2019-01-01

Development Of Ligand-Kinase Binding Affinity Predictive Model

Govinda Kc
University of Texas at El Paso, kcgovindap@gmail.com

DEVELOPMENT OF LIGAND-KINASE BINDING AFFINITY PREDICTIVE MODEL

GOVINDA BAHADUR KC

Master's Program in Computational Science

APPROVED:

Suman Sirimulla, Ph.D., Chair

Mahesh Narayan, Ph.D.

Amy Wagler, Ph.D.

Xianyi Zeng, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

to my Parents

with love

DEVELOPMENT OF LIGAND-KINASE BINDING AFFINITY PREDICTIVE MODEL

by

GOVINDA BAHADUR KC

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Master's Program in Computational Science

THE UNIVERSITY OF TEXAS AT EL PASO

May 2019

Acknowledgements

Whether it is a doctoral dissertation or a masters thesis, any research work involves efforts from multiple members of the research community. Even though I am the author of this thesis, it would not have been possible for me to accomplish the work without the support from different scholars from my department and my family members.

First and foremost, my heartfelt appreciation goes to my dissertation advisor Prof. Dr. Suman Sirimulla who walked me through the research process from the day I began to formulate research questions to the day I concluded my research. Thank you very much Dr. Sirimulla for constantly encouraging and supporting me.

My immense appreciation also goes to Mr. Md Mahmudulla Hassan and all the lab members who helped me in different stages of my research work.

Finally, I would like to thank my wife Bibechana Basnet who devotedly supported me during the process of research work. Thank you very much Bibechana for being so supportive and patient even when I spent countless hours working on thesis both at home and at school.

Abstract

Prediction of interaction between drugs or drug like compounds and targets, is of high importance in drug discovery process as it provides important insights into therapeutic potential and possible adverse effects. As the experimental testing would be highly expensive, laborious and time consuming, screening the molecules computationally before performing experiments would be cost effective, faster and convenient as a method of approach. In this study, I have developed computational models, leveraging machine learning techniques, to predict drug-kinase binding affinities. The predictive model is mainly built using the Random Forest (RF) machine learning method. This study is focused on kinases because of their importance as drug targets for therapeutic use. The dataset encompasses the kinases and ligands binding information collected from Drug Target Commons (DTC) and Pharos. The data was split into a training set (75%) and a test set (25%). The performance of the model was evaluated using several metrics and the best model achieved a correlation coefficient (R) of 0.86, root mean square error (RMSE) of 0.52, concordance index (CI) of 0.81, and Area Under a receiver operating characteristic Curve (AUC) of 0.95 during the internal 10-fold cross validation. An additional blind test was also performed on synapse IDG-DREAM Challenge, which is a Drug-Kinase Binding Prediction Challenge and RF model achieved AUC of 0.68. I demonstrated that RF model has the potential to predict the binding affinity for the interaction of ligand and kinase based on structural, physicochemical and atom pair based two-dimensional pharmacophore fingerprints. I also compared the results based on grid search and random search methods. I observed that there was no significant difference in model performance. However, random search reduced the model building time significantly. I aim to build the better model based on other machine learning approaches with more data in the future.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Discovering a new drug is a long, expensive, and often haphazard process. The approaches and the methodologies used in the drug discovery process have advanced over time. Out of large number [10^{60} – 10^{100}] of synthetically feasible molecules [40], we have to find a compound satisfying the list of cirteria such as drug metabolism, bioactivity, synthetic accessibility and pharmacokinetic profile and many more. According to the most recent analysis by the Tufts Center for the Study of Drug Development (CSDD) [29], the average cost to develop and approve a new drug is approximately 2.6 billion dollars. The Pharmaceutical Research and Manufacturers of America (PhRMA) claims that the overall process for the drug to go into the market takes at least 10 years, of which around six to seven of those years are spent undergoing clinical trials and usually less than 12 percent of drugs that enter clinical trials get approved [3]. Thus, the modern Artificial Intelligence, especially machine learning (ML) approach plays a crucial role in the drug discovery process. It is a new trending field which uses certain statistical algorithms with the help of computers but without being explicitly programmed. If the success rate of drugs using ML approach can just be raised by a very small amount, that would save billions of dollars.

Finding a compound that binds to a particular target is a one of the most challenging parts of the drug discovery process. One of the recent studies by Ding et al. has reviewed the importance of machine learning approach on predicting the drug target interactions on a large scale using the information about the compounds and proteins. Proteins are the good targets in drug design [5] and get activated or inhibited by drug compounds. The interaction between drugs and targets facilitates the drug side effect prediction [34], drug

repurposing [27] and many others. Biochemical experiment methods for drug target interactions are found highly costly and take a lot of time [45], where as computational methods are efficient, faster and more convenient [15]. Docking and machine learning approaches are two widely used methods to find the pharmacological profiles of drugs under development; however, docking needs the high resolution X-ray crystal structures of proteins [15]. In contrast, ML approach does not need the protein structures, thus it is considered as an alternative choice. It relies on the chemical structures of potential ligands that are responsible for the binding. There are mainly three types of interactions studied in computer aided drug discovery, namely compound (i.e., ligand) based, target based, and the most recent system based frameworks. Quantitative structure-activity relationship (QSAR) is a widely used ligand based technique that uses the statistics and analytical tools to analyse the relationship between the structures of ligands and their corresponding effects [16]. Similarly, the target based method is solely based on the target information such as protein sequence information for drug target prediction [20]. Another method is the system based frameworks based on information of both compounds and targets. A recent study by Cichonska et al. [9] implemented the regression ML model for prediction and verification of compound target bioactivity profiles. In the present study, I am focused on kinase target family as they are used extensively to transmit signals and regulate complex processes in cells. Dysfunction of protein kinases cause many diseases such as cancer and inflammatory diseases. I have developed the prediction model for binding affinity prediction based on the interaction profiles of kinases and their ligands. The prediction binding affinity values of drug target complexes play a crucial role in the drug discovery process. It is usually measured by dissociation constant (K_d), inhibition constant (K_i) and half maximal concentration (IC_{50}). Binding affinity values are continuous values, however, most of the studies about it are in binary classification problems. Some of the recent studies have done crucial work developing the regression models to predict the binary affinity values. Pahikkala et al [33] employed the algorithm named as KronRLS and later He et al. [14] proposed the SimBoost to predict the binding affinity values but both use the similarity information

obtained from 2D representations of compounds and used the traditional machine learning method. Another recent work by Indra et al. [18] has implemented the regression models using the Weka [13] tool to predict the protein ligand binding affinities and reported the determination coefficient (R^2) and RMSE as 0.76 and 1.31 respectively but rmse values still high. In our study, I propose the Random forest regressor which is trained on an extensively large data set obtained from the IDG-DREAM challenge [1] which is originally from Drug Target Commons (DTC) [43]. I combined it with the data set obtained from Pharos [31] . The final data set I have used contains the 97564 instances with 196 protein kinases interacting with 6792 compounds. I have also evaluated our model on blind 17258 drug-kinase pairs that I filtered from the Metz data set [26].

1.1 Protein Kinases

Protein kinases are one of the largest families of genes in eukaryotes and covers around 2% of the genome [23]. They have been intensively investigated as drug targets. Phosphorylase kinase was the first protein kinase characterized biochemically which later led to discovery of cAMP-dependent protein kinase which catalyzes the phosporylation and activation of phosphorylase kinase [38]. Protein kinases modify the function of other proteins by attaching phosphate groups to them. They are key controllers of most biochemical pathways and important in health and disease. By adding the phosphate groups to substrate proteins, they direct the activity, localization and overall functions of proteins. They are particularly prominent in signal transduction and coordination of complex functions such as cell cycle. Out of 518 human protein kinases, 478 belong to a single superfamily whose catalytic domains are related in sequence, which can be clustered into groups, families and sub-families, of increasing sequence similarity and biochemical domains [10].

1.2 Classification of Kinases

Protein kinase is mainly comprised of AGC, CAMK, CK1, CMGC, MAPK, STE, TK, TKL [12].

AGC: The AGC group of serine/threonine kinases is named after their homology to protein kinases A, G and C. It contains 60 members, including PKA, PKG and PKC. The family comprises some intensely examined protein kinases (such as Akt, S6K, RSK, MSK, PDK1 and GRK) as well as many less well-studied enzymes (such as SGK, NDR, LATS, CRIK, SGK494, PRKX, PRKY and MAST)[35].

CAMK: CAMK stands for Calcium/Calmodulin-dependent kinase of enzymes. CAM kinases are divided into two groups: substrate restricted and substrate multifunctional, based on substrate specificity. Their activation are modulated by increase in the concentration of intracellular calcium ions (Ca2+) and activated CAMK is involved in the phosphorylation of transcription factors. The concentration of free intracellular Ca2+ that range between basally 50 nM to stimulated levels around 1-10 μM depending up on the cell type [44].

CK1: CK1 is an abbreviation of Casein Kinase 1. It is a small group of kinases that are very distinct from other kinase groups but are very similar to each other in sequence. These kinases are serine/threonine kinases with a preference for acidic substrates.

