Hüseyin GÜNER A Master’s Thesis AGU 2022

MOLECULAR RECOGNITION OF PROTEIN-LIGAND COMPLEXES VIA CONVOLUTIONAL NEURAL NETWORKS

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Hüseyin GÜNER
January 2022

MOLECULAR RECOGNITION OF PROTEIN-LIGAND COMPLEXES VIA CONVOLUTIONAL NEURAL NETWORKS

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE OF
ABDULLAH GUL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Hüseyin GÜNER
January 2022

SCIENTIFIC ETHICS COMPLIANCE

I hereby declare that all information in this document has been obtained 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 materials and results that are not original to this work.

Name-Surname: Hüseyin GÜNER
Signature :

REGULATORY COMPLIANCE

M.Sc thesis titled MOLECULAR RECOGNITION OF PROTEIN-LIGAND COMPLEXES VIA CONVOLUTIONAL NEURAL NETWORKS has been prepared in accordance with the Thesis Writing Guidelines of the Abdullah Gül University, Graduate School of Engineering & Science.

Prepared By
Hüseyin GÜNER
Signature

Advisor
Assoc. Prof. Dr. Zafer AYDIN
Signature

Head of the Electrical and Computer Engineering Program
Assoc. Prof. Dr. Kutay İÇÖZ

ACCEPTANCE AND APPROVAL

M.Sc. thesis titled MOLECULAR RECOGNITION OF PROTEIN-LIGAND COMPLEXES VIA CONVOLUTIONAL NEURAL NETWORKS and prepared by Hüseyin GÜNER has been accepted by the jury in the Electrical and Computer Engineering Graduate Program at Abdullah Gül University, Graduate School of Engineering & Science.

19 /12 /2021

(Thesis Defense Exam Date)

JURY:

Advisor : (Assoc. Prof. Dr. Zafer AYDIN)

Member : (Asst. Prof. Dr. İsmail AKÇOK)

Member : (Prof. Dr. Servet ÖZCAN)

APPROVAL:

The acceptance of this M.Sc thesis has been approved by the decision of the Abdullah Gül University, Graduate School of Engineering & Science, Executive Board dated ….. /….. / ….. and numbered …………..……. .

……….. /……….. / ………..

(Date)

Graduate School Dean
Prof. Dr. İrfan ALAN

ABSTRACT

MOLECULAR RECOGNITION OF PROTEIN-LIGAND COMPLEXES VIA CONVOLUTIONAL NEURAL NETWORKS

Hüseyin GÜNER

MSc. in Electrical and Computer Engineering
Advisor: Assoc. Prof. Zafer AYDIN

January 2022

As a sub-discipline of Artificial Intelligence, deep neural networks have received enormous interest in research and industrial applications over the last decades owing to their highly successful performance in addressing and solving broad areas of problems. Hence, especially hitherto achievements in computer-aided drug design brought an extra impetus with the novel deep learning approaches in structure-based drug design etiology. Our group offers a novel convolutional neural network model, deepMLR, that casts insight into the molecular recognition of ligand molecules and a receptor protein molecule. Having compared our model and a few other existing models with a case study of a traditional approach, herein, we present the success story of a deep learning model straight.

Keywords: Molecular Recognition, Structure Based Drug Design, Deep Convolutional Neural Networks, Protein-Ligand Affinity Prediction, Structure Based Virtual Screening

ÖZET

PROTEİN-LİGAND KOMPLEKSLERİNİN KONVOLÜSYENEL SİNİR AĞLARI İLE MOLEKÜLER TANINMASI

Hüseyin GÜNER

Elektrik ve Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans
Tez Yöneticisi: Doç. Dr. Zafer AYDIN

Ocak 2022

Yapay Zeka'nın bir alt disiplini olarak derin sinir ağları, geniş spektrumdaki problem alanlarını ele alma ve çözmedeki son derece başarılı performansları nedeniyle, son on yılda (özellikle) araştırma ve endüstriyel uygulamalarda büyük bir ilgi görmeye başladı. Özellikle son zamanlardaki, bilgisayar destekli ilaç tasarımındaki başarıları nedeniyle, yapı tabanlı ilaç tasarımı etiyolojislerindeki yeni derin öğrenme yaklaşımlarına karşı ekstra bir ivme kazanmıştır. Grubumuz, ligand moleküllerinin ve bir reseptör protein molekülünün moleküler olarak tanınması hakkında bir fikir veren yeni bir konvolüsyonel sinir ağı modeli sunmaktadır. Diğer mevcut modellerle ve modelimizle geleneksel bir yaklaşımın örnek çalışmasıyla karşılaştırıldığında, burada derin bir öğrenme modelinin başarı hikayesini sunuyoruz.

Anahtar kelimeler: Moleküler Tanıma, Yapı Tabanlı İlaç Tasarımı, Derin Konvolüsyonel Sinir Ağları, Protein-Ligand Bağlanma Tahmini, Yapı Tabanlı Sanal Tarama

Acknowledgements

I had greatly appreciated my thesis supervisor Dr Zafer AYDIN's inspirational mentorship and guidance during and even before my graduate studies commenced. He has guided me throughout my research in a passionate manner and helped me envision the subject matter in more detail. His keen observations have contributed much to my inadequacies related to my personal and professional life.

Dr Ismail Akcok and Dr Emel Basak Akcok allowed me to deal with a perfect example to test with our genuine approach. I am delighted to work with them, and they helped me connect with the domain knowledge of the computational problem. It is fun to exchange ideas and discuss with both, and it made me happy to spend time with them.

My beloved family sacrificed their valuable time throughout my journey in graduate studies, and I am utterly thankful for their support and encouragement. Without my sons and my wife's support, my thesis could have turned into a dull work. Every weekend and weekdays, I always felt their presence as my primary motivation and inspiration.

I am thankful to the Dean of the Faculty of Nature and Life Sciences, Dr. Alaattin Sen, for providing us the computational resources, faculty's high-performance cluster (HPC), to facilitate our scientific computations.

I would like to sincerely devote my work to my beloved father, Yilmaz Guner. I will always set my mind to achieve the greater possibility since he made me think I could try and succeed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Networks
CADD	Computer-Aided Drug Design
CNN	Convolutional Neural Networks
DL	Deep Learning
DMTA	Design-Make-Test-Analyze
DTI	Drug-Target Interactions
GNN	Graph Neural Networks
HTS	High Throughput Screening
MD	Molecular Dynamics
ML	Machine Learning
MM/PBSA	Molecular Mechanics/Poisson−Boltzmann Surface Area
MM/GBSA	Molecular Mechanics/Generalized Born Surface Area
OPLS	Optimized Potential for Liquid Simulation
OPLS-AA	Optimized Potential for Liquid Simulations with An All-Atom
PDB	Protein Data Bank
RNN	Recurrent Neural Networks
SAR	Structure-Activity Relationship
SBDD	Structure-Based Drug Design/Discovery
SBVS	Structure-Based Virtual Screening
SF	Scoring Functions
SGD	Stochastic Gradient Descent
SHAP	SHapley Additive exPlanations
VS	Virtual Screening

Sevgili Babam
Yılmaz GÜNER'in
aziz hatırasına…

Chapter 1

Introduction

Even in prehistoric times, drugs -as a chemical substance- were primarily extracted from different plants and were used as therapeutics. Based on the archaeological remains, the first known medicines that were extracted from plants date back to 60,000 BC [1]. Those natural products were obtained and processed by traditional experts like shamans in certain regions of the world. The process of finding a potent small chemical can be simply described as drug discovery[2]. Instead of a shaman, scientists from different disciplines are trying to articulate newer methodologies to find the most suitable candidate for a certain target macromolecule, most of the time a protein of interest which is believed to be the biomarker of a certain disease.

Advancements in organic chemistry, genomics, proteomics, and crystallographic experiments started a new era of successful findings into the drug discovery process [2]. The biological importance of the target molecule’s molecular functions and its structure is becoming the main motive of investigating newer or alternative small chemicals. High throughput screening (HTS) [3] experiments of hundreds of thousands of chemicals can be tested using biochemical assays set up to determine the biological activity of compounds of interest. It is estimated that the overall discovery process costs more than a billion dollars [4]. Nevertheless, those expensive and lengthy techniques can be complemented or replaced with a computational method of different types to reduce the time and the expenses. Computer-aided drug design (CADD) provides a cheaper and faster alternative in finding the leads and testing their performance before trying them clinically [5].

Theoretical chemists and many other counterpart scientists from different research areas have designed highly successful molecular mechanics methodologies to describe the nature of interactions of all constituents of drug-target interactions (DTI). Their success in building the foundations of understanding the molecular interactions and the dynamic systems fueled the computational scientists working in Artificial Intelligence (AI) to thrive into building models to tackle similar problems. From its early days of the fifties to our time, AI has been used in everyday applications and its success, especially in image processing, speech recognition and natural language processing, inspired everyone in the industry and research. After a long winter season of AI studies, newly invented algorithms, and rapid advancements in computer infrastructure and hardware strived the scientists and professionals of the industry to put more effort in solving all sorts of problems using AI and Computational Science [6], [7].

In this thesis work, we aimed to introduce a few existing AI methods and compare them to a real case study to showcase how effective and successful their methodology is. We have additionally proposed a similar and extensive model using the same AI method they have used. Our model is based on an analogy with the structure-based drug design/discovery (SBDD) approach of different combinations of available in vitro, in vivo biochemical techniques and heavy computational analysis methods of molecular interactions.

In chapter 2, we will scrutinize the conceptual framework of SBDD, starting with molecular recognition. We have hesitated to recount the theory and applications of SBDD and merely focused on the field's core concepts to build a solid theoretical basis for our computational tool. It will be out of the scope of our work to deliver the theoretical chemist's agenda to our readers. We have included an overview of structure-based virtual screening (SBVS) as a subsection in that chapter. The next chapter outlines the structure of the AI method in detail and explains the ingredients of model construction and outputs of the procedure. Chapter 4 is dedicated to presenting a showcase of selected models and our model and included the traditional CADD elaboration for a case study and compared all of them on that specific example work. In the final chapter, we discussed the prospects and use and benefit of our model in detail.

Chapter 2

Structure Based Drug Discovery

Structural biology was introduced as a direct method in drug discovery as early as the seventies, and the first successful applications appeared at the beginning of the nineties [2]. Even in industry and the research environments, structural biology and SBDD became essential operational tools in drug discovery. As depicted in Fig. 1, SBDD can be utilized at different stages of the discovery process, and it can be used as a benchmark for other steps [7]. HTS is performed in-vitro, whereas virtual screening (VS) is an in-silico operation to screen large chemical libraries [8].

When the target in question has a known three-dimensional structure, VS is called structure-based VS (SBVS). The output of HTS experiments yielded a massive accumulation of structure-activity relationship (SAR) datasets [9], [10]. Therefore, as a direct result of the collected data, the protein-ligand docking became the most frequently applied methodology of SBVS to predict the poses and strength of binding affinity for a predetermined binding pocket. The results of docking experiments can be employed to evaluate further and rank the potential drug candidates for selectivity and potency [11]. Indeed, the predicative compound can have off-target effects and may undoubtedly interact with other macromolecules. Thus, it leads to unwanted side effects.

The total number of possible drug-like small chemicals is estimated to be between 10^{30}–10^{60} [12]. The discovery of novel biologically active compounds will flourish our understanding of biological processes and improve therapeutic methods. Expanding the size of the chemical spaces to be screened is also becoming a significant issue in SBVS. A cycle of design-make-test-analyze (DMTA) includes many consecutive syntheses of ligands and the construction of biological assays [13].

The biological activity of a potent drug could be measured by the extent of its molecular interaction strength with a known target macromolecule, mostly a protein, and it is the primary mediator of interruption of a disease mechanism. Indinavir is a recently found protease inhibitor, blocking the enzyme's biological function that is sold in the market to treat AIDS/HIV patients [14]. It was approved last decade and identified before preclinical tests by employing SBDD using CADD mainly. Hence, there are a growing number of successful cases accomplished by computational approaches in drug discovery.

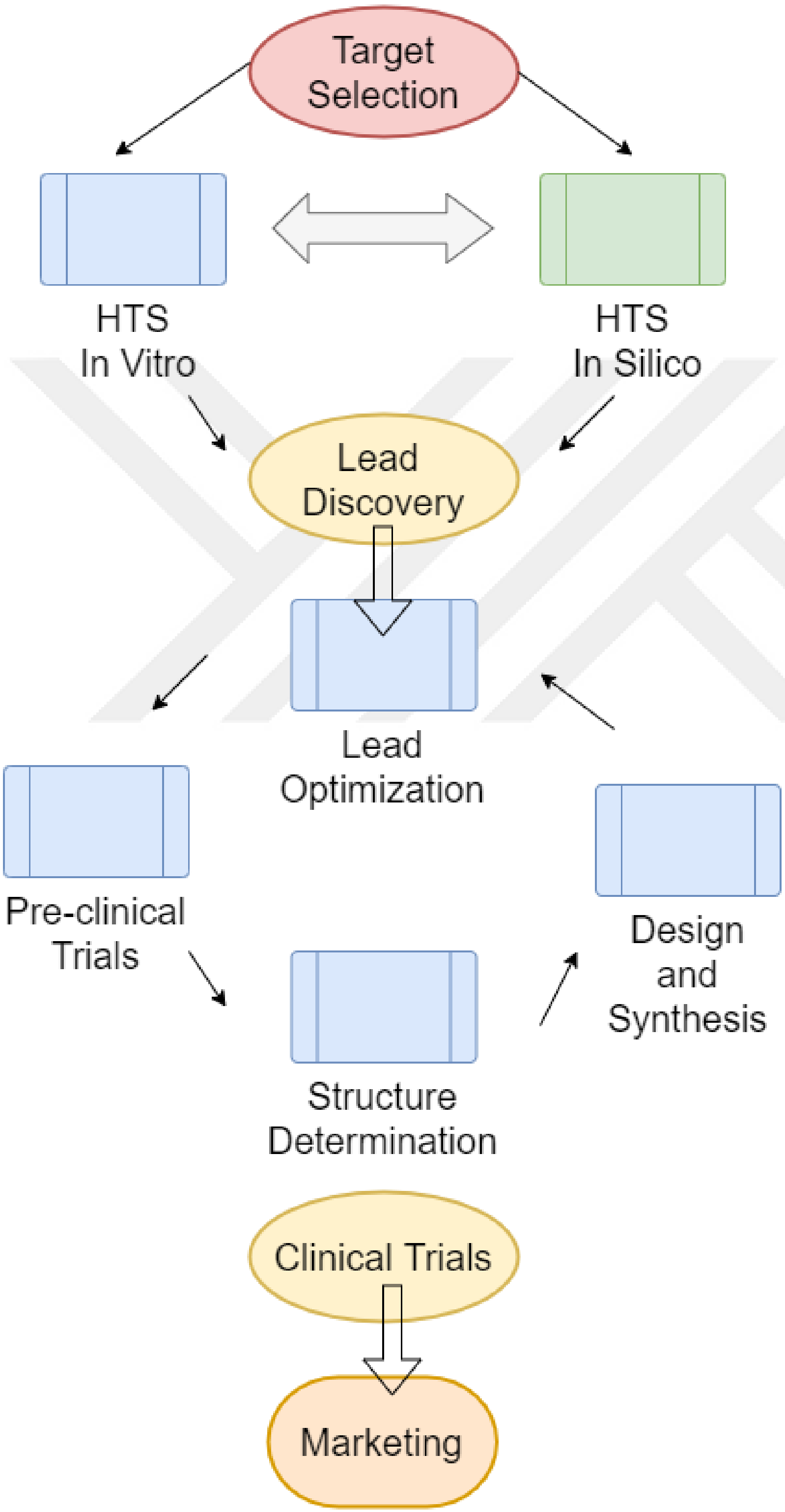

Figure 2.1 Schema of structure-based drug discovery

2.1 Molecular Recognition

Advancements in protein purification methods and crystallography techniques have caused the vast availability of target molecules to be determined structurally. Ideally, the availability of protein structures provides an excellent opportunity to elaborate on quantitative and qualitative aspects of protein-ligand complexes [15].

Recent studies show that the formation of such complexes is a direct product of a somewhat complex process of thermodynamic equilibrium [16]. The ligand and protein molecules are solvated into several conformers of their own and coexist in equilibrium. Water molecules, as a solvent residing at the binding site, are forced to be displaced from their position by the ligand to yield an ultimate solvated complex. A stable complex will be formed if the individual interaction strengths of both protein and ligand with solvent are smaller than that of the complex and the solvation medium. Thermodynamics evaluations do not foresee a complex formation favorable due to the vanishing conformational degrees of freedom. Moreover, spatial degrees of freedom of rotation and translations will be exterminated after a complex formation. To overcome the drawbacks of a formation, specific contacts between the ligand and protein should compensate for the net effect. Fig 2.2 provides the steps of the formation of the complex itself [16].

