PREDICTION OF DRUGGABILITY PROPERTIES OF CHEMICALS USING MACHINE LEARNING TECHNIQUES

by

Gamze Ege Kahya

B.S., Computer Engineering, Boğaziçi University, 2015

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering

Boğaziçi University

2020

ACKNOWLEDGEMENTS

I would like to thank my advisor Assoc. Prof. Arzucan Özgür for her constant motivation and support for this work. Her deep knowledge, guidance and positive thoughts helped me greatly in all the time of research and writing of this thesis. I would like to thank my co-advisor Prof. Kutlu Ülgen for her guidance, motivation, support with her ideas and showing me the directions. I thank Prof. Haluk Bingöl, Prof. Olcay Taner Yıldız and Assoc. Prof. Burak Alakent for participating in my thesis jury and giving feedback.

I would like to express my deepest gratitude to my beloved husband, for his constant support and encouragement, never ending patience and love. He has always been the greatest supporter of my education, helping me to pursue my excitement and willingness in every moment of our lives. I would like to thank my little son for being the joy and sun of my life. His beautiful presence makes my life extremely cheerful and challenging at the same time. Thank you my little man, for making me believe in myself to achieve anything in life.

I am very much grateful to my parents who are constantly supporting me, sharing my burden in life, cheering me up when I am down, praying for me. I am very much thankful to them for their endless love, encouragement and sacrifices for my education and my life. I would like to express my very great appreciation to my sister, for her warmly embracing and motivating thoughts, enlightening me whenever I lost my way in life. She has been always my best companion and friend and always will be.

I am also very grateful to my best friend, Zeynep Doğru, for her constant motivating words, understanding and supportive ideas and being my personal counselor whenever I needed. She always shows me the things I have achieved, and makes me believe that I will also achieve the next challenge.

ABSTRACT

PREDICTION OF DRUGGABILITY PROPERTIES OF CHEMICALS USING MACHINE LEARNING TECHNIQUES

Drug discovery is the process of designing and developing new medicine. The research and clinical experiments for new drug proposals are costly and take a long time. Although there has been proposed a lot of drug-like molecules, the number of drugs that are confirmed by regulating bodies and released to the market is very low. That is because most of the drug candidate molecules have low pharmacokinetic properties. Therefore, early assessments of ADMET properties have gained extreme importance for pharmaceutical industry, to be able to avoid costly failures. Here, our aim is to come up with an approach that reliably predicts druggability features of drug candidate molecules as well as to point out the relations between ADMET properties and molecular descriptors.

In this thesis study, we examine and compare 4 different molecule representations to predict druggability features of molecules, using 3 different machine learning algorithms; namely k-nearest neighbor, support vector machine classifier and random forest on 9 ADMET property datasets. We conclude that among all molecular representations, morgan fingerprint performs better in terms of accuracy and F-measure, however run time for parameter tuning and train is longer with fingerprint representations. As far as the machine learning algorithms are concerned, SVM classifier with morgan fingerprint performs better with higher accuracy and F-measure. With descriptor vector representation, we examine a set of molecular descriptors and using RF classifier, we evaluate most effective molecular descriptor for each ADMET property. For some datasets, we add the most contributive descriptor to morgan fingerprint representation and report an increase on evaluation metrics.

ÖZET

Makina Öğrenmesi Tekniklerini Kullanarak Kimyasalların İlaç Olabilirliğinin Tahmin Edilmesi

İlaç keşfi, yeni ilaç tasarlama ve geliştirme sürecidir ancak bu alanda yapılan araştırmalar maliyetlidir ve uzun zaman almaktadır. Piyasada ilaç benzeri bir çok molekül önerilmiş olmasına rağmen, onaylanan ve piyasaya sürülen ilaçların sayısı çok düşüktür, çünkü ilaç aday moleküllerinin çoğu düşük farmakokinetik özelliklere sahiptir. Bu nedenle, ilaç benzeri moleküllerin Emilim, Dağılım, Metabolizma, Boşaltım, Zehirlilik özelliklerinin erken değerlendirilmesi, maliyetli ve başarısız çalışmalardan kaçınılabilinmesi açısından ilaç endüstrisi için oldukça önemlidir. Amacımız, ilaç aday moleküllerinin ilaç olabilme özelliklerini öngören bir yaklaşım ortaya koymak ve ADMET özellikleri ile moleküler tanımlayıcılar arasındaki ilişkileri belirtmektir.

