the relative movement of a residue upon ligand binding. MD simulations were also used to build the training dataset, DS-PIE, to cover the pairwise interaction energy of a ligand-residue pair, which represents the non-covalent interactions of the pair. Specifically, average dielectric constants of different types of residues reported by Li et al. in 2013 [10], were adopted for the pairwise interaction energy calculation.

Four single-target algorithms including k-nearest-neighbors instance-based learning algorithm, elastic net and lasso linear regression algorithms, and random forest regression algorithm were used to build models to predict $\log_{10}k_{on}$ and $\log_{10}k_{off}$ separately. Additionally, one multi-target random forest classification algorithm and one multi-target lasso regression algorithm were used, for the first time, to build models to predict $\log_{10}k_{on}$ and $\log_{10}k_{off}$ simultaneously, as $k_{on}$ and $k_{off}$ are dependent.

The results of this study reveal that (1) kinetic properties are more important than thermal dynamic characteristics in determining protein-ligand binding kinetics; (2) NMA
is an efficient method to capture conformational dynamics features for the large scale modeling of protein-ligand binding; (3) multi-target algorithm is potentially valuable tool for modeling PLI because single-target algorithm and multi-target algorithm have performed equivalently on the accuracy of prediction.

The Welch’s t-test conducted in this study identified eleven residues which could possess dual characteristics in PLI depending on the values of the kinetic rate constants. In addition to these eleven residues, seven more residues were selected in the feature selection process. This process was conducted to identify residues significant to PLI with frequency of occurrence greater than 25% in the Leave-One-Out (LOO) cross-validation experiments of the three random forest classifiers trained with DS-PIE, DS-RMLR, and DS-RMRR with iteration number = 500. Among the eighteen residues, nine are residues of protease inhibitor resistance mutation (PIRM).
Chapter 2
Materials and Methods

In this study, thirty-nine ligand-bound HIV-1 protease complexes were used as training samples to build three principal datasets including DS-RMLR, DS-RMRR, and DS-PIE. Four ML single-target algorithms and two ML multi-target algorithms were used for the prediction of kinetic rate constants ($k_{\text{on}}$ and $k_{\text{off}}$). The four single-target algorithms include k-nearest-neighbors instance-based learning algorithm, elastic net and lasso linear regression algorithms, and random forest regression algorithm. The two multi-target algorithms include multi-target random forest classification algorithm and multi-target lasso regression algorithm.

The binding between a ligand molecule and HIV-1 protease is conducted by induced fit mechanism; thus, the ligand dissociation ($k_{\text{off}}, \tau = 1/k_{\text{off}}$) is determined by the retrograde induced fit mechanism [3,11]. In order to obtain accurate prediction on the kinetic rate constants, the feature attributes in the principal datasets cover the thermal dynamic characteristics of the ligand-residue pair, and the kinetic characteristics of the induced fit mechanism as well as its retrogradation. The thermal dynamic characteristics are described by the feature attributes in the dataset DS-PIE, which was constructed using molecular dynamics simulations. The kinetic characteristics are expressed by the feature attributes in the datasets DS-RMRR and DS-RMLR, which were built using normal mode analysis technique. Specifically, DS-PIE covers the pairwise interaction energy (PIE) of a ligand-residue pair; DS-RMRR covers the relative movement of a residue upon ligand binding; and DS-RMLR covers the relative movement of the ligand-residue pair.
Generating 3D ligand molecular structure:
2D Ligand Molecular Structure
SMILES
1D SMILES String of Ligand
Frog
3D Ligand Structure

Docking using eHiTS

Generating 3D structure of ligand-HIV-1 protease

Solvent Accessible Surface Area
with probe radius of 1.4 and 2.1 Å

Searching for residues close to ligand

Building three datasets containing kinetic and thermal dynamic data of ligand-HIV-1 interaction

Molecular Dynamics Simulations
Normal Mode Analysis
Normal Mode Analysis

machine learning model induction

Cross Validation

Figure 2.1 Methodological process of this study
Figure 2.1 depicts the methodological process of this study, which includes four phases: Phase 1 concerns the structure construction of 3D ligand-bound HIV-1 protease complex. Phase 2 addresses the identification of residues which are close to the ligand. Phase 3 targets the construction of the three principal datasets. Phase 4 is machine learning computation. In this study, ML was the tool used to predict the kinetic rate constants of PLI. Then, the results were analyzed for useful information extraction. It is not only to identify the major findings, but also to improve the predicting accuracy of the model by adding the missing characteristics of PLI in the model to eliminate the drawbacks of the model.

### 2.1 Phase 1 - 3D Structure of Ligand-Bound HIV-1 Complex

In 2002, Markgren et al. reported the kinetic rate constants ($k_{on}$ and $k_{off}$) of thirty-nine ligand-bound HIV-1 complexes [12] using the technique of Surface Plasmon Resonance Based (SPR) Biosensor (Table 2.1) [13]. Thirty-three of them were classified into five structural categories (Table 2.2) in reference to the 2D molecular structure of B206 as shown in Figure 2.2. The five categories include non-B268 analogues, P1/P1’ analogues of B268, P2/P2’ analogues of B268, cyclic ureas and cyclic sulfamides. Using standard nomenclature, $P_1...P_n$, $P_1’...P_n’$ is used to designate amino acid residues of peptide substrates in the enzyme-substrate interactions (Figure 2.3). In this study, ten 3D molecular structures of the thirty-nine complexes were collected from RCBS Protein Data (PDB) Bank [14]. They include DMP-1QBS, AMP-3EKV, B435-1D4H, B369-1EBY, B409-1EC1, B388-1EBZ, B425-1D4I, Nelfinavir-3EKX, Ritonavir-1HXW, and U75875-1HIV. The remaining twenty-nine complex structures were obtained from a three-step building process. First, a 2D molecular ligand structure was transformed into...
its SMILES string [15]. Figure 2.4 depicts the transformation from the 2D molecular structure of B268 into its SMILES string. Second, the ligand SMILES string was converted into a 3D molecular ligand structure using Frog [16] as shown in Figure 2.5. Third, Electronic High Throughput Screening (eHiTS) program [17-20] was used to dock a target ligand into the active site of a wild type HIV-1 protein. The receptor was chosen from one of the five ligand-bound HIV-1 complexes (1QBS, 1EBW, 1AJV, 1EC2, and 1D4H) with the co-crystallized ligand structure similar to the target ligand structure. Table 2.3 shows the results of the docking process.

2.1.1 Detection Limits of SPR Biosensor

Due to the baseline stability of SPR biosensor and the detection limit exerted by the diffusion rate of an ligand to its binding partner which is immobilized on the biosensor surface, SPR biosensor is only capable of measuring association (kon) and dissociation (koff) kinetic rate constants in the range of $10^2$ to $10^8\ M^{-1}\ s^{-1}$ and 1 to $10^{-6}\ s^{-1}$, respectively [21]. As shown in Table 2.1, the kon and koff values of DMP, B376, and A008 are beyond the upper detection limit and the kon values of B277 and A016 are proximate to the lower detection limit.

2.1.2 Simplified Molecular Input Line Entry Specification (SMILES)

SMILES is a language with a few grammar rules for specifying a chemical structure on a single line (1D structure). One important property of SMILES is that it is quite compact compared to most other methods of representing structure. There are five generic SMILES encoding rules that correspond to the specification of atoms, bonds, branches, ring closures, and disconnections. Additionally, SMILES isomer specification rules allow chirality to be completely specified for any known structure.
<table>
<thead>
<tr>
<th>Ligand</th>
<th>$k_{\text{off}}$</th>
<th>$k_{\text{on}}$</th>
<th>$\log_{10}(k_{\text{off}})$</th>
<th>$\log_{10}(k_{\text{on}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>U75875</td>
<td>0.00544</td>
<td>6760000</td>
<td>-2.2644</td>
<td>6.8299</td>
</tr>
<tr>
<td>Saquinavir</td>
<td>0.00023</td>
<td>817000</td>
<td>-3.644</td>
<td>5.9122</td>
</tr>
<tr>
<td>Ritonavir</td>
<td>0.00216</td>
<td>3920000</td>
<td>-2.6655</td>
<td>6.5933</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>0.00067</td>
<td>663000</td>
<td>-3.1752</td>
<td>5.8215</td>
</tr>
<tr>
<td>Indinavir</td>
<td>0.00158</td>
<td>1530000</td>
<td>-2.8013</td>
<td>6.1847</td>
</tr>
<tr>
<td>DMP</td>
<td>83.3</td>
<td>252000000000</td>
<td>1.9206</td>
<td>10.4014</td>
</tr>
<tr>
<td>BEA409</td>
<td>0.00043</td>
<td>348000</td>
<td>-3.3645</td>
<td>5.5416</td>
</tr>
<tr>
<td>B440</td>
<td>0.0003</td>
<td>477000</td>
<td>-3.5186</td>
<td>5.6785</td>
</tr>
<tr>
<td>B439</td>
<td>0.00163</td>
<td>81100</td>
<td>-2.7878</td>
<td>4.909</td>
</tr>
<tr>
<td>B435</td>
<td>0.00653</td>
<td>101000</td>
<td>-2.1851</td>
<td>5.0043</td>
</tr>
<tr>
<td>B429</td>
<td>0.00037</td>
<td>323000</td>
<td>-3.4283</td>
<td>5.5092</td>
</tr>
<tr>
<td>B425</td>
<td>0.234</td>
<td>666000</td>
<td>-0.6308</td>
<td>5.8235</td>
</tr>
<tr>
<td>B412</td>
<td>0.00082</td>
<td>181000</td>
<td>-3.0878</td>
<td>5.2577</td>
</tr>
<tr>
<td>B408</td>
<td>0.00169</td>
<td>889000</td>
<td>-2.7721</td>
<td>5.9489</td>
</tr>
<tr>
<td>B388</td>
<td>0.0227</td>
<td>5970000</td>
<td>-1.644</td>
<td>6.776</td>
</tr>
<tr>
<td>B376</td>
<td>13.7</td>
<td>20500000000</td>
<td>1.1367</td>
<td>8.3118</td>
</tr>
<tr>
<td>B369</td>
<td>0.0133</td>
<td>6390000</td>
<td>-1.8761</td>
<td>6.8055</td>
</tr>
<tr>
<td>B365</td>
<td>0.0309</td>
<td>304000</td>
<td>-1.51</td>
<td>5.4829</td>
</tr>
<tr>
<td>B355</td>
<td>0.373</td>
<td>1080000</td>
<td>-0.4283</td>
<td>6.0334</td>
</tr>
<tr>
<td>B347</td>
<td>0.027</td>
<td>9200</td>
<td>-1.5686</td>
<td>3.9638</td>
</tr>
<tr>
<td>B322</td>
<td>0.0677</td>
<td>1850000</td>
<td>-1.1694</td>
<td>6.2672</td>
</tr>
<tr>
<td>B295</td>
<td>0.436</td>
<td>902000</td>
<td>-0.3605</td>
<td>5.9552</td>
</tr>
<tr>
<td>B277</td>
<td>0.00485</td>
<td>134</td>
<td>-2.3143</td>
<td>2.1271</td>
</tr>
<tr>
<td>B268</td>
<td>0.00367</td>
<td>355000</td>
<td>-2.4353</td>
<td>5.5502</td>
</tr>
<tr>
<td>B249</td>
<td>0.273</td>
<td>41000</td>
<td>-0.5638</td>
<td>4.6128</td>
</tr>
<tr>
<td>AMP</td>
<td>0.00488</td>
<td>4430000</td>
<td>-2.3116</td>
<td>6.6464</td>
</tr>
<tr>
<td>A047</td>
<td>0.0697</td>
<td>188000</td>
<td>-1.1568</td>
<td>5.2742</td>
</tr>
<tr>
<td>A045</td>
<td>0.263</td>
<td>499000</td>
<td>-0.58</td>
<td>5.6981</td>
</tr>
<tr>
<td>A038</td>
<td>0.00049</td>
<td>29300</td>
<td>-3.3125</td>
<td>4.4669</td>
</tr>
<tr>
<td>A037</td>
<td>0.00037</td>
<td>204000</td>
<td>-3.4377</td>
<td>5.3096</td>
</tr>
<tr>
<td>A030</td>
<td>0.042</td>
<td>512000</td>
<td>-1.3768</td>
<td>5.7093</td>
</tr>
<tr>
<td>A024</td>
<td>0.0685</td>
<td>221000</td>
<td>-1.1643</td>
<td>5.3444</td>
</tr>
<tr>
<td>A023</td>
<td>0.139</td>
<td>200000</td>
<td>-0.857</td>
<td>5.301</td>
</tr>
<tr>
<td>A021</td>
<td>0.0273</td>
<td>687000</td>
<td>-1.5638</td>
<td>5.837</td>
</tr>
<tr>
<td>A018</td>
<td>0.474</td>
<td>348000</td>
<td>-0.3242</td>
<td>5.5416</td>
</tr>
<tr>
<td>A017</td>
<td>0.179</td>
<td>436000</td>
<td>-0.7471</td>
<td>4.6395</td>
</tr>
<tr>
<td>A016</td>
<td>0.0605</td>
<td>172</td>
<td>-1.2182</td>
<td>2.2355</td>
</tr>
<tr>
<td>A015</td>
<td>0.938</td>
<td>109000</td>
<td>-0.0278</td>
<td>5.0374</td>
</tr>
<tr>
<td>A008</td>
<td>43.8</td>
<td>70600000000</td>
<td>1.6415</td>
<td>9.8488</td>
</tr>
</tbody>
</table>

Table 2.1 Association and dissociation rate constants ($k_{\text{on}}, k_{\text{off}}$) of the interactions between thirty-nine inhibitors and HIV-1 protease. Surface Plasmon Resonance Based Biosensor was used for the measurement of the constants. The $k_{\text{on}}$ and $k_{\text{off}}$ values in green are beyond the upper detection limit and the $k_{\text{off}}$ values in red are proximate to the lower detection limit.
Table 2.2 Structural classification of thirty-three ligand-bound HIV-1 complexes in reference to the structure of B268.

<table>
<thead>
<tr>
<th>non-B268 analogues</th>
<th>P1/P1’ analogues of B268</th>
<th>P2/P2’ and central hydroxy analogues</th>
<th>cyclic urea compound</th>
<th>cyclic sulfamide compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>B295</td>
<td>B277</td>
<td>A017</td>
<td>A008</td>
<td>A021</td>
</tr>
<tr>
<td>B355</td>
<td>B268</td>
<td>A016</td>
<td>DMP323</td>
<td>A024</td>
</tr>
<tr>
<td>B408</td>
<td></td>
<td>A015</td>
<td></td>
<td>A047</td>
</tr>
<tr>
<td>B409</td>
<td></td>
<td>B376</td>
<td></td>
<td>A045</td>
</tr>
<tr>
<td>B440</td>
<td></td>
<td>A018</td>
<td></td>
<td>A030</td>
</tr>
<tr>
<td>B429</td>
<td></td>
<td>B322</td>
<td></td>
<td>A023</td>
</tr>
<tr>
<td>B412</td>
<td></td>
<td>B365</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A037</td>
<td></td>
<td>B347</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B439</td>
<td></td>
<td>B388</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A038</td>
<td></td>
<td>B369</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B425</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B435</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>B249</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.2 2D molecular structure of B268. R1 and R2 are benzene rings.

Figure 2.3 Structure of interaction between aspartyl protease and peptide substrate.
Figure 2.4 Transformation from the 2D molecular structure of B268 into its SMILES string

Figure 2.5 Conversion of B268 SMILES string into its 3D molecular structure. Color scheme: Turquoise, blue and red represent carbon, nitrogen and oxygen atoms respectively.
Table 2.3 eHiTS docking results. Target ligands and the HIV-1 protease of the complex in the same column were adopted in the eHiTS docking process to generate the corresponding ligand-bound HIV-1 complex structures.

<table>
<thead>
<tr>
<th>Complex</th>
<th>1QBS</th>
<th>1EBW</th>
<th>1AJV</th>
<th>1EC2</th>
<th>1D4H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Ligand</td>
<td>A008</td>
<td>A015</td>
<td>A021</td>
<td>A037</td>
<td>B295</td>
</tr>
<tr>
<td></td>
<td>A016</td>
<td>A023</td>
<td>A038</td>
<td>B355</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A017</td>
<td>A024</td>
<td>B412</td>
<td>Indinavir</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A018</td>
<td>A030</td>
<td>B429</td>
<td>Saquinavir</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B249</td>
<td>A045</td>
<td>B439</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B268</td>
<td>A047</td>
<td>B440</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B277</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B322</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B347</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B365</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B376</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B408</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.1.3 Frog

Frog is a free online service aimed at generating 3D conformations for drug-like compounds starting from their 1D SMILES format or 2D SDF (structure data file). The conformations are approached by Monte-Carlo steps, and scored using van der Waals and simple Coulomb interactions based on the Merck Molecular Force Field.

2.1.4 Electronic High Throughput Screening (eHiTS)

eHiTS is a systemic docking algorithm that takes a five steps divide and conquer approach to the docking problem. Step 1: dividing the ligands into rigid fragments and connecting flexible chains. Step 2: docking each fragment independently in the active site. Step 3: reconstructing the original ligand by finding all compatible fragment sets. Step 4: optimizing the ligand structure within the active site. Step 5: ranking conformational poses based on scoring function values.
The scoring functions employed in protein-ligand docking can be divided into three major classes: knowledge-based, empirical, and force field-based scoring functions.

Knowledge-based scoring function concentrates on following the rules and general principles statistically derived that aim to reproduce experimentally determined structures. It models interactions in protein-ligand complexes using interaction surface points (ISPs). There are 23 types of ISPs as defined in eHiTS. They are allocated on the surface of the ligand and the binding pocket. (Table 2.4).

Empirical scoring function evaluates the binding affinity of a protein-ligand pair by counting standard types of interactions and assuming an average contribution for each to the free energy of the system. The interactions include salt bridges, hydrogen bonds and solvent-accessible surface area.

Force field-based scoring function takes the atom coordinates of a molecular system and computes its potential energy by explicitly modeling intermolecular forces including van der Waals and electrostatic forces.

The scoring function of eHiTS combines components from both knowledge-based and empirical approaches. In total, the various contributions to the score are given as a weighed sum of 20 terms:

\[ \text{eHiTS-Score} = W_1E_1 + \ldots + W_jE_j + \ldots + W_{20}E_{20} \quad (1 \leq j \leq 20) \]

where \( E_j \) is the energy term associated with an empirical or knowledge-based contribution, and \( W_j \) is a weight parameter.

In this study, a two stage eHiTS docking process was adopted. First, the target ligand was docked onto the HIV-1 receptor of which its active binding pocket was defined by the co-crystallized ligand. The ligand pose with the lowest eHiTS-Score was
chosen. Second, the ligand pose obtained from the first stage was used to define the active binding pocket and subsequently docked onto the receptor. The ligand pose with the lowest rmsd in reference to the ligand pose obtained from the first stage was chosen. Additionally, accuracy level was set at 6 in both stages. The commands for the docking process of ligand A021 used in both stages are as follows:

1st stage:
```
ehits --ligand A021.pdb --receptor 1AJV_pro_water.pdb --clip NMB_ligand.pdb --out A021_1AJV_process1.sdf --clean --accuracy 6
```

2nd stage:
```
ehits --ligand A021pose0_1AJV_process1.pdb --receptor 1AJV_pro_water.pdb --clip A021pose0_1AJV_process1.pdb --out A021_1AJV_process2.sdf --clean --accuracy 6 --rms A021pose0_1AJV_process1.pdb
```

The process of protein-ligand docking was performed using eHiTS program in the ENIAC system of the Computer Science Department at Hunter College, CUNY.

### 2.2 Phase 2 – Identify Residues close to Ligand

Solvent accessible surface area (SASA) procedures [22] with probe radii of 1.4 and 2.1 Å, written in TCL script were implemented on Visual Molecular Dynamics (VMD) [23] platform to identify residues close to the ligand in a ligand-protein complex. The steps of the SASA computation are as follows: First, calculate the SASA for each residue in the ligand-protein complex. Second, calculate the SASA for each residue in the isolated protein molecule. Third, subtract the results from the above two steps for the same residue. Residues that are close to the ligand have non-zero values after subtraction. The results of the examinations of the three-nine complexes reveal that 44 residues out of 198 are close to the ligands within 4.2 Å. Figure 2.6 shows the TCL script of SASA with probe radius of 1.4 Å for the A045-1AJV complex.
<table>
<thead>
<tr>
<th>Surface Point Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>METAL</td>
<td>positively charged metal ion point</td>
</tr>
<tr>
<td>CHARGED_HPLUS</td>
<td>positively charged hydrogen, e.g. Arginine</td>
</tr>
<tr>
<td>PRIMARY_AMINE_HLP</td>
<td>primary amine hydrogen/one-pair, e.g. –NH₂ or –NH₃⁺</td>
</tr>
<tr>
<td>HDONOR</td>
<td>strong (primary) hydrogen bond donor H (polar-atom-H)</td>
</tr>
<tr>
<td>WEAK_HDONOR</td>
<td>weak (secondary) hydrogen bond donor H (polarized C-H)</td>
</tr>
<tr>
<td>CHARGED_LONEPAIR</td>
<td>lone pair of negatively charged group, e.g. PO₃⁻</td>
</tr>
<tr>
<td>ACID_LONEPAIR</td>
<td>lone pair of an acid group, e.g. carboxylate</td>
</tr>
<tr>
<td>LONEPAIR</td>
<td>strong (primary) hydrogen bond acceptor lone pair</td>
</tr>
<tr>
<td>WEAK_LONEPAIR</td>
<td>weak (secondary) hydrogen bond acceptor lone pair</td>
</tr>
<tr>
<td>AMBITANT_HLP</td>
<td>donor H OR acceptor Lp depending on protonation state</td>
</tr>
<tr>
<td>ROTATABLE_H</td>
<td>rotatable-hydroxy donor H</td>
</tr>
<tr>
<td>ROTATABLE_LP</td>
<td>rotatable-hydroxy acceptor Lp</td>
</tr>
<tr>
<td>HYDROPHOB</td>
<td>H on aliphatic (chain) hydrophobic carbon</td>
</tr>
<tr>
<td>H_AROM.EDGE</td>
<td>H on hydrophobic carbon in aromatic ring (non-polarized)</td>
</tr>
<tr>
<td>WS_LIPO</td>
<td>H on weak secondary hydrophobic atom (e.g. carbon next to polar)</td>
</tr>
<tr>
<td>NEUTRAL</td>
<td>H on neutral atom (no recognized activity)</td>
</tr>
<tr>
<td>PI_AROMATIC</td>
<td>π electron of an aromatic ring</td>
</tr>
<tr>
<td>PI_RESON_POLAR</td>
<td>π electron on polar atom (N/O) in resonance chain, e.g. amide</td>
</tr>
<tr>
<td>PI_RESON_CARBON</td>
<td>π electron on carbon atom in resonance chain, e.g. amide</td>
</tr>
<tr>
<td>PI_SP2_POLAR</td>
<td>π electron on sp² polar atom (N/O) (non-resonating, non-atom)</td>
</tr>
<tr>
<td>PI_SP3_CARBON</td>
<td>π electron on sp³ carbon atom (non-resonating, non-atom)</td>
</tr>
<tr>
<td>HALOGEN</td>
<td>lone electron pair of a halogen atom (F/Cl/Br)</td>
</tr>
<tr>
<td>SULFUR</td>
<td>lone electron pair of a sulfur atom</td>
</tr>
</tbody>
</table>

**Table 2.4 The 23 Surface Point Types as Defined in eHiTS**

**2.3 Phase 3 – Principal Dataset Construction**

Three principal training datasets including DS-PIE, DS-RMLR, and DS-RMRR were constructed for the ML prediction of kinetic rate constants ($\log_{10}k_{on}$ and $\log_{10}k_{off}$). They cover both the kinetic characteristics and the thermal dynamic properties of PLI. Each dataset comprises thirty-nine feature vectors, and each vector contains forty-four training attributes and two target attributes ($\log_{10}k_{off}$ and $\log_{10}k_{on}$). DS-PIE covers PIE of a residue-ligand pair. Molecular dynamics simulations were conducted to compute the energy. DS-RMLR concerns the relative movement of a residue-ligand pair and...
# sasa of chain A, B, X. chain X is ligand
set outfilef0 [open A045_1AJV_f0_f02_r14.dat w]
set outfilef1 [open A045_1AJV_f1_f12_r14.dat w]
set outfilef02 [open A045_1AJV_f2_f02_r14.dat w]
set outfilef12 [open A045_1AJV_f2_f12_r14.dat w]

set complex02 [atomselect top "chain A or chain X"]
set complex12 [atomselect top "chain B or chain X"]
set frag0 [atomselect top "chain A"]
set frag1 [atomselect top "chain B"]
set frag2 [atomselect top "chain X"]
set x 1
set y 1
set z 1
set d 1

set residlistf0 [lsort -integer -unique [$frag0 get residue]]
foreach r $residlistf0 {
    set sel [atomselect top "residue $r"]
    set rsasa1($x) [measure sasa 1.4 $complex02 -restrict $sel]
    set rsasa2($x) [measure sasa 1.4 $frag0 -restrict $sel]
    set diff0 [expr $rsasa2($x) - $rsasa1($x)]
    $sel delete
    puts $outfilef0 "residue $r, complex-f02: $rsasa1($x) single-f0 $rsasa2($x) difference: $diff0"
    set x [expr $x + 1]
}
puts $outfilef0 " "

set residlistf1 [lsort -integer -unique [$frag1 get residue]]
foreach r $residlistf1 {
    set sel [atomselect top "residue $r"]
    set rsasa3($y) [measure sasa 1.4 $complex12 -restrict $sel]
    set rsasa4($y) [measure sasa 1.4 $frag1 -restrict $sel]
    set diff1 [expr $rsasa4($y) - $rsasa3($y)]
    $sel delete
    puts $outfilef1 "residue $r, complex-f12: $rsasa3($y) single-f1 $rsasa4($y) difference: $diff1"
    set y [expr $y + 1]
}
puts $outfilef1 " "

Figure 2.6 SASA TCL script with probe radius of 1.4Å for the A045-1AJV complex
DS-RMRR pertains to the relative movement of a residue upon ligand binding. Normal mode analysis was performed to compute the attribute values in DS-RMLR and DS-RMRR.

### 2.3.1 Training dataset DS-PIE

DS-PIE covers the PIE of a residue-ligand pair. The energy comprises two terms:

- **van der Waals energy** + **electrostatic energy** (kcal/mol)

  - **van der Waals energy** = \( C_{12}/r^{12} - C_{6}/r^6 \), where \( r \) is the distance between the two atoms’ nuclei, \( C_{12} \) and \( C_{6} \) are constants, whose values depend on the depth of the energy well and the equilibrium separation of the two atoms’ nuclei.

  - **electrostatic energy** = \( Cq_{i}q_{j}/\varepsilon r_{ij} \), where \( C \) is the Coulomb constant; \( q_{i} \) and \( q_{j} \) are point charges \( i \) and \( j \); \( \varepsilon \) is the dielectric constant, and \( r_{ij} \) is the distance between the two points. Since \( \varepsilon \) can range from 1 to 80 in a protein environment, a reasonable \( \varepsilon \) value is important to the correctness of the electrostatic energy calculation, and determines the accuracy of the PIE calculation. Average dielectric constants of different types of residues (Figure 2.7) reported by Li et al. [10], were adopted in this study to calculate the electrostatic energy. The dielectric constants range from 11.0 to 25.6 and are physically sound. Charged amino acids (Lys, Arg, Glu, and Asp) are associated with the highest average dielectric values. They also tend to be loosely packed on the protein surface, leaving room for structural rearrangement. On the other hand, hydrophobic residues (Cys, Ile, Phe, Val) are assigned with low dielectric values and they tend to be found in the protein core.

MD simulations of the thirty-nine ligand-HIV-1 models were carried out using the Nanoscale Molecular Dynamics (NAMD) [24] program with CHARMM27 force field for
HIV-1 protein and CHARMM general force field for the ligands [25]. All models proceeded through a minimization process of at least 4 ps and an equilibrium process of 200 ps. During the simulations, temperature was set at 310 K, and the generalized born implicit solvent method was used with the ionConcentration, GBISDelta, GBISBeta, GBISGamma and alphaCutoff set at 0.15, 0.8, 0.0, 2.90912 and 14, respectively [26-28]. After loading the 200 ps trajectory file produced from MD simulations into VMD, the value of PIE of each unique residue-ligand pair was calculated using the NAMDEnergy plugin in VMD. Figure 2.8 shows the NAMDEnergy graphical user interface for the computation of the PIE between ligand A045 and leucine of chain A. The expression of “(chain A) and (resid 10)” on the tab of Selection 1 identifies leucine of chain A with the dielectric constant of 11.8 entered on the dielectric tab and the expression of “resid 501” on the tab of Selection 2 identifies ligand A045. MD simulations were performed at the High Performance Computing Center at the College of Staten Island, CUNY.