CMGC: The group name CMGC comes from the CDK family (Cyclin-Dependent Kinase), the MAPK family (Mitogen-Activated Protein Kinase), the GSK family (Glycogen Synthase Kinase), the DYRK family (Dual specificity Tyrosine Regulated Kinase), and the dual specificity CLK family (CDC2-Like Kinase).

MAPK: The name MAPK is an abbreviation of a mitogen-activated protein kinase that get activated within the protein kinase cascades specific to the amino acids serine and threonine and involved in a variety of fundamental cellular processes such as differentiation, proliferation, stress response, apoptosis, motility, and survival. Each cascade is initiated by specific extracellular cues and leads to activation of a particular MAPK following the successive activation of a MAPK kinase (MAPKKK) and a MAPK kinase (MAPKK)[28].

TK: TK stands for Tyrosine Kinases. These are major signaling enzymes in the process of cell signal transduction, which catalyzes the transfer of ATP-gammaphosphate to the tyrosine residues of the substrate protein. They are involved in several steps of neoplastic development and progression. Their signaling pathways are often genetically or epigenetically altered in cancer cells to impart a selection of advantage to the cancer cells.
TKL: TKL stands for tyrosine kinase like and they are named so because of their close sequence similarity to tyrosine kinases. These are diverse groups of families that resemble both serine/threonine and tyrosine kinases, and function as dual specificity kinases.
STE: These are homologs of the yeast STE7, STE11, and STE20, which sequentially activate the MAPK family. The p21-activated kinases (Paks) are the prominent members of these families, which are critical regulators of diverse signaling pathways. Their alterations in Pak expression in human tumors makes them as an attractive new therapeutic agent [12].

1.3 Structure of Protein Kinases

The structure of protein kinase was first observed by X-ray crystallography in 1971 [17]. A typical protein kinase core consists of two lobes (or subdomains): the larger carboxy terminal lobe (green) and the smaller amino terminal lobe (red). They are commonly known as C-terminal and N-terminal. N terminal contains a series of β and one important α helix, where as C terminal mainly consists of α helix. The lobes are connected by a short polypeptide chain, which is known as the linker, or the hinge region as shown in figure(1.1).

1.4 Importantance of Protein Kinases

Protein kinases are important for a wide range of the cellular processes, mostly involved in the signal transduction. They transfer a phosphate group from a molecule of adenosine triphosphaste (ATP) onto a protein or other substrate which is known as phosphorylation.

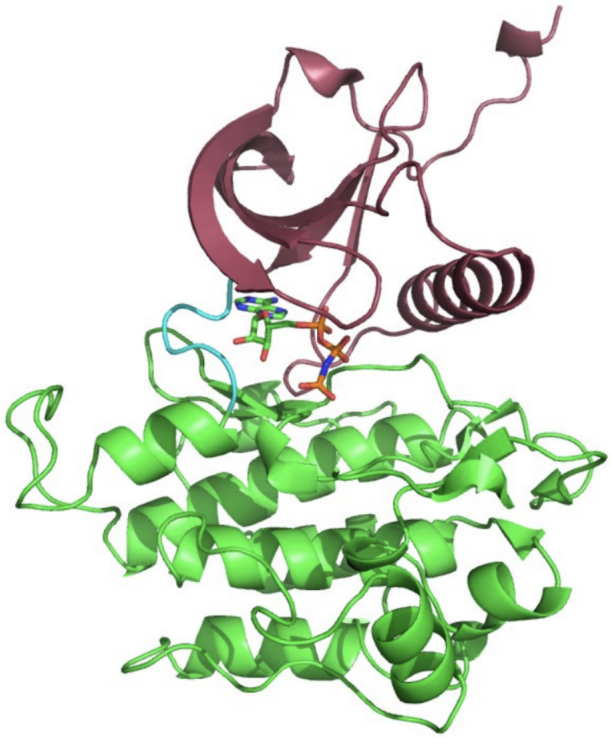


Figure 1.1: A typical protein kinase (Adopted from PDB 1IR3)

Phosphorylation plays an important role in various cellular processes such as cell division, metabolism, survival and apoptosis. Deregulation of protein kinases cause many diseases such as chronic myelogenous leukaemia, gastrointestinal stromal tumours and cancers as well as non-malignant disorders. Therefore, kinases are attractive targets for both biological research and drug development. The protein kinases undergo the following enzymatic reaction.

$$\mathrm{MgATP}^{1-} + \mathrm{protein-O:H} \longrightarrow \mathrm{protein-O:PO}_3^{2-} + \mathrm{MgADP} + \mathrm{H}^+ \qquad (1.1)$$

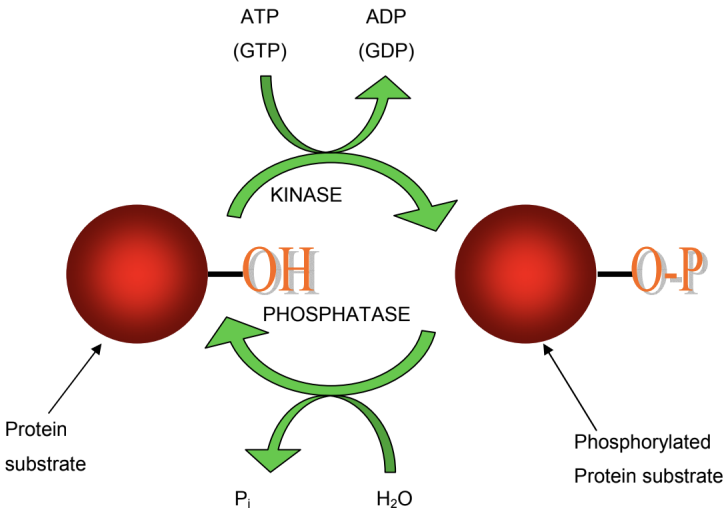


Figure 1.2: Enzymatic activity of a kinase (from, Tesi Di Dottorato, Padua@research)

1.5 Kinase Inhibitor

A protein kinase inhibitor is a type of enzyme inhibitor that blocks the action of one or more protein kinases. They are important research tools to study cell functions and human diseases. Targeted therapies with small molecule kinase inhibitors (KIs) are one of the cornerstones in the treatment of many cancers. Most of the kinase inhibitors are ATP-competitive and are called type I inhibitors because they compete with the nucleotide phosphodonor substrate in the catalytic site of the enzyme. There are 48 small molecule protein kinase inhibitors approved by United States Federal Drug Adminstration (US FDA) and nearly all of which are orally effective except netarsudil [39].

Dar and Shokat have defined three classes of small molecule protein kinase inhibitor, which are labeled as types I, II, and III [11]. Type I inhibitor is defined as a small molecule that binds to the active conformation of a kinase in the ATP pocket. Similarly, the type II inhibitor is defined as a small molecule that binds to an inactive conformation of a kinase, and the type III inhibitor is defined as a non-ATP competitive inhibitor or allosteric inhibitor.

Table 1.1: Names of some diseases and involved kinases

Diseases	**Kinases**	**Diseases**	**Kinases**
Cancer	Aurora, CK1d, CK2, RTK, NRTK	Essential hypertension	ERK, P38
Parkinson	JNK, DYRK1a	Pulmonary hypertension	ALK1
Inflamation	ERK, P38, JNK	Insulin dependent diabeties	GSK3, PKC
Cardiovascular disease	ERK	Insulin independent diabeties	GSK3, PKC
Down's syndrome	DYRK1a	Alzheimer	GRSK3, PKC, ERK, CDK5, CK1d
Craniosynostosis	FGFR	Infectious and parasitical disease	CK2, CK1
Obesity	PKA	Schizophrenia and depression	CAMKII

1.6 Enzyme Kinetics and Inhibition

Enzymes are protein catalysts which increase the rate of reaction without being consumed in the process. The enzyme kinetics refers to the study of rate at which an enzyme works. The study of enzyme kinetics help us to understand the catalytic mechanism of an enzyme, its role in metabolism and also the drugs and poisons inhibition in its activity.

1.6.0.1 Substrates and Inhibitors

A substrate is a compound, for example, a drug, which is metabolized by an enzyme where as an enzyme inhibitor is a molecule or compound that binds to an enzyme and binds its activity and also may decrease the metabolism of substrates. Although enzymes are

absolutely essential for life, abnormally high enzyme activity can lead to disease conditions. So, manipulation of enzyme catalysis with inhibitors is critical for prevention of infectious diseases, cell growth, control of inflammatory response and more. The dissociation constant

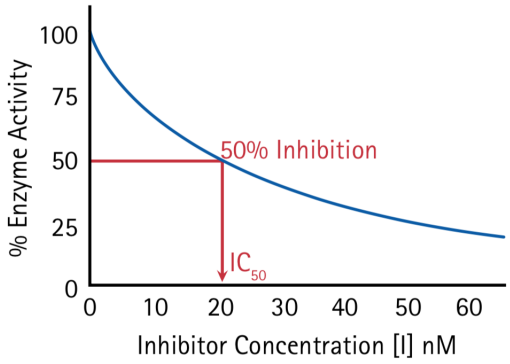


Figure 1.3: Relation between Inhibitor concentration and IC_{50} (from, Chandra Mohan, 2014)

K_d measures the binding affinity of the substrate for enzymes. The smaller the value of K_d the greater the biniding affinity is. Similarly, K_i is the inhibition constant of the enzyme-inhibitor complex or the reciprocal of the binding affinity of the inhibitor to the enzyme and the smaller the value of K_i, the smaller the amount of the medication needed in order to inhibit activity of the enzyme. Another parameter IC_{50} is the concentration of an inhibitor relative to substrate concentration producing the 50 percent inhibition to enzyme, that is, binding is reduced by 50 percent. In enzymatic inhibiton assay, IC_{50}, the relation between concentration at which the inhibitor causes a 50% enzymatic activity and inhibitor constant (K_i) can be expressed using the Cheng Prausoff model [8]

$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_m}} \tag{1.2}$$

where [S] is the experimental substrate concentration and K_m is the concentration of the substrate at which the enzyme activity is half minimal.