Bu tez çalışmasında, k en yakın komşu, destek vektör makineleri ve rastgele orman olmak üzere 3 farklı makine öğrenme algoritması kullanarak, moleküllerin ilaç olabilme özelliklerini tahmin etmek için 4 farklı molekül temsilini, 9 ADMET özelliği veri kümesi üzerinde inceleyip karşılaştırıyoruz. Tüm moleküler gösterimler arasında, morgan parmak izinin doğruluk ve F-ölçümü açısından daha iyi performans gösterdiğine, ancak parametre ayarlama ve algoritma eğitim süresinin parmak izi gösterimleri ile daha uzun olduğu sonucuna vardık. Makine öğrenme algoritmaları söz konusu olduğunda, morgan parmak izi kullanılan DVMS, daha yüksek doğruluk ve F-ölçümü göstermektedir. Moleküler tanımlayıcılardan oluşan vektör temsili ile RF sınıflandırıcısını kullanarak her ADMET özelliği için en etkili moleküler tanımlayıcıyı değerlendiriyoruz. Bazı veri kümeleri için, morgan parmak izi temsiline en çok katkıda bulunan tanımlayıcıyı ekleyerek yaptığımız deneylerde değerlendirme metriklerinde artış gözlemekteyiz.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZET v
LIST OF FIGURES ix
LIST OF TABLES xi
LIST OF SYMBOLS xiii
LIST OF ACRONYMS/ABBREVIATIONS xiv
1. INTRODUCTION 1
1.0.1. Drug Discovery 1
1.0.2. Motivation 2
1.0.3. Our Contribution 3
1.0.4. Thesis Outline 3
2. RELATED WORK 5
3. BACKGROUND 7
3.1. Machine Learning Methods 7
3.1.1. K-Nearest Neighbor Algorithm 7
3.1.2. Support Vector Machine (SVM) Classifier 8
3.1.3. Random Forest 10
4. CHEMICAL DRUGGABILITY FEATURES 12
4.1. ADMET Properties 12
4.1.1. Chemical Mutagenicity (AMES) 12
4.1.2. Blood Brain Barrier 13
4.1.3. Cytotoxicity (HepG2) 14
4.1.4. Drug-Induced Liver Injury 14
4.1.5. hERG Blockers 15
4.1.6. Human Liver Microsomal Stability 15
4.1.7. Mitochondrial Membrane Potential Disruption 15
4.1.8. Pgp Substrates and Inhibitors 16

4.2. Chemical Descriptors 17
4.2.1. Molecular Weight 18
4.2.2. The Water Octanal Partition Coefficient 18
4.2.3. Topological Polar Surface Area 19
4.2.4. Number of Hydrogen Bond Donors and Acceptors 20
4.2.5. Number of Rotatable Bonds 20
4.2.6. Molar Refractivity 21
4.2.7. Balaban's J Value 21
4.2.8. Number of Valence Electrons 22
4.2.9. Number of Aromatic Rings 22
4.2.10. Atom Numbers 23
4.2.11. Acidic and Basic Group Counts 23
4.2.12. Eccentric Connectivity Index 23
4.2.13. Molecular Diameter and Radius 24
4.2.14. Petitjean Index 24
4.2.15. Atomic and Bond Polarizability 25
4.3. Chemical Fingerprints 25
4.3.1. Topological Fingerprints 26
4.3.2. Morgan Fingerprints 28
5. EXPERIMENTS AND RESULTS 29
5.1. Datasets 29
5.2. Molecule Representation 31
5.2.1. Fingerprinting 31
5.2.2. SmilesVec Representation 31
5.2.3. Descriptor Vector Representation 32
5.3. Parameter Tuning 33
5.4. Results 34
5.4.1. Performance Measures 35
5.4.2. Chemical mutagenicity Results 35
5.4.3. Blood Brain Barrier Results 38
5.4.4. Cytotoxicity Results 42

5.4.5. Drug Induced Liver Injury Results 45
5.4.6. hERG Blockers Results 47
5.4.7. Human Liver Microsomal Stability Results 51
5.4.8. Mitochondrial Membrane Potential Distruption Results 53
5.4.9. Permeability Glycoprotein Inhibitors Results 56
5.4.10. Permeability Glycoprotein Substrates Results 59
6. DISCUSSION 64
7. CONCLUSION AND FUTURE WORK 68
REFERENCES 69

LIST OF FIGURES

Figure 3.1. An illustration KNN algorithm for k values of 4 and 12 8

Figure 3.2. Illustration of hyperplanes that separates 2 classes 8

Figure 3.3. Optimal hyperplane with maximum margin 9

Figure 3.4. Internal structure of random forests 10

Figure 4.1. A Molecular Substructure Fingerprint with predefined SMARTS . . 26

Figure 4.2. Hashed Fingerprints 27

Figure 5.1. SmilesVec Representation 32

Figure 5.2. Data split for train and test 33

Figure 5.3. Parameter tuning on training set, by 10 fold cross validation. 33

Figure 5.4. Experiments on test set 10 times, by using 90% of the training set each time. 34