2.3.2 Training Datasets DS-RMLR and DS-RMRR

The training attributes of DS-RMLR and DS-RMRR illustrate the relative movement of a ligand-residue pair and the relative movement of a residue upon ligand binding respectively. The NMA system, iMOD [29] developed by Lopez-Blanco et al. in 2011 was adopted to compute the displacement vectors of the residues and the ligands in the models. iMOD uses internal coordinates instead of Cartesian coordinates and defines the potential energy as follows:
Figure 2.7 Average dielectric constants of different types of amino acids.

Figure 2.8 NAMDEnergy graphical user interface.
Potential energy = $\sum_{i<j} F_{ij} (r_{ij} - r_{0ij})^2 + s \sum_\alpha (\theta_\alpha - \theta_0^\alpha)^2$, where

- $r_{ij}$ is the distance between atom $i$ and $j$, and the super-index 0 indicates the initial equilibrium conformation.
- $F_{ij}$ is the matrix whose elements describe the force constant associated with each atom pair. $F_{ij} = k/(1 + (r_{0ij}/r_o)^p)$ if $r_{0ij} < r_{\text{cut}}$, otherwise $F_{ij} = 0$ and $k$, $r_o$, $p$ and $r_{\text{cut}}$ were set to 1, 3.8Å, 6 and 10Å respectively.
- The second term of the energy equation is added for tip effect prevention. $\theta_\alpha$ is the dihedral angle [30].

Two applications, imode and imodview of iMOD were used in this study. First, imode program was used with –save_cart option to produce Cartesian normal modes in the output file with .evec extension. Second, the output file was used as an input file for imodview to compute the 3D vector sets of residues and ligands for the ten lowest frequency modes ($n = 1$ to 10). Residue/ligand molecule was set to be the averaging level to compute the arrow of eigenvector (level =1). According to recent studies, the first ten lowest frequency modes cover nearly 90% of protein conformational change, and thus, it is necessary for the training attributes consisting of information from the first 10 lowest frequency modes [30]. The NMA Training Attribute Value is defined as follows:

\[ \text{NMA Training Attribute Value (NTAV)} = (\text{DPVV}_1^2 + ... + \text{DPVV}_j^2 + ... + \text{DPVV}_{10}^2)^{1/2}, \]

where $j = 1$ to 10 is the normal mode index. DPVV is either the dot product of ligand displacement vector after normalization and residue displacement vector in DS-RMLR or the dot product of two displacement vectors of a residue upon ligand binding in DS-RMRR. Specifically, after aligning to the corresponding ligand-bound complex, closed-
flap HIV-1 protease, 3IXO (PDB ID code), was used as the unbound structure of HIV-1 protease.

2.4 Phase 4 - Machine Learning

- Machine learning is composed of a training phase and a predicting phase. In unsupervised learning, there is no outcome and the goal is to describe associations and patterns among a set of input variables. In supervised learning, the goal is to build models to predict the value of an outcome variable based on a number of input variables. Supervised learning has two categories: classification and regression. In classification, objects are placed into one or several predefined discrete classes based on a training set of data that contains observations whose class membership is known. Regression is similar to classification, but maps objects to a real-valued outcome variable. The algorithm (estimator) that implements classification or regression is known as classifier or regressor, respectively. Typically, either a single-target classifier or a single-target regressor builds models to predict only one output variable. But because the two kinetic rate constants \( (k_{off}, k_{on}) \) are correlated, we used a multi-target classifier and a multi-target regressor to train coherent binary-output models to increase the predictive performance for the two constants. Thus, a multi-target random forest classification algorithm of Clus [31] and a multi-target lasso regression algorithm of Scikit-Learn [32] were used to predicted the two kinetic rate constants simultaneously. Additionally, a single-target k-nearest-neighbors instance-based learning algorithm of Scikit-Learn and three single-target regression algorithms of Scikit-Learn including a random forest algorithm, a lasso algorithm and an elastic net algorithm, were adopted to predict the two kinetic rate constants seperately.
2.4.1 Cross-Validation

A model induced by a machine learning algorithm should be validated to verify the predictive ability of the model and to select the most appropriate method for further tuning and refinement. In cross-validation, a subset of the data provided is kept aside and the remaining data, the training set, is used by the algorithm to build a model. In k-fold cross-validation, the dataset is row-wise randomly divided into k subsets of equal size. Each subset is left out once and a model is induced from the remaining k -1 subsets. The data in the subset that is left out is used to assess the predictive performance of the model. The performance of the model on each test set is averaged to compute the overall performance of the algorithm. The advantage of k-fold cross-validation is that each instance of the dataset gets to be in the test set once and it can be used to evaluate the performance of the model within a single dataset. This approach becomes computationally demanding for large amounts of data and is, therefore, better suited for small sized datasets. In this study, leave-one-out (LOO) cross-validation is used. Thus, k is equal to 39 because there are thirty-nine feature vectors in each of the three datasets.

2.4.2 Performance Measurement

The predictive quality measure depends on whether the modeling task is classification or regression. A classification algorithm predicts the class value taken by the output attribute for a given instance in the test set. Prediction results are represented as a confusion matrix, with rows corresponding to actual values and columns corresponding to predicted values for the output attributes. Each block in the confusion matrix gives the number of times the actual class is predicted as the class given by the column. The numbers in the diagonal blocks give the number of time the predicted class
value was equal to the actual class value. Thus, the sum of entries along the diagonals divided by the total number of instances in the test set gives the percentage of the number of correctly classified instance. In the case of multiple n-fold cross-validations, the confusion matrices obtained for each test set seen are averaged to obtain a confusion matrix with the mean values.

Regression algorithms are used to determine the value taken by the output attribute in the given instance, based on an equation or mathematical operations. The performance of a regression algorithm can be determined by the difference between the actual value and the predicted value, which gives the amount of error in the prediction mode. In the case of n-fold cross-validation, the error value is averaged across all the test sets seen.

### 2.4.2.1 Quality Measures for Classification

Figure 2.9 depicts a confusion matrix for a binary classifier with two outcomes, high and low binding affinity (1,0). The targets that are correctly classified are denoted as true positives (TP) and true negatives (TN), and the targets that are misclassified are denoted as false positives (FP) and false negatives (FN). Sensitivity and specificity are the true positive (TP) and true negative (TN) rate respectively.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>High (1)</td>
<td>TP</td>
</tr>
<tr>
<td>Low (0)</td>
<td>FN</td>
</tr>
</tbody>
</table>

**Figure 2.9 A confusion matrix for a binary classifier.**
Common measures are accuracy and error rate. Accuracy is the percentage of a test set that is correctly classified and error rate is simply the percentage of a test set that is misclassified. They are computed as:

sensitivity = TP / (TP + FN)

specificity = TN / (TN + FP)

accuracy = 100 x (TP + TN) / (TP + FN + TN + FP) %

error rate = 100 x (FP + FN) / (TP + FN + TN + FP) %

2.4.2.2 Quality Measures for Regression

Two quality measures are used for regression analysis in this study.

1. Pearson’s Correlation Coefficient (PC or r) is a measure of the correlation of linear dependence between two variables X and Y, giving a value between +1 and -1 inclusive. It is defined as the covariance of the two variables divided by the product of their standard deviations. A value of 1 suggests that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of -1 indicates that all data points lie on a line for which Y increases as X decreases. A value of 0 implies that there is no linear correlation between the variables.

\[ r = \frac{1}{n - 1} \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right) \]

where \( \bar{X} \), \( \bar{Y} \), and \( s_X \) are the standard score, sample mean, and sample deviation; \( n \) is the number of samples.

2. %deviation: it is a measure in percentage of how far a predicted value deviates from the actual value and is calculated as:

\[ 100 \times \frac{|\text{predicted value} - \text{actual value}|}{|\text{actual value}|} \]
2.4.2.3 Quality Measure of Multi-Target Prediction

There are two measurements for multi-target prediction:

1. Multi-target measure for classification prediction:

\[
\text{MM-Accuracy} = \left( \text{kon-Accuracy}^2 + \text{koff-Accuracy}^2 \right)^{1/2}
\]

where \( \text{kon-Accuracy} \) is the accuracy of \( k_{on} \) prediction and \( \text{koff-Accuracy} \) is the accuracy of \( k_{off} \) prediction.

2. Multi-target measure for regression prediction:

\[
\text{MM-%deviation} = \left( \text{kon-%deviation}^2 + \text{koff-%deviation}^2 \right)^{1/2}
\]

where \( \text{kon-%deviation} \) is the %deviation of \( k_{on} \) prediction and \( \text{koff-%deviation} \) is the %deviation of \( k_{off} \) prediction.

2.5 Feature Evaluation

In this study, two techniques were used to evaluate the characteristics of the features in the principal training datasets. They are two-tailed Welch’s t-test and feature selection. The feature selection technique is intended to reduce overfitting and computational cost, and improve interpretability. The purpose of the two-tailed Welch’s t-test technique is to evaluate the behaviors of the features in PLI in response to the change of the kinetic rate constants.

2.5.1 Two-tailed Welch’s t-test

Evenly splitting the thirty-nine training records in each of the three principal datasets (DS-PIE, DS-RMLR, and DS-RMRR) into two subsets (X and Y) using one of the three criteria: \( \log_{10} k_{on} \geq 5.5502, k_{off} \geq 0.00653, \) and \( \log_{10} k_D \leq -7.9856 \). In total, nine binary subsets were generated from the three principal datasets and the three criteria. Table
2.5 shows the results of the splitting. Two-tailed Welch’s t-test was conducted on these nine pairs of binary subsets to identify the features having p-value < 0.05.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>log₁₀ₐₖᵋₚ ≥ 5.5502</th>
<th>kᵋᵧ ≥ 0.00653</th>
<th>log₁₀ₐₖᵋₜ ≤ -7.9856</th>
</tr>
</thead>
</table>

Table 2.5 Results of splitting. Using the three criteria: log₁₀ₐₖᵋₚ ≥ 5.5502, kᵋᵧ ≥ 0.00653, and log₁₀ₐₖᵋₜ ≤ -7.9856, nine different pairs of binary subsets were produced from the three principal datasets, DS-PIE, DS-RMLR, and DS-RMRR.

2.5.2 Feature Selection

Statistical experiment was conducted to identify the training features in DS-PIE, DS-RMLR, and DS-RMRR with frequency of occurrence greater than 25% in the LOO cross-validation experiment of the binary-target random forest classification algorithm with iteration number = 500. The features selected from DS-PIE, DS-RMLR, and DS-RMRR were assigned to build new datasets F-PIE, DS-RMLR, and DS-RMRR respectively.

2.6 Using Receiver Operating Characteristic Curve to Differentiate PIRM from Non-PIRM Residue

Protease inhibitor resistance mutation is the mutation that code for the changes of the protein conformational shape facilitate resistance of HIV to protease inhibitors. There are twenty-six protease inhibitor resistance mutations (PIRM) reported by the World Health Organization in 2013 [33].
A receiver operating characteristic (ROC) curve is a graphical plot that depicts the performance of a binary classifier system. The curve is created by plotting the true positive rate (sensitivity) versus false positive rate (1 – specificity), as some discrimination threshold is varied. An area under the ROC curve close to 1.0 means that the classifier is able to perfectly map objects into classes. An area close to 0.5 means that the classifier does not perform better than random guessing.

A ROC curve is created for each of the three training datasets, F-PIE, DS-RMLR, and DS-RMRR according to the following procedure:

1. Sort the score of importance of the residues in F-PIE, DS-RMLR, and DS-RMRR in descending order.
2. Choose different values of score of importance as different thresholds for the curve construction.
3. Compute True Positive Rate and False Positive Rate as follows:
   
   True Positive Rate = number of cumulative PIRM residues / 24

   False Positive Rate = number of cumulative non-PIRM residues / 20  
   
   where 24 and 20 are the number of PIRM residues and the number of non-PIRM residues among the 44 training features.

3. Construct the ROC curve by plotting True Positive Rate versus False Positive Rate.
2.7 Clustering

K-Means cluster algorithm with various k values (k = 4, 6 and 8) was used to cluster the three principal datasets after normalization. The datasets were normalized according to the following equation:

\[ X_i = \frac{(S_i - M)}{SD} \]

where

\( S_i \) is a sample in the column containing \( n \) samples. \( M \) and \( SD \) are the mean and standard deviation of the \( n \) samples in the column.
Chapter 3

Molecular Dynamics Simulations

In 1977, McCammon et al. [34] reported the first molecular dynamics simulations of bovine pancreatic trypsin inhibitor by solving the equations of motion for the atoms with an empirical potential energy function. The results reveal the correlations, and magnitude of fluctuations about the average structure and suggest that the protein interior is fluid-like in which the atom motions have a diffusional property. Since then, MD simulation has been developed at a very fast pace.

There are three major applications of MD simulation in biology. First, MD simulations are used in ligand-docking applications; second, MD simulations give insights into the natural dynamics on different time scales of bio-molecules in solution; and third, MD simulations give thermal averages of molecular properties [35,36].

Figure 3.1 depicts a simplified description of an MD simulation algorithm. The simulation proceeds iteratively by alternatively calculating forces and solving the equations of motion based on the accelerations obtained from the new forces [37].

MD simulation begins with an initial set of atomic coordinates. The coordinates can be obtained from X-ray crystallographic or NMR structure data. The structure is first refined using an iterative minimization algorithm to relieve local stresses due to overlaps of non-bonded atoms and bond length distortions. Next, atoms are assigned velocities \(v\) taken at random from a Maxwellian distribution for a low temperature. Then, a simulation is performed for a few picoseconds. The equilibration is continued by alternating new velocity assignments, chosen from Maxwellian distributions for temperatures that are successively increased to a chosen value, with intervals of
dynamical relaxation. The temperature $T$ of the system is measured by the mean kinetic energy,

$$(1/2)\sum_{i=1}^{N} M_i (V_i^2) = 3/2 (Nk_B T)$$

where $M, V$ are the mass and the velocity of $i$th atom respectively, $N$ is the number of atoms in the system, and $k_B$ is the Boltzmann constant. The equilibrium period is considered finished when the temperature is stable for longer than about 10 ps.

![Diagram of molecular dynamics simulation algorithm](image)

**Figure 3.1** Molecular dynamics simulation algorithm. $r, v, a, \text{ and } F$ represent position, velocity, acceleration, and force respectively.

### 3.1 Ergodic Hypothesis

According to the ergodic hypothesis [38], one can simulate a single molecule with its surroundings for a period of time and get time-averaged molecular properties that approach the experimentally measurable ensemble averages. The ergodic hypothesis states: Ensemble average ($<A>_{\text{ensemble}}$) = Time average ($<A>_{\text{time}}$)
The basic idea is that if one allows the system to evolve in time indefinitely that system will eventually pass through all possible states. Hence, one of the goals of MD simulations is to generate enough representative conformations such that this equality is satisfied.

In statistical mechanics, average values are defined as an ensemble average. The ensemble average is given by

\[ \langle A \rangle_{\text{ensemble}} = \int \int dp^N dr^N A(p^N, r^N) \rho(p^N, r^N) \]

where

\[ A(p^N, r^N) \] is the observation of interest and is expressed as a function of the number of particles \( N \), the momenta \( p \), and the position \( r \), of the system. Integration is over all possible variables of \( r \) and \( p \). The probability density of the ensemble is given by

\[ \rho(p^N, r^N) = \left( \frac{1}{Q} \right) \exp\left[ -\frac{H(p^N, r^N)}{k_BT} \right] \]

where

\[ H \] is the Hamiltonian, \( T \) is the temperature, \( k_B \) is Boltzmann’s constant, and \( Q \) is the partition function

\[ Q = \int \int dp^N dr^N \exp\left[ -\frac{H(p^N, r^N)}{k_BT} \right] \]

This integral is generally extremely difficult to calculate because one must calculate all possible states of the system. In an MD simulation, the points in the ensemble are calculated sequentially over time, so to compute an ensemble average, the MD simulations must pass through all possible states corresponding to the particular thermodynamic constraints.

In an MD simulation, the time average of \( A \) is expressed as

\[ \langle A \rangle_{\text{time}} = \lim_{\tau \to \infty} \left( \frac{1}{\tau} \int_{t=0}^{\tau} A(p^N(t), r^N(t)) dt \right) \approx \left( \frac{1}{S} \right) \sum_{t=1}^{S} A(p^N, r^N) \]

where \( \tau \) is the simulation time, \( S \) is the number of time steps in the simulation, and \( A(p^N, r^N) \) is the instantaneous value of \( A \).
3.2 Potential Energy Function

Molecular dynamics simulations begin with knowledge of the energy of the system as a function of the atomic coordinates, \( R \), of all the atoms in the system. The forces, acting on the atoms, which are related to the first derivatives of the potential energy (PE) with respect to the atom positions, can be used to compute the dynamic behaviour of the system by solving Newton’s equations of motion for the atoms as a function of time. The value of the energy is computed as the sum of internal or bonded terms called \( E_{\text{bonded}} \) that describes the bonds, angles and bond rotations in a molecule, and the sum of external or non-bonded terms called \( E_{\text{non-bonded}} \). These terms account for interactions between non-bonded atoms or atoms separated by three or more covalent bonds.

\[
\text{PE}(R) = E_{\text{bonded}} + E_{\text{non-bonded}}
\]

The \( E_{\text{bonded}} \) term is a sum of three terms:

1. \( E_{\text{bond-stretch}} = \sum_{1,2 \text{ pairs}} K_b (b - b_0)^2 \)

\( E_{\text{bond-stretch}} \) is a harmonic potential representing the interaction between atomic pairs, where atoms are separated by one covalent bond (1,2-pairs). This is the approximation to the energy of a bond as a function of displacement from the ideal bond length, \( b_0 \). The force constant \( K_b \) determines the strength of the bond. Both \( b_0 \) and \( K_b \) are specific for each pair of bound atoms, that is, depend on the chemical type of atoms-constituents.

2. \( E_{\text{bond-bend}} = \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 \)

\( E_{\text{bond-bend}} \) is a harmonic potential representing the interaction between atomic pairs, where atoms are separated by one covalent bond (1,2-pairs). This is the approximation to the energy of a bond as a function of displacement from the ideal bond length, \( b_0 \). The force constant \( K_\theta \) determines the strength of the bond. Both \( b_0 \) and \( K_\theta \) are specific for each pair of bound atoms, that is, depend on the chemical type of atoms-constituents.
E_{bond-bend} is associated with the alteration of bond angles $\theta$ from ideal values $\theta_0$, which is represented by a harmonic potential. Values of $\theta_0$ and $K_\theta$ depend on the chemical type of atoms constituting the angle.

$$(iii) \ E_{\text{rotate-along-bond}} = \sum_{1,4 \ \text{pairs}} K_\phi (1 - \cos (n\phi))$$

$E_{\text{rotate-along-bond}}$ represents the torsion angle potential function, which models the presence of steric barriers between atoms separated by three covalent bonds (1,4 pairs). The motion associated with this term is a rotation, described by a dihedral angle and coefficient of symmetry around the middle bond.

$E_{\text{non-bonded}}$ has two components, the van der Waals interaction energy and the electrostatic interaction energy. It represents the contribution of non-bonded interactions.

$$(II) \ E_{\text{non-bonded}} = E_{\text{van-der-Waals}} + E_{\text{electrostatic}}$$

The van der Waals interaction between two atoms arises from a balance between attractive and repulsive forces. The attractive force arises from fluctuations in the charge distribution in the electron clouds. The repulsive force arises at short distances where the electron-electron interaction is strong.

$$(i) \ E_{\text{van-der-Waals}} = \sum_{\text{nonbonded pairs}} \left[ (K_{12}/r^{12}) - (K_6/r^6) \right]$$

where $r$ is the distance between the two atoms' nuclei. $K_{12}$ and $K_6$ are constants whose values depend on the depth of the energy well and the equilibrium separation of the two atoms' nuclei.

$$(ii) \ E_{\text{electrostatic}} = C q_i q_j / \varepsilon r_{ij}$$

where $C$ is the Coulomb constant; $q_i$ and $q_j$ are point charges $i$ and $j$; $\varepsilon$ is the dielectric constant; and $r_{ij}$ is the distance between the two points.
3.3 Average Dielectric Constants of Different Types of Residues

Average dielectric constants of different types of residues (Figure 2.7) reported by Alexov et al. were adopted in this study to calculate the electrostatic energy. The author used a Gaussian-based approach to deliver a smooth dielectric function for the space domain of the protein and its water environment. There are three steps in the computation. First, for a protein molecule with N atoms, the Gaussian distribution of the density of an atom $i$ is given by:

$$\rho_i(r) = \exp\left[-\frac{r_i^2}{(\sigma^2 R_i)^2}\right]$$  \[3.3-1\]

where $\rho_i(r)$ is the density at position $r$, $r_i$ is the distance between the center of the atom $i$ and position $r$, $R_i$ is the vdW radius of atom $i$ and $\sigma$ is the variance in the Gaussian distribution. Second, the density in the overlapping area occupied by multiple atoms is calculated by

$$\rho_{mol}(r) = 1 - \prod_i [1 - \rho_i(r)]$$  \[3.3-2\]

where the $\rho_{mol}(r)$ is the density at position $r$ coming from multiple atoms; $\rho_i(r)$ is the density of a single atom $i$, which is obtained from equation [3.3-1]. The density $\rho_{mol}(r)$ always stays between 1 and 0. Third, the smooth dielectric function is represented by a linear function:

$$\varepsilon = \rho \cdot \varepsilon_{in} + (1 - \rho)\varepsilon_{out}$$  \[3.3-3\]

where $\varepsilon$ on the left denotes the dielectric distribution function; $\varepsilon_{in}$ denotes the reference dielectric value when the density is 1 ($\varepsilon_{in} = 4$); $\varepsilon_{out}$ denotes the reference dielectric value for the water phase ($\varepsilon_{out} = 80$); and $\rho$ is the density obtained from equation [3.3-2].
The dielectric values per residue type were computed on 91 protein structures and the results were averaged per amino acid type. The 91 protein structures were selected from a large set of diverse proteins taken from the PDB bank using a three-step filtering process. First, only protein structures determined by X-ray experiments with a resolution less than 1.5Å were chosen. Second, the protein structures with a sequence similarity more than 30% were removed. Third, the protein structures with cofactors which are not made of regular residues were also removed.

For each residue, the average dielectric values of all side chains were calculated as described in equation [3.3-3]. The dielectric values ranged from 11.0 to 25.6 and thus, are physically sound. Charged amino acids (Lys, Arg, Glu, and Asp) are associated with the highest average dielectric values. They tend to be loosely packed on the protein surface, leaving room for structural rearrangement. On the other hand, hydrophobic residues (Cys, Ile, Phe, Val) are assigned with low dielectric values. They tend to be found in the protein core [10].

### 3.4 Integration Algorithms

Potential energy is a complicated function of the atomic positions of all atoms in the system and can be solved numerically by the following numerical integration algorithms:

- Verlet algorithm
- Leap-frog algorithm
- Velocity Verlet
- Beeman’s algorithm
All the positions, velocities, and accelerations in the integration algorithms are approximated by a Taylor series expansion:

\[
\begin{align*}
    r(t + \delta t) &= r(t) + v(t)\delta t + (1/2)a(t)\delta t^2 + \ldots \\
v(t + \delta t) &= v(t) + a(t)\delta t + (1/2)b(t)\delta t^2 + \ldots \\
a(t + \delta t) &= a(t) + b(t)\delta t + \ldots 
\end{align*}
\]

where \( r \) is the position, \( v \) is the velocity (the first derivative with respect to time), and \( a \) is the acceleration (the second derivative with respect to time).

### 3.5 NAnoscale Molecular Dynamics - NAMD

NAMD [39,40] is a scalable classical molecular dynamics package developed by the Theoretical and Computational Biophysics Group at the University of Illinois at Urbana-Champaign. It was first introduced in 1995 by Nelson et al. as a parallel molecular dynamics code enabling interactive simulation by linking to the visualization code VMD [41].

NAMD is implemented using the Converse runtime system. The major components of NAMD are written in Charm++, noted for its parallel efficiency and often used to simulate large systems of millions of atoms. Converse provides a machine-independent interface to all popular computers as well as workstation clusters. Converse consists of five major components including a machine interface (that supports communication, timers, and other operating system calls), scheduler queues, a threads package, a message manager, and a load-balancing package.

Charm++ is a parallel, object-oriented programming language based on C++ and developed in the Parallel Programming Laboratory at the University of Illinois. It is designed with the goal of enhancing programmer productivity by providing a high-level
abstraction of a parallel program while at the same time delivering good performance on a wide variety of underlying hardware platforms. Programs written in Charm++ are decomposed into a number of cooperating message-driven objects called chares. When a programmer invokes a method on a chare, the Charm++ runtime system sends a message to the invoked chare. The chare may reside on the local processor or on a remote processor in a parallel computation. This message triggers the execution of code within the chare to handle the message asynchronously.

3.5.1 Velocity Verlet Integration Algorithm

The Velocity Verlet algorithm was proposed by Swope et al. in 1982. The algorithm computes the particle position and velocity at the next time step \((r_{n+1}, v_{n+1})\) from the current one \((r_n, v_n)\), assuming the force \(F_n = F(r_n)\) is already computed, in the following ways:

Step 1: half-kick
\[
v_{n+1/2} = v_n + \frac{m^{-1}F_n}{2} \Delta t
\]  
[3.5.1-1]

Step 2: drift
\[
r_{n+1} = r_n + v_{n+1/2} \Delta t
\]  
[3.5.1-2]

Step 3: computing force
\[
F_{n+1} = F(r_{n+1})
\]  
[3.5.1-3]

Step 4: half-kick
\[
v_{n+1} = v_{n+1/2} + \frac{m^{-1}F_{n+1}}{2} \Delta t
\]  
[3.5.1-4]

where \(\Delta t\) is a small time increment, \(m\) is the particle mass, and \(F_n\) is the total force acting on the particle at time step \(n\). Note that by implementing the algorithm one generates a phase space trajectory, in which a sequence of snapshots for particle coordinates and velocities at all intermediate times \(t_j = j \times \Delta t\) \((j = 1, 2, \ldots, n)\) is generated. MD simulations thus provide the sequence of microscopic configurations through which the model system passes in time. Such detailed microscopic information allows scientists to compute the result of an ensemble average measurement according to the time average measurement.
3.5.2 Implicit Solvation - $GB^{OBC}$ Model

Implicit solvation [28,42,43,44] is a method that represents a solvent as a continuous medium instead of as the individual explicit solvent molecules used in MD simulations. Implicit solvent models are continuum models that attempt to capture the average effect of the water on a solute. They have five advantages over the explicit water representation in MD simulations. First, the computational cost is lower. Second, sampling of conformational space is enhanced, due to the absence of viscosity associated with the explicit water environment. Third, instantaneous dielectric response from the solvent eliminates the need for lengthy equilibration of water that is typically necessary in explicit water simulations. Fourth, it is more effective for free energy estimation. Since solvent degrees of freedom are taken into account implicitly, estimating the free energies of solvated structures is more straightforward than with explicit water models. Fifth, the continuum model corresponds to solvation in an infinite volume of solvent, thereby avoiding possible artifacts of the replica interactions that occur in the periodic systems typically used for explicit water calculations.

Solvation free energy ($\Delta G_{solv}$) consists of two terms including electrostatic ($\Delta G_{el}$) and nonelectrostatic ($\Delta G_{surf}$) parts: $\Delta G_{solv} = \Delta G_{el} + \Delta G_{surf}$ where $\Delta G_{surf}$ is the free energy of solvating a molecule from which all charges have been removed (i.e., the partial charges of every atom are set to zero); and $\Delta G_{el}$ is the free energy of first removing all charges in the vacuum, and then adding them back in the presence of the solvent environment. $\Delta G_{surf}$ is approximated by taking to be proportional to the total solvent-accessible surface area (SASA) of the molecule, $\Delta G_{surf} \approx \sigma \times \text{SASA}$,
with the proportionality constant \((\sigma)\) derived from experimental solvation energies of small non-polar molecules. Relative to the Poisson-Boltzmann (PB) model, the analytical Generalized Born (GB) model is an approximate way to calculate the electrostatic part of the solvation free energy, \(\Delta G_{el}\). Due to its relative simplicity and computational efficiency, the GB model has become popular in MD simulations.

\[
\Delta G_{el} \approx \Delta G_{GB} = (-1/2) \sum_{ij} \left(\frac{q_i q_j}{f_{GB}(r_{ij}, R_i, R_j)} \right) \left[1 - \exp\left(-\frac{K f_{ij}^{GB}/\varepsilon_w}{1}\right)\right] \tag{3.5.2-1}
\]

\[
f_{GB} = \left[r_{ij}^2 + R_i R_j \exp\left(-\frac{r_{ij}^2}{4R_i R_j}\right)\right]^{1/2} \tag{3.5.2-2}
\]

Within the \(GB^{OBC}\) model (OBC represents the authors, Onufriev, Bashford, and Case), each atom in a molecule is a sphere of radius \(\rho_i\) with a charge \(q_i\) at its center. The interior of the atom is assumed to be filled uniformly with material of dielectric constant 1. The molecule is surrounded by a solvent of a high dielectric value \(\varepsilon_w\) (80 for water at 300K). \(r_{ij}\) is the distance between atom \(i\) and \(j\). \(R_i\) is the effective Born radius of atom \(i\), which reflects the degree of the atom’s burial inside the molecule. For an isolated ion, \(R_i\) is equal to its van der Waals (VDW) radius \(\rho_i\), while for a deeply buried one, \(R_i \gg \rho_i\).