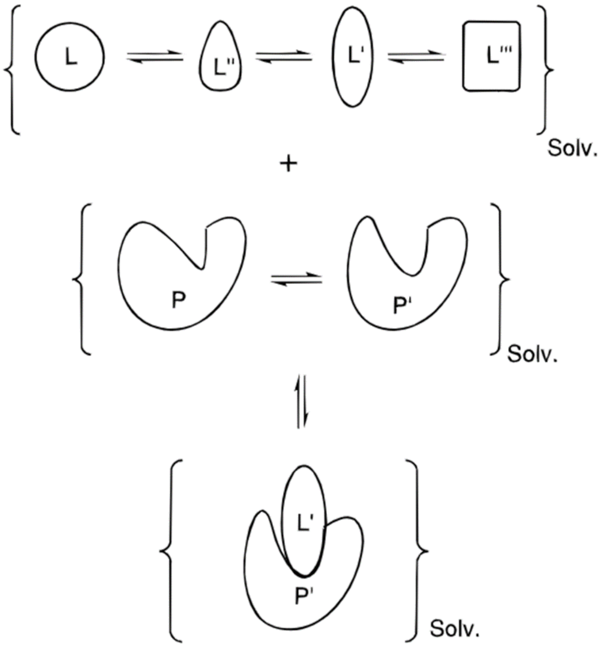

Figure 2.2 Formation of a complex

Crystallographic structure determinations of a vast variety of protein-ligand complexes can help elucidate the qualitative features of molecular recognition of complex formation. Nevertheless, the resolution of structural determination is not yet at the atomic scale, and the exact positions of atoms cannot be determined. Different types of non-covalent bonding between the host and guest molecule could represent molecular recognition. For example, hydrogen bonding, weak van der Walls forces, strong π-π interactions, hydrophobic forces and metal coordination forces can be accounted as non-covalent interactions [17]. In addition, the solvent molecules also have an indirect contribution to seizing the complementarity of complex molecular interactions between the host and guest. Recent developments in the Physics of molecular recognition boosted the computational approaches to handle the plethora of available experimental data at hand. Now, we have a better picture of a microscopic and dynamic snapshot of the experimental phenomena.

2.1.1 Thermodynamic Entities

A statistical ensemble's probabilistic nature of microstates can be closely associated with the amount of free energy in the thermodynamic system [16].

Let's take a system of *N* particles residing inside a volume of *V*, at temperature *T*, the system's free energy *F* can be defined as

$$F_{NVT} = \ (N!\,b^{3N})^{-1} \iint e^{-\mathcal{H}(\boldsymbol{r},\boldsymbol{p})/k_BT}\, d\,\boldsymbol{p}d\boldsymbol{r}\ , \tag{2.1}$$

where the factorial of *N* is disregarded if we take the particles as indistinguishable, and b stands for Planck's constant. The system's free energy can be calculated by taking the integral over all particle coordinates ***r*** and momenta ***p*.** The Boltzmann factor of the system Hamiltonian becomes

$$\mathcal{H}(\boldsymbol{r},\boldsymbol{p}) = \ \mathcal{V}(\boldsymbol{r};\boldsymbol{s}) + \mathcal{K}(\boldsymbol{r},\boldsymbol{p}), \tag{2.2}$$

the direct sum of potential $\mathcal{V}(\boldsymbol{r};\boldsymbol{s})$, and kinetic energy $\mathcal{K}(\boldsymbol{r},\boldsymbol{p})$ of the whole system. The potential energy can be estimated from empirical force fields and depends on a parameter set ***s*** yet is independent of the momenta.

It will become a necessity to express the standard relationship of free energy as,

$$F = U - T.S \tag{2.3}$$

where U is energy or H enthalpy, and S is entropy. One can calculate the change in free energy of the whole ensemble after complex formation at a constant temperature as,

$$\Delta F = \Delta U - T.\Delta S \tag{2.4}$$

The information provided by Eq. (2.4) is used as a thermodynamic signature of the molecular recognition process of interest. Identifying the terms of the equation is often regarded as an enthalpy-entropy compensation issue [18].

2.2 Ligand-Protein Affinity Scoring Functions

To quantitatively assess the affinity of the ligand and receptor molecules, a multitude of scoring functions (SF) have been created and implemented for the SBDD paradigm. Indeed, the prime objective of using SFs is to obtain a quick evaluation of binding affinity; they are not designed to ultimately reveal the physics of the interaction. In essence, they create a key balance between the overall efficiency and accuracy of the SBDD process [19].

The current SFs are classified as force-field or physics-based, empirical, or regression-based, potential of mean force or knowledge-based, and descriptor or machine-learning based, depending on the source of derivation. [19]. In the early seventies, a group of scientists had brought in the concept of computing the interactions between a ligand and a receptor by the known force fields. Force-field contains non-covalent energy terms, and they are the key elements used in this specific SF [20].

The second type of SF was first introduced by Böhm [21], and other well-known examples were also developed by others, like X-Score, Glide-Score, and Chemcore [22]–[24]. Empirical SF is a sum of individual contributions by different energy terms of simpler non-functional forms, fitted to available empirical data. On the contrary, the physics-based SF comprises sophisticated functional forms of each energetic addition with some parameters of interest. Empirical SF is limited to the capacity of available experimental training data sets, and their experimental values are mostly inconsistent [25].

In 1996, SMoG, a drug design software, offered the first known implementation of a knowledge-based SF.[26]. This SF is defined to be the summation of pairwise statistical potential between the contact atoms of protein and ligand molecules. Known examples of protein-ligand complexes from the PDB database make up the training set to construct the contact-distance potentials based on the knowledge of frequency of occurrence of each pair of atoms in the same dataset [27].

The final type of SF is based on a machine learning model that generates features from chemical descriptors of protein and ligand molecules using a well-established method known as quantitative structure-activity/property relationship (QSAR) analysis. The most well-known examples are the RF-Score, NNScore, and SFScore [28]–[30].

2.3 Molecular Dynamics Simulations

Even though molecular dynamics simulation (MD) is a computationally expensive method for assessing the strength of protein-ligand interactions in SBVS experiments, it is a widely utilized technique to supplement and validate early docking results in an SBDD campaign. As a novel successor to Monte Carlo simulations, which was originated back in the 18th century, MD entered the realm of research at the beginning of the 1950s [31]. Fermi and his colleagues' work were executed on a special analog computer at Las Alamos lab, MANIAC I, aimed to solve the dynamics of the time evolution of a many-body ensemble under many different force fields, and their work was named as Fermi-Pasta-Ulam-Tsingou problem [32]. In a nutshell, their study explained the behavior of a nonlinear physical system where the nonlinearity was a product of a small perturbation onto an initially linear system of interest. Since there were no direct analytic solutions to solve the equations, a heuristic-based numerical analysis was performed on that computer.

After its successful elaborations in theoretical physics, MD attracted other researchers from materials science. Moreover, first applications of MD in biochemistry and biophysics started at the early seventies. To simulate the motion of biological macromolecules, MD was applied, and it conveyed the mechanisms of their interactions with other molecules of interest. MD experimental runs has several limitations of initial settings and inadequacies in representing the entropic energy contributions. Nevertheless, it is a superior technique to all available quick snapshots, and most importantly, owing to the improvements in force-fields and computing hardware.

2.3.1 Force fields in protein-ligand complexes

An MD run is governed by an initial selection of a certain type of force field buried with various empirical parameters related to the atom types, types of chemical bonds, dihedral angular movements etc. which basically describes the interaction of particles at and near atomic scale. There are plenty of review articles overviewing the available force fields especially designed for biological macromolecules [33]. A specific type of water model is also assigned before starting an MD simulation [34].

The optimized potential for liquid simulations (OPLS) with an all-atom (OPLS-AA) is a favorable force field for the description of the dynamics of proteins and their interactions with other molecules, since proteins and organic liquids have common functional groups [35].

2.4 The MM/PBSA and MM/GBSA methods

Implicit/continuum solvation methods are also applied in and after MD simulations as an alternative to force-field based ones [36]. It is used to determine free energy of solute-solvent interactions in molecular recognition of protein-ligand complex formation(s). The Molecular Mechanics/Poisson−Boltzmann Surface Area (MM/PBSA) and the Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) methods are both applied vastly in most of the SBDD campaigns [37], [38].

Using solvent accessible surface area, continuum electrostatics calculations, or combinations of other entities, one can design different implicit solvent models. The accessible surface area of the solute molecule and the free energy transferred are linked linearly. The solvation mechanism plays a crucial role in this method. On the other hand, the mechanistic continuum approach is based on the enthalpy of free energy.

The energy change of a solvated molecule can be calculated by summing the solvent-accessible surface area of every atom multiplied with an empirical parameter of solvation. The measure of penetration of every atom into the solvent molecules is dependent on the Born radius. The Generalized Born model is equipped with solvent accessible surface area, and it is the most widely used method in implicit solvent approaches.

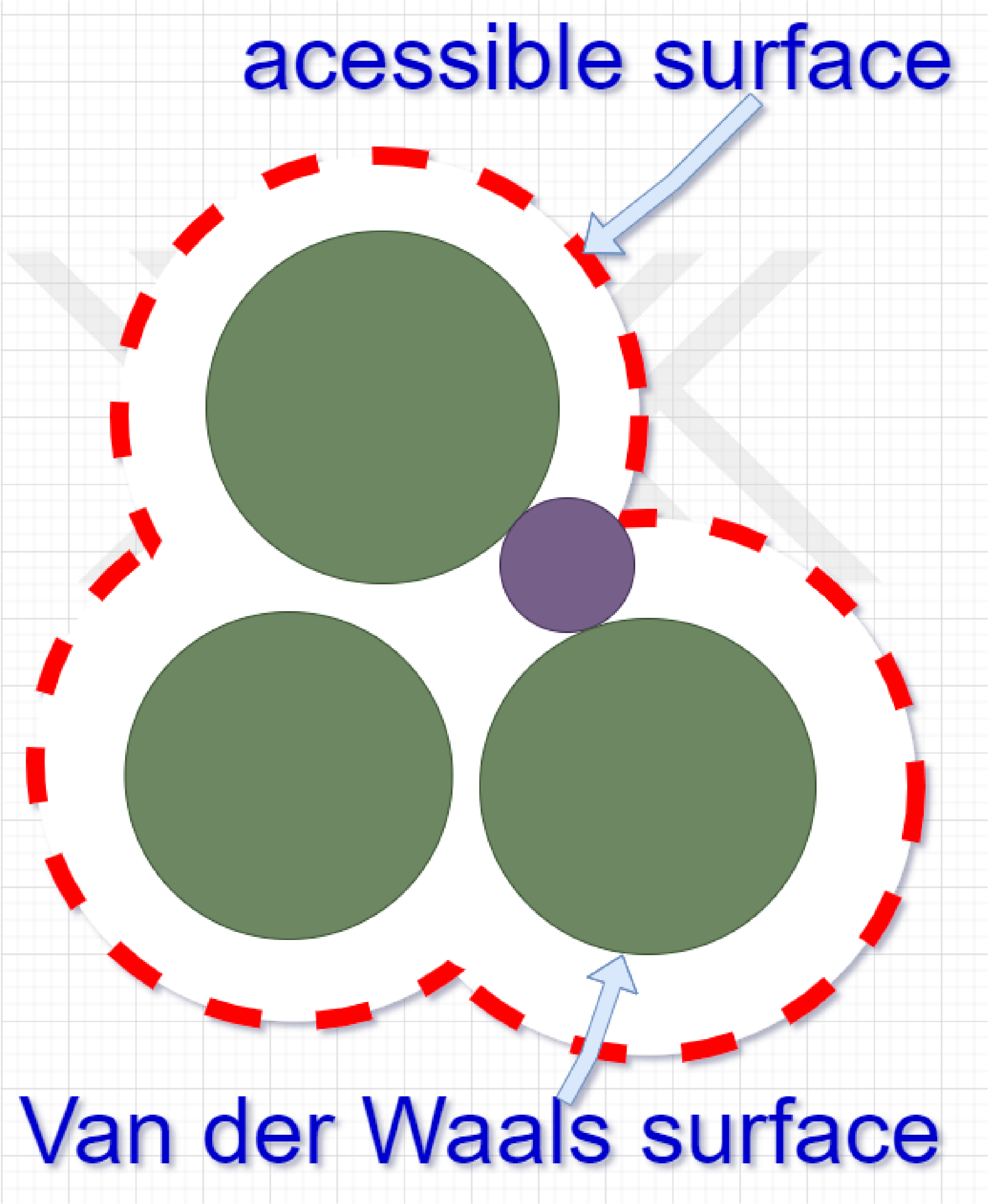

Figure 2.3 Accessible surface depiction

Chapter 3

Deep Learning Methods

3.1 Artificial Intelligence in Drug Discovery

Pioneering work by Alan Turing exemplified the possibility of a machine's ability to think or perform tasks done by human-like intelligence. His groundbreaking paper published in the philosophy journal "Mind" offers the idea of programming an electronic device to act like an intelligent being and provides the depiction of his famous "imitation game" that we call as Turing's Test [39]. His work is devoted to mathematical logic and philosophical assertations of what an AI would look like.

Pioneers from other disciplines, claimed novel ideas giving birth to the realm of AI [40]. To name a few, we present the following works. Wiener worked on cybernetics, which covers feedback and control. McCulloch asserted neural networks' resemblance to nervous system of simpler organisms. Newell and Simon studied the experimental psychology and many others contributed to the fields of communication theory, linguistics, game theory and statistical learning theories to shape the fundamentals of AI.

The approval process of a new drug is becoming more and more hectic, and the cost of discovering novel therapeutics is increasing at the same time. Failures in clinical trials can be overcome if we can employ better preclinical tests to check the efficacy and toxicity of drug candidates more effectively. The cost of discovery will be reduced substantially. Drug discovery has been revolutionized by the eminent changes in AI for the last decade.

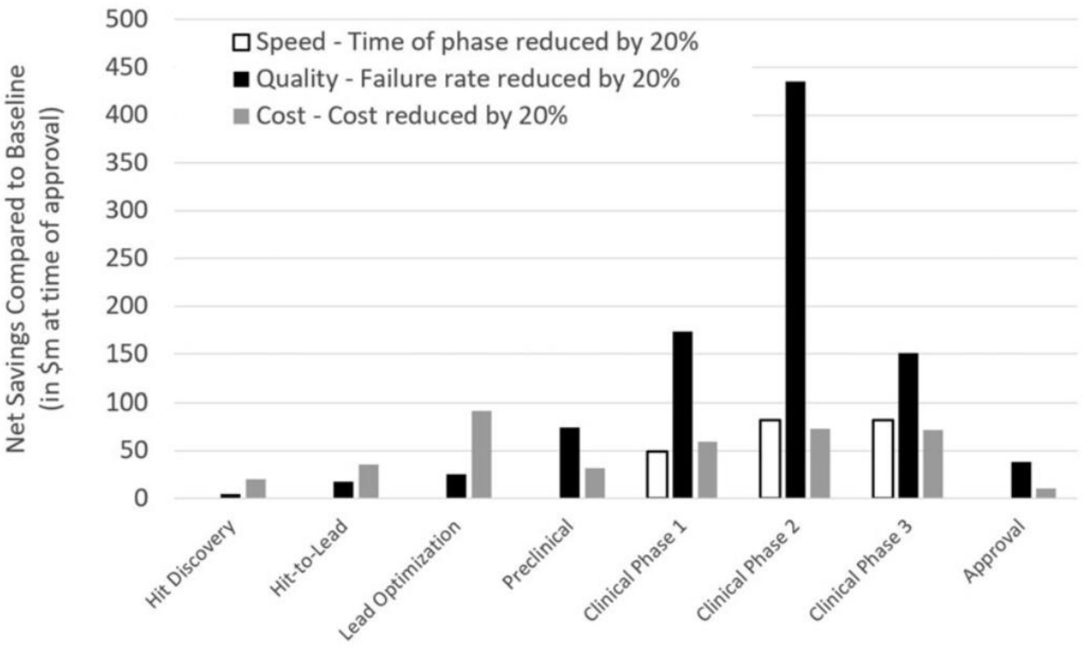

Figure 3.1 Cost, time, and quality factors. Adopted from [7]

The most common applications that AI took part in include VS, reaction prediction, de novo drug/protein design, and others. Here we can categorize those recent innovative approaches into generative and predictive AI tasks. A broad selection of AI methods and models derived from old school machine learning (ML) frameworks to artificial neural networks (ANN) were used to accomplish listed tasks. For example, convolutional neural networks (CNN), recurrent neural networks (RNN), graph neural networks (GNN) are the most applied ANN models in drug discovery. Nevertheless, our main interest is to provide a complete account of CNNs, and the rest of the ANN models and traditional ML methods is out of the scope of our work.

3.1.1 Feed-forward neural networks

Deep learning (DL) is a subclass of machine learning algorithms based on ANN with representation learning. DL has now matured into a highly successful framework for supervised learning algorithms. As a result, DL covers various application fields like drug design, bioinformatics, computer vision, health informatics, text processing but is not limited to these areas. Indeed, in almost all these fields, DL outperformed the human opponents drastically.