Chapter 2

Theoretical Background

2.1 Machine Learning

Machine Learning (ML) is a branch of artificial intelligence based on the idea that a system learns from the data, looks for the patterns and makes decisions with minimal human intervention. They are not being explicitly programmed by the people. Machine learning is widely used now a days because of its ability to apply complex mathematical calculations to big data automatically, quickly and iteratively. It has tremendously reduced the time and expenses of the people in the modern world. It is so fascinating and effective in it's performance. There are many applications of it from virtual assistance to video surveillance, customer support, social media services and much more. Some of the great examples of machine learning are the self-driving Google car, recommendations on the side of web pages, friend suggestions on Facebook, cyber fraud detection, etc. It is also becoming a powerful and flexible tool in the field of medicine to analyze and predict the outcomes from the biological and clinical data. There are two widely used machine learning methods: Supervised learning and unsupervised learning.

2.1.1 Unsupervised Machine Learning Method

In this method, data has not any historical labels. So, the system is not told the output and the algorithm must figure out what is going on in the data. The goal of this method is to find the structure within the data. For example, for the data that contains the dogs and cats images which are never seen, the machine has no idea about the features of dogs, but

it can categorize into two clusters where the first may contain all the pictures of dogs and the second may contain all the pictures of cats. Unsupervised learning method is further categorized into algorithms, namely Clustering and Associating.

2.1.2 Supervised Machine Learning Method

In this method, algorithms are trained using the known input and output values. For example, suppose when size, color and shape of a particular fruit is given. If the size is big, color is red, and shape is rounded with a depression at the top, one can confirm that the fruit is an apple. So, the structure of the data is already known, and the goal is to assign the new data to the correct classes. For the input variables x and output variable Y, the mapping function from the input to the output is as follows.

$$Y = f(x) \tag{2.1}$$

Supervised learning is further grouped into classification, regression, and forecasting.

Classification: If the data are being used to predict the categorical variables such as 'Yes' or 'No', '1' or '0', then supervised learning is also called classification. When the output variables consist s of only two labels, this is called binary classification whereas the output variable with more than two labels, is known as the multi-class classification. The prediction of site of metabolism for FMO is an example of classification problem.

Regression: In this problem, the predicting values are continuous, for example, in between 0 to 1. A continuous output variable is a real-value. They are either an integer or floating point value. I have used the regression method for prediction of the kinases' inhibitors' activity and drug metabolic properties.

Forecasting: In this method, predictions about the future are made based on the past and the present data. It is mostly used to analyze the trends. For example, forecasting of rainfall based on the current and previous year's rainfall.

Following are the some of the examples of machine learning.

(I) Support Vector Machine:

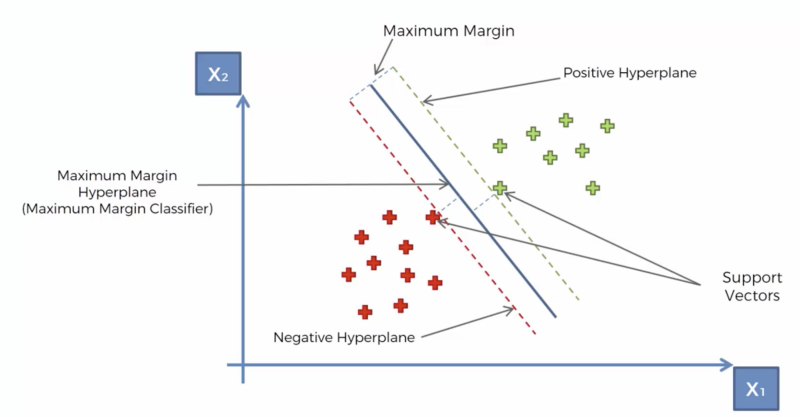


Figure 2.1: Support vector machine (from, https://udemy.com)

Support vector machine (SVM) is the supervised machine learning that is based on the concept of a decision plane that defines decision boundaries. A decision plane is the plane that separates the set of objects have into different class memberships. SVM can be used for both classification or regressions problems. For example, we have the data points in two-dimensional space as shown below. For the new data in the future, we dont know where they will fall either in the green area or in the red area. So, we have to separate those data points with the help of a decision plane. This plane can be drawn in many ways such as vertically or horizontally or diagonally as shown in Figure 2.1. However, we need to find the optimal line among those all possible lines so that data will either fall in the red zone or green zone. This overall process is known as the support vector machine. This optimal line is the best decision boundary. SVM searches this line through the maximum margin between two groups and the distance between the line and each one of these points are in equi-distance. Therefore, the sum of these two distances has to be maximized in order for the line to be the result of the SVM. Two points as shown in the figure are called the support vectors.

(II) Random Forests

Random forest is a method running the decision tree method multiple times, giving us the random forests. It can be used for both the classification and regression. A forest is

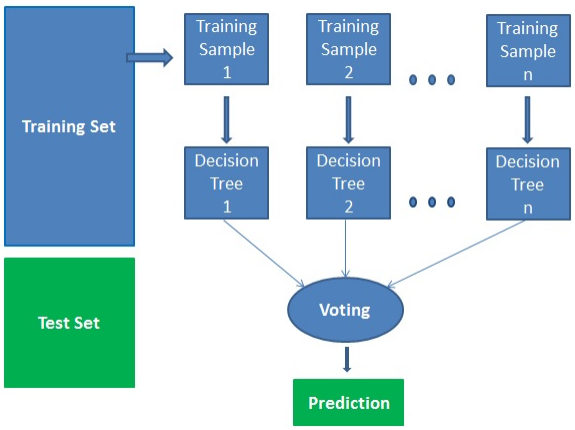


Figure 2.2: Random forests

comprised of many trees. In general, it is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, get predictions from each tree and selects the best solution by means of voting as shown in figure 2.2. It is widely used for a variety of applications, such as recommendation engines, image classification and feature selection. Since random forests are based on the decision trees, let's talk a little bit about the decision trees. Decision trees are predictive models that use a set of binary rules to calculate a target value. There are of also two types, classification and regression trees. Classification trees are used to create categorical and regression trees are used for continuous data sets.

(III) Artificial Neural Network

As the name suggests, artificial neural networks (ANN) are the brain-inspired systems which are intended to replicate the way that we humans learn. They are comprised of artificial neurons, also known as nodes. ANN has anywhere from a dozen to millions of artificial neural neurons arranged in a series of layers. ANN consist of input layers, output layers, and hidden layers as shown in figure 2.3 most of the time. Input layer is the layer of input signals also known as synapses that are passed to the neurons in the hidden layers. Synapses are assigned with weights. Weights are crucial in ANN because that is how ANN learns showing which signal is important, and which is not. Signals reached into the

neurons are summed to form weighted sum. This weighted sum is applied to the activation function. Based on the activation function neuron will either pass on the signal.

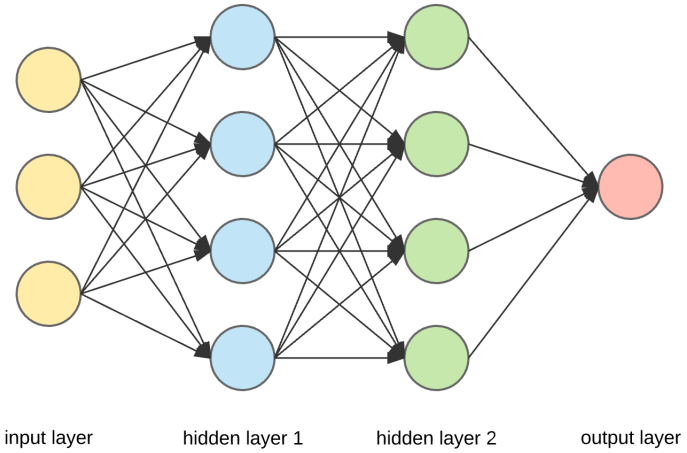


Figure 2.3: Artificial neural network

Activation Functions:

There are the many types of the activation functions. Following are the three important activation functions that are widely used in artificial neural network.

(i) Rectifier Function

The mathematical form of rectifier (relu) function is:

$$\varphi(x) = \max(x, 0) \tag{2.2}$$

It is one of the most used activation functions and the reason is it's sparsity. Only the positive values are allowed. Since the negative values are not passed, it will negate the possibility of occurrence of a dead neuron and therefore speeds up the process. The drawback of the ReLU is that the function is zero for the negative values of x and therefore gradient hits zero at those negative values. This will result in not adjusting the weights of the neurons during back propagation.