Additionally, the electrostatic screening effects of monovalent salt are incorporated into formula [3.5.2-1] via the Debye-Huckel screening parameter \(\kappa (\AA^{-1}) \approx 0.316/[/salt](\text{mol/L})\). The effective Born radius is defined as follows:

\[
R_i^{-1} = \rho_i^{-1} - \rho_i^{-1} \tanh(\alpha \Psi - \beta \Psi^2 + \gamma \Psi^3) \quad \text{where}
\]

\[
\rho_i = \rho_i - 0.09\AA \quad \text{and} \quad \Psi = I \varrho_i \quad \text{where}
\]

\[
I = \left(1/4\pi\right) \int_{\text{VDW}} \theta(|r| - \varrho_i)(1/r^4)d^3r \quad \text{where}
\]

\(\theta\) is a step function that ensures the volume of atom \(i\) itself is excluded from the integration, and \(r\) is the coordinate vector. There are two sets of \(\{\alpha, \beta, \gamma\}\):
GB$^{\text{OBC}}$ I: $\alpha = 0.8, \beta = 0, \gamma = 2.91$

GB$^{\text{OBC}}$ II: $\alpha = 1.0, \beta = 0.8, \gamma = 4.85$

Both parameter sets I and II were verified to achieve equivalent accuracy in the experiment of $\Delta G_{el}$ computation for protein-A at neutral pH [28]. GB$^{\text{OBC}}$ I was adopted for the MD simulations in the study.

### 3.5.3 Scalability of Decomposition Strategy

NAMD2 uses hybrid spatial/force decomposition (QSD and FD) strategies with communication cost per node of $O(N/P)$. $N$ is the number of atoms and $P$ is the number of processors. FD is the force decomposition and QSD stands for quantized spatial decomposition.

FD is a technique that distributes the sparse force matrix in a block-wise fashion across processors. Each processor’s share of the $N \times N$ force matrix is a block of size $(N/\sqrt{P}) \times (N/\sqrt{P})$. To compute this block the processor needs the coordinates of $2N/\sqrt{P}$ atoms, which come from $\sqrt{P}$ different processors. The communication cost per processor is thus $O(N/\sqrt{P})$.

QSD assigns nearby atoms to the same processor by partitioning sample space into fixed-size boxes, with dimension larger than the cutoff distance, requiring communication only between neighboring boxes. The communication cost per node is $O(N\sqrt{P})$.

### 3.5.4 Load Balancing

The following procedure was used by NAMD to perform load balancing for MD simulations. First, the simulation runs for a small number of steps that typically lasts a few minutes. Migratable object times ($t_{\text{migratable}}$) are measured during this time. In
addition, the Converse system provides a means of measuring idle time \( t_{idle} \) for each processor, so non-migratable load is determined by 
\[ t_{non-migratable} = t_{total} - t_{idle} - t_{migratable}. \]

Second, an average load is computed from the total measured load. Then each migratable object is examined, starting with the most expensive. The load balancer considers three possible situations for the computation: the least-loaded processor can receive two proxies, one proxy, or no proxies. A proxy is a representative of the home patch containing the coordinates of patches in a processor. The object is assigned to the two-proxy processor if the resulting load on the processor is less than the average load times an overload factor. Otherwise, the one-proxy processor is considered, and finally, if necessary, the no-proxy processor. Then, the load balancer adds the object load to the selected processor’s total load, and updates the proxy map. This procedure repeats until all computational objects are assigned.

Third, the load balancer executes a refinement procedure. It examines the processor assignments again. A smaller overload factor is selected. Any processors whose load exceeds the average by more than the overload factor has a number of force objects moved to lighter loaded processors. Finally, the simulation runs for a few more steps. Then it stops for a second load balancing to handle the changes in the background load due to the changes in the communication patterns caused by the previous step.

### 3.5.5 Stages in a Typical NAMD MD Simulation

Figure 3.2 depicts the five stages in a typical NAMD MD simulation.

(A) Initial stage: Four input files are required to start a NAMD MD simulation. They include a configuration file, a parameter file, a PDB file and a PSF file. The flowchart in Figure 3.3 indicates the role of files as used by NAMD and VMD.
(a) The configuration file is given to NAMD on the command line and specifies what
dynamics options and values that NAMD should use, such as the number of timesteps to
perform, initial temperature, what features are active or inactive, etc. The options and
values in this file control how the system will be simulated. The following parameters in
the configuration file are required for NAMD MD simulation:

1. numsteps: The number of timesteps will be performed in the simulation.

2. coordinates (PDB file): The PDB file contains initial position coordinate data.

3. structures (PSF file): The X-PLOR format PSF file describes the molecular
   system to be simulated.

4. parameters (parameter file): A CHARMM19, CHARMM22, or CHARMM27
   parameter file that defines all or part of the parameters necessary for the
   molecular system to be simulated.

5. exclude: This parameter specifies which pairs of bonded atoms should be
   excluded from non-bonded interactions.

6. outputname: At the end of every simulation, NAMD writes two files, one
   containing the final coordinates and another containing the final velocities of all
   atoms in the simulation. The position coordinates will be saved to a file named as
   this prefix with .coor appended. The velocities will be saved to a file named as
   this prefix with .vel appended.

7. one of the following three: temperature, velocities or binvelocities.

The configuration file of MD simulation for A045-bound HIV-1 protein complex is
shown in Figure 3.4.
(b) The PDB file contains ATOM and HETATM records, which describe the coordinates of the protein and any waters, ions, or other heterogeneous atoms in the crystal molecule.

(c) The PSF file, also called protein structure file, contains all of the molecule-specific information needed to apply a particular force field to a molecular system. Given a PDB file and a topology file, a PSF file can be generated using the psfgen program in VMD.

(d) The Parameter file supplies numerical values needed to evaluate forces and energies, given a PSF structure file and atomic coordinates (PDB file).

(B) Minimization stage: The purpose of the minimization stage is to adjust the structure to the force field and the particular distribution of solvent molecules, and to relax possible steric clashes created by guessing coordinates of atoms during generation of PSF file. When the energy change from step to step is less than 0.001 kcal/mol, the structure is sufficiently minimized.

(C) Heating stage: During the heating stage, the temperature of the system is linearly increased from 0K to the assigned temperature. At each integration step, velocities are reassigned from a new Maxwell distribution and the temperature is incremented.

(D) Equilibration stage: This stage is designed for the equilibration between kinetic and potential energies, i.e., to distribute the kinetic energy added into the system during heating among all degrees of freedom. In other words, the kinetic energy must be transferred to potential energy. As soon as potential energy levels off the equilibration stage is completed.

(E) Production stage: In this stage, the MD simulation samples the structural characteristics and dynamics of the model at the assigned temperature.
Figure 3.2 Stages in a NAMD MD simulation.

Figure 3.3 Flowchart indicating the role of files as used by NAMD and VMD
structure    A045_1AJV.psf
Coordinates  A045_1AJV.pdb
set temperature  310
set outputname A045_1AJV_eq
firsttimestep  0
paraTypeCharmm on
parameters     par_all27_prot_na.prm
mergeCrossterms yes
parameters     par_all36_A045_cgenff.prm
temperature $temperature
exclude        scaled1-4
1-4scaling    1.0
GBIS           on
ionConcentration 0.15
GBISDelta      0.8
GBISBeta       0.0
GBISGamma      2.90912
alphaCutoff    14
switching      on
Switchdist     15
cutoff         16
pairlistdist   17.5
timestep       2
rigidBonds     all
nonbondedFreq  2
fullElectFrequency 4
langevin on ;
langevinDamping 5 ;
langevinTemp $temperature
langevinHydrogen off ;
outputName $outputname
restartfreq    1000 ;
dcdfreq        500
outputEnergies 500
outputPressure 500
minimize       2000
reinitvels $temperature
run            100000 ;

Figure 3.4 Configuration file of MD simulation for A045-bound HIV-1 protein complex
3.6 The Role of Molecular Dynamic Simulation in this Study

In this study, thirty-nine 200 ps MD trajectories were created using NAMD for PIE calculation, i.e. one for each model of a ligand-bound HIV-1 complex. Figure 3.5 depicts the 3D molecular model of the A045-bound HIV-1 complex. The 200 ps equilibrium trajectory is shown in Figure 3.6. Since the rmsd of backbone atoms of the complex levels off at 100 ps, the last 50 ps (from 150 ps to 200 ps) were used for the PIE calculation. The procedure of the PIE calculation for a ligand-residue pair is as follows:

1. Load the dcd trajectory file and its corresponding psf file in VMD
2. Use the NAMDEnergy function in the Analysis program
3. Specify the resname of ligand and residue
4. Specify the dielectric constant of the residue type
5. Start the calculation by clicking the “Run NAMDEnergy” function
Figure 3.5 Molecular model of A045-bound HIV-1 complex. HIV-1 protein is depicted in gray NewCartoon. The colorful molecule at the center is ligand A045.

Figure 3.6 Equilibrium of A045-bound HIV-1 complex model. X-coordinate is the time frame of trajectory in ps. Y-coordinate is the rmsd (Å) of the backbone atoms of the model.
3.7 Clustering

DS-PIE, DS-RMLR, and DS-RMRR in the format of 39 rows (records/ligands) and 44 columns (attributes/residues) were normalized to build N-DS-PIE, N-DS-RMLR and N-DS-RMRR. K-Means cluster algorithm with various k values (k = 4, 6 and 8) was used to cluster the three normalized principal datasets: N-DS-PIE, N-DS-RMLR and N-DS-RMRR. The program was iterated for 100 times and the results from the experiment with the lowest inertia value were selected. The value of inertia is computed as follows:

\[ \sum_{i=0}^{N} \min_{U_j \in C} (\| X_i - U_j \|^2) \]

where

C represents the k disjoint clusters, N is the number of X samples in the set, and U_j is the mean of the samples in the cluster.

As shown in Figure 3.7, the inflection point of the N-DS-RMRR curve is at k = 6. Thus, the results from the clustering experiment with k = 6 were adopted in this study.

![Figure 3.7 inertia value versus k-value](image)
The ligand names and their corresponding identities, as well as the residue names and their corresponding identities are shown in Table 3.1.

<table>
<thead>
<tr>
<th>Ligand Name</th>
<th>Ligand I.D.</th>
<th>Residue Name</th>
<th>Residue I.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>U75875</td>
<td>L1</td>
<td>ChainA-R8</td>
<td>S1</td>
</tr>
<tr>
<td>Saquinavir</td>
<td>L2</td>
<td>ChainA-L10</td>
<td>S2</td>
</tr>
<tr>
<td>Ritonavir</td>
<td>L3</td>
<td>ChainA-L23</td>
<td>S3</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>L4</td>
<td>ChainA-D25</td>
<td>S4</td>
</tr>
<tr>
<td>Indinavir</td>
<td>L5</td>
<td>ChainA-G27</td>
<td>S5</td>
</tr>
<tr>
<td>DMP</td>
<td>L6</td>
<td>ChainA-A28</td>
<td>S6</td>
</tr>
<tr>
<td>BEA409</td>
<td>L7</td>
<td>ChainA-D29</td>
<td>S7</td>
</tr>
<tr>
<td>B440</td>
<td>L8</td>
<td>ChainA-D30</td>
<td>S8</td>
</tr>
<tr>
<td>B439</td>
<td>L9</td>
<td>ChainA-V32</td>
<td>S9</td>
</tr>
<tr>
<td>B435</td>
<td>L10</td>
<td>ChainA-K45</td>
<td>S10</td>
</tr>
<tr>
<td>B429</td>
<td>L11</td>
<td>ChainA-I47</td>
<td>S11</td>
</tr>
<tr>
<td>B425</td>
<td>L12</td>
<td>ChainA-G48</td>
<td>S12</td>
</tr>
<tr>
<td>B412</td>
<td>L13</td>
<td>ChainA-G49</td>
<td>S13</td>
</tr>
<tr>
<td>408</td>
<td>L14</td>
<td>ChainA-I50</td>
<td>S14</td>
</tr>
<tr>
<td>B388</td>
<td>L15</td>
<td>ChainA-G52</td>
<td>S15</td>
</tr>
<tr>
<td>B376</td>
<td>L16</td>
<td>ChainA-F53</td>
<td>S16</td>
</tr>
<tr>
<td>B369</td>
<td>L17</td>
<td>ChainA-I54</td>
<td>S17</td>
</tr>
<tr>
<td>B365</td>
<td>L18</td>
<td>ChainA-I76</td>
<td>S18</td>
</tr>
<tr>
<td>B355</td>
<td>L19</td>
<td>ChainA-T80</td>
<td>S19</td>
</tr>
<tr>
<td>B347</td>
<td>L20</td>
<td>ChainA-P81</td>
<td>S20</td>
</tr>
<tr>
<td>B322</td>
<td>L21</td>
<td>ChainA-V82</td>
<td>S21</td>
</tr>
<tr>
<td>B295</td>
<td>L22</td>
<td>ChainA-I84</td>
<td>S22</td>
</tr>
<tr>
<td>B277</td>
<td>L23</td>
<td>ChainB-R8</td>
<td>S23</td>
</tr>
<tr>
<td>B268</td>
<td>L24</td>
<td>ChainB-L10</td>
<td>S24</td>
</tr>
<tr>
<td>B249</td>
<td>L25</td>
<td>ChainB-L23</td>
<td>S25</td>
</tr>
<tr>
<td>AMP</td>
<td>L26</td>
<td>ChainB-D25</td>
<td>S26</td>
</tr>
<tr>
<td>A047</td>
<td>L27</td>
<td>ChainB-G27</td>
<td>S27</td>
</tr>
<tr>
<td>A045</td>
<td>L28</td>
<td>ChainB-A28</td>
<td>S28</td>
</tr>
<tr>
<td>A038</td>
<td>L29</td>
<td>ChainB-D29</td>
<td>S29</td>
</tr>
<tr>
<td>A037</td>
<td>L30</td>
<td>ChainB-D30</td>
<td>S30</td>
</tr>
<tr>
<td>A030</td>
<td>L31</td>
<td>ChainB-V32</td>
<td>S31</td>
</tr>
<tr>
<td>A024</td>
<td>L32</td>
<td>ChainB-K45</td>
<td>S32</td>
</tr>
<tr>
<td>A023</td>
<td>L33</td>
<td>ChainB-I47</td>
<td>S33</td>
</tr>
<tr>
<td>A021</td>
<td>L34</td>
<td>ChainB-G48</td>
<td>S34</td>
</tr>
<tr>
<td>A018</td>
<td>L35</td>
<td>ChainB-G49</td>
<td>S35</td>
</tr>
<tr>
<td>A017</td>
<td>L36</td>
<td>ChainB-I50</td>
<td>S36</td>
</tr>
<tr>
<td>A016</td>
<td>L37</td>
<td>ChainB-G52</td>
<td>S37</td>
</tr>
<tr>
<td>A015</td>
<td>L38</td>
<td>ChainB-F53</td>
<td>S38</td>
</tr>
<tr>
<td>A008</td>
<td>L39</td>
<td>ChainB-I54</td>
<td>S39</td>
</tr>
</tbody>
</table>

Table 3.1 Ligand and residue names and their corresponding identities
3.7.1 N-DS-PIE Clustering

DS-PIE in the format of 39 rows (records/ligands) and 44 columns
(attributes/residues) was normalized to build N-DS-PIE and then clustered by using K-Means cluster algorithm with k = 6.

Table 3.2 shows the data analysis of the inertia results. The clustering result of minimum inertia value = 961.11 was adopted.

<table>
<thead>
<tr>
<th></th>
<th>inertia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>975.30</td>
</tr>
<tr>
<td>minimum</td>
<td>961.11</td>
</tr>
<tr>
<td>maximum</td>
<td>997.54</td>
</tr>
<tr>
<td>standard deviation</td>
<td>6.44</td>
</tr>
</tbody>
</table>

Table 3.2 Data analysis of the inertia results of N-DS-PIE clustering (k = 6) based on the ligand identity

As shown in Figure 3.8 A-F, there are 11, 9, 13, 2, 1, and 3 samples in cluster-0, cluster-1, cluster-2, cluster-3, cluster-4, and cluster-5 respectively (Table 3.3).

<table>
<thead>
<tr>
<th>cluster 0</th>
<th>cluster 1</th>
<th>cluster 2</th>
<th>cluster 3</th>
<th>cluster 4</th>
<th>cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>U75875</td>
<td>Ritonavir</td>
<td>Saquinavir</td>
<td>A018</td>
<td>Indinavir</td>
<td>B388</td>
</tr>
<tr>
<td>Nelfinavir</td>
<td>B440</td>
<td>DMP</td>
<td>A015</td>
<td></td>
<td>B277</td>
</tr>
<tr>
<td>BEA409</td>
<td>B435</td>
<td>B429</td>
<td></td>
<td></td>
<td>A008</td>
</tr>
<tr>
<td>B439</td>
<td>B376</td>
<td>B412</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B425</td>
<td>B369</td>
<td>B408</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B365</td>
<td>B355</td>
<td>B347</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A045</td>
<td>B295</td>
<td>B322</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A037</td>
<td>B268</td>
<td>AMP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A030</td>
<td>B249</td>
<td>A047</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A023</td>
<td></td>
<td></td>
<td></td>
<td>A038</td>
<td></td>
</tr>
<tr>
<td>A017</td>
<td></td>
<td></td>
<td>A024</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A021</td>
<td></td>
<td>A016</td>
</tr>
</tbody>
</table>

Table 3.3 Results of the N-DS-PIE clustering based on the ligand identity
Figure 3.8 A & B Results of the N-DS-PIE clustering based on the ligand identity
Figure 3.8 C & D Results of the N-DS-PIE clustering based on the ligand identity
Figure 3.8 A to F Results of the N-DS-PIE clustering based on the ligand identity. The x-axis, y-axis, and legend represent the residue identity, the normalized pairwise interaction energy kcal/mol, and the ligand identity respectively.
Normal Mode Analysis

Normal mode analysis is a powerful computational method for identifying and characterizing the slowest molecular deformational motions with large amplitude that are widely involved in biological functions of macromolecules, but inaccessible by other methods. In NMA, it is assumed that the lowest frequency modes are the ones that are functionally relevant. Recently, studies on NMA have revealed that functionally important transition pathways of bio-molecules often follow the trajectories of one or a few low-frequency normal modes. Examples include the study of the collective motions in HIV-1 reverse transcriptase conducted by Bahar et al. in 1999 [45]; the analysis of the functional motions of influenza virus hemagglutinin carried out by Isin et al. in 2002 [46] as well as the experiment of the study for ligand docking operated by Cavasotto et al. in 2005 [47].

A standard NMA requires a set of coordinates, a force field describing the interactions between constituent atoms, and software to perform the required calculations. The performance of NMA in Cartesian coordinate space requires three main computational steps: (1) minimization of the conformational potential energy as a function of the atomic Cartesian coordinates; (2) the calculation of the Hessian matrix, which is the second derivatives of the potential energy with respect to the mass weighted atomic coordinates; and (3) the diagonalization of the Hessian matrix. This final step yields eigenvalues and eigenvectors.

In the model of standard normal mode analysis, the potential energy is written in terms of bonded and non-bonded energy terms. In 1996, Tirion [48] proposed a
simplified model referred to as the Elastic Network Model (ENM). In ENM, the interaction between two atoms is described by their Hookean pairwise potential (Figure 4.1), where the distances are taken to be at the minimum, avoiding minimization. In 1997, Haliloglu et al. [49] proposed a Gaussian Network Model (GNM) by applying a coarse-grained (Cα) protein representation in ENM. GNM was later extended to a 3D vectorial Anisotropic Network Model (ANM) by Atilgan et al. in 2001 [50]. Through the diagonalization of the Hessian matrix, ANM provides eigenvalues and eigenvectors that describe the frequencies and shapes of the normal modes, and their directions.

The potential energy function $E_p(r)$ of a protein molecule around a minimum $r^0$ (atomic Cartesian coordinates) can be expanded by Taylor polynomials:

$$E_p(r) = E_p(r^0) + \sum_i \frac{\partial E_p}{\partial r_i} |_{r=r^0} (r_i - r_i^0) + \frac{1}{2!} \sum_{ij} \frac{\partial^2 E_p}{\partial r_i \partial r_j} |_{r=r^0} (r_i - r_i^0)(r_j - r_j^0) + \ldots$$

Since the $r^0$ is a minimum of the energy function,

$$\frac{\partial E_p}{\partial r_i} (r^0) = 0$$

In addition, without loss of generality, the potential energy can be defined in connection with this structure as

$$E_p(r^0) = 0$$

In considering sufficiently small displacements, terms greater than the second order may be neglected. Consequently, the approximate potential energy function is given by the harmonic approximation as follows:

$$E_p(r) \approx \frac{1}{2!} \sum_{ij} \frac{\partial^2 E_p}{\partial r_i \partial r_j} |_{r=r^0} (r_i - r_i^0)(r_j - r_j^0)$$

This equation shows that the protein is fluctuating as if it is governed by a harmonic potential energy around the equilibrium position. Hence, the equation can be solved analytically through the singular value decomposition of the Hessian matrix. If the
protein has n atoms, then the dimension of the Hessian matrix, H, is 3n and there will be 3n sets of eigenpair solutions. Among those solutions, only 3n-6 normal modes are meaningful because the first six smallest normal modes have an eigenvalue equal to 0. They correspond to 3 translational and 3 rotational motions of the whole system.

![Figure 4.1 Hookean pairwise potential.](image)

4.1 Standard Normal Mode Analysis

In standard normal mode analysis, the potential energy has the form

\[
E_p = \frac{1}{2} \sum_{\text{bonds}} K_b (b-b_0)^2 + \frac{1}{2} \sum_{\text{angles}} K_\theta (\theta-\theta_0)^2 + \frac{1}{2} \sum_{\text{dihedrals}} K_\phi (1+\cos(n\Phi-\delta)) + \sum_{\text{nonbondedpairs}} [\frac{A}{r^{12}} - \frac{B}{r^6} + q_1 q_2 / Dr]
\]

The first three terms describe the energy cost in the distortion of bond lengths, bond angles, and dihedral angles, and the last term represents steric repulsions, van der Waals attractions, and electrostatic interactions between non-bonded atoms. The various bonded constants, \(K_b, b_0, K_\theta, \) etc., are specific for each type of covalent interaction. The non-bonded constants, A and B are specific for every type of interacting atom pairs. There are two disadvantages of this potential definition. First, the necessary initial energy minimization requires a great deal of computer time and memory, and is practically
impossible for large proteins with a reasonable degree of accuracy. This inevitably leads to unstable modes and casts doubt on the validity of the analysis. Second, because the minimization process is carried out in vacuo, the final configuration disagrees with the known crystallographic structure, complicating the interpretation of the NMA results.

4.2 Elastic Network Model

In 1996, Tirion proposed a simplified model with the potential energy defined as follows:

\[ E(r_{a,b}) = \frac{C}{2}(|r_{a,b} - r_{0_{a,b}}|^2) \]

Here, \( r_{a,b} = r_a - r_b \) denotes the vector connecting atoms \( a \) and \( b \), and the zero superscript indicates the given initial configuration where the distances are taken to be at minimum, avoiding the minimization process. Expanding to second order about \( r_{0_{a,b}} \) yields

\[ E(r_{a,b}) = \frac{C}{2}(r_{0_{a,b}} \cdot \Delta r_{a,b}/|r_{0_{a,b}}|^2) \]

where \( \Delta r = r - r_{0} \)

The potential energy within a molecule is then given by

\[ E_p = \sum_{(a,b)} E(r_{a,b}) \]

The sum is restricted to atom pairs separated by less than \( R_{vdW}(a) + R_{vdW}(b) + R_c \), where \( R_{vdW} \) refers to the van der Waals radii, and \( R_c \) is an arbitrary cutoff parameter. \( R_c \) determinates the total number of interacting atom pairs contributing to the potential energy of the system. The strength of potential \( C \) is a phenomenological constant, assumed to be the same for all interaction pairs. This is because slow vibrational modes involve coherent motion of large groups of atoms. The combined effect of these interactions approaches a universal form, governed by the central limit theorem, regardless of the details of individual pair-potentials.
4.3 Gaussian Network Model

The idea of the Elastic Network Model was further extended to use coarse-grained (Cα) protein representation by Haliloglu et al. in 1997. In GNM, the position of the nodes of ENM are defined by the Cα-atom coordinates, and the springs connecting the nodes are representative of the bonded and non-bonded interactions between the pairs of residues located within an interaction range, or cutoff distance, of \( r_c \) (Figure 4.2).

![Figure 4.2 Description of GNM. Schematic representation of the equilibrium positions of the \( i^{th} \) and \( j^{th} \) nodes, \( R_{0i} \) and \( R_{0j} \). \( R_{0ij} \) is the equilibrium distance between nodes \( i \) and \( j \). The instantaneous fluctuation vectors, \( \Delta R_i \) and \( \Delta R_j \), are shown by the dashed arrows, along with the instantaneous separation vector \( R_{ij} \) between the positions of the two residues.](image)

The potential of GNM is defined as

\[
V_{\text{GNM}} = \frac{\gamma}{2} \left[ \sum_{i,j} (R_{ij} - R_{0ij})^2 D(r_c - d_{ij}) \right]
\]

where

\[
D(r_c - d_{ij}) = 1, \text{ if the argument is positive and zero otherwise. Hence, } D(r_c - d_{ij}) \text{ selects all residue pairs within the cutoff separation of } r_c. \gamma \text{ is the force constant taken to be uniform for all network springs. } R_{ij} \text{ and } R_{0ij} \text{ are vectors represent the instantaneous and}
\]
The major disadvantage of GNM is that it cannot provide information regarding the individual components: \( \Delta X_i, \Delta Y_i, \) and \( \Delta Z_i \), of the deformation vectors \( \Delta R_i \) associated with normal mode, \( k \). It only provides information regarding the magnitude, \( |\Delta R_i| \), induced by such deformations.

### 4.4 Anisotropic Network Model

In 2001, Atilgan et al. extended the GNM model to a 3D vectorial Anisotropic Network Model (ANM). Through diagonalization of the Hessian matrix, ANM provides eigenvalues and eigenvectors that not only describe the frequencies and shapes of the normal modes, but also their directions. The potential of ANM is defined as

\[
V_{\text{ANM}} = \gamma/2 \sum_{i,j} (R_{ij} - R_{ij}^0)^2 D(r_c - R_{ij})
\]

where \( D(r_c - R_{ij}) \) equals to 1, if the argument is positive and zero otherwise. Hence, \( D(r_c - R_{ij}) \) selects all residue pairs within the cutoff separation of \( r_c \). \( R_{ij} \) and \( R_{ij}^0 \) are scalar values of the instantaneous and equilibrium distance between nodes \( i \) and \( j \), respectively.

Compared to GNM, ANM has been observed to produce excessively high fluctuations, and hence necessitates the use of a higher cutoff distance for interactions. With a higher cutoff distance, each residue is connected to more neighbors in a more constrained and consolidated network.

### 4.5 Normal Mode Analysis in Internal Coordinate

The major advantage of operating NMA in internal coordinate (IC) are twofold. First, compared to NMA using Cartesian coordinates as variables, performing NMA in IC requires at least one-third fewer degrees of freedom and hence reduces both
computational time and memory usage [29]. Second, it is possible to prevent the tip effect that makes many eigenvectors of some low-frequency modes irrational. For example, in 2006, Lu et al. [30] reported a method for coarse-grained elastic normal-mode analysis using IC to overcome the tip effect problem (Figure 4.3).

Figure 4.3 Motional patterns for the fourth mode of lysozyme (PDB code: 3lzt). (a) From the conventional elastic NMA, the lower-right portion has abnormal motions. (b) From the method proposed by Lu et al., the motions for lower-right portion are much more realistic.

4.6 iMod

In this study, iMod was adopted to conduct the NMA analysis. iMod, developed by Pablo Chacón et al. [29], is a multipurpose tool chest to explore the conformational flexibility of both protein and nucleic acid structures using NMA in IC. IC is defined by the backbone dihedral angles $\Phi$ and $\Psi$ in proteins (Figure 4.4). To avoid ring closure problems, $\Phi$ is fixed for proline. The first $\Phi$ angles and the last $\Psi$ of the chains are also
not considered. The remaining dihedral angles and all covalent bond lengths and angles are assumed to be fixed, thereby preserving the underlying covalent structure.

**Figure 4.4 Internal coordinate system of protein.** \( \Phi \) (phi) involves the backbone atoms \( C'_{i-1} - N_i - C\alpha_i - C'_{i} \), and \( \Psi \) (psi) involves the backbone atoms \( N_i - C\alpha_i - C'_{i-1} - N_{i+1} \). Thus, \( \Phi \) controls \( C'_{i-1} - C'_{i} \) distance, and \( \Psi \) controls the \( N_i - N_{i+1} \) distance.