With DL, any supervised, semi-supervised or unsupervised learning method can be successfully applied. In unsupervised learning DL distinct patterns inside the training dataset can be identified without any external reference, unlike the supervised version requires a preprocessing of the dataset to classify it into the known patterns.

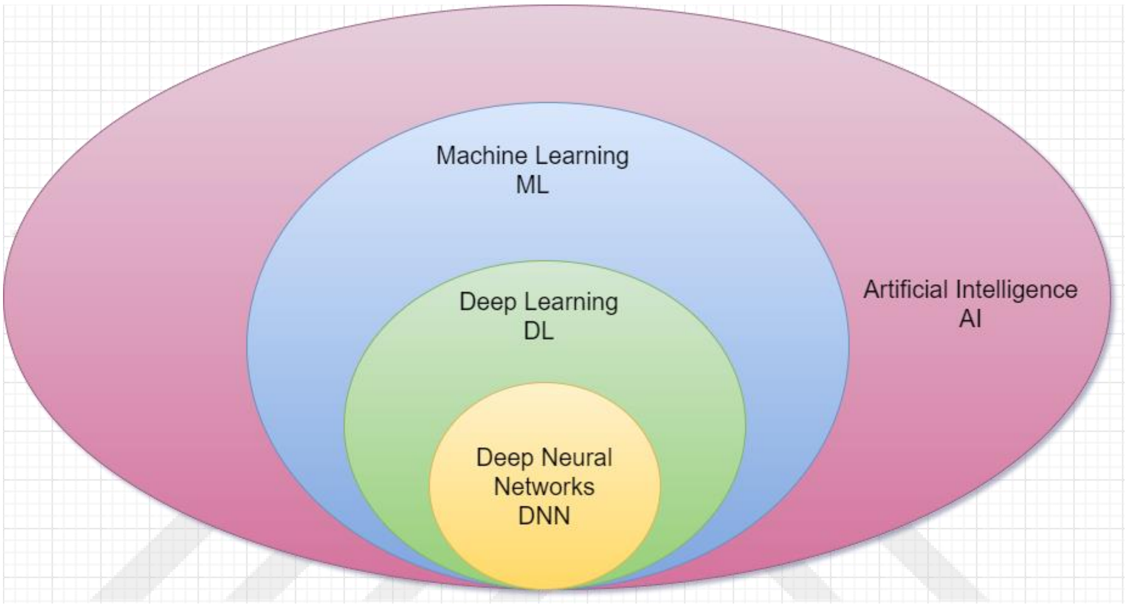

Figure 3.2 Schema of AI world

Initially ANNs were inspired from theories of neuroscience [41], but further advancements owe much to algorithmic improvements on the stochastic gradient descent (SGD) optimization method [42]. Interconnected nets of artificial neurons, like the neurons in an animal's nervous system, send information through each connection in the same way that synapses transmit signals. Many neurons are cast into layers that are assigned with transforming the initial input signal and traversing it to different layers. Neurons in the same layer are not dependent on one other but have a direct relationship to neurons in other layers of various types via largely non-linear transformations.

An ordinary ANN, sometimes called vanilla, contains input and output layers and at least one hidden layer, and all layers' neurons are connected fully. Unlike a recurrent one, a feed-forward neural network or multilayer perceptron (MLP) is built on an acyclic graph-like structure with input and output nodes connected only in one direction. Each node passes a computation to all nodes in the next layer starting from the input to the final output layer without any loops. The feedback connections which bring the outputs of the network back to the input layer is missing in vanilla networks. If we add feedback

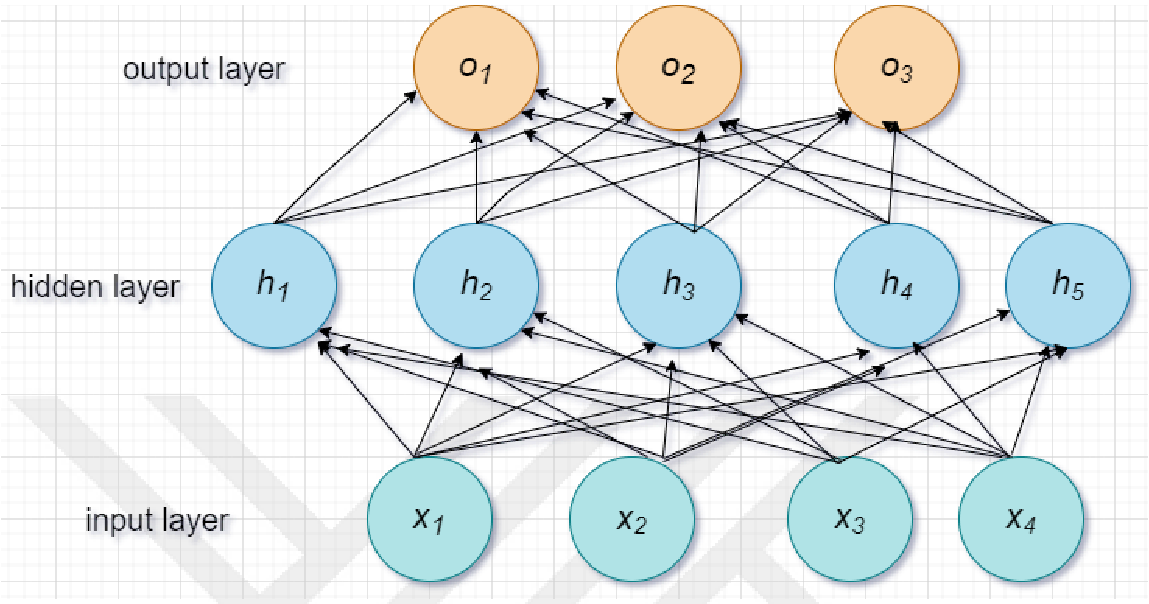

Figure 3.3 Standard CNN

connections to the model it becomes a recurrent neural network (RNN).

Each single neuron in an ANN does compute the weighted sum of the inputs from the previous nodes and subsequently triggers a non-linear transformation to yield its output value. Assume that a_j is the output of j^{th} neuron and $w_{i,j}$ is the sum of weights from neuron i to j ; thereby we will get,

$$a_j = g_j\left(\sum_i w_{i,j} a_i \right) \equiv g_j\left(in_j \right) \tag{3.1}$$

where g_j is the activation function which guarantees non-linearity of our model. If we want to write the equation (3.1) in vector form:

$$a_j = g_j\left(\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x} \right) \tag{3.2}$$

where **w** is the vector of all weights fed-forward and **x** is the vector of inputs. The most frequently used non-linear activation function is ReLU (Rectified Linear Unit) function, and GELU (Gaussian Error Linear Unit) is also becoming popular because of its surpassing effect on the vanishing gradient terms in the back-propagation algorithm:

$$ReLU(x) = \max\,(0, x) \qquad \textbf{(3.3)}$$

$$GELU(x) = \mathrm{x}\,\Phi(\mathrm{x}) \qquad \textbf{(3.4)}$$

where Φ(x) is distribution function of a gaussian distribution.

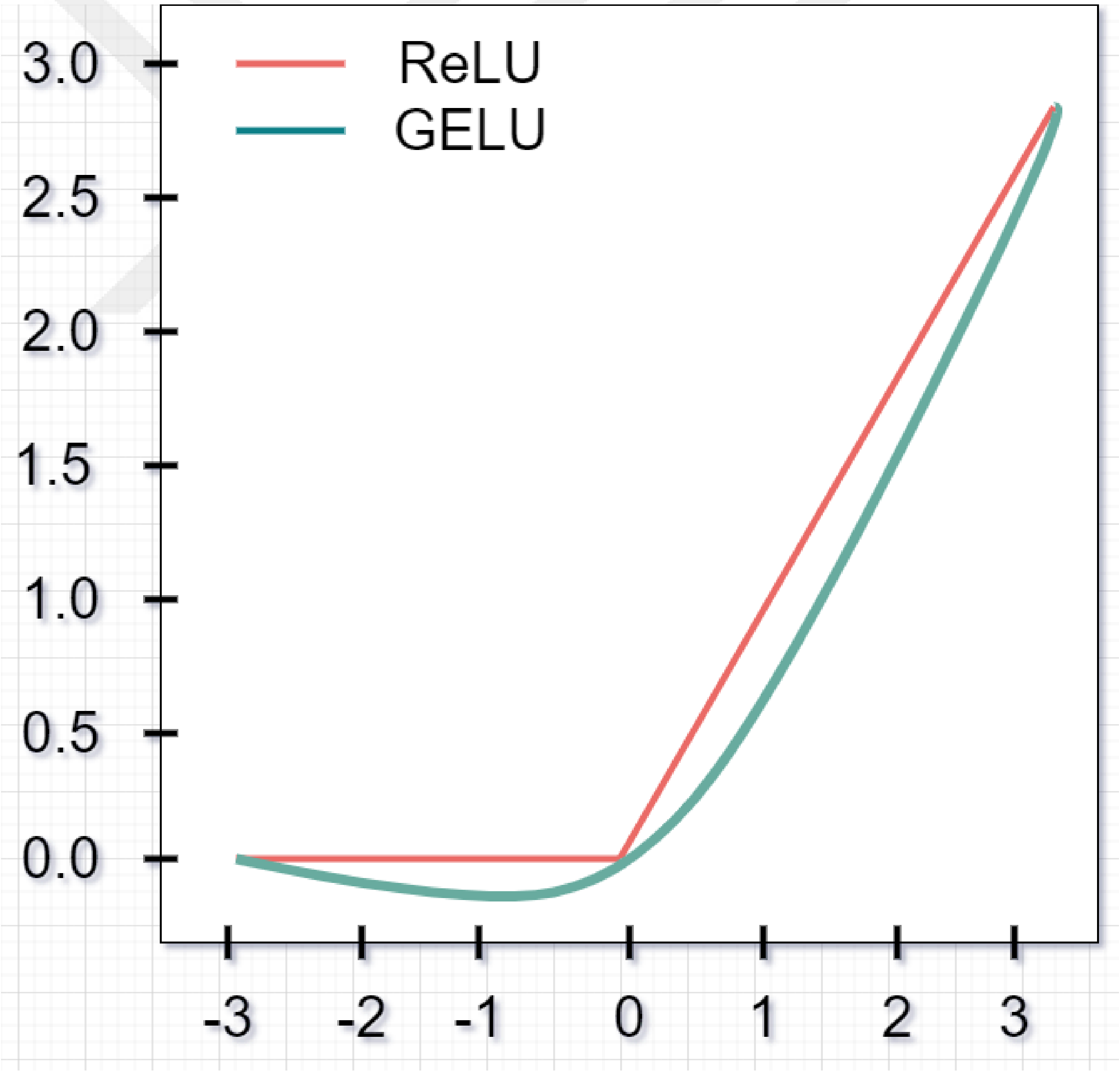

Figure 3.4 ReLU and GELU activation functions

Training is the process of learning the best parameters of the network and has been made possible by the back-propagation technique of calculating the derivatives of computational graphs in an automated fashion [43]. The partial derivatives of the cost function with respect to weights and biases of each unit needs to be calculated. The chain rule of a derivative is applied orderly from the output layer to the hidden layer(s).

We do not favor back-propagation only because of its superior speed in the learning process. Indeed, it provides a handful of insights on inter-dependencies of a minor change in the weights and/or biases of neurons and the whole network. Optimization of parameters and weights of each unit is achieved by a mini-batch stochastic gradient descent algorithm more efficiently.

The essential ingredients of training an ANN with a supervised learning scheme should have a data set of inputs with its targets, a definite architecture of the network, a non-linear activation function, a loss function, a computational technique to handle derivatives and an optimizer to control the overall procedure.

3.1.2 Convolutional Neural Networks

Convolutional neural network (CNN, ConvNet), inspired from neuroscience, is the most widely used ANN technique. At its earlier times, CNNs were initially constructed from their resemblance to a set of biological processes (BP) of the visual part of the animal cerebral cortex [44]. More specifically**,** Hubel and Wiesel's work by recording the electrical activity of neurons in the visual cortex of cat brain, showed that neurons at different layers detects varying complexity of visual images. The neurons on the first layer receives primitive features of the images and the second layer of neurons captures more advanced forms shaped by the combination of previous layer. As they have pointed out in their experimental work, more sophisticated features can be obtained by combining the simpler features at first hand.

Scaling, translation, and rotational invariance of the spatial data is a needy feature in designing a model's architecture. For the same token, a set of features of having local receptive fields of convolutional operations, sharing the weights between the units of the same layer, and sub-sampling through the layers are inherent to the convolutional neural networks. To train and learn to recognize handwritten character images, a CNN architecture, a LeNet-5 as shown in Fig 3.4, was successfully adopted. Fully connected MLP could have been safely applied. However, the parameter overload and requirement for additional feature identifications were the main reasons for choosing a CNN for such an image recognition task. In addition, overall learning accuracy is superior in the case of a CNN application than that of a vanilla ANN.

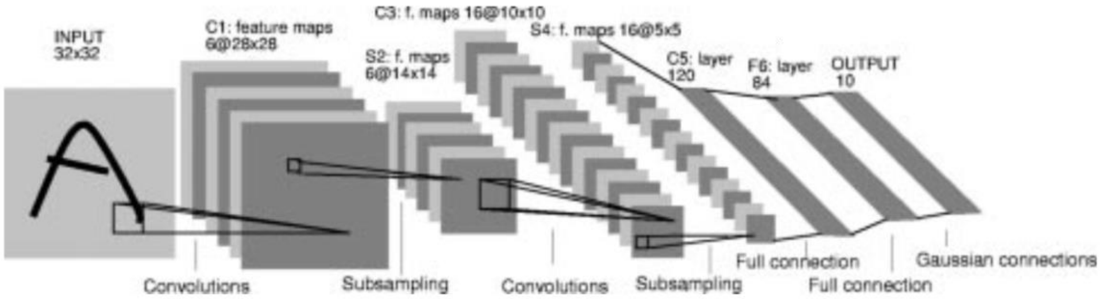

Figure 3.5 LeNet-5 Model Architecture [45]

The LeNet-5 model structure is built on seven layers, three of which is convolutional, and there exist two average pooling layers. Two dense layers are added on top of feature generating convolutional layers with a final SoftMax activated output layer. Input images of handwritten and typewritten characters with a grey colour channel have the input size of 32x32, and the convolutional layer of feature maps of size 28x28 are obtained by applying six kernels of size 5x5. The layer of sub-sampling S2 contains six additional feature maps of size 14x14, a product of average downsampling. Convolution layer C3 is constructed with sixteen filters of size 5x5 and generates a higher level of feature mappings of size 10x10. The size of layer C3 is dropped to its halve, and a new set of sixteen feature maps of size 5x5 is created after average pooling. The final convolution layer contains 120 units of size 1x1 and comprises convolution filters of size 5x5 from S4. Even though C5 and S4 are fully connected for that provided input size, C5 will be sparsely connected and serve its role for a more extensive network. Hidden layers are activated by tanh function [45].

From its early designs of simpler structures to today's novel CNN architectural designs, each model's representational learning power has been strengthened with every type of enhancement or advancement. Fig 3.5. shows a rich list of models designed in the history of CNN. Based on their architectural similarities, an evolutionary tree is constructed to present the urging developments in the architecture of CNN models. Each design element is shown as a different branch, and the tree's stems are the initiators of innovative approaches [46].

Table 3.1 lists the available DL frameworks where CNN architectures can be implemented and perform the data representational learning tasks of all sorts. Tensorflow , which is developed by Google, has the best user ratings from Github visitors and developers and it is easier to use and develop.

