PREDICTING THE BINDING AFFINITIES OF DRUG-PROTEIN INTERACTION BY ANALYZING THE IMAGES OF BINDING SITES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZLEM ERDAŞ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

JULY 2013

Approval of the thesis:

PREDICTING THE BINDING AFFINITIES OF DRUG-PROTEIN INTERACTION BY ANALYZING THE IMAGES OF BINDING SITES

submitted by **ÖZLEM ERDAŞ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Prof. Dr. Ferda Nur Alpaslan _____
Supervisor, **Computer Engineering Dept., METU**

Prof. Dr. Erdem Büyükbingöl _____
Co-supervisor, **Faculty of Pharmacy, Ankara University**

Examining Committee Members:

Prof. Dr. Sibel Süzen
Faculty of Pharmacy, Ankara University _____

Prof. Dr. Ferda Nur Alpaslan
Computer Engineering Dept., METU _____

Assoc. Prof. Dr. Ahmet Coşar
Computer Engineering Dept., METU _____

Assoc. Prof. Dr. Tolga Can
Computer Engineering Dept., METU _____

Assist. Prof. Dr. Ahmet Oğuz Akyüz
Computer Engineering Dept., METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ÖZLEM ERDAŞ

Signature :

ABSTRACT

PREDICTING THE BINDING AFFINITIES OF DRUG-PROTEIN INTERACTION BY ANALYZING THE IMAGES OF BINDING SITES

Erdaş, Özlem
Ph.D., Department of Computer Engineering
Supervisor : Prof. Dr. Ferda Nur Alpaslan
Co-Supervisor : Prof. Dr. Erdem Büyükbingöl

July 2013, 78 pages

Analysis of protein-ligand interactions plays an important role in designing safe and efficient drugs, contributing to drug discovery and development. Recently, machine learning methods have been found useful in drug design, which utilize intelligent techniques to predict unknown protein-ligand interactions by learning from specific properties of known protein-ligand interactions. The aim of this thesis is to propose a novel computational model, Compressed Images for Affinity Prediction (CIFAP), to predict binding affinities of structurally related protein-ligand complexes. The novel method presented here is based on a protein-ligand model from which computational affinity information is obtained by utilizing 2D electrostatic potential images determined for the binding site of the proteins with its inhibitors. The patterns obtained from the 2D images were used for building a predictive model whose strength was tested using Partial Least Squares Regression (PLSR), Support Vector Regression (SVR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) in comparison. The experiments were conducted on two distinct protein-ligand complex systems, which were complexes of CHK1-thienopyridine derivatives and CASP3-isatin sulfonamide derivatives. It is observed that the pixels of the images which are close to the surfaces of the interaction site have better explanation of the binding affinity. Moreover, PLSR is found to be the most promising prediction method for CIFAP as compared to SVR and ANFIS with the lowest error and the highest correlation between the observed and experimental binding affinities. The computational algorithm presented here is proposed to

have a great potential in pharmacophore-based drug design, especially in prediction of binding related properties.

Keywords: protein-drug interactions, binding affinity prediction, feature selection, regression algorithms

ÖZ

BAĞLANMA ALANLARININ GÖRÜNTÜLERİNİ İNCELEYEREK İLAÇ-PROTEİN ETKİLEŞİMİNİN BAĞLANMA EĞİLİMİNİN TAHMİN EDİLMESİ

Erdaş, Özlem

Doktora, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi : Prof. Dr. Ferda Nur Alpaslan

Ortak Tez Yöneticisi : Prof. Dr. Erdem Büyükbingöl

Temmuz 2013 , 78 sayfa

Protein-ligand etkileşimlerinin analizi güvenli ve etkili ilaçların tasarımında, ilaç keşfi ve geliştirilmesinde önemli rol oynamaktadır. Yakın zamanda, bilinen protein-ligand etkileşimlerinin belirli özelliklerini öğrenerek bilinmeyen protein-ligand etkileşimlerinin tahmininde akıllı yöntemler kullanan makine öğrenimi metotları ilaç tasarımı konusunda yararlı bulunmuştur. Bu tezin amacı, Eğilim Tahmini için Sıkıştırılmış Görüntüler (CIFAP) adında yapısal benzerlik taşıyan protein-ligand komplekslerinin bağlanma eğilimlerinin tahmininde kullanılmak üzere yeni hesaplamalı bir model geliştirmektir. Burada sunulan yeni metot, hesaplamalı eğilim bilgisinin proteinlerin inhibitörleri ile bağlandıkları bölgede belirlenen iki boyutlu elektrostatik potansiyel görüntüleri kullanarak elde edilen bir protein-ligand modeline dayalıdır. İki boyutlu görüntülerden elde edilen örüntüler tahminsel bir model etmekte kullanıldı ve bu modelin gücü Kısmi En Küçük Kareler Regresyonu (PLSR), Destek Vektör Regresyonu (SVR) ve Adaptif Nöro-Bulanık Çıkarsama Sistemi (ANFIS) yöntemleri ile test edildi. Deneyler iki farklı protein-ligand kompleks sisteminde gerçekleştirildi. Bu sistemler, CHK1-tienopiridin türevleri ve CASP3-izatin sülfonamid türevleri bileşikleriydi. Görüntü piksellerinden bağlanma yüzeylerine yakın olanlarının bağlanma eğilimini açıklamakta daha iyi olduğu gözlendi. Bununla birlikte, SVR ve ANFIS ile karşılaştırıldığında deneysel ve tahmin edilen bağlanma eğilimleri arasında en düşük hata ve en yüksek korelasyonu

sağlayan PLSR'ın CIFAP için en umut verici yöntem olduğu tespit edildi. Burada sunulan algoritmanın farmakofora dayalı ilaç tasarımında, özellikle bağlanma eğilimlerinin tahmininde, büyük bir potansiyele sahip olduğu görülmektedir.

Anahtar Kelimeler: protein-ilaç etkileşimleri, bağlanma eğilimi tahmini, özellik seçimi, regresyon algoritmaları

To my family

ACKNOWLEDGMENTS

I am greatly indebted to many people without whom this study might not come to an end.

Foremost, I am grateful to my supervisor Prof. Dr. Ferda Nur Alpaslan for believing in me, guiding me and motivating me through the process. I am also indebted to my co-supervisor Prof. Dr. Erdem Büyükbingöl for being patient of my pessimism, helping my mind open to creative ideas, pushing me to the limit. They both are more than supervisors to me; they are my mentors, guides and friends.

I am also thankful to my thesis committee members, Prof. Dr. Sibel Süzen and Assoc. Prof. Dr. Tolga Can, for their motivation and comments during this study.

I wish to express my sincere thanks to Assist. Prof. Dr. Cenk A. Andaç who helped me overcome the chemical part of the problem and gave me inspiration about academic writing. I would also like to thank A. Selen Gürkan-Alp for helping me prepare the data used in this study. Without them, it would have taken much more time for me to complete this study.

My sincere thanks also goes to my friends, Pınar who helped me start the experiments, Özge who gave me hope in my darkest moments, Leyla, Deniz, and Elif who supported and motivated me. I would also like to thank my ex-colleagues, Alev, Serdar, Mine, Ayşe Gül, Zerrin, and Hilal for being there for me whenever I needed.

I take this opportunity to record, my sincere gratitude to Scientific and Technological Research Council of Turkey (TÜBİTAK) for providing me Ph.D. fellowship (2211).

Last but not the least, I would like to thank my family, especially my parents Vicdan and Muammer Erdaş, for supporting me unconditionally throughout my life. I would also thank my brother Özgür, and my cousins Sebahat and Neslihan for being a part of my life.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xvi

LIST OF ABBREVIATIONS xxi

CHAPTERS

1 INTRODUCTION 1

1.1 Motivation 1

1.2 Basic Concepts 2

1.3 Problem Statement 3

1.4 Contributions 3

1.5 Organization of the Thesis 4

2 RELATED WORK 7

2.1 Classification: To Bind or Not To Bind 7

2.1.1 SVM with Different Kernels 7

2.1.2 Ensemble Learning 8

2.1.3 SVM vs. Other Machine Learning Methods 9
2.2 Prediction: Strength of the Interaction 10
2.2.1 Geometrical Descriptors 10
2.2.2 Quantitative Logical Rules 10
2.2.3 Selection of Promising Features 11
2.2.4 Random Forests 12
2.2.5 Image Processing for Prediction 13
3 DATA MODELLING METHODS 15
3.1 Ligand Preparation and Docking 15
3.2 Obtaining 3D Electrostatic Potential Grid Maps 15
3.3 Compressing 3D Cube into 2D Image 16
3.4 Feature Selection 16
3.4.1 Sequential Forward Selection 16
3.4.2 Sequential Backward Elimination 17
3.4.3 Sequential Forward Floating Selection 17
4 PREDICTION METHODS 19
4.1 Selection of training and test sets 19
4.2 Regression Algorithms 20
4.2.1 Multiple Linear Regression 20
4.2.2 Partial Least Squares Regression 20
4.2.3 Support Vector Regression 21
4.2.4 Adaptive Neuro-Fuzzy Inference System 25
4.3 Statistical Analysis 27

5 EXPERIMENTAL RESULTS 29

5.1 Checkpoint Kinase 1 and its inhibitors 29

5.1.1 Chemical Structures 29

5.1.2 Data Modelling Phase 33

5.1.3 Prediction Phase 36

5.2 Caspase 3 and its inhibitors 49

5.2.1 Chemical Structures 49

5.2.2 Data Modelling Phase 49

5.2.3 Prediction Phase 54

6 DISCUSSION AND CONCLUSION 67

REFERENCES 71

CURRICULUM VITAE 77

LIST OF TABLES

TABLES

Table 5.1 R^2 and RMSE values for an average and the best 3 ANFIS determination of random subsampling set selection method out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CHK1-ligand complexes. 38

Table 5.2 Optimal values of the SVR parameters C, ε, and γ for CHK1-ligand complexes. 40

Table 5.3 R^2 and RMSE values for an average and the best three SVR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CHK1-ligand complexes. 42

Table 5.4 RMSE comparison between observed binding affinities (pIC_{50}) for 57 CHK1 inhibitors, published by Zhao et al.[1], and the corresponding binding affinities (pIC_{50}) predicted by the PLSR, SVR, and ANFIS determination of leave-one-out cross validation for the Z-feature vectors of the testing data sets. 46

Table 5.5 R^2 and RMSE values for an average and the best three PLSR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CHK1-ligand complexes. 48

Table 5.6 R^2 and RMSE values for an average and the best 3 ANFIS determination of random subsampling set selection method out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes. 56

Table 5.7 Optimal values of the SVR parameters C, ε, and γ for CASP3-ligand complexes. 58

Table 5.8 R^2 and RMSE values for an average and the best three SVR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes. 60

Table 5.9 RMSE comparison between observed binding affinities (pIC_{50}) for 35 CASP3 inhibitors, published by Wang et. al. [2], and the corresponding binding affinities (pIC_{50}) predicted by the PLSR, SVR, and ANFIS determination of leave-one-out cross validation for the X-feature vectors of the testing data sets. 63

Table 5.10 R^2 and RMSE values for an average and the best three PLSR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes. 64

LIST OF FIGURES

FIGURES

Figure 1.1 Flow chart of the proposed method. 5

Figure 4.1 ε-tube, variables for noisy data, ξ, ξ^* and loss function. 22

Figure 4.2 The ANFIS architecture 25

Figure 5.1 SAR at 4-position of thienopyridine 30

Figure 5.2 SAR at 2-position of thienopyridine 31

Figure 5.3 SAR of core modification of thienopyridine 32

Figure 5.4 An exemplary illustration of the 3D electrostatic potential (EP) grid for the CHK1-compound 70 complex and the corresponding compressed 2D images. A view of the EP grid through (a) the X-axis (top) and the corresponding compressed X-image (bottom), (b) Y-axis (top) and the corresponding compressed Y-image (bottom), and (c) Z-axis (top) and the corresponding compressed X-image (bottom). The color scales for the compressed images are shown on the right side. 34

Figure 5.5 Two dimensional X-, Y- and Z-pattern images of CHK1-ligand complexes obtained by Sequential Forward Selection, SFS. 35

Figure 5.6 Two dimensional X-, Y- and Z-pattern images of CHK1-ligand complexes obtained by Sequential Floating Forward Selection, SFFS. 35

Figure 5.7 Affinity correlation plots constructed upon ANFIS determination of leave-one-out cross validation. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 57 different testing sets selected from the X-pattern (a), Y-pattern (b), and Z-pattern (c) images of the CHK1-ligand complexes. . 37

Figure 5.8 Affinity correlation plots constructed by ANFIS determination of random subsampling set selection. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.1) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 57 CHK1-ligand complexes, 10 of which were used for testing (red dots) and the rest were used for training (blue circles). 39

Figure 5.9 Affinity correlation plots constructed upon SVR determination of leave-one-out cross validation. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the optimal C, ε and γ values given in Table 5.2 and 57 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CHK1-ligand complexes. 41

Figure 5.10 Affinity correlation plots constructed by SVR determination of random subsampling set selection. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.3) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 57 CHK1-ligand complexes, 10 of which were used for testing (red dots) and the rest were used for training (blue circles). 43

Figure 5.11 Affinity correlation plots constructed upon PLSR determination of leave-one-out cross validation. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 57 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CHK1-ligand complexes. 44

Figure 5.12 Affinity correlation plots constructed by PLSR determination of random subsampling set selection. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.5) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 57 CHK1-ligand complexes, 10 of which were used for testing (red dots) and the rest were used for training (blue circles). 47

Figure 5.13 Chemical structures and binding affinities of Caspase3 inhibitors . . 50

Figure 5.14 Chemical structures and binding affinities of Caspase3 inhibitors (continued) 51

Figure 5.15 An exemplary illustration of the 3D electrostatic potential (EP) grid for the Caspase3-compound 1 complex and the corresponding compressed 2D images. A view of the EP grid through (a) the X-axis (top) and the corresponding compressed X-image (bottom), (b) Y-axis (top) and the corresponding compressed Y-image (bottom), and (c) Z-axis (top) and the corresponding compressed X-image (bottom). The color scales for the compressed images are shown on the right side. 52

Figure 5.16 Two dimensional X-, Y- and Z-pattern images of CASP3-ligand complexes obtained by Sequential Forward Selection, SFS. 53

Figure 5.17 Two dimensional X-, Y- and Z-pattern images of CASP3-ligand complexes obtained by Sequential Floating Forward Selection, SFFS. 53

Figure 5.18 Affinity correlation plots constructed upon ANFIS determination of leave-one-out cross validation. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 35 different testing sets selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CASP3-ligand complexes. . 55

Figure 5.19 Affinity correlation plots constructed by ANFIS determination of random subsampling set selection. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (random-1 in Table 5.6) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 35 CASP3-ligand complexes, 7 of which were used for testing (red dots) and the rest were used for training (blue circles). 57

Figure 5.20 Affinity correlation plots constructed upon SVR determination of leave-one-out cross validation. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the optimal C, ε and γ values given in Table 5.7 and 35 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CASP3-ligand complexes. 59

Figure 5.21 Affinity correlation plots constructed by SVR determination of random subsampling set selection. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.8) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 35 CASP3-ligand complexes, 7 of which were used for testing (red dots) and the rest were used for training (blue circles). 61

Figure 5.22 Affinity correlation plots constructed upon PLSR determination of leave-one-out cross validation. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 35 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CASP3-ligand complexes. 62

Figure 5.23 Affinity correlation plots constructed by PLSR determination of random subsampling set selection. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.8) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 35 CASP3-ligand complexes, 7 of which were used for testing (red dots) and the rest were used for training (blue circles). 65

LIST OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
ANFIS	Adaptive Neuro-Fuzzy Inference System
CASP3	Caspase 3
CHK1	Checkpoint Kinase 1
CIFAP	Compressed Images for Affinity Prediction
IC_{50}	Half maximal inhibitory concentration
LOOCV	Leave-one-out cross validation
MLR	Multiple Linear Regression
PLSR	Partial Least Squares Regression
QSAR	Quantitative structure-activity relationship
RMSE	Root mean square error
SBE	Sequential ackward Elimination
SFFS	Sequential Floating Forward Selection
SFS	Sequential Forward Selection
SVILP	Support Vector Inductive Logic Programming
SVM	Support Vector Machines
SVR	Support Vector Regression

CHAPTER 1

INTRODUCTION

1.1 Motivation

Scientists have made progress in understanding the insights of diseases by developing new techniques in the fields of genomics, proteomics, and medicine in recent years. As long as the knowledge at a molecular level is gathered and blended, finding safe and efficient drugs will be promising in medical treatment [3]. However, drug discovery and development is an expensive process which takes almost 20 years and costs billions of dollars. The process includes many steps as follows [4]:

- authenticating the drug targets (the proteins which the drug might affect)
- finding the promising compound to interact with the selected target
- testing the novel drug in the laboratories
- conducting clinical trials
- getting the approval
- offering the drug to the market

There are a lot of failures in drug discovery. Among the thousands of compounds which are studied, only one can get the approval. Understanding protein-ligand interactions are important to design new drugs which are safe and efficient and to help drug discovery and development. Computational methods, especially molecular docking, are useful for investigating protein-ligand interactions. However, scoring functions of docking programs which are used for predicting the strength of the interaction are not always reliable. Recently, intelligent methods have become popular in drug design. It is possible to gather information from known interactions, to search for or predict specific properties of the interactions, and design new drugs with the help of bioinformatics and machine learning methods [5].

1.2 Basic Concepts

The terminology of protein, ligand and interaction should be defined in order to understand protein-ligand interactions.

Proteins are large molecules composed of an exact sequence of amino acids which are small molecules composed of an amino group (NH_2), a carboxyl group (COOH), and a hydrogen atom attached to a carbon at the center [6, 7, 8, 9]. There are 20 groups of amino acids. Although all proteins are constructed by the union the same 20 groups, they are dissimilar in terms of amino acid arrangement, meaning that some proteins consist of an excess amount of one amino acid while other proteins may be deficient in some members of the group[8]. The arrangement of amino acids causes the protein to be folded into a specific three dimensional geometry, so called conformation. It should be noted that proteins are chemical and flexible structures which allow them to carry on their vital functions in the cells [6].

The molecule which binds to the protein in order to form a complex and to accomplish a duty like inhibiting an enzymes activity is called a ligand. The word “ligand” comes from “ligare” which means “band” or “tie” [6, 10]. A ligand may either be a single atom, a molecule or a protein [6]. However, the drug-like molecules will be referred as ligands during this thesis. The protein to which a ligand binds will also be addressed as "target" or "receptor".

There should be a simultaneous set of weak bonds such as hydrogen bonds, ionic bonds, Van der Waals bonds besides a hydrophobic interaction between the ligand and its protein for the formation of a strong binding. Such a strong binding may not occur unless the surface profile of the ligand strictly fits to the target like “a hand in a glove” [6]. The proteins contain a smaller region called the binding site which allows ligands to bind selectively. The binding of the ligands can be detected by drug design methods such as Nuclear Magnetic Resonance (NMR) or by X-ray crystallography[11, 9]. The protein changes its shape for helping the ligand easily dock to the binding site. This docking is generally reversible [6].

The strength of the protein-ligand interaction is measured by binding affinity which is affected by thermodynamical and chemical forces. In general, high-affinity ligand binding results from greater intermolecular force between the ligand and its receptor than that of low-affinity ligand binding. Moreover, ligands with high-affinity resides in the binding site longer than the ligands with low-affinity. The understanding of the principles of binding thermodynamics and the calculation of binding affinity are difficult because it is based on calculating the binding free energy of the complex which is really smaller than the individual free energies of the ligand and the protein [12]. Because of the difficulty in calculation of binding affinity, the “half maximal inhibitory concentration” (IC_{50}) is used as a standard of the impact of a ligand in inhibiting biological function of a protein. This numerical scale states how much of a specific

ligand is required "to inhibit a given biological process by half" [13].

1.3 Problem Statement

Calculation of the binding affinity is a difficult task with today's technology using the computational methods of the drug design. Novel methods should be developed for estimating the strength of the protein-ligand interaction without experimenting on thousands of potential molecules. These methods should represent the interaction data well enough in order to obtain precise estimations.

Data representation schemes in drug design, which are explained in details in Chapter 2, are mainly based on three groups of features: ligand-based, target-based and binding pocket related [14, 15, 16, 17, 18, 19, 20, 21, 22]. Reviewed studies either used one of these types of representations or a combination of them. It is observed in the study of Li et al. [20] that the combined representations of ligands and binding pockets are more informative than the other combinations. In the same study, it is also stated that some of the molecular and geometrical descriptors such as electrostatic and hydrophobic properties of the ligand are more promising in predicting the binding affinity of protein-ligand complexes. Moreover, it is required that both prediction and classification algorithms work fast in order to search huge biological and molecular spaces. Also, the chosen algorithms should be tuned easily to handle noise and to achieve optimal goals.

The main problem addressed in this thesis is the absence of an efficient data model to represent the binding affinity of drug molecules to the protein of interest. The aim of this thesis is to propose a novel data model of protein-ligand complexes in order to predict binding affinity of the complex. The proposed data model uses the structural information of the complexes and be represented by 2D images of the electrostatic potential. Utilizing machine learning methods for analysis of the 3D structure of protein-ligand complexes would be helpful in the quest of finding the most promising drug candidates. The thesis addresses the second problem as to provide a methodology for predicting the binding affinity of protein-ligand complexes using the proposed data model. A combination of feature selection and regression methods is provided as the solution.

1.4 Contributions

The primary goal in this thesis is to propose and establish a groundwork for a novel data modelling and prediction methodology, so called Compressed Images For Affinity Prediction, CIFAP, to predict binding affinities of novel compounds which may be used in medical treatment.

The contributions of the thesis are summarized as follows:

- A novel data model of protein-ligand interactions is proposed. The data representation utilizes the 3D geometrical information and electrostatic potential energy of protein-ligand complexes.
- A visualization in form of 2D images is provided for interaction data by compressing 3D grid cube of electrostatic potential map.
- It is revealed that the most important portion of a protein-ligand complex lies in the vicinity of the ligand surface when the aim is to predict the binding affinity.
- A prediction system is built for forecasting the binding affinity of protein-ligand complexes using the 2D images of the binding site of the complex.
- Linear regression explains the relationship between the binding affinity and geometrical/electrical structure of protein-ligand complexes more accurate than non-linear or fuzzy regression.

1.5 Organization of the Thesis

Previous machine learning studies, which will be explained in Chapter 2 in details, analyze the protein-ligand pairs by handling ligand and protein separately. Only molecular docking studies use the structural information but their scoring calculations take excessive time. The aim of this study is to construct a novel data representation and methodology, the so-called Compressed Images for Binding Affinity Prediction (CIFAP), for predicting binding affinity of protein-ligand interaction by analyzing the 2D images of the binding sites of complexes which uses the geometrical and electrical information of protein-ligand complexes.