Figure 5.5. Feature importances of custom descriptor set for AMES dataset . . 38

Figure 5.6. Feature importances of custom descriptor set for BBB dataset 41

Figure 5.7. Feature importances of custom descriptor set for Cytotoxicity dataset 44

Figure 5.8. Feature importances of custom descriptor set for DILI dataset . . . 47

Figure 5.9. Feature importances of custom descriptor set for hERG blockers dataset 50

Figure 5.10. Feature importances of custom descriptor set for HLM stability dataset 53

Figure 5.11. Feature importances of custom descriptor set for MMP distruption dataset 56

Figure 5.12. Feature importances of custom descriptor set for Pgp inhibitors dataset 59

Figure 5.13. Feature importances of custom descriptor set for Pgp substrates dataset 62

LIST OF TABLES

Table 5.1. Hyperparameters for AMES 36

Table 5.2. Ames mutagenicity Results (with standard deviations in parenthesis) 37

Table 5.3. Hyperparameters for BBB 39

Table 5.4. BBB Penetration Results (with standard deviations in parenthesis) 40

Table 5.5. Results for morgan fingerprints and updated morgan fingerprints (2 descriptors) added with Random Forest classifier 42

Table 5.6. Hyperparameters for Cytotoxicity 42

Table 5.7. Cytotoxicity Results (with standard deviations in parenthesis) 43

Table 5.8. Hyperparameters for DILI 45

Table 5.9. Drug Induced Liver Injury Results (with standard deviations in parenthesis) 46

Table 5.10. Hyperparameters for hERG 48

Table 5.11. Human Ether a-go-go Related Gene Blockers Results (with standard deviations in parenthesis) 49

Table 5.12. Results of Random Forest classifier with morgan fingerprints and updated morgan fingerprints (Balaban's J index descriptor added) . 51

Table 5.13. Hyperparameters for HLM 51

Table 5.14. Human Liver Microsomal Stability Results (with standard deviations in parenthesis) 52

Table 5.15. Hyperparameters for MMP 54

Table 5.16. Mitochondrial Membrane Potential Distruption Results (with standard deviations in parenthesis) 55

Table 5.17. Hyperparameters for Pgp inhibitors 57

Table 5.18. Permeability Glycoprotein Inhibitors Results (with standard deviations in parenthesis) 58

Table 5.19. Hyperparameters for Pgp substrates 60

Table 5.20. Permeability Glycoprotein Substrates Results (with standard deviations in parenthesis) 61

Table 5.21. Results of Random Forest classifier with morgan fingerprints and updated morgan fingerprints (diameter descriptor is added) 63

Table 6.1. Comparison of VNN and Random Forest classifier using morgan fingerprints (AMES, Cytotoxicity, DILI, HLM, MMP, Pgp inhibitors) or updated morgan fingerprint (BBB, hERG, Pgp substrates) representation 64

Table 6.2. Top 3 most important features for each dataset 65

LIST OF SYMBOLS

B	Number of bonds in the molecule
C	Number if rings in the molecule
D	Diameter of molecule
$E(i)$	Eccentricity of atom i
I_G	Gini impurity
I_{PJ}	Petitjean index
J	Balaban's J index
M	Molecular weight
$P(C_i\|t)$	Probability of Class i given the condition t
R	Radius of molecule
$[R]$	Molar refractivity
x_k	kth training sample
x_q	Query sample
$V(i)$	Degree of Vertex
γ	Curliness of decision boundary
ξ^c	Eccentiric Connectivity Index
σ	Light density
σ_i	Vertex distance

LIST OF ACRONYMS/ABBREVIATIONS

0D	0 Dimensional
1D	1 Dimensional
2D	2 Dimensional
3D	3 Dimensional
ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
Apol	Atomic Polarizability
AROM	Number of Aromatic Rings
BBB	Blood Brain Barrier
Bpol	Bond Polarizability
DILI	Drug Induced Liver Injury
ECIndex	Eccentric Connectivity Index
ECFP	Extended Connectivity Fingerprint
FP	Fingerprint
HBA	Hydrogen Bond Acceptors
HBD	Hydrogen Bond Donors
hERG	Human Ether a-go-go Related Gene
HLM	Human Liver Microsomal Stability
kNN	K-Nearest Neighbor
LogP	Water-Octanol Partition Coefficient
MACCS	Molecular ACCess System
MMP	Mitochondrial Membrane Potential
MolMR	Molar Refractivity
MW	Molecular Weight
nAcid	Number of Acidic Groups
nB	Number of Boron Atoms
nBr	Number of Bromine Atoms
nBase	Number of Basic Groups
nC	Number of Carbon Atoms

nCl	Number of Chlorine Atoms
nF	Number of Fluorine Atoms
nH	Number of Hydrogen Atoms
nI	Number of Iodine Atoms
nN	Number of Nitrogen Atoms
nO	Number of Oxygen Atoms
nP	Number of Phosphorus Atoms
nS	Number of Sulfur Atoms
Pgp	Permeability Glycoprotein
QSAR	Quantitative Structure-Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RBF	Radial Basis Function
RF	Random Forest
ROTB	Number of Rotatable Bonds
SMARTS	A Language For Describing Molecular Patterns
SVM	Support Vector Machine
TPSA	Topological Polar Surface Area