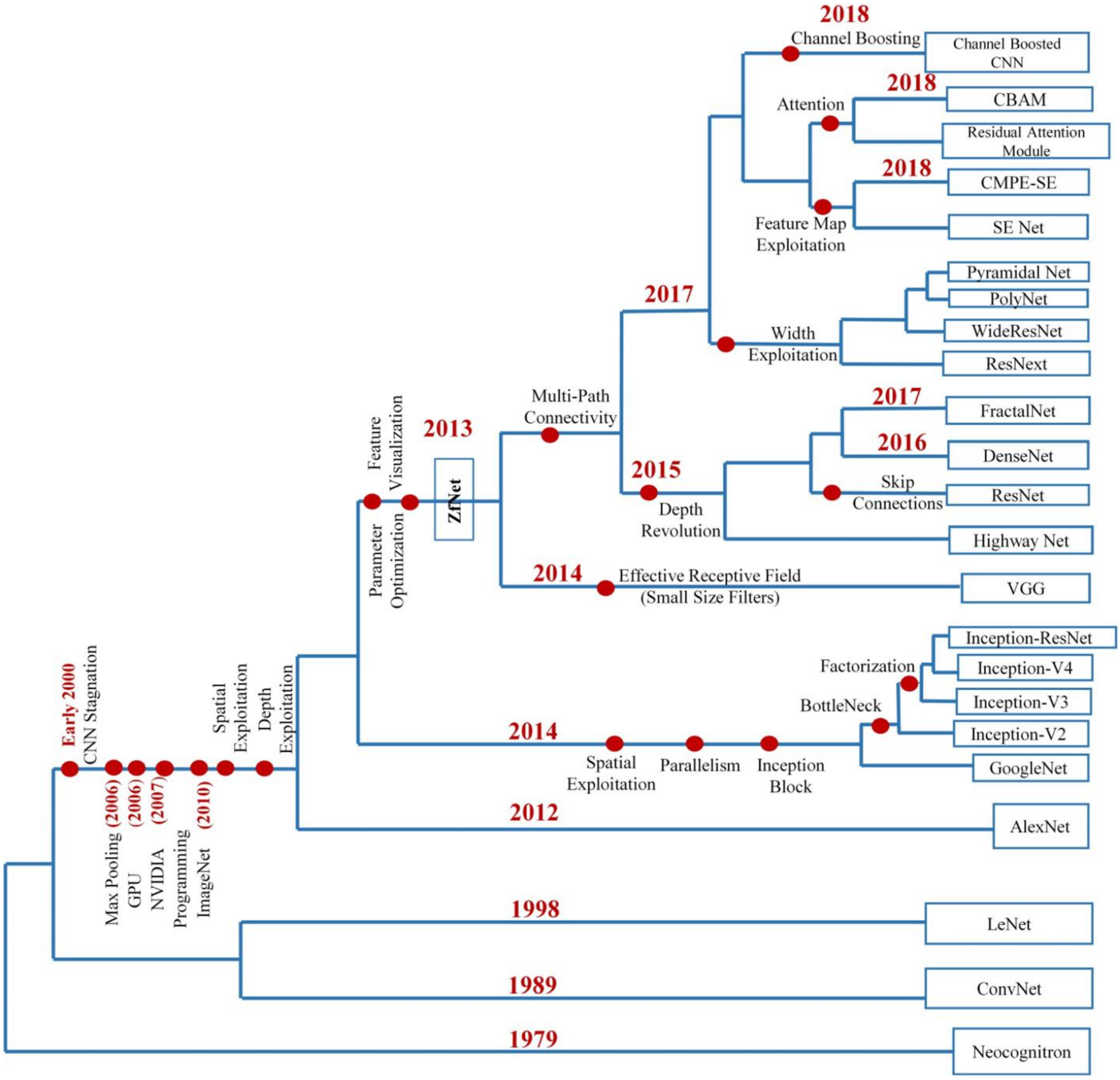

Figure 3.6 Evolutionary tree of known CNN architectures [46]

Table 3.1 Popular Frameworks of CNN

Brand	Copyright	Implemented in	Released
TensorFlow	Open Source	Python and C++	2015
Keras	Open Source	Python	2015
Caffe	Open Source	C++	2015
MatConvNet	Oxford	MATLAB	2014
MXNet	Open Source	C++	2015
CNTK	Open Source	C++	2016
Theano	Open Source	Python	2008
Torch	Open Source	C and Lua	2002
DL4j	Open Source	Java	2014

Chapter 4

Experimental Results

4.1 Computational Resources

Experiments were carried out on a bare-metal high-performance cluster (HPC) with one monitoring and four compute nodes. The operating system, scientific computation software(s), and DL framework(s) were all installed from scratch. Each computational node is built on four Tesla K60 GPUs, 28 CPUs, and 256 GB RAM.

4.2 Conventional/Traditional SBDD Campaign

We have employed a full spectrum SBDD campaign on a selected protein, Tyrosine-protein phosphatase non-receptor type 11, which is coded by the gene PTPN11 of Homo Sapiens organism, screening against the approved and about to be approved drugs and substances. We are going to present the bare minimum of the outline of VS and MD, and free energy experiments conducted and the selected significant results only.

4.2.1 Structural Data Files and Preparation for Docking Experiments

We selected PDB entry of 2SHP as the target molecule's X-Ray crystallography model file. The Broad Institute's drug repurposing hub provides a list of chemicals to screen potential drugs. Their repository contains a collection of FDA-approved drugs, drugs at clinical trials and pre-clinical compounds, making a complete list of 13,553 substances of 6798 unique compounds. We have downloaded the structure files of matching items of selected compounds from PubChem. We also retrieved a similar drug repurposing list of 5811 substances from the Zinc15 repository by browsing the in-trials subset collection. All the available structure files were downloaded accordingly.

Because the three-dimensional structural files for some of the ligands were not available, they were processed using RDKIT. After adding hydrogens and minimizing the 3D structure of each ligand in the screening library, it was ready to dock. The receptor was also prepared for docking using PDBFixer. We have cleaned the receptor from existing water molecules, added missing residues and hydrogens.

The Fpocket tool was used to inspect and identify potentially druggable pockets. Pocket 15 is chosen as a potential protein-ligand binding site. We used Pymol to visually inspect the druggable pockets and identify the binding area encompassing the biologically essential binding domain, N-SH2.

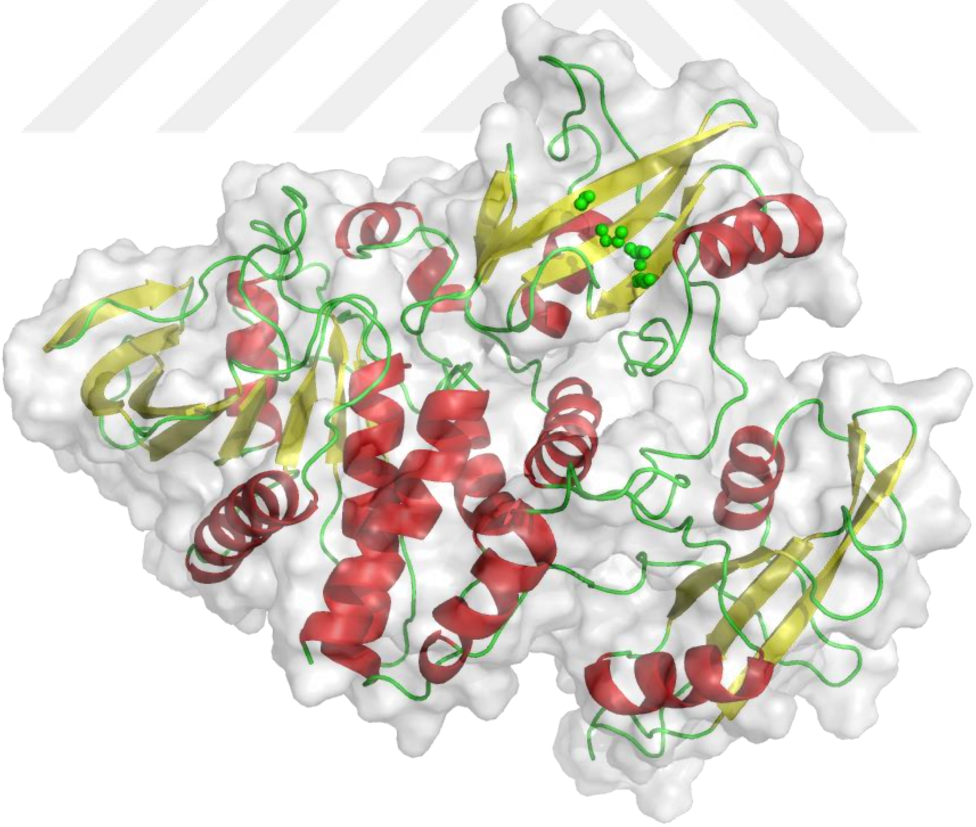

Figure 4.1 2SHP_fixed and pocket 15 at N-SH2 domain

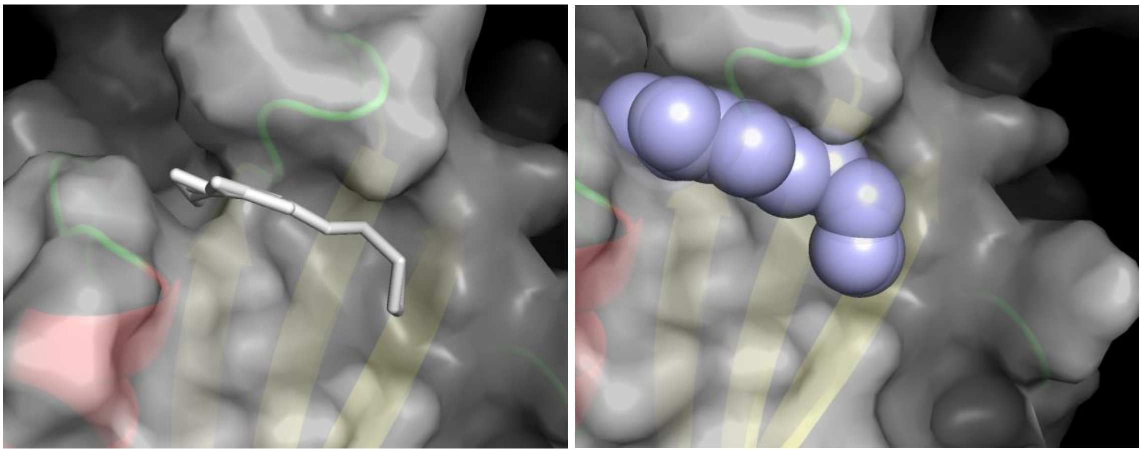

Figure 4.2 Closer view of ligand binding site of pocket 15 and a representative ligand in the form of sticks and spheres

4.2.2 Docking experiments

We used Smina [47], a variant of Autodock Vina [48], as our docking scoring utility. We selected the grid box of a rectangular parallelogram with the coordinates (-2.5, -38.8, 39.3) and axis sizes (14.0, 14.6,19.3) plus 4 Angstrom (Å) on each axis as the search space for the best binding poses. The exhaustiveness parameter, number of MonteCarlo chains was set as 16, and the seed was selected as 43. We ran docking experiments against both chemical spaces prepared and saved the resulting ligands with the best nine conformers' structure file.

We selected the best ranking chemicals with a minimum Smina score of -8.00. They were tabulated and selected only those without any biological activity with targets of any kinase enzymes for further analysis. Table 4.1 shows the list of compounds selected as the lead molecules of interest.

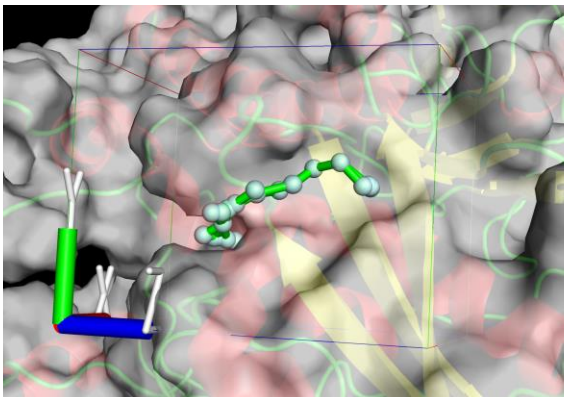

Figure 4.3 Grid box of search space

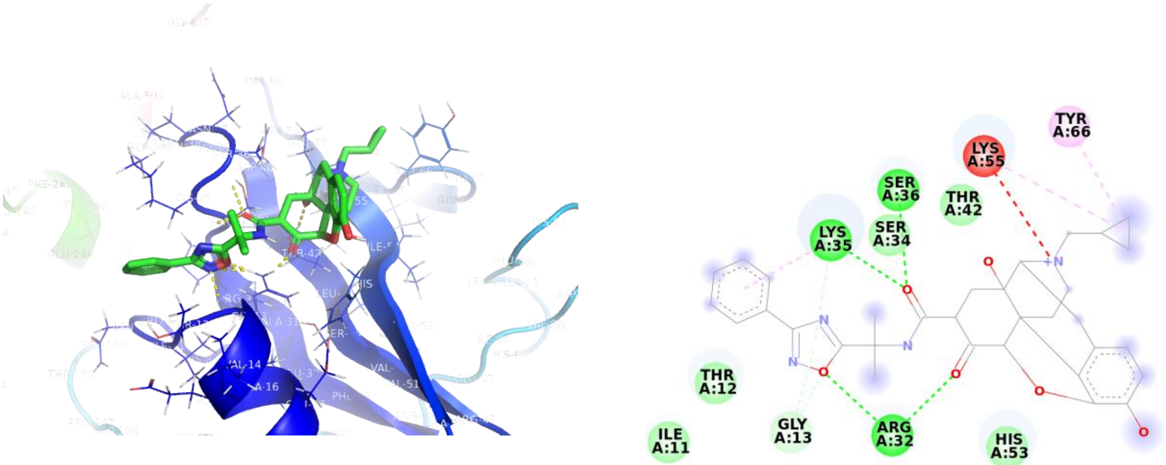

Figure 4.4 Lead molecule with PubChem id 54732242, best docking pose and 2D ligand interactions with residues of the receptor

Table 4.1 Lead molecules

ZINC	PubChem	Name	Affinity
ZINC01612996	60838	Irinotecan	-9.3
ZINC68267814	51049968	Rimegepant	-9.2
ZINC252670820	12940973		-9.1
ZINC11679756	135449332	Eltrombopag	-9.1
ZINC53073961	68723	Antrafenine	-8.9
ZINC100378061	54732242	Naldemedine	-8.8
ZINC100054519	3082214	4-Cis-Hydroxycyclohexyl Glyburide	-8.8
ZINC208938373	73774610		-8.7
ZINC40165257	151223	Estriol 3-sulfate 16-glucuronide	-8.7
ZINC85552271	21252309	Cholic acid glucuronide	-8.7
ZINC113459996	53340771	Glpg-0187	-8.7

4.2.3 MD Experiments

To test the stability of each complex formed with the leading drug candidates, MD simulations were carried out using Gromacs ver. 2021.03 [49]. Protein structure was cleaned and prepared for running simulations processing the script provided by the Chimera package. We selected the OPLS-AA [50]as the force field and SPC216 as the explicit water models to describe the ensemble. Ligand topologies and their force field parameter files for the force field selected were calculated by LigParGen [51] command-line tool. We exerted a dodecahedral unit cell, and the complex was subject to solvate inside it. If necessary, we have added ions to neutralize the whole system.

We ran a 50,000 step of the steepest descent minimization of the complex formed between the crystal structure of the protein and the initial docking confirmation of selected ligand. The minimization is required to keep the system free of steric clashes and inappropriate geometries of molecule groups. The NVT equilibration of position restrained molecules with a reference temperature of 300 K was applied to the system to

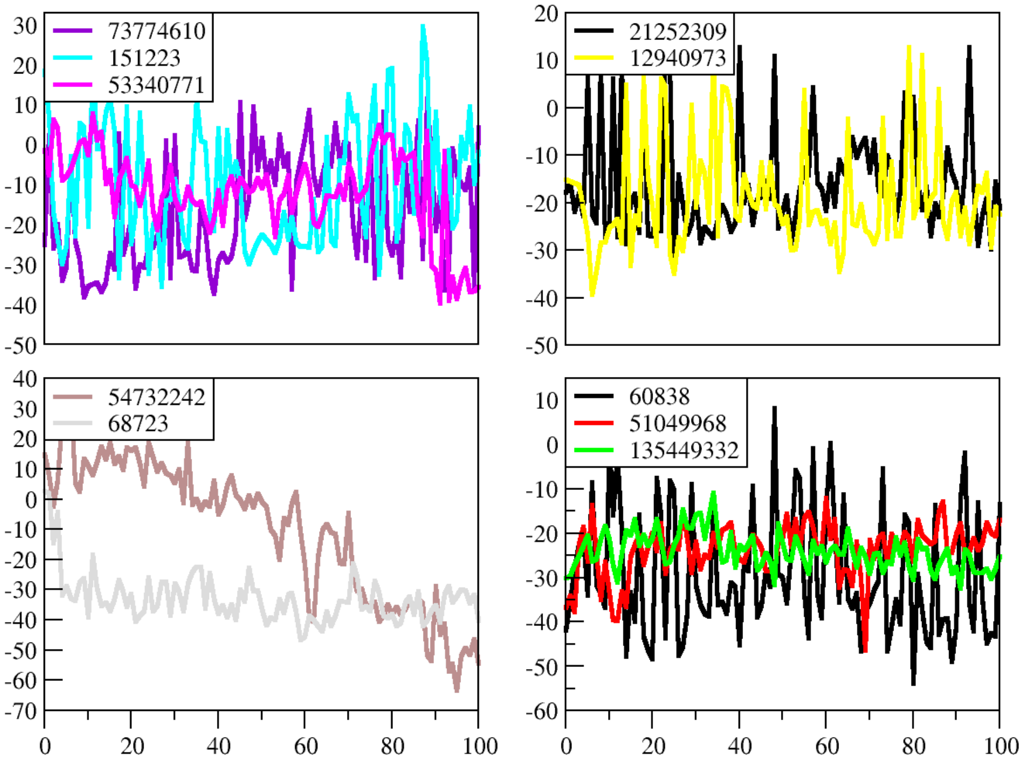

Figure 4.4 MM./PBSA free energy scores

stabilize the ensemble thermally. An NPT equilibration with a reference pressure of 1 bar was also employed to achieve an isobaric ensemble stabilization.