(ii) Sigmoid Function

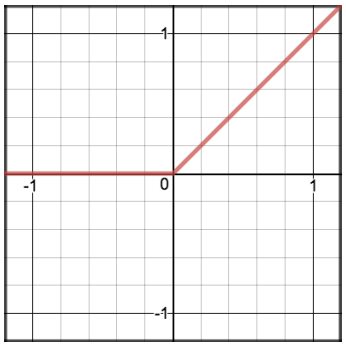


Figure 2.4: Rectifier function

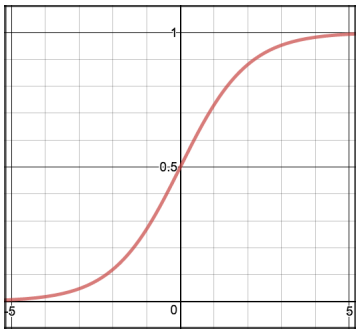


Figure 2.5: Sigmoid function

The mathematical form of sigmoid function is:

$$\varphi(x) = \frac{1}{1 + \exp(-x)} \tag{2.3}$$

It is a non-linear function. When the weighted sum is applied in the place of x, the output of this function is in between 0 and 1. The large negative numbers are scaled towards 0 and large positive numbers are scaled towards 1. It is continuously differentiable, monotonic, and has a fixed output range. One of the drawbacks of this function is that it has the problem of vanishing gradients. As you can see from the figure, when x, the input value to the function, is really small, that is, towards $-\infty$, the output of the sigmoid function will be closer to zero. Conversely, when x is really big, that is towards ∞, the output of the sigmoid function will be closer to 1. In those regions, the gradient is going to be very small and even vanishes. This is problem is known as "vanishing gradients" problem.

(iii) Hyperbolic Tangent:

The mathematical form of hyperbolic tangent function is:

$$\varphi(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)} \tag{2.4}$$

The output of this function is from -1 to 1. Since its output is zero centered unlike to sigmoid function, it's more preferred than sigmoid function. The drawback of this function is it also has the vanishing gradient problem.

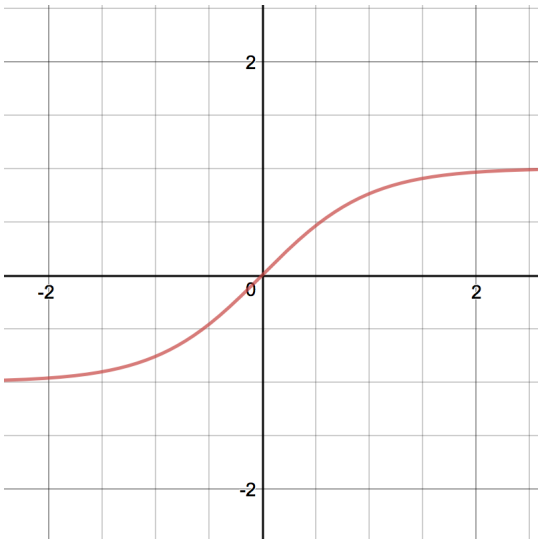


Figure 2.6: Hyperbolic tangent

2.2 Evaluation Metrics of ML Models

Models evaluation in machine learning is an essential part of the machine learning model development process. There are several ways to measure the performance of the developed model. We have summarized here the evaluation metrics of the classification as well as regression models.

2.2.1 Classification ML Evaluation Metrics

(i) Confusion Matrix: As the name suggests, the output of the ML model is in the form of the matrix. For example, let's say we have a binary problem that has two outputs, 1 for 'Yes' and 0 for 'No'. Then, the confusion matrix is as follows:

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 2.7: Confusion matrix

True Positive(TP): This is the case where we predicted 'Yes' and the actual is 'Yes'.

True Negatives (TN): This is the case where we predicted 'No' and the actual is 'No'.

False Positive(FP): This is the case where we predicted 'Yes' and the actual is 'No'.

False Negatives(FN): This is the case where we predicted 'No' and the actual is 'Yes'.

(ii) Accuracy: The Accuracy in terms of above parameters can be written as:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{2.5}$$

(iii) Sensitivity: It is the number of true positives divided by the number of true positives and the number of false negatives. It is also known as recall. The range of sensitivity is in between 0 and 1. The best value is 1 and worst is 0.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{2.6}$$

(iv) Specificity: It is calculated as the number of correct negative predictions divided by the total number of negatives. It is also known as True negative rate. The range of specificity is also 0 to 1. The best value is 1 and worst is 0.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{2.7}$$

(v) Precision (Positive Predictive Value): Precision is also known as positive predictive value. It is calculated as the number of correct positive predictions divided by the total number of positive predictions. The best precision is 1.0, whereas the worst is 0.0.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2.8}$$

(vi) F1-Score: It is the harmonic mean of the recall and precision and can be written as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.9}$$

F1 score is calculated as the weighted average of precision and recall. Therefore, both false positives and false negatives are taken into account. F1 is usually more useful than accuracy, especially if we have an uneven class distribution. Accuracy works best if false positives

and false negatives have similar cost. If the cost of false positives and false negatives are very different, its better to look at both precision and recall.

(vii) Mathew Correlation Coefficient (MCC): It is calculated as,

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})} \tag{2.10}$$

It is used to measure the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 represents no better than random prediction and -1 indicates disagreement between prediction and observation (source: Wikipedia).

2.2.2 Regression ML Evaluation Metrics

(i) Mean Square Error(MSE): It measures average squared error of our predictions. It calculates square difference between the predictions and the target for every input and then averages those values. It can be obtained as (write equation):

$$\text{MSE} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} (\text{y}_\text{i} - \hat{\text{y}}_\text{i})^2 \tag{2.11}$$

The lesser the value, the better the model is. It is non-negative, since were squaring the individual prediction-wise errors before summing them. Its value is zero for the perfect model.

(ii) Root Mean Square Error(RMSE): It is just the value obtained from the squared root of root mean square error. The square root is introduced to make scale of the errors to be the same as the scale of targets. So, it is written as:

$$\text{RMSE} = \sqrt{\frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} (\text{y}_\text{i} - \hat{\text{y}}_\text{i})^2} \tag{2.12}$$

(iii) Pearson Correlation Coefficient: It is a widely used correlation statistic that measures the degree of relationship between linearly related variables. For two variables x and y, it can be calculated as:

$$R = \frac{N(\sum xy) - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}} \tag{2.13}$$

(iv) Coefficient of Determination: Another way to evaluate the model is to measure the coefficient of determination, which is written as R^2. Mathematical form of R-squared is as follows:

$$R^2 = 1 - \frac{SSE}{SST} \tag{2.14}$$

where SSE is the sum of squared errors of our regression model and SST is the sum of squared errors of our baseline model, they are calculated as

$$SSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.15}$$

$$SST = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2 \tag{2.16}$$

where $\hat{y}$ is the expected value and $\bar{y}$ is the mean of the observed data. It is the square of the correlation (R) between predicted scores and actual scores. Thus, its values ranges from 0 to 1. $R^2 = 0$ means that the dependable variable can not be predicted from the independent variable. $R^2 = 1$ means the dependent variable can be predicted perfectly from the independent variable. R^2 in between 0 and 1 is the measurement of how likely the dependable variable is predictable.

(v) Concordance Index: CI measures the probability of two randomly drawn drug-target pairs with different label values are in the correct order. That means the prediction for the larger affinity is larger than the prediction of smaller affinity value. If f_i, f_j, y_i and y_j are the predictions and actual affinity values in the case of larger and smaller values respectively, CI can be obtained as:

$$\mathrm{CI} = \frac{1}{\mathrm{Z}} \sum_{\mathrm{y_i} > \mathrm{y_j}} \mathrm{h}(\mathrm{f_i} - \mathrm{f_j}) \tag{2.17}$$

where Z is the normalization constant, h(u) is the step function. h(u) is 1.0, 0.5 and 0.0 for $u > 0$, $u = 0$ and $u < 0$ respectively. The CI value lies in between 0.5 and 1.0. The value 0.5 is for the random predictor where as 1.0 is for the perfect prediction.