The potential energy is formulated as follows:

\[
V = \sum_{i<j} F_{ij} (r_{ij} - r_{ij}^0)^2 + s \sum_\alpha (\theta_\alpha - \theta_\alpha^0)^2 ,
\]

where

- \( r_{ij} \) is the distance between atom i and j, and the super-index 0 indicates the initial equilibrium conformation.
- \( F_{ij} \) is the matrix whose elements describe the force constant associated with each atom pair. \( F_{ij} = k / (1 + (r_{ij}^0 / r_o)^p) \) if \( r_{ij}^0 < r_{\text{cut}} \), otherwise \( F_{ij} = 0 \) and \( k, r_o, p \) and \( r_{\text{cut}} \) were set to 1, 3.8Å, 6 and 10Å respectively.
- The second term of the energy equation is added for tip effect prevention. \( \theta_\alpha \) is the dihedral angle, and s is related to each dihedral angle.

Note that in IC, all modes are internal. This means that they do not contain external components of motion. The three rotational and the three translational motions of a rigid
body are eliminated from Cartesian coordinate NMA using Eckart conditions of linear and angular momentum conservations.

### 4.7 Conformational Change Covered by NMA

In order to quantify how well a conformational change is described by normal mode \( j \), \( I_j \) is defined as follows [51]:

\[
I_j = \Delta x \cdot y_j = \sum \Delta x_k y_{kj} / (\sum \Delta x_k^2)^{1/2},
\]

where

\[
\Delta x = \{\Delta x_1, ..., \Delta x_k, ..., \Delta x_{3N}\}
\]

is the conformational change observed by crystallographers; \( y_j = \{y_{1j}, ..., y_{kj}, ..., y_{3Nj}\} \) is the \( j \)th normal mode of the protein; \( \Delta x_k = x_k^o - x_k^c \), \( x_k^o \) and \( x_k^c \) are, respectively, the \( k \)th atomic coordinate of the protein in the open crystallographic structure and in the closed one. Since \( y_j \) is normalized, a \( I_j \) value of \( \pm 1 \) for the overlap means that the direction given by \( y_j \) is identical or opposite to \( \Delta x \). Note that \( \Theta_m \), the cumulative overlap, calculated as:

\[
\Theta_m = 100 \sum_{j=1}^{n} I_j^2
\]

\( \Theta_m \) is equal to 100% when \( n = 3N \); that is, when all modes are taken into account since the 3N modes form a complete basis set. According to recent studies, the first ten lowest frequency modes cover nearly 90% of protein conformational change [51]. Figure 4.5 shows the conformational change covered by each individual mode as well as the cumulative overlap (\( \Theta_m \)) of the first twenty modes of maltose-binding protein (MBP) (panel a) and intracellular nucleotide-binding domains (NBD) (panel b). In the case of MBP, the \( \Theta_m \)s of the first two and twenty low-frequency normal modes are 0.96 and 0.97 respectively. For NBD, the \( \Theta_m \)s of the first four and twenty low-frequency normal modes are 0.96 and 0.98 respectively [52].
Figure 4.5 Cumulative overlap. 
(a and b) Open and closed crystal structures of MBP and NBD (top panels), and individual projections (bars) as well as cumulative overlaps (red dotted line) of the 20 low-frequency normal modes of the conformational changes of MBP and NBD (bottom panels).

4.8 Utilities and Limitations

Despite the enormous success of NMA applications to the study of large-amplitude molecular deformational motions that are widely involved in biological functions of macromolecules, there are still concerns regarding the validity of NMA. This section addresses the following questions: Can the harmonic NMA modes be used to model activated non-equilibrium conformational transitions triggered by ligand binding? How do we address the timescale of protein motion in solvent by using normal modes calculated in vacuum? What is the usefulness of coarse-grained NMA [53]?
4.8.1 Harmonic Deformation

From numerous studies of NMA, functionally important deformation of biomolecules usually has large-amplitude, low-frequency motions that are highly anharmonic because of the rugged energy landscape. In other words, there is a disagreement between the amplitude of thermal fluctuations of harmonic oscillators described by normal modes and the amplitude of bio-molecular deformation that is experimentally observed.

At thermal equilibrium, $A^2 = k_B T/k$, where $A$ is the average amplitude of thermal fluctuation of any harmonic oscillator; $k_B$ is the Boltzmann constant; $T$ is the temperature; and $k$ is the force constant for the harmonic oscillator. For a low-frequency normal mode calculated based on the CHARMM force field, $k_B T/k$ is less than $1\text{Å}^2$ at 300K. However, in reality, a conformational change induced by ligand binding can have an amplitude much larger than the equilibrium thermal fluctuation. The explanation of this discrepancy is that a conformational transition induced by ligand binding is usually an activated process and the binding of the ligand brings in additional energy. Figure 4.6 schematically illustrates the changes in the energy landscape during ligand binding to a protein. Figure 4.6A shows the potential energy of a protein before ligand binding. The protein oscillates around the ligand-free equilibrium state. Figure 4.6B illustrates that the energy barrier along the binding is significantly decreased because the energy landscape is gradually tipped over toward the final bound state by going through the continuous energetically favorable engagement of the ligand with the protein. The actual transition takes place upon ligand binding, along with the gain of binding energy that shifts the equilibrium toward the ligand-bound state (Figure 4.6C).
Figure 4.6 Changes of energy landscape upon ligand binding. The vertical dashed line indicates the ligand-bound conformation for the protein. The horizontal axis gives the progression of the reaction coordinate. (A) the ligand-free structure. (B) the conventional view of an energy barrier during ligand binding. (C) the energy landscape tilting toward the final ligand-bound state.

4.8.2 Timescales of Harmonic Motions

Since the presence of solvent damping dramatically slows down the large-amplitude motions of bio-molecules, the timescales of molecular motions in reality are much longer than what can be estimated from the eigenvalues ($\lambda = \omega^2$) of NMA that are calculated in vacuum. In other words, solvent damping causes a discrepancy on a timescale between NMA and real molecular motions.

The study conducted by Ma et al. in 2005 revealed that the presence of solvent has a much larger impact on eigenvalues than on eigenvectors, which are determined by the potential surface only. In fact, the information provided by the eigenvectors for the
directionality of conformational transitions has wider applications than the information provided by eigenvalues.

**4.8.3 Usefulness of Coarse-grained NMA**

Ligand binding and unbinding events are often on a long-time scale ranging from milliseconds to days, far beyond the current capability of MD simulations. Coarse-grained NMA has the advantage of allowing a scientist to extract important dynamic information with a much extended capacity so as to make it possible to study systems with sizes completely beyond conventional NMA. This is particularly useful when only the low-frequency modes are important because low-frequency motions are typically delocalized throughout the system and involve mainly collective movements of residues. The study performed by Durand et al. in 1994 showed that method of rigid-body motions of blocks (RTB method) yields accurate approximations for the low-frequency normal modes of proteins of various sizes and various folds. In their study, for a protein of n residues and N atoms, the RTB method requires the diagonalization of an n x n matrix, whereas standard procedure require the diagonalization of a 3N x 3N matrix [54].

**4.9 The Role of Normal Mode Analysis in this Study**

Two training datasets including DS-RMLR and DS-RMRR were built using normal mode analysis technique. Both DS-RMLR and DS-RMRR cover the ten lowest frequency modes and comprise forty-four attributes selected by the SASA program with probe radius of 1.4/2.1Å. The training attributes of DS-RMLR illustrate the relative movement of a ligand-residue pair and the training attributes of DS-RMRR illustrate relative movement of a residue upon ligand binding. Both DS-RMLR and DS-RMRR
comprise thirty-nine records, and each record comprises 44 attributes. In total, there are 1716 attribute values in both datasets.

In section 2.3.2, NTAV was defined as:

$$(DPVV_1^2 + ... + DPVV_j^2 + ... + DPVV_{10}^2)^{1/2},$$

where $j = 1$ to 10 is the normal mode index. In DS-RMLR, DPVV is the dot product of ligand displacement vector after normalization and residue displacement vector. In DS-RMRR, DPVV is the dot product of two displacement vectors of a residue upon ligand binding.

4.9.1 iMOD

Two programs of iMOD system [29], imode and imodview, were used to produce the eigenvectors of the ligand-HIV-1 protease at residue level. First, imode was used to produce Cartesian normal modes and the results were written to the file with .evec extension. Second, the .evec output file was used as an input file for imodview to compute the 3D vector sets of residues and ligands for the ten lowest frequency modes. The commands used for the generation of the 3D vector sets of DMP bound HIV-1 protease are listed as follows:

1. `imode: imode DMP-1QBS.pdb –o DMP-1QBS-imode –save_cart`

DMP-1QBS.pdb is the structure of the DMP bound HIV-1 protease in PDB format. Two options were specified in the command. –o defines the basename of the output files and –save_cart is used to generate normal modes in Cartesian coordinates. The default values of all other basic options were adopted. This command produced four output files including DMP-1QBS-imode.log, DMP-1QBS-imode_model.pdb, DMP-1QBS-imode_ic.evec and DMP-1QBS-imode_cart.evec.
Three options were specified in the command. –n identifies the normal mode index. –color sets the color of the eigenvector arrows displayed on the VMD screen. –level sets the averaging level for the computation of the eigenvector arrow (1 = residue). The output file DMP-1QBS_n1.vmd contains the Cartesian coordinates of the eigenvectors of 1st normal mode. The other nine .vmd output files for the normal modes ranging from 2 to 10 were created by repeating the imodview command with –n set to the corresponding mode index value. For each of the first ten normal modes, forty .vmd files were created. They are the thirty-nine files for each of the thirty-nine ligand bound HIV-1 protease structures and the file for the 3IXO structure after alignment. In total, four hundred .vmd files were created for the first ten normal modes. The data in the .vmd files were used to compute the DPVV values.

Figure 4.7 depicts the superposition of the DMP bound HIV-1 protease structure (PDB code: 1QBS blue) and the ligand free HIV-1 protease structure (PDB code: 3IXO red) after aligning to 1QBS. The aligned 3IXO structure was used to compute the eigenvector sets of the ligand free HIV-1 protease structure. Figure 4.8 depicts the superposition of the 44 residues of the aligned 3IXO and the DMP bound 1QBS of 1st normal mode. It illustrates the shift of the eigenvectors of the 44 residues of HIV-1 protease upon ligand binding. Figure 4.9 illustrates the relative displacements the 44 ligand-residue pairs in the DMP bound HIV-1 protease.
Figure 4.7 Superposition of ligand free HIV-1 protease (red) and DMP (green) bound HIV-1 protease (blue)

Figure 4.8 Superposition of the 44 residues of the aligned 3IXO (red) and the DMP bound 1QBS (blue) of 1st normal mode
4.10 Clustering

K-Means cluster algorithm with $k = 6$ was used to cluster normalized DS-RMLR based on ligand identity, and normalized DS-RMRR based on both the ligand identity and residue identity respectively.

4.10.1 N-DS-RMLR

DS-RMLR in the format of 39 rows (records/ligands) and 44 columns (attributes/residues) was normalized to build normalized DS-RMLR (N-DS-RMLR). N-DS-RMLR was clustered based on ligand identity by using K-Means algorithm with $k = 6$. The program was iterated for 100 times. Table 4.1 shows the data analysis of the inertia values. The results from the clustering experiment with the minimum inertia value
= 696.50 were adopted. As shown in Figure 4.10 A-F, there are 10, 11, 10, 1, 5, and 2
samples in cluster-0, cluster-1, cluster-2, cluster -3, cluster-4 and cluster-5 respectively
(Table 4.2).

<table>
<thead>
<tr>
<th></th>
<th>inertia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>703.18</td>
</tr>
<tr>
<td>minimum</td>
<td>696.50</td>
</tr>
<tr>
<td>maximum</td>
<td>711.40</td>
</tr>
<tr>
<td>standard deviation</td>
<td>3.76</td>
</tr>
</tbody>
</table>

Table 4.1 Data analysis of the inertia results of N-DS-RMLR clustering based on the
ligand identity

<table>
<thead>
<tr>
<th>cluster 0</th>
<th>cluster 1</th>
<th>cluster 2</th>
<th>cluster 3</th>
<th>cluster 4</th>
<th>cluster 5</th>
<th>cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nelfinavir</td>
<td>B440</td>
<td>Saquinavir</td>
<td>A008</td>
<td>B277</td>
<td>U75875</td>
<td></td>
</tr>
<tr>
<td>DMP</td>
<td>B435</td>
<td>Indinavir</td>
<td>B268</td>
<td>Ritonavir</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B408</td>
<td>B425</td>
<td>BEA409</td>
<td>B249</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B388</td>
<td>B412</td>
<td>B439</td>
<td>A018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B376</td>
<td>B355</td>
<td>B429</td>
<td>A017</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B365</td>
<td>B295</td>
<td>B369</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B347</td>
<td>A047</td>
<td>AMP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B322</td>
<td>A045</td>
<td>A037</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A038</td>
<td>A030</td>
<td>A016</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A021</td>
<td>A024</td>
<td>A015</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A023</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2 Results of the N-DS-RMLR clustering based on the ligand identity
Figure 4-10 A & B Results of the N-DS-RMLR clustering based on the ligand identity
Figure 4-10 C & D Results of the N-DS-RMLR clustering based on the ligand identity
Figure 4-10 A to F Results of the N-DS-RMLR clustering based on the ligand identity. The x-axis, y-axis, and legend represent the residue identity, the normalized NTAV value ($\text{Å}^2$), and the ligand identity respectively.
4.10.2 N-DS-RMRR

4.10.2.1 Clustering Based on the Ligand Identity:

DS-RMRR in the format of 39 rows (records/ligands) and 44 columns (attributes/residues) was normalized to build N-DS-RMRR and then clustered by using K-Means algorithm with k =6. The program was iterated for 100 times. Table 4.3 shows the data analysis of the inertia results. The results from the experiment with the minimum inertia value = 348.91 were adopted.

<table>
<thead>
<tr>
<th></th>
<th>inertia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>350.64</td>
</tr>
<tr>
<td>minimum</td>
<td>348.91</td>
</tr>
<tr>
<td>maximum</td>
<td>353.07</td>
</tr>
<tr>
<td>standard deviation</td>
<td>1.41</td>
</tr>
</tbody>
</table>

Table 4.3 Data analysis of the inertia results of N-DS-RMRR clustering based on the ligand identity

As shown in Figure 4.11 A-F, there are 11, 2, 4, 3, 9, and 10 samples in cluster-0, cluster-1, cluster-2, cluster-3, cluster-4, and cluster-5 respectively (Table 4.4).
Figure 4-11 A & B Results of the N-DS-RMRR clustering based on the ligand identity
Figure 4-11 C & D Results of the N-DS-RMRR clustering based on the ligand identity
Figure 4-11 A to F Results of the N-DS-RMRR clustering based on the ligand identity. The x-axis, y-axis, and legend represent the residue identity, the normalized NTAV value, and the ligand identity respectively.
Table 4.4 Results of the N-DS-RMRR clustering based on the ligand identity

4.10.2.2 Clustering Based on the Residue Identity:

DS-RMRR in the format of 44 rows (attributes/residues) and 39 columns (records/ligands) was normalized to build N-DS-RMRR-R and then clustered based on residue identity by using K-Means algorithm with k = 6. The program was iterated for 100 times. Table 4.5 shows the data analysis of the inertia results. The results from the experiment with the lowest inertia value = 33.49 were adopted.

<table>
<thead>
<tr>
<th>inertia value</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
</tr>
<tr>
<td>minimum</td>
</tr>
<tr>
<td>maximum</td>
</tr>
<tr>
<td>standard deviation</td>
</tr>
</tbody>
</table>

Table 4.5 Data analysis of the inertia results of N-DS-RMRR-R clustering based on the residue identity

As shown in Figure 4.12 A-F, there are 6, 7, 12, 4, 5, and 10 samples in cluster-0, cluster-1, cluster-2, cluster-3, cluster-4 and cluster-5 respectively (Table 4.6).
Figure 4-12 A & B Results of the N-DS-RMRR-R clustering based on the residue identity
Figure 4-12 C & D Results of the N-DS-RMRR-R clustering based on the residue identity
Figure 4-12 A to F Results of the N-DS-RMRR-R clustering based on the residue identity. The x-axis, y-axis, and legend represent the ligand identity, the normalized NTAV value (Å²), and the residue identity respectively.
<table>
<thead>
<tr>
<th>cluster 0</th>
<th>cluster 1</th>
<th>cluster 2</th>
<th>cluster 3</th>
<th>cluster 4</th>
<th>cluster 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChainB-R8</td>
<td>ChainA-G49</td>
<td>ChainA-D30</td>
<td>ChainB-G52</td>
<td>ChainA-P81</td>
<td>ChainA-G27</td>
</tr>
<tr>
<td>ChainB-I47</td>
<td>ChainA-I50</td>
<td>ChainA-V32</td>
<td>ChainB-F53</td>
<td>ChainB-G48</td>
<td>ChainA-D29</td>
</tr>
<tr>
<td>ChainB-I54</td>
<td>ChainB-K45</td>
<td>ChainA-I76</td>
<td>ChainB-P81</td>
<td>ChainA-V82</td>
<td>ChainA-V82</td>
</tr>
<tr>
<td>ChainB-T80</td>
<td>ChainB-G49</td>
<td>ChainA-I84</td>
<td>ChainB-P81</td>
<td>ChainA-V82</td>
<td>ChainA-V82</td>
</tr>
<tr>
<td>ChainB-I50</td>
<td>ChainB-L23</td>
<td>ChainB-D25</td>
<td>ChainB-G27</td>
<td>ChainB-D29</td>
<td></td>
</tr>
<tr>
<td>ChainB-K45</td>
<td>ChainB-A28</td>
<td>ChainB-V32</td>
<td>ChainB-D29</td>
<td>ChainB-V82</td>
<td></td>
</tr>
<tr>
<td>ChainB-D30</td>
<td>ChainB-V32</td>
<td>ChainB-D29</td>
<td>ChainB-V82</td>
<td>ChainB-V82</td>
<td></td>
</tr>
<tr>
<td>ChainB-I76</td>
<td>ChainB-I76</td>
<td>ChainB-I76</td>
<td>ChainB-I76</td>
<td>ChainB-I76</td>
<td></td>
</tr>
<tr>
<td>ChainB-I84</td>
<td>ChainB-I84</td>
<td>ChainB-I84</td>
<td>ChainB-I84</td>
<td>ChainB-I84</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6 Results of the N-DS-RMRR-R clustering based on the residue identity

- **Relative displacement of individual residue upon ligand binding**

  Figure 4.13 depicts the average normalized NTAVs of the 39 ligand bound complexes at the ground state (E₂L) for each of the 44 residues. As shown in Table 4.7, upon ligand binding, G52, F53, I50, G49, K45, G48, P81, I54, T80, and I47 of both chains A and B conducted significant displacements with total normalized NTAVs ≥ 1.15 Å².
Figure 4.13 Average normalized NTAVs (Å²) of the 39 ligand bound complexes at the ground state (E₂L)

Table 4.7 Residues in the descending order of the average normalized NTAVs (Å²).
Chapter 5

Machine Learning

Machine learning (ML) techniques for the prediction of compounds with pharmacological activity, specific pharmacodynamics and ADMET (absorption, distribution, metabolism, excretion and toxicity) properties have been applied increasingly in drug discovery and evaluation. ML techniques have been explored recently for predicting inhibitors, antagonists, blockers, agonists, activators and substrates of proteins related to specific therapeutic targets. All these predictions, however, are based on static characteristics of ligands such as ADMET and/or static properties of the PLI model such as QSAR (quantitative structure-activity relationship) modeling, and are insufficient to achieve an accurate prediction on kinetic rate constants of PLI [8,9,55,56,57]. In the ligand-HIV-1 interaction, ligand dissociation follows the induced fit retrograde mechanism. Consequently, residence time depends on the kinetic equilibrium between the ground state and the excited state of the ligand-bound protein complex. Thus, PLI kinetic properties are the optimal feature attributes for ML prediction of the kinetic rate constants.

In this study, three principal training datasets, covering kinetic properties and thermal dynamic properties of PLI, were constructed for the prediction of kinetic rate constants of PLI ($\log_{10}k_{on}$ and $\log_{10}k_{off}$). Machine learning algorithms were used. They include one multi-target classification algorithm, one multi-target regression algorithm, three single-target regression algorithms, and one single-target instance-based learning algorithm.
Machine learning is a sub-field of artificial intelligence whose concern is the development, understanding and evaluation of algorithms and techniques to allow a computer to define. ML intertwines with other disciplines such as statistics, human psychology and brain modeling. Human psychology and neural models obtained from brain modeling help in understanding the workings of the human brain, and especially its learning process, which can be used in the formulation of ML algorithms. Since many ML algorithms use analysis of data for building models, statistics plays a major role in this field.

Generally, an ML algorithm needs a dataset, which constitutes the knowledge base, to build a model of the domain. The dataset is a collection of instances from the domain. Each instance consists of a set of attributes which describe the properties of that example from the domain. An attribute takes in a range of values based on its attribute type, which can be discrete or continuous. A discrete attribute comes from a finite or countably infinite set (i.e. integers) and can be either numeric or categorical. A continuous attribute comes from an infinite set and is numeric. Each instance consists of a set of input attributes and an output attribute. The input attributes are the information given to the learning algorithm and the output attribute contains the feedback of the activity on that information. The value of the output attribute is assumed to depend on the values of the input attributes. The attribute, along with the value assigned to it, defines a feature, which makes an instance a feature vector. The model built by an algorithm can be seen as a function that maps the input attributes in the instance to a value of the output attribute.
In supervised learning approach for ML, the goal is to learn a mapping from input x to output y, given a labeled set of input-output pairs \(D = \{(x_i, y_i)\}_{i=1}^N\). Here, D is called the training set, and N is the number of training examples. In the simplest setting, each training input \(x_i\) represents an \(H\)-dimensional feature vector, \(x_i = (x_{1i}, \ldots, x_{Hi}) \in \mathbb{R}^H\), and each dimension is a feature. The form of the output or response variable \(y_i\) can be a categorical variable from some finite set \(y_i \in \mathcal{Y} = \{1, \ldots, C\}\) or a real-valued scalar. For classification problems, a classifier \(t\) is a mapping \(t: \mathbb{R}^H \rightarrow \mathcal{Y}\), while for regression problems, \(y_i = s(x_i) + \epsilon_i\), where \(s\) is a regression function and \(\epsilon_i\) are the residuals or errors.

One way to formalize the problem is as function approximation. Assume that \(y = f(x)\) for some unknown function \(f\), and the goal of learning is to estimate the function \(f\) given a training set, and then to make predictions using \(y' = f'(x)\) (apostrophe is used to denote an estimate). The accuracy of a model can be estimated from the difference between the predicted and actual value of the target attribute in the test set.

### 5.1 Probabilistic Prediction

There are two broad categories of probability interpretation: frequentist interpretation and Bayesian interpretation. Frequency probabilities are associated with random physical systems such as rolling dice. In this view, probabilities represent long run frequencies of events. In the Bayesian view, probability is fundamentally related to information rather than repeated trials. One big advantage of the Bayesian interpretation is that it can be used to model our uncertainty about events that do not have long term frequencies. It can be assigned to any statement whatsoever, even when no random process is involved, as a way to represent the degree to which the statement is supported by the available evidence.
**Basic Rules of Probability**

Probability of an event A is denoted by \( p(A) \). It is measured on a scale between 0 and 1 inclusive. If A is impossible \( p(A) = 0 \), if A is certain then \( p(A) = 1 \).

- **Probability of a Union of Two Events:**

  Given two events, A and B, the probability of A or B is defined as follows:

  \[
p(A \cup B) = p(A) + p(B) - p(A \cap B) \quad \text{or} \quad = p(A) + p(B) \quad \text{if A and B are mutually exclusive.}
\]

2. **Joint Probabilities:**

   The probability of the joint event A and B is defined as follows:

   \[
p(A,B) = p(A \cap B) = p(A|B)p(B) \quad \text{(product rule)}
   \]

   where

   \[
p(B) = \sum_a p(B|A=a)p(A=a) \quad \text{(sum rule)}.
   \]

   The product rule can be applied multiple times to yield the chain rule of probability:

   \[
p(X_{1:D}) = p(X_1)p(X_2|X_1)p(X_3|X_2,X_1)\ldots p(X_D|X_{1:D-1}) \quad \text{where}
   \]

   \( 1:D \) denotes the set \( \{1,2,\ldots,D\} \).

3. **Conditional Probability:**

   Given that event B is true, the conditional probability of event A is defined as follows:

   \[
p(A|B) = p(A,B)/p(B) \quad \text{if } p(B) > 0.
   \]

4. **Bayes Rule:**

   Combining the definition of conditional probability with the product and sum rules yields Bayes rule:

   \[
p(X = x|Y = y) = p(X = x,Y = y)/p(Y = y)
   \]

   \[
   = p(X = x)p(Y = y|X = x)/\sum_{x'} p(X = x')p(Y = y|X = x')
   \]
Classification Learning

Given the input vector \( x \) and training set \( D \), the probability distribution over possible labels is denoted by \( p(y|x, D) \). This notation makes explicit that the probability is conditional on the test input \( x \), as well as the training set \( D \), by putting these terms on the right hand side of the conditioning bar \( | \). When choosing between different models, the notation is \( p(y|x,D,M) \), where \( M \) denotes the model. Given a probabilistic output, the best guess can be computed using

\[
y' = f'(x) = \text{argmax}_c p(y=c|x, D)
\]

This corresponds to the most probable class label, and is called the mode of the distribution \( p(y|x,D) \); it is also known as a MAP (maximum a posterior) estimate.

Regression Learning

Regression is just like classification except the response variable is continuous. Using linear regression model as an example, the response is a linear function of the inputs:

\[
y(x) = w^T x + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon, \quad \text{where}
\]

\( w^T x \) represents the inner or scalar product between the input vector \( x \) and the model’s weight vector \( w \), and \( \epsilon \) is the residual error between the linear prediction and the true response. Assume that \( \epsilon \) has a Gaussian distribution. \( \epsilon \) is \( \sim N(\mu,\sigma^2) \), where \( \mu \) is the mean and \( \sigma^2 \) is the variance. To make the connection between linear regression and Gaussian more explicit, the model can be rewritten in the following form:

\[
p(y|x, \theta) = N(y|\mu(x),\sigma^2(x))
\]
This makes it clear that the model is a conditional probability density. In the simplest case, \( \mu \) is a linear function of \( x \), so \( \mu = w^T x \), and the noise is fixed, \( \sigma^2(x) = \sigma^2 \). In this case, \( \theta = (w, \sigma^2) \) are the parameters of the model.

5.2 Multi-Target Learning

Multi-target learning (MTL) is defined as follows: Given a set of learning examples \( D \) of the form \( (x, y) \), where \( x = (x_1, x_2, \ldots, x_k) \) is a vector of \( k \) descriptive attributes and \( y = (y_1, y_2, \ldots, y_t) \) is a vector of \( t \) target attributes, learn a model that, given a new unlabeled example \( x \), can predict the values of all target attributes \( y \) simultaneously. When \( y_i \) is categorical, the problem is known as classification; and when \( y_i \) is real-valued, the problem is known as regression.

Kocev et al. [58] reported that there are three main advantages of MTL application compared with single-target learning (STL) application in target prediction.

1. An MTL model is usually much smaller than the total size of the individual models for all target attributes.
2. An MTL model analyzes dependencies among the different target attributes.
3. The predictive performance of a MTL model is similar or slightly better than the predictive performance of a STL model.

5.3 Neural Networks and Multi-Layer Perceptron

Generally, neural networks consist of layers of interconnected nodes, each node producing a non-linear function of its input. In the simplest networks, the output from one node is fed into another node in such a way as to propagate messages through layers of interconnecting nodes. More complex behavior may be modeled by networks in which
the final output nodes are connected with earlier nodes, and then the system has the characteristics of a highly nonlinear system with feedback.

In 1959, Rosenblatt introduced the perceptron structure. Perceptron is a single-layer network and designed for the study of the relationships between the organization of a nerve net, the organization of its environment, and the psychological performances of which it is capable. It calculates a linear combination of its input and outputs a 1 if the result is greater than some threshold and a -1 if it is not [59].

In 1969, Minsky and Papert showed that perceptrons [60] could not model the exclusive-or function, because its outputs are not linearly separable. Two classes of outputs are linearly separable if and only if they can be separated by a straight line in two dimensions. They proposed a two-layer perceptron structure to solve the problem (Figure 5.1) [61]. This structure is widely used today, although the Perceptron Learning Rule (Delta Rule) could not be generalized to find weights for this structure.

The delta rule is a learning rule for a network with a continuous activation function. It attempts to minimize the cumulative error over a data set as a function of the weights in the network:

\[ \text{Delta}(w_{ji}) = c(d_i - O_i)F'(net_i)x_j \]

where \( c \) is the learning rate, \( d_i \) and \( O_i \) are the desired and actual outputs for the \( i^{\text{th}} \) node, and \( F'(net) \) is the derivative of the activation function for the \( i^{\text{th}} \) node, and \( x_j \) is the \( j^{\text{th}} \) input to the \( i^{\text{th}} \) node. One popular activation function is the sigmoidal function, such as the logistic function:

\[ F(\text{net}) = \frac{1}{1 + e^{-1.4\text{net}}} \]

where
L is lambda, a parameter for squashing the function, and net is the output or sum of the weights.

In 1985, Rumelhart et al. proposed the Generalized Delta Rule [62] which defines a notion of back-propagation of error derivatives through the network, and enables a large class of models with hidden layers to be trained.