After completion of two equilibration runs, we conducted position restrains free MD runs of 100ns with the complexes formed with 10 selected ligands. After a quick evaluation of the resulting MM/PBSA [52] total binding energies of the complexes we have eliminated five of them. An extension of 150ns long MD runs was processed for the selected five successful candidates, shown in bottom panels of Fig 4.5.

Table 4.2 MM/PBSA scores

Pubchem IDs	**Binding energy kcal/mol (0-100ns)**
60838	-29.39
68723	-32.79
151223	-10.17
12940973	-19.06
21252309	-16.06
51049968	-23.70
53340771	-11.82
54732242	-10.86
73774610	-16.92
135449332	-24.07

4.3 A 3D CNN model by Pafnucy

Pafnucy [53] is selected as one of the representative 3D CNN models build to compare with ours. Their model uses the information of 3D cartesian coordinates of the ligand and receptor molecules forming a complex, and a list of chemical descriptors of them. Their spatial data is obtained from a 20 Angstrom cube of volume centered on the geometric center of the ligand encapsulationg the neighboring residue atoms of the receptor. Only the 3D positions of heavy atoms 1 A apart was selected as data-points There are 19 additional features resembling the color channels of image data.As shown in Fig 4.6, 4D tensor of input data has an initial shape of 21x21x21x19. Additional descriptive features can be checked at their publication. [54]

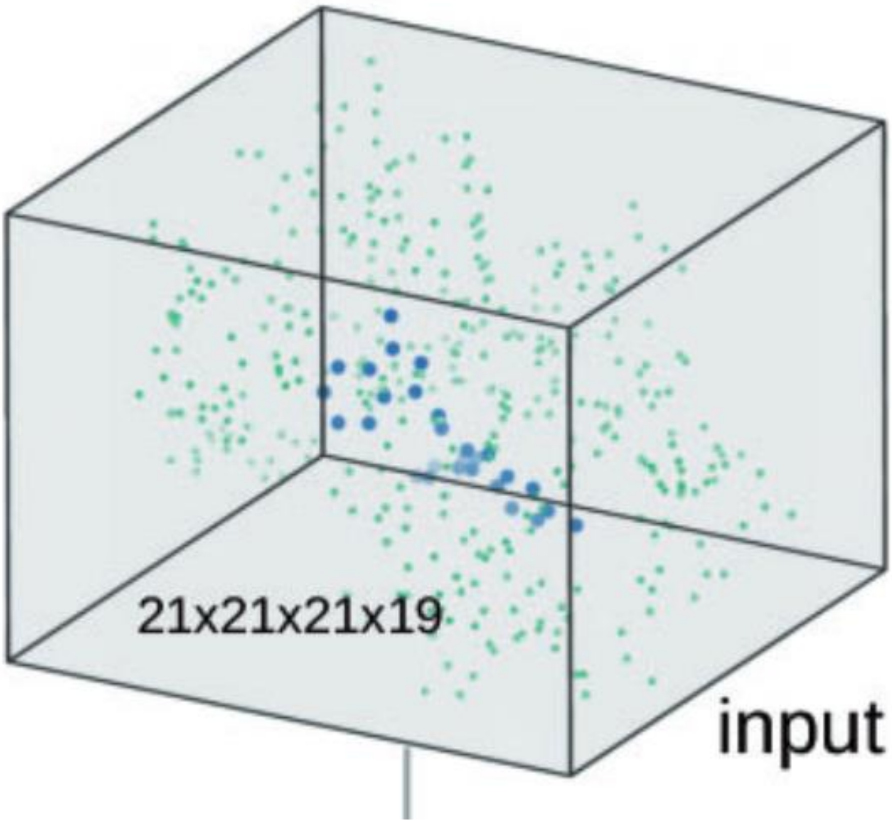

Figure 4.5 Representative depiction of input data

4.3.1 Datasets

The latest available version of PDBbind database [55], version 2020, is downloaded and split into 4 different datasets, namely, training, validation, test and core2013. The database contains 19,443 protein-ligand complexes in total. The binding affinity of ligand-protein complexes is expressed as the values of pK_a, calculated as minus value of the logarithm of the dissociation or inhibition constants. Each entry of the complex is curated and processed from their original references and the structure files of the complex and binding pockets as the interaction regions of the molecules is provided separately. We have used the same datasets for our model training and testing as well.

4.3.2 Architecture of the network

Their model designed to predict the binding affinity score of complexes. The final dense layer contains a single unit to create the output of prediction. Initial part of the model is situated with 3 layers of convolutional networks that serves as machinery of feature generation. A subsequent block of three dense layers is added to process the new features obtained to yield the final output.

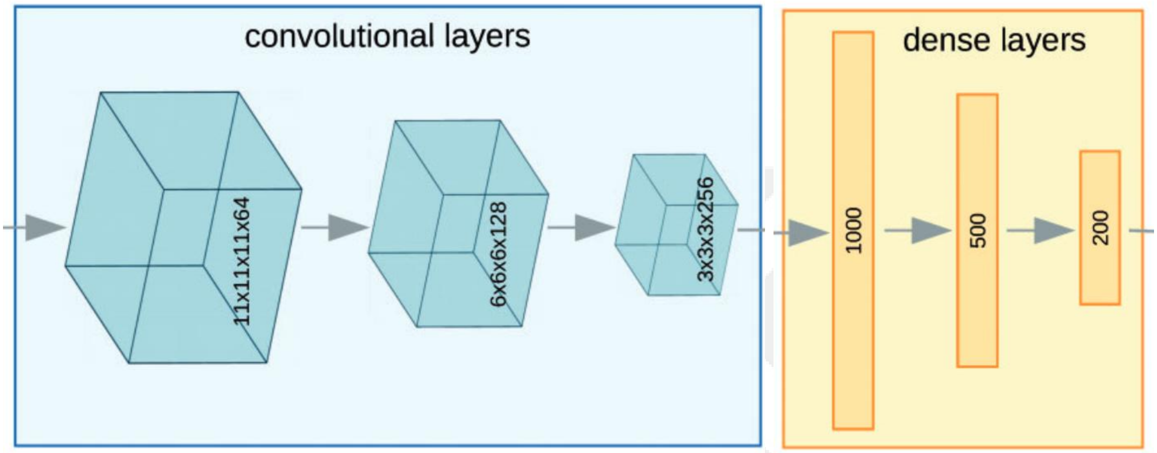

Figure 4.6 Architecture of Pafnucy

The spatial data with 19 attributes is introduced as a 4D tensor and transformed through a block of convolutional layers of 64, 128 and 256 kernels. Subsampling is applied after each convolutional layer with a max pooling operation of a cubic filter from 5-A to 2-A. Flattening of the last layer in the convolutional block will provide an input to the next layer of fully connected units. There are 1000, 500, and 200 units at each layer in this block. Dropout operation with a rate of 0.5 is applied to each dense layer to increase the generalization power of the model. All layers in both convolutional and dense block is transformed non-linearly by a ReLU activation.

4.3.3 Training with Back-propagation

Weight initialization for the convolutional block was applied using a truncated normal distribution with a zero mean and 0.01 standard deviation. All the biases were set as 0.1. Similarly, the initialization of weights for the dense layers was employed using the same method with a zero mean and the standard deviation with a value of inverse of the square-root of the total number of incoming neurons. Finally, all the biases have the same value for the dense block also.

Adam optimizer with a learning rate of 10^{-5} was the method of optimization used instead of a regular SGD. Mini-batch size was set as 5. The dropout and L2 regularization with a rate of 0.001 was applied to reduce the overfitting in training process. To overcome the rotational variance of the selected grid-box of the input data, 24 different orientations were augmented to increase the prediction rate.

4.3.4 Results and metrics of back-propagation

Only the first rotational orientation was selected for calculation of the errors, RMSE, in the training and validation sets. As can be seen in Fig 4.7, at epoch number 18 the model yields the best results and selected as the best weights for testing the model. Both metrics of RMSE and MAE was calculated and the R value of Pearson’s correlation between the predicted and experimental values was also computed. As proposed in CASF [55], standard deviation (SD) in the approximation line was also measured for each prediction.

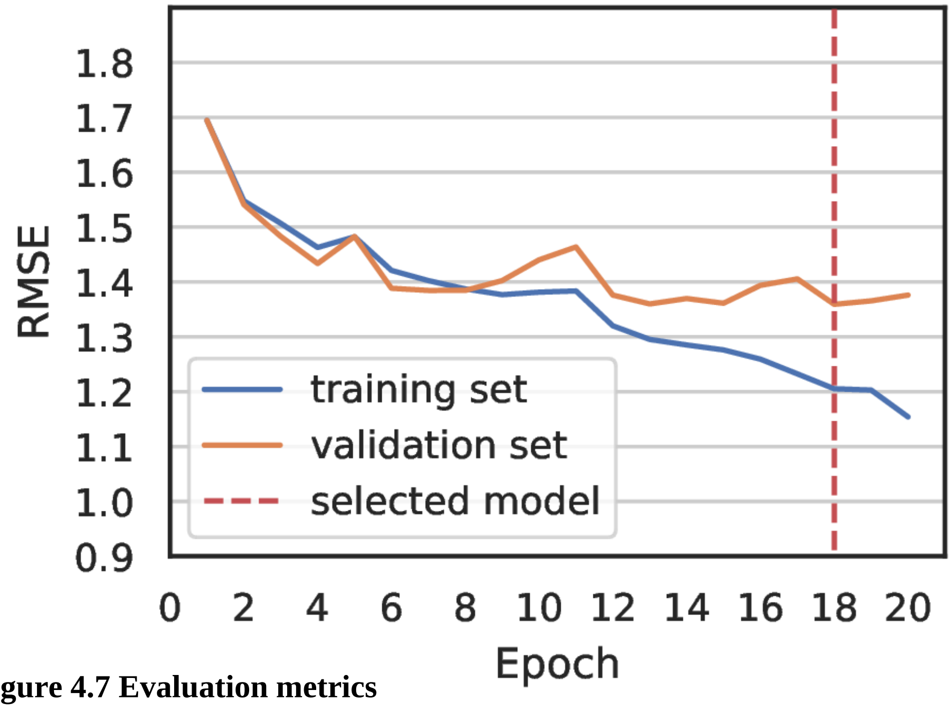

Figure 4.7 Evaluation metrics

The training dataset outcomes the lowest errors in both RMSE and MAE methods. The untrained sets yielded the accuracies of 0.70 in core2013 and 0.75 in test set prepared from PDBbind ver.2020.

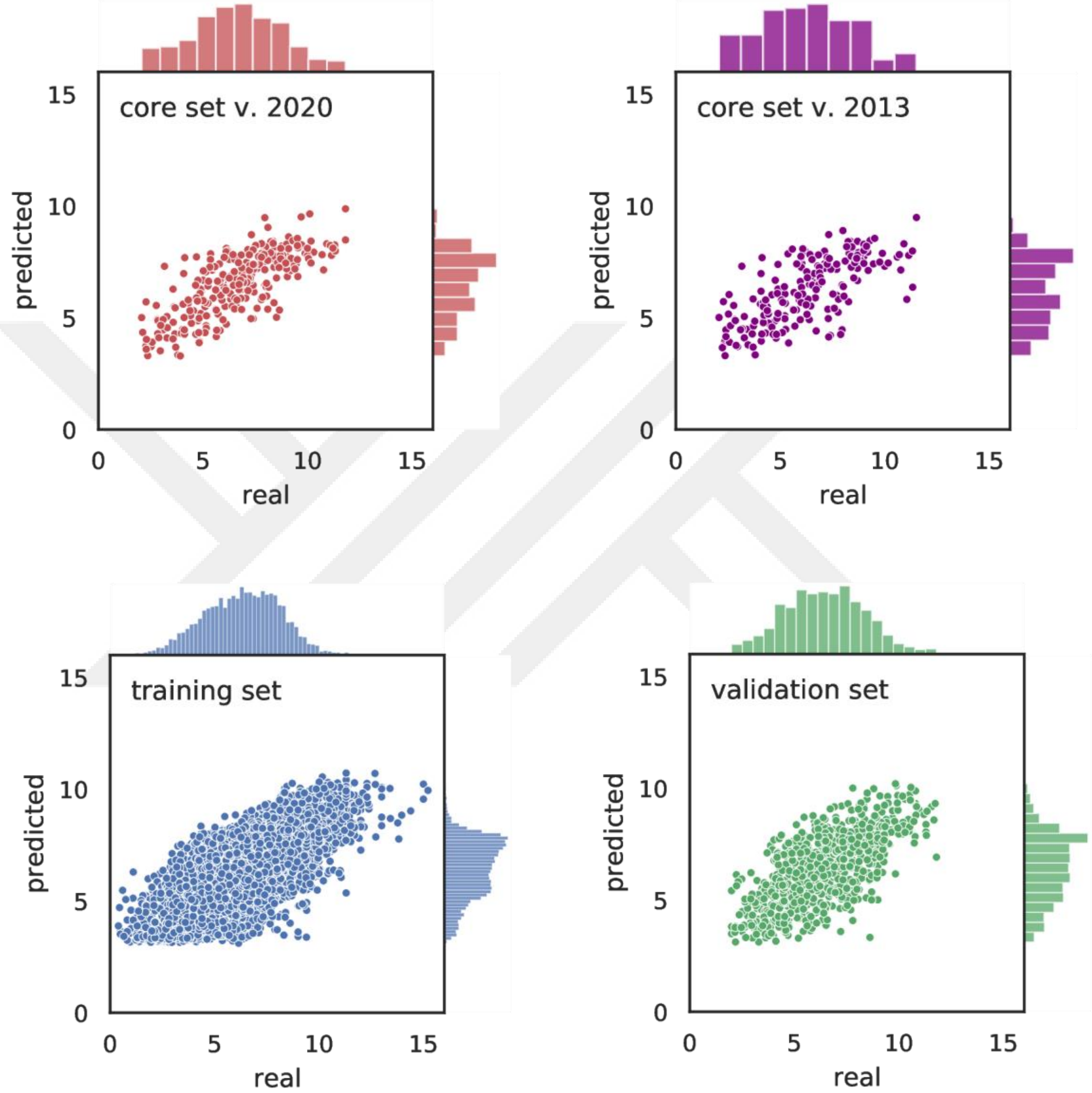

Figure 4.8 Correlations between the predictions made by Pafnucy and the experimental data

Table 4-3 Scores of evaluations

Data Set	**Size**	**RMSE**	**MAE**	**R**	**SD**
Training	18082	1.205	0.946	**0.76**	1.848
Validation	1000	1.359	1.071	**0.71**	1.906
Test	266	1.437	1.146	**0.75**	2.156
Core2013	195	1.618	1.261	**0.70**	2.154

4.3.5 Binding affinity of 2SHP with the leads

Results of the binding affinity scores of predicted by the model is shown in the Table 4.5. Three of the lead molecules, which were prioritized after MD simulations and MM/PBSA free energy calculation results, has also ranked the same order, and got the best results.

Table 4.4 Predictions made by Pafnucy

Pubchem IDs	**Binding Affinity Predicted by Pafnucy**
12940973	6.305908
135449332	6.588825
151223	6.551441
21252309	6.439477
51049968	6.950965
53340771	6.91218
54732242	6.675267
60838	6.902526
68723	7.04899
73774610	6.045316

4.4 Our Model deepMLR

First version of our model has been built and designed as a 1D CNN. Instead of atoms' exact positional 3D coordinates, we have used both binding pocket and ligand molecules' every atom's total area of solvent accessible surface. We have used RDKIT [56]to generate features of related physiochemical attributes of selected heavy atoms from both the ligand molecule and the pocket region. Cheminformatics feature generation failed for some of the entries in our datasets and they were excluded from training procedure. We obtained 13342 entries for training, 697 for validation and 225 for the test sets. RDKIT provides a library, rdkit.Chem.rdFreeSASA , to calculate solvent accessible surface area of each atomic element. Their library provides a method to generate our primary feature of interest to describe the interaction of interacting atoms of both partners. The list of features generated is presented in Table 4.7.