1. INTRODUCTION

1.0.1. Drug Discovery

In pharmacology, a drug is defined as a chemical substance that is used for treatment, diagnosis or prevention of a particular disease, or for enhancing the well-being of the body. A drug molecule has the ability of binding to a target protein in the body. The scaffold of the molecule should be limited in a way that all the functional groups in the molecule only interact with the target protein. Otherwise due to unexpected bindings with unrelated structures may cause harmful side effects or toxicity.

Drug discovery is the process of identifying potential new medicine and it requires continuous development and effort of expert chemists and pharmacologists. Basically, drug discovery has 4 stages; research & development, preclinical studies, clinical trials and review & approval [1]. First stage of drug discovery is research and development. It starts way before the clinical trials and approval of drug. In R&D phase, initial step is to determine target body part, illness or protein. Then research is about looking for a molecule that actually aims to affect the target by screening up to 10s of thousands of different molecules. A set of candidate molecules are identified and may be modified to improve effects on target. This R&D phase takes 3 to 6 years.

Second stage in drug discovery is preclinical studies. In this phase, lead molecules are further examined with different types of surrounding molecules or cells. The efficiency and toxicity of the molecules are deeply searched before the human trialsç This phase is called in vitro study, latin for "in glass" study. Molecules that are successful in vitro studies, are further examined in in vivo studies ("in a living"). In vivo study stage, molecules are tested on animals. Standards require testing for toxicity on at least two mammals for a drug to continue its development in clinical trials. Second stage is approximately one year laboratory study.

Third stage of drug discovery is clinical trials, in which drug candidate molecules are tested by human participants. This stage itself has three phases. Phase 1 is for examining the metabolism and excretion of the drug as well as the side effects. Phase 2 is for testing the efficiency of drug candidate molecules on patients who are actually suffered from the disease that drug is designed to cure. Phase 3 is done on larger number of participants to increase the reliability of drug candidate. This stage takes 4 to 7 years.

Last stage of the drug discovery is review&approval which takes 1 to 2 years. It includes submission to regulating bodies and monitoring duration after it is released to the market. The time until the process comes to the approval and release is mostly for research and development and clinical trials. This process takes 10 to 15 years in total for a valid conclusion. Despite advances in understanding biological systems and technology, drug discovery is still a time consuming, expensive and inefficient process with low rate of new medicine discovery.

1.0.2. Motivation

An important question arises in drug discovery. Which properties of drugs make a difference that are discriminating drugs from other chemicals? Answer of the question is proposed by Chris Lipinski, defining the "drug-likeness" or "druggability" of a molecule as having acceptable ADMET (Absorption, Distribution, Metabolism, Excreation, Toxicity) properties that makes the molecule continue its development until the completion of phase 1 trails in clinical trials [2]. Recently, there have been a significant amount of increase in the number of compounds that are possible drugs. However, the number of drugs in marketplace has not been increased accordingly. That is because these drug candidates do not meet desired pharmacokinetic, in other words druggability properties. Another factor for drug candidates to fail is due to high toxicity.

About 2 decades ago, 50% of drug candidates failed to fulfil desired properties and removed from the market or never proposed. But for pharmaceutical companies,

success depends on canalizing finite resources to the project that have the potential and most of the times costly failures cannot be tolerated. Hence, early assessments of compounds' properties related to absorption, distribution, metabolism, excretion and toxicity (ADMET) have become very significant to be able to avoid costly failures [3]. Clinical assessment of ADMET related properties is costly and time consuming , that is why computational methods for predicting ADMET related properties have gained importance over time.

1.0.3. Our Contribution

In this study, we examine 9 ADMET properties namely, chemical mutagenicity (AMES), blood-brain barrier penetration (BBB), cytotoxicity, drug-induced liver injury (DILI), the human ether a-go-go related gene (hERG) blockers, human liver microsomal (HLM) stability, mitochondrial toxicity (MMP distruption), pgp (permeability glycoprotein) substrates and pgp inhibitors. Prediction is done by using commonly used machine learning methods k-nearest neighbors (kNN), support vector machine classification (SVMC) and random forest (RF). Our initial aim is to train our models using fingerprints, namely topological and morgan fingerprints which are commonly used molecular representations in chemoinformatics and compare our results with other works. Secondly, we propose a custom set of descriptors and compare our scores with the others. Here, our main contribution is for each ADMET property, we extract most important molecular descriptor(s) that can be used for enhancing the feature vectors, hence enhancing the prediction scores.