The structure of a two-layer perception is shown in Figure 5.1. The inputs form the input nodes of the network; the outputs are taken from the output nodes. The middle layer of nodes, visible to neither the inputs nor the outputs, is termed the hidden layer, and its size in not fixed. The operation of this network is specified by

\[
y_i^{(H)} = F^{(H)}\left(\sum_j w_{ij}^{(HI)} x_j \right) \\
y_i^{(O)} = F^{(T)}\left(\sum_j w_{ij}^{(TH)} y_j^{(H)} \right)
\]  

[5.3-1]

This specifies how input pattern vector \(x\) is mapped into output pattern vector \(y^{(O)}\), via the hidden pattern vector \(y^{(H)}\), in a manner parameterized by the two layers of weights \(w^{(HI)}\) and \(w^{(TH)}\). The univariate functions \(F(.)\) are typically each set to

\[F(x) = 1/(1 + e^{-x})\]

which varies smoothly from 0 at \(-\infty\) to 1 at \(\infty\), as a threshold function would do abruptly.

The multilayer perceptron can act as either a feedforward network or a recurrent network.

- **Feedforward Network:**

In a feedforward network, the output vector \(y\) is a function of the input vector \(x\) and some parameters \(w\). It could be written

\[y = F(x;w)\]

for some vector function \(F\) given in detail by [5.3-1] in the 2-layer case.
• Recurrent network:

It is also possible to define a recurrent network by feeding the outputs back to the inputs. The general form of a recurrent perceptron is

$$y_i(t + 1) = F(\sum_j w_{ij} y_j(t)),$$

which could be written as

$$y(t + 1) = F(y(t); w)$$

for a discrete-time model.

![Multi-layer Perceptron Structure](image)

**Figure 5.1 Multilayer Perceptron Structure**

**5.4 Decision Trees**

A decision tree is constructed with a recursive partitioning algorithm known as Top-Down Induction of Decision Trees (TDIDT) from a training set of records. TDIDT starts by selecting a test for the root node. Based on this test, the training set is partitioned into subsets (two subsets for a binary tree) according to the test outcomes. This procedure is recursively repeated to construct the subtrees. The partitioning process stops when a stopping criterion is satisfied. Each internal node represents a test on an attribute. Each
branch extending from a node represents one of the possible alternatives available at that point. The set of alternatives must be mutually exclusive and collectively exhaustive. Each terminal node has an associated terminal value. Each terminal value measures the result of a scenario: the sequence of decisions and events on a unique path leading from the initial decision node to a specific terminal node [63].

**Information Gain**

Information gain is a measure used to evaluate the effectiveness of an attribute in classifying the training data. It is the expected reduction in entropy caused by partitioning the examples according to this attribute. If the target attribute takes on c different values in a set of collection of examples $S$, then the entropy of $S$ relative to this c-wise classification is defined as

$$\text{Entropy}(S) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

where $P_i$ is the proportion of $S$ belonging to class $i$.

Specifically, the information gain, $\text{Gain}(S,A)$ of an attribute $A$, relative to $S$, is defined as

$$\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{\nu \in \text{Values}(A)} \left( \frac{|S_\nu|}{|S|} \right) \text{Entropy}(S_\nu)$$

where $\text{Values}(A)$ is the set of all possible values for attribute $A$, and $S_\nu$ is the subset of $S$ for which attribute $A$ has value $\nu$ (i.e., $S_\nu = \{ s \in S \mid A(s) = \nu \}$). Note the first term in the equation for Gain is just the entropy of the original collection $S$ and the second term is the expected value of the entropy after $S$ is partitioned using attribute $A$. The expected entropy described by this second term is simply the sum of the entropies of each subset $S_\nu$, weighted by the fraction of examples $|S_\nu|/|S|$ that belong to $S_\nu$. $\text{Gain}(S,A)$ is therefore the expected reduction in entropy caused by knowing the value of attribute $A$. 

95
**Pruning**

One of the questions that arises in a decision tree algorithm is the optimal size of the final tree. It is hard to tell when a tree algorithm should stop because it is impossible to tell if the addition of a single extra node will dramatically decrease error. A common strategy is to grow the tree until each node contains a small number of instances then use pruning to remove nodes that do not provide additional information. In 1989, Mingers conducted a study to examine five of the principal methods for pruning decision trees [64]. The five methods include error-complexity pruning, critical value pruning, minimum-error pruning, reduced-error pruning and pessimistic error pruning. As a result of the study, he drew the following conclusions:

1. Minimum-error pruning produced markedly different levels of pruning even on essentially the same set of data. It was the least accurate method.
2. Pessimistic pruning was the quickest and did not need a separate test data set, but it gave bad results on certain data sets and should be treated with caution.
3. Critical value, error-complexity, and reduced-error methods all performed well, producing consistently low error rates over all the data sets.

**Multi-Target Decision Trees**

Multi-target decision trees generalize regression trees to the prediction of several target attributes simultaneously. The leaves of a multi-target decision tree store a vector, instead of storing a single value. Each component of this vector is a prediction for one of the target attributes. It is either the majority vote among trees in classification or the mean value of the target attribute calculated over the records that are sorted into the leaf in regression. In the multi-target regression trees algorithm of CLUS, the heuristic function
used for the attribute tests at each internal node to guide the algorithm towards a smaller
tree with good predictive performance is the intra-cluster variance summed over the
subsets induced by the test. Intra-cluster variance is defined as
\[ N \cdot \sum_{t=1}^{T} \text{Var}[y_t] \]
with \( N \) representing the number of examples in the cluster, \( T \) the number of target
variables, and \( \text{Var}[y_t] \) the variance of target variable \( y_t \) in the cluster. Lower intra-subset
variance results in predictions that are more accurate.

5.5 Random Forest

A Random Forest (RF) is an ensemble of trees. An ensemble method constructs a
set of predictive models. It gives a prediction for a new data instance by combining the
predictions of its models for that instance. The predictions can either take the average of
the outputs of the models in a regression approach or the majority votes of the models in
a classification approach. RFs correct for decision trees’ habit of overfitting to their
training set.

Random Forest algorithm

The random forests algorithm [65] is as follows:

1. For a RF of \( T \) trees, draw \( T \) bootstrap samples from the original data. The
bootstrap algorithm is as follows: Given a training set of \( N \) records, sample, with
displacement, \( N \) training records from the training set.

2. For each of the bootstrap samples, grow an un-pruned classification or regression tree.
At each node, randomly sample \( M \) attributes from the \( P \) inputs. In this study, \( M \) was set to
\( P \).
3. Predict new data by aggregating the predictions of the T trees (i.e., majority votes for classification, average for regression).

**The Out-of-Bag Estimate of Error Rate**

In the forest building process, when bootstrap sample set is drawn by sampling with replacement for each tree, about 1/3 of the original instances are left out. This set of instances is called OOB (out-of bag) data. Each tree has its own OOB dataset, which is used to estimate the prediction error for each tree. Aggregate the OOB predictions and calculate the error rate. It is called the OOB estimate of error rate.

**Variable Importance**

The importance of an attribute can be estimated by looking at how much prediction error increases when OOB data for that attribute is permuted while all others are left unchanged. The most useful measures for the score of importance of a given attribute are mean square error (MSE) for regression and misclassification rate for classification. The procedure of the measure of variable importance is as follows:

1. Fit a random forest classifier to a dataset, \( D_N = \{(X_i, Y_i)\}_{i=1}^{10N} \), where \( N \) is the number of records in the dataset. During the fitting process the OOB error for each data point is recorded and averaged over the forest.

2. To measure the importance of the \( j^{th} \) feature after training, the values of the \( j^{th} \) feature are permuted among the training data and the OOB error is again computed on this perturbed dataset. The importance score for the \( j^{th} \) feature is computed by averaging the difference in OOB error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

The larger the score value, the more important the feature is [66].
5.5.1 Multi-Target Random Forest Classification Algorithm of Clus

Eight Random Forest models for binary-class prediction were built with the number of iteration set at 1, 10, 25, 50, 100, 200, 250, and 500. Default values were chosen for all other parameters except FTest, which was set at 0.1. FTest sets the stopping criterion for regression; a node will only be split if a statistical FTest indicates a significant reduction of variance inside the subsets.

To be able to apply a classification algorithm on the datasets, the target attributes (\(\log_{10} k_{\text{off}}\), \(\log_{10} k_{\text{on}}\)) must be mapped from continuous values into binary classes: (0,0), (0,1), (1,0), and (1,1) based on the criteria of \(\log_{10} k_{\text{off}} = -2\) and \(\log_{10} k_{\text{on}} = 5.6\) (Figure 5.2 and Table 5.1).

![Figure 5.2 Discretization. Results of the discretization based on the criteria set at \(\log_{10} k_{\text{off}} = -2\) (x-axis) and \(\log_{10} k_{\text{on}} = 5.6\) (y-axis). Thirty-nine training records were discretized into four binary classes: (0,0), (0,1), (1,0), and (1,1). Each blue diamond represents one record.](image-url)
Table 5.1 Results of the discretization. There are 9, 8, 10, and 12 training records in the binary classes of (0,0), (0,1), (1,0), and (1,1), respectively. Each record was identified by its corresponding ligand name.

5.5.2 Single-Target Random Forest Regression Algorithm of Scikit-Learn

Random Forest Regressors were built using Scikit-Learn software package with n_estimators (the number of trees) set at 200. All the other parameters were chosen by default which included bootstrapping and MSE. All attributes in the training set were used for the test at each internal node.

5.6 Linear Regression Algorithm

A linear regression algorithm performs least squares regression to identify linear relations in the training data. The algorithm calculates a regression equation to predict the output (y) for a set of input attributes $x_1, x_2, ..., x_p$. The equation to calculate the output is expressed in the form of a linear combination of input attributes with each attribute associated with its respective weight $w_0, w_1, ..., w_p$, where $w_1$ is the weight of $x_1$ and $w_0$ is the intercept. An equation takes the form of

$$y = w_0 + w_1 x_1 + \ldots + w_p x_p = w_0 + X W,$$
where $W$ and $X$ are vectors. $W = (w_1,…,w_p)$ and $X = (x_1,…,x_p)$

A set of training instances is used to update the weights based on the minimization of $\hat{W}$.

$$\hat{W} = \min(\|y - XW\|^2)$$

5.6.1 Single-Target Linear Regression Algorithm

Two types of single-target linear regression algorithms were used in this study. They are elastic net and lasso.

A. Single-Target Elastic Net Linear Regression Algorithm

Elastic Net integrates L1 and L2 loss functions into the ordinary linear regression algorithm. Mathematically it solves the linear regression problem of the form:

$$\hat{W} = \min(\|y - XW\|^2 + \lambda_2|W|^2 + \lambda_1|W|)$$

When $\lambda_2 = 0$, it is lasso (L1 loss function) and in the case of $\lambda_1 = 0$, it is ridge (L2 loss function).

Three elastic net models were built with the following parameter sets:

Set 1: $\alpha = [0.05,0.1,0.25]$, l1_ratio = 0.75, max_iter = 10000

Set 2: $\alpha = [0.25,0.5,0.75]$, l1_ratio = 0.75, max_iter = 10000

Set 3: $\alpha = [0.25,0.5,0.75]$, l1_ratio = [0.25,0.5,0.75], max_iter = 10000

The penalty is controlled by $(a * L1 + b * L2)$ where $\alpha = a + b$ and l1_ratio = $a/(a+b)$. max_iter is the maximum number of iterations.

B. Single-Target Lasso Linear Regression Algorithm

Single-target linear lasso models were built with the following parameters:

alpha = 0.5/1.0/1.5, copy_X = True, fit_intercept = True, max_iter = 10000, normalize = False, tol = 0.0001, and warm_start = False. All other parameters were set to their default values. Alpha is the constant that multiplies the L1 term. Max_iter specifies the
maximum of iterations. Tol is the tolerance used for optimization. The updates will continue until it is smaller than tol.

### 5.6.2 Multi-Target Lasso Linear Regression Algorithm

Binary-target lasso linear models for the predictions of $\log_{10}k_{on}$ and $\log_{10}k_{off}$ were built with the following parameters: alpha = 0.5/1.0/1.5, copy_X = True, fit_intercept = True, max_iter = 10000, normalize = False, tol = 0.0001, and warm_start = False. All other parameters were set to their default values.

### 5.7 K-Nearest Neighbors Algorithm

In this study, two k-nearest neighbors (k-NN) models were built with n_neighbors = 2 or 5. Since the size of the training set is small, a brute force algorithm was adopted in the computation.

The k-nearest neighbors (k-NN) algorithm is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification. Instead of performing explicit generalization, instance-based learning compares new problem instances with instances seen in training, which have been stored in memory. When attempting to predict the value of the target variable for a new instance, the k-NN algorithm picks the k instances that are most similar to the new instance, and uses them for the prediction. The output depends on whether k-NN is used for classification or regression [67]:

- In k-NN classification, an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.
- In k-NN regression, the value of an object is the average of the values of its k nearest neighbors.
The similarity of two instances $X$ and $Y$ is given by the formula

$$\text{Similarity} (X, Y) = \sum_{i=1}^{d} (1 - |X_i - Y_i|)$$

where $d$ is the number of attributes of each instance, and $X_i$ is the normalized value of the $i^{th}$ attribute of $X$. “Normalized” means divided by the range of the attribute, so that $X_i \in [0,1]$ for every attribute $i$ and instance $X$. Similarity between two instances varies between 0 and $d$, with a similarity of $d$ indicating the two instances are identical.

Both for classification and regression, it can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant one. The weight of an instance $X$ is the similarity between $X$ and the test instance.

### 5.8 Results of Machine Learning Computation

For the simultaneous predictions of $\log_{10}k_{on}$ and $\log_{10}k_{off}$, the multi-target random forest classification algorithm from Clus and the multi-target lasso linear regression algorithm from Scikit-Learn were used for classification and regression, respectively. In addition, four single-target algorithms from Scikit-Learn were used to compute the values of $\log_{10}k_{off}$ and $\log_{10}k_{on}$ separately. They are the elastic net linear regression algorithm, the lasso linear regression algorithm, the random forest regression algorithm, and the $k$-NN instance-based learning algorithm.

#### 5.8.1 Classification Results

The experiment of multi-target random forest classification was conducted on eight various iteration numbers (1, 10, 25, 50, 100, 200, 250, 500) and trained with six different training datasets, for the simultaneous predictions of $\log_{10}k_{off}$ and $\log_{10}k_{on}$. The six training datasets are DS-PIE, DS-RMLR, DS-RMRR, DS-PIE + DS-RMLR, DS-RMRR + DS-RMLR, and DS-
RMLR + DS-RMRR, and DS-PIE + DS-RMLR + DS-RMRR. There are 44, 44, 44, 88, 88, and 132 training attributes in the feature vectors of the six datasets respectively.

Tables 5.2 and 5.3 show the results of all the classifiers for the binary classification predictions of log$_{10}$k$_{off}$ and log$_{10}$k$_{on}$, in terms of sensitivity, specificity, accuracy, and error rate from the LOO cross-validation experiments. The average accuracy of log$_{10}$k$_{off}$ prediction is 63.62% and the average accuracy of log$_{10}$k$_{on}$ prediction is 60.54.

<table>
<thead>
<tr>
<th>Iteration Number (log$<em>{10}$k$</em>{off}$)</th>
<th>Training Dataset</th>
<th>Measurement</th>
<th>1</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>250</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DS-PIE</td>
<td>Sensitivity</td>
<td>0.73</td>
<td>0.86</td>
<td>0.91</td>
<td>0.91</td>
<td>0.86</td>
<td>0.86</td>
<td>0.91</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.53</td>
<td>0.35</td>
<td>0.47</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy %</td>
<td>64.10</td>
<td>64.10</td>
<td>71.79</td>
<td>66.67</td>
<td>64.10</td>
<td>64.10</td>
<td>66.67</td>
<td>61.54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error rate %</td>
<td>35.90</td>
<td>35.90</td>
<td>28.21</td>
<td>33.33</td>
<td>35.90</td>
<td>35.90</td>
<td>33.33</td>
<td>38.46</td>
</tr>
<tr>
<td></td>
<td>DS-RMLR</td>
<td>Sensitivity</td>
<td>0.50</td>
<td>0.73</td>
<td>0.82</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.53</td>
<td>0.12</td>
<td>0.41</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy %</td>
<td>51.28</td>
<td>46.15</td>
<td>64.10</td>
<td>69.23</td>
<td>69.23</td>
<td>69.23</td>
<td>69.23</td>
<td>71.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error rate %</td>
<td>48.72</td>
<td>53.85</td>
<td>35.90</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
<td>28.21</td>
</tr>
<tr>
<td></td>
<td>DS-RMRR</td>
<td>Sensitivity</td>
<td>0.59</td>
<td>0.73</td>
<td>0.77</td>
<td>0.73</td>
<td>0.77</td>
<td>0.77</td>
<td>0.77</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.35</td>
<td>0.53</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy %</td>
<td>48.72</td>
<td>53.85</td>
<td>35.90</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error rate %</td>
<td>51.28</td>
<td>53.85</td>
<td>35.90</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
<td>30.77</td>
</tr>
<tr>
<td></td>
<td>DS-PIE + DS-RMLR</td>
<td>Sensitivity</td>
<td>0.45</td>
<td>0.86</td>
<td>0.73</td>
<td>0.86</td>
<td>0.73</td>
<td>0.91</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.24</td>
<td>0.29</td>
<td>0.41</td>
<td>0.35</td>
<td>0.29</td>
<td>0.24</td>
<td>0.29</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy %</td>
<td>35.90</td>
<td>53.85</td>
<td>58.97</td>
<td>65.10</td>
<td>65.10</td>
<td>65.10</td>
<td>65.10</td>
<td>71.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error rate %</td>
<td>64.10</td>
<td>43.59</td>
<td>41.03</td>
<td>33.33</td>
<td>28.21</td>
<td>28.21</td>
<td>28.21</td>
<td>28.21</td>
</tr>
<tr>
<td></td>
<td>DS-RMRR + DS-RMRR</td>
<td>Sensitivity</td>
<td>0.68</td>
<td>0.91</td>
<td>0.95</td>
<td>0.86</td>
<td>0.91</td>
<td>0.86</td>
<td>0.86</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Specificity</td>
<td>0.41</td>
<td>0.47</td>
<td>0.41</td>
<td>0.41</td>
<td>0.47</td>
<td>0.53</td>
<td>0.53</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy %</td>
<td>56.41</td>
<td>71.79</td>
<td>71.79</td>
<td>66.67</td>
<td>71.79</td>
<td>71.79</td>
<td>71.79</td>
<td>66.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error rate %</td>
<td>43.59</td>
<td>28.21</td>
<td>28.21</td>
<td>33.33</td>
<td>28.21</td>
<td>28.21</td>
<td>28.21</td>
<td>33.33</td>
</tr>
</tbody>
</table>

Table 5.2 Results of the LOO cross-validation experiments of the binary-target classifiers of Clus for log$_{10}$k$_{off}$ prediction.
Table 5.3 Results of the LOO cross-validation experiments of the binary-target classifiers of Clus for log\textsubscript{10}k\textsubscript{on} prediction.

For the binary-target classification prediction, MM-Accuracy is used to measure the quality of the prediction. Table 5.4 reveals that the binary-target classifiers with DS-RMLR + DS-RMRR as the training dataset produced the highest MM-Accuracy value (92.40) in which the accuracy of log\textsubscript{10}k\textsubscript{off} is 68.59 and the accuracy of log\textsubscript{10}k\textsubscript{on} is 61.91. Thus, this model was used to evaluate the relationship between iteration number and accuracy. As shown in Figure 5.3, the accuracies of log\textsubscript{10}k\textsubscript{off} and log\textsubscript{10}k\textsubscript{on} reach the top of the curve at the iteration number of 200, and then begin to level off.
Table 5.4 Quality measure of binary-target classification prediction

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>log(<em>{10})k(</em>{\text{off}})</th>
<th>log(<em>{10})k(</em>{\text{on}})</th>
<th>MM-Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE</td>
<td>65.38</td>
<td>62.50</td>
<td>90.45</td>
</tr>
<tr>
<td>DS-RMLR</td>
<td>63.78</td>
<td>63.78</td>
<td>90.20</td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>60.90</td>
<td>45.83</td>
<td>76.22</td>
</tr>
<tr>
<td>DS-PIE + DS-RMLR</td>
<td>58.97</td>
<td>63.78</td>
<td>86.87</td>
</tr>
<tr>
<td>DS-RMLR + DS-RMRR</td>
<td>68.59</td>
<td>61.91</td>
<td>92.40</td>
</tr>
<tr>
<td>DS-PIE + DS-RMLR + DS-RMRR</td>
<td>64.10</td>
<td>65.38</td>
<td>91.57</td>
</tr>
</tbody>
</table>

MM-Accuracy = (k\(_{\text{on}}\)-Accuracy\(^2\) + k\(_{\text{off}}\)-Accuracy\(^2\))\(^{1/2}\). The binary-target classifiers trained with DS-RMLR + DS-RMRR generated the highest MM-Accuracy = 92.40.

Figure 5.3 Iteration number versus accuracy. Results of log\(_{10}\)k\(_{\text{off}}\) and log\(_{10}\)k\(_{\text{on}}\) prediction were given by the binary-target Random Forest classifier of Clus trained with DS-RMLR + DS-RMRR.

5.8.2 Feature Selection

In supervised learning problems involving very high dimensional data, it is often desirable to reduce the number of features given to the learning machine for the following reasons: 1. Removal of irrelevant variable may improve the performance of the learning machine. 2. Identifying only those features that are important for ML prediction may help in the interpretation of the model. 3. Reducing the number of features may provide faster and more cost-effective predictors.
In this study, statistical experiments were conducted to identify the training attributes with frequency of occurrence greater than 25% in the LOO cross-validation experiments of binary-target random forest classification algorithm with iteration number = 500. The procedure of the feature selection is as follows:

1. For each classifier with iteration number = 500, the results of node test, including which feature was selected, and the score of importance of the feature, were written to a text file. Since three principal training datasets, each of which comprises 39 records, were used in this study, in total, there are 117 forests (3*39) and 58500 trees were constructed (500*3*39 = 58500).

2. Count the number of occurrences (N) each feature in each forest.

3. Calculate the frequency of occurrence as follows: 100*(N /500) %

4. Calculate the average of the frequency of the 39 results given by the LOO experiment.

5. Select the features with frequency of occurrence > 25%

6. Calculate the average of the score of importance for each of the selected features.

Table 5.5 shows the selected features in the descending order of score of importance. Consequently, fourteen, sixteen, and fifteen features were selected from DS-PIE, DS-RMLR and DS-RMRR and assigned to build new datasets F-PIE, F-RMLR, and F-RMRR respectively. In addition, there are eleven residues are common to the three datasets F-PIE, F-RMLR, and F-RMRR. They are: R8, L10, L23, D25, G27, A28, D29, D30, V32, K45, and A52.
Table 5.5 Attribute selection. The selection criterion is: frequency of attribute occurrence ≥ 25%. Most of the residues are in chain A, and only two residues in green color are in chain B. Residues of protease inhibitor resistance mutation are in red.

### 5.8.3 Regression Results

A multi-target lasso regression algorithm, a single-target k-nearest-neighbors instance-based learning algorithm, and three single-target regression algorithms including a random forest algorithm, a lasso algorithm and an elastic net algorithm, were used to predict the two kinetic rate constants.

#### 5.8.3.1 Single-Target Elastic Net Linear Regression Algorithm

Table 5.6 shows the %deviation and the Pearson’s correlation coefficient (PC) from the result analysis of the LOO cross-validation experiments of single-target elastic net linear regression algorithm. Seven different training datasets (DS-PIE, DS-RMLR, DS-RMRR, DS-PIE+DS-RMLR+DS-RMRR, F-PIE, F-RMLR and F-RMRR) were used to train the regressors.
5.8.3.2 Single-Target Lasso Linear Regression Algorithm

Single-target lasso linear regressors were conducted on three different alpha values, 0.5, 1.0, and 1.5 and trained with four training datasets: DS-PIE, DS-RMLR, DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR. Table 5.7 shows the predicted results of log10koff and log10kon from the LOO cross-validation experiments of the single-target lasso linear regressors.

5.8.3.3 Binary-Target Lasso Linear Regression Algorithm

Table 5.8 shows the predicted results of log10koff and log10kon from the LOO cross-validation experiments of the binary-target lasso linear regressors. Four training datasets including DS-PIE, DS-RMLR, DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR were used to train the regressors on three alpha values, 0.5, 1.0, and 1.5.

5.8.3.4 Single-Target Random Forest Algorithm

There are two stages in the experiment with the random forest regression algorithm. First, the optimal iteration number for the algorithm was determined. Second, the random forest regressors trained with different training datasets were conducted at the optimal iteration number for the log10kon and log10koff predictions.

- Optimal Iteration Number

Random forest algorithm was run with the number of iteration set at 200, 300 and 500 for the log10kon and log10koff predictions. As shown in Table 5.9, Figure 5.4 and 5.5, the regressors with iteration number = 200 performed the best.
Figure 5.4 Results of log₁₀kₒₜ prediction given by the random forest regressors. The regressors were trained with DS-PIE, DS-RMLR, and DS-RMRR and conducted at three different iteration numbers: 200, 300 and 500.
Figure 5.5 Results of $\log_{10}k_{\text{off}}$ prediction given by the random forest regressors. The regressors were trained with DS-PIE, DS-RMLR, and DS-RMRR and conducted at three different iteration numbers: 200, 300 and 500.