The CIFAP includes two phases. The first phase is the data modelling phase which will be explained in Chapter 3. The data modelling phase involves docking process of the ligands into the binding site of the X-ray coordinates of the target protein, generation of an electrostatic potential grid box covering the binding pocket of the complex, and compression of the grid points through the X, Y, and Z dimensions into 2D electrostatic images of the binding site. After obtaining the compressed images, the aim becomes to find certain patterns by filtering the 2D images via the Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) techniques to avoid redundant features in the prediction. The second phase is called the prediction phase as mentioned in Chapter 4, which initially aims to apply regression and learning procedures on the filtered 2D-images. Here, Partial Least Squares Regression (PLSR), Support Vector Regression (SVR) and Adaptive Neuro-Fuzzy Inference System (ANFIS) methods, which are thought to be promising prediction tools in drug discovery, were applied to test the regression and learning features of CIFAP, and the strength

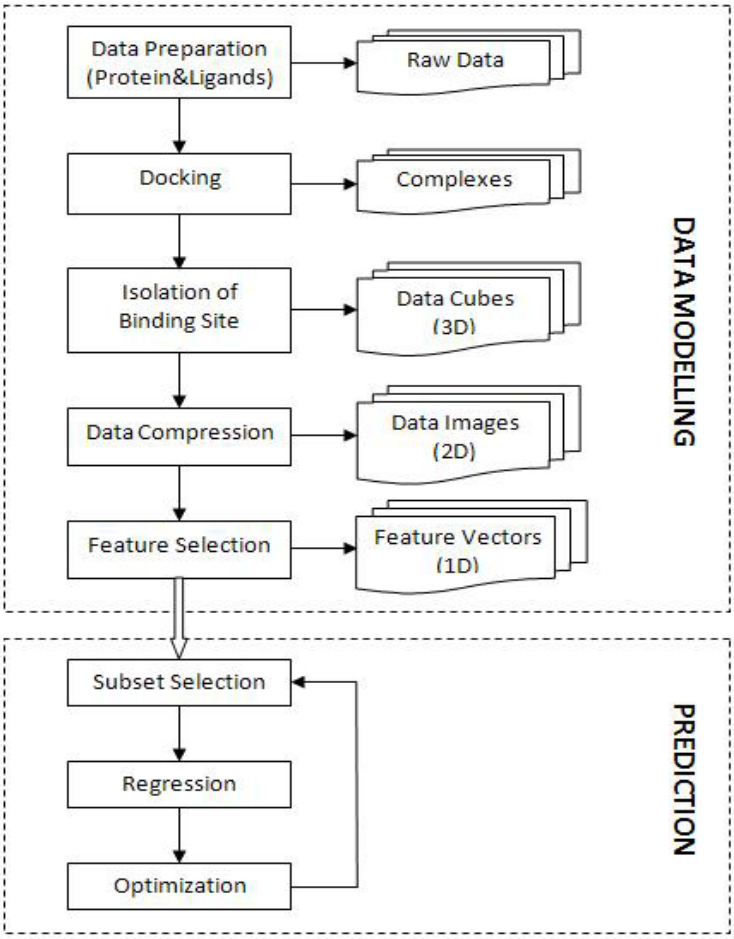


Figure 1.1: Flow chart of the proposed method.

of these methods are discussed in Chapter 5. Figure 1.1 shows the flow chart of the proposed method.

Protein-ligand interaction data which belong to a single protein in complex with multiple ligands were selected to be tested in this study. Moreover, the ligands with similar pharmacophore structures were preferred for building a consistent model. Checkpoint 1 kinase (CHK1) [1] and Caspase 3 (CASP3) inhibitors, which will be mentioned in Chapter 5 in details, were utilized for the experimental work.

CHAPTER 2

RELATED WORK

This section provides an overview of recent studies which use machine learning techniques for classification and prediction of binding properties of protein-ligand interactions.

2.1 Classification: To Bind or Not To Bind

2.1.1 SVM with Different Kernels

In silico (computer-based) methods in drug design are categorized into ligand-based and structure-based (docking) approaches. Ligand-based methods use a candidate ligand in comparison with the known ligands of the receptor for predicting specific properties using machine learning algorithms while structure-based methods utilize the geometrical structure of the receptor for finding out how well each candidate interacts with the receptor. The classical approaches are not applicable for a given target with unknown 3D structure and unknown ligands [15].

Jacob and Vert [15] used the idea of chemogenomics in their study. The goal of chemogenomics is searching through the entire chemical space of all small molecules for interactions with the biological space of all proteins, especially drug targets. This study attacks to the problem of deciding whether a given protein-ligand pair interacts.

Compound interaction data were collected from the KEGG BRITE Database including three receptor types; enzyme, GPCR, ion channel; and a list of known ligands for each class. For simplification, data were classified as positive (interacting) and negative (not interacting). Hence, a positive data set was constructed from the known interactions while a negative data set was constructed from targets and ligands which are not known to interact experimentally. This resulted in 2436 data points for enzymes (1218 known enzyme-ligand pairs and 1218 generated negative points) representing interactions between 675 enzymes and 524 compounds [15].

In this research, separate kernels for targets and ligands were calculated and were

combined by tensor product. Ligands were represented as 2D graphs. For ligands, Tanimoto kernel which is formulated using binarized vector representation of graphs was calculated by ChemCPP program. For targets, five different kernels were constructed:

1. Dirac kernel simply represents different targets as orthonormal vectors. For instance, Dirac kernel is 1 if two targets are equal and 0 if not.
2. The multitask kernel takes away the orthogonality of different proteins letting information be shared. However, the kernel cannot measure how well currently known interactions contribute to the model.
3. The mismatch kernel compares target proteins and determines the dissimilarities.
4. The local alignment kernel calculates scores for each target and determines how similar the proteins are with respect to their primary sequences.
5. The hierarchy kernel, which was defined by the authors, between two proteins of the same family is the number of common ancestors in the corresponding hierarchy summed by one.

Support Vector Machines (SVM) was trained on the data sets with the kernels proposed. The performance measure was the area under the ROC curve (AUC). Hierarchy kernel was observed to give rise to the best results with 90% accuracy of classification. The results on experimental ligand data sets pointed that target kernels sharing information across the targets improved the prediction to a great extent, particularly considering targets with few known ligands.

2.1.2 Ensemble Learning

Quantitative structure-activity relationship (QSAR) analysis plays a crucial role in the drug discovery process by facilitating the search for new drugs. QSAR assumes that there is a relationship between the structural or molecular properties of a ligand and its biological activity. The goal of QSAR analysis is to determine these relationships for calculating the activity of novel molecules with respect to their physiochemical features [16, 17].

Ensemble learning techniques work on the training samples to create a collection of classifiers. The selected learner and a training data set serve as the input to the learning algorithm which executes the base learner numerous times on the different distributions of the training set samples. The newly created classifiers are then incorporate to generate a final classifier which is utilized to classify the test set. In this study of Liu, two of the most popular techniques for constructing ensembles, Bagging and AdaBoost, were investigated [16].

Bagging (bootstrap aggregating) [23] tries to carry out the training data by replacing the original training data by randomly selected items. Each replacement training set, so called bootstrap replicate, has 63.2% of the original training set, meaning that some training samples may appear multiple times while some training samples do not appear at all. The final classifier is generated by combining the constructed classifiers by voting.

AdaBoost [24] is based on equally weighing every sample. In each iteration, it attempts to reduce the weighted error on the training set and generates a classifier. The weighted error of every classifier is calculated in order to update the weights on the training samples. The weight of each sample adapts according to its effects on the classifier's outcome. A misclassified sample gains more weight while a correctly classified sample loses weight. The final classifier is generated by a weighted vote of the previously constructed classifiers according to their performance based on the weighted training set.

For this study, two data sets including 74 instances of Pyrimidines with 28 features and 186 instances of Triazines with 61 features were collected from the literature. Features were arranged according to the positions of possible substitutions and contained molecular descriptors like polarity, size, flexibility, hydrogen-bond donor, hydrogen-bond acceptor, π donor, π acceptor, polarizability, σ effect, branching and biological activity [16].

Decision Tree C4.5, 1-R ($p < 0.05$), Naïve Bayesian (NB) and 1-Nearest Neighbor (1-NN) methods were trained with AdaBoost and Bagging on these data sets. It was observed that ensemble learning methods improved the performance of single learning methods like C4.5 and 1-R which are not statistically stable. However, NB and 1-NN were not affected from ensembling. The accuracy was around 80% for both data sets and for the algorithms which used ensemble methods. It was also reported that the ensemble learning method worked better on the pyrimidine data set because of its simpler structure.

2.1.3 SVM vs. Other Machine Learning Methods

In this study of Burbidge et. al.[17], the problem was to predict the inhibition of dihydrofolate reductase by pyrimidines. The biological activity was measured as $log(1/K_i)$, where K_i is the equilibrium constant for the interaction of the ligand to dihydrofolate reductase. Data were the pyrimidine data as mentioned in the previous research [16]. However, 55 instances were selected among 74 pyrimidine compounds.

The aim here was to transform the regression problem into a classification problem. The prediction task became learning the relationship $great(d_n, d_m)$ which declared that the n^{th} drug's activity is higher than that of the m^{th} drug. Each data sample

contained a couple of drugs having 54 features, and a label 'true' or 'false', indicating the value of the relationship $great()$.

For this classification problem, SVM with Gaussian Kernel, multi-layer perceptron (MLP), radial basis function (RBF) networks and C5.0 decision trees were used for comparison. It was observed that SVM outperformed the other methods except MLP which is however 10 times slower than SVM.

2.2 Prediction: Strength of the Interaction

2.2.1 Geometrical Descriptors

The goal of the study of Deng et. al.[18] was to predict the binding affinity of ligand-protein interactions. The algorithm presented in this was based on the number of the appearance of each atom type pair that included atoms from both ligand and binding site of the receptor within a certain distance range. This information was exploited to arrange features for quantitative structure activity/-property relationships (QSAR/QSPR) analysis. Each feature value is then defined as the appearance of a special atom pair within a specified distance bin of 1 Å width.

Two distinct data sets containing 61 and 105 co-crystalized complexes were collected from RSCB Protein Data Bank as training sets. The experimental dissociation values for each complex were obtained from the literature. Two distinct test sets containing 6 and 10 complexes were also constructed externally as test set. The 1445 features were the number of the appearance of protein-ligand atom pairs within a specified distance bin. Since the number of features was much greater than the number of complexes, Genetic algorithms (GAFEAT) were used for avoiding redundant features and "curse of dimensionality". After feature selection, the kernel partial least squares (K-PLS) method with RBF kernel was applied 100 times for better generalization. The results were compared to those of multiple adaptive regression splines (MARS) method and were found to be similar. It was reported that the proposed feature selection method improved the accuracy of prediction [18].

2.2.2 Quantitative Logical Rules

Most of the docking programs are considered to be notable at correctly locating the ligand in the binding site as compared to X-ray structures. On the other hand, the problem of predicting the binding affinity of a ligand to a particular target is hard to solve [19].

Amini et al. [19] studied on X-Ray structures of five inhibitors; HIV protease, carbonic anhydrase II (CA II), trypsin, thrombin, and factor Xa whose binding affinities were

taken from the literature. Tanimoto coefficient which is based on the number of similar fragments two molecules share was calculated for every molecule in one-to-all manner.

Inductive logic programming (ILP) is a qualitative method, which manipulates logic to form rules describing particular properties of each instance of a data set. In this work [19], support vector inductive logic programming (SVILP) which is a quantitative version of ILP was utilized for predicting binding affinities of collected protein-ligand complexes.

According to the proposed algorithm [19], each ligand molecule was fragmented by determining a central non-hydrogen atom to which other atoms were bonded as a fragment. The distance was computed between the atom at the center of each fragment and all residual protein atoms having at least one atom within 5 Å radius of a compound atom. The data set was splitted into positives and negatives with respect to their activity for letting ILP construct rules specifying the distances with predictive power. Support vector machine (SVM) assessed ILP rules using the SVILP methodology. A model was developed in form of a matrix having the activity of each molecule versus each rule. "1" corresponded to the occurrence of a rule for a molecule while "0" corresponded to the absence of the rule. A resembling matrix was built for testing molecules with unknown activities and the model obtained from training matrix was used for quantitative prediction of these molecules.

The results of SVILP were compared to results of comparative molecular field analysis (CoMFA), comparative molecular similarity analysis (CoMSIA), GoldScore, and DrugScore. It was shown that SVILP gave rise to lower mean squared error (MSE) values than GoldScore and DrugScore, and could also be compete with CoMFA and CoMSIA methods. Moreover, the outcome of SVILP was more human interpretable by humans and could be helpful for calculating rescoring functions on different systems with the same procedure.

2.2.3 Selection of Promising Features

In the research of Li et. al. [20], it is stated that the binding affinity of a ligand to a given protein can be calculated experimentally by NMR spectrometry, microcalorimetry, and surface plasmon resonance, but these methods are not often feasible considering the time and money. Therefore, many in silico ways are proposed for predicting the binding affinity which make use of the structural and chemical properties of protein and ligand. The most successful in silico methods are docking and scoring methods which use mainly three classes of scoring functions; such as force field-based (e.g. DOCK, GOLD, SIE, and LIE), knowledge-based (e.g., DrugScore, DFIRE, DDFT, PMF, BLEEP, ITScore, and M-Score), and the empirical scoring functions with some varieties of statistical techniques (X-Score, FlexX Score, VALIDATE, SCORE1 (LUDI), SCORE, Chem-Score, SMoG, GEMDOCK, and SODOCK). Lately, other in

silico methods including statistical and machine learning have been developed as an alternative to docking and scoring. These methods have proved that they have some particular benefits such as ease of use, speed, and good generalization ability and usability as a fast filter in the virtual screening of large chemical databases [20].

In this study [20], 1300 refined protein-ligand complexes were collected from PDBBind 2007 database to be used as data set in which 493 samples have the binding affinity of dissociation constant (K_d) value and 807 samples have inhibition constant (K_i) value.

The protein-ligand complexes were described by three blocks of descriptors: sequence information of protein, ligand structural information, and binding pocket structural information. For proteins, 1497 descriptors were calculated which represented their structural and physicochemical properties. Ligands which were drawn using Hyper-Chem had 1664 features calculated by Dragon program. Binding pockets which were minimized using Tripos force field in Sybyl software had 125 descriptors. The data were divided into training and test sets based on Euclidean distance.

The number of descriptors were firstly reduced by removing the redundant variables who have a pair correlation higher than 0.9. Then, features were selected using ReliefF method which assigns a weight to each feature by sampling an instance multiple times and determining the value of the given property for the nearest samples belonging to the same and different classes. After ranking the features, 35 features having the best correlation values were selected by LS-SVM and leave-one-out cross validation techniques. Two models of LS-SVM were generated to handle K_d and K_i separately. It was found to be interesting that none of the selected features were from protein structure block. Also, it was seen that electrostatic and hydrophobic properties were very essential information for protein-ligand interaction. Although the hydrophobic effect of the ligand was important for K_i, K_d was dominated by the hydrophobic effect of the binding-pocket. Considering the geometrical descriptors, aromacity of the ligand seemed to be the most important property for both K_d and K_i models. The proposed method performed better than other similar published studies in terms of the trade off between predictive ability and model complexity. Because of its satisfied performance, the predictive model was reported to be a promising method as a fast filter for the rapid virtual screening of large chemical databases.

2.2.4 Random Forests

Molecular docking is an in silico method whose goal is to determine whether and how a certain ligand will tightly bind to a receptor. Ballester and Mitchell [21] define the two stages of molecular docking as: "docking molecules into the target's binding site (pose identification), and predicting how strongly the docked conformation binds to the target (scoring)."

In this study[21], a new scoring function was proposed for docking using Random Forest (RF) [25] which is based on a collection of decision trees constructed from bootstrap instances of training data, with predictions computed by general agreement of all trees. This non-parametric machine learning technique because it is difficult to model the docking procedure when other resampling methods such as bagging and cross validation cannot guarantee the generalization capability of parameter estimation for scoring functions.

Data collected from PDBbind 2007 database were refined by removing protein-protein and protein-nucleic acid complexes. The refined complexes had known dissociation and inhibition constants and contained ligand molecules which consisted of only the common heavy atoms (C, N, O, F, P, S, Cl, Br, I). The final step in refinement of the data set was to cluster the data into 65 clusters with a 90% similarity cutoff using BLAST sequence similarity. In order to distribute the binding affinity uniformly over the clusters, three complexes having the highest, median and lowest binding affinity were chosen for each cluster. 36 features were obtained each of which is the number of appearances of a specific protein-ligand atom type pair interacting within a particular extent [21].

A modified algorithm for Random Forests was used in the study [21]. First, bootstrap samples were used for growing each tree without pruning instead of using the same training data. Second, a small number of randomly selected features were used instead of the whole feature set. The RF-Score was compared with the top scoring functions such as X-Score, DrugScoreCSD, ChemScore and DS-PLP1. As a result, it was observed that RF-Score was highly correlated with experimental binding affinities.

2.2.5 Image Processing for Prediction

Saghaie et al. [22] implemented the multivariate image analysis method for investigating quantitative structure activity relationship of Cyclin dependent kinase 4 (CDK4) inhibitors. In the proposed multivariate image analysis (MIA) QSAR method, descriptors were pixels of bitmaps of molecules which resulted in large number of descriptors and the problem of high correlation between them. To solve the problem of collinearity, partial least squares (PLS) and radial basis function neural networks with principal component analysis (PC-RBFNN) were preferred as regression algorithms.

Compounds collected from the literature consisted of 94 indenopyrazole derivatives of which inhibitory activity in terms of $logIC_{50}$ were also provided. The two dimensional structures of 94 molecules were drawn using ChemDraw 7.0, and then saved in form of bitmaps which were set to 940×600 pixels and were fixed by selecting a common pixel. The number of features were reduced to 14775 by eliminating features with the zero variance. 20 samples out of 94 were selected as the test set while the remaining samples established the training set. The test set was constructed with a rational

heuristic requiring that each test sample were close to at least one training sample.

The performance of developed models namely PLS and PC-RBFNN were tested by well-known statistical measures such as root mean squared error (RMSE) and correlation coefficient (R^2) between observed and predicted inhibitory activity. The resulting PLS model had a higher statistical quality than PC-RBFNN for predicting the activity of the compounds. Because of high correlation between values of predicted and observed activity, MIA-QSAR was found to be a highly promising approach in prediction of inhibitory activity. It was also indicated that the linear method (PLS) had a higher performance than the nonlinear method (PC-RBFNN) in prediction of activity of studied CDK4 inhibitors.

CHAPTER 3

DATA MODELLING METHODS

3.1 Ligand Preparation and Docking

The ligands are drawn and minimized by the MM2 force field using the HyperChem 5.1 [26], which are then saved in MOL2 format. The ligands are then saved in PDB format by Discovery Studio Visualizer v.1.7[27]. X-ray coordinates for the selected receptor, in complex with the reference compound, generally the compound with the highest affinity, are obtained from the Protein Data Bank[28].

The reference compound initially removed from the binding site of the receptor. MGL Tools v.1.5.4[29] is used for preparing the ligands and the receptor, for which non-polar hydrogens are removed and the ligands and the receptor are saved in PDBQT format for docking. AutoDock Vina v.1.1.2[30] is used for docking the ligands flexibly into the binding site of the rigid coordinates of the receptor. Docking of the ligands is implemented in a confined grid box determined by MGL Tools v.1.5.4. The most suitable docked poses of the ligands are selected based on the pharmacophore showing the best superposition with that of the X-ray coordinates of the reference compound. Protons of the best poses for docked ligands are added by MGL Tools v.1.5.4, which are then saved in PDB format.

3.2 Obtaining 3D Electrostatic Potential Grid Maps

A cubic frame centering docked ligands and the inner boundaries of the binding site of the receptor is set by MGLTools v.1.5.4, with size by $37 \times 37 \times 37$. The center coordinates of the cube are determined by averaging the center coordinates of all ligands. Electrostatic potential grid map files in ASCII format are generated using the coordinates and size of the cubic frame, and the PDBQT coordinates of the docked ligands and the receptor by the AutoGrid4 module of AutoDock v4.2 suite of programs[31]. The cubic grids contain 37 grid points in each dimension, each separated by 0.5 Å. Each point has the electrostatic potential values corresponding to the coordinates of that point. The cube is preferred to be small because the images seem more significant

when the cube is focused on the ligand and the binding site.

3.3 Compressing 3D Cube into 2D Image

3D electrostatic potential matrices for the binding site of the complexes are constructed by MATLAB[32] using the corresponding electrostatic potential grid map files as input. The matrices are compressed into 2D images by summing up the electrostatic potential values at the grid points through the X, Y, and Z directions, resulting in three 2D images for each complex, so called the X-image, the Y-image, and the Z-image, respectively. Each image possesses a total of 1369 pixels of the compressed electrostatic potential values, which are used as feature sets in the feature selection step.

3.4 Feature Selection

The success of the learning system is determined by the representation and quality of data. Irrelevant and redundant features affect the quality of data and make learning complicated. Although, there are usually many features in a real-world problem, only some of them are related to the solution. The features are mainly distinguished in three categories: relevant, irrelevant, and redundant. The relevant features affect the output directly, so that any other feature cannot mimic their role. On the other hand, the irrelevant features do not affect the result at all, meaning that removal of them do not change the output. The redundant features have similar influence on the output as other features [33].

After achieving the 2D images, each compressed 2D image defined by 37 x 37 pixels is further processed to generate the corresponding feature vector (X/Y/Z-vector), having a total of 1369 compressed features, by sequentially lining up each of the 37 rows of the compressed 2D image next to each other. The Sequential Forward Selection (SFS)[34] and the Sequential Floating Forward Selection (SFFS)[35] methods, which are explained in details in the succeeding subsections, are then applied to reduce the number of features in the vectors to avoid the irrelevant and redundant features.

3.4.1 Sequential Forward Selection

SFS starts with an empty feature set, to which new features are added greedily so as to not to be deleted at a later stage, yielding a newly learned model with the best generalization ability at each step. Addition of new features by SFS ends at a point where the lowest root mean square error, RMSE, converges [36]. SFS with Multiple Linear Regression[37] is applied with leave-one-out cross validation, which

will be explained in Chapter 4, in order to find the best feature subset with the lowest RMSE.

One of the drawbacks of SFS is the failure of adding the interdependent features since it includes only one feature at each step. Furthermore, SFS cannot remove a feature, once it is added to the set. On the other hand, the algorithm has several advantages like being fast and achieving features which are operative and few in number.

Algorithm 1 :Sequential Forward Selection

Input: $P = \emptyset$ - initial feature set
Q - the full set of features
J - criterion function to minimize
Output: P - final feature set
repeat
for all $x \in Q$ **do**
set $P' \leftarrow P \cup \{x\}$
calculate $J(P')$
end for
set $P \leftarrow P \cup \{x^+\}$ where $x^+ = argmin[J(P')]$
set $Q \leftarrow Q \setminus \{x^+\}$
until no further improvement in J

3.4.2 Sequential Backward Elimination

Sequential Backward Elimination (SBE) which was introduced by Marill and Green [38] should be mentioned in order to explain SFFS. SBE starts with all features in the initial feature set and greedily removes one feature according to the criterion function at each step. SBE extracts the feature whose removal generates a feature set with the best generalization. Like SFS, SBE removes a feature permanently so that it cannot add it to the feature set later [39].

The algorithm works slower when the number of initial features increases. Moreover, SBE is not convenient to use with linear regression methods when the number features exceeds the number of data samples. Nevertheless, it is able to construct feature sets with good generalization ability since it can keep the interdependent features in the same feature set.

3.4.3 Sequential Forward Floating Selection

Sequential Forward Floating Selection (SFFS) [35] is a combination of SFS and SBE. The algorithm starts with an empty feature set and applies a forward selection step to add a single feature. Then, it performs variable steps of backward elimination. The

Algorithm 2 :Sequential Backward Elimination

Input: P - the full set of feature set
J - criterion function to minimize
Output: P - final feature set
repeat
for all $x \in P$ **do**
set $P' \leftarrow P \setminus \{x\}$
calculate $J(P')$
end for
set $P \leftarrow P \setminus \{x^-\}$ where $x^- = argmax[J(P')]$
until no further improvement in J

forward selection and backward elimination steps continue until there is no further improvement in the criterion function, which is RMSE in this study.

The algorithm works slower than SFS but faster than SBE. The main advantage of SFFS is that it overcomes the problem of "nesting"[35, 40, 41]. "Nesting", which is a common problem for both SFS and SBE, is defined as the assumption that the subset of the k best features selected contains the subset of the $k-1$ best features. However, the subset of the k best features may not include the subset of the $k-1$ best features in the real-world problems [41].