1.0.4. Thesis Outline

The outline of thesis is as follows. In chapter 2, we overview the related work in the literature. In chapter 3 machine learning algorithms that are used for ADMET prediction are explained. In chapter 4, chemical druggability features are overviewed. In the first subsection, 9 ADMET properties that we examined for prediction are discussed. In the second subsection, chemical descriptors which are used to represent

molecules are described. We looked for special relationships between these chemical descriptor and ADMET properties. In the third subsection, chemical fingerprints that are also used for representation of molecules are described in detail. In chapter 5, our experimental setups for each dataset and results for all of the classifiers using different molecular representations are explained. In chapter 6, we discuss some aspects of the results. In chapter 7, a conclusion is made and future work directions are outlined.

2. RELATED WORK

In traditional drug discovery processes, main purpose was to identify new structures that show good pharmacological properties. After that phase, the structures were examined for other physicochemical properties such as solubility, toxicity, permeability, excretion, distribution, metabolism, absorption (ADMET properties) one by one. This process is time consuming and inefficient for reaching the optimal drug molecule since the pipeline of examinations is one after another. Now, the multiple examinations on a drug candidate molecule at the same time has been receiving a great interest because all of the properties contribute to the final decision for proceeding the development of the drug molecule.

For data modelling, quantitative structure-activity relationship (QSAR) and quantitative structure–property relationship (QSPR) approaches are used for prediction problems in pharmacology. With these data modellings, an effort to find a correlation between a structure and a property is performed [4]. With these molecular representations, various techniques like multiple linear regression and partial least squares, as well as the machine learning algorithms such as k nearest neighbors, decision tree, support vector machine for classification and regression and random forest are used to predict ADMET properties.

There are many QSAR studies in the literature. In Chembench [5] which is an online tool, they allow users for creating and validating QSAR models, using MACCS keys and different ML algorithms and some visualization and virtual screening functionalities are provided. A slightly different approach was made by Schyman et al [6]. In their study for ADMET prediction they propose that a variable nearest neighbor algorithm in which for all nearest neighbors there is a threshold similarity constraint. They suggest, with this approach, the model is allowed to make predictions based on molecules only similar to query instance, and building the model is fast. The molecule representation used in vNN is extended connectivity fingerprints (ECFP). We compare

our results with their scores in chapter 5.

AdmetSAR [7] [8] is yet another web tool for predicting ADMET properties. Initial version makes prediction with MACCS keys using only SVM. In the last version, they implemented models selecting the best results obtained using 6 different types of molecular fingerprints, for SVM classification and regression, kNN and RF. System is comprehensive with 47 models for prediction. SwissADME [9] is another web based tool for ADMET prediction which calculates a set of physicochemical descriptors and a small set of pharmacokinetic properties, namely BBB, Pgp substrates and Cyp inhibitors.

There are many other studies for prediction of pharmacokinetic properties of molecules, some of online tools are OCHEM [10], FAF-Drugs [11], ADME-AP [12], PK/DB [13] and lazar [14]. However there are limitations about chemical space coverage of these tools which are mostly constructed on small datasets. Main purpose of these tools are to give as much information as they can provide for a query molecule, including physical properties such as hydrogen bonding information, molecular weight, logP etc. that can be calculated using molecular descriptors libraries, some visualization functionalities, search according to name, id or similarity and many more.

On the other hand, in this study, our main perspective is different than ADMET predictors above. We try to find relations between physicochemical properties of the molecule with the ADMET property, while predicting ADMET property of a molecule with a high reliability score. Our findings about relations can be used for further generalizations. Instead of using bulky vector representations of molecules, a set of most related descriptors can represent the molecule more efficiently and modelling onto these small but efficient vectors make it faster to train and build the system.

3. BACKGROUND

3.1. Machine Learning Methods

In this chapter, machine learning methods that are used to predict druggability properties of molecules are explained.

3.1.1. K-Nearest Neighbor Algorithm

KNN is a supervised machine learning algorithm in which labeled data instances are represented in a high dimensional feature space. It is used for classification problems as well as regression. When an unlabeled query data point is considered, algorithm assumes that the nearest data samples are most similar to the query. By majority voting, label for the query sample is decided by the 'k' nearest data points.

Given a query sample x_q, Mitchell [15] described the algorithm:

- For x_1, x_2, ... x_k that are closest k training samples to x_q,
- Return $\hat{f}(x_q) = \underset{v \in V}{\operatorname{argmax}} \sum_{i=1}^{k} \delta(v, f(x_i))$

where $\delta(A,B) = 1$ if $a = b$ where $\delta(A,B) = 0$ otherwise.