- $\log_{10}k_{\text{on}}$ and $\log_{10}k_{\text{off}}$ Predictions

A single-target random forest regression algorithm was run with iteration number = 200 for $\log_{10}k_{\text{off}}$ and $\log_{10}k_{\text{on}}$ predictions. Five training datasets were used to train the random forest regressors. Table 5.10 shows the %deviation and PC values from the result analysis of the LOO cross-validation experiments of the regressors.
<table>
<thead>
<tr>
<th>Training dataset</th>
<th>Parameter Set</th>
<th>%deviation</th>
<th>PC</th>
<th>%deviation</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DS-PIE</strong></td>
<td>Set 1</td>
<td>20.81</td>
<td>0.0525</td>
<td>201.27</td>
<td>0.1488</td>
</tr>
<tr>
<td></td>
<td>Set 2</td>
<td>20.72</td>
<td>-0.5765</td>
<td>204.74</td>
<td>0.1933</td>
</tr>
<tr>
<td></td>
<td>Set 3</td>
<td>20.72</td>
<td>-0.5765</td>
<td>204.74</td>
<td>0.1933</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>20.75</td>
<td></td>
<td>203.59</td>
<td></td>
</tr>
<tr>
<td><strong>DS-RMLR</strong></td>
<td>Set 1</td>
<td>22.59</td>
<td>0.0616</td>
<td>273.28</td>
<td>-0.0455</td>
</tr>
<tr>
<td></td>
<td>Set 2</td>
<td>22.60</td>
<td>-0.2161</td>
<td>267.31</td>
<td>-0.2753</td>
</tr>
<tr>
<td></td>
<td>Set 3</td>
<td>21.75</td>
<td>0.0723</td>
<td>268.93</td>
<td>0.0567</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>22.31</td>
<td></td>
<td>269.84</td>
<td></td>
</tr>
<tr>
<td><strong>DS-RMRR</strong></td>
<td>Set 1</td>
<td>20.20</td>
<td>0.4366</td>
<td>284.09</td>
<td>0.2691</td>
</tr>
<tr>
<td></td>
<td>Set 2</td>
<td>19.65</td>
<td>0.4488</td>
<td>263.83</td>
<td>0.2292</td>
</tr>
<tr>
<td></td>
<td>Set 3</td>
<td>19.67</td>
<td>0.4451</td>
<td>281.26</td>
<td>0.1355</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>19.84</td>
<td></td>
<td>276.39</td>
<td></td>
</tr>
<tr>
<td><strong>DS-PIE +</strong></td>
<td>Set 1</td>
<td>21.85</td>
<td>0.3553</td>
<td>211.47</td>
<td>0.3796</td>
</tr>
<tr>
<td><strong>DS-RMLR +</strong></td>
<td>Set 2</td>
<td>19.74</td>
<td>0.4516</td>
<td>248.93</td>
<td>0.2009</td>
</tr>
<tr>
<td><strong>DS-RMRR</strong></td>
<td>Set 3</td>
<td>19.74</td>
<td>0.4516</td>
<td>246.91</td>
<td>0.2576</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>20.44</td>
<td></td>
<td>235.77</td>
<td></td>
</tr>
<tr>
<td><strong>F-PIE</strong></td>
<td>Set 1</td>
<td>20.06</td>
<td>0.1781</td>
<td>179.35</td>
<td>0.2359</td>
</tr>
<tr>
<td></td>
<td>Set 2</td>
<td>20.11</td>
<td>0.0439</td>
<td>179.19</td>
<td>0.2388</td>
</tr>
<tr>
<td></td>
<td>Set 3</td>
<td>20.44</td>
<td>0.0584</td>
<td>185.32</td>
<td>0.1993</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>20.20</td>
<td></td>
<td>181.29</td>
<td></td>
</tr>
<tr>
<td><strong>F-RMLR</strong></td>
<td>Set 1</td>
<td>19.50</td>
<td>0.4543</td>
<td>250.47</td>
<td>0.3221</td>
</tr>
<tr>
<td></td>
<td>Set 2</td>
<td>21.39</td>
<td>-0.0407</td>
<td>245.33</td>
<td>0.0783</td>
</tr>
<tr>
<td></td>
<td>Set 3</td>
<td>20.21</td>
<td>0.2002</td>
<td>247.05</td>
<td>0.2695</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>20.37</td>
<td></td>
<td>247.62</td>
<td></td>
</tr>
<tr>
<td><strong>F-RMRR</strong></td>
<td>Set 1</td>
<td>20.23</td>
<td>0.3619</td>
<td>284.21</td>
<td>0.0776</td>
</tr>
<tr>
<td></td>
<td>Set 2</td>
<td>20.26</td>
<td>0.2664</td>
<td>273.28</td>
<td>0.1194</td>
</tr>
<tr>
<td></td>
<td>Set 3</td>
<td>20.26</td>
<td>0.2647</td>
<td>273.30</td>
<td>0.1134</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>20.25</td>
<td></td>
<td>276.93</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target elastic net regressors
Set 1: alpha = [0.05,0.1,0.25], l1_ratio = 0.75, max_iter = 10000
Set 2: alpha = [0.25,0.5,0.75], l1_ratio = 0.75, max_iter = 10000
Set 3: alpha = [0.25,0.5,0.75], l1_ratio = [0.25,0.5,0.75], max_iter = 10000
The penalty is controlled by \((a \times L1 + b \times L2)\) where alpha = a + b and l1_ratio = a/(a+b). max_iter is the maximum number of iterations.
<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Alpha</th>
<th>(%\text{deviation})</th>
<th>PC</th>
<th>(%\text{deviation})</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE</td>
<td>0.5</td>
<td>20.61</td>
<td>-0.2373</td>
<td>195.37</td>
<td>0.2342</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>20.66</td>
<td>-0.9999</td>
<td>239.25</td>
<td>-0.9974</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.61</td>
<td>-0.2373</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>20.63</td>
<td></td>
<td>224.59</td>
<td></td>
</tr>
<tr>
<td>DS-RMLR</td>
<td>0.5</td>
<td>21.15</td>
<td>-0.4123</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>20.66</td>
<td>-0.9999</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.66</td>
<td>-0.9999</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>20.82</td>
<td></td>
<td>239.15</td>
<td></td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>0.5</td>
<td>19.80</td>
<td>0.4398</td>
<td>251.57</td>
<td>0.2794</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>19.80</td>
<td>0.4298</td>
<td>252.59</td>
<td>0.2937</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.15</td>
<td>0.2870</td>
<td>250.27</td>
<td>0.3008</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>19.92</td>
<td></td>
<td>251.47</td>
<td></td>
</tr>
<tr>
<td>DS-PIE+DS-RMLR+DS-RMRR</td>
<td>0.5</td>
<td>19.80</td>
<td>0.4398</td>
<td>224.16</td>
<td>0.3138</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>19.80</td>
<td>0.4298</td>
<td>252.59</td>
<td>0.2937</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.15</td>
<td>0.2870</td>
<td>250.27</td>
<td>0.3008</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>19.92</td>
<td></td>
<td>242.34</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7 \%\text{deviation} and Pearson’s correlation coefficient of the predicted results given by the single-target lasso regressors

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Alpha</th>
<th>(%\text{deviation})</th>
<th>PC</th>
<th>(%\text{deviation})</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE</td>
<td>0.5</td>
<td>20.53</td>
<td>-0.1696</td>
<td>208.04</td>
<td>0.2154</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>20.66</td>
<td>-0.9993</td>
<td>239.26</td>
<td>-0.9868</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.66</td>
<td>-0.9999</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>20.62</td>
<td></td>
<td>228.82</td>
<td></td>
</tr>
<tr>
<td>DS-RMLR</td>
<td>0.5</td>
<td>21.30</td>
<td>-0.3000</td>
<td>250.40</td>
<td>-0.5469</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>20.66</td>
<td>-0.9999</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.66</td>
<td>-0.9999</td>
<td>239.15</td>
<td>-0.9999</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>20.87</td>
<td></td>
<td>242.90</td>
<td></td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>0.5</td>
<td>19.54</td>
<td>0.4548</td>
<td>268.63</td>
<td>0.2480</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>19.82</td>
<td>0.4209</td>
<td>255.42</td>
<td>0.2575</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>20.13</td>
<td>0.3151</td>
<td>252.00</td>
<td>0.2397</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>19.83</td>
<td></td>
<td>258.68</td>
<td></td>
</tr>
<tr>
<td>DS-PIE+</td>
<td>0.5</td>
<td>19.89</td>
<td>0.4344</td>
<td>235.31</td>
<td>0.3184</td>
</tr>
<tr>
<td>DS-RMLR+</td>
<td>1.0</td>
<td>19.82</td>
<td>0.4209</td>
<td>255.42</td>
<td>0.2575</td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>1.5</td>
<td>20.13</td>
<td>0.3151</td>
<td>252.00</td>
<td>0.2397</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>19.95</td>
<td></td>
<td>247.58</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.8 \%\text{deviation} and Pearson’s correlation coefficient of the predicted results given by the binary-target lasso regressors
Table 5.9 Results of \(\log_{10}k_{\text{on}}, \log_{10}k_{\text{off}}\) predictions given by the random forest regressors. The regressors were trained with DS-PIE, DS-RMLR, and DS-RMRR and conducted at three different iteration numbers: 200, 300 and 500.

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>(\log_{10}k_{\text{on}}) - %deviation</th>
<th>(\log_{10}k_{\text{off}}) - %deviation</th>
<th>MM-%deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE-200</td>
<td>25.49</td>
<td>201.13</td>
<td>202.74</td>
</tr>
<tr>
<td>DS-PIE-300</td>
<td>25.86</td>
<td>202.54</td>
<td>204.18</td>
</tr>
<tr>
<td>DS-PIE-500</td>
<td>25.82</td>
<td>203.38</td>
<td>205.01</td>
</tr>
<tr>
<td>DS-RMLR-200</td>
<td>19.03</td>
<td>226.29</td>
<td>227.09</td>
</tr>
<tr>
<td>DS-RMLR-300</td>
<td>19.12</td>
<td>227.62</td>
<td>228.42</td>
</tr>
<tr>
<td>DS-RMLR-500</td>
<td>19.20</td>
<td>228.45</td>
<td>229.26</td>
</tr>
<tr>
<td>DS-RMRR-200</td>
<td>22.36</td>
<td>201.60</td>
<td>202.84</td>
</tr>
<tr>
<td>DS-RMRR-300</td>
<td>22.40</td>
<td>212.25</td>
<td>213.43</td>
</tr>
<tr>
<td>DS-RMRR-500</td>
<td>23.14</td>
<td>202.66</td>
<td>203.98</td>
</tr>
</tbody>
</table>

Table 5.10 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target random forest regressors (iteration number = 200).

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>(\log_{10}k_{\text{on}}) %deviation</th>
<th>PC</th>
<th>(\log_{10}k_{\text{off}}) %deviation</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE</td>
<td>25.49</td>
<td>-0.1839</td>
<td>201.13</td>
<td>-0.0649</td>
</tr>
<tr>
<td>DS-RMLR</td>
<td>19.03</td>
<td>0.4914</td>
<td>226.29</td>
<td>0.3560</td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>22.36</td>
<td>0.0949</td>
<td>201.60</td>
<td>0.2200</td>
</tr>
<tr>
<td>DS-PIE+DS-RMLR+DS-RMRR</td>
<td>22.54</td>
<td>0.1104</td>
<td>220.72</td>
<td>0.1335</td>
</tr>
<tr>
<td>F-PIE+F-RMLR+F-RMRR</td>
<td>21.06</td>
<td>0.3383</td>
<td>184.96</td>
<td>0.2973</td>
</tr>
</tbody>
</table>

5.8.4 K-Nearest Neighbors

A single-target K-Nearest Neighbor algorithm was conducted on \(k = 2\) or \(k = 5\). Four training datasets including DS-PIE, DS-RMLR, DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR were used for the training. Table 5.11 shows the %deviation and Pearson’s correlation coefficient from the result analysis of the LOO cross-validation experiments of the regressors.
5.9 Impact of Individual Training Dataset on the Accuracy of Prediction

In this study, 10 training datasets have been used to train the ML models for the predictions of $\log_{10}k_{\text{on}}$ and $\log_{10}k_{\text{off}}$. As shown in Table 5.12, there are 6, 7, 4, 4, 5, and 4 training datasets used for binary-target random forest classification, single-target elastic net linear regression, binary-target lasso linear regression, single-target lasso linear regression, single-target random forest regression, and single-target K-nearest neighbors instance-based learning respectively. This section analyzes the impact of individual training dataset on the predicting accuracy of $\log_{10}k_{\text{on}}$ as well as $\log_{10}k_{\text{off}}$. 

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>N</th>
<th>$%$deviation</th>
<th>PC</th>
<th>$%$deviation</th>
<th>PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE</td>
<td>2</td>
<td>27.07</td>
<td>-0.3226</td>
<td>280.39</td>
<td>0.0332</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>23.80</td>
<td>-0.1377</td>
<td>291.67</td>
<td>-0.0121</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>25.43</td>
<td></td>
<td>286.03</td>
<td></td>
</tr>
<tr>
<td>DS-RMLR</td>
<td>2</td>
<td>24.01</td>
<td>0.2284</td>
<td>90.38</td>
<td>0.2214</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>22.73</td>
<td>-0.0642</td>
<td>171.23</td>
<td>-0.1610</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>23.37</td>
<td></td>
<td>130.81</td>
<td></td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>2</td>
<td>21.67</td>
<td>0.4086</td>
<td>248.72</td>
<td>0.0483</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>20.64</td>
<td>0.2045</td>
<td>190.25</td>
<td>0.1470</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>21.15</td>
<td></td>
<td>219.48</td>
<td></td>
</tr>
<tr>
<td>DS-PIE+</td>
<td>2</td>
<td>20.54</td>
<td>0.4443</td>
<td>239.79</td>
<td>0.1117</td>
</tr>
<tr>
<td>DS-RMLR+</td>
<td>5</td>
<td>19.83</td>
<td>0.3226</td>
<td>215.18</td>
<td>0.1756</td>
</tr>
<tr>
<td>DS-RMRR</td>
<td></td>
<td>20.19</td>
<td></td>
<td>227.49</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.11 %deviation and Pearson’s correlation coefficient of the predicted results given by the single-target k-NN regressors
Table 5.12 Training datasets used for ML models for the predictions of log_{10}k_{on} and log_{10}k_{off}

5.9.1 Binary-Target Random Forest Classification Algorithm of Clus

Six training datasets were used to train random forest classifiers for log_{10}k_{off} and log_{10}k_{on} predictions. They are DS-PIE, DS-RMLR, DS-RMRR, DS-PIE+DS-RMLR, DS-RMLR+DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR. Figure 5.6 shows that the classifiers trained with DS-RMLR+DS-RMRR produced the largest MM-Accuracy (92.40) with average accuracy of log_{10}k_{off} = 68.59% and average accuracy of log_{10}k_{on} = 61.91%. Combining DS-PIE with DS-RMLR+DS-RMRR increases the average accuracy of log_{10}k_{on} (+3.47%), but decreases the average accuracy of log_{10}k_{off} (-4.49%) possibly due to over-fitting. In addition, all predicted values of average accuracy are ≥ 58.97%, except for the average accuracy of log_{10}k_{on} of DS-RMRR (45.83%).
Figure 5.6 Results of binary-target random forest classification

5.9.2 Single-Target Elastic net Linear Regression Algorithm of Scikit-Learn

Seven training datasets including DS-PIE, DS-RMLR, DS-RMRR, DS-PIE+DS-RMLR+DS-RMRR, F-PIE, F-RMLR, and F-RMRR were used to train the single-target elastic net regressors.

For the log\(_{10}\)\(k_{on}\) prediction, the minimum, maximum, standard deviation, and average %deviation are 19.84, 22.31, 0.80, and 20.59. DS-RMRR performed the best with %deviation = 19.84 (Figure 5.7).

For the log\(_{10}\)\(k_{off}\) prediction, the minimum, maximum, standard deviation, and average %deviation are 181.29, 276.93, 37.43 and 241.63, respectively (Figure 5.8).
Figure 5.7 Results of $\log_{10} k_{on}$ prediction of single-target elastic net regression
Figure 5.8 Results of $\log_{10}k_{\text{off}}$ prediction of single-target elastic net regression

5.9.3 Single-Target Lasso Linear Regression Algorithm of Scikit-Learn

Four training datasets including DS-PIE, DS-RMLR, DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR were used to train the single-target lasso linear regressors.

For the $\log_{10}k_{\text{on}}$ prediction, the minimum, maximum, standard deviation, and average %deviation are 19.92, 20.82, 0.471, and 20.32, respectively. Two datasets, DS-RMRR and DS-PIE+DS-RMLR+DS-RMRR, performed equally well with %deviation = 19.92. (Figure 5.9).

For the $\log_{10}k_{\text{off}}$ prediction, the minimum, maximum, standard deviation, and average %deviation are 224.59, 251.47, 11.16, and 239.38, respectively (Figure 5.10).
Figure 5.9 Results of log_{10}k_{on} prediction of single-target lasso linear regression

Figure 5.10 Results of log_{10}k_{off} prediction of single-target lasso linear regression
5.9.4 Single-Target K-Nearest Neighbors Instance-Based Learning

Four training datasets, DS-PIE, DS-RMLR, DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR, were used to train the single-target K-Nearest Neighbors models.

For the log10k_on prediction, the minimum, maximum, standard deviation, and average %deviation are 20.19, 25.43, 2.34, and 22.53, respectively. DS-PIE+DS-RMLR+DS-RMRR performed the best with %deviation = 20.19 (Figure 5.11).

For the log10k_off prediction, the minimum, maximum, standard deviation, and average %deviation are 130.81, 286.03, 64.04, and 215.95, respectively (Figure 5.12).

![Figure 5.11 Results of log10k_on prediction of single-target K-nearest neighbors instance-based learning](image-url)
5.9.5 Single-target Random Forest Regression

Five training datasets including DS-PIE, DS-RMLR, DS-RMRR, DS-PIE+DS-RMLR+DS-RMRR and F-PIE+F-RMLR+F-RMRR were used to train the single-target random forest regressors (iteration number = 200).

For the $\log_{10}k_{\text{on}}$ prediction, the minimum, maximum, standard deviation, and average %deviation are 19.03, 25.49, 2.35, and 22.09, respectively. DS-RMLR performed the best with %deviation = 19.03 (Figure 5.13).

For $\log_{10}k_{\text{off}}$ prediction, the minimum, maximum, standard deviation, and average %deviation are 184.96, 226.29, 16.65, and 206.94, respectively (Figure 5.14).
Figure 5.13 Results of log₁₀$k_{on}$ prediction of single-target random forest regression

Figure 5.14 Results of log₁₀$k_{off}$ prediction of single-target random forest regression
5.9.6 Binary-Target Lasso Regression Algorithm of Scikit-Learn

Four datasets including DS-PIE, DS-RMLR, DS-RMRR, and DS-PIE+DS-RMLR+DS-RMRR were used to train the binary-target lasso linear regressors.

For the log$_{10}$kon prediction, the minimum, maximum, standard deviation and average %deviation are 19.83, 20.87, 0.506, and 20.31, respectively. DS-RMRR performed the best in the prediction of log$_{10}$kon with %deviation = 19.83 (Figure 5.15).

For the log$_{10}$koff prediction, the minimum, maximum, standard deviation and average %deviation are 228.82, 258.68, 12.36 and 244.49, respectively (Figure 5.16).

![Figure 5.15 Results of log$_{10}$kon prediction of binary-target lasso linear regression]
Figure 5.16 Results of $\log_{10}k_{\text{off}}$ prediction of binary-target lasso linear regression
5.9.7 Multi-Target Lasso versus Single-Target Lasso

As the MM-%deviation shown in Figure 5.17, single-target lasso regression algorithm performs slightly better than binary-target lasso regression algorithm in the $\log_{10}k_{on}$ and $\log_{10}k_{off}$ predictions.

![Figure 5.17 MM-%deviation. Binary-target lasso linear regression versus Single-target lasso linear regression](Image)
5.10 Feature Evaluation

In addition to the feature selection process (Section 5.8.2) which was carried out for the identification of the features signification to PLI, two techniques were used to evaluate the characteristics of the features in the principal training datasets. They are Welch’s t-test and receiver operating characteristic curve. The Welch’s t-test (two-tailed) was conducted to evaluate the roles of the features in PLI versus kinetic rate constants. The ROC curves using the data in F-PIE, F-RMLR, and F-RMRR were constructed to identify residues of protease inhibitor resistance mutation.

5.10.1 Protease Inhibitor Resistance Mutation

In 2004, Ceccherini-Silberstein et al. reported that in naive patients, the amino acid sequence of HIV-1 showed conservation in sixty-eight out of ninety-nine residues. But under drug pressure the conserved residues reduced to forty-five [68].

The emergence of drug-resistant and cross-resistant mutants of HIV protease has impeded the effectiveness of inhibitors, rendering AIDS with no definitive cure. There are twenty-six protease inhibitor resistance mutations (PIRM) reported by the World Health Organization in 2013. They are L10, V11, K20, L23, L24, D30, V32, L33, K43, M46, I47, G48, I50, F53, I54, Q58, A71, G73, T74, L76, V82, N83, I84, N88, L89, and L90. The fourteen residues in red color are major PIRM, and the twelve residues in green color are non-polymorphic accessory mutation. PIRM is significant because it changes drug activity by inducing alteration of the conformational shape of the protease. Referring to the structure of Ritonavir-HIV-1 Protease complex (Figure 5.18), twelve among the twenty-six mutations are active site mutations with the distance between the mutant residue and Ritonavir less than or equal to 4.2Å approximately. The rest of the
fourteen mutations are non-active site mutations with the distance between the mutant residue and Ritonavir ≥ 8.29Å (Table 5.13). Active site mutations are capable of directly changing the interactions of the inhibitors with the protease through steric hindrance and pairwise interaction. Non-active mutations are considered to affect by using other mechanisms, like influencing dimer stability (I50V) and allosteric effect. Additionally, in 1999 Tsai et al. reported that the consequence of non-active mutation could be a phenomenon of shifts in energy landscapes in the events of folding and binding cascades [69,70].

One of the strategies to tackle the problem of drug resistance is to bind ligand to the backbone atoms of the active site, as there is minimal change in the backbone conformation between wild type and mutant proteases. For example, Darunavir was designed using its bis-THF (bis-tetrahydrofuranylurethane) group to form hydrogen bonds to bind to the backbone atoms of the S2 sub-site of HIV-1 protease.

Figure 5.18 Active site PIRM and non-active site PIRM in Ritonavir-bound HIV-1 complex (PDB code 1HXW). HIV-1 protease is shown as transparent blue NewCartoon. Ritonavir is shown as green lines. Ten active site mutant residues are shown as blue lines. Sixteen non-active site mutant residues are shown as red lines.
Residue of PIRM

<table>
<thead>
<tr>
<th>active site mutation</th>
<th>non-active site mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>L10</td>
<td>V11 [13.84]</td>
</tr>
<tr>
<td>L23</td>
<td>K20 [15.39]</td>
</tr>
<tr>
<td>D30</td>
<td>L24 [9.40]</td>
</tr>
<tr>
<td>V32</td>
<td>L33 [11.17]</td>
</tr>
<tr>
<td>I47</td>
<td>K43 [22.43]</td>
</tr>
<tr>
<td>G48</td>
<td>M46 [15.82]</td>
</tr>
<tr>
<td>I50</td>
<td>Q58 [18.29]</td>
</tr>
<tr>
<td>F53</td>
<td>A71 [20.53]</td>
</tr>
<tr>
<td>I54</td>
<td>G73 [19.93]</td>
</tr>
<tr>
<td>L76</td>
<td>T74 [18.15]</td>
</tr>
<tr>
<td>V82</td>
<td>N83 [8.29]</td>
</tr>
<tr>
<td>I84</td>
<td>N88 [15.69]</td>
</tr>
<tr>
<td>L89</td>
<td>L90 [15.86]</td>
</tr>
<tr>
<td>L90</td>
<td>L90 [15.23]</td>
</tr>
</tbody>
</table>

Table 5.13 Residues of PIRM. The numeric values in blanket are the distances (Å) between residues and Ritonavir.

5.10.2 Receiver Operating Characteristic Curve

As shown in Figure 5.19, the areas under the ROC curves of F-RMLR, F-PIE, and F-RMRR are 0.592, 0.47, and 0.304 respectively. F-RMLR outperformed F-RMRR and F-PIE in the capability of the differentiation between PIRM residue and non-PIRM residue.
Figure 5.19 ROC curves of F-PIE, F-RMLR, and F-RMRR. The areas under the ROC curves of F-RMLR, F-PIE, and F-RMRR are 0.592, 0.47, and 0.324 respectively.

5.10.3 Welch’s T-Test

Splitting the thirty-nine training records in each of the three principal datasets (DS-PIE, DS-RMLR, and DS-RMRR) into two subsets (X and Y) using each of the three criteria: \( \log_{10} k_{\text{on}} \geq 5.5502 \), \( k_{\text{off}} \geq 0.00653 \), and \( \log_{10} k_{D} \leq -7.9856 \). Table 5.14 shows the results of the splitting. Using these three criteria, three pairs of binary subsets were produced from each principal dataset and a total of nine binary subsets were generated. Two-tailed Welch’s t-test was conducted on these nine pairs of binary subsets. Table 5.15 lists the number of occurrence of the features having p-value < 0.05 in each principal training dataset. The results indicate that there are eleven residues having p-values < 0.05, including R8, L10, L23, D25, A28, D29, D30, V32, G49, I50, and P81.
Table 5.14 Results of splitting. Using the three criteria, three different pairs of binary subsets were produced from each principal dataset.

<table>
<thead>
<tr>
<th>Residue I.D.</th>
<th>DS-PIE</th>
<th>DS-RMLR</th>
<th>DS-RMRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>32</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>47</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>52</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>54</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>76</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>82</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>84</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5.15 Number of occurrence of attributes with two-tailed p-value < 0.05 in each principal training dataset.
Chapter 6

Structural Determinants of Protein-Ligand Binding Kinetics

HIV-1 is an aspartic protease comprising two monomers with only one active site. Each monomer is composed of ninety-nine amino acids. The catalytic site has the characteristic Asp25-Thr26-Gly27 sequence common to all aspartic proteases. There are three important regions in the protease structure. They are the active site cavity, the flexible flaps, and the dimer interface. The active site cavity comprises residues Arg8, Leu23, Asp25, Gly27, Ala28, Asp29, Asp30, Val32, Lys45, Ile47, Met46, Gly48, Gly49, Ile50, Phe53, Leu76, Thr80, Pro81, Val82, and Ile84 (Figure 6.1A). The majority of the residues forming the substrate binding site are hydrophobic except Asp25 and Asp29, which form hydrogen bonds with peptide main chain groups, and Arg8, Asp30 and Lys45, which can interact with polar side chains or distal main chain groups in longer peptides [71].

The flap region comprises residues Lys43, Pro44, Lys45, Met46, Ile47, Gly48, Gly49, Ile50, Gly51, Gly52, Phe53, Ile54, Lys55, Val56, Arg57, and Gln58 (Figure 6.1B) [72]. The flaps were observed in closed and open conformation in the crystal structures of inhibitor bound and free protease [73, 74].

L24I, I50V, and F53L are the mutations showing significant impacts on dimer stability in the urea test. The mutants are only ~50% stable i.e. half of the activity is lost at a urea concentration that is 50% lower than seen for the wild type protease (Figure 6.1C) [75, 76].
Figure 6.1 Structure of HIV-1 protease (PDB code: 1HXW, Ritonavir is not shown)  
(A) Active site region comprises twenty residues (circles). Residues on chain A/B  
are in red/green respectively. (B) The flap region comprises sixteen residues (resid  
43-58, green NewCartoon). (C) Significant residues to dimmer stability: F53, I50,  
and L24 are in order from top to bottom (residues on chain A/B are in red/blue).
Several mechanisms for aspartic protease have been proposed. The most widely accepted mechanism is the one proposed by Suguna et al [77]. In the proposed mechanism, a water molecule is activated by the negative aspartate side chain. Then, it applies strain on the carbonyl group in the substrate scissile amide bond, causing it to rotate out of plane and lose its double bond character. This enhances its vulnerability towards hydrolysis (Figure 6.2).

Figure 6.2 Catalytic mechanism for aspartic protease proposed by Suguna et al.
6.1 Elementary Aspects of Ligand Protein Binding Kinetics

Dissociation constant (K\text{d}) or its proxy, half maximal inhibitory concentration (IC\text{50}), is predicated on the assumption that affinity is a suitable representative for in vivo efficacy. But in an open, in vivo system the concentration of a drug may vary on a timescale faster than binding and unbinding to its receptor. Thus, equilibrium binding affinity is not an accurate measure for efficacy in vivo, but instead the rates of drug-receptor association and dissociation, as reflected by the rate constants k\text{on} and k\text{off}, are more appropriate for the measure. In 2013, Pan et al. reported that residence time is highly correlated with functional efficacy of a series of agonists of the A\textsubscript{2A} adenosine receptor (r\textsuperscript{2} = 0.95), but there is little correlation with binding affinity (r\textsuperscript{2} = 0.15) (Figure 6.3 A & B) [4,5]. Thus, the residence time of a drug-receptor complex, \( \tau = 1/ k\text{off} \), is a better predictor of efficacy than binding affinity is. Assuming that target selectivity is important, a drug with a longer time on one receptor can select kinetically for that receptor over another, even when the affinity for both receptors is comparable. Moreover, a faster-binding drug might target a short-lived receptor more effectively.

On the other hand, the on-target side effect could be reduced by reducing the drug residence time. Thus, a drug with optimal efficacy and safety profile should have a balanced k\text{on} and k\text{off}. Since K\text{d} and IC\text{50} depend on the measurement of the combined effect of k\text{on} and k\text{off}, they are actually insufficient to explain the impact of conformational dynamics on PLI, as the same value of K\text{d} can come from infinite number of combinations of k\text{on} and k\text{off}. Additionally, since K\text{d} is dependent on the free energy difference between the bound and unbound states but is independent on the transition state of PLI, it is inadequate to elucidate the binding kinetics of PLI (Figure 6.4).
Figure 6.3 (A) log (residence time, $\tau$) versus functional efficacy. (B) log(binding affinity, $K_d$) versus functional efficacy.

Figure 6.4 depicts a free energy profile of a drug (D) binding to a receptor (R) to form a drug-receptor complex (DR) with the reaction described as follows:

\[
D + R \quad \xrightleftharpoons[k_{\text{off}}]{k_{\text{on}}} \quad DR
\]

where

\[
K_d = [D][R]/[DR] = k_{\text{off}}/k_{\text{on}}
\]

The free energy difference between the bound and unbound states, $\Delta G_d$, determines the binding affinity as $K_d = e^{-\Delta G_d/RT}$. The association rate constant, $k_{\text{on}} = e^{-\Delta T_{\text{on}}/RT}$, depends on the free energy difference, $\Delta T_{\text{on}}$, between the bound state and the transition state. The dissociation rate constant, $k_{\text{off}} = e^{-\Delta T_{\text{off}}/RT}$, depends on the free energy difference, $\Delta T_{\text{off}}$, between the unbound state and the transition state. Here, R is the ideal gas constant and T is the temperature in Kelvin.
6.2 Ligand-Protein Binding Models

In 1958, Daniel Koshland [78] proposed the induced fit hypothesis that built upon Emil Fisher’s key-lock principal and suggested that the binding of the ligand may induce the conformational change of the protein. An example of protein-ligand binding interaction conducted by induced fit mechanism is the binding of 2-phoglycolate (PGA) inhibitor to phosphoenolpyruvate carboxykinase (PEPCK) enzyme. The binding induces a closure of the flap region to occlude the active site from bulk solvent [79].

In 1999, Tsai et al. [69] presented an alternative mechanism, selected fit based on the energy-landscape picture of protein folding. According to selected fit, a ligand selects and stabilizes a complementary protein conformation from the equilibrium of low-energy and higher-energy conformations. Consequently, this shifts the conformational density
towards the complex structure upon binding. Practically, support for selected fit in a specific PLI is based mainly on finding bound-like conformations of proteins in the corresponding unbound ensembles of structures. For example, in 2008, using nuclear magnetic resonance, Gsponer et al. [80] proposed that Ca\textsuperscript{2+}-bound calmodulin samples the conformational space of calmodulin bound to myosin light chain kinase.

In 2004, Grünberg et al. [81] presented a third mechanism in which both selected fit and induced fit mechanisms play a role in protein-ligand binding with an initial conformational selection step followed by induced fit rearrangements. They proposed that protein-ligand binding follows a three-step mechanism of diffusion, free conformer selection, and refolding. This three-step mechanism was also supported by the results of the study performed by Wlodarski and Zagrovic in 2009 [82]. The study involved an extensive structural analysis of a large set of unbound and bound ubiquitin conformers. The purpose of the study was to understand the impact of the induced fit mechanism on the residues after conformational selection in the protein-ligand binding process.