Algorithm 3 :Sequential Forward Floating Selection

Input: $P = \emptyset$ - initial feature set
Q - the full set of features
J - criterion function to minimize
Output: P - final feature set
repeat
Step 1. Select the best feature $x^+ = argmin[J(P \cup \{x^+\})]$
set $P \leftarrow P \cup \{x^+\}$
Step 2. Select the worst feature $x^- = argmax[J(P \setminus \{x^-\})]$
if $J(P \setminus \{x^-\}) < J(P)$ **then**
set $P \leftarrow P \setminus \{x^-\}$
go to Step 2
else
go to Step 1
end if
until no further improvement in J

CHAPTER 4

PREDICTION METHODS

Regression analysis is commonly utilized in prediction. A concerning problem in regression is to anticipate the functional dependence of the variable $y \in \Re$ on an m-dimensional independent variable $\mathbf{x}$. Thus, a mapping from $\Re^m$ to $\Re$ leads to an approximation to the real valued function of $f(x) = y$[42]. A data set of n input-output pairs can be described as $D = \{[\mathbf{x}_i, y_i] | \mathbf{x}_i \in \Re^m, y_i \in \Re, i = 1, ..., n\}$. In this study, $\mathbf{x}_i$'s and y_i's relate to feature vectors extracted from compressed 2D images and the corresponding experimental (observed) pIC_{50} values. The following sections describe the sample set selection methods, the prediction methods Multiple Linear Regression (MLR), Partial Least Squares Regression (PLSR), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Support Vector Regression (SVR), and statistical analysis used to test the CIFAP algorithm on protein-ligand complexes.

4.1 Selection of training and test sets

The prediction phase of CIFAP utilizes two different sample set selection methods that are applied through the regression analysis methods, described in the following sections, of pattern vectors of the protein-ligand complex systems: Leave-one-out cross validation [43] and repeated random subsampling [43]. The leave-one-out cross validation method used a X-feature, Y-feature, and Z-feature vector for testing and the rest of the X-feature, Y-feature, and Z-feature vectors for training. The leave-one-out cross validation tests were implemented by the regression analysis methods for the number of vectors times by picking up a different X-feature, Y-feature, and Z-feature vector as a new testing data set at each cycle.

The repeated random subsampling method shuffled the X-feature, Y-feature, and Z-feature vectors of the protein-ligand complexes, then used the first 18%-20% of X-feature, Y-feature, and Z-feature for test set and the rest of the X-feature, Y-feature, and Z-feature vectors as a training set. The repeated random subsampling validation was implemented by the regression analysis methods for 1000 times by reshuffling the X-feature, Y-feature, and Z-feature vectors and using the aforementioned number of

vectors at each cycle.

4.2 Regression Algorithms

4.2.1 Multiple Linear Regression

Multiple linear regression (MLR) is a method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is sometimes called the predictand, and the independent variables the predictors. MLR assumes that there is a linear relationship between predictand and predictors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_D x_{iD} \tag{4.1}$$

where β_0 is the constant variable and β_1 to β_D are coefficients.

MLR is based on least squares: the model is fit such that the sum-of-squares of differences of observed and predicted values is minimized.

The details of implementation: MLR computations utilizing leave-one-out cross validation were implemented for predicting the binding affinities of protein-ligand complexes in the feature selection process, for which SFS and SFFS algorithms described in the data modelling methods section. RMSE values of the observed and predicted binding affinity for each test vector was calculated in order to determine which features of the X-, Y-, and Z-images were meaningful for prediction.

4.2.2 Partial Least Squares Regression

Partial Least Squares Regression (PLSR) was published by Herman Wold for use in social sciences specifically in financial sciences. Nevertheless, computational chemistry researchers found PLSR useful after '80s [38].

PLSR mainly intends to search for "dependent variable" $\mathbf{Y}$ given "independent variables" $\mathbf{X}$ and take out the statistical attributes which are similar. The regression problem can be solved by multiple regression when the dependent variable is a vector and independent variables form a matrix with the "maximum number of linearly independent columns". The multiple regression cannot handle the problem when there are more independent variables than examples because of multicollinearity[1]. The techniques like Principal Component Regression (PCR) may eliminate multicollinearity by

[1] "When the correlation between the predictors is of high degree, multicollinearity occurs. In this case, effect of the predictors over prediction becomes hard to separate."

applying Principal Component Analysis (PCA) of $\mathbf{X}$ and removing some of them as a result. The application of PCA yields the principal components of $\mathbf{X}$ which will be utilized for predicting $\mathbf{Y}$. The orthogonality of principal components prevents the occurrence of multicollinearity. However, the method is a little problematic because of the selection of the best subset of predictors. It is hard to choose the suitable components for $\mathbf{Y}$. However, PLSR aims to discover the appropriate components of $\mathbf{X}$ which describe the covariance between $\mathbf{X}$ and $\mathbf{Y}$ as a recommended feature. This feature shows the generalization ability of PCA. At last, solution of $\mathbf{X}$ let $\mathbf{Y}$ be predicted[38].

The details of implementation: PLSR computations used the leave-one-out cross-validation [43] and repeated random subsampling [43] methods, which are described in details in the training/test set selection subsection of the prediction methods section. The feature vectors obtained from X-pattern, Y-pattern, and Z-pattern images by SFFS method were inputs to compute a PLSR model which calculates the linear coefficients of each variable for each training set. The coefficients computed by PLSR were used for predicting the pIC_{50} values of the test vectors which were selected by leave-one-out cross-validation and repeated random subsampling methods.

4.2.3 Support Vector Regression

The Support Vector Regression is an extension of Support Vector Machines which was developed by Vapnik et. al. In the basic sense, the Support Vector Regression, SVR, attempts to approximate the function in Equation 4.2

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b \tag{4.2}$$

where $\mathbf{x}$ is a feature vector of input data, $\mathbf{w}$ is the weight vector, and b is the bias. SVR computes the error of estimation instead of the margin as in Support Vector Classification (SVC)[44]. The use of a loss function differs SVR from the old-school regression techniques. ε-insensitive loss function[44] constructed by Vapnik, which describes an tube with a radius of ε, so called the ε-tube, is defined by Equation 4.3

$$E(\mathbf{x}, y, f) = |y - f(\mathbf{x})|_{\varepsilon} = \begin{cases} 0 & if\ |y - f(\mathbf{x})| \leq \varepsilon \\ |y - f(\mathbf{x})| - \varepsilon & otherwise \end{cases} \tag{4.3}$$

where $f, \mathbf{x}$, and y are the function between input and output data, the input vector of independent variables and the output value to be predicted, respectively. The loss in Equation 4.3 equals to zero if the estimated value is within the boundaries of the tube whose radius is ε. When the estimated value is outside the boundaries of the tube, the loss becomes the difference between the estimated value and the radius of ε-tube, ε. For the reduction of the loss in Equation 4.3, the summation term in Equation 4.4 should be minimized,

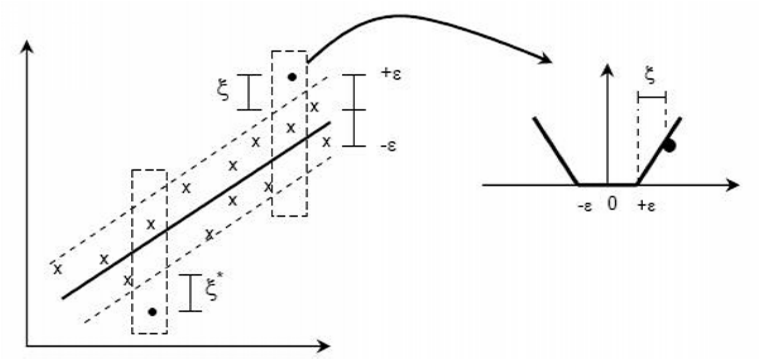


Figure 4.1: ε-tube, variables for noisy data, ξ, ξ^* and loss function.

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}|y - f(\mathbf{x})|_\varepsilon \tag{4.4}$$

where C is a constant for a "trade-off between the error of approximation and model complexity".

If data used are noisy, for eliminating the results of the noise in the background as represented in Figure 4.1, the new slack variables ξ_i and ξ_i^*[45] are introduced as shown in Equation 4.5,

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{4.5}$$

subject to

$$\begin{aligned} &y_i - f(\mathbf{x}) \leq \varepsilon + \xi_i \\ &f(\mathbf{x}) - y_i \leq \varepsilon + \xi_i^* \\ &\xi_i, \xi_i^* \geq 0 \end{aligned} \tag{4.6}$$

where $i = 1, ..., n$. It should be noted that a large C results in smaller slack variables and decreases error. Here, ε is the ε-tube's radius which regulates the quantity of support vectors which are placed on or outside the ε-tube. If ε becomes higher, the quantity of support vectors decrease eventually to zero, invalidating the prediction[46].

A Lagrangian function is calculated for the solution of the problem and optimized for finding the minimum/maximum problems. By exchanging the Karush-Kuhn-Tucker (KKT) conditions into the related function, the problem is transformed into a dual problem as in Equation 4.7

$$\begin{aligned} L_d(\alpha_i, \alpha_i^*) \quad = \quad & -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(\mathbf{x}_i \cdot \mathbf{x}_j) \\ & -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)y_i \end{aligned} \tag{4.7}$$

subject to

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \tag{4.8}$$

$$0 \leq \alpha_i, \alpha_i^* \leq C \tag{4.9}$$

where $i = 1, ..., n$.

When the learning process ends, n pairs of Lagrangian multipliers (α_i, α_i^*) are yielded as a result, and the multipliers which are different than zero determines the quantity of support vectors. It is important to remind that the dimensions of the data are irrelevant for finding the quantity of support vectors.

For finding the best solution, the KKT conditions below should be compensated:

$$\alpha_i(\mathbf{w} \cdot \mathbf{x}_i + b - y_i + \varepsilon + \xi_i) = 0 \tag{4.10}$$

$$\alpha_i^*(-\mathbf{w} \cdot \mathbf{x}_i - b + y_i + \varepsilon + \xi_i^*) = 0 \tag{4.11}$$

$$\beta_i\xi_i = (C - \alpha_i)\xi_i = 0 \tag{4.12}$$

$$\beta_i^*\xi_i^* = (C - \alpha_i^*)\xi_i^* = 0 \tag{4.13}$$

It is clear that the slack variables become zero if $0 < \alpha_i, \alpha_i^* < C$. Moreover, the combination of the first and second conditions could be written as follows:

$$b = y_i - \mathbf{w} \cdot \mathbf{x}_i - \varepsilon \quad for \quad 0 < \alpha_i < C \tag{4.14}$$

$$b = y_i - \mathbf{w} \cdot \mathbf{x}_i + \varepsilon \quad for \quad 0 < \alpha_i^* < C \tag{4.15}$$

The calculation of b is possible with the aforementioned equations. Nevertheless, it is indicated in [42] that b should be computed by calculating the mean over the "*free support vectors*" since the computation of b is sensitive. The "*free support vectors*" are the data vectors where Lagrange multipliers are not zero and smaller than C. However, when $\alpha_i = C$ or $\alpha_i^* = C$ for the data vectors outside the ε-tube, these vectors are named "*bounded support vectors*".

In SVR prediction, the best weight vector, $\mathbf{w}$, is yielded upon calculation of Lagrangian multipliers α_i and α_i^*[45] as in Equation 4.16.

$$\mathbf{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{4.16}$$

The data vectors which satisfy the condition of $0 < \alpha_i, \alpha_i^* < C$ are called support vectors. As a result, the best regression hyperplane can be expressed as in Equation 4.17

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}_i \cdot \mathbf{x} + b \tag{4.17}$$

where $\mathbf{x}_i$'s are the input vectors used for building the predictive model and $\mathbf{x}$ is the input vectors used for testing the model.

In a non-linear situation, the data is transformed into a space with higher dimension by using the "kernel function" $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j)$, eventually leading to Equation 4.18

$$f(\mathbf{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b \tag{4.18}$$

The details of implementation: LibSVM library which was developed by Chang and Lin [47] was used for the implementation of SVR. Radial basis function was used as the kernel shown in Equation 4.19

$$K(\mathbf{x}_i, \mathbf{x}) = e^{-\gamma|\mathbf{x}-\mathbf{x}_i|^2} \tag{4.19}$$

where γ is the width of the RBF-kernel which and regulates the organization of the independent variables in the data[48].

Obtaining reliable results by SVR was found to greatly depend on the optimization of the internal parameters C, ε and γ, which was implemented by a computationally time-consuming grid search [46, 45]. Before applying the grid search, the feature vectors were scaled in the range $[-1, +1]$ in order to avoid domination of some features and to simplify the numerical calculations [47]. The grid search method utilized the leave-one-out cross-validation method[43], which is described in details in the training/test set selection subsection of the prediction methods section. At first, a coarse grid values which were $\{2^0, 2^1, ..., 2^{15}\}$ for C, $\{2^0, 2^1, ..., 2^{15}\}$ for C, $\{2^{-15}, 2^{-14}, ..., 2^3\}$ for γ, and $\{2^{-15}, 2^{-14}, ..., 2^3\}$ for γ were set as recommended by Chang and Lin [47]. After determining the coarse parameters, a fine grid search was applied for which the values of the parameters were iterated by 1 for the parameter C, and 0.0001 for the parameters ε and γ. It should be noted that RMSE values tended to decrease as the value of the parameter C increased. However, high values of C causes the model to be complex and to have low generalization ability. The increase in the parameter C was cut at the point when the decrease of RMSE was no more than 0.0001. The optimal C, ε and γ values obtained by the grid search were also used in the SVR determination of the repeated random subsampling method[43], which is described in details in the training/test set selection subsection of the prediction methods section.

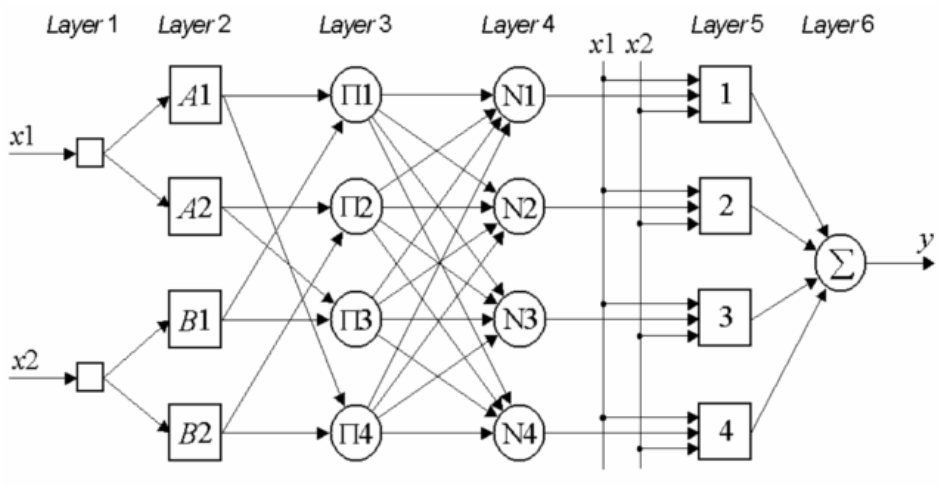


Figure 4.2: The ANFIS architecture

4.2.4 Adaptive Neuro-Fuzzy Inference System

Fuzzy inference is a method of construction of mapping from a given input to an output using the fuzzy logic. Then, the mapping forms a basis for decision making or recognizing patterns. There are successful applications of fuzzy inference systems in fields such as control, classification, decision support systems, expert systems, and computer vision[49].

ANFIS, which is based on Sugeno model, was proposed by Jang in 1993 [50]. Assume that two fuzzy IF-THEN rules of the first-order Sugeno kind are as follows:

1. R_1: IF x is A_1 and y is B_1, THEN $f_1 = p_1x + r_1y + t_1$
2. R_2: IF x is A_2 and y is B_2, THEN $f_2 = p_2x + r_2y + t_2$

where A_1, B_1, A_2 and B_2 are linguistic labels of fuzzy sets and p_1, r_1, t_1, p_2, r_2, and t_2 are parameters. If the consequent (THEN) part is chosen as a constant, then the rule is said to be a zeroth-order Sugeno type. ANFIS utilizes two types of parameters which belong to the membership functions in the antecedent (IF) and the polynomial functions in the consequent (THEN).

A six-layered ANFIS architecture can be defined as follows (Figure 4.2):

Layer 1: The first layer is the "input layer". Neurons of the "input layer" simply pass extraneous non-fuzzy signals to Layer 2.

Layer 2: Inputs are fuzzified in this layer. Adaptive nodes of this layer generate the parameters for the bell-shaped membership functions as in Equation 4.20

$$\mu_{A_i}(x) = \frac{1}{1 + [\frac{x - c_i}{a_i}]^{b_i}} \tag{4.20}$$

where μ_{A_i} is the membership function of A_i specifying the degree to which the given x satisfies the linguistic label A_i with the premise parameters a_i, b_i, and c_i. Although, bell-shaped membership functions are commonly used in this level, any continuous and piecewise differentiable functions, such as trapezoidal and triangular-shaped membership functions can be selected as node functions in this layer [50].

Layer 3: This layer is called the "rule layer". Here, every single neuron matches to a single "Sugeno-type fuzzy rule". A rule neuron gets inputs from the corresponding neurons of Layer 2 and computes the "firing strength" of the rule it represents. In an ANFIS, the conjunction of the rule antecedents is evaluated by the operator product [50]. Thus, the output of neuron i in Layer 3 is obtained as in Equation 4.21,

$$y_i^{(3)} = \prod_{j=1}^{k} x_{ij}^{(3)} \tag{4.21}$$

where $x_{ij}^{(3)}$ is the output of neuron j in Layer 2 which is connected to the neuron i in Layer 3.

Layer 4: This layer is referred to as the "normalization layer". Each neuron in this layer receives inputs from all neurons in the rule layer, and calculates the normalized firing strength of a given rule. The normalized firing strength is the ratio of the firing strength of a given rule to the sum of firing strengths of all rules. It represents the contribution of a given rule to the final result [50]. Thus, the output of neuron i in Layer 4 is determined as in Equation 4.22,

$$y_i^{(4)} = \frac{x_{ii}^{(3)}}{\sum_{j=1}^{n} x_{ij}^{(4)}} = \frac{\mu_i}{\sum_{j=1}^{n} \mu_j} = \overline{\mu_i} \tag{4.22}$$

Layer 5: This is the "defuzzification layer". Each neuron of Layer 5 make a connection to the corresponding normalization neuron, and besides gets initial inputs, i.e. x_1 to x_n. A defuzzification neuron calculates the weighted consequent value of a given rule as in Equation 4.23,

$$y_i^{(5)} = x_i^{(5)}[k_{i0} + k_{i1}x_1 + ... + k_{in}x_n] = \overline{\mu_i}[k_{i0} + k_{i1}x_1 + ... + k_{in}x_n] \tag{4.23}$$

where $x_i^{(5)}$ is the input and $y_i^{(5)}$ is the output of "defuzzification neuron" i in Layer 5, and $k_{i0}, k_{i1}, ..., k_{in}$ form a group of resulting parameters of the i^{th} rule, R_i.

Layer 6: It is the layer possessing a single summation neuron which sums up the outputs of all "defuzzification neurons" and calculates the final ANFIS output, y, as in Equation 4.24,

$$y = \sum_{i=1}^{n} x_i^{(6)} = \sum_{i=1}^{n} \overline{\mu_i}[k_{i0} + k_{i1}x_1 + ... + k_{in}x_n] \tag{4.24}$$

ANFIS does not need any prior knowledge for rule consequent parameters. It learns parameters and tunes up membership functions. ANFIS commonly uses a hybrid learning algorithm that combines the least-squares estimator and the gradient descent method, in which each epoch includes a forward pass and a backward pass. In the forward pass, a data set of training patterns is given to the ANFIS to compute neuron outputs layer by layer and to identify the resulting rule parameters. In the backward pass, the back-propagation algorithm is performed for propagating the error signals from the output back to the input. During propagation, the antecedent parameters are updated according to the chain rule [50].

The details of implementation: ANFIS computations utilized the leave-one-out cross-validation [43] and repeated random subsampling [43] methods, which are described in details in the training/test set selection subsection of the prediction methods section. The feature vectors obtained from X-, Y-, and Z-images by SFFS method were given seperately as inputs to compute a distinct ANFIS model for each feature set. By using training feature vectors, the fuzzy inference system was generated by subtractive clustering [51] which reduces the extensive requirements of time and memory. The observed pIC_{50} values were supplied as output to the fuzzy inference system for tuning up the bell-shaped membership functions and consequent parameters in the rules. Backpropagation[50] method was preferred as the learning system in the backward pass. ANFIS was trained for 10 epochs in order to avoid overfitting. The computed ANFIS models were used for predicting the pIC_{50} values of the test vectors which were selected by leave-one-out cross-validation and repeated random subsampling methods.

4.3 Statistical Analysis

The performance measures for prediction of pIC_{50} are root mean square error, RMSE, in Equation 4.25 and coefficient of determination, R^2, in Equation 4.26:

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{4.25}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{4.26}$$

where $y_i, \hat{y}_i$, and $\overline{y}$ are actual, observed, and mean response variables, respectively. In regression, the R^2 is a statistical measure of how well the regression function approximates the real data points. An R^2 of 1.0 indicates a perfect fit of a regression line

with observed data while R^2 of 0.0 shows no correlations at all. Tropsha et. al.[52, 53] reported two criteria indicating that a regression model produced by a QSAR study is predictive if the following conditions hold:

- $R^2_{LOOCV} > 0.5$ for leave-one-out cross-validation, and
- an average of $R^2 > 0.6$ for random subsampling which is also referred as leave-many-out cross-validation [53].

CHAPTER 5

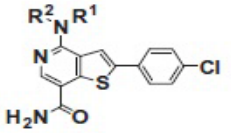
EXPERIMENTAL RESULTS

5.1 Checkpoint Kinase 1 and its inhibitors

The objective of cancer therapy is to specifically destroy cancerous cells while protecting healthy cells. The cell cycle of a healthy cell is arrested at the G1 and the G2/M cell cycle checkpoints by the p53 tumor suppressor protein upon DNA damage[54]. The cell cycle arrest prepares a cell for DNA repair, senescence or apoptosis[55, 56]. The function of free p53 in cell is down-regulated by mouse/human double minute-2, (M/H)DM2, oncoprotein[57]. If free p53 level is extremely lowered by a mutation in a cell, then the cell itself never arrest the cell cycle, transforming the cell into a cancerous cell with uncontrollable mitotic cell divisions. Interestingly, inhibition or knockout of checkpoint kinase 1 (CHK1), a serine / threonine protein kinase, arrests the G2 or the S cell cycle checkpoint in p53-deficient cancer cells[58, 59], potentiating the efficacy of DNA damaging anticancer agents such as 5-fluorouracil or doxorubicin. In that regards, some CHK1 inhibitors have already been developed and patented[60]. Considering that almost half the human cancers are p53-deficient, it then becomes an attractive choice to guide the current computer-assisted drug discovery technology through the development of novel and more potent inhibitors of CHK1 in cancer therapy.

5.1.1 Chemical Structures

In the experiments, 57 checkpoint kinase 1 (CHK1) inhibitors derived from the thienopyridine pharmacophore were adopted from Zhao et al.[1], whose chemical structures and the corresponding IC_{50} values in nM concentration were listed in the Figures 5.1, 5.2, and 5.3. The corresponding figures were obtained from [1]. Briefly, the IC_{50} values for the inhibitors vary between 1 nM and $2.7x10^4$ nM. The observed IC_{50} for CHK1 inhibitors were converted to nanomolar pIC_{50} values by $-log_{10}IC_{50} \times 10^{-9}$[61] for CIFAP computations.