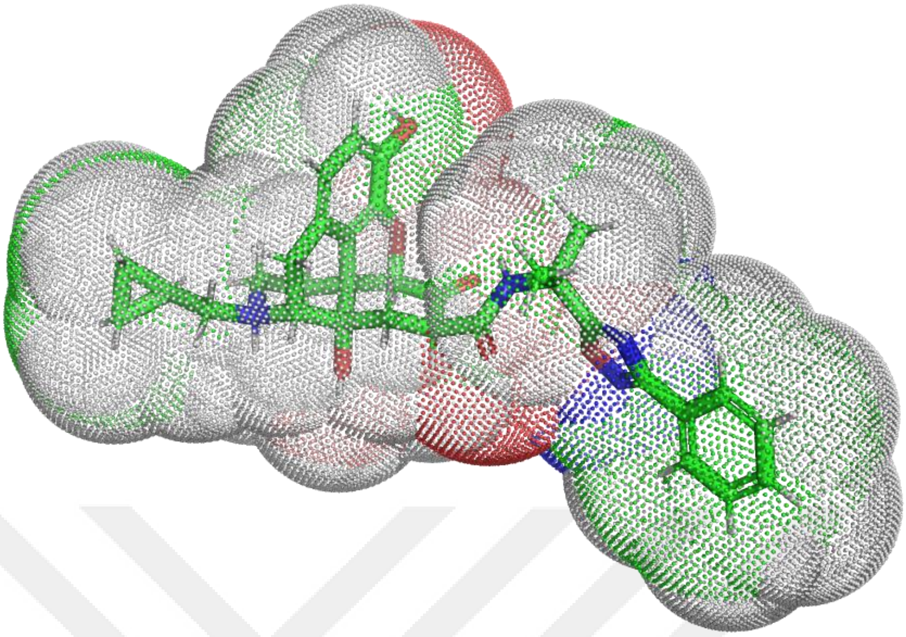

Figure 4.9 Solvent accessible surface area of the lead 54732242

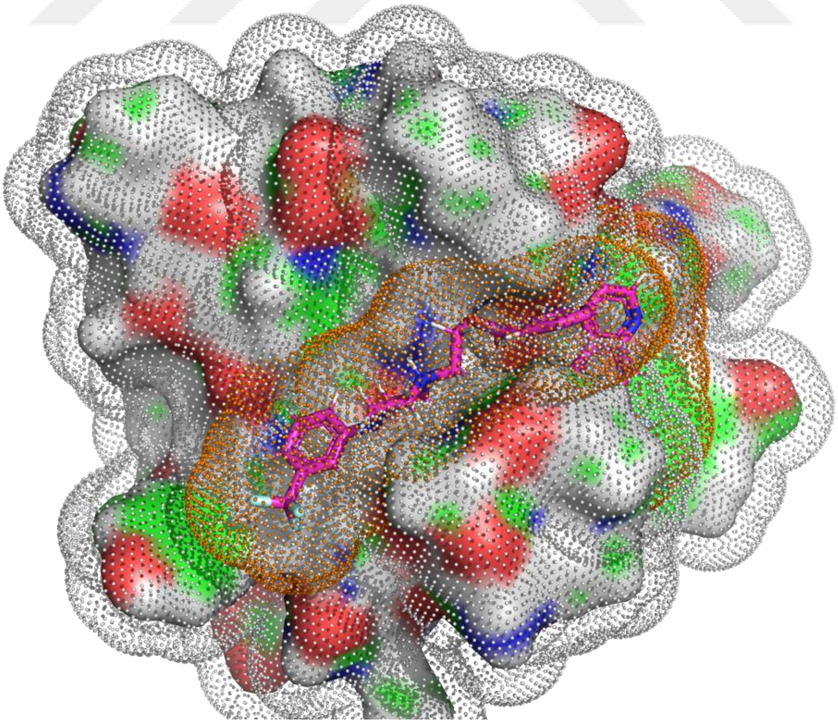

Figure 4.10 SASA depiction of the lead 68723 and its binding pocket on 2SHP

Table 4.5 Cheminformatics feature set

Position	Feature Name
0	SASA Value
1	SASA Class
2	Gasteiger Charge
3	Explicit Valence
4 - 12	Atom Types
13 - 15	Chiral Tags
16 - 19	Hybridization types
20	Degree
21	Aromatic
22	Ring
23	Neighbors
24	Explicit H's
25	Implicit H's
26	Radical Electrons
27	Residue
28	Hetero Atom

4.4.1 Architecture of DeepMLR

Cheminformatics feature matrices of ligand and pocket molecules were both introduced to an embedding layer. Input layers of our model, DeepMLR, have a cutoff dimension representing the total number of atoms in both ligands and pocket molecules. We set 750 for the number of pocket atoms and 175 for the ligand atoms, respectively, to represent their interactions. The output of the input layer was subjected to a transpositional transformation before feeding into the next hidden layers of convolutional nets. We have designed our model to generate new features by adding three successive 1D convolutional layers with a GELU nonlinear activation and batch normalization at each step and a final down-sampling operation via an adaptive max-pooling. Each layer has a convolution filter value of 3 with a stride value of 1 and an increasing number of channels of three consecutive values of 32, 64 and 128. After down-sampling, generated feature maps were subjected to a concatenation operation. We augmented vanilla layers with three fully connected nets to process the newly found features by convolutional operations to predict the last output value of our model. The trainable parameters of the model add up to 136,257 and a typical training of 40 epochs lasted around 15 minutes on a single GPU and 7 cores of CPU accommodated by our HPC.

Table 4.6 Model architecture and the number of parameters of DeepMLR

Layer (type)	Input Shape	# of Param.
Linear-1	[N, 175, 29]	3,840
Conv1d-2	[N, 128, 175]	12,320
BatchNorm1d-3	[N, 32, 173]	64
GELU-4	[N, 32, 173]	0
Conv1d-5	[N, 32, 173]	6,208
BatchNorm1d-6	[N, 64, 171]	128
GELU-7	[N, 64, 171]	0
Conv1d-8	[N, 64, 171]	24,704
BatchNorm1d-9	[N, 128, 169]	256
GELU-10	[N, 128, 169]	0
AdaptiveMaxPool1d-11	[N, 128, 169]	0
Squeeze-12	[N, 128, 1]	0
Linear-13	[N, 750, 29]	3,840
Conv1d-14	[N, 128, 750]	12,320
BatchNorm1d-15	[N, 32, 748]	64
GELU-16	[N, 32, 748]	0
Conv1d-17	[N, 32, 748]	6,208
BatchNorm1d-18	[N, 64, 746]	128
GELU-19	[N, 64, 746]	0
Conv1d-20	[N, 64, 746]	24,704
BatchNorm1d-21	[N, 128, 744]	256
GELU-22	[N, 128, 744]	0
AdaptiveMaxPool1d-23	[N, 128, 744]	0
Squeeze-24	[N, 128, 1]	0
Dropout-25	[N, 256]	0
Linear-26	[N, 256]	32,896
Dropout-27	[N, 128]	0
GELU-28	[N, 128]	0
Linear-29	[N, 128]	8,256
Dropout-30	[N, 64]	0
GELU-31	[N, 64]	0
Linear-32	[N, 64]	65
GELU-33	[N, 1]	0
	Total params:	**136,257**

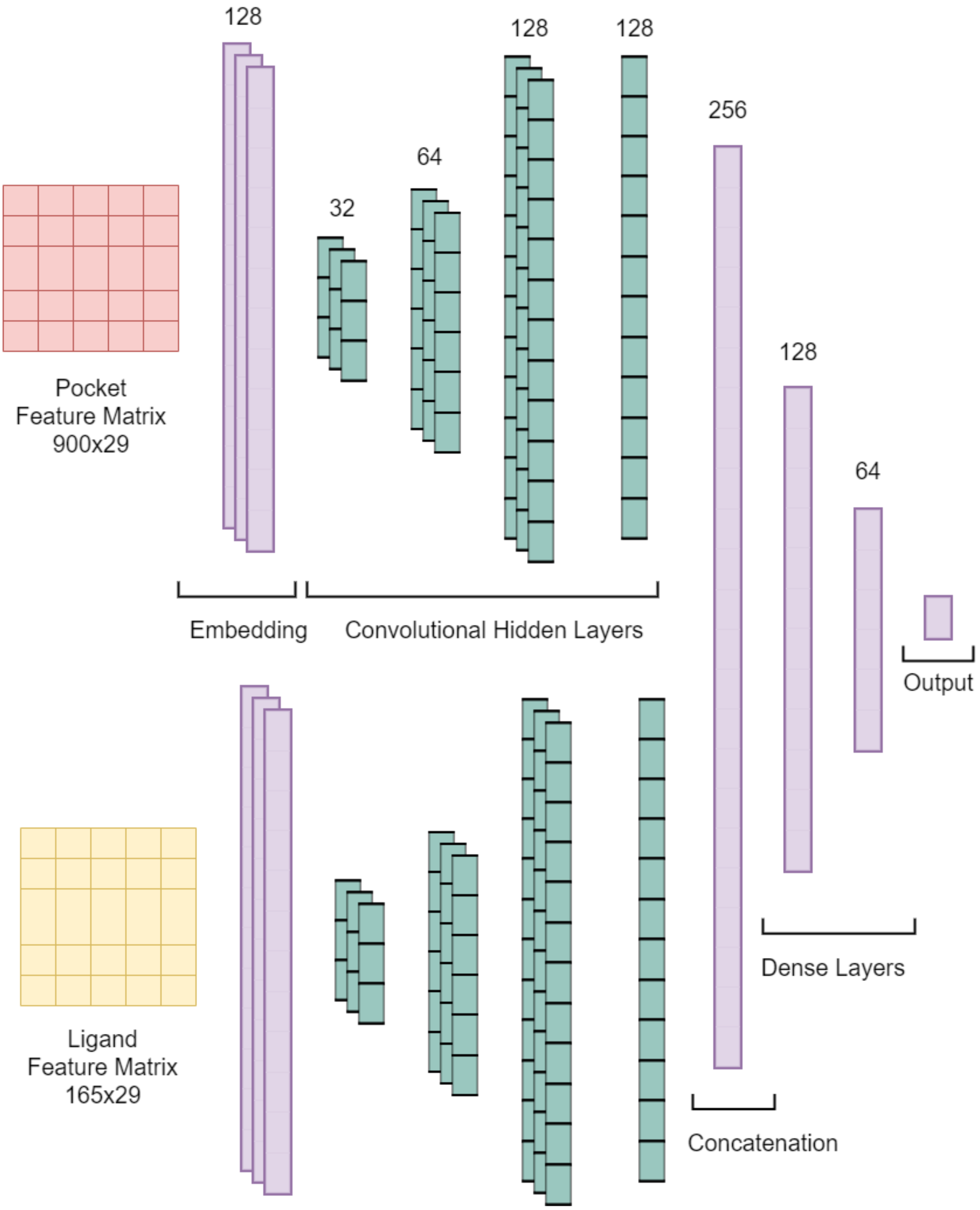

Figure 4.11 Outline of the model architecture

4.4.2 Training and evaluation of DeepMLR

We used the same datasets that were used with the Pafnucy model. Our cheminformatics feature creation process generated a repository of text data files divided into three datasets, each of which contained ligand and pocket features in the form of comma-separated data files of 2D numerical arrays. The DataLoader class of PyTorch

[57] deep learning suite Ver 1.11 was used to collect the data needed for training and testing our model. To achieve the best results, batch sizes of 10 and 20 were selected, as well as different learning rate parameters. The optimization algorithms AdamW, NAdam, and RAdam were applied to improve the performance of our CNN model. The set value of 0.00071 was chosen as the optimal learning rate parameter with the AdamW optimizer.

4.4.3 Evaluation Metrics

Predictions made by our model, DeepMLR, were compared to their experimental affinity values provided with the original datasets, during and after the training to assess the best weight and bias terms of the model parameters. The root mean square error (RMSE) and mean square error (MAE) values were calculated to check the prediction power of our model. We also incorporated the metrics to determine the Pearson correlation (CORR) between the experimental findings and model predictions and their standard deviations (SD). In addition, our model was also tested against the concordance index (CI) to test its ranking power.

4.4.4 Results of experiments run by DeepMLR

Although DeepMLR is equipped with such a basic architecture and relatively shallow structure, it performed better than the Pafnucy and several other state-of-the-art models introduced earlier. Our best model was selected after the 26th epoch, which yielded a Pearson correlation value over 79 percent for both the training and test datasets and similar concordance index values for both datasets. The best model is determined by the criterion of the lowest validation loss calculated at an individual epoch. Table 4.6 lists the resulting metrics obtained with DeepMLR, and figure 4.12 depicts both the correlation and root mean squared error values obtained through 40 epochs in total.

Correlation plots of our model’s predictions and the affinity scores provided by PDBBIND ver.2020 are presented in Fig. 4.13. DeepMLR’s experimental results are set as our new baseline to evaluate our model further and improve the structure and its performance accordingly.

Table 4.7 Model performance metrics of DeepMLR

DATASET	SIZE	LOSS	C_INDEX	RMSE	MAE	SD	CORR
Training Set	13342	1.3005	**0.7973**	1.1404	0.8925	1.1310	**0.7928**
Validation Set	697	1.7282	**0.7729**	1.3146	1.0265	1.3074	**0.7362**
Test Set	225	1.7963	**0.7967**	1.3402	1.0973	1.3038	**0.7944**