Chapter 3

Approach and Methodology

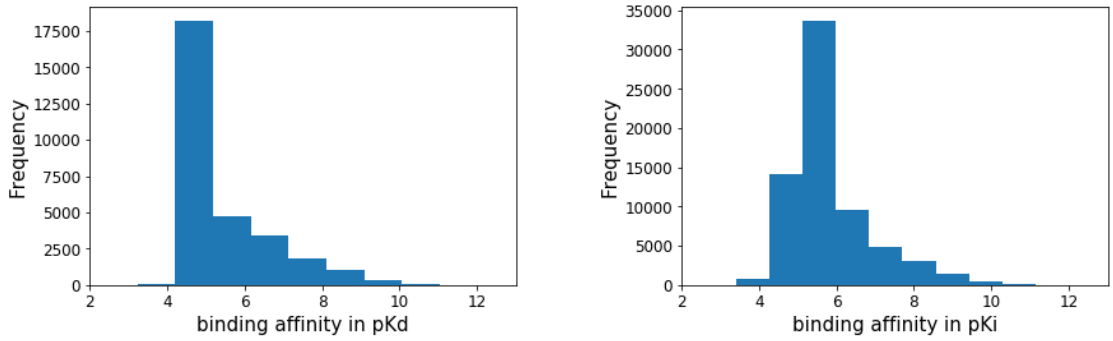
3.1 Data Preparation

This study is built on two drug target interaction data sets 1) Drug Target Commons generated by Tang et al. [43] and 2) Target Central Resource Database generated by UNM/NIH [30]. Both of these databases are publicly available drug target interaction databases. The combined data set not only contained the the drug-target but also ligand-protein interaction information. It contained the compound id, standard inchiKey, compound names, synonyms, target id, target preference name, gene names, wildtype or mutant, mutation information, pubmed id, standard type, standard value, standard units, assay format, compound concentration, substrate type, substrate value, assay description, and many other information. I have removed the duplicate lines found in the data set. The database had the K_i (inhibition constant), K_d (dissociation constant), and IC_{50} (half maximal inhibitory constant) values representing the binding affinities of kinase inhibitors complexes. I have focused only on K_i and K_d values. The data set had 5,77,061 lines at the beginning. This combined data set had 2,207 targets. Since I am focused only on kinases, I used the drug target interactions data set for 196 kinases. The final data set contained 97,564 instances obtained from 196 protein kinases and 6,792 compounds and also corresponding binding affinities. K_d or K_i values are transformed into log space, pK_d or pK_i using the following equation. All the values are divided by 10^9 converting nano Molar (nM) to Molar (M).

Table 3.1: Summary of data Sets

Data Sets	Kinases	Compounds	Interactions
DTC and Pharos	196	6792	97496
Metz	148	240	17258

$$pK_d = -\log10\left(\frac{K_d}{1e^9}\right) \tag{3.1}$$



(a) binding affinity values in pkd

(b) binding affinity values in pki

Figure 3.1: Distribution of data sets

3.2 Molecular Features (Descriptors)

3.2.1 Protein Features

Features of all 196 protein kinases are obtained from a web server 'PROFEAT' [21] using protein sequences. Protein sequences for all the kinases are extracted from NCBI [32]. The FASTA format or RAW format of sequence of protein is subjected as an input in a window provided in the web server. Multiple sequences of proteins can be supplied as a input in the form of text or csv file so that features of many protein sequences can be

generated and saved conveniently. An input sequence of protein with less than eight amino acids is not accepted by the server. Each generated file for each protein contains the 10 sets of commonly used structural and physicochemical features including 51 descriptors and 1447 descriptive values. These total features can be divided into six groups which are (i) amino acid, dipeptide composition (ii) Normalized Moreau-Broto autocorrelation (iii) Moran autocorrelation (iv) Geary autocorrelation (v) Composition, transition, distribution (vi) sequence order. Each group has been treated as an independent set of features.

3.2.2 Ligand Features

Molecular descriptors are obtained using the perl scripts 'Topological Pharmacophore Atom Pairs Fingerprints' (TPAPF) and 'Calculate Physicochemical Properties' through MayaChemTools [42]. TPAPF is based on the topological cross correlation of generalized atom type which is a simple molecular descriptor that leads to a compact, molecular size independent description of potential pharmacophores [2]. This representation scheme counts the distance between atom pairs and then coded into the histograms representing the exhaustive fingerprint of the molecule. The distances are expressed as the number of bonds of the shortest path connecting two nodes of non-hydrogen atoms in the molecular graph. Each atom on node contains a specific type from one of hydrogen-bond donor(D), hydrogen-bond acceptor (A), positively charged (P), negatively charged (N) or lipophilic (L). The number of occurences of all possible pharmocophore point pairs (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) for each molecule is determined. Distances of up to ten bonds for molecular representation, that is 15×10, leads to 150 dimensional vectors [41]. Large numbers of virtual compound libraries for similar structures can be searched based on the correlation-vector representation. Each library molecule is compared with the query vector. A measure of Euclidean distance can be used to measure the similarity between each library molecule and query vector. The Euclidean distance between two molecules A

and B is obtained by the following equation.

$$D(A, B) = \sqrt{\sum_{i=1}^{150}(v_i^A - v_j^B)} \tag{3.2}$$

where v_i^A and v_j^B represent the correlation vectors for A and B respectively. I have also used to the physicochemical features of the compounds using the mayachem tools. The properties of compounds that I have used as features are MolecularWeight, HeavyAtoms, MolecularVolume, RotatableBonds, HydrogenBondDonor, HydrogenBondAcceptors, SLogP, TPSA.

Molecular weight: Molecular weight is one of the most significant descriptors. It has been correlated with a number of key parameters in drug discovery. The molecular weights of the marketed drugs or drug candidates has been increasing [37]. It has also been found that molecular weights in oral drugs approved in between 1983 and 2002 had 14 percent greater molecular weight than those approved before 1983 [19]. Higher number of failures in early clinical trials are found with the physical properties with low molecular weight.

Polar Surface Area: Molecules having a large polar surface area (PSA) may face difficulty in transiting biological membranes [4].

Rotatable Bonds: Molecular flexibility of a compound is an important property that is frequently optimized in drug discovery process. The number of rotatable bonds are used in regard to molecular flexibility.

Hydrogen Bond Donors or Acceptors: Lipinski [22] showed that most of marketed drugs have fewer than 5 hydrogen bond donors and fewer than 10 hydrogen bond acceptors. Hydrogen bonds are found to play key roles in drug discovery processes. The location of the hydrogen bond is important to determine the potency and selectivity of a compound. Similarly heavy atoms, Molecular volume, and SLogP are also key elements in drug discovery program.

3.2.3 Machine Learning Model Development

Different modeling techniques have been developed to depict various aspects of the mathematical modeling of data. In the past decades, many statistical methods and stochastic approaches have been proposed for prediction such as in one of the recent article [24], the authors used stochastic models to describe a unique type of dependence of measurements on time different types of data and obtained very low errors of estimated parameters. In this study, I have used the machine learning approach aiming to predict the binding affinity values of ligand-kinase interactions with low predictive errors. Since the binding affinity values are quantitative values, I have used the Random Forest Regressor (RFR), which is an ensemble of various decision Trees [7]. The model is optimized by grid search method with 10 fold cross validation based on scikit-learn library machine learning library for python [36]. Cross validation describes the process of splitting the whole data set into x (say) parts and using each one of them sequentially as the test data set while combining the others to the training data. I also developed model based random search method [6] and compared the results with the grid search method. For RFR, I have tuned the hyperparmeters: n_estimators, max_features, and mini_samples_split as they are found effective while developing the model for both grid search and random search methods.

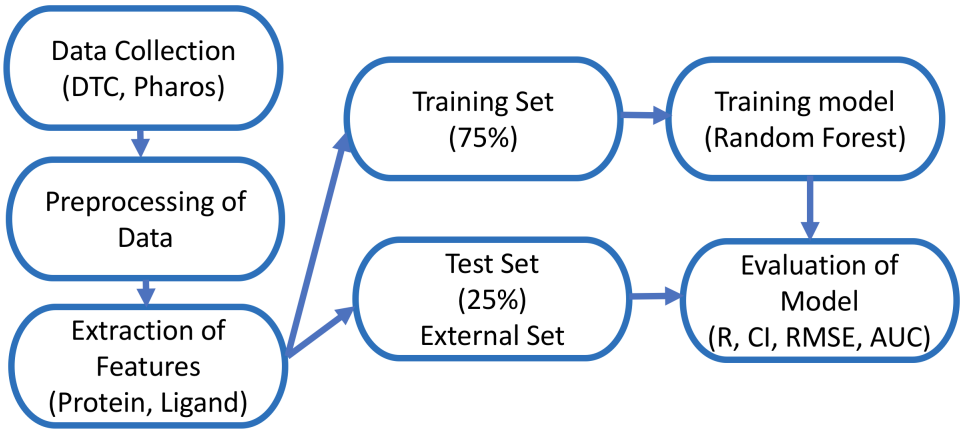


Figure 3.2: Project work flow

3.2.4 Grid Search Vs Random Search:

Hyperparameters are model specific properties which are fixed even before the model is trained or tested on the data. Hyperparameter optimization is considered as the trickiest part of the machine learning model as well as artificial intelligence in models development processes. The main purpose of it is to find the best spot in the available space so that the best set of hyperparameter leads to high precision and accuracy. There are several parameter tuning techniques but grid search and random search are the most popular methods.

Grid Search Method:

In this method, every combination of a preset of values of the hyperparameters are used to train the model and among them, the best combination is chosen which ends up with best results. It is usually considered as a good choice when we have a small set of parameters to optimize. For each hyperparameter, we give a set of candidate values to explore from range of all possible values. We train and evaluate the data for each combination and the at the end we keep one which yields the best results. The main problem of the grid search is that time increases exponentially as the number of parameters grow. For example, if there are p parameters and each parameter can take r number of values, then running time is calculated as $O(r^p)$. But, when each parameter can take the different number of values, then we simply multiply the all of them.