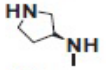
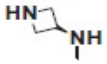
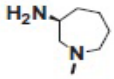
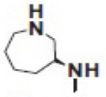
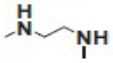
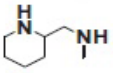
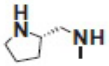
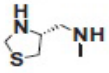
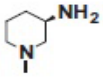
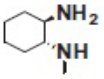
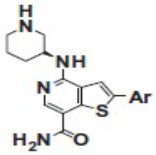
Compound	R^1	R^2	CHK1 IC_{50} (nM)
1	—	—	2
2a	H		3
2b	H		14
2c	H		707
2d			746
2e	H		3
2f	H		7
2g	H		292
2h	H		197
2i	H		937
2j			148
2k	H		211
2l	H		27,143

Figure 5.1: SAR at 4-position of thienopyridine



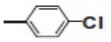
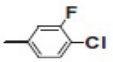
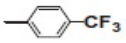
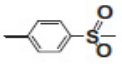
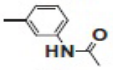
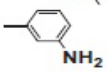
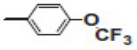
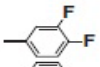
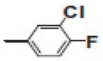
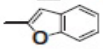
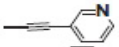
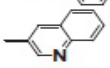
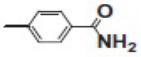
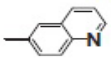
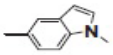
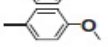
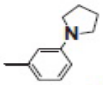
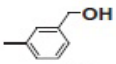
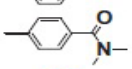
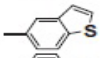
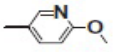
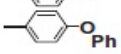
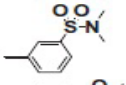
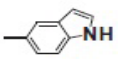
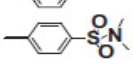
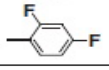
Compound	Ar	IC_{50} (nM)	Compound	Ar	IC_{50} (nM)
2a		3	39		9
19		24	40		5
20		5	41		2
21		4	42		7
22		2	43		5
23		2	44		9
24		3	45		60
25		2	46		4
26		5	47		2
27		4	48		67
28		21	49		34
29		5	50		1
30		19	51		5
31		3	52		8
32		3	53		3
33		3	54		1
34		4	55		3
35		4	56		20
36		60	57		16
37		16	58	Br	56
38		9	59	H	717

Figure 5.2: SAR at 2-position of thienopyridine

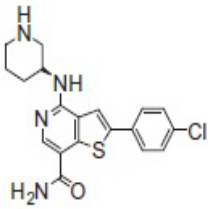
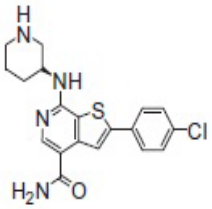
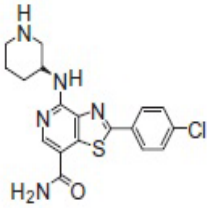
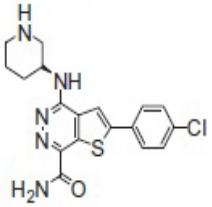
Compound	Structure	CHK1 IC_{50} (nM)
2a		3
60		4582
69		9
70		1

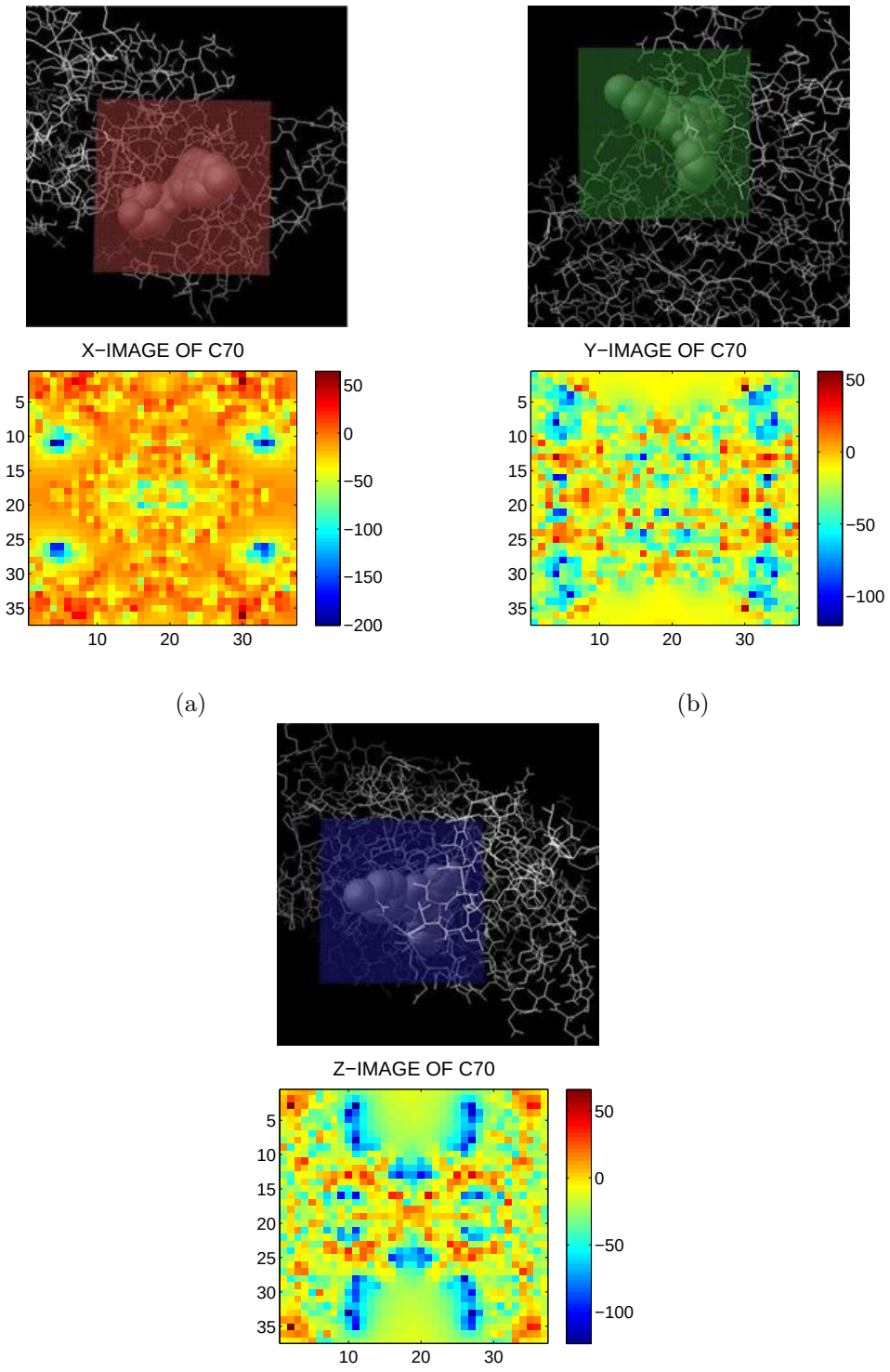
Figure 5.3: SAR of core modification of thienopyridine

5.1.2 Data Modelling Phase

As described in the data modelling methods section, 3D electrostatic potential grid cubes with center coordinates of (20, –3, 11) and a size of $37 \times 37 \times 37$ for 57 ligands in complex with CHK1 were compressed by summing up electrostatic potential values at grid points in orthogonal (X, Y and Z) directions into three 2D images with 37×37 pixels for each complex. The compression process for the complex of compound 70 and CHK1 is exemplified in Figure 5.4 which shows the grid cubes through the X-, Y-, and Z-axis from left to right at the top of the figure, and the corresponding 2D images at the bottom of the figure. It should be noted here that the X-Ray, PDB ID: 3PA3[1], and docked coordinates of bound compound 70 generated almost identical 2D compressed images. Therefore, only the docked coordinates of bound compound 70 was taken as granted and correlated with its experimental binding affinity in the following sections. It is also essential that the cubic grid be as small as possible and cover the inner boundaries of the binding site of the receptor as well as the ligand itself in order to gain more meaningful information from compressed 2D images. The

Before the prediction phase of CIFAP, 2D compressed images were initially converted to linear vectors, which were then used for feature selection to generate feature vectors by SFS and SFFS. It should be noted that BFS is not applicable in this case since the number of features is much higher than the number of samples. The best RMSE values determined by MATLAB using Equation 4.25 and data obtained from SFS computations were found to be 0.6355 for leave-one-out cross-validation of the X-images, 0.5298 for leave-one-out cross-validation of Y-images, and 0.6297 for leave-one-out cross-validation of Z-images. In order to visualize the patterns, the feature vectors obtained by SFS were then converted into 2D X-pattern, Y-pattern, and Z-pattern images, shown in Figure 5.5, by breaking the feature vectors into 37 fragments, each having 37 points and the selected features (black pixels in the figures), and sequentially stacking the fragments. The white area in each image of Figure 5.5 represents the sum of electrostatic potential values through that direction which were not found to be informative and were not used in regression analysis. SFS algorithm selected 14, 18, and 16 features for X-pattern, Y-pattern, and Z-pattern images respectively. SFFS implementation gave rise to better RMSE values by replacing the redundant features with more informative ones. As demonstrated in Figure 5.6, SFFS algorithm yielded 15 features for X-pattern images, 21 features for Y-pattern images, and 25 features for Z-pattern images. The best RMSE values obtained by SFFS implementation were found to be 0.6355 for leave-one-out cross-validation of the X-images, 0.4749 for leave-one-out cross-validation of Y-images, and 0.3962 for leave-one-out cross-validation of Z-images. It is clearly seen that the RMSE values of Y- and Z-images were highly improved while the RMSE value of X-image stayed the same.

The lowest RMSE values which were obtained by Y-images and Z-images shows that the information conveyed by selected features of Y-images and Z-images has the better



(a) (b)

(c)

Figure 5.4: An exemplary illustration of the 3D electrostatic potential (EP) grid for the CHK1-compound 70 complex and the corresponding compressed 2D images. A view of the EP grid through (a) the X-axis (top) and the corresponding compressed X-image (bottom), (b) Y-axis (top) and the corresponding compressed Y-image (bottom), and (c) Z-axis (top) and the corresponding compressed X-image (bottom). The color scales for the compressed images are shown on the right side.

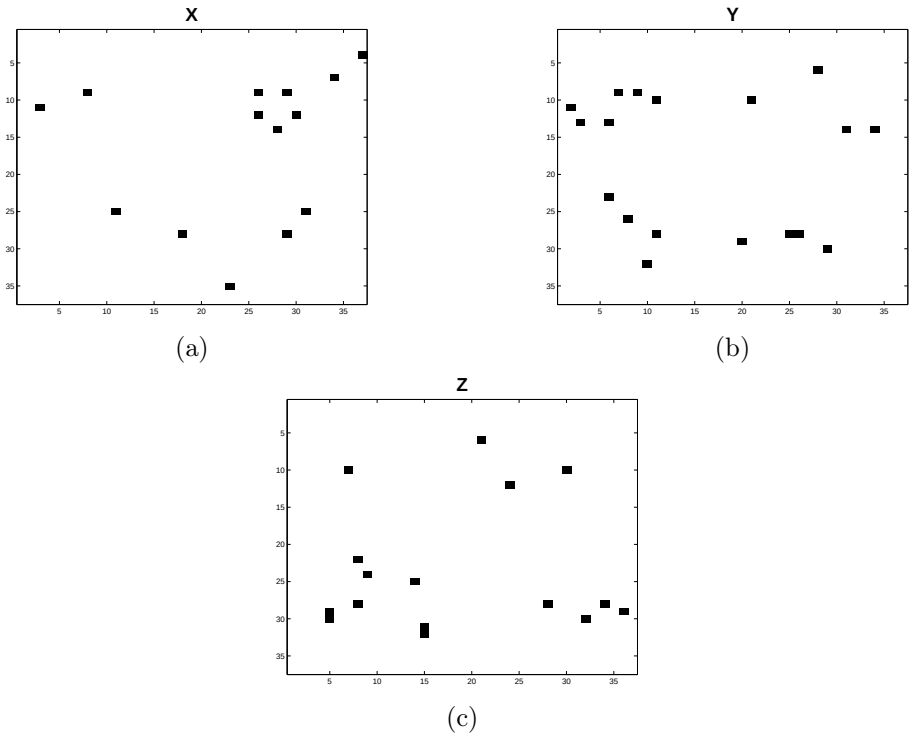


Figure 5.5: Two dimensional X-, Y- and Z-pattern images of CHK1-ligand complexes obtained by Sequential Forward Selection, SFS.

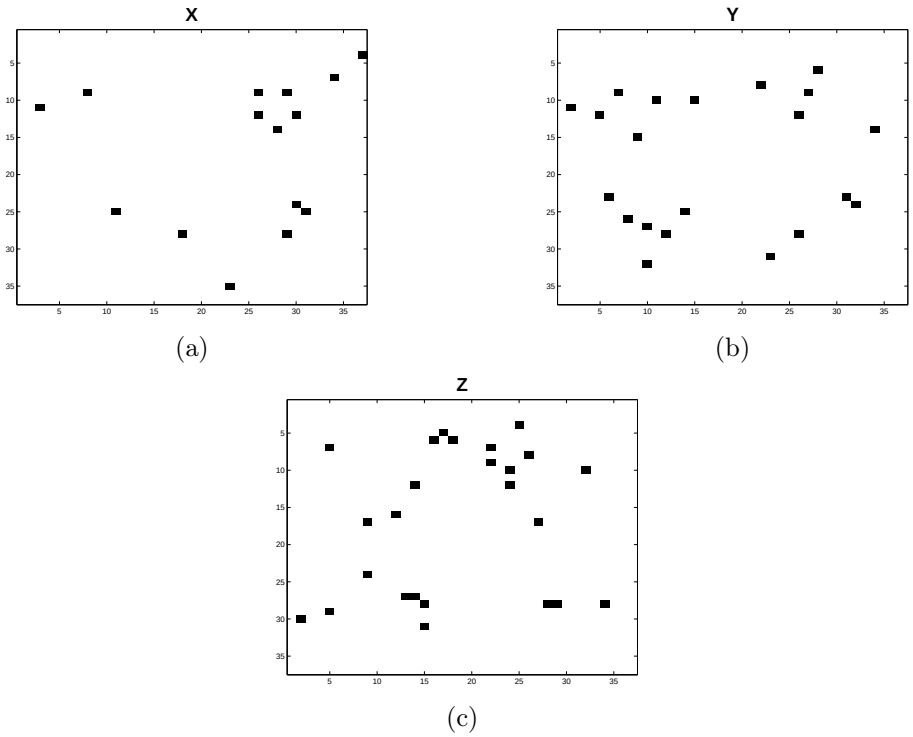


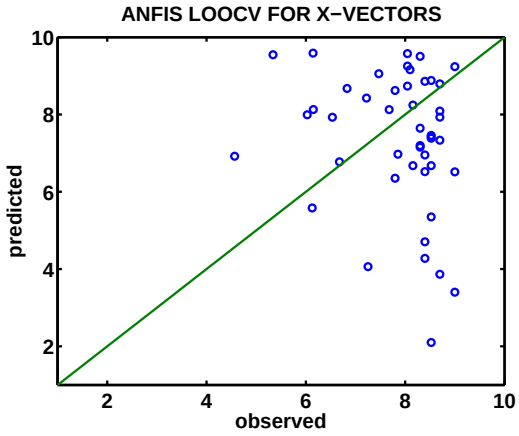
Figure 5.6: Two dimensional X-, Y- and Z-pattern images of CHK1-ligand complexes obtained by Sequential Floating Forward Selection, SFFS.

explanation of the pIC_{50} values. It is meaningful that the patterns are grouped around the margins of the images, which correspond to the binding interface between the ligand and the binding site of the receptor. As far as the CHK1-compound 70 complex is concerned, the lowest RMSE values are obtained with the 2D Y-images and Z-images due most likely to the greatest 2D area occupied by the ligand when looking at the cubic grid of the binding site through the Y-axis and Z-axis as in Figures 5.4c and 5.4b (top), while the ligand seems to take up less space when looking at the cubic grid through the X-axis, Figure 5.4a (top). Ideally an angle showing a bound ligand with a shallower and broader size, especially as in the Z-axis view of bound compound 70 shown in Figure 5.4c (top), should provide more detailed information on drug-receptor interactions in the form of compressed electrostatic potentials.

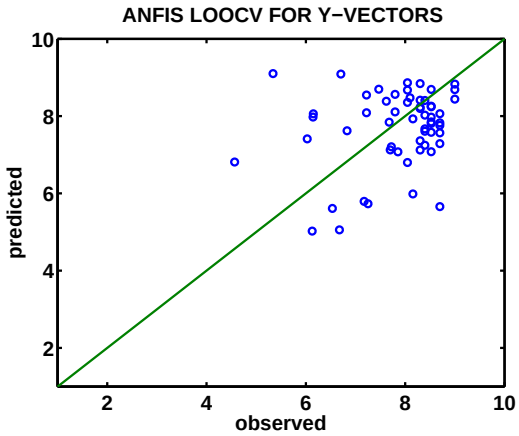
5.1.3 Prediction Phase

The features selected by SFFS were then used for prediction by three independent regression analysis methods; Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Regression (SVR), and Partial Least Squares Regression (PLSR). Figure 5.7 shows 2D affinity correlation plots showing the observed binding affinities on the x-axis in pIC_{50} versus the binding affinities on the y-axis in pIC_{50}, which were predicted by the ANFIS applying the leave-one-out cross-validation method, using 57 different sets of test vectors selected from the X-feature, Y-feature ,Z-feature vectors of the CHK1-ligand complexes. Although the Y-affinity correlation graph, Figure 5.7b, and the Z-affinity correlation graph, Figure 5.7c, seem to have less data distribution and possibly better correlations as compared to that of the X-affinity correlation plot in Figure 5.7a, as a matter of fact an RMSE of 1.0868 obtained for the Z-feature vectors, and an RMSE of 1.1999 obtained for the Y-feature vectors indicate a weak correlation between the predicted and observed binding affinities, suggesting that the leave-one-out cross validation applied by ANFIS may not be suitable for optimal correlations in all dimensions.

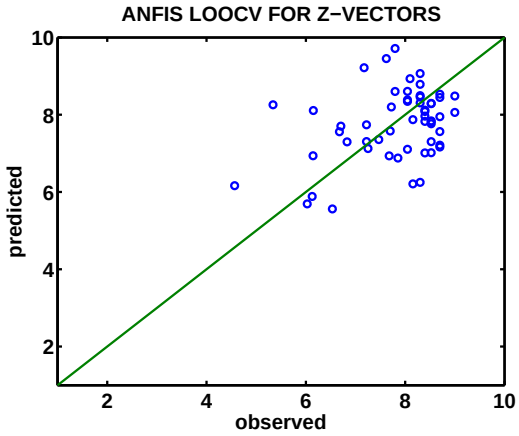
Aside from the leave-one-out implementation of ANFIS, ANFIS was also carried out by applying the random subsampling set selection method, which used 1000 different training sets, having 47 feature vectors in each set, and test sets, having 10 feature vectors in each set, by shuffling 57 X-feature, Y-feature, and Z-feature vectors of the CHK1-ligand complexes separately. RMSE and R^2 values for the best three ANFIS results (Random-1, Random-2 and Random-3) and average RMSE and R2 values out of 1000 randomly selected training and test sets are given in Table 5.1 for the X-feature ,Y-feature, and Z-feature vectors. As seen in Table 5.1, the Y-feature and Z-feature vectors gave rise to the lowest RMSE and the highest R^2 values than those of the X-feature vectors, suggesting that the Y-feature and Z-feature vectors provide more useful information in correlating the predicted binding affinities with the corresponding observed binding affinities. This finding is also visualized in Figure 5.8 showing affinity



(a)



(b)



(c)

Figure 5.7: Affinity correlation plots constructed upon ANFIS determination of leave-one-out cross validation. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 57 different testing sets selected from the X-pattern (a), Y-pattern (b), and Z-pattern (c) images of the CHK1-ligand complexes.

Table5.1: R^2 and RMSE values for an average and the best 3 ANFIS determination of random subsampling set selection method out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CHK1-ligand complexes.

ANFIS	test		train	
X	R^2	RMSE	R^2	RMSE
Random-1	0.4157	1.0237	0.9801	0.1185
Random-2	0.3197	1.0077	0.9895	0.0884
Random-3	0.2711	0.7131	0.9823	0.1292
Average	0.0947	3.5604	0.9832	0.1190

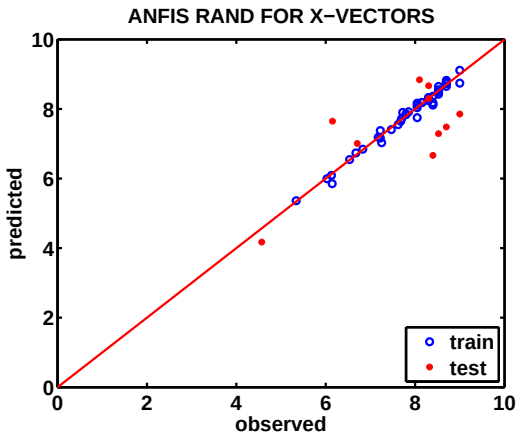
ANFIS	test		train	
Y	R^2	RMSE	R^2	RMSE
Random-1	0.7379	0.6525	0.9495	0.1948
Random-2	0.6813	0.7306	0.9472	0.1975
Random-3	0.6084	0.7333	0.9456	0.2083
Average	0.1762	1.2035	0.9544	0.2008

ANFIS	test		train	
Z	R^2	RMSE	R^2	RMSE
Random-1	0.7082	0.7052	0.9245	0.2354
Random-2	0.6992	0.6881	0.8455	0.3419
Random-3	0.6948	0.6500	0.8774	0.3123
Average	0.1945	1.4817	0.8903	0.3084

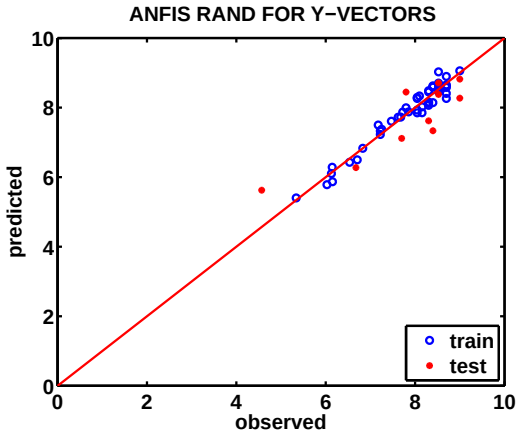
correlations plotted for the observed (x-axis) versus predicted (y-axis) binding affinities (pIC_{50}) using the best random subsampling set selection (Random-1 in Table 5.1) for the X-feature, Figure 5.8a, Y-feature, Figure 5.8b, and Z-feature, Figure 5.8c, vectors of 57 CHK1-ligand complexes, 10 of which were used for testing and the rest were used for training.

As compared to the ANFIS determination of leave-one-out cross validation, the ANFIS determination of the random subsampling set selection gave rise to more correlative results with the Y-feature and Z-feature vectors which are able to explain 70% of variance in pIC_{50} values. In terms of the ANFIS application of the CHK1-ligand complexes and possibly of other complex systems, it should be known that learning depends not only on the nature of the features but also design of the training and test sets as well. Therefore, it is essential that the use of the features as well as the training and test sets be optimized for a given complex system.