The selected fit and induced fit mechanisms can be represented by a four-state model of protein-ligand binding (Figure 6.5). In this model, the protein has two dominant conformations E\textsubscript{1} and E\textsubscript{2}, where E\textsubscript{1} is the ground state conformation and E\textsubscript{2} is the excited state conformation. On the induced fit route, E\textsubscript{1} \rightarrow E\textsubscript{1}L \rightarrow E\textsubscript{2}L, the protein first binds the ligand in conformation E\textsubscript{1}, which causes the transition into conformation E\textsubscript{2}. On the selected fit route, E\textsubscript{1} \rightarrow E\textsubscript{2} \rightarrow E\textsubscript{2}L, the protein binds the ligand in the higher-energy conformation E\textsubscript{2}. 

138
6.2.1 Kinetic Rate Model of Induced Fit Mechanism

The induced fit mechanism is illustrated by the route $E_1 \rightarrow E_1L \rightarrow E_2L$ as shown in Figure 6.5. The protein first binds the ligand in conformation $E_1$, which induces the transition into conformation $E_2$. Figure 6.6 depicts the association and dissociation processes of the induced fit kinetic rate model proposed by Weikl and von Deuster in 2008 [11]. According to this model, the induced fit on-rate per mole ligand is

$$r_{on} \approx r_b$$

Thus, the rate of association is limited by the diffusional rate of encounter complex formation of the proteins in their unbound conformational ensemble. Likewise, the induced fit off-rate is

$$r_{off} \approx r_{12}r_u / (r_{21} + r_u)$$

For fast conformational relaxation into the bound ground state with rate $r_{21} \gg r_u$, the induced fit off-rate is approximately

$$r_{off} \approx r_u / K_b$$

where $K_b = r_{21}/r_{12} = (E_2L) / (E_1L)$. Thus, the induced fit dissociation process is dependent on the conformational equilibrium between $E_2L$ and $E_1L$ complexes.
Figure 6.5 Four-state model of protein-ligand binding. Without the ligand $L$, $E_1$ and $E_2$ are the ground-state conformation and the excited-state conformation of protein respectively. When the ligand is bound, $E_2L$ is the ground state, and $E_1L$ is the excited-state.

\[ K_u = \frac{(E_2)}{(E_1)} \ll 1, \quad K_b = \frac{(E_2L)}{(E_1L)} \gg 1, \quad K_1 = \frac{(E_1L)}{(E_1)(L)}\text{ and } K_2 = \frac{(E_2L)}{(E_2)(L)}\]

---

Figure 6.6 Kinetic rate model of the induced fit mechanism proposed by Weikl and von Deuster. Here, $r_b$ is the binding rate of conformation $E_1$ per mole ligand, and $r_u$ is the unbinding rate. $r_{12}$ and $r_{21}$ are the rates for the conformational transitions in the bound state. $r_{on}$ and $r_{off}$ are the induced fit on-rate and off-rate respectively.
6.2.2 Kinetic Rate Model of Selected Fit Mechanism

The selected fit mechanism is illustrated by the route $E_1 \rightarrow E_2 \rightarrow E_2L$ as shown in Figure 6.5. The ligand binds the protein at the excited state ($E_2$) with higher energy conformation. The association and dissociation processes of the selected fit kinetic rate model proposed by Weikl and von Deuster is depicted in Figure 6.7.

Based on this model, the selected fit on-rate per mode ligand is

$$s_{on} \approx K_u s_b$$

where $K_u = (E_2)/(E_1)$. Hence, the selected fit association process is dependent on the equilibrium between the conformations of $E_1$ and $E_2$. Likewise, the selected fit off-rate is

$$s_{off} \approx s_u.$$

The selected fit dissociation process is identical with the rate $s_u$ for the bottleneck stage, the unbinding process from $E_2L$ to $E_2$.

![Kinetic rate model of the selected fit mechanism proposed by Weikl and von Deuster. Here $s_{21}$ and $s_{12}$ are the rates for the conformational transitions in the unbound state, $s_b$ is the binding rate of conformation $E_2$ per mole ligand, and $s_u$ is the unbinding rate. $s_{on}$ and $s_{off}$ are the selected fit on-rate and off-rate respectively.](image)

141
6.2.3 Kinetic Rate Model of Three-Step Mechanism

In 2004, Grünberg et al. revealed that both the selected fit and induced fit models play a role in receptor-ligand binding with an initial conformational selection step followed by induced fit rearrangements. They proposed a three-step mechanism of diffusion, free conformer selection, and refolding for receptor-ligand binding as shown in Figure 6.8 [81]. Receptor-ligand association begins with the diffusional encounter of the two free structure ensembles (Rf and Lf), which leads to a micro-collision with correct orientation of receptor and ligand (Rf•Lf). From the free conformation ensembles of receptor and ligand, specifically matching conformations will choose each other and form a recognition complex (Rf*Lf*). The recognition barrier (barrier II in Figure 6.8) departs from the energy of the encounter complex Rf•Lf by a loss of conformational entropy because it can only be crossed by a subset of free conformation ensembles of receptor and ligand. The recognition complex (Rf*Lf*) stabilized by desolvation as well as short-range electrostatic and van der Waals interactions transits into the native complex (RbLb) after passing over the refolding barrier (barrier III in Figure 6.8). If the free-energy cost of selecting matching conformers is much lower than the free-energy cost of finding the correct orientation (k1 « k2), the model will be reduced into a two-step mechanism including diffusion and refolding. On the other hand, if recognition requires bound conformers (k2 « k3), the refolding barrier would be absent.
Figure 6.8 Three-step process of diffusion, free conformer selection, and refolding. $R_f$ and $L_f$ are the free structure ensembles of receptor and ligand, respectively. $R_f^*$ and $L_f^*$ are subsets of the free receptor and ligand ensembles. $R_bL_b$ is the native complex.

6.3 Residence Time

Copeland et al. [83] defined the term residence time ($\tau$) as the period for which a receptor is occupied by a ligand in vivo. Residence time ($\tau$) is quantified through experimental measurements of the reciprocal of the dissociation rate constant ($\tau = 1/k_{off}$) or the dissociation half-life ($\tau_{1/2} = -\ln(0.5)/k_{off} = 0.693/k_{off}$).

There exists a common feature of ligand-protein binding in the induced fit mechanism: the closure of a flap region to occlude the protein active site from bulk solvent upon ligand binding. In this manner, a protein forms a lid over the ligand-bound pocket to block the escape trajectory of the ligand from the protein. This mechanism is seen, for example, upon ligand binding to a variety of kinases, HIV integrase, HIV protease, methionine adenosyltransferase, and Hepatitis NS3 protease [3].
Figure 6.9 shows the superposition of Darunavir-bound HIV-1 (PDB code: 2IEN) and two ligand-free HIV-1 structures (PDB code: 1HHP and 2G69). The two ligand-free structures exhibit opened flap conformations and the Darunavir-bound structure reveals closed flap conformation. As a result of the binding to Darunavir, the HIV-1 protein was induced to advance a conformational change that closed the flap region [78].

The duration of ligand-protein occupany is determined by the rate of ligand dissociation in vivo. Ligand dissociation is a unimolecular process dependent on the concentration of the binary complex and not on the concentrations of ligand and protein (Recall that in Section 6.2.1, $r_{off} \approx r_u / K_b$: the induced fit dissociation process is dependent on the conformational equilibrium between $E_2L$ and $E_1L$ complexes). Thus, any conformational change that must accompany ligand dissociation most likely occurs through the equivalent of a retrograde induced fit mechanism. Figure 6.10 depicts the reaction coordinate diagram of the escape trajectory for ligand dissociation following a retrograde induced fit mechanism. The retrograde mechanism requires the conversion of the closed flap conformation ($E_2L$) back to the opened flap conformation ($E_1L$) before the dissociation of the ligand and recovery of the free protein ($E_1$). As shown in Figure 6.11, both of these conversions force the system to overcome a significant energy barrier to transiently attain two sequential transition states: $E_2L^\dagger$ and $E_1L^\dagger$. Once the system has reached the $E_1L$ state, it can either overcome the $E_2L^\dagger$ transition state to return to $E_2L$ state or surmount the $E_1L^\dagger$ transition state to complete the ligand escape process. Thus, the residence time of a ligand-protein complex relates directly to the relative stabilities of
the E$_2$L and E$_1$L states, and hence the most optimal for prolonged residence time is to destabilize the E$_1$L state and/or stabilize the E$_2$L state.

Figure 6.9 Superposition of ligand-bound and ligand-free HIV-1 protease complexes.  
Red color: Darunavir-bound HIV-1 (2IEN).  
Green and blue color: ligand-free HIV-1 proteases. (1HHP, 2G69).  
The ligand-free structures exhibit opened flap conformation; Darunavir-HIV-1 complex reveals closed flap conformation.

Figure 6.10 Reaction coordinate diagram of the escape trajectory for ligand dissociation following a retrograde induced-fit mechanism
Figure 6.11 Free-energy diagram of the escape trajectory for ligand dissociation following a retrograde induced fit mechanism. The system begins at the $E_2L$ state and must surmount the energy barrier to attain the first transition state $E_2L^\dagger$ in order to reach the intermediate state $E_1L$. The system next must overcome another energy barrier to attain a second transition state $E_1L^\dagger$ before decaying to the unbound protein $E_1$.

6.4 Metadynamics

The application of MD simulations is limited by the time scales that can be routinely sampled. Recently, metadynamics has emerged as a powerful approach for accelerating rare events and computing multidimensional free energy surfaces. Metadynamics is not only used to improve sampling in molecular dynamics simulations of a system where ergodicity is hindered by the form of the system's energy landscape, but is also adopted as a powerful technique for reconstructing the free-energy surface as a function of a few selected degrees of freedom, referred to as collective variables (CVs). In metadynamics, an external history-dependent bias potential (a function of the CVs) is added to the Hamiltonian of the system to discourage the system from revisiting configurations that have already been sampled. This bias potential can be written as a
sum of Gaussian potentials deposited along the system trajectory in the CVs space.

Figure 6.12 shows the effect of bias potential on a one-dimensional potential system containing three local minima A, B, C [84]. The system is prepared in the local minimum B. In an MD simulation, the system would remain stuck in this minimum because barriers are larger than thermal fluctuations. In metadynamics simulation, as time goes by, Gaussian potentials are deposited causing the underlying bias potential to grow until, around \( t = 135 \) (\( t \) is the measurement of the number of Gaussian potential added), the system is pushed out of basin B into basin A. Here, the Gaussian potential accumulation starts again. The system is trapped in A until the underlying free-energy basin is completely filled at \( t = 430 \). Starting from \( t = 810 \), the system can easily access the region of C. Finally, when this basin is also compensated by the bias potential (\( t = 1650 \)), the system evolution resembles a random walk on the flattened free-energy surface.

The bias potential \( V_G \) provides an unbiased estimate of the underlying free energy

\[
V_G(S, t \rightarrow \infty) = -F(S) + C
\]

where

\( C \) is an irrelevant additive constant and \( S \) is a set of \( d \) functions of the microscopic coordinates \( R \) of the system: \( S(R) = S_1(R), \ldots, S_d(R) \). \( F(S) \) is the free energy and is defined as:

\[
F(S) = -(1/\beta) \ln(\int \delta(S - S(R))e^{-\beta U(R)})
\]

where

\( \beta = (k_B T)^{-1} \), \( k_B \) is the Boltzmann constant, \( T \) is the temperature of the system, and \( U(R) \) is the potential energy function.

At time \( t \), the metadynamics bias potential can be written as

\[
V_G(S, t) = \int_0^t dt' \omega \exp(-\sum_{i=1}^d ((S_i(R) - S_i(R(t'))^2)/2\sigma_i^2))
\]

where
σᵢ is the width of the Gaussian for the iᵗʰ CV. ω is an energy rate and \( \omega = W/\tau_G \) (\( W \) = Gaussian potential height and \( \tau_G \) = the deposition stride).

Recently, metadynamics simulation has been successfully applied in different fields. In 2009, Pfaendtner et al. studied the effect of different nucleotides (ATP, ADP-Pi, and ADP) on the conformational free-energy landscape of actin by means of all-atom MD simulations in explicit solvent and metadynamics [85]. In 2010, Limongelli et al. conducted metadynamics technique to simulate the full dissociation process of a highly potent and selective inhibitor, SC-558, in both COX-1 and COX-2 and discovered a previously unreported alternative binding mode in COX-2 explaining the time-depending inhibition exhibited by this class of inhibitors and consequently their long residence time inside this isoform [86]. In 2014, Tiwary et al. using the technique of metadynamics not only successfully discovered the unbinding pathways and the rate-limiting steps, but also accurately estimated the kinetic rate constants in the paradigmatic case of the trypsin-benzamidine system [87].
Figure 6.12 Example of metadynamics simulation in a one-dimensional model potential.
(Top) Time evolution of the collective variables during the simulation. $t$ is the number of Gaussians deposited.
(Bottom) Schematic representation of the deposit of Gaussian potentials on the system along the trajectory. Thick line is the underlying potential. The sum of the underlying potential and of the metadynamics bias is shown at different times (thin lines).
6.5 Analysis of the log$_{10}$k$_{on}$ and log$_{10}$k$_{off}$ Results

In this study, six ML algorithms in three different categories including classification, regression, and instance-based learning were used for the predictions of the kinetic rate constants, log$_{10}$k$_{on}$ and log$_{10}$k$_{off}$. The six algorithms comprise two multi-target algorithms including a binary-target random forest classification algorithm and a binary-target lasso regression algorithm, and four single-target algorithms including a random forest regression algorithm, an elastic net linear regression algorithm, a lasso linear regression algorithm and a k-nearest-neighbors instance-based learning algorithm.

6.5.1 Impact of Sample Distribution on log$_{10}$k$_{on}$ Prediction

Table 6.1 was built to show the distribution of the log$_{10}$k$_{on}$ predicted results with respect to the log$_{10}$k$_{on}$ values of the ligands. First, the mean %deviation of log$_{10}$k$_{on}$ prediction of each record was calculated for 42 predicted results in all the LOO cross-validation experiments. Second, the log$_{10}$k$_{on}$ axis ranging from 2.13 to 10.4 was divided into 10 bins with bin width of 0.827. Third, the values of mean %deviation were classified into one of the 10 bins based on the log$_{10}$k$_{on}$ values of ligands.

The k$_{on}$ values of A016 and B277 are 172 (log$_{10}$k$_{on}$ = 2.24) and 134 (log$_{10}$k$_{on}$ = 2.13), both of which are proximate to the lower detection limit of SPR Biosensor. Additionally, the k$_{on}$ values of DMP, B376, and A008 are 2.52 x 10$^{10}$ (log$_{10}$k$_{on}$ = 10.401), 2.05 x 10$^{8}$ (log$_{10}$k$_{on}$ = 8.311), and 7.06 x 10$^{9}$ (log$_{10}$k$_{on}$ = 9.849) respectively, and they are all beyond the upper detection limit of SPR Biosensor. Thus, if the predicted log$_{10}$k$_{on}$ results for these five ligands were ignored, the mean %deviation for 34 ligands is 11.19%. Comparing to the mean %deviation for all 39 ligands (21.21%), it is 10.02% more accuracy.
Figure 6.13 depicts the relationship between number of records and %deviation of $\log_{10}k_{on}$ prediction in each bin. The accuracy (the smaller the %deviation, the higher the accuracy) of $\log_{10}k_{on}$ prediction is proportional to the number of training records in the bin. Specifically, the average %deviation in the bin[5.438] with 16 training records is as low as 6.34%. In other words, the on-rate binding properties of PLI are significantly covered by the three principal training datasets: DS-PIE, DS-RMLR, and DS-RMRR.
<table>
<thead>
<tr>
<th>Bin Width</th>
<th>Log10kon</th>
<th>0.827</th>
<th>0.827</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.827</td>
<td>2.130</td>
<td>2.957</td>
<td>3.784</td>
</tr>
<tr>
<td>2.130</td>
<td>4.611</td>
<td>5.438</td>
<td>6.265</td>
</tr>
<tr>
<td>4.611</td>
<td>7.092</td>
<td>7.919</td>
<td>8.746</td>
</tr>
<tr>
<td>7.092</td>
<td>9.573</td>
<td>10.400</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ligand</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>B277</td>
<td>A038</td>
</tr>
<tr>
<td>182.73</td>
<td>A037</td>
</tr>
<tr>
<td>A016</td>
<td>B347</td>
</tr>
<tr>
<td>156.54</td>
<td>B412</td>
</tr>
<tr>
<td>B249</td>
<td>B440</td>
</tr>
<tr>
<td>20.47</td>
<td>U75875</td>
</tr>
<tr>
<td>B365</td>
<td>3.09</td>
</tr>
<tr>
<td>A024</td>
<td>19.14</td>
</tr>
<tr>
<td>6.47</td>
<td>14.77</td>
</tr>
<tr>
<td>A047</td>
<td>11.26</td>
</tr>
<tr>
<td>11.35</td>
<td>16.72</td>
</tr>
<tr>
<td>A023</td>
<td>7.37</td>
</tr>
<tr>
<td>7.69</td>
<td></td>
</tr>
<tr>
<td>A017</td>
<td>17.16</td>
</tr>
<tr>
<td>5.93</td>
<td></td>
</tr>
<tr>
<td>A015</td>
<td>8.92</td>
</tr>
<tr>
<td>4.45</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1 Distribution of log10kon predicted results over 10 bins with bin width 0.827 along log10kon axis ranging from 2.13 to 10.4. The values on the top row are the lower bound of the bin interval. Each cell comprises two components: ligand name and %deviation of log10kon prediction.
Figure 6.13 %deviation of log$_{10}$kon prediction versus number of records in bins. The numeric value inside bracket is the lower bound of the bin interval; the value on the left of bracket is the number of records in the bin.

6.5.2 Impact of the Excited State of Ligand Bound Complex on log$_{10}$k$_{off}$ Prediction

Using the same process that created Table 6.1 for log$_{10}$kon, Table 6.2 was created for log$_{10}$k$_{off}$. Table 6.2 contains ten bins with bin width = 0.556 covering log$_{10}$k$_{off}$ from 1.9206 to -3.644. The average %deviation of log$_{10}$k$_{off}$ prediction of twenty-seven ligands with k$_{off}$ ≤ 0.0697 is 37.82, and the average %deviation of log$_{10}$k$_{off}$ of twelve ligands with 0.0697 < k$_{off}$ ≤ 83.3 is 662.56. The overall average %deviation of log$_{10}$k$_{off}$ for thirty-nine ligands is 230.04. Compared to the results of log$_{10}$kon prediction (average %deviation = 11.19), the results of log$_{10}$k$_{off}$ prediction are unacceptable and indicate that deficiency exists in the ML models of k$_{off}$ prediction.
Figure 6.14 depicts the plot of %deviation of $\log_{10}k_{\text{off}}$ prediction versus $\log_{10}k_{\text{off}}$ (%deviation of A015 is out of scope (5492.8)). It reveals that the accuracy deteriorates as the dissociation rate constant ($k_{\text{off}}$) increases.

### 6.5.3 Analysis

The experimental results of $k_{\text{on}}$ and $k_{\text{off}}$ predictions obtained from this study, strongly support the association and dissociation processes of the induced fit kinetic rate model {Section 6.2.1}. First, the kinetic properties of the ligand bound HIV-1 complex at the ground state alone is sufficient to achieve acceptable accuracy for $\log_{10}k_{\text{on}}$ prediction, because the rate of association of the induced fit model is simply controlled by the diffusional rate of encounter complex formation of the proteins in their unbound conformational ensemble. Second, according to the induced fit model, the rate of dissociation is equivalent to:

$$\approx r_u * \frac{E_1L}{E_2L}$$

where

$r_u$ is the unbinding rate, and $E_2L$ and $E_1L$ are the ground state complex and the excited state complex respectively. The absence of the kinetic characteristics of $E_1L$ in the two kinetic principal datasets causes the inaccuracy of $\log_{10}k_{\text{off}}$ prediction. Third, the accuracy of $\log_{10}k_{\text{off}}$ prediction deteriorates as $k_{\text{off}}$ increases, because the rate of dissociation is directly proportional to the concentration of the excited state complex ($E_1L$).

Conclusively, in order to accurately predict the kinetic rate constants ($k_{\text{on}}$ and $k_{\text{off}}$) of the ligand-HIV-1 systems, it is significant to use the technique of metadynamics to study the kinetics of the systems, including the binding and unbinding pathways, the rate-limiting steps, the intermediates, the depths of the free energy basins, and specifically, the role of water molecules in the binding and unbinding processes.
Table 6.2 Distribution of log<sub>10</sub><i>k_{off}</i> prediction over 10 bins with bin width 0.5567 along log<sub>10</sub><i>k_{off}</i> axis ranging from 1.9206 to -3.644. The values on the top row are the lower bound of the bin interval. Each cell comprises of two components: ligand name and %deviation of log<sub>10</sub><i>k_{off}</i> prediction.

<table>
<thead>
<tr>
<th>log&lt;sub&gt;10&lt;/sub&gt;&lt;i&gt;k_{off}&lt;/i&gt;</th>
<th>Ligand Name</th>
<th>%deviation of log&lt;sub&gt;10&lt;/sub&gt;&lt;i&gt;k_{off}&lt;/i&gt; prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.64</td>
<td>Saq</td>
<td>45.06</td>
</tr>
<tr>
<td>-3.08</td>
<td>Ind</td>
<td>32.55</td>
</tr>
<tr>
<td>-2.52</td>
<td>B408</td>
<td>53.1</td>
</tr>
<tr>
<td>-1.97</td>
<td>B369</td>
<td>24.4</td>
</tr>
<tr>
<td>-1.41</td>
<td>A030</td>
<td>45.81</td>
</tr>
<tr>
<td>-0.86</td>
<td>A023</td>
<td>102.48</td>
</tr>
<tr>
<td>-0.30</td>
<td>A015</td>
<td>5492.82</td>
</tr>
<tr>
<td>0.25</td>
<td>B376</td>
<td>250.62</td>
</tr>
<tr>
<td>0.80</td>
<td>A008</td>
<td>175.66</td>
</tr>
<tr>
<td>1.36</td>
<td>DMP</td>
<td>166.21</td>
</tr>
<tr>
<td>1.92</td>
<td>B440</td>
<td>53.93</td>
</tr>
<tr>
<td>2.52</td>
<td>B439</td>
<td>24.23</td>
</tr>
<tr>
<td>3.08</td>
<td>Rit</td>
<td>31.48</td>
</tr>
<tr>
<td>3.64</td>
<td>B388</td>
<td>23.97</td>
</tr>
<tr>
<td>4.15</td>
<td>A016</td>
<td>47.49</td>
</tr>
<tr>
<td>4.62</td>
<td>A017</td>
<td>102.75</td>
</tr>
<tr>
<td>5.15</td>
<td>A037</td>
<td>50.6</td>
</tr>
<tr>
<td>5.62</td>
<td>B268</td>
<td>49.4</td>
</tr>
<tr>
<td>6.15</td>
<td>B347</td>
<td>22.43</td>
</tr>
<tr>
<td>6.62</td>
<td>B322</td>
<td>22.42</td>
</tr>
<tr>
<td>7.15</td>
<td>B425</td>
<td>200.51</td>
</tr>
<tr>
<td>7.62</td>
<td>B429</td>
<td>40.27</td>
</tr>
<tr>
<td>8.15</td>
<td>B277</td>
<td>47.22</td>
</tr>
<tr>
<td>8.62</td>
<td>A021</td>
<td>14.99</td>
</tr>
<tr>
<td>9.15</td>
<td>A024</td>
<td>33.72</td>
</tr>
<tr>
<td>9.62</td>
<td>A045</td>
<td>250.89</td>
</tr>
<tr>
<td>10.15</td>
<td>B409</td>
<td>52.02</td>
</tr>
<tr>
<td>10.62</td>
<td>Amp</td>
<td>53.26</td>
</tr>
<tr>
<td>11.15</td>
<td>B365</td>
<td>24.4</td>
</tr>
<tr>
<td>11.62</td>
<td>A047</td>
<td>27.42</td>
</tr>
<tr>
<td>12.15</td>
<td>B249</td>
<td>166.06</td>
</tr>
<tr>
<td>12.62</td>
<td>A038</td>
<td>41.78</td>
</tr>
<tr>
<td>13.15</td>
<td>U75875</td>
<td>24.45</td>
</tr>
<tr>
<td>13.62</td>
<td>B355</td>
<td>358.12</td>
</tr>
<tr>
<td>14.15</td>
<td>Nel</td>
<td>74.5</td>
</tr>
<tr>
<td>14.62</td>
<td>B435</td>
<td>26</td>
</tr>
<tr>
<td>15.15</td>
<td>B295</td>
<td>356.98</td>
</tr>
<tr>
<td>15.62</td>
<td>33.86</td>
<td>327.73</td>
</tr>
</tbody>
</table>

Figure 6.14 %deviation of log<sub>10</sub><i>k_{off}</i> prediction versus log<sub>10</sub><i>k_{off}</i>. %deviation of A015 is too large and is depicted in red color.
6.6 Results of Feature Selection

In order to evaluate the significance of the impacts of the residues in the datasets F-PIE, F-RMLR, and F-RMRR to PLI, the residues were assigned into four classes: active site, flap region, dimmer stability and PIRM residue.

(A) F-RMLR (Figure 6.15): Table 6.3 shows the results of assignment for the residues in F-RMLR. Significant features revealed from the table are summarized as follows:

1. Among the thirteen residues in the active site, five of them are located in the flap region. A52 and L10 are adjacent to but excluded from the active site.

2. The impact of F53 covers the four classes. It is not only able to directly interact with ligand in the active site, but also plays a role in the closure of the flap region upon ligand binding. Its mutant, F53L, affects the dimmer stability of HIV-1 [78] and is also one of the 26 major PIRM.

3. I47V/A, G48V/M, L23I, D30N, V32I, and L10F are PIRMs.

(B) F-RMRR: Except for I50, fourteen among the fifteen residues in F-RMRR are common to F-RMLR and F-RMRR. The fourteen common residues are R8, L10, L23, D25, G27, A28, D29, D30, V32, K45, I47, G48, G49, and A52. I50 is an active site residue located in the flap region. I50L/V is a PIRM residue.

(C) F-PIE: Except for L76 and P81 (Figure 6.16), eleven among the thirteen residues in F-PIE are common to F-RMLR, F-RMRR, and F-PIE. The eleven common residues are R8, L10, L23, D25, G27, A28, D29, D30, V32, K45, and A52. Table 6.4 lists the common residues shared by DS-RMLR, DS-RMRR, and DS-PIE.
Summary

- The fourteen residues common to both F-RMLR and F-RMRR are significant to the two kinetic properties of the ligand-bound HIV-1 complex at ground state. 1st kinetic property: the relative directionality of normal modes between ligand and residue. 2nd kinetic property: the change of directionality of normal modes of binding site residues upon the ligand binding.
- F53 is only significant to the 1st kinetic property.
- I50 is only significant to the 2nd kinetic property.
- The eleven residues common to F-RMLR, F-RMRR and F-PIE are significant to both the two kinetic properties and the thermal dynamic characteristic- pairwise energy interaction of a ligand-residue pair.

<table>
<thead>
<tr>
<th>F-RMLR</th>
<th>Active Site</th>
<th>Flap Region</th>
<th>Dimmer Stability</th>
<th>PIRM</th>
</tr>
</thead>
<tbody>
<tr>
<td>F53</td>
<td>√</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>I47</td>
<td>√</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>G48</td>
<td>√</td>
<td>√</td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>L23</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D30</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>V32</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K45</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G49</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R8</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D25</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G27</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A28</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D29</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A52</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3 Results of assignment for the residues in F-RMLR.
Figure 6.15 Residues of F-RMLR. Beads represent the 15 residues of chain A in F-RMLR. Gray/green NewRibbon represents the chain A/B of HIV-1 protease.

Figure 6.16 L76 and P81 in the HIV-1 active site. L76 and P81 are in red and green CPK respectively.
<table>
<thead>
<tr>
<th>F-RMLR</th>
<th>F-RMRR</th>
<th>F-PIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>R8</td>
<td>R8</td>
<td>R8</td>
</tr>
<tr>
<td>L10</td>
<td>L10</td>
<td>L10</td>
</tr>
<tr>
<td>L23</td>
<td>L23</td>
<td>L23</td>
</tr>
<tr>
<td>D25</td>
<td>D25</td>
<td>D25</td>
</tr>
<tr>
<td>G27</td>
<td>G27</td>
<td>G27</td>
</tr>
<tr>
<td>A28</td>
<td>A28</td>
<td>A28</td>
</tr>
<tr>
<td>D29</td>
<td>D29</td>
<td>D29</td>
</tr>
<tr>
<td>D30</td>
<td>D30</td>
<td>D30</td>
</tr>
<tr>
<td>V32</td>
<td>V32</td>
<td>V32</td>
</tr>
<tr>
<td>K45</td>
<td>K45</td>
<td>K45</td>
</tr>
<tr>
<td>A52</td>
<td>A52</td>
<td>A52</td>
</tr>
<tr>
<td>I47</td>
<td>I47</td>
<td>L76</td>
</tr>
<tr>
<td>G48</td>
<td>G48</td>
<td>P81</td>
</tr>
<tr>
<td>G49</td>
<td>G49</td>
<td></td>
</tr>
<tr>
<td>F53</td>
<td>I50</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4 Common residues shared by F-RMLR, F-RMRR, and F-PIE. The fourteen residues common to F-RMLR and F-RMRR are in red. The eleven residues common to F-RMLR, F-RMRR, and F-PIE are in green.