For example; a random forest regressor has a list of hyperparameters and among them, some are found effective while developing the machine learning model such as n_estimators, max_features, min_samples_split. These three are found helpful to boost the performance of the model. So, the possible number of ways that the grid search run can be calculated is as follows: if n_estimators = [100,200,300,400,500], max_features = ['auto', 'sqrt', 'log2', None], and min_samples_split = [2, 5, 10], Then total number of ways grid search run is equal to (5*4*3)=60 and if we also use the cross validation as CV = 10, then 60*10 = 600. That means the model runs for 600 times (i.e., 600 iterations).

Random Search Method:

This idea was proposed by James Bergstra and Yoshua Bengio[6]. In this method, we normally provide the statistical distribution of each hyperparameters rather than any specific set of parameters. The values are randomly picked on each iteration to train the model. It is more generally a process of choosing a representative sample of parameter from the entire space to know about the whole data. It is good in testing wide ranges of values and normally ends to give the best results, however best parameters are not always guaranteed. It is based on the notion that each hyperparameters are not equally important for the model development. Random search is popular for large data sets. In high dimensional parameter space, grid search performs worse with the same iterations because points become sparse. As all the hyperparameters are not equally important, grid search wastes a lot of iteration whereas random search is faster and saves a lot of time. For example; if we use the same set of hyperparameters just like in grid search, then the number of iterations in random search are actually determined by the value chosen for n_iter and k-folds cross validation. If we have chosen n_iter = 10, and CV = 10, then the total number of iterations are only 10*10 = 100. So, if we compare with the grid search it is way faster than grid search.

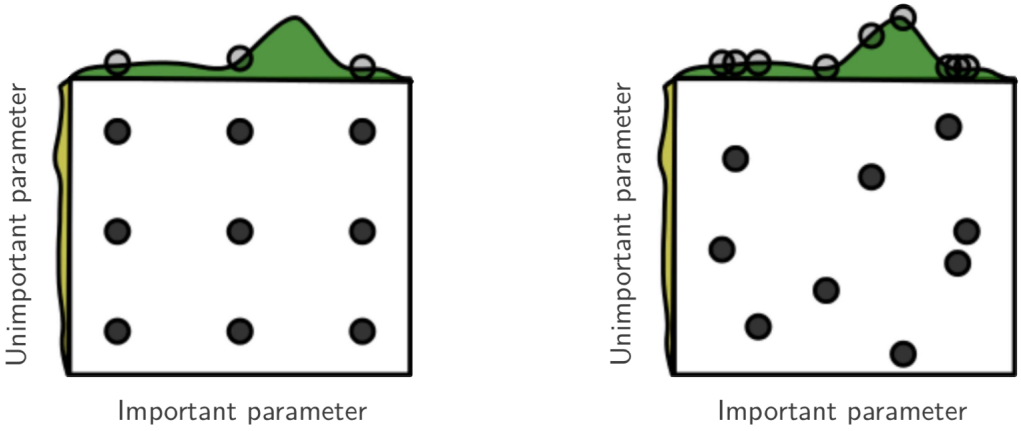


Figure 3.3: Grid search vs random search (from, Random Search for Hyper-Parameter Optimization[6], James Bergstra, Yoshua Bengio)

3.2.5 K-Fold Cross Validation

Cross validation (CV) is an important technique for tuning parameters and producing robust measurements of model performance. There are several types of cross validation techniques such as k-fold, leave one out cross validation (LOOV), nested, etc depending upon the nature of data. For example, nested cross validation is used for time series data where K-fold CV is not valid option while tuning the ML model. It is because the time series data takes into account the temporal dependencies of the measurements [25]. Our model is based on the K-fold cross validation. At first the data is divided into train and test set. Test set is reserved for evaluation after the model selection. We have used k-folds validation with CV=10. k-folds cross validation means the data is divided into k different subsets where (k-1) are used to train the model and 1 remaining is used to test the model. For each unique group, hold it as test set and the remaining sets as training set. So, the model is trained with training set and evaluated by test set. If k = 10, then there are 10 unique processes as every time 1 unique test is used to evaluate the model and other remaining sets are used to train the model. The overall evaluation of the model is the average value of all values.

Figure 3.4: K-fold cross validation (from, Karl Rosaen, http://karlrosaen.com)

3.3 Evaluation of Machine Learning Models

This is an important part of the machine learning where we evaluate the developed models on the basis of some statistical parameters. Since this is a regression problem, we have evaluated the model using Root Mean Square Error (RMSE), Pearson's correlation coefficient (R), Concordance Index (CI), and Area Under the ROC Curve (AUC). We used the tools that are made avilable by IDG-DREAM challenge [1] to evualute the model. Some of them are as summarized below.

3.3.1 Evaluation Metrics

Concordance Index: Since the binding affinities in drug target interactions are continuous values, one of the evaluation metrics we used was the Concordance Index (CI) for the prediction accuracy [33]. CI measures the probability of two randomly drawn drug-target pairs with different label values are in the correct order. That means the prediction for the larger affinity is larger than the prediction of smaller affinity value. If d_i, d_j, x_i and x_j are the predicted and actual affinity values in the case of larger and smaller binding affinities respectively, CI can be obtained as:

$$CI = \frac{1}{Z} \sum_{x_i > x_j} h(d_i - d_j) \tag{3.3}$$

where Z is the normalization constant, h(u) is the step function. h(u) is 1.0, 0.5 and 0.0 for $u > 0$, $u = 0$ and $u < 0$ respectively. The CI values lies in between 0.5 and 1.0. The value 0.5 for the random predictor where as 1.0 for the perfect prediction.

Root Mean Squre Error(RMSE): If B_a and B_p are the actual and predicted binding affinities. Error can be calculated as:

$$Error(E_i) = B_{a_{(i)}} - B_{p_{(i)}} \tag{3.4}$$

Mean Square Error (MSE) is calculated as mean of differences of actual binding affinity values and calculated binding affinity values.

$$\mathrm{MSE} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{E}_{\mathrm{i}}^{2} \tag{3.5}$$

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{3.6}$$

RMSE value depends up on the nature of data sets. We always try to minimize it. It should be similar for both training and test set. If values are higher in test than the training set, it is likely that we have badly over fit the data.

Correlation Coefficient Coefficient (R):

$$\overline{\mathrm{B_a}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{B_{a_{(i)}}} \tag{3.7}$$

$$\overline{\mathrm{B_p}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{B_{p_{(i)}}} \tag{3.8}$$

$$\mathrm{R} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{B_{a_{(i)}}} - \overline{\mathrm{B_a}})(\mathrm{B_{p_{(i)}}} - \overline{\mathrm{B_p}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{B_{a_{(i)}}} - \overline{\mathrm{B_a}})^2 \sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{B_{p_{(i)}}} - \overline{\mathrm{B_p}})^2}} \tag{3.9}$$

R measures the linear relationship between the actual and predicted binding affinity scores. It lies in between -1 and 1.

Area Under the Curve (AUC): The area under the Receiver Operating Characteristic (ROC) curve is generally adopted in binary classification problems. However, it can also be used to measure the regression problems by converting the quantitative values into binary values by selecting thresholds. There are different ways to measure AUC for regression problems. We have measured the average AUC converting the actual compound kinase interaction values (i.e, binary affinities) into binary labels given certain interaction thresholds. In roc curve the True Positive Rate (TPR) (i.e., sensitivity) is plotted in function of the False Positive Rate (FPR) (i.e., 1 – specificity) for different cut-off points.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{3.10}$$

$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})} \tag{3.11}$$

Accuracy is measured by the area under the ROC curve. An area of 1 (i.e, 100%) repre-

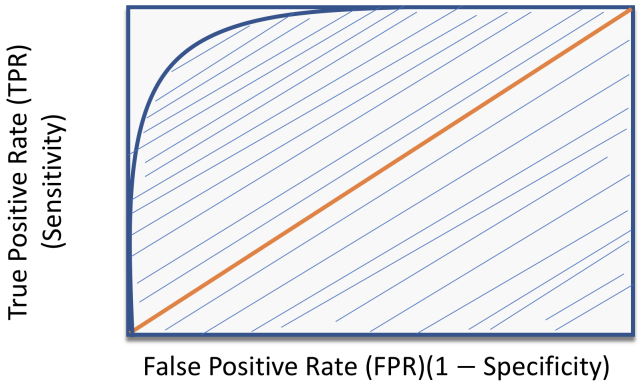


Figure 3.5: Area under ROC curve

sents a perfect test; an area of .5 (i.e, 50%) represents a worthless test (or random test). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

Chapter 4

Results and Discussions

4.1 Results

The performance of RFR model by grid search method and random search method is shown in table 4.1 and 4.2. The performance of model with increasing number of trees for both test and external sets are shown tables 4.3 and 4.4.