It is a straight forward, easy to implement algorithm which does not require additional parameters to tune. One important aspect of kNN algorithm is to specify 'k' value because too high or too low k values may cause high rates of misclassification. Small values of k causes the results of the algorithm to be more unstable whereas large values of k increases the stability of the algorithm due to majority voting, may increase error rate of classification. Hence, when deciding the value of k, the point where the validation error is minimized is chosen.

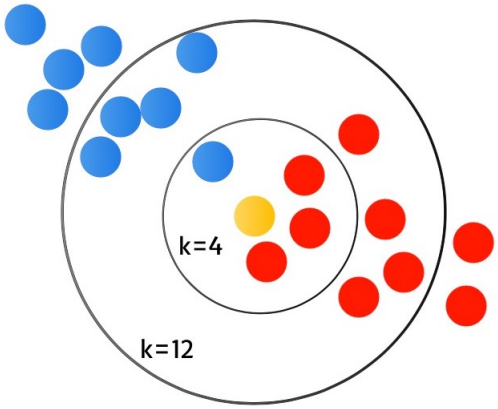

Figure 3.1. An illustration KNN algorithm for k values of 4 and 12

In this work, one of the machine learning algorithms that we use to predict the druggability features of drug-like molecules is kNN. kNN parameters and results of each dataset are explained in the results section.

3.1.2. Support Vector Machine (SVM) Classifier

Support vector machine is a machine learning algorithm proposed by Vapnik et al. in 1995 [16]. Its main objective is to find a hyperplane in an N-dimensional vector space, that discriminates two classes of elements.

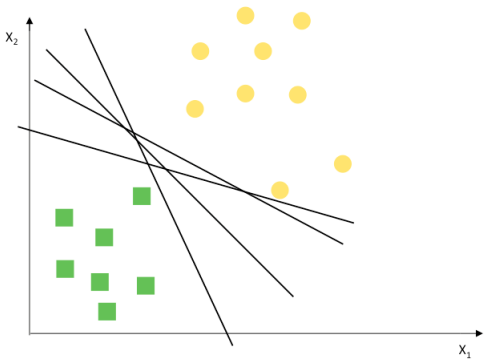

Figure 3.2. Illustration of hyperplanes that separates 2 classes

There may be more than one hyperplanes that separates two classes, in that case the one with the maximum margin distance is chosen because large margin strengthen the prediction for test samples. Points that are closest to the decision boundary are

called support vectors and these vectors affect the position of the decision boundary.

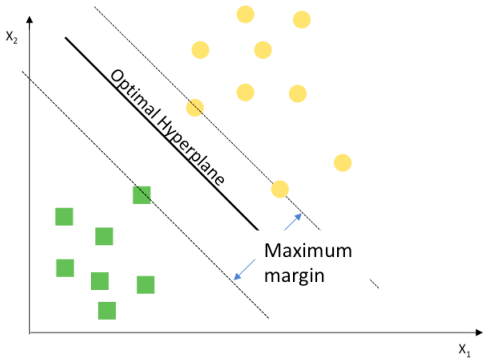

Figure 3.3. Optimal hyperplane with maximum margin

In SVM, learning of the boundary plane is done by transforming the problem using different types of kernels. In this work, we used different settings of linear, polynomial and radial basis functions (RBF) kernels, which can be expressed as follows. Linear kernel function is expressed as the dot product of two vectors:

$$K_{linear}(x, y) = x^t y + c \tag{3.1}$$

Polynomial kernel is defined in a way that it distinguishes nonlinear and curved input spaces. d is the degree of the polynomial.

$$K_{polynomial}(x, y) = (\alpha x^t y + c)^d \tag{3.2}$$

Radial basis function is another popularly used SVM function which can express data in an infinite dimensional vector space. The two main parameters for RBF kernel are C and gamma. C is the penalty for misclassification and gamma can be thought the measure of how curvy is the decision boundary. Its value is between 0 and 1. For low values of gamma, decision boundary is not very curvy, for the values closer to one decision boundary is more like a curve. For very large values of gamma, algorithm

overfits.

$$K_{rbf}(x,y) = exp(-\gamma \|x-y\|^2) \tag{3.3}$$

In this work, we use SVM classifier for druggability property prediction, for each dataset we examine different kernels and use which fits best to the dataset. We explain results of SVM for each database in chapter 5.

3.1.3. Random Forest

Random forest is an ensemble learning algorithm in which a number of classifiers are used to get a final decision for the test prediction. It can be thought as a set of decision trees. The aim of using many classifiers is to increase the accuracy. Every classifier makes a decision about the test sample and by majority voting [17], the final decision is declared. In random forest algorithm which can be visualized as a bunch of decision trees bundled together, bagging ensemble method is used [18] in which every classifier is built by a randomly drawn set of training data and gets equal vote for labelling unlabeled data.