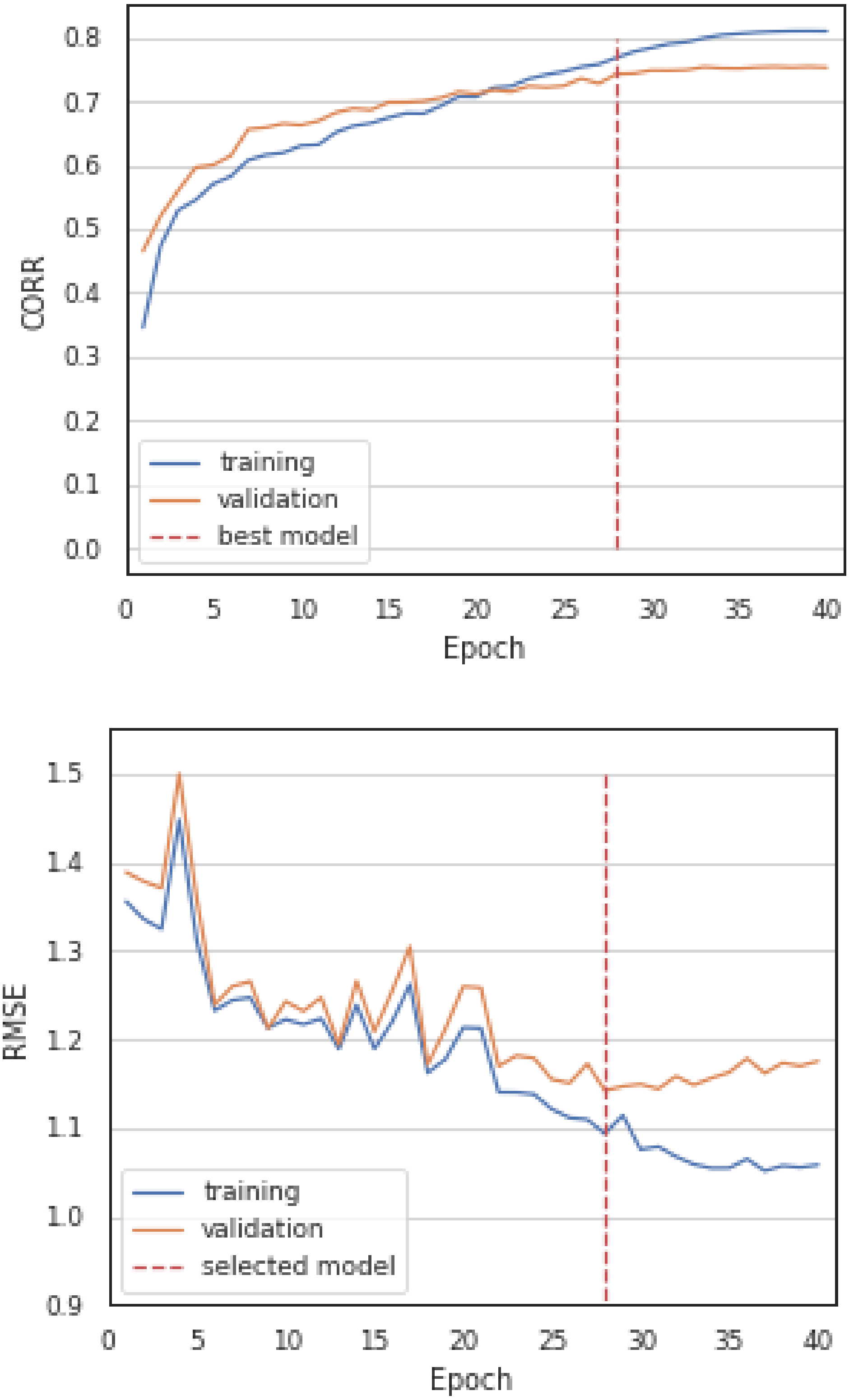

Figure 4.13 CORR and RMSE plots of best performing model of DeepMLR

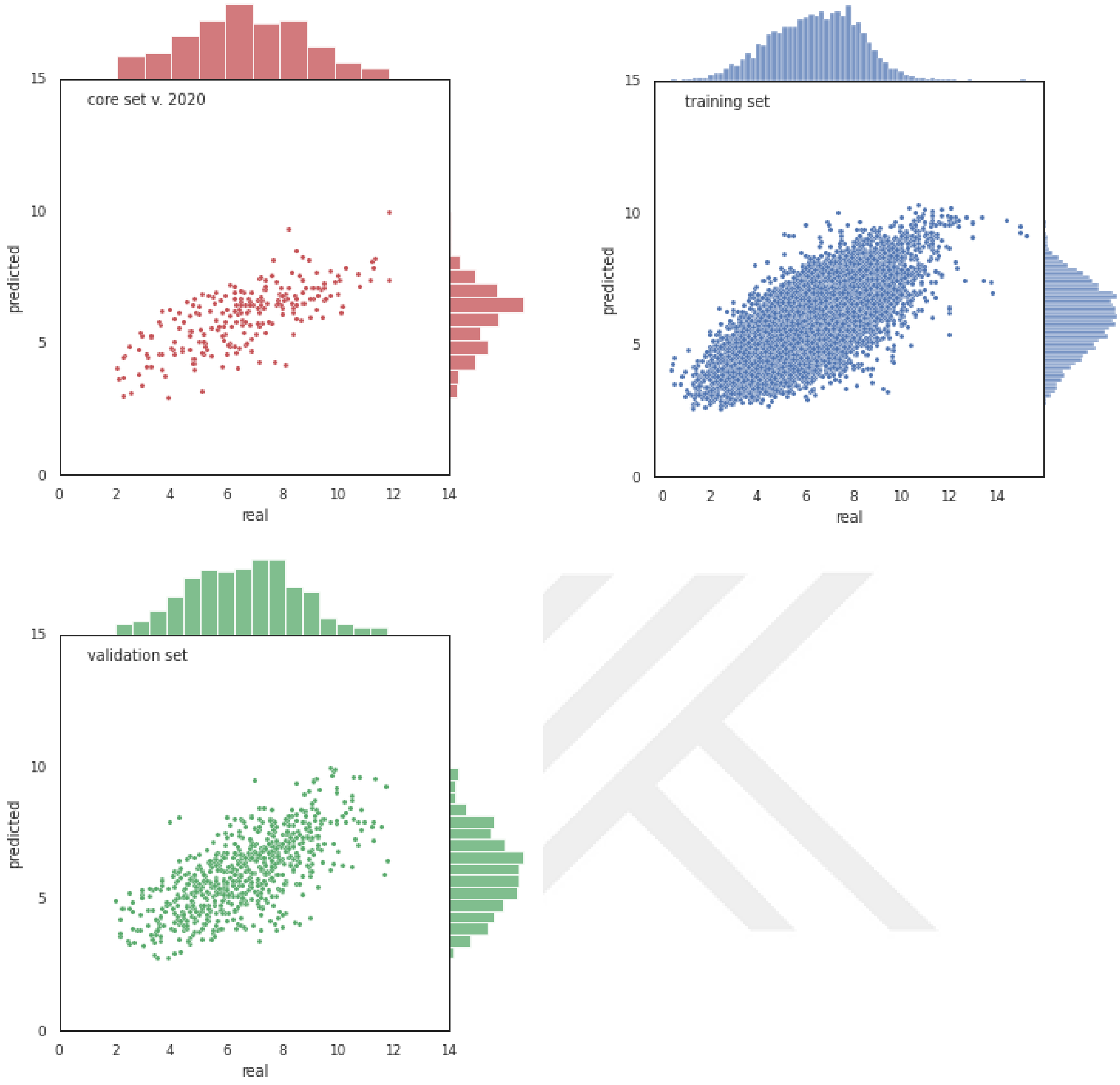

Figure 4.14 Correlations between the predictions made by deepMLR and the experimental data

We have applied two distinctive feature attribution methods to determine the most critical features of interest. Gradient SHapley Additive Predictions (SHAP) [58] is a fast and accurate method to prioritize the features contributing to a better predictive performance. Integrated Gradients is also a computationally efficient method to mark the best features that played a crucial role in learning outcomes. In all scenarios, SASA value is the most important feature by far attribution value. The number of neighboring atoms and the number of implicit Hydrogen atoms was also highly selective in attributing features.

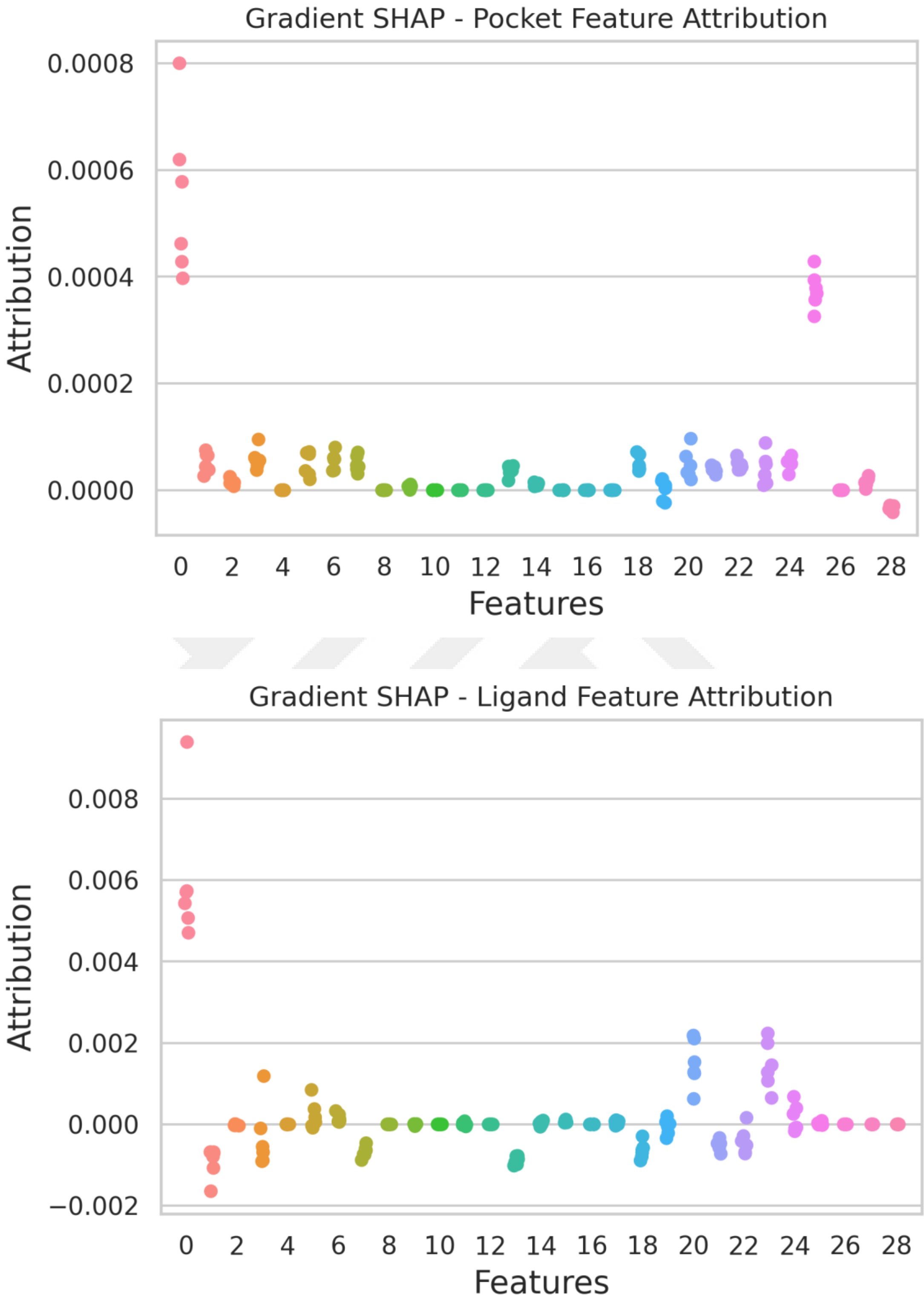

Figure 4.15 Gradient SHAP method Ligand and Pocket feature attributions

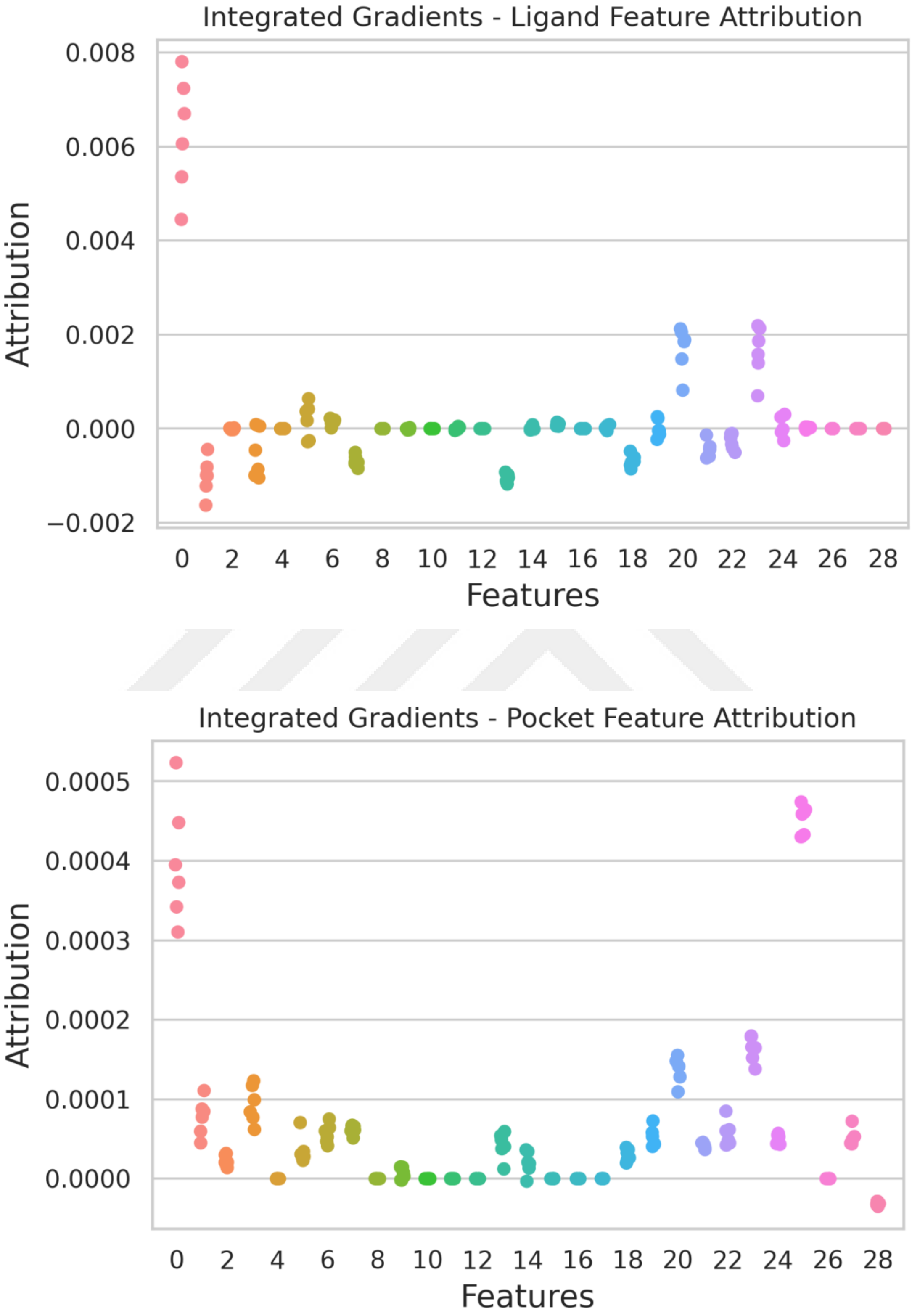

Figure 4.16 Integrated Gradients method Ligand and Pocket feature attributions

Chapter 5

Conclusions and Future Prospects

5.1 Conclusions

The deep learning model we have designed was making use of a critical feature of solvation energy models, namely, the solvent accessible area of each atom in a protein-ligand complex. Solvent accessible area of individual atoms in closer contact will mimic the molecular recognition mechanisms. The deep learning model will be able to learn not only a singular time point of a variety of chemical/physical interactions but also the evolution of a longer duration and can perfectly simulate the possible confirmations. Another advantage is that its value is translationally and rotationally invariant under spatial transformations. We wanted to construct a simpler preparation before conducting the learning experiments. For a more extensive library of chemical space in question, ligands' chemical representations can be easily turned into a matrix used as the input of our deep learning model, deepMLR.

The accuracy results of the training and test dataset's representational learning are promising for further evaluation. Most of the conflicting results that were not captured after VS experiments of both the traditional and other deep learning models were successfully captured by our model. It is mainly due to the representational input data that we have selected in our DL model. Even though our approach was employed only locally, without considering the global picture of the system, the results are closer and even better than most of the other models built with globular descriptors.

5.2 Societal Impact and Contribution to Global Sustainability

In 2015, at a United Nations General Assembly (UN-GA) meeting, seventeen bullet points of future goals were determined to be achieved by 2030 for a brighter and more sustainable world for all its inhabitants. The sustainable development goals (SDGs) are, in total, 17 interconnected issues related to the globe-wide improvement of conditions on our environment and societies at first. Health is ranked and mentioned chiefly as the first in most reports about the progress of worldwide efforts in addressing the SDGs and their achievements. Global health directly connects to the first six items of the list of seventeen goals targeted.

Drugs in our modern world is a societal and an economical mediator that stabilizes the adverse effects of human health problems. The global health industry is dependent heavily on pharmaceutical therapeutic agents, chemical drugs to keep the wealth and well-being of modern humans. Improvements in VS of potential drugs will directly impact the quality and quantity of health conditions of individuals and the ecosystem they live in.

Our DL model aims to identify the molecular recognition of the ligand-receptor complex directly and quickly. Therefore, it will be better in addressing the lack of discovering highly potent drug candidates with lesser side effects. The time and cost of investigating newer and better drugs will have multiple outcomes to increase the number of new drugs and the final price. Access to newer drugs worldwide will be cheaper, and the overall quality of world health will positively impact all six interconnected SDGs. More importantly, the number of critical illnesses that lacks a proper drug due to the cost of investments will be lowered, and there will be more drugs invested even if it will have less use.

Every advancement in the design and application of the AI world will also positively impact the progress of SDGs goals in general. Our model development process can be a small but essential showcase for others to inspire and follow the quest to develop better DL models.

5.3 Future Prospects

The feature of solvent accessible surface area is the main chemical descriptor used in modeling and running of our DL model, deepMLR. We are planning to extend our model with the similar features describing the solvation chemistry. We have trained our model on the datasets fulfilling redocking scenario, but we are also going to include cross-docking experiments into our training setup. Concatenation of different models and applying alternative learning strategies will also be considered. A coding repository on the Github is being developed and our codes will be available to public free of charge. Biologists will be able to accommodate our model for their VS experiments as a primary scoring function or as a complement to validate their traditional molecular docking experiments. We are also interested in dealing with small scale case-studies to test the generalization capacity of our model.

BIBLIOGRAPHY

[1] “What The Shanidar Cave Burials Tell Us About Neanderthals – Wonderful Things Heritage.” https://wonderful-things.org/2014/01/28/what-the-shanidar-cave-burials-tells-us-about-neanderthals/ (accessed Nov. 20, 2021).

[2] G. Scapin, “Structural Biology and Drug Discovery,” 2006.

[3] D. A. Pereira and J. A. Williams, “Origin and evolution of high throughput screening,” *British Journal of Pharmacology*, vol. 152, no. 1, pp. 53–61, Sep. 2007, doi: 10.1038/SJ.BJP.0707373.

[4] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, “Innovation in the pharmaceutical industry: New estimates of R&D costs,” *Journal of Health Economics*, vol. 47, pp. 20–33, May 2016, doi: 10.1016/J.JHEALECO.2016.01.012.

[5] S. S. Ou-Yang, J. Y. Lu, X. Q. Kong, Z. J. Liang, C. Luo, and H. Jiang, “Computational drug discovery,” *Acta Pharmacologica Sinica 2012 33:9*, vol. 33, no. 9, pp. 1131–1140, Aug. 2012, doi: 10.1038/aps.2012.109.

[6] J. Deng, Z. Yang, I. Ojima, D. Samaras, and F. Wang, “Artificial Intelligence in Drug Discovery: Applications and Techniques,” pp. 1–65, 2021, [Online]. Available: http://arxiv.org/abs/2106.05386

[7] A. Bender and I. Cortés-Ciriano, “Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet,” *Drug Discovery Today*, vol. 26, no. 2. Elsevier Ltd, pp. 511–524, Feb. 01, 2021. doi: 10.1016/j.drudis.2020.12.009.