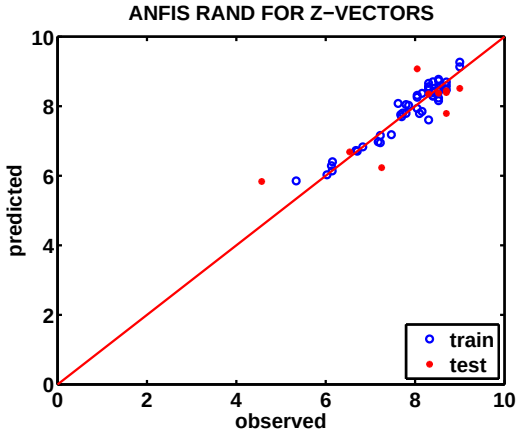
Another method of regression analysis, that was applied to predict the binding affinities of the CHK1 inhibitors, is the Support Vector Regression, SVR, which is described in details in the SVR subsection of the prediction methods section. As mentioned in the SVR subsection of the Prediction Methods section, the C parameter is a trade-off between error tolerance and model complexity, and its value should be optimized along with the other internal parameters ε, the radius of the ε-tube, and γ, the width



(a)



(b)



(c)

Figure 5.8: Affinity correlation plots constructed by ANFIS determination of random subsampling set selection. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.1) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 57 CHK1-ligand complexes, 10 of which were used for testing (red dots) and the rest were used for training (blue circles).

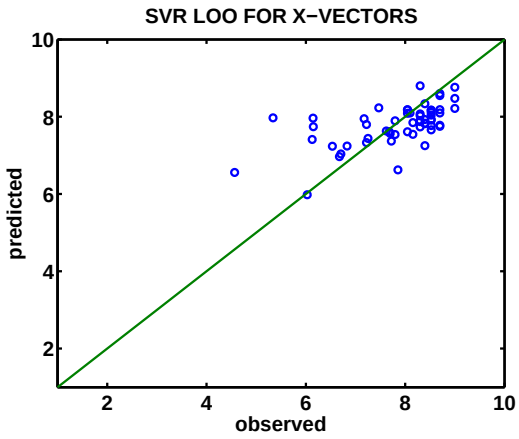
Table5.2: Optimal values of the SVR parameters C, ε, and γ for CHK1-ligand complexes.

	C	γ	ε
X-image	2000	0.002	0.0005
Y-image	2000	0.001	0.0078
Z-image	2000	0.0009	0.0313

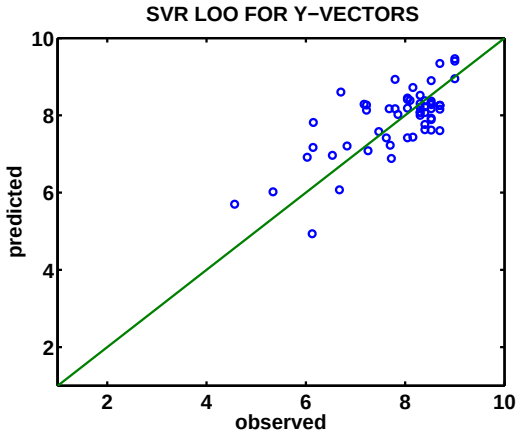
of the RBF-kernel to regulate the distribution of independent variables in the features. In order to minimize the error of approximation regardless of the complexity of the CHK1-ligand complex systems, a range of C values between 1 and 32,768 (2^{15}) were included in the grid search by the SVR determination of leave-one-out cross validation, which was carried out for each of the 57 feature vectors to determine the lowest RMSEs that give rise to optimal values for the internal parameters C, ε, and γ. In general, deviations in RMSE caused by slight variations in C were observed to be greater than those of ε and γ. It was found that a value of 2000 for C afforded the lowest RMSE and the highest R^2 values as well as the optimal C, ε, and γ values for the X-feature, Y-feature, and Z-feature vectors, which are presented in Table 5.2.

Using the optimal C, ε, and γ values given in Table 5.2 and 57 X-feature, Y-feature, and Z-feature vectors of the CHK1-ligand complexes, the binding affinities of 57 bound ligands were predicted by the SVR determination of leave-one-out cross validation. The resulting predicted affinities (pIC_{50}) obtained for the testing feature vectors were then correlated with the corresponding observed (experimental) binding affinities (pIC_{50}) of the CHK1-ligand complexes as shown in Figure 5.9a for the X-feature vectors, Figure 5.9b for the Y-feature vectors, and Figure 5.9c for the Z-feature vectors. Figure 5.9c clearly shows that the best correlations between the predicted and observed binding affinities were obtained with the testing data obtained from the Z-feature vectors. Interestingly, the affinity correlation profile for the Z-feature vectors in Figure 5.10c looks much better than the affinity correlation profile obtained by the ANFIS determination of leave-one-out cross validation, shown in Figure 5.8c. An RMSE of 0.5940 obtained for the Z-feature vectors, expresses a better correlation between the predicted and observed binding affinities, suggesting that the leave-one-out cross validation applied by SVR may be a better prediction system than the leave-one-out cross validation applied by ANFIS. Moreover, an R^2 of 0.6098 obtained for the Z-feature vectors, shows that the model using SVR with Z-feature vectors is more predictive than the model using ANFIS with Z-feature vectors according to the first Tropsha criterion [52, 53].

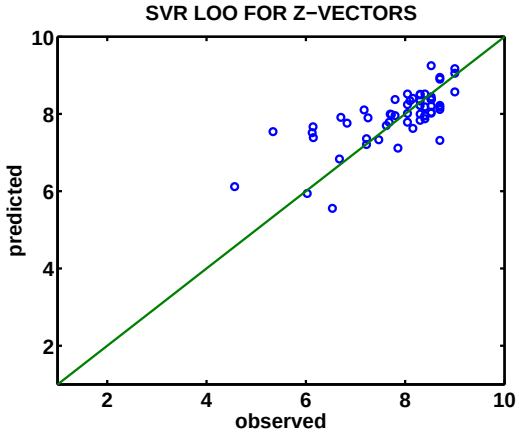
SVR computations were also implemented by utilizing the random subsampling set selection method, described in details in the selection of training and test sets subsection of the prediction methods section. The SVR determination of random subsampling set selection used the optimal C, ε and γ values given in Table 5.2. R^2 and RMSE values for the best three SVR results (Random-1, Random-2 and Random-3) and average RMSE and R^2 values out of 1000 randomly selected training and test data



(a)



(b)



(c)

Figure 5.9: Affinity correlation plots constructed upon SVR determination of leave-one-out cross validation. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the optimal C, ε and γ values given in Table 5.2 and 57 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CHK1-ligand complexes.

Table5.3: R^2 and RMSE values for an average and the best three SVR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CHK1-ligand complexes.

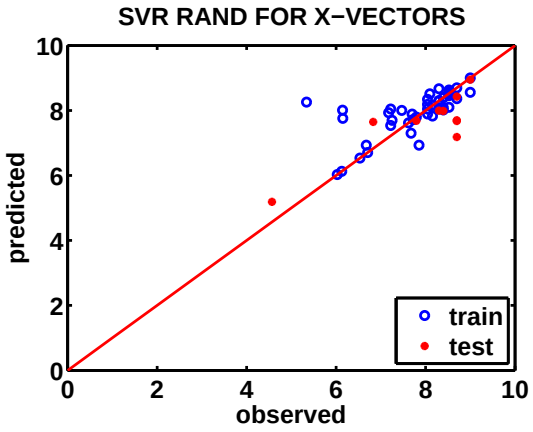
SVR	test		train	
X	R^2	RMSE	R^2	RMSE
Random-1	0.8875	0.2886	0.5897	0.6204
Random-2	0.8853	0.4016	0.5320	0.6109
Random-3	0.8859	0.4418	0.4757	0.6135
Average	0.5112	0.6464	0.6488	0.5764

SVR	test		train	
Y	R^2	RMSE	R^2	RMSE
Random-1	0.8938	0.2272	0.7468	0.4981
Random-2	0.8512	0.4077	0.7325	0.4793
Random-3	0.8814	0.4199	0.6646	0.5099
Average	0.5245	0.6789	0.7780	0.4556

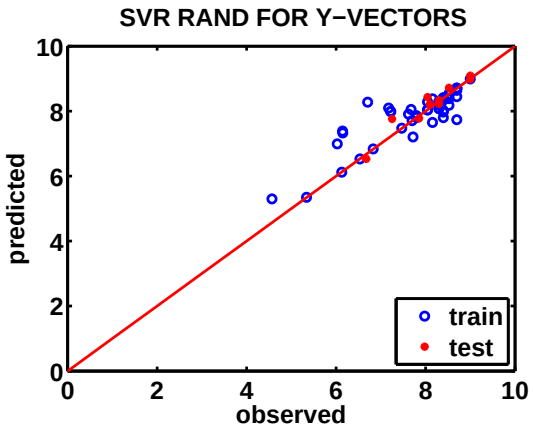
SVR	test		train	
Z	R^2	RMSE	R^2	RMSE
Random-1	0.9276	0.1889	0.8309	0.3994
Random-2	0.9215	0.2715	0.8146	0.4082
Random-3	0.9087	0.2818	0.8196	0.4001
Average	0.5636	0.6271	0.8358	0.3826

sets are listed in Table 5.3 for the X-feature, Y-feature and Z-feature vectors. In a similar manner to the RMSE results shown in Table 5.1 for the ANFIS determination of random subsampling set selection, the results presented in Table 5.3 for the SVR determination of random subsampling set selection indicate that the Z-feature vectors yield more correlative predicted binding affinities with an average RMSE of 0.6271 determined out of 1000 test sets and an RMSE of 0.1889 for the best test set (Random-1). A correlation of the predicted binding affinities (pIC_{50}) belonging to the Random-1 training and test sets in Table 5.3, which gave the lowest RMSEs by the SVR determination of random subsampling set selection, versus the observed binding affinities (pIC_{50}) of 57 CHK1 inhibitors, published by Zhao et al.[1], are shown in Figures 5.10a-5.10c for the X-feature, Y-feature, and Z-feature vectors, respectively, of 57 CHK1-ligand complexes, 10 of which were used for testing and the rest were used for training. The affinity correlation profile shown in Figure 5.10c for 57 CHK1 inhibitors is indeed in good agreement with the low RMSE values given in Table 5.3 for the Random-1 training/test sets selected from the Z-feature vectors.

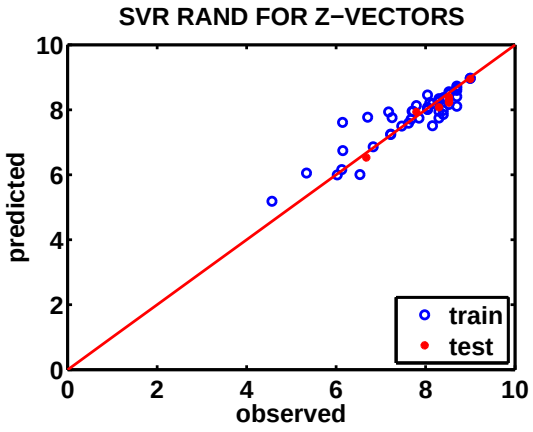
The last method of regression analysis, that was applied to predict the binding affinities of for the CHK1 inhibitors, is Partial Least Squares Regression, PLSR, which is a linear regression method and described in details in the PLSR subsection of the prediction methods section. The correlation between the observed (x-axis) and predicted (y-axis) binding affinities obtained by the PLSR determination of the leave-one-out



(a)

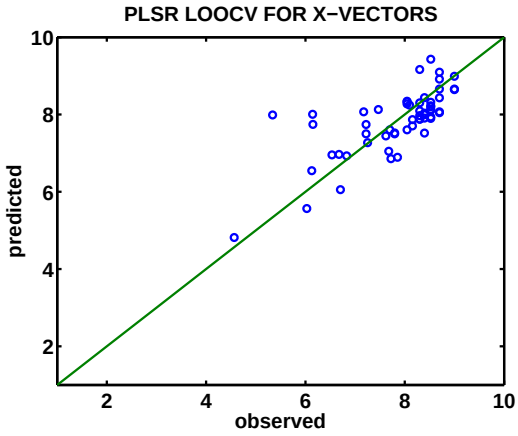


(b)

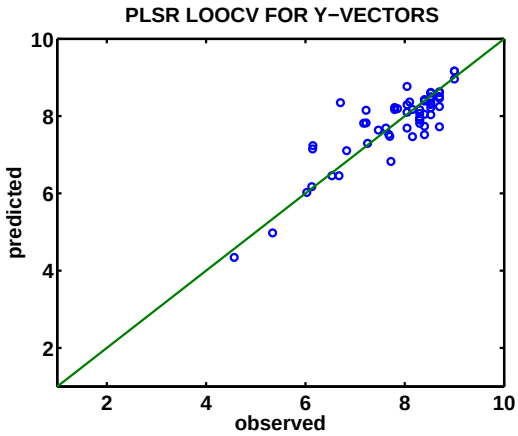


(c)

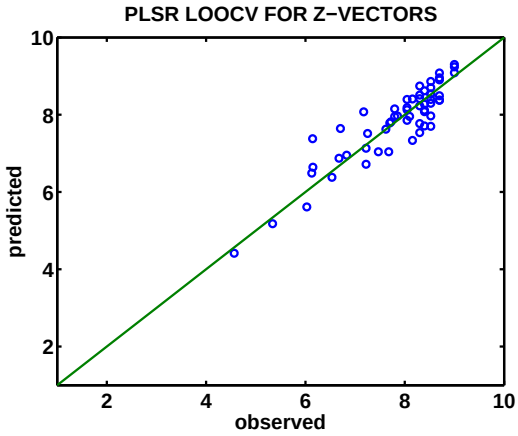
Figure 5.10: Affinity correlation plots constructed by SVR determination of random subsampling set selection. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.3) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 57 CHK1-ligand complexes, 10 of which were used for testing (red dots) and the rest were used for training (blue circles).



(a)



(b)



(c)

Figure 5.11: Affinity correlation plots constructed upon PLSR determination of leave-one-out cross validation. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 57 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CHK1-ligand complexes.

cross-validation, using 57 different sets of test vectors selected from the X-feature, Y-feature, and Z-feature vectors of the CHK1-ligand complexes, is demonstrated in Figure 5.11. As in other regression models discussed earlier in this chapter, Z-feature vectors resulting in an RMSE of 0.4109 and an R^2 of 0.8133 are found to be more informative than X-feature and Y-feature vectors.

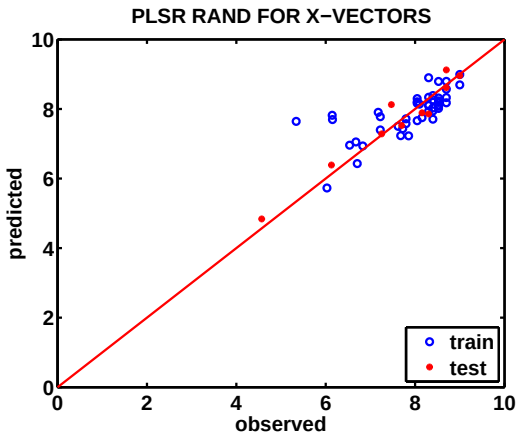
Indeed, the RMSE values given in Table 5.4 to compare the experimental binding affinities (pIC_{50}) of 57 CHK1 inhibitors, published by Zhao et al.[1], with those predicted by the PLSR, SVR, and ANFIS determination of leave-one-out cross validation, which used the Z-feature vectors, further support the aforementioned observation, strongly suggesting that the PLSR determination of leave-one-out cross validation gave rise to more reliable predicted binding affinities by an RMSE of 0.4109 as compared to both the SVR and ANFIS determination of leave-one-out cross validation, which resulted in RMSE values of 0.5940 and 1.0868, respectively. Moreover, high correlation ($R^2 = 0.8133$) yielded by PLSR determination of the leave-one-out cross-validation using Z-feature vectors indicates that the relationship between the data model and binding affinity is more likely linear since PLSR, a linear regression algorithm, outperformed two strong regression methods SVR which is a non-linear method with the use of RBF kernel, and ANFIS which is a fuzzy regression method.

In addition to leave-one-out cross-validation implementation, PLSR was applied by using the random subsampling set selection method by randomly selecting 47 vectors of CHK1-ligand complexes for training and 10 vectors of CHK1-ligand complexes for testing without repetition. Table 5.5 reports the R^2 and RMSE values for the best three SVR results (Random-1, Random-2 and Random-3) and average RMSE and R^2 values out of 1000 randomly selected training and test data sets belonging to the X-feature, Y-feature and Z-feature vectors. Along with the results of SVR and ANFIS determination of random subsampling set selection, PLSR determination of random subsampling set selection found out that the Z-feature vectors generate better predictors with an average RMSE of 0.5388 determined out of 1000 test sets and an RMSE of 0.1717 for the best test set (Random-1). In Figures 5.12a-5.12c for the X-feature, Y-feature, and Z-feature vectors, the correlation between the observed and predicted binding affinities (pIC_{50}) belonging to the Random-1 training and test sets in Table 5.5, is illustrated. Although the training/test sets selected from the Y-feature vectors also yielded a low average RMSE of 0.5535 and a high average R^2 of 0.6373, Z-feature vectors provided a better explanation of the variability in the model considering a lower average RMSE of 0.5388 and a higher average R^2 of 0.7141.

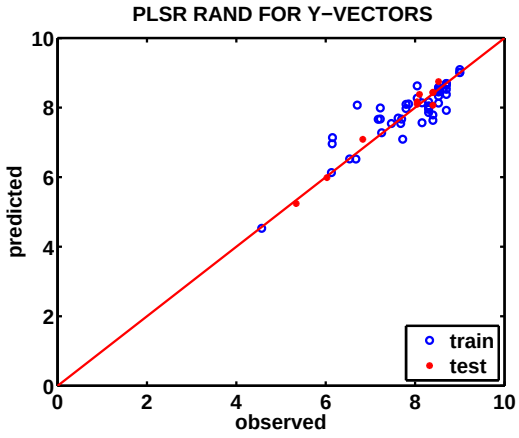
Based on the analysis of all the regression and set selection methods applied on the 57 CHK1-ligand systems, it was observed that it was indeed the Z-feature vectors obtained from the Z-images that presented more reliable and correlative information to address the learning and prediction features of CIFAP. Therefore, it is more plausible to opt the Z-images as the major sources of collecting information for learning and thus predicting binding affinities of novel thienopyridine analogs. Although the concept of

Table5.4: RMSE comparison between observed binding affinities (pIC_{50}) for 57 CHK1 inhibitors, published by Zhao et al.[1], and the corresponding binding affinities (pIC_{50}) predicted by the PLSR, SVR, and ANFIS determination of leave-one-out cross validation for the Z-feature vectors of the testing data sets.

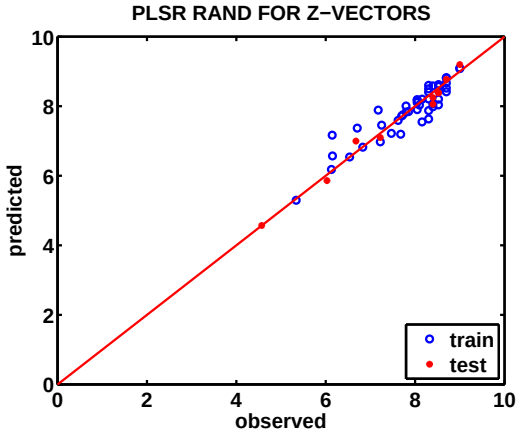
No.	**pIC_{50}**	**PLS**	**SVR**	**ANFIS**	**No.**	**pIC_{50}**	**PLS**	**SVR**	**ANFIS**
1	8.6990	9.0802	8.5814	7.9482	35	8.3979	8.1192	7.9728	7.9640
19	7.6198	7.6262	7.6354	9.454	36	7.2218	6.7173	7.2321	7.3062
20	8.3010	8.4158	8.4011	9.065	37	7.7959	8.1512	8.3521	8.6034
21	8.3979	8.0800	7.9025	7.012	38	8.0458	8.1836	8.2123	8.6053
22	8.6990	8.9547	8.1661	7.165	39	8.0458	8.3955	8.5173	8.3451
23	8.6990	8.3763	8.1170	7.564	40	8.3010	7.7712	8.0305	8.4584
24	8.5229	8.7002	8.4405	7.767	41	8.6990	8.9038	8.9185	8.5310
25	8.6990	8.3792	8.2262	8.446	42	8.1549	8.4085	8.3877	7.8738
26	8.3010	7.5363	7.8269	8.494	43	8.3010	8.5079	8.4389	8.3080
27	8.3979	8.6195	8.4182	8.120	44	8.0458	8.1289	7.7554	7.1065
28	7.6778	7.0414	7.8371	6.935	45	7.2218	7.1325	7.5735	7.7385
29	8.3010	8.7442	8.5743	6.250	46	8.3979	7.7111	7.9410	8.0852
2a	8.5229	7.7007	8.3175	8.294	47	8.6990	8.4916	7.8556	7.2124
2b	7.8539	7.9671	7.2053	6.879	48	7.1739	8.0757	8.1005	9.2178
2c	6.1506	6.6419	7.0371	8.109	49	7.4685	7.0416	7.3380	7.3563
2d	6.1273	6.4896	7.2529	5.885	50	9.0000	9.2396	9.1386	8.0614
2e	8.5229	8.8622	8.6881	7.303	51	8.3010	8.2312	8.1397	8.7869
2f	8.1549	7.3352	7.4241	6.209	52	8.0969	7.9556	8.3526	8.9328
2g	6.5346	6.3794	5.7454	5.561	53	8.5229	8.5333	8.6080	7.0164
2h	6.7055	7.6442	7.8961	7.705	54	9.0000	9.3011	8.5512	11.1004
2i	6.0283	5.6129	5.7120	5.691	55	8.5229	8.2958	8.3550	8.2865
2j	6.8297	6.9519	7.6451	7.298	56	7.6990	7.7859	8.1327	7.5784
2k	6.6757	6.8728	6.4813	7.558	57	7.7959	7.9553	8.1287	9.7156
2l	4.5663	4.4159	6.1559	6.164	58	7.2518	7.5155	7.7758	7.1252
30	7.7212	7.8154	7.9624	8.201	59	6.1445	7.3787	7.6180	6.9379
31	8.5229	8.3959	8.3149	7.841	60	5.3389	5.1800	7.1871	8.2565
32	8.5229	8.5356	8.2682	7.818	69	8.0458	7.8547	8.0412	8.3827
33	8.5229	7.9691	7.9267	10.4276	70	9.0000	9.0852	8.5874	8.4834
34	8.3979	8.2727	8.1189	7.8277	**RMSE**		**0.4109**	**0.5940**	**1.0868**



(a)



(b)



(c)

Figure 5.12: Affinity correlation plots constructed by PLSR determination of random subsampling set selection. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.5) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 57 CHK1-ligand complexes, 10 of which were used for testing (red dots) and the rest were used for training (blue circles).

Table5.5: R^2 and RMSE values for an average and the best three PLSR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CHK1-ligand complexes.

PLSR	test		train	
X	R^2	RMSE	R^2	RMSE
Random-1	0.9363	0.3260	0.5280	0.5850
Random-2	0.9218	0.3206	0.5808	0.5850
Random-3	0.9169	0.3760	0.5402	0.5810
Average	0.5163	0.6868	0.6894	0.5252

PLSR	test		train	
Y	R^2	RMSE	R^2	RMSE
Random-1	0.9707	0.1842	0.7861	0.4228
Random-2	0.9636	0.2595	0.7210	0.4179
Random-3	0.9607	0.2302	0.7781	0.4198
Average	0.6373	0.5535	0.8388	0.3773

PLSR	test		train	
Z	R^2	RMSE	R^2	RMSE
Random-1	0.9805	0.1717	0.8542	0.3150
Random-2	0.9669	0.1755	0.8906	0.3129
Random-3	0.9563	0.3077	0.8313	0.3136
Average	0.7141	0.5388	0.9129	0.2776

the Z-images varies with different coordinates and angles of a 3D grid map for the same or different protein-ligand complexes, in this particular CHK1-ligand complex systems they are called the Z-images.