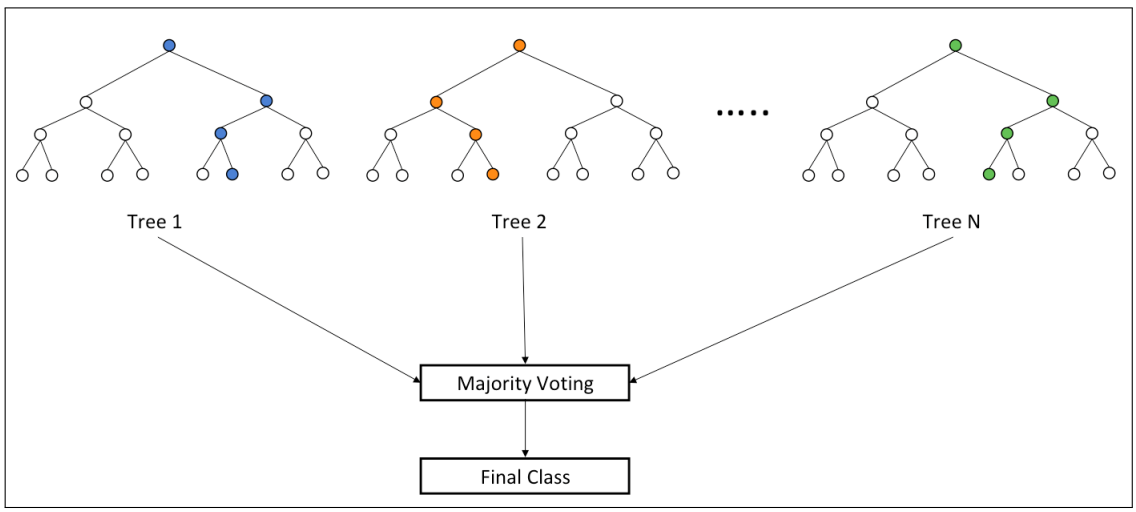

Figure 3.4. Internal structure of random forests

Random forest algorithm also uses randomization when deciding to split on the

node that is most appropriate. It is simply the $\sqrt{F}$ where F is number of features in dataset. Another randomization is used when deciding on which feature to split the node. For that, gini impurity defined as follows is used:

$$I_G(t) = 1 - \sum_{i=1}^{N} P(C_i|t)^2 \tag{3.4}$$

where t is a condition, N the number of classes in the data set, and C_i is the ith class label in the data set [19]. At each node, algorithm looks for the feature that reduces the gini impurity the most.

Feature Importance : For a random forest, feature importances indicate the sum of all gini impurity reduction over all the nodes that split on that feature. It displays the features that play a significant role on decisions. Since we interpret the gini impurity of a node as the probability of a randomly chosen sample in the node is misclassified, we draw the conclusion that the most reductive feature of gini impurity contributes to true classification the most.

RF is another classification method that we use in this work. Most importantly, we draw a set of specific descriptors that are quite contributive to unique representation of that particular dataset. We discuss most important features of each dataset and results in last chapter.

4. CHEMICAL DRUGGABILITY FEATURES

4.1. ADMET Properties

In drug discovery, for a drug to be successful, it must reach its therapeutic target in the body. Pharmacodynamics analyzes effects of the drug on the body, in other words the toxicology and the activity of the drugs. The main question in pharmacodynamics is that "What does the drug do to body?" On the other hand, pharmocokinetics examines the question the opposite way: "What does the body do to the drug?" The effects of the body specifies transportation and chemical transformation of the drug. In other words, the response of the body determines absorbtion, distribution, metabolism and excretion of the drug molecule. Therefore, the candidate drug molecule must have high quality of pharmacodynamic and pharmacokinetic properties (ADMET), to be able to reach to the target body part and have a desired chemical interaction with the target.

In this section, we explain a set of ADMET related properties that we chose to study on.

4.1.1. Chemical Mutagenicity (AMES)

In toxicology, mutagenicity is a very important end point because mutagenic substances are often carcinogen, therefore cause to cancer. In literature, widely employed AMES test [20], after Dr. Bruce Ames, is done using bacteria strains, to examine whether the given chemical can damage DNA and cause mutations. A positive result from the test indicates that the given substance is mutagenic.

In drug discovery, a positive AMES test result for a drug candidate molecule can end the development of that particular molecule as drug because possible mutagenicity of a molecule would be a serious problem in later phases of drug discovery. However, the test needs a laboratory setup and time, making the computational pre-

diction of AMES mutagenicity an attractive subject. In this study, we try to predict AMES mutagenicity of molecules, learning from a set of molecules which took the test before and make a prediction for new samples.

4.1.2. Blood Brain Barrier

Blood Brain Barrier (BBB) is one of the most important mechanism that protects central nervous system from circulatory system agents that can be harmful for neural functioning of the brain. [21] It is a highly selective barrier, allowing only a few molecules to penetrate into brain; like, glucose, water, amino acids and water-soluble lipid molecules. Nearly all of the large molecules cannot get passage through it.