6.7 Molecular Determinants of Ligand Binding and Unbinding

Kinetics

Several recent studies have suggested that multiple factors determine the ligand binding and unbinding kinetics. They include the size and flexibility of ligand [4,88], conformational flexibility of receptor [89,90], electrostatic interactions between the ligand and the receptor [91,92], water effect on the hydrophobic interaction and hydrogen-bonding [93,94], and conformational ensembles of protein-ligand complex [71,73,81].

- Size and flexibility of ligand: A survey of over 2000 drugs binding to G-protein-coupled receptors, protein kinases and other enzymes found that higher molecular weight
drugs tend to have lower off-rates, and a greater number of rotatable bonds tend to have longer residence time [4,89].

In this study, as shown in Figure 6.17, the molecular weights of the 39 ligands make no significant differences on the potency of $k_{on}$ and $k_{off}$ determination, but as shown in Figure 6.18, the values of $\log_{10}k_{on}$ and $\log_{10}k_{off}$ decrease as the number of the rotatable bond of ligand increases.

![Figure 6.17 Ligand molecular weight versus log\(_{10}\)kon/log\(_{10}\)koff](image)

Figure 6.17 Ligand molecular weight versus log\(_{10}\)kon/log\(_{10}\)koff
Conformational flexibility of receptor: The binding of a ligand to a protein requires shape and property complementarity. In the course of binding they have to adapt to each other in order to achieve a successful recognition process. For example, binding of ligands to InhA converses a disordered loop into an ordered α helix at the active site; this large conformational change impact the rate of inhibitor binding and unbinding significantly [95].

All-atom MD simulations have shown the conformational dynamics of flap region plays a key role in the binding process of HIV-1 protease [76,77]. Consistent with this observation, the statistical analysis of the clustering results [Figure 4.13, Table 4.7] in this study reveals that upon ligand binding, ten HIV-1 residues including G52, F53, I50, G49, K45, G48, P81, I54, T80, and I47 on both chains A and B, conduct significant
conformational displacements with total normalized NTAVs ≥ 1.15 Å². Additionally, 8 residues (in red color) among them are located in the flap region [Figure 6.19].

Figure 6.19 Ten active site residues (K45, I47, G48, G49, I50, G52, F53, I54, T80, P81) conduct significant conformational displacements upon ligand binding. The eight residues in red are located in the flap region (red bead). At the center is the ligand, Ritonavir (gray cpk) (PDB code: 1HXW).

In addition, multiple evidence from statistical analysis of this study suggest that several residues that are directly involved in the ligand-receptor interactions may be as important as the flap region. Interestingly, the directionality of ligand binding site residue movement has stronger corrections with the kinetic constants than the intensity of the ligand-receptor interactions. The coherent movement between the ligand and the receptor may play a critical role in determining the ligand binding and unbinding kinetics.

As shown in Figure 6.20, even when two protein-ligand complexes have the same non-covalent interactions with the same intensity (e.g. the same binding affinity), they may
have different kinetic constants due to the different relative movements between the ligand and the receptor.

\[ \text{Figure 6.20 Ligand-receptor interaction} \]

(A) Attractive interaction between ligand I and receptor A. The dot product of the two eigenvectors is:

\[ \text{Vector}_{\text{ligand I}} \cdot \text{Vector}_{\text{receptor A}} \cos \theta = \text{Vector}_{\text{ligand I}} \cdot \text{Vector}_{\text{receptor A}} \text{ with } \theta = 0. \]

(B) Repulsive interaction between ligand II and receptor B. The dot product of the two eigenvectors is:

\[ \text{Vector}_{\text{ligand II}} \cdot \text{Vector}_{\text{receptor B}} \cos \theta = \text{Vector}_{\text{ligand II}} \cdot \text{Vector}_{\text{receptor B}} \text{ with } \theta = 180. \]
Electrostatic interactions between the ligand and the receptor: Electrostatic interactions between a charged drug and a charged receptor impact the kinetic rate constants. Specifically, $k_{\text{on}}$ is sensitive to long-range electrostatic interaction, and $k_{\text{off}}$ tends to be influenced more by short-range interactions such as hydrogen bonds, salt bridges and van der Waals contacts [96].

Table 6.5 lists the molecular characteristics of the 22 residues selected in the SASA process. The majority of the residues are hydrophobic; the exceptions are the catalytic D25 and D29, which are able to form hydrogen bonds with the main chain groups of substrate peptides, and R8, D30 and K45 which can interact with polar side chains or distal main chain groups in longer substrate peptides.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Side-chain polarity</th>
<th>Side-chain charge (pH 7.4)</th>
<th>Hydropathy index</th>
</tr>
</thead>
<tbody>
<tr>
<td>R8</td>
<td>basic polar</td>
<td>positive</td>
<td>-4.5</td>
</tr>
<tr>
<td>L10</td>
<td>nonpolar</td>
<td>neutral</td>
<td>3.8</td>
</tr>
<tr>
<td>L23</td>
<td>nonpolar</td>
<td>neutral</td>
<td>3.8</td>
</tr>
<tr>
<td>D25</td>
<td>acidic polar</td>
<td>negative</td>
<td>-3.5</td>
</tr>
<tr>
<td>G27</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-0.4</td>
</tr>
<tr>
<td>A28</td>
<td>nonpolar</td>
<td>neutral</td>
<td>1.8</td>
</tr>
<tr>
<td>D29</td>
<td>acidic polar</td>
<td>negative</td>
<td>-3.5</td>
</tr>
<tr>
<td>D30</td>
<td>acidic polar</td>
<td>negative</td>
<td>-3.5</td>
</tr>
<tr>
<td>V32</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.2</td>
</tr>
<tr>
<td>K45</td>
<td>basic polar</td>
<td>positive</td>
<td>-3.9</td>
</tr>
<tr>
<td>I47</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.5</td>
</tr>
<tr>
<td>G48</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-0.4</td>
</tr>
<tr>
<td>G49</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-0.4</td>
</tr>
<tr>
<td>I50</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.5</td>
</tr>
<tr>
<td>G52</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-0.4</td>
</tr>
<tr>
<td>F53</td>
<td>nonpolar</td>
<td>neutral</td>
<td>2.8</td>
</tr>
<tr>
<td>I54</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.5</td>
</tr>
<tr>
<td>I76</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.5</td>
</tr>
<tr>
<td>T80</td>
<td>polar</td>
<td>neutral</td>
<td>-0.7</td>
</tr>
<tr>
<td>P81</td>
<td>nonpolar</td>
<td>neutral</td>
<td>-1.6</td>
</tr>
<tr>
<td>V82</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.2</td>
</tr>
<tr>
<td>I84</td>
<td>nonpolar</td>
<td>neutral</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Table 6.5 Residue molecular properties
Water effect on the hydrophobic interaction and hydrogen-bonding: In 2011, Schmidtke et al. reported that when a ligand and a receptor interact via hydrogen bonds shielded from water by surrounding hydrophobic regions, the resulting complex tends to be more kinetically stable than if the hydrogen bonds were less shielded [97].

DMP4550 is a cyclic urea inhibitor developed by the Dupont Merck group. The four benzene rings could help shield the hydrogen bonds from water by surrounding hydrophobic regions (Figure 6.21) [80].

Figure 6.21 (A) Structure of cyclic urea inhibitor. (B) 2D structure of DMP4550. (C) Schematic interaction of a cyclic urea based inhibitor with HIV-1 protease.
Conformational ensembles of protein-ligand complex: There are three different kinetic rate models of ligand-protein interaction. They are the model of induced fit mechanism (section 6.2.1), the model of selected fit mechanism (section 6.2.2), and the model of three step mechanism (section 6.2.3). The mechanism of the model determines the on-rate and off-rate equations. For example, the induced fit on-rate ($r_{on} \approx r_b$) is limited by the diffusional rate of encounter complex formation of the proteins in their unbound conformational ensemble, but the off-rate ($r_{off} \approx r_a (E_1 L)/(E_2 L)$) is dependent on the equilibrium between the ground state complex ($E_2 L$) and the excited state complex ($E_1 L$). Because the training datasets DS-RMLR and DS-RMRR only cover the kinetic characteristics of the ground state complex, but omit the kinetic properties of the excited state complex, in the ML experiments for the predictions of the two kinetic rate constants, only the results of $\log_{10} k_{on}$ prediction is accurate.
Chapter 7

Summary, Conclusion and Future Work

7.1 Summary

The predicted results of $\log_{10}k_{on}$ and $\log_{10}k_{off}$ for classification and regression, as well as the results of feature selection and Welch’s t-tests are listed as follows:

7.1.1 Machine Learning Prediction

Tables 7.1 and 7.2 summarize the $\log_{10}k_{on}$ and $\log_{10}k_{off}$ results for classification and regression respectively.

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Target</th>
<th>Average</th>
<th>Minimum</th>
<th>Maximum</th>
<th>stdev</th>
<th>MM-Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-PIE</td>
<td>$\log_{10}k_{on}$</td>
<td>62.50</td>
<td>53.85</td>
<td>69.23</td>
<td>4.73</td>
<td>90.45</td>
</tr>
<tr>
<td></td>
<td>$\log_{10}k_{off}$</td>
<td>65.38</td>
<td>64.10</td>
<td>71.79</td>
<td>3.06</td>
<td></td>
</tr>
<tr>
<td>DS-RMLR</td>
<td>$\log_{10}k_{on}$</td>
<td>63.78</td>
<td>51.28</td>
<td>69.23</td>
<td>5.72</td>
<td>90.20</td>
</tr>
<tr>
<td></td>
<td>$\log_{10}k_{off}$</td>
<td>63.78</td>
<td>46.15</td>
<td>71.79</td>
<td>9.64</td>
<td></td>
</tr>
<tr>
<td>DS-RMRR</td>
<td>$\log_{10}k_{on}$</td>
<td>45.83</td>
<td>38.46</td>
<td>51.28</td>
<td>4.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\log_{10}k_{off}$</td>
<td>60.90</td>
<td>48.72</td>
<td>64.10</td>
<td>5.08</td>
<td>76.22</td>
</tr>
<tr>
<td>DS-PIE+DS-RMLR</td>
<td>$\log_{10}k_{on}$</td>
<td>63.78</td>
<td>53.85</td>
<td>69.23</td>
<td>5.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\log_{10}k_{off}$</td>
<td>58.97</td>
<td>35.90</td>
<td>71.79</td>
<td>10.61</td>
<td>86.87</td>
</tr>
<tr>
<td>DS-RMLR+DS-RMRR</td>
<td>$\log_{10}k_{on}$</td>
<td>61.91</td>
<td>51.72</td>
<td>69.23</td>
<td>6.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\log_{10}k_{off}$</td>
<td>68.59</td>
<td>56.41</td>
<td>71.79</td>
<td>5.44</td>
<td>92.40</td>
</tr>
<tr>
<td>DS-PIE+DS-RMLR+DS-RMRR</td>
<td>$\log_{10}k_{on}$</td>
<td>65.38</td>
<td>58.97</td>
<td>69.23</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\log_{10}k_{off}$</td>
<td>64.10</td>
<td>58.97</td>
<td>66.67</td>
<td>2.38</td>
<td>91.57</td>
</tr>
</tbody>
</table>

Table 7.1 Summary of classification prediction
As shown in Figure 7.1, the classifiers trained with DS-RMLR+DS-RMRR produced the largest MM-Accuracy (92.40) with average accuracy of log_{10}k_{off} = 68.59% and average accuracy of log_{10}k_{on} = 61.91%.

For the regression predictions of log_{10}k_{on}, the most accurate prediction (%deviation = 19.03) was given by the single-target random forest regressor trained with DS-RMLR. In fact, the kinetic training datasets, DS-RMLR, and DS-RMRR, outperformed the thermal dynamic training dataset, DS-PIE, in all regression predictions. Specifically, DS-RMLR was even better than DS-RMRR in the log_{10}k_{on} predictions. Thus, the relative movement of a residue-ligand pair is the optimal training feature in k_{on} prediction (Table 7.2).

The predicted results of log_{10}k_{off} are unacceptable. The lowest error (%deviation = 130.81) given by the single-target K-nearest neighbors model trained with DS-RMLR is still too high to be accepted (Table 7.2).

### Table 7.2 Summary of regression prediction

<table>
<thead>
<tr>
<th></th>
<th>Regression</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-target</td>
<td></td>
<td></td>
<td></td>
<td>Binary-target</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elastic Net</td>
<td>Lasso</td>
<td>Random Forest</td>
<td>KNN</td>
</tr>
<tr>
<td>log_{10}k_{off}</td>
<td>minimum</td>
<td>181.29</td>
<td>224.59</td>
<td>184.96</td>
<td>130.81</td>
</tr>
<tr>
<td></td>
<td>maximum</td>
<td>276.93</td>
<td>251.47</td>
<td>226.29</td>
<td>286.03</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>246.63</td>
<td>239.38</td>
<td>206.94</td>
<td>215.95</td>
</tr>
<tr>
<td></td>
<td>stdev</td>
<td>37.43</td>
<td>11.16</td>
<td>16.65</td>
<td>64.04</td>
</tr>
<tr>
<td>log_{10}k_{on}</td>
<td>minimum</td>
<td>19.84</td>
<td>19.92</td>
<td>19.03</td>
<td>20.19</td>
</tr>
<tr>
<td></td>
<td>maximum</td>
<td>22.31</td>
<td>20.82</td>
<td>25.49</td>
<td>25.43</td>
</tr>
<tr>
<td></td>
<td>average</td>
<td>20.59</td>
<td>20.32</td>
<td>22.09</td>
<td>22.53</td>
</tr>
<tr>
<td></td>
<td>stdev</td>
<td>0.8</td>
<td>0.47</td>
<td>2.35</td>
<td>2.34</td>
</tr>
<tr>
<td>best performed training dataset</td>
<td>[DS-RMLR]</td>
<td>{DS-RMRR}</td>
<td>{DS-PIE+DS-RMLR+DS-RMRR}</td>
<td>{DS-PIE+DS-RMLR+DS-RMRR}</td>
<td>[DS-RMRR]</td>
</tr>
</tbody>
</table>
7.1.2 Feature Selection, Welch’s t-test and PIRM

Table 7.3 summaries the results of the feature selection, the Welch’s t-test and PIRM.

<table>
<thead>
<tr>
<th>Feature Selection</th>
<th>Welch's t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-value &lt; 0.05</td>
</tr>
<tr>
<td>F-RMLR</td>
<td>✓</td>
</tr>
<tr>
<td>F-RMRR</td>
<td>✓</td>
</tr>
<tr>
<td>F-PIE</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>p-value &lt; 0.05</td>
</tr>
<tr>
<td>R8</td>
<td>✓</td>
</tr>
<tr>
<td>L10</td>
<td>✓</td>
</tr>
<tr>
<td>L23</td>
<td>✓</td>
</tr>
<tr>
<td>D25</td>
<td>✓</td>
</tr>
<tr>
<td>G27</td>
<td>✓</td>
</tr>
<tr>
<td>A28</td>
<td>✓</td>
</tr>
<tr>
<td>D29</td>
<td>✓</td>
</tr>
<tr>
<td>D30</td>
<td>✓</td>
</tr>
<tr>
<td>V32</td>
<td>✓</td>
</tr>
<tr>
<td>K45</td>
<td>✓</td>
</tr>
<tr>
<td>A52</td>
<td>✓</td>
</tr>
<tr>
<td>I47</td>
<td>✓</td>
</tr>
<tr>
<td>G48</td>
<td>✓</td>
</tr>
<tr>
<td>G49</td>
<td>✓</td>
</tr>
<tr>
<td>I50</td>
<td>✓</td>
</tr>
<tr>
<td>F53</td>
<td>✓</td>
</tr>
<tr>
<td>L76</td>
<td>✓</td>
</tr>
<tr>
<td>P81</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 7.3 Summaries of the results of feature selection, Welch’s t-test and PIRM. The values under the Welch’s t-test column are the number of occurrence of p-value < 0.05 in the Welch’s t-test.

- Among the eighteen residues selected from the feature selection process, eleven residues are common to F-PIE, F-RMLR, and F-RMRR indicating that they are significant to both the kinetic and thermal dynamic properties of the ligand-bound HIV-1 complex at the ground state. Specifically, they all have significant impact on the kinetic characteristics of the coupling moving of residue-ligand pair and the coupling movement of a residue upon ligand binding, and the thermal dynamic characteristic of pairwise
interaction energy. Moreover, all the eleven residues except G27, have a p-value < 0.05 in the Welch’s t-test revealing that as the kinetic rate constants vary, these ten residues could alter their roles in PLI.

- I47, G48, and G49 are common to F-RMLR and F-RMRR. They are significant to the kinetic properties of the coupling movement of residue-ligand pair and the coupling movement of a residue upon ligand binding.
- F53 belongs to F-RMLR exclusively. It is only significant to the kinetic property of the coupling movement of residue-ligand pair.
- I50 belongs to F-RMRR exclusively. It is only significant to the kinetic property of the coupling movement of a residue upon ligand binding,
- L76 and P81 belong to F-PIE exclusively. They are only significant to the thermal dynamic characteristic of pairwise interaction energy.
- L10, L23, V32, I47, G48, I50, F53, and L76 are the PIRM residues.
- As shown in Figure 2.3, HIV-1 protease recognizes peptides of six residues from P3 to P3’ and catalyzes the hydrolysis of the P1-P1’. The peptide is bound between the active site residues 25-29 and two flaps by means of hydrogen bonds. Specifically, D25 and D29 form hydrogen bonds with peptide. The D25 – peptide hydrogen bond is as strong as ≥ 4kcal/mol.

**Area Under ROC curve**

The areas under the ROC curves shown in Figure 5.5 (Section 5.9.2) reveal that F-RMLR is better able than F-RMRR and F-PIE to distinguish PIRM residues from non-PIRM residues. Thus, the kinetic property of the coupling movement of residue-ligand
pair outperforms the thermal dynamic property of pairwise interaction energy of ligand-residue pair in its ability to differentiate between PIRM residues and non-PIRM residues.

**Welch’s t-test**

Twenty-two residues (forty-four for two chains) were selected from the SASA process in Section 2.2. Table 7.4 shows the probability of the twenty-two residues having p-value < 0.05 in the Welch’s T-test. Eight residues including D29, L10, L23, R8, D25, A28, D30, and V32, are significant to PLI for the following two reasons. First, the probability (≥ 16.67) of these residues having p-value < 0.05 is high. Second, they were also selected in the feature selection process (Section 5.8.2).

<table>
<thead>
<tr>
<th>Residue</th>
<th>Probability %</th>
</tr>
</thead>
<tbody>
<tr>
<td>D29</td>
<td>50.00</td>
</tr>
<tr>
<td>L10</td>
<td>27.78</td>
</tr>
<tr>
<td>L23</td>
<td>27.78</td>
</tr>
<tr>
<td>R8</td>
<td>22.22</td>
</tr>
<tr>
<td>D25</td>
<td>22.22</td>
</tr>
<tr>
<td>A28</td>
<td>22.22</td>
</tr>
<tr>
<td>D30</td>
<td>22.22</td>
</tr>
<tr>
<td>V32</td>
<td>16.67</td>
</tr>
<tr>
<td>V82</td>
<td>11.11</td>
</tr>
<tr>
<td>I84</td>
<td>11.11</td>
</tr>
<tr>
<td>G49</td>
<td>5.56</td>
</tr>
<tr>
<td>I50</td>
<td>5.56</td>
</tr>
<tr>
<td>G52</td>
<td>5.56</td>
</tr>
<tr>
<td>F53</td>
<td>5.56</td>
</tr>
<tr>
<td>P81</td>
<td>5.56</td>
</tr>
<tr>
<td>G27</td>
<td>0.00</td>
</tr>
<tr>
<td>K45</td>
<td>0.00</td>
</tr>
<tr>
<td>I47</td>
<td>0.00</td>
</tr>
<tr>
<td>G48</td>
<td>0.00</td>
</tr>
<tr>
<td>I54</td>
<td>0.00</td>
</tr>
<tr>
<td>I76</td>
<td>0.00</td>
</tr>
<tr>
<td>T80</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 7.4 Probability of having p-value < 0.05 in the Welch’s T-Test. Residues in green color are the PIRM residues.
**Clustering**

The statistical analysis of the results given by K-Means cluster algorithm with \( k = 6 \) [Figure 4.13, Table 4.7] reveals that upon ligand binding, ten HIV-1 residues including G52, F53, I50, G49, K45, G48, P81, I54, T80, and I47 on both chains A and B, conduct large conformational displacements with the total normalized NTAVs ≥ 1.15 Å². Additionally, 8 residues (in red color) among them are located in the flap region [Figure 6.19].

**7.2 Conclusion**

The major findings of this thesis include:

- Acceptable accuracy of \( \log_{10}k_{\text{on}} \) prediction. Average %deviation is 11.19 and reduces to 6.34% if enough training records are provided {Section 6.5.1}.
- Multi-target classification and multi-target regression are potentially valuable tools for modeling PLI. Multi-target random forest classification algorithm yielded accuracy of 68.59% for \( \log_{10}k_{\text{off}} \) prediction and 61.91% for \( \log_{10}k_{\text{on}} \) prediction. In addition, there exists no significant difference between the MM-%deviations given by the multi-target lasso linear regression algorithm and single-target lasso linear regression algorithm.
- NMA is an efficient method to capture conformational dynamics features for the large scale modeling of protein-ligand binding. As supported by the results of \( \log_{10}k_{\text{on}} \) prediction, coherent conformational dynamics coupling between protein and ligand were proven to be more significant than pairwise binding energy between ligand and residue in the predictions of kinetic rate constants. Hence, kinetic characteristics are more important than thermal dynamic properties in determining protein-ligand binding kinetics.
R8, L10, L23, D25, G27, A28, D29, D30, V32, K45, A52, I47, G48, G49, I50, F53, L76, and P81 are the eighteen residues selected from the feature selection process and significant to PLI. Among them, 50% are PIRM residues including L10, L23, D30, V32, I47, G48, I50, F53, and L76. The eleven residues in green are common to F-PIE, F-RMLR, and F-RMRR indicating that they are significant to both the kinetic properties of the coupling movement of a residue-ligand pair and the coupling movement of a residue upon ligand binding, and the thermal dynamic property of pairwise interaction energy between ligand and residue.

Eight residues in the flap region, including K45, I47, G48, G49, I50, G52, F53, and I54, and two residues in the loop, including T80 and P81, on both chains A and B, conduct significant conformational displacements with total normalized NTAVs ≥ 1.15 Å² upon ligand binding [Figure 6.19].

There is also a deficiency in the model. As revealed by induced fit mechanism, \( k_{off} \) is regulated by the unbinding rate of the excited state of ligand bound complex, and the concentrations of the ground state and the excited state of the ligand bound complexes. Missing the kinetic properties of the excited state of the ligand bound complex cause large deviations in the prediction of \( k_{off} \). Moreover, the accuracy of \( k_{off} \) prediction deteriorates as the value of \( k_{off} \) increases.
7.3 Future Work

7.3.1 Metadynamics Simulation for the Kinetic Role of Water

HIV-1 protease inhibitors were designed to mimic the transition state of the protease’s peptide substrate. The peptide linkage (-NH-CO-) in the substrate is replaced by a nonhydrolyzable moiety in the inhibitor such as hydroxyethylene, hydroxyethylamine, ketoamide or phosphonamide (Figure 2.3).

Depending on the structure of the inhibitor, water molecules play different roles in PLI. As shown in Figure 7.1, an inhibitor with α-keto amide core structure binds to HIV-1 protease in its hydrated form after hydration. Moreover, a water molecule near the flaps forms four hydrogen bonds with the two I50 residues on the flaps and the two carbonyl oxygen atoms on the inhibitor [77].

DMP4550 is a cyclic urea inhibitor developed by the Dupont Merick group. The cyclic urea core replaces the flap water molecule and could lead to better binding energy due to the positive entropic effect that is provided by the cyclic urea core. Additionally, the inhibitor also contains the diol functionality as a transition state mimic to interact with the catalytic aspartates (Figure 6.21) [77].

In 2014, Tiwary et al. studied the unbinding of the inhibitor bezamidine from trypsin, a serine protease protein using metadynamics and found that water molecules played a significant role in the unbinding process of PLI. Specifically, the solvent promotes unbinding by facilitating the breakage of shielded hydrogen bonds through the formation of water bridge interactions [87,88].

Obviously, the role of water molecules is significant in the ligand-HIV-1 protease interaction and worth studying with metadynamics simulation. Additionally, the feature
selection process and the Welch’s t-test of this study as well as the PIRM information reported by the World Health Organization reveal the significant residues in the ligand-HIV-1 interaction, and thus provide the choice of CVs in metadynamics simulation for the evaluation of the role of water molecules in the interaction.

7.3.2 $k_{on}$ and $k_{off}$ Prediction

In this study, the result of the log$_{10}k_{off}$ prediction from the regression experiment was unacceptable. As the value of $k_{off}$ rises, the %deviation of log$_{10}k_{off}$ prediction deteriorates. On the other hand, the result of log$_{10}k_{on}$ prediction from the regression experiment is acceptable, but there is still room for improvement. Four major elements should be included in the future work to improve accuracy.

- The characteristics of the excited state of ligand bound complex:

  Ligand dissociation ($k_{off}$) in vivo (residence time) depends on the equilibrium between the ground state (E$_2$L) and the excited state (E$_1$L) of the ligand bound complex. Thus, any conformational changes that must accompany ligand dissociation most likely occur through the equivalent of a retrograde induced fit mechanism. Ligand binding induces the closure of the flap lid of the ligand bound HIV-1 protease (E$_2$L), which traps the ligand in the active site. Following the retrograde mechanism, re-opening of the flap lid is the required conformational change (E$_1$L) for the escape of the ligand to bulk solvent. Figures 6.10 and 6.11 illustrate the conformational changes required to open an escape trajectory to bulk solvent for the bound ligand. As shown in Figures 6.6 and 6.11, $k_{off}$ depends on the equilibrium between the ground state and the excited state of the ligand bound complex. As $k_{off}$ increases, the importance of the excited state of the ligand bound complex in $k_{off}$ determination becomes more significant. Because training datasets DE-
RMLR and DS-RMRR only cover the kinetic characteristics of the ground state of the ligand bound complex, the kinetic properties of the excited state of the ligand bound complex need to be added to the training datasets for higher accuracy of prediction. Using alignment and docking techniques, the excited state of ligand bound complex can be constructed in two steps from 1HHP (PDB code): a single chain 3D crystal structure of ligand free HIV-1 protease with open flap lid. First, a two-chain 3D structure will be built from 1HHP according to the symmetric characteristics of HIV-1 protease. Second, minimization and equilibrium will be conducted using MD simulation. Then the kinetic and thermal dynamic characteristics of the complex will be collected using NMA and MD simulations.

- More training records:

Thirty-nine pairs of the kinetic constants ($k_{on}$ and $k_{off}$) of ligand bound HIV-1 complexes were reported [12] but, due to the detection limit problems of SPR Biosensor, only thirty-four training vectors are available for $\log_{10}k_{on}$ prediction {Section 2.1.1}. As shown in Figure 6.13, the accuracy of prediction rises as the number of training records in the bin increases. Thus, more training records should be collected.

- Non-active-site PIRM residues:

In this study, forty-four residues with distance of $\leq 4.2\text{"Å} from ligand were chosen for training attribute construction. According to the report issued by the World Health Organization in 2013, there are fourteen major PIRM residues and twelve residues of non-polymorphic accessory mutation. Among them, fifteen are non-active-site residues having distances of $\geq 8.29\text{"Å} from the ligand. They are L10, V11, K20, L24, L33, K43,
M46, Q58, A71, G73, T74, N83, N88, L89, and L90. Adding the characteristics of the non-active-site PIRM residues in the training dataset is future work.

- New mathematical definition is needed:

In order to cover 90% of protein function, kinetic properties of the ten lowest frequency modes of NMA were collected in NMA Training Attribute Value:

\[(DPVV_1^2 + DPVV_2^2 + \ldots + DPVV_{10}^2)^{1/2},\]

where \(j = 1\) to 10 is the normal mode index. \(DPVV_j^2\) is the square of the dot product of two vectors A and B. Since

\[(A\cdot B\cdot \cos\Theta)^2 = (A\cdot B\cdot \cos(180-\Theta))^2,\]

the negative correlation of \(A\cdot B\cdot \cos\Theta = -A\cdot B\cdot \cos(180-\Theta)\) is eliminated in the NMA Training Attribute Value.

Figure 7.1 Mechanism for the hydration of the keto amide inhibitor by HIV-1 protease
Bibliography


www.ks.uiuc.edu/Research/vmd/current/ug.pdf


http://psych.stanford.edu/~jlm/papers/PDP/Volume%201/Chap8_PDP86.pdf


oz.berkeley.edu/~breiman/randomforest2001.pdf


www.pnas.org/cgi/doi/10.1073/pnas.1424461112