It appears that the PLSR determination of random subsampling set selection, which gave the lowest average RMSE, 0.5388 for test sets and 0.2776 for training sets, for the Z-feature vectors in Table 5.5, should be preferred to the SVR determination of random subsampling set selection, which led to an average RMSE of 0.6514 for testing and 0.3848 for training with Z-feature vectors in Table 5.3, and also ANFIS determination of random subsampling set selection, which led to an average RMSE of 1.2035 for testing and 0.2008 for training with Z-feature vectors in Table 5.1. In addition, it should be noted that PLSR implementation is faster than SVR implementation which requires an exhausting process of parameter optimization, and than ANFIS implementation which requires neural network training. As a result, the utilization of PLSR for the Z-feature vectors should be the best choice of methodology for reliability on learning and prediction in the development of more potent novel CHK1 inhibitors (the thienopyridine analogs) by CIFAP.

5.2 Caspase 3 and its inhibitors

Apoptosis is the mechanism of the cell death which has a major role in many biological processes like growth, demolition of unhealthy cells, and immune system activities[62]. Caspases play important role in regulating apoptosis [63]. Caspase activity is related with a number of diseases, including neurodegenerative diseases, stroke, cardiomyopathy, ischemia and cancers [63, 2, 64]. For instance, Caspase3 becomes active in Alzheimer's disease and affects apoptosis of neurons [63]. Controlling cell death by inhibiting caspases is thought to be very helpful in the therapy of above diseases [64]. Recent studies have claimed that isatin sulfonamide analogues could be promising inhibitors of Caspase3 in medical therapy [65, 66, 67].

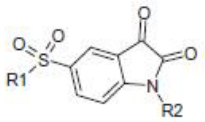
5.2.1 Chemical Structures

35 Caspase 3 (CASP3) inhibitors derived from isatine sulfonamide pharmacophore were collected from Wang et. al. [2] who, in fact, studied on 59 isatine sulfonamide analogues. 14 cyano compounds [68] were eliminated immediately since it is reported by Hasegawa et. al. [69] that cyano compounds interact with CASP3 through a different biological mechanism. The chemical structures and calculated pIC_{50} values which change within limits from 5.84 to 8.44 were listed in the Figures 5.13 and 5.14, adopted from the study of Hasegawa et. al [69].

5.2.2 Data Modelling Phase

The reference ligand in this series of experiments was compound 1 which was at the top of the list in Figure 5.13. 3D X-Ray structure of CASP3-compound 1 complex, PDB ID: 1GFW [2], was used for modelling 35 inhibitors according to the best pose and the minimum energy level obtained from docking process. The coordinates of the docked compounds and associated experimental binding affinities were utilized in the modelling and prediction experiments. For each complex, 2D images with 37×37 pixels were acquired by compressing 3D electrostatic potential grid cubes with center coordinates (39,36,27) and size by $37 \times 37 \times 37$ through the summation of electrostatic potential values at grid points in orthogonal (X, Y and Z) directions. Figure 5.15 demonstrates the views of the grid cube of docked CASP3-compound 1 complex through the X-, Y-, and Z-axis from left to right at the top of the figure, and the related 2D images below each view.

The feature selection using SFS and SFFS algorithms took place before the prediction phase in order to eliminate redundant features and to find out the most informative features. For this purpose, X-, Y-, and Z-images were converted into feature vectors as explained in the feature selection subsection under the data modelling section. The



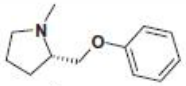
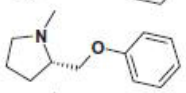
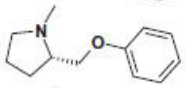
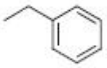
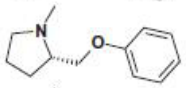
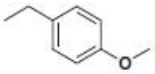
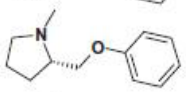
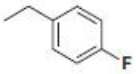
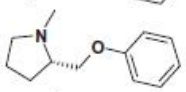
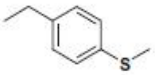
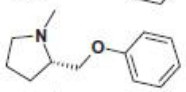
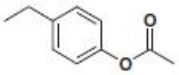
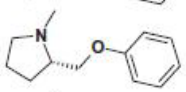
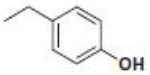
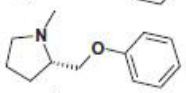
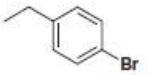
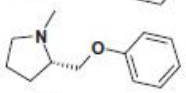
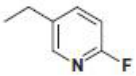
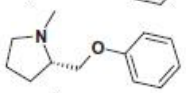
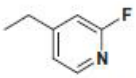
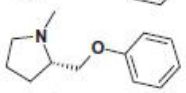
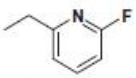
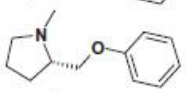
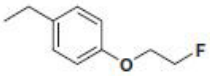
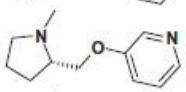
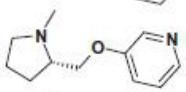
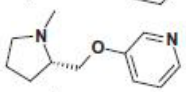
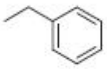
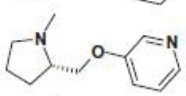
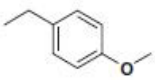
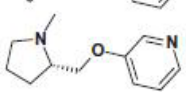
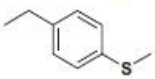
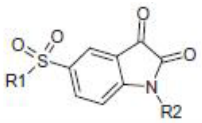
No.	R1	R2	pIC_{50}
1		$-CH_3$	6.92
2		–H	6..62
3			7.91
4			7.84
5			7.92
6			7.91
7			7.92
8			7.87
9			8.01
10			7.99
11			7.67
12			8.04
13			8.01
14		–H	7.23
15		$-CH_3$	7.63
16			8.28
17			8.41
18			8.36

Figure 5.13: Chemical structures and binding affinities of Caspase3 inhibitors



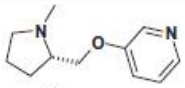
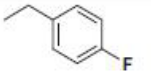
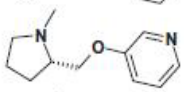
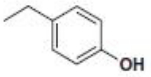
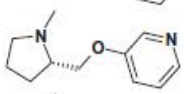
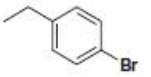
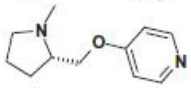
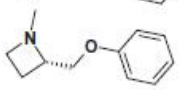
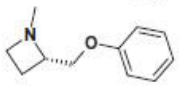
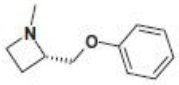
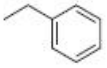
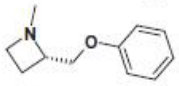
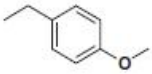
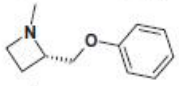
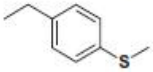
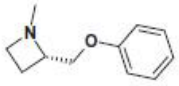
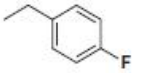
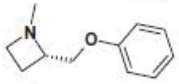
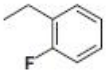
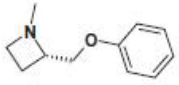
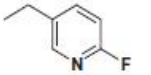
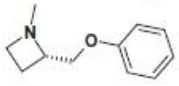
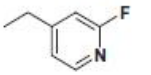
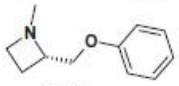
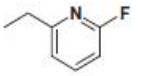
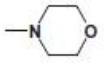
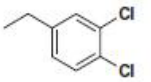
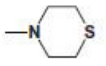
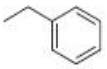
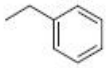
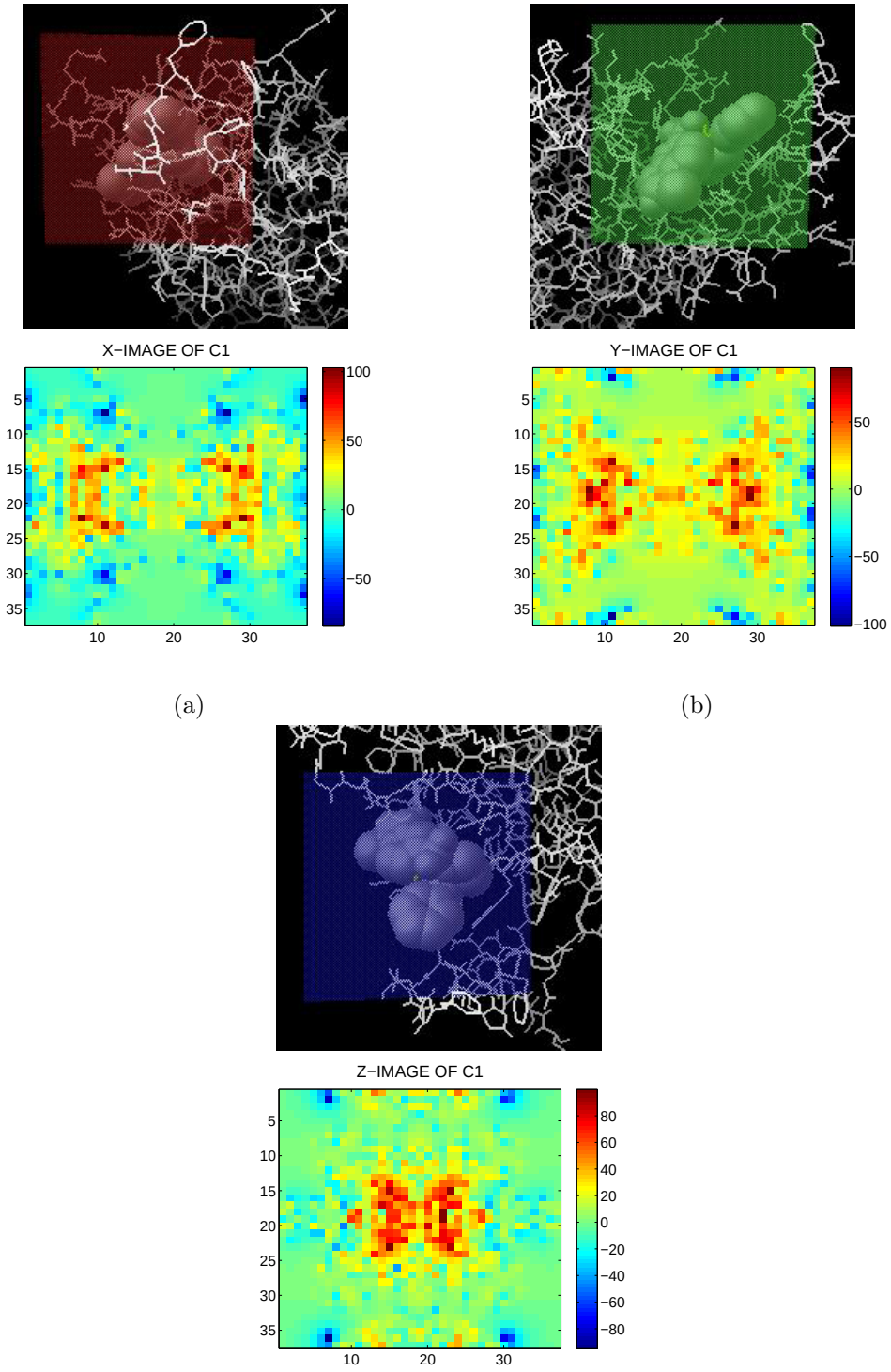
No.	R1	R2	pIC_{50}
19			8.08
20			8.41
21			8.44
22		–H	7.69
23		–H	6.54
24		$-CH_3$	7.04
25			8.01
26			8.08
27			7.95
28			8.06
29			8.03
30			7.96
31			7.53
32			8.24
33			5.84
34			5.99
35			6.94

Figure 5.14: Chemical structures and binding affinities of Caspase3 inhibitors (continued)



(a) (b)

(c)

Figure 5.15: An exemplary illustration of the 3D electrostatic potential (EP) grid for the Caspase3-compound 1 complex and the corresponding compressed 2D images. A view of the EP grid through (a) the X-axis (top) and the corresponding compressed X-image (bottom), (b) Y-axis (top) and the corresponding compressed Y-image (bottom), and (c) Z-axis (top) and the corresponding compressed X-image (bottom). The color scales for the compressed images are shown on the right side.

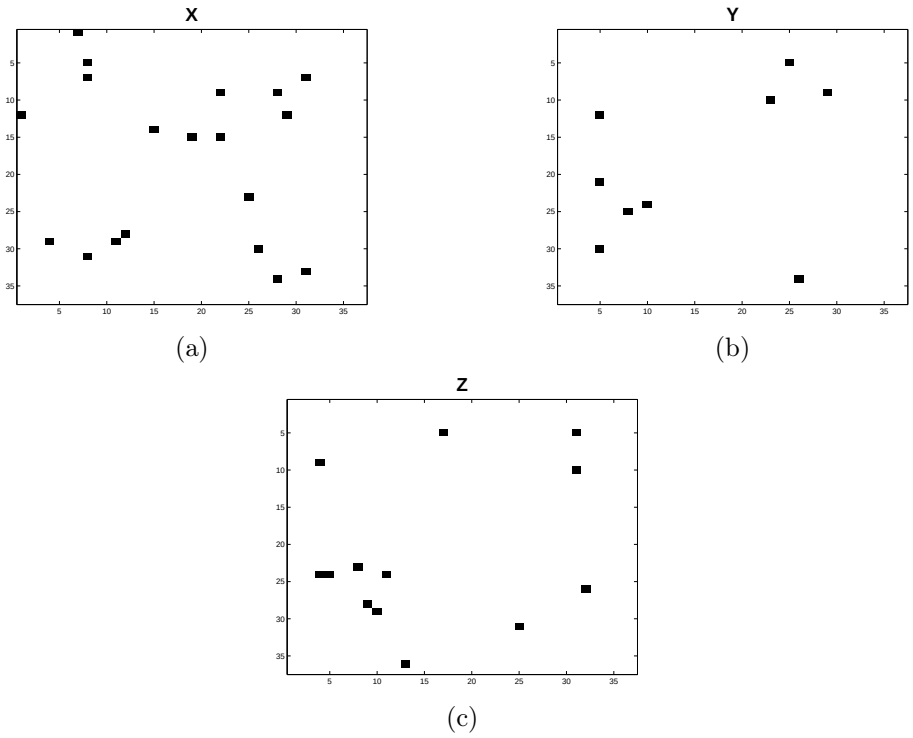


Figure 5.16: Two dimensional X-, Y- and Z-pattern images of CASP3-ligand complexes obtained by Sequential Forward Selection, SFS.

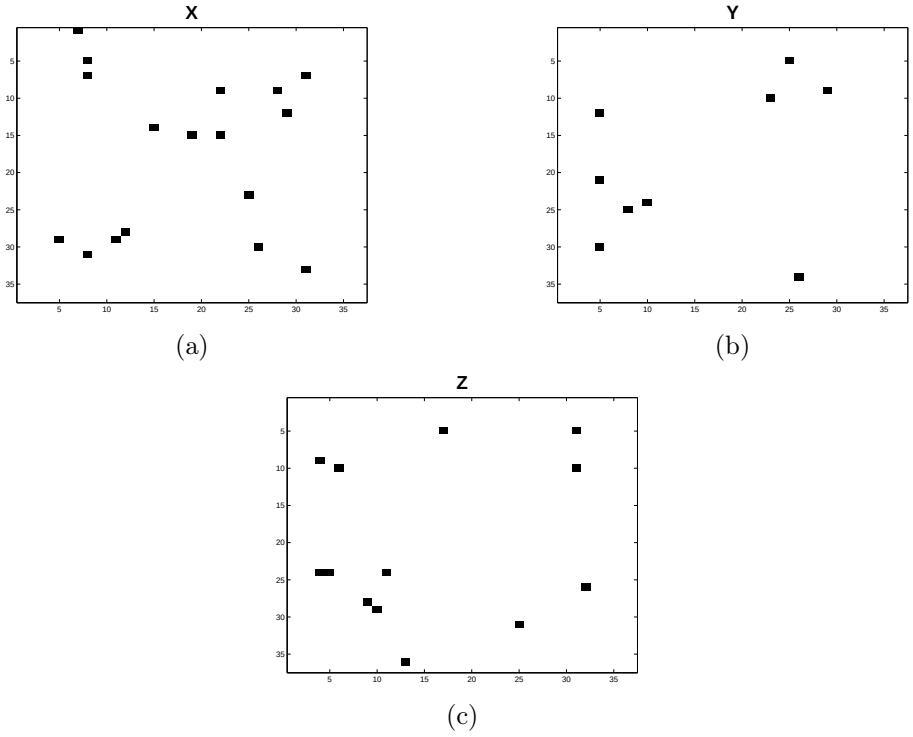


Figure 5.17: Two dimensional X-, Y- and Z-pattern images of CASP3-ligand complexes obtained by Sequential Floating Forward Selection, SFFS.

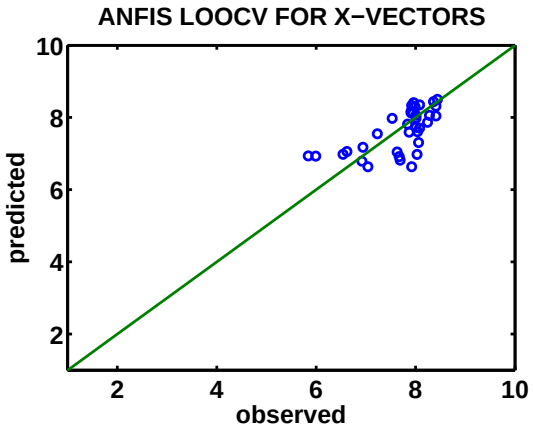
best RMSE values yielded by SFS computations were 0.1025 for leave-one-out cross-validation of the X-images, 0.1797 for leave-one-out cross-validation of Y-images, and 0.1442 for leave-one-out cross-validation of Z-images. The patterns obtained by SFS were then transformed into 2D pattern images for visualization, as in Figure 5.16. The black pixels represent the informative features which were found to be informative for predicting pIC_{50} values. SFS algorithm found out that 19 features of X-images, 9 features of Y-images, and 13 features of Z-images were valuable to be used in regression analysis. SFFS implementation slightly improved RMSE values which were found to be 0.0964 for leave-one-out cross-validation of the X-images, 0.1797 for leave-one-out cross-validation of Y-images, and 0.1421 for leave-one-out cross-validation of Z-images. Figure 5.17 demonstrates the features selected by SFFS algorithm which resulted in 17 features for X-images, 9 features for Y-images, and 13 features for Z-images.

When Figures 5.16 and 5.17 are compared, it is observed that two more redundant features were eliminated from X-images, the number and the location of the features belonging to Y-images did not change at all, and a redundant feature was replaced by an informative feature of Z-images. Unlike the results achieved from data modelling experiments on CHK1 and its inhibitors, X-images of isatine sulfonamide analogues in complex with CASP3 gave rise to the lowest RMSE values because the grid cube along the X-direction probably revealed a better appearance of the ligand into the binding site of CASP3 than the grid cube along the Y- and Z-directions. Once more, the meaningful features were located away from the center, except a few patterns in the X-pattern image (Figure 5.17a), determining the contact area of the ligand and the protein which were compound 1 and CASP3 in this case.

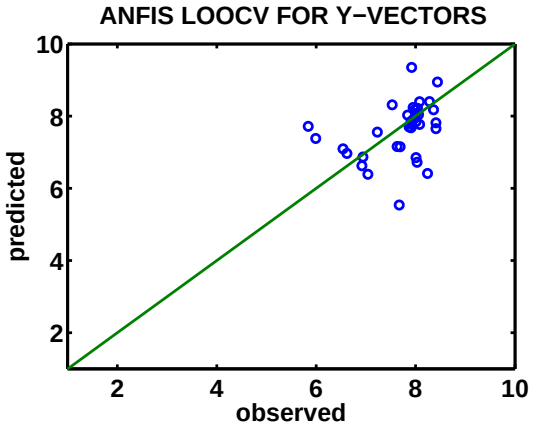
5.2.3 Prediction Phase

The binding affinities (pIC_{50}) of CASP3-ligand complexes were predicted by ANFIS, SVR, and PLSR using the X-feature, Y-feature, and Z-feature vectors obtained by SFFS algorithm. Three scatter plots in Figure 5.18 show the correlation between the experimental and predicted affinity when ANFIS was applied with leave-one-out cross-validation which utilized 35 distinct test sets for each of the X-feature, Y-feature, and Z-feature vectors of CASP3-ligand complexes. X-scatter plot in Figure 5.18a and Y-scatter plot in 5.18b achieve better correlation than that of Z-scatter plot in 5.18c which has data distributed away from the $y = x$ line, the so-called identity line. Moreover, the RMSE values obtained by ANFIS determination of leave-one-out cross-validation, which are 0.5299 for X-feature vectors, 0.7992 for Y-feature vectors, and 2.7216 for Z-feature vectors, denote that ANFIS failed to predict the binding affinities of CASP3-ligand complexes by using the feature vectors obtained from Z-images.

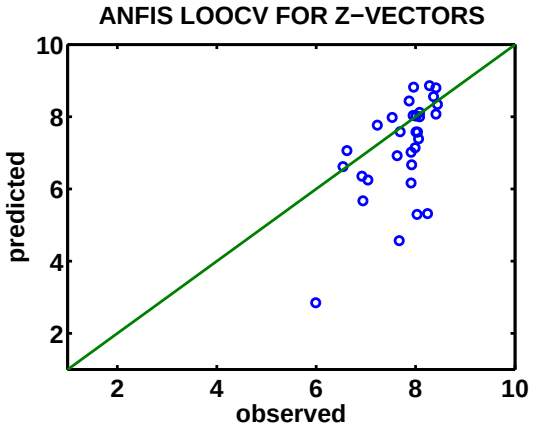
ANFIS was also implemented using the random subsampling set selection method which generates 1000 random training sets of 28 feature vectors, and test sets of 7 feature vectors by mixing up 35 X-feature, Y-feature, and Z-feature vectors of the



(a)



(b)



(c)

Figure 5.18: Affinity correlation plots constructed upon ANFIS determination of leave-one-out cross validation. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 35 different testing sets selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CASP3-ligand complexes.

Table5.6: R^2 and RMSE values for an average and the best 3 ANFIS determination of random subsampling set selection method out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes.