Neuropharmaceutics is one of the largest components of pharmaceutical industry and has a big potential to grow [22]. The main problem in neuropharmaceutics is to come up with a candidate drug which can penetrate into this highly selective barrier. However most of the drug candidate molecules cannot pass through BBB [21].That's why it has a great importance to know a designed molecule is BBB permeable or not.

The most common numeric value of BBB permeability is defined as follows:

$$\log BB = \log \frac{c_{brain}}{c_{blood}} \tag{4.1}$$

where c_{brain} is the concentration of chemical in brain and c_{blood} is the concentration of chemical in blood. Because the experimental calculation of BBB permeability value is high-cost and time-consuming process, we have a limited number of BBB permeability value records in literature [23]. That makes computational prediction of BBB permeability value even more important. In this study, we also examined a set of BBB-permameable and BBB-nonpermeable molecules, to predict the BBB permeability of a new test sample.

4.1.3. Cytotoxicity (HepG2)

Cytotoxicity is one of the most important assays in drug discovery. It means the ability of a chemical to destroy living cells. To be able to accurately screen cytotoxicity of drug candidates is very useful to detect compounds that can be risky for the human health.

Understanding cytotoxicity assay is also very crutial for designing anti-cancer drugs. By determining the cytotoxicity of cancer cells or any other cell causing certain kind of diseases, anti-cancer medications can be designed in a way that the problematic cells can be blocked from cell division or getting enough nutrition to grow, eventually die. In this study, we examined the cytotoxicity assays of HepG2 cells which is known as human liver cancer cell line.

4.1.4. Drug-Induced Liver Injury

Drug-induced liver injury also known as drug-induced hepatoxicity is a disease related to prescripted drugs or dietary supplements which cause damage to the human livers that may be fatal, ending up in the need of liver transplanation. Studies showed that DILI is the main cause of drug withdrawal from the market and drug failure in clinical trial phase of drug discovery [24].

Predicting hepatoxicity of a drug candidate is very desirable, to be able to avoid costly failures in clinical trials. However, laboratory experiments assessing the hepatoxicity is challenging and every experiment are not necessarily conclusive. Some experiments done on animals showed that DILI causing drugs do no always cause the same disease on animals [6]. That is why we examined DILI and made predictions for new substances.

4.1.5. hERG Blockers

The human ether a-go-go related gene (hERG) is responsible for production of a potassium ion channel that provides a healthy repolarization of hearth. Blocking the functionality of hERG can be fatal, leading to arrhythmia, fainting, drowning or death [25]. Recently, there have been a considerable number of drugs, that were withdrawed from the market due to dangerous cardiotoxicity effects [26]. Hence, assessment of hERG blockers in early stage of drug development is very crutial for any drug candidate.

4.1.6. Human Liver Microsomal Stability

In pharmaceutical researches, it is very important for a drug to have metabolic stability in the body. Due to high metabolizing functionality of the human liver, most chemicals are cleared out by the liver very rapidly due to functioning of hepatic enzymes. For a drug to reach its adequate density in the circulatory system and reach its therapeutic target with the acceptable amount of daily dosage, it should not be metabolized by the liver very quickly. Therefore, in the drug discovery phases, researchers use human liver microsomal stability assay to identify molecules with low metabolic stability [6].

In the literature, there are many recordings of HLM assay of the molecules. Making use of these datasets in the literature, to be able to extract molecules with low metabolic stability before the actual syntasis stage in drug discovery is of a considerable interest. Thus, in our study, we also examined and predict HLM assay of molecules.

4.1.7. Mitochondrial Membrane Potential Disruption

Mitochondria is one of the most important organelle in an organism that produce cellular energy. Chemicals including environmental polluters and prescribed drugs

may cause mitochondrial toxicity which eventually leads mitochondria dysfunction. Many diseases including diabetes, cancer, Alzheimer's disease, Parkinson's disease and heart-related diseases are associated with dysfunction of mitochondria, to some extend. Hence, to able to protect people from toxicants that harm mitochondria and avoid costly failures in drug discovery, it is significant to predict mitochondrial toxicity levels of drug candidate molecules.

To be able to predict mitochondrial toxicity, mitochondrial membrane potential disruption assay is used for monitoring the chemicals because toxicants that affect mitochondria is likely to decrease mitochondial membrane potential. We examined a dataset consisting of MMP distruption values of molecules.

4.1.8. Pgp Substrates and Inhibitors

Permeability glycoprotein is a cell membrane protein that is responsible for extracting foreign substances from the cell or keeping them outside. Although, it is a protection over toxicants for the tissues, P-glycoprotein may pump the drugs into the lumen, decreasing their absorbtion in the body cells, hence decreasing bioavailability [27