[8] S. Kim, “Expert Opinion on Drug Discovery Getting the Most out of PubChem for Virtual Screening Review: Getting the Most out of PubChem for Virtual Screening,” 2016, doi: 10.1080/17460441.2016.1216967.

[9] Y. Wang *et al.*, “PubChem BioAssay: 2017 update,” *Nucleic Acids Research*, vol. 45, pp. 955–963, 2016, doi: 10.1093/nar/gkw1118.

[10] T. Sterling and J. J. Irwin, “ZINC 15 – Ligand Discovery for Everyone,” *Journal of Chemical Information and Modeling*, vol. 55, no. 11, p. 2324, Nov. 2015, doi: 10.1021/ACS.JCIM.5B00559.

[11] Y. Hu and J. Bajorath, “Compound promiscuity: what can we learn from current data?,” *Drug Discovery Today*, vol. 18, no. 13–14, pp. 644–650, Jul. 2013, doi: 10.1016/J.DRUDIS.2013.03.002.

[12] C. M. Dobson, “Chemical space and biology,” *Nature*, vol. 432, no. 7019, pp. 824–828, Dec. 2004, doi: 10.1038/NATURE03192.

[13] G. Schneider, “Automating drug discovery,” *Nature Reviews Drug Discovery 2017 17:2*, vol. 17, no. 2, pp. 97–113, Dec. 2017, doi: 10.1038/nrd.2017.232.

[14] B. D. Dorsey *et al.*, “L-735,524: The Design of a Potent and Orally Bioavailable HIV Protease Inhibitor,” *Journal of Medicinal Chemistry*, vol. 37, no. 21, pp. 3443–3451, Oct. 1994, doi: 10.1021/JM00047A001/SUPPL_FILE/JM00047A001_SI_001.PDF.

[15] C. Reviews and C. Reviews, “Introduction : Molecular Recognition,” vol. 97, no. 5, 1997.

[16] R. Baron and J. A. McCammon, “Molecular recognition and ligand association,” *Annual Review of Physical Chemistry*, vol. 64, pp. 151–175, 2013, doi: 10.1146/annurev-physchem-040412-110047.

[17] I. Cosic, “Macromolecular Bioactivity: Is It Resonant Interaction Between Macromolecules?—Theory and Applications,” *IEEE Transactions on Biomedical Engineering*, vol. 41, no. 12, pp. 1101–1114, 1994, doi: 10.1109/10.335859.

[18] J. D. Dunitz, “Win some, lose some: enthalpy-entropy compensation in weak intermolecular interactions,” *Chemistry & biology*, vol. 2, no. 11, pp. 709–712, 1995, doi: 10.1016/1074-5521(95)90097-7.

[19] J. Liu and R. Wang, “Classification of current scoring functions,” *Journal of Chemical Information and Modeling*, vol. 55, no. 3, pp. 475–482, 2015, doi: 10.1021/ci500731a.

[20] M. Karplus and D. L. Weaver, “Protein-folding dynamics,” *Nature*, vol. 260, 1976.

[21] H. J. Böhm, “The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure,” *Journal of Computer-Aided Molecular Design 1994 8:3*, vol. 8, no. 3, pp. 243–256, Jun. 1994, doi: 10.1007/BF00126743.

[22] R. A. Friesner *et al.*, “Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy,” 2004, doi: 10.1021/jm0306430.

[23] C. W. Murray, T. R. Auton, and M. D. Eldridge, “Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model,” *Journal of Computer-Aided Molecular Design 1998 12:5*, vol. 12, no. 5, pp. 503–519, 1998, doi: 10.1023/A:1008040323669.

[24] R. Wang, L. Lai, and S. Wang, “Further development and validation of empirical scoring functions for structure-based binding affinity prediction,” *Journal of Computer-Aided Molecular Design 2002 16:1*, vol. 16, no. 1, pp. 11–26, 2002, doi: 10.1023/A:1016357811882.

[25] C. Kramer, T. Kalliokoski, P. Gedeck, and A. Vulpetti, “The Experimental Uncertainty of Heterogeneous Public K i Data,” 2012, doi: 10.1021/jm300131x.

[26] B. A. Grzybowski, A. v. Ishchenko, J. Shimada, and E. I. Shakhnovich, “From Knowledge-Based Potentials to Combinatorial Lead Design in Silico,” *Accounts of Chemical Research*, vol. 35, no. 5, pp. 261–269, 2002, doi: 10.1021/AR970146B.

[27] A. Ben-Naim, “Statistical potentials extracted from protein structures: Are these meaningful potentials?,” *Journal of Chemical Physics*, vol. 107, no. 9, pp. 3698–3706, Sep. 1997, doi: 10.1063/1.474725.

[28] P. J. Ballester and J. B. O. Mitchell, “A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking,” *Bioinformatics*, vol. 26, no. 9, pp. 1169–1175, May 2010, doi: 10.1093/BIOINFORMATICS/BTQ112.

[29] J. D. Durrant and J. A. Mccammon, “NNScore 2.0: A Neural-Network ReceptorÀLigand Scoring Function,” *J. Chem. Inf. Model*, vol. 51, pp. 2897–2903, 2011, doi: 10.1021/ci2003889.

[30] D. Zilian and C. A. Sotriffer, “SFCscore RF : A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein−Ligand Complexes,” *J. Chem. Inf. Model*, vol. 53, 2013, doi: 10.1021/ci400120b.

[31] B. J. Alder and T. E. Wainwright, “Phase transition for a hard sphere system,” *The Journal of Chemical Physics*, vol. 27, no. 5. pp. 1208–1209, 1957. doi: 10.1063/1.1743957.

[32] E. Fermi, P. Pasta, S. Ulam, and M. Tsingou, “STUDIES OF THE NONLINEAR PROBLEMS,” May 1955, doi: 10.2172/4376203.

[33] J. A. Lemkul, “Pairwise-additive and polarizable atomistic force fields for molecular dynamics simulations of proteins,” *Progress in Molecular Biology and Translational Science*, vol. 170, pp. 1–71, Jan. 2020, doi: 10.1016/BS.PMBTS.2019.12.009.

[34] A. v. Onufriev and S. Izadi, “Water models for biomolecular simulations,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 8, no. 2, p. e1347, Mar. 2018, doi: 10.1002/WCMS.1347.

[35] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, “Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids,” 1996, Accessed: Nov. 24, 2021. [Online]. Available: https://pubs.acs.org/sharingguidelines

[36] C. J. Cramer and D. G. Truhlar, “Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics,” *Chemical Reviews*, vol. 99, no. 8, pp. 2161–2200, 1999, doi: 10.1021/CR960149M.

[37] S. Genheden and U. Ryde, “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities,” *Expert Opinion on Drug Discovery*, vol. 10, no. 5, p. 449, May 2015, doi: 10.1517/17460441.2015.1032936.

[38] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo, “Recent developments and applications of the MMPBSA method,” *Frontiers in Molecular Biosciences*, vol. 4, no. JAN, p. 87, Jan. 2018, doi: 10.3389/FMOLB.2017.00087/BIBTEX.

[39] A. M. Turing, “M IND A QUARTERLY REVIEW OF PSYCHOLOGY AND PHILOSOPHY I.-COMPUTING MACHINERY AND INTELLIGENCE,” 1950.

[40] B. G. Buchanan, “A (Very) Brief History of Artificial Intelligence,” 2005.

[41] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics 1943 5:4*, vol. 5, no. 4, pp. 115–133, Dec. 1943, doi: 10.1007/BF02478259.

[42] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the importance of initialization and momentum in deep learning,” 2013.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature 1986 323:6088*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.

[44] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *The Journal of Physiology*, vol. 148, no. 3, pp. 574–591, Oct. 1959, doi: 10.1113/JPHYSIOL.1959.SP006308.

[45] Y. LeCun *et al.*, “Backpropagation Applied to Handwritten Zip Code Recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: 10.1162/neco.1989.1.4.541.

[46] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review 2020 53:8*, vol. 53, no. 8, pp. 5455–5516, Apr. 2020, doi: 10.1007/S10462-020-09825-6.

[47] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, “Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise,” *Journal of Chemical Information and Modeling*, vol. 53, no. 8, pp. 1893–1904, Aug. 2013, doi: 10.1021/CI300604Z/SUPPL_FILE/CI300604Z_SI_002.PDF.

[48] O. Trott and A. J. Olson, “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.,” *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, Jan. 2010, doi: 10.1002/JCC.21334.

[49] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, “GROMACS: Fast, flexible, and free,” *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/JCC.20291.

[50] M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen, “Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field,” *Journal of Chemical Theory and Computation*, vol. 11, no. 7, pp. 3499–3509, Jun. 2015, doi: 10.1021/ACS.JCTC.5B00356/SUPPL_FILE/CT5B00356_SI_004.XLSX.

[51] W. L. Jorgensen and J. Tirado-Rives, “Potential energy functions for atomic-level simulations of water and organic and biomolecular systems,” 2005, Accessed: Feb. 17, 2022. [Online]. Available: www.pnas.orgcgidoi10.1073pnas.0408037102

[52] R. Kumari, R. Kumar, and A. Lynn, “G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations,” *Journal of Chemical Information and Modeling*, vol. 54, no. 7, pp. 1951–1962, Jul. 2014, doi: 10.1021/CI500020M/SUPPL_FILE/CI500020M_SI_001.PDF.

[53] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction,” *Bioinformatics*, vol. 34, no. 21, p. 3666, Nov. 2018, doi: 10.1093/BIOINFORMATICS/BTY374.

[54] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, and P. Siedlecki, “Development and evaluation of a deep learning model for protein–ligand binding affinity prediction,” *Bioinformatics*, vol. 34, no. 21, pp. 3666–3674, Nov. 2018, doi: 10.1093/bioinformatics/bty374.

[55] M. Su *et al.*, “Comparative Assessment of Scoring Functions: The CASF-2016 Update,” *Journal of Chemical Information and Modeling*, vol. 59, no. 2, pp. 895–913, Feb. 2019, doi: 10.1021/ACS.JCIM.8B00545/SUPPL_FILE/CI8B00545_SI_001.PDF.

[56] “RDKit.” http://www.rdkit.org/ (accessed Feb. 17, 2022).

[57] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” 2019.

[58] S. M. Lundberg, P. G. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017, Accessed: Feb. 08, 2022. [Online]. Available: https://github.com/slundberg/shap

CURRICULUM VITAE

1991 – 1997	B.Sc., Department of Physics, Middle East Technical University, Ankara, TURKEY
2007 - 2015	Bioinformatics Specialist, University of Wisconsin-Madison, Madison, Wisconsin, USA
2018 – Present	Lecturer, Department of Molecular Biology and Genetics, Abdullah Gül University, Kayseri, TURKEY

SELECTED PUBLICATIONS AND PRESENTATIONS

J1) Hitit Mustafa, Özbek Mehmet, Ayaz Güner Şerife, **Güner Hüseyin**, Öztuğ Merve, Bodu Mustafa, Kırbaş Mesut, Bülbül Bülent, Bucak Mustafa Numan, Ataman Mehmet Bozkurt, Memili Erdoğan, Kaya Abdullah (2021). Proteomic fertility markers in ram sperm. *Animal Reproduction Science,* 235, Doi: 10.1016/j.anireprosci.2021.106882

J2) Acar Mustafa B., Aprile Domenico, Ayaz Güner Şerife, **Güner Hüseyin**, Tez Coşkun, Di Bernardo Giovanni, Peluso Gianfranco, Özcan Servet, Galderisi Umberto (2021). Why Do Muse Stem Cells Present an Enduring Stress Capacity? Hints from a Comparative Proteome Analysis. *International Journal of Molecular Sciences,* 22(2064), 1-20., Doi: 10.3390/ijms2

J3) Acar Mustafa Burak, Aprile Domenico, Ayaz Güner Şerife, **Güner Hüseyin**, Tez Coşkun, Bernardo Giovanni Di, Peluso Gianfranco, Özcan Servet, Galderisi Umberto (2021). Why Do Muse Stem Cells Present an Enduring Stress Capacity? Hints from a Comparative Proteome Analysis. *International Journal of Molecular Sciences*, 22, 1-22., Doi: 10.3390/

J4) Güner Hüseyin, Cai Wenxuan, Gregorich Zachery R.,Chen Albert J.,Ayaz-Guner Serife,Peng Ying,Valeja Santosh G.,Liu Xiaowen,Ge Ying (2016). MASH Suite Pro: A

Comprehensive Software Tool for Top-Down Proteomics. *Molecular Cellular Proteomics*, 15(2), 703-714., Doi: 10.1074/mcp.O115.054387

J5) Gregorich Zachery R.,Peng Ying,Lane Nicole M.,Wolff Jeremy J.,Wang Sijian,Guo Wei, **Güner Hüseyin**, Doop Justin,Hacker Timothy A.,Ge Ying (2015). Comprehensive Assessment Of Chamber-Specific And Transmural Heterogeneity İn Myofilament Protein Phosphorylation By Top-Down Mass Spectrometry. *Journal Of Molecular And Cellular Cardiology*, 87(null), 102-112., Doi: 10.1016/j.yjmcc.2015.08.007

J6) Chang Ying-Hua,Ye Lei,Cai Wenxuan,Lee Yoonkyu, **Güner Hüseyin**, Lee Youngsook, Kamp Timothy J., Jianyi Zhang,Ge Ying (2015). Quantitative proteomics reveals differential regulation of protein expression in recipient myocardium after trilineage cardiovascular cell transplantation. *Proteomics*, 15(15), 2560-2567., Doi: 10.1002/pmic.201500131

J7) Chang Ying-Hua, Gregorich Zachery R., Chen Albert J., Hwang Leekyoung, **Güner Hüseyin**, Yu Deyang,Jianyi Zhang,Ge Ying (2015). New Mass-Spectrometry-Compatible Degradable Surfactant for Tissue Proteomics. *Journal Of Proteome Research*, 14(3), 1587-1599., Doi: 10.1021/pr5012679

J9) Valeja Santosh G., Xiu Lichen,Gregorich Zachery R., **Güner Hüseyin,** Jin Song,Ge Ying (2015). Three Dimensional Liquid Chromatography Coupling Ion Exchange Chromatography/Hydrophobic Interaction Chromatography/Reverse Phase Chromatography For Effective Protein Separation İn Top-Down Proteomics. *Analytical Chemistry*, 87(10), 5363-5371., Doi: 10.1021/acs.analchem.5b00657

J10) Güner Hüseyin, Patrick L. Close,Cai Wenxuan,Zhang Han,Peng Ying,Gregorich Zachery R.,Ge Ying (2014). MASH Suite: A User-Friendly And Versatile Software Interface For High-Resolution Mass Spectrometry Data Interpretation And Visualization. *Journal Of The American Society For Mass Spectrometry*, 25(3), 464-470., Doi: 10.1007/s13361-013-0789-4

J11) Peng Ying, Gregorich Zachery R.,Valeja Santosh G.,Zhang Han,Cai Wenxuan,Chen Yi-Chen, **Güner Hüseyin,** Chen Albert J.,Schwahn Denise J.,Hacker Timothy A.,Liu

Xiaowen,Ge Ying (2014). Top-Down Proteomics Reveals Concerted Reductions İn Myofilament And Z-Disc Protein Phosphorylation After Acute Myocardial Infarction. *Molecular Cellular Proteomics*, 13(10), 2752-2764., Doi: 10.1074/mcp.M114.040675

J12) Xu Fangmin,Xu Qingge,Dong Xintong,Guy Moltu J., **Güner Hüseyin**, Hacker Timothy A.,Ge Ying (2011). Top-Down High-Resolution Electron Capture Dissociation Mass Spectrometry For Comprehensive Characterization Of Post-Translational Modifications İn Rhesus Monkey Cardiac Troponin I. *International Journal Of Mass Spectrometry*, 305(2-3), 95-102., Doi: 10.1016/j.ijms.2010.09.007

J13) Zhang Jiang,Guy Moltu J.,Norman Holly S.,Chen Yi-Chen,Xu Qingge,Dong Xintong, **Güner Hüseyin**, Wang Sijian,Kohmoto Takushi,Young Ken H.,Moss Richard L.,Ge Ying (2011). Top-Down Quantitative Proteomics Identified Phosphorylation Of Cardiac Troponin I As A Candidate Biomarker For Chronic Heart Failure. *Journal Of Proteome Research,* 10(9), 4054-4065., Doi: 10.1021/pr2002S8m

C1) Kaplan Oktay İsmail, Torun Muhammet Furkan, **Güner Hüseyin**, Çevik Kaplan Sebiha (2019). Conserved Clinical Variation Visualization Tool (Convart*). European Journal Of Human Genetics* (27), 1697-1698.

C2) Acar Mustafa Burak, Alessio Nicola, Ayaz Güner Şerife, **Güner Hüseyin**, Karakükcü Musa, Galderisi Umberto, Özcan Servet (2020). Proteomic Profile of Metformin Treated Senescent Mesenchymal Stem Cell Secretome. *US HUPO 16th Annual Conference 2020*