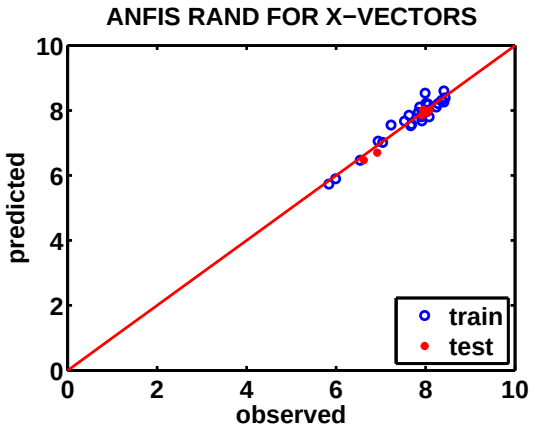
ANFIS	test		train	
X	R^2	RMSE	R^2	RMSE
Random-1	0.9552	0.1177	0.9255	0.1829
Random-2	0.9494	0.1229	0.9033	0.2044
Random-3	0.9462	0.1468	0.8958	0.2087
Average	0.4806	0.5360	0.9177	0.2010

ANFIS	test		train	
Y	R^2	RMSE	R^2	RMSE
Random-1	0.9389	0.1283	0.9287	0.1810
Random-2	0.8915	0.1880	0.9327	0.1728
Random-3	0.8750	0.2188	0.9330	0.1686
Average	0.3344	0.9569	0.9508	0.1655

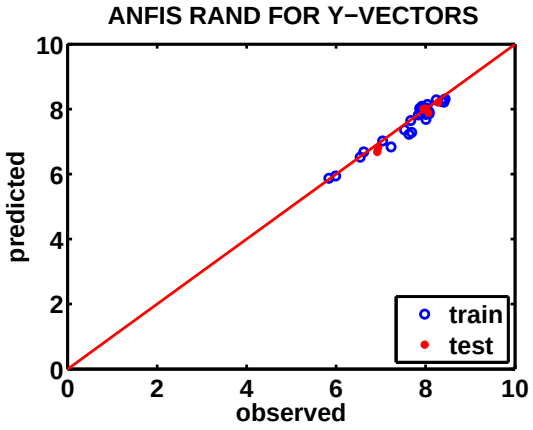
ANFIS	test		train	
Z	R^2	RMSE	R^2	RMSE
Random-1	0.7711	0.2877	0.9836	0.0824
Random-2	0.7538	0.3734	0.9870	0.0701
Random-3	0.6680	0.4235	0.9809	0.0863
Average	0.1969	2.2922	0.9842	0.0792

CASP3-ligand complexes without repetition. Table 5.6 shows the best three and average RMSE and R^2 values of ANFIS determination of random subsampling set selection on X-feature, Y-feature, and Z-feature vectors. As indicated in Table 5.6, X-feature vectors provided the lowest errors and the highest correlations between observed and predicted binding affinities compared to Y-feature and Z-feature vectors when the average of the results are considered. However, the results are found to be similar in the best training and test sets (Random-1 in Table 5.6) yielding the best RMSE and R^2 values when the scatter plots in Figures 5.19a-5.19c of the best training and test sets of X-feature, Y-feature, and Z-feature vectors are investigated. It is also observed that Z-feature vectors produce the best training errors although the test errors of Z-feature vectors are higher than that of X-feature and Y-feature vectors, which indicates an overfitting for ANFIS determination of random subsampling set selection using Z-feature vectors. Furthermore, even the highest R^2 of 0.4806 obtained by averaging the results of X-feature vectors does not satisfy the condition $R^2 > 0.6$ for ANFIS determination of random subsampling set selection which indicates that ANFIS is not a proper regression algorithm for CIFAP.

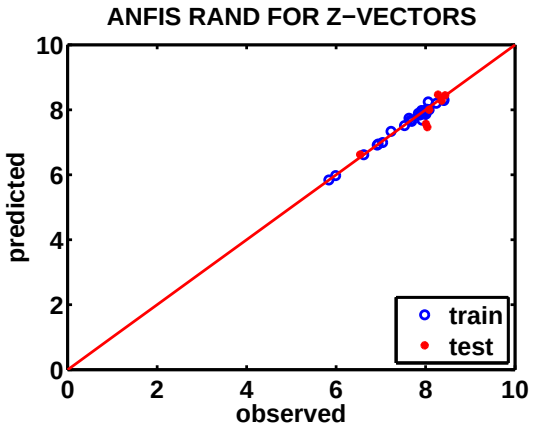
In order to predict binding affinities of the CASP3-ligand complexes, SVR with RBF-kernel was applied to X-feature, Y-feature, and Z-feature vectors produced in data modelling phase. Before the prediction of binding affinities, SVR parameters C, the trade-off value between error tolerance and model complexity, ε, the radius of the



(a)



(b)



(c)

Figure 5.19: Affinity correlation plots constructed by ANFIS determination of random subsampling set selection. The plots show ANFIS correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (random-1 in Table 5.6) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 35 CASP3-ligand complexes, 7 of which were used for testing (red dots) and the rest were used for training (blue circles).

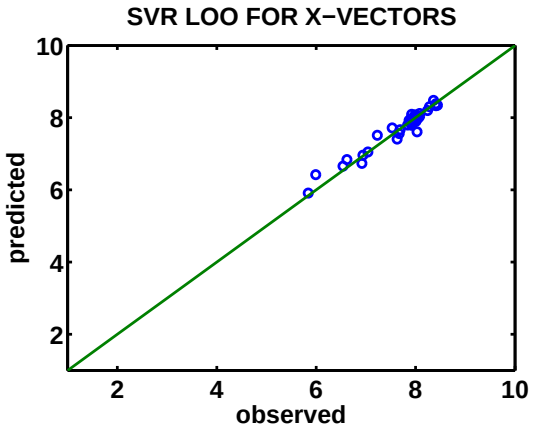
Table5.7: Optimal values of the SVR parameters C, ε, and γ for CASP3-ligand complexes.

	C	γ	ε
X-image	100	0.0080	0.0063
Y-image	100	0.0032	0.0611
Z-image	100	0.0143	0.0249

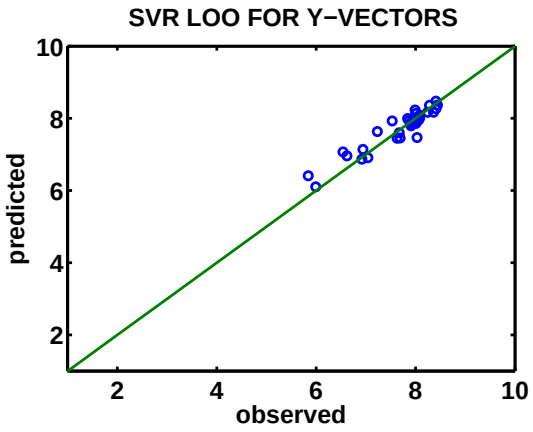
ε-tube, and γ, the width of the RBF-kernel, were optimized by a grid search using leave-one-out cross-validation. The optimal C, ε, and γ values for the X-feature ,Y-feature, and Z-feature vectors which provided the lowest RMSE and the highest R^2 values are listed in Table 5.7. Unlike the high C parameter of CHK1-ligand complexes in the previous subsection, the optimal C parameter for CASP3-ligand complexes is found to be 100 which is more preferable because a low value for C parameter causes the predictive model to be smooth and general.

SVR with RBF-kernel was implemented for predicting the binding affinities (pIC_{50}) of 35 CASP3 inhibitors, using leave-one-out cross-validation and the tuned C, ε, and γ values listed in Table 5.7. The correlations between the observed (x-axis) and predicted (y-axis) binding affinities were plotted in Figure 5.20a for the X-feature vectors, Figure 5.20b for the Y-feature vectors, and Figure 5.20c for the Z-feature vectors of CASP3-ligand complexes. It is observed that feature vectors in all directions resulted in good correlations, however, X-feature and Z-feature vectors lead to better correlations than that of Y-feature vectors with a small margin as seen in Figure 5.20. All three SVR models are capable of explaining at least 87% of the variability in the data. In addition, the R^2 of 0.9476 obtained for the X-feature vectors, 0.8791 obtained for the Y-feature vectors, and 0.9213 obtained for the Z-feature vectors verify that all three SVR models are predictive considering the first Tropsha criterion which is explained in the statistical analysis subsection of the prediction methods section.

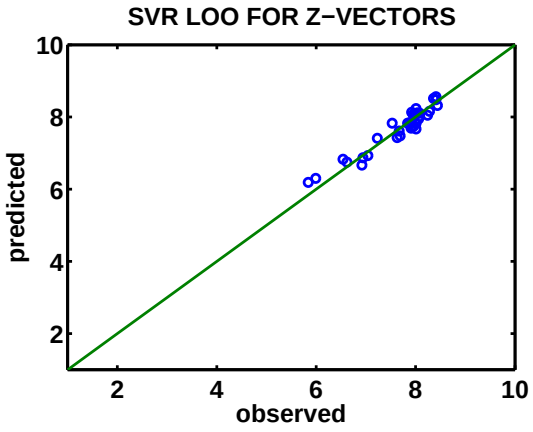
The tuned C, ε and γ values given in Table 5.7 were also utilized by the SVR determination of random subsampling set selection which shuffles the 35 X-feature, Y-feature, and Z-feature vectors 1000 times and reserves 7 feature vectors for the test set and 28 feature vectors for training set. The best three (Random-1, Random-2 and Random-3) and average results in terms of RMSE and R^2 values were presented in Table 5.8 for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes. According to the results in Table 5.8, the X-feature vectors yield the minimum error between the observed and predicted binding affinities with an average RMSE of 0.1966 determined out of 1000 test sets and an RMSE of 0.0478 for the best test set (Random-1). In addition to the RMSE values, the average R^2 values for the test sets of X-feature, Y-feature, and Z-feature vectors are higher than 0.8, satisfying the second criterion of Tropsha. Unlike the ANFIS determination of random subsampling set selection method, the results of SVR models indicate that Z-feature vectors also provide valuable information to be used in the prediction of binding affinities. The scatter plots



(a)



(b)



(c)

Figure 5.20: Affinity correlation plots constructed upon SVR determination of leave-one-out cross validation. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the optimal C, ε and γ values given in Table 5.7 and 35 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CASP3-ligand complexes.

Table5.8: R^2 and RMSE values for an average and the best three SVR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes.

SVR	test		train	
X	R^2	RMSE	R^2	RMSE
Random-1	0.9967	0.0478	0.9914	0.0548
Random-2	0.9967	0.0508	0.9915	0.0530
Random-3	0.9964	0.0485	0.9913	0.0562
Average	0.8891	0.1966	0.9940	0.0478

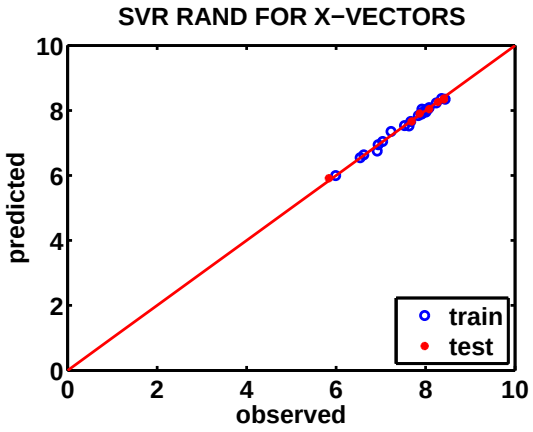
SVR	test		train	
Y	R^2	RMSE	R^2	RMSE
Random-1	0.9879	0.0807	0.9202	0.1770
Random-2	0.9860	0.0829	0.9209	0.1789
Random-3	0.9876	0.0925	0.9048	0.1838
Average	0.8210	0.2578	0.9395	0.1557

SVR	test		train	
Z	R^2	RMSE	R^2	RMSE
Random-1	0.9856	0.0739	0.9756	0.10274
Random-2	0.9804	0.1211	0.9679	0.10432
Random-3	0.9733	0.1365	0.9688	0.10451
Average	0.8497	0.2418	0.9793	0.0900

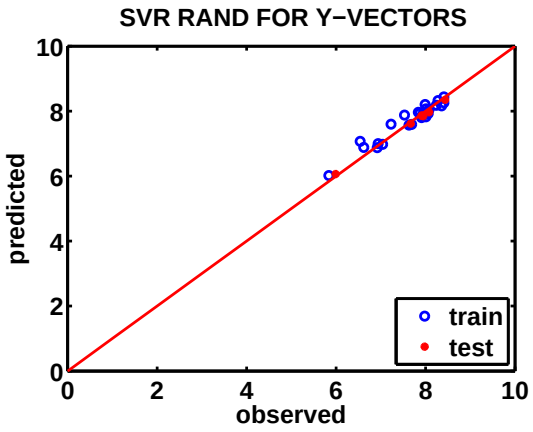
in Figures 5.21a-5.21c for the best test sets (Random-1) of X-feature, Y-feature, and Z-feature vectors give the correlation of the observed binding affinities versus the predicted binding affinities. The affinity correlation profile belonging to the X-feature vectors of 35 CASP3-ligand complexes in Figure 5.21c also verifies the results listed in Table 5.8.

The final method of regression analysis is Partial Least Squares Regression (PLSR), which was applied to predict the binding affinities of the CASP3 inhibitors using the X-feature, Y-feature, and Z-feature vectors produced by SFFS method. Figure 5.22 shows the correlation between the x-axis (observed) and y-axis (predicted) binding affinities acquired by the PLSR determination of the leave-one-out cross-validation, using 35 different sets of test vectors chosen from the X-feature, Y-feature, and Z-feature vectors of the CASP3-ligand complexes. The X-feature vectors ending in an RMSE of 0.1097 and an R2 of 0.9717 are discovered more informative than Y-feature vectors having an RMSE of 0.2576 and an R2 of 0.8429, and Z-feature vectors RMSE of 0.1562 and an R2 of 0.9422.

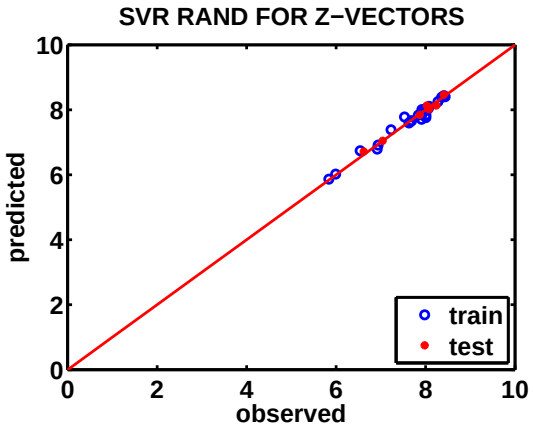
All three regression models mentioned earlier in this chapter agree that X-feature vectors of CASP3 inhibitors are more valuable considering the information they provide. Table 5.9 shows the observed binding affinities (pIC_{50}) of 35 CASP3 inhibitors published by Wang et al.[2] and the predicted binding affinities obtained by PLSR, SVR,



(a)

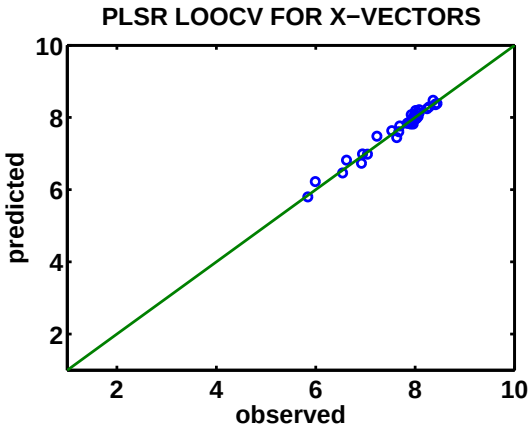


(b)

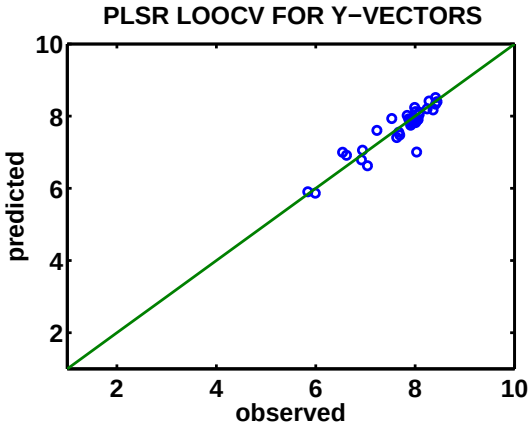


(c)

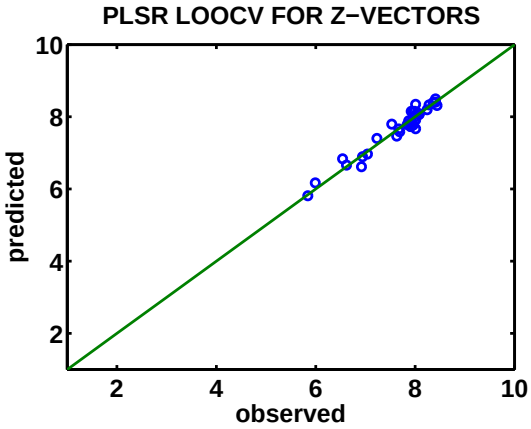
Figure 5.21: Affinity correlation plots constructed by SVR determination of random subsampling set selection. The plots show SVR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.8) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 35 CASP3-ligand complexes, 7 of which were used for testing (red dots) and the rest were used for training (blue circles).



(a)



(b)



(c)

Figure 5.22: Affinity correlation plots constructed upon PLSR determination of leave-one-out cross validation. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using 35 different testing sets (each including only one feature vector) selected from the X-feature (a), Y-feature (b), and Z-feature (c) vectors of the CASP3-ligand complexes.

Table5.9: RMSE comparison between observed binding affinities (pIC_{50}) for 35 CASP3 inhibitors, published by Wang et. al. [2], and the corresponding binding affinities (pIC_{50}) predicted by the PLSR, SVR, and ANFIS determination of leave-one-out cross validation for the X-feature vectors of the testing data sets.

No.	**pIC_{50}**	**PLS**	**SVR**	**ANFIS**	**No.**	**pIC_{50}**	**PLS**	**SVR**	**ANFIS**
1	6.92	6.7321	6.7343	6.8130	19	8.08	8.2158	8.1158	8.4373
2	6.62	6.8182	6.8415	6.7979	20	8.41	8.3566	8.3876	7.8571
3	7.91	7.8197	7.8416	7.9236	21	8.44	8.3868	8.3458	8.2555
4	7.84	7.8378	7.7962	8.1298	22	7.69	7.7639	7.6676	6.7577
5	7.92	8.0773	8.0937	7.8266	23	6.54	6.4630	6.6568	6.9685
6	7.91	7.8444	7.9712	8.0251	24	7.04	6.9840	7.0516	6.6491
7	7.92	7.8539	7.7916	6.5414	25	8.01	7.9913	8.0443	7.7015
8	7.87	7.8612	7.9271	7.8615	26	8.08	8.1287	8.0406	7.8752
9	8.01	7.9563	7.9009	8.0539	27	7.95	7.9886	7.9694	7.8291
10	7.99	8.0808	8.0790	8.1730	28	8.06	8.0264	8.0028	6.9894
11	7.67	7.6068	7.5624	7.4812	29	8.03	7.9798	7.6096	6.9992
12	8.04	8.1252	7.9809	7.7134	30	7.96	7.8192	7.8758	8.4012
13	8.01	8.1930	7.9631	7.7549	31	7.53	7.6330	7.7175	7.8267
14	7.23	7.4823	7.5135	7.5139	32	8.24	8.2430	8.1979	7.9824
15	7.63	7.4434	7.4080	7.1605	33	5.84	5.7997	5.9116	6.8290
16	8.28	8.2919	8.3048	8.1005	34	5.99	6.2235	6.4212	7.2851
17	8.41	8.3718	8.3344	8.0606	35	6.94	6.9903	6.9621	6.9803
18	8.36	8.4774	8.4788	8.5570	**RMSE**		**0.1097**	**0.1488**	**0.5299**

and ANFIS determination of leave-one-out cross validation, which used the X-feature vectors. It is clearly seen that the predictions of PLSR determination of leave-one-out cross validation were closer to the actual binding affinities by an RMSE of 0.1097 as compared to both the SVR and ANFIS determination of leave-one-out cross validation, which resulted in RMSE values of 0.1488 and 0.5299, respectively. PLSR determination of the leave-one-out cross-validation using Z-feature vectors also provided the highest correlation with $R^2 = 0.9717$ among the other two regression models, SVR and ANFIS.

Besides the leave-one-out cross-validation implementation, PLSR was implemented by using the random subsampling set selection method by randomly selecting 28 vectors of CASP3-ligand complexes for training and 7 vectors of CASP3-ligand complexes for testing without repetition. The R2 and RMSE values for the best three SVR results (Random-1, Random-2 and Random-3) and average RMSE and R2 values out of 1000 randomly selected training and test data sets related to the X-feature, Y-feature and Z-feature vectors are listed in Table 5.10. When compared to results of SVR and ANFIS

Table5.10: R^2 and RMSE values for an average and the best three PLSR determination of random subsampling set selection out of 1000 random training/test sets for the X-feature, Y-feature and Z-feature vectors of CASP3-ligand complexes.

PLSR	test		train	
X	R^2	RMSE	R^2	RMSE
Random-1	0.9966	0.0877	0.9938	0.0552
Random-2	0.9962	0.1092	0.9947	0.0502
Random-3	0.9954	0.0772	0.9927	0.0581
Average	0.9041	0.1623	0.9937	0.0503

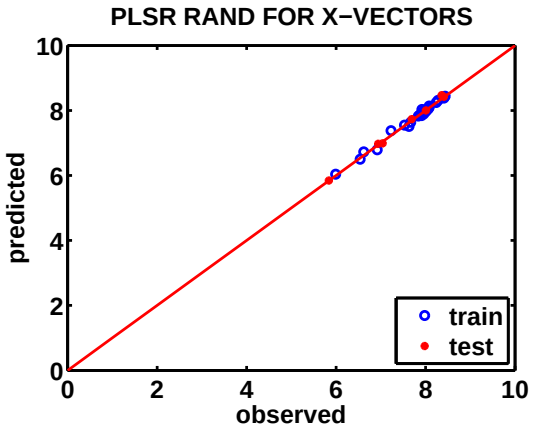
PLSR	test		train	
Y	R^2	RMSE	R^2	RMSE
Random-1	0.9924	0.1160	0.8991	0.1573
Random-2	0.9901	0.1273	0.9282	0.1590
Random-3	0.9887	0.1109	0.9026	0.1592
Average	0.8235	0.2558	0.9527	0.1378

PLSR	test		train	
Z	R^2	RMSE	R^2	RMSE
Random-1	0.9963	0.0910	0.9548	0.1206
Random-2	0.9962	0.0888	0.9667	0.1208
Random-3	0.9950	0.0862	0.9518	0.1188
Average	0.8913	0.1842	0.9741	0.1021

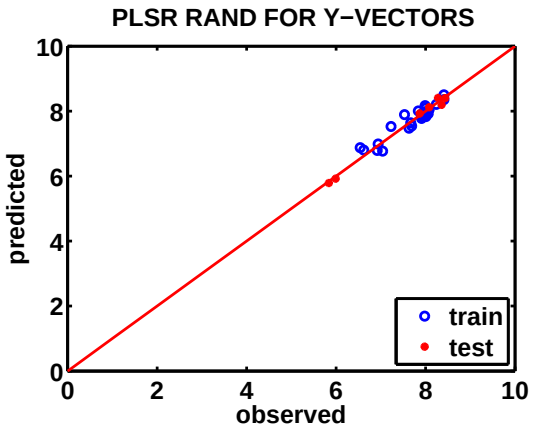
determination of random subsampling set selection, PLSR determination of random subsampling set selection reveals that the X-feature vectors produce more predictive model with an average RMSE of 0.1613 determined out of 1000 test sets and an RMSE of 0.0877 for the best test set (Random-1). Figures 5.23a-5.23c for the X-feature, Y-feature and Z-feature vectors show the correlation between the observed and predicted biding affinities of the Random-1 training and test sets in Table 5.10. X-feature vectors allows a better description of the variability in the model regarding a lower average RMSE of 0.1613 and a higher average R^2 of 0.9041 even though the test/training sets chosen from the Z-feature vectors also resulted in a low average RMSE of 0.1842 and a high average R^2 of 0.8913.

The analysis of all the regression and set selection methods implemented on the 35 CASP3-ligand complexes inform that the X-feature vectors produced from the X-images revealed more reliable and correlative information to address the learning and prediction abilities of CIFAP. As a result, the Z-images is a better choice for forecasting binding affinities of recently developed isatine sulfonamide derivatives in order to inhibit CASP3 protein. Although the coordinate system and the angular view may vary from one system to another system, the X-images are accepted as the source of information for CIFAP algorithm regarding the CASP3-ligand complexes.