Table 4.1: Performance of model developed by grid search method

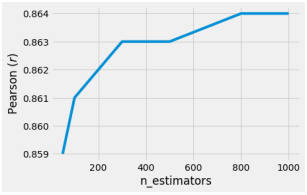
Grid search method							
Test data set				Metz data set			
R	RMSE	CI	avg. AUC	R	RMSE	CI	avg. AUC
0.864	0.523	0.814	0.959	0.725	0.532	0.74	0.938

Table 4.2: Performance of model developed by random search method

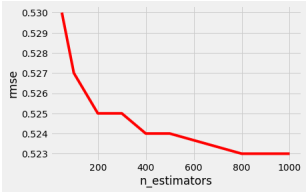
Random search method							
Test data set				Metz data set			
R	RMSE	CI	avg. AUC	R	RMSE	CI	avg. AUC
0.862	0.525	0.813	0.959	0.723	0.533	0.738	0.935

Table 4.3: Performance of model on test set with increasing the the number of trees

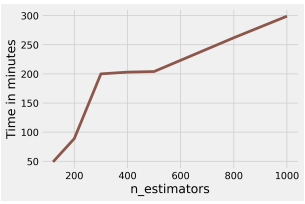
n-estimators	Pearson cor. coeff (R)	RMSE	Concordance Index(CI)	avg. AUC
50	0.859	0.531	0.809	0.957
100	0.861	0.527	0.812	0.959
200	0.862	0.525	0.813	0.959
300	0.863	0.524	0.813	0.959
400	0.863	0.524	0.813	0.959
500	0.863	0.523	0.814	0.959
800	**0.864**	**0.523**	**0.814**	**0.959**
1000	0.864	0.523	0.814	0.959



(a) R vs number of trees



(b) RMSE vs number of trees

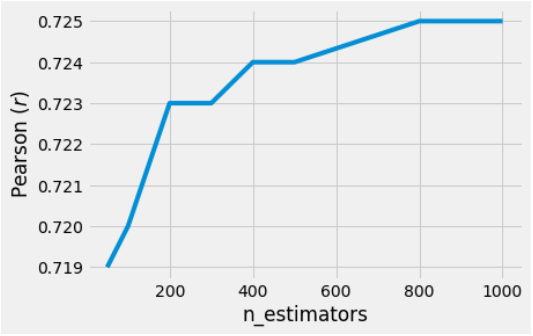


(c) time vs number of trees

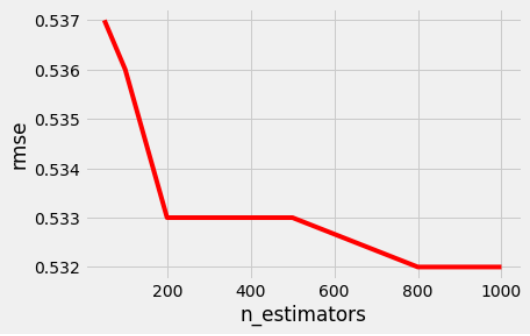
Figure 4.1: Plots showing the performance of RFR model on test data set

Table 4.4: Performance of RFR model on external data set with increasing the the number of trees

n-estimators	Pearson cor. coeff (R)	RMSE	Concordance Index(CI)	avg. AUC
50	0.719	0.537	0.736	0.932
100	0.72	0.536	0.737	0.934
200	0.723	0.533	0.739	0.935
300	0.723	0.533	0.739	0.936
400	0.724	0.533	0.74	0.936
500	0.724	0.533	0.74	0.937
800	**0.725**	**0.532**	**.74**	**0.938**
1000	0.725	0.532	0.74	0.938

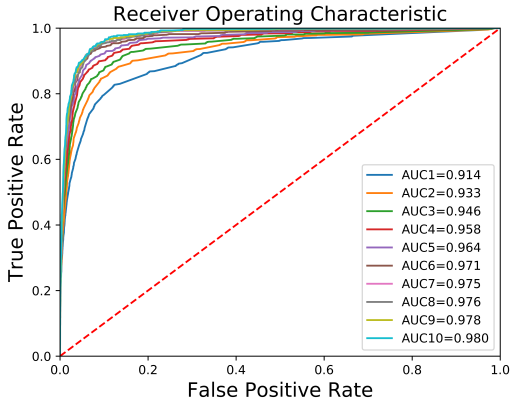


(a) R vs number of trees

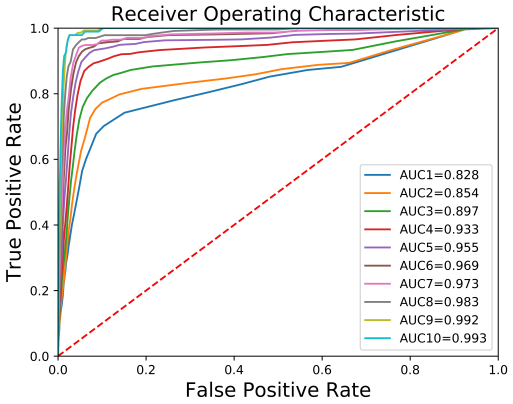


(b) RMSR vs number of trees

Figure 4.2: Plots showing the performance of RFR model on external data set

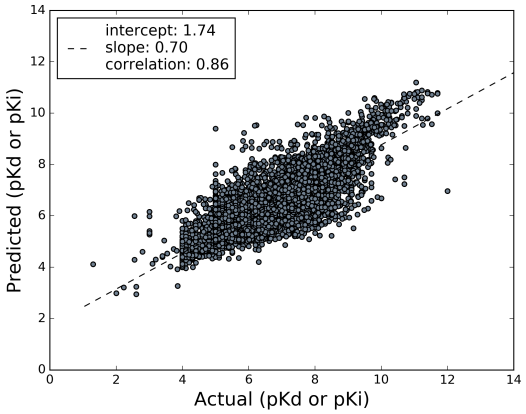


(a) Area under the ROC curve for test data set (avg. AUC=0.959)

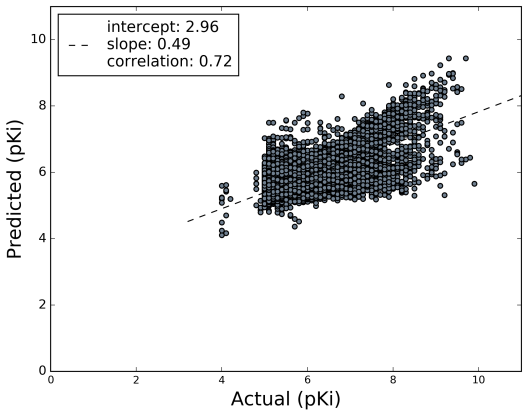


(b) Area under the ROC curve for Metz data set (avg. AUC=0.938)

Figure 4.3: Plots showing the area under the ROC curve.



(a) Actual vs predicted binding affinity of test data set



(b) Actual vs predicted binding affinity of Metz data set

Figure 4.4: Scatter plots for actual vs predicted binding affinity of test and external data sets.

Table 4.5: Results of IDG DREAM Drug-Kinase Binding Prediction Challenge for our model and best model in challenge

Model	**Round 1b**						**Round 2**		
	Pearson (R)	Spearman (ρ)	CI	AUC	F1	rmse	Spearman (ρ)	AUC	rmse
Our model	0.368	0.316	0.60	0.689	0.283	1.223	0.291	0.688	1.14
Best model based on AUC	0.544	0.556	0.711	0.78	0.40	1.025	0.527	0.794	0.897

4.2 Discussion of Results

The agreement between the binding affinity values were evaluated in three different ways. First off, the model was evaluated by blind test set (25%) and showed the highest consistency with a Pearson correlation coefficient of 0.86. Other evaluated scores were rmse 0.52, concordance index score 0.81, and average AUC 0.95. Since the model was evaluated by internal 10 folds cross validation where each different test out of 10 sets are internally evaluated, it is a purely unbiased model. I tested the model on another blind data set by Metz et al. I compared and excluded the compound target pairs which are common with our data set. There were 17,258 completely new compound kinase pairs containing the binding affinity (pK_i) values. The scores obtained for Metz data set were slightly lower than the test data set with R 0.72, RMSE 0.53, CI 0.74 and avg. AUC 0.938; however, the RMSE and avg. AUC scores were still promising. The low value of RMSE shows that predicted values were close to actual values and the high value of avg-auc indicates the power of the model to discriminate between the true positives and negatives. I observed an increase in the correlation coefficient with the increase in number of trees; however, it was not improving much after 800, so I stopped at 800. Increase in number of trees raised the performance of model to a certain level but after that it only increased the computational cost.

The model was developed based on both grid search method as well as random search method with in same parameter space. I did this because I wanted to compare the computational cost and model performance developed under these two widely popular methods. The hyperparameters in RFR which were effective to improve the model performance were namely, n_estimators (or number of trees), max_features and min_samples_split. The performance of model was almost same for both grid search and random search methods [6]. However, the computational cost was significantly reduced by random search method. Thus, random search method gave the similar results in a short interval of time compared to grid search method.

The model was also tested in IDG-DREAM challenge and scores obtained in round 1b and round 2 are as shown in table 4.5. Test set of round 1b contains the 430 pK_d values between 207 kinases and 25 compounds and round 2 test set contains the 394 pK_d values between 199 kinases and 70 compounds. I was at 13th (based on AUC) and top 29th (based on Pearson correlation coefficient) at round 1b out of 216 submitted scores and at top 41th (based on Spearman correlation coefficient) and 32th (based on AUC) position out of 101 submitted scores. The scores obtaied were lower than expected. However, I worked more even after the challenge to increase the performance of the model. Basically, more relevant features were added that actually boosted up the model performance to some extent.