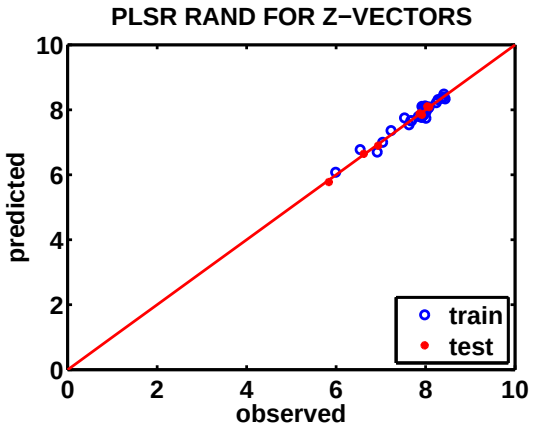
It seems that both PLSR and SVR methods that produce lower RMSE and higher



(a)



(b)



(c)

Figure 5.23: Affinity correlation plots constructed by PLSR determination of random subsampling set selection. The plots show PLSR correlations between the observed (x-axis) and predicted (y-axis) binding affinities (pIC_{50}), determined using the best randomly selected training/test sets (Random-1 in Table 5.8) for the X-feature (a), Y-feature (b), and Z-feature (c) vectors of 35 CASP3-ligand complexes, 7 of which were used for testing (red dots) and the rest were used for training (blue circles).

R^2 for the X-feature vectors are preferable to ANFIS method which did not produce reliable predictions for any of the X-feature, Y-feature, and Z-feature vectors. All the models generated by PLSR and SVR methods are found to be predictive since they satisfy the two criteria of Tropsha [52, 53] which are $R^2_{LOO} > 0.5$ for the leave-one-out cross-validation and an average $R^2 > 0.6$ for random subsampling set selection. Besides the better results of PLSR, the fast implementation of PLSR approves the fact that the utilization of PLSR for the X-feature vectors is superior to SVR and ANFIS methods, to be chosen for predicting the binding affinities of CASP3 inhibitors by CIFAP method and for developing novel ligands based on isatine sulfonamide pharmacophore.

CHAPTER 6

DISCUSSION AND CONCLUSION

A novel data representation method, CIFAP, is proposed in this study to predict the binding affinities of protein-ligand complexes whose ligands share a common pharmacophore with the published empirical IC_{50} or pIC_{50} values. The application of the CIFAP algorithm follows two phases: the data modelling phase and the prediction phase. The first phase, the data modelling phase, of CIFAP involves the compression of 3D electrostatic cubic grid maps of the binding site of the selected protein, in complex with docked coordinates of its inhibitors into 2D images in orthogonal X, Y and Z directions. 2D images of the binding site of the protein-ligand complexes step through the feature selection process implemented by Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS) algorithms in order to eliminate the redundancies in the images and to obtain informative feature vectors. The second phase, the prediction phase, forecasts the binding affinities, pIC_{50}, of the protein-inhibitor complexes by application of promising statistical machine learning methods, Partial Least Squares Regression (PLSR) [38], Support Vector Regression (SVR) [44] and Adaptive Neuro-Fuzzy Inference System (ANFIS) [50], on filtered 2D-images in training and test data sets. Training and test data sets were selected by the leave-one-out cross validation[43] and repeated random sub sampling[43] methods.

CIFAP was applied to two distinct protein-ligand complex systems for evaluation. The first system included the Checkpoint kinase 1 (CHK1), which is a primary target in the cancer therapy, and its 57 inhibitors based on thienopyridine derivatives, whose chemical structures and binding affinities (IC_{50}) were published by Zhao et. al.[1]. In a general sense, it was observed that 2D data points compressed through the surface boundaries of docked ligands seem to possess more informative values in correlation with empirical pIC_{50} values. It is observed that SFFS method is superior over SFS method for feature selection of CIFAP because of the improved RMSE and R^2 values obtained from X-pattern, Y-pattern, and Z-pattern images and the capability of SFFS to remove previously selected features which are found to be redundant or insignificant later on. Furthermore, Z-feature vectors obtained from the Z-images gave rise to more plausible data points to correlate with the empirical pIC_{50} values of the CHK1 inhibitors. The accumulation of the most useful information in the Z-direction is

thought to reflect the ligand surface boundaries that interact with the binding site of CHK1, see Figure 5.4.

Application of the CIFAP method on 57 CHK1-ligand complexes revealed that the PLSR determination of leave-one-out cross-validation provided the best results, an RMSE of 0.4109 and an R^2 of 0.8133, while its competitors SVR yielded an RMSE of 0.5940 and an R^2 of 0.6098, and ANFIS yielded an RMSE of 1.0868 and an R^2 of 0.1705 by applying leave-one-out cross-validation for the Z-feature vectors of CHK1-ligand complexes. Moreover, the PLSR implementation of random subsampling set selection yielded the best prediction profile with the lowest average RMSE values, 0.5388 for testing and 0.2776 for training, and the greatest R^2 values, 0.7141 for testing and 0.9129 for training, for the Z-feature vectors. These findings strongly suggest that the PLSR method should be preferentially chosen to enrich the library of the thenopyridine derivatives and thus develop more potent novel CHK1 inhibitors.

The second system analysed by CIFAP method contained Caspase 3 (CASP3), which plays a role in apoptosis of neurons, and its 35 inhibitors based on isatin sulfonamide analogue, whose chemical structures were published by Wang et. al. [2] and related pIC_{50} values were published by Hasegawa et. al. [69]. In the case of CASP3-ligand complexes, X-images obtained by compressing 3D electrostatic grid map through X-direction achieved better correlations with binding affinities as compared to Y-images and Z-images. Moreover, the X-feature vectors produced from X-images using SFFS method express that the margins of the interaction site provide significant information related to binding affinities as can be seen in Figure 5.17.

Once more, the PLSR determination of leave-one-out cross-validation was discovered as the most successful prediction method because of the lowest RMSE of 0.1097 and the highest R^2 of 0.9717 for X-feature vectors of CASP3-ligand complexes. On the other hand, ANFIS determination of leave-one-out cross-validation yielded the worst results, an RMSE of 0.3347 and an R^2 of 0.5299, while the SVR determination of leave-one-out cross-validation could be acknowledged as a promising method besides PLSR for X-feature vectors of CASP3-ligand complexes with an RMSE of 0.1488 and an R^2 of 0.94756. In addition to leave-one-out cross validation, the PLSR implementation of random subsampling set selection achieved the best correlations with the lowest average RMSE values, 0.1623 for testing and 0.0503 for training, and the highest R^2 values, 0.9041 for testing and 0.9937 for training, for the X-feature vectors.

The choice of the prediction algorithms of CIFAP was made for representing three different regression paradigms which are linear (PLSR), non-linear (SVR with RBF kernel), and neuro-fuzzy (ANFIS). The results obtained so far indicate that the relationship between the proposed data model and the binding affinities is more likely linear since the best predictions were generated by PLSR method. ANFIS, which is a rule based system using fuzzy relations and neural networks, is found to be complicated for explaining the relation between the compressed binding site images and the binding

affinities. Meanwhile, SVR is also a successful predictor although it was outperformed by PLSR. It should be noted here that the CIFAP method has not been yet tested on a variety of receptor-ligand systems other than the 57 CHK1-ligand complexes and the 35 CASP3-ligand complexes. Therefore, it is currently unwise to generalize the effect of the CIFAP method on all receptor-ligand systems. As far as other receptor-ligand systems are concerned, it would perhaps be more useful to apply other linear and non-linear regression methods in the literature beside the methods described in this dissertation to determine which method would yield optimal RMSE and R^2 values.

Another point to raise here is that a pharmacophore-based docking algorithm was applied to dock CHK1 and CASP3 inhibitors into the binding site of the mentioned proteins, assuming that the thienopyridine/isatine sulfonamide phamacophore of all ligands possess binding coordinates very similar to that of the X-ray coordinates of a thienopyridine/isatine sulfonamide derivative, which is compound 70 (PDB ID: 3PA3)[1] in complex with CHK1 and compound 1 (PDB ID: 1GFW)[2] in complex with CASP3. As a matter of fact, the absence of X-ray crystallographic coordinates for the other bound CHK1/CASP3 inhibitors disables us to make a comparison between docked and empirical coordinates. It is therefore not known yet to what extent the RMSE values will differ when different bound conformations of the CHK1/CASP3 inhibitors are likely. Nevertheless, the docked coordinates of the bound CHK1/CASP3 inhibitors studied here gave rise to the best predicted values in correlation with empirical pIC_{50} values. As far as the development of novel CHK1 inhibitors is concerned, one would make certain that a new substituent added to the thienopyridine/isatine sulfonamide pharmacophore should not substantially change the bound conformation of the pharmacophore in order for the CIFAP algoritm not to fail or to yield a false positive or false negative result. It is therefore strongly suggested that a library of compounds with a certain pharmacophore of interest be enriched as much as possible and that the docked coordinates of the novel compounds be compared to experimental bound coordinates when possible in order for CIFAP to be used more reliably in novel drug development.

To conclude, the affinity between a protein and the corresponding ligand is an important attribute for determining the success of interaction. It is difficult to calculate the binding affinity because the computations involve the calculation binding free energy which is a very small quantity. Moreover, docking programs perform calculations assuming the medium is air, which is in fact, water. Analyzing 3D structure of protein-ligand complexes in combination with electrical properties of the complex may be helpful for the predicting of the activity of novel compounds. In this thesis, a novel data model was presented which uses both 3D structure and electrical properties of protein-ligand interactions. The results of binding affinity prediction which uses the proposed data model were found to be more useful than the docking programs which provide only the best conformation of the ligands with minimum energy rather than the actual binding affinity. As a future work, the data model can be improved

by using other important electrical properties such as hydrophobicity. Moreover, it would be beneficial to obtain the images from a different angle of view by rotating the protein-ligand complexes.

REFERENCES

[1] L. Zhao, Y. Zhang, C. Dai, T. Guzi, D. Wiswell, W. Seghezzi, D. Parry, T. Fischmann, and M.A. Siddiqui. Design, synthesis and sar of thienopyridines as potent chk1 inhibitors. *Bioorg. Med. Chem. Lett.*, 20:7216–7221, 2010.

[2] Q. Wang, R.H. Mach, and D.E. Reichert. Docking and 3d-qsar studies on isatin sulfonamide analogues as caspase-3 inhibitors. *J. Chem. Inf. Model.*, 49(8):1963–1973, 2009.

[3] J.A. DiMasi, R.W. Hansen, H.G. Grabowski, et al. The price of innovation: new estimates of drug development costs. *J. Health. Econ.*, 22(2):151–186, 2003.

[4] P.L. Herrling. The drug discovery process. In *Imaging in Drug Discovery and Early Clinical Trials*, pages 1–14. Springer, 2005.

[5] M. Andrabi, C. Nagao, K. Mizuguchi, and S. Ahmad. *Bioinformatics Approaches for Analysis of Protein-Ligand Interactions*, chapter 9, pages 267–291. In Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2009.

[6] B. Alberts, A. Johnson, and J. Lewis. *Molecular Biology of the Cell*, chapter 3. Garland Science, New York, USA, 2002. Available from: http://www.ncbi.nlm.nih.gov/books/NBK26911/.

[7] R. Altman and J. Dugan. Defining bioinformatics and structural bioinformatics. *Structural Bioinformatics*, pages 3–14, 2003.

[8] David Whitford. *Proteins: structure and function.* Wiley, 2005.

[9] Ilme Schlichting. X-ray crystallography of protein-ligand interactions. In *Protein-Ligand Interactions*, pages 155–165. Springer, 2005.

[10] J. Krumrine, F. Raubacher, N. Brooijmans, and I. Kuntz. Principles and methods of docking and ligand design. *Structural Bioinformatics*, 44:441–476, 2005.

[11] D.R. Hall, D. Kozakov, and S. Vajda. Analysis of protein binding sites by computational solvent mapping. In *Computational Drug Discovery and Design*, pages 13–27. Springer, 2012.

[12] S.A. Hassan, L. Gracia, G. Vasudevan, and P.J. Steinbach. Computer simulation of protein-ligand interactions: Challenges and applications. In *Protein-Ligand Interactions: Methods and Applications*, volume 305 of *Methods in Molecular Biology*, pages 451–492. Springer, March 2005.

[13] Y. Cheng and W.H. Prusoff. Relationship between the inhibition constant (k1) and the concentration of inhibitor which causes 50 per cent inhibition (i50) of an enzymatic reaction. *Biochem. Pharmacol.*, 22:3099–3108, 1973.

[14] Keith James. The evolution of quantitative drug design. In David J. Livingstone and Andrew M. Davis, editors, *Drug Design Strategies - Quantitative Approaches*. Royal Society of Chemistry, Oxford, 2012.

[15] L. Jacob and J.P. Vert. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, 24(19):2149–2156, 2008.

[16] Ying Liu. Drug design by machine learning: ensemble learning for qsar modeling. In *Proceedings of the Fourth International Conference on Machine Learning and Applications*, pages 187–193, Los Angeles, CA, USA, 2005. IEEE Computer Society.

[17] R. Burbidge, M. Trotter, B. Buxton, and S. Holden. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Computers&Chemistry*, 26(1):5–14, 2001.

[18] W. Deng, C. Breneman, and M.J. Embrechts. Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J. Chem. Inf. Model.*, 44(2):699–703, 2004.

[19] A. Amini, P.J. Shrimpton, and S.H. Muggleton. A general approach for developing system-specific functions to score protein-ligand docked complexes using support vector inductive logic programming. *Proteins*, 69:823–831, 2007.

[20] S. Li, L. Xi, C. Wang, J. Li, B. Lei, H. Liu, and X. Yao. A novel method for protein-ligand binding affinity prediction and the related descriptors exploration. *J. Comput. Chem.*, 30:900–909, 2009.

[21] P.J. Ballester and J.B.O. Mitchell. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26:1169–1175, 2010.

[22] L. Saghaie, M. Shahlaei, and A. Madadkar-Sobhani. Application of partial least squares and radial basis function neural networks in multivariate imaging analysis-quantitative structure activity relationship: Study of cyclin dependent kinase 4 inhibitors. *J. Mol. Graphics Modell.*, Oct 2010.

[23] L. Breiman. Bagging predictors. *Mach. Learn.*, 24:123–140, 1996.

[24] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proceeding of 13th International Conference on Machine Learning*, pages 148–156, 1996.

[25] L. Breiman. Random forests. *Mach. Learn.*, 45:5–32, 2001.

[26] *HyperChem 5.1.* Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA, http://www.hyper.com [last accessed September 2011].

[27] *Accelrys Software Inc., Discovery Studio Modeling Environment, Release 1.7, San Diego: Accelrys Software Inc.* http://www.accelrys.com/ [June 2011].

[28] F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.E. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112(3):535–542, 1977.

[29] Michel F. Sanner. Python: A programming language for software integration and development. *J. Mol. Graphics Modell.*, 17:57–61, 1999.

[30] O. Trott and A. J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comp. Chem.*, 31:455–461, 2010.

[31] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. Autodock4 and autodocktools4: automated docking with selective receptor flexiblity. *J. Comput. Chem.*, 16:2785–2791, 2009.

[32] MATLAB. *version 7.12 (R2011a)*. The MathWorks Inc., Natick, Massachusetts, 2011.

[33] S.B. Kotsiantis, I.D. Zaharakis, and P.E. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[34] A. Wayne Whitney. A direct method of nonparametric measurement selection. *Computers, IEEE Transactions on*, 100(9):1100–1103, 1971.

[35] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125, 1994.

[36] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery Data Mining*. Kluwer Academic Publishers, Boston, 1998.

[37] S. Chatterjee and A.S. Hadi. *Regression Analysis by Example*. John Wiley & Sons., 4 edition, 2006.

[38] T. Marill and D. Green. On the effectiveness of receptors in recognition systems. *Information Theory, IEEE Transactions on*, 9(1):11–17, 1963.

[39] R. Caruana and D. Freitag. Greedy attribute selection. In *Proceedings of the eleventh international conference on machine learning*, pages 28–36. Citeseer, 1994.

[40] P. Somol, P. Pudil, J. Novovičová, and P. Paclık. Adaptive floating search methods in feature selection. *Pattern recognition letters*, 20(11):1157–1163, 1999.

[41] S. Nakariyakul and D.P. Casasent. Improved forward floating selection algorithm for feature subset selection. In *Wavelet Analysis and Pattern Recognition, ICWAPR'08. International Conference on*, volume 2, pages 793–798. IEEE, 2008.

[42] V. Kecman. *Support Vector Machines - An introduction*, volume 177 of *Studies in Fuzziness and Soft Computing*, chapter 1, pages 1–49. Springer-Verlag Berlin Heidelberg, 2005.

[43] R. Picard and D. Cook. Cross-validation of regression models. *J. Am. Stat. Assoc.*, 79(387):575–583, 1984.

[44] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, 6:155–161, 1996.

[45] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[46] B. Üstün, W.J. Melssen, M. Oudenhuijzen, and L.M.C. Buydens. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Anal. Chim. Acta*, 544(1–2):292–305, 2005.

[47] C.C. Chang and C.J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm [September 2012].

[48] V. Cherkassky and Y. Ma. Practical selection of svm parameters and noise estimation for svm regression. *Neural Networks*, 17(1):113–126, 2004.

[49] Z.-L. Sun, K.-F. Au, and T.-M. Choi. A neuro-fuzzy inference system through integration of fuzzy logic and extreme learning machines. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 37(5):1321–1331, 2007.

[50] J.S.R. Jang. Anfis: Adaptive-network-based fuzzy inference systems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 23(3):665–685, 1993.

[51] R.R. Yager and D.P. Filev. Generation of fuzzy rules by mountain clustering. *J. Intelligent and Fuzzy System*, 2:209–219, 1994.

[52] A. Golbraikh and A. Tropsha. Beware of q2! *J. Mol. Graphics Modell.*, 20(4):269, 2002.

[53] A. Tropsha, P. Gramatica, and V.K. Gombar. The importance of being earnest: validation is the absolute essential for successful application and interpretation of qspr models. *QSAR & Combinatorial Science*, 22(1):69–77, 2003.

[54] M.L. Agarwal, A. Agarwal, W.R. Taylor, and G.R. Stark. p53 controls both the g2/m and the g1 cell cycle checkpoints and mediates reversible growth arrest in human fibroblasts. *Proc. Natl. Acad. Sci. U.S.A.*, 92(18):8493–8497, 1995.

[55] K.H. Vousden and X. Lu. Live or let die: the cell's response to p53. *Nat. Rev. Cancer*, 2:594–604, 2002.

[56] S.L. Harris and A.J. Levine. The p53 pathway: positive and negative feedback loops. *Oncogene*, 24(17):2899–2908, 2005.

[57] E.R. Rayburn, S.J. Ezell, and R. Zhang. Recent advances in validating mdm2 as a cancer target. *Anti-Cancer Agents Med. Chem.*, 9:882–903, 2009.

[58] Z. Xiao, J. Xue, T.J. Sowin, and H. Zhang. Differential roles of checkpoint kinase 1, checkpoint kinase 2, and mitogen-activated protein kinase-activated protein kinase 2 in mediating dna damage-induced cell cycle arrest: implications for cancer therapy. *Mol. Cancer Ther.*, 8:1935–1943, 2006.

[59] Z.F. Tao and N.H. Lin. Chk1 inhibitors for novel cancer treatment. *Anti-Cancer Agents Med. Chem.*, 4:377–388, 2006.

[60] Michelle Prudhomme. Novel checkpoint 1 inhibitors. *Recent Pat. Anti-Cancer Drug Discovery*, 1:55–68, 2006.

[61] C. Selvaraj, S.K. Tripathi, K.K. Reddy, and S.K. Singh. Tool development for prediction of pic50 values from the ic50 values - a pic50 value calculator. *Curr. Trends Biotechnol. Pharm.*, 5(2):1104–1109, 2011.

[62] H.A. Harrington, K.L. Ho, S. Ghosh, and K.C. Tung. Construction and analysis of a modular model of caspase activation in apoptosis. *Theor. Biol. Med. Model.*, 5(26), 2008.

[63] B. Fang, P.I. Boross, J. Tozser, and I.T. Weber. Structural and kinetic analysis of caspase-3 reveals role for s5 binding site in substrate recognition. *J. Mol. Biol.*, 360(3):654–666, 2006.

[64] I.T. Weber, B. Fang, and J. Agniswamy. Caspases: structure-guided design of drugs to control cell death. *Mini Reviews in Medicinal Chemistry*, 8(11):1154–1162, 2008.

[65] D. Lee, S.A. Long, J.H. Murray, J.L. Adams, M.E. Nuttall, D.P. Nadeau, K. Kikly, J.D. Winkler, C.M. Sung, M.D. Ryan, M.A. Levy, P.M. Keller, and W.E. DeWolf. Potent and selective nonpeptide inhibitors of caspases 3 and 7. *J. Med. Chem.*, 44(12):2015–2026, 2001.

[66] W. Chu, J. Zhang, C. Zeng, J. Rothfuss, Z. Tu, Y. Chu, D.E. Reichert, M.J. Welch, and R.H. Mach. N-benzylisatin sulfonamide analogues as potent caspase-3 inhibitors: synthesis, in vitro activity, and molecular modeling studies. *J. Med. Chem.*, 48(24):7637–7647, 2005.

[67] W. Chu, J. Rothfuss, A. d'Avignon, C. Zeng, D. Zhou, R.S. Hotchkiss, and R.H. Mach. Isatin sulfonamide analogs containing a michael addition acceptor: a new class of caspase 3/7 inhibitors. *J. Med. Chem.*, 50(15):3751–3755, 2007.

[68] E. Gail, S. Gos, R. Kulzer, J. Lorösch, A. Rubo, M. Sauer, R. Kellens, J. Reddy, N. Steier, and W. Hasenpusch. Cyano compounds, inorganic. *Ullmann's Encyclopedia of Industrial Chemistry*, pages 1–37, 2000.

[69] K. Hasegawa and K. Funatsu. New description of protein-ligand interactions using a spherical self-organizing map. *Bioorg. Med. Chem.*, 20:5410–5415, 2012.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Erdaş, Özlem
Nationality: Turkish (TC)
Date and Place of Birth: April 15, 1981, Bursa, Turkey
Marital Status: Single
Phone: 0532 704 4335

EDUCATION

Degree	Institution	Year of Graduation
Ph.D.	Department of Computer Engineering, METU	2013
M.S.	Department of Computer Engineering, METU	2007
B.S.	Department of Mathematics, ITU	2003
High School	Vefa Anatolian High School, Istanbul	1999

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2010-Present	Republic of Turkey Prime Ministry	Engineer
2007-2010	Dept. of Computer Engineering, METU	Teaching Assistant
2005-2006	Republic of Turkey Ministry of Finance	System Analyst/Programmer

PUBLICATIONS

International Journal Publications

1. O. Erdas, C.A. Andac, A.S. Gurkan-Alp, F.N. Alpaslan, E. Buyukbingol, Compressed Images for Affinity Prediction (CIFAP): A Study on Predicting Binding Affinities for Checkpoint Kinase 1 Protein Inhibitors, Journal of Chemometrics, 27(6):155-164, 2013.

2. O. Erdas, E. Buyukbingol, F.N. Alpaslan, A. Adejare, Modeling and Predicting Binding Affinity of Phencyclidine-Like Compounds Using Machine Learning Methods, Journal of Chemometrics, 24(1):1-13, 2010.

International Conference Publications

1. O. Erdas, C.A. Andac, A.S. Gurkan-Alp, F.N. Alpaslan, E. Buyukbingol, Compressed Images for Affinity Prediction (CIFAP): A Novel Machine Learning Methodology on Protein-Ligand Interactions, International Conference on Omics Studies, Orlando, FL, September 4-6, 2013.

2. O. Erdas, C.A. Andac, A.S. Gurkan-Alp, F.N. Alpaslan, E. Buyukbingol, Predicting Binding Affinities of Drug-Protein Interactions By Analysis of Binding Site Images, 10^{th} International Symposium on Pharmaceutical Sciences (ISOPS-10), Ankara, Turkey, April 26-29, 2012.

AWARD AND SCHOLARSHIP

- Ph.D. Fellowship by TUBITAK (2008-2013)
- First honors degree, Department of Mathematics, Faculty of Science and Letters, ITU (2003)