Chapter 5

Conclusion

I proposed Random Forest Regressor (RFR) machine learning model to predict the ligand kinase binding affinity using the drug or drug like compounds and targets information. The performance of model showed that RFR model based on structural, physicochemical and atom pair based two-dimensional pharmacophore fingerprints has the potential to predict the binding affinity values. The model was developed using both grid search and random search methods under the same parameter space. I observed that hyperparameters namely, n_estimators, max_features and min_samples_split in RFR were found effective to tune the best model. The running time and performance of the model based on grid search and random search methods were compared. The results based on two methods were not significantly different. However, random search reduces the model building time significantly. I also observed that increase in number of trees hardly improved accuracy of model but increased the computational time instead. As database contained both dissociation constant (K_d) and inhibition constant (K_i), the performance of model showed that these values can be combined to develop the machine learning model. As a future work, I will focus on building better model based on other machine learning methods with more data.

References

[1] https://www.synapse.org/#!Synapse:syn15667962/wiki/590573.

[2] J. Klein, C. W. Lehmann, H.-W. Schmidt, W. F. Maier,Angew. Chem.1998,110, 3557 3561;Angew. Chem. Int. Ed.1998,37, 3369 3372.

[3] Biopharmaceutical Research & Development, the process behind new medicines. http://phrma-docs.phrma.org/sites/default/files/pdf/rd_brochure_022307.pdf.

[4] Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. Pharm Res. 1999 Oct; 16(10): 15141519. doi: 10.1023/A:1015040217741.

[5] A. C. Anderson. The process of structure-based drug design. *Chemistry & Biology*, 10(9):787 – 797, 2003.

[6] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *JMLR*, page 305, 2012.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[8] R. Z. Cer, U. Mudunuri, R. Stephens, and F. J. Lebeda. IC50–to–Ki: a web-based tool for converting IC50 to Ki values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Research*, 37(suppl-2):W441–W445, 04 2009.

[9] A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu, and T. Aittokallio. Computational-experimental approach to drug-target interaction mapping: A case study on kinase inhibitors. *PLOS Computational Biology*, 13:1–28, 08 2017.

[10] F. Clementi and G. Fumagalli. *General and molecular pharmacology: principles of drug action*. John Wiley and Sons, Inc., 2015.

[11] A. C. Dar and K. M. Shokat. The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling. *Annual Review of Biochemistry*, 80(1):769–795, 2011. PMID: 21548788.

[12] E. D. G. Fleuren, L. Zhang, J. Wu, and R. J. Daly. The kinome 'at large' in cancer. *Nature Reviews Cancer*, 16:83 EP –, Jan 2016.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[14] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester. Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *J Cheminform*, 9(1):24–24, Apr 2017. 29086119[pmid].

[15] A. L. Hopkins. Predicting promiscuity. *Nature*, 462:167 EP –, Nov 2009.

[16] H.-J. Huang, H. W. Yu, C.-Y. Chen, C.-H. Hsu, H.-Y. Chen, K.-J. Lee, F.-J. Tsai, and C. Y.-C. Chen. Current developments of computer-aided drug design. *Journal of the Taiwan Institute of Chemical Engineers*, 41(6):623 – 635, 2010.

[17] D. R. Knighton, J. H. Zheng, L. F. Ten Eyck, N. H. Xuong, S. S. Taylor, and J. M. Sowadski. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science*, 253(5018):414–420, Jul 1991.

[18] I. Kundu, G. Paul, and R. Banerjee. A machine learning approach towards the prediction of proteinligand binding affinity based on fundamental molecular properties. *RSC Adv.*, 8:12127–12137, 2018.

[19] P. D. Leeson and A. M. Davis. Time-related differences in the physical property profiles of oral drugs. *Journal of Medicinal Chemistry*, 47(25):6338–6348, Dec 2004.

[20] Q. Li and L. Lai. Prediction of potential drug targets based on simple sequence properties. *BMC Bioinformatics*, 8:353–353, Sep 2007. 17883836[pmid].

[21] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen. Profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*, 34(Web Server issue):W32–W37, Jul 2006. 16845018[pmid].

[22] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1):3 – 25, 1997. In Vitro Models for Selection of Development Candidates.

[23] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[24] M. C. Mariani, M. A. M. Bhuiyan, and O. K. Tweneboah. Estimation of stochastic volatility by using ornsteinuhlenbeck type models. *Physica A: Statistical Mechanics and its Applications*, 491:167 – 176, 2018.

[25] M. C. Mariani, M. A. M. Bhuiyan, O. K. Tweneboah, H. Gonzalez-Huizar, and I. Florescu. Volatility models applied to geophysics and high frequency financial market data. *Physica A: Statistical Mechanics and its Applications*, 503:304 – 321, 2018.

[26] J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, and P. J. Hajduk. Navigating the kinome. *Nature Chemical Biology*, 7:200 EP –, Feb 2011.

[27] F. Moriaud, S. B. Richard, S. A. Adcock, L. Chanas-Martin, J.-S. Surgand, M. Ben Jelloul, and F. Delfaud. Identify drug repurposing candidates by mining the Protein Data Bank. *Briefings in Bioinformatics*, 12(4):336–340, 04 2011.

[28] D. K. Morrison. Map kinase pathways. *Cold Spring Harb Perspect Biol*, 4(11):a011254. 23125017[pmid].

[29] R. Mullin. Cost to develop new pharmaceutical drug now exceeds $2.5b. *Scientific American*, 2014.

[30] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, G. Liu, A. Ma'ayan, G. Mandava, S. Mani, S. Mehta, J. Overington, J. Patel, A. D. Rouillard, S. Schürer, T. Sheils, A. Simeonov, L. A. Sklar, N. Southall, O. Ursu, D. Vidovic, A. Waller, J. Yang, A. Jadhav, T. I. Oprea, and R. Guha. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res*, 45(D1):D995–D1002, Jan 2017. 27903890[pmid].

[31] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. Juhl Jensen, A. Karlsson, G. Liu, A. Ma'ayan, G. Mandava, S. Mani, S. Mehta, J. Overington, J. Patel, A. D Rouillard, S. Schurer, and R. Guha. Pharos: Collating protein information to shed light on the druggable genome. 45, 11 2016.

[32] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 11 2015.

[33] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, and T. Aittokallio. Toward more realistic drug-target interaction predictions. *Brief Bioinform*, 16(2):325–337, Mar 2015. 24723570[pmid].

[34] E. Pauwels, V. Stoven, and Y. Yamanishi. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, 12(1):169, May 2011.

[35] L. R. Pearce, D. Komander, and D. R. Alessi. The nuts and bolts of agc protein kinases. *Nature Reviews Molecular Cell Biology*, 11:9 EP –, Jan 2010. Review Article.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] J. R. Proudfoot. The evolution of synthetic oral drug properties. *Bioorganic & Medicinal Chemistry Letters*, 15(4):1087 – 1090, 2005.

[38] R. Roskoski. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacological Research*, 100:1 – 23, 2015.

[39] R. Roskoski. Properties of fda-approved small molecule protein kinase inhibitors. *Pharmacological Research*, 144:19 – 50, 2019.

[40] G. Schneider and U. Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649–663, 2005.

[41] G. Schneider, W. Neidhart, T. Giller, and G. Schmid. scaffold-hopping by topological pharmacophore search: A contribution to virtual screening. *Angewandte Chemie International Edition*, 38(19):2894–2896, 1999.

[42] M. Sud. Mayachemtools: An open source package for computational drug discovery. *Journal of Chemical Information and Modeling*, 56(12):2292–2297, Dec 2016.

[43] J. Tang, Z. ur Rehman Tanoli, B. Ravikumar, Z. Alam, A. Rebane, M. Vh-Koskela, G. Peddinti, A. J. van Adrichem, J. Wakkinen, A. Jaiswal, E. Karjalainen, P. Gautam, L. He, E. Parri, S. Khan, A. Gupta, M. Ali, L. Yetukuri, A.-L. Gustavsson, B. Seashore-Ludlow, A. Hersey, A. R. Leach, J. P. Overington, G. Repasky, K. Wennerberg, and T. Aittokallio. Drug target commons: A community effort to build a consensus knowledge base for drug-target interactions. *Cell Chemical Biology*, 25(2):224 – 229.e2, 2018.

[44] G. A. Wayman, H. Tokumitsu, M. A. Davare, and T. R. Soderling. Analysis of camkinase signaling in cells. *Cell Calcium*, 50(1):1–8, Jul 2011. 21529938[pmid].

[45] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban. Keynote review: In vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discovery Today*, 10(21):1421 – 1433, 2005.

Curriculum Vitae

Govinda KC was born in Nepal. He Joined The University of Texas at El Paso (UTEP) in 2014 as a graduate student in Department of Physics. He got Master's in Physics in summer, 2016 and joined the Computational Science Program in August, 2016. He is working as a teaching assistant in Department of Mathematics at UTEP.

Permanent address: 4110 Westcity Ct
El Paso, Texas 79912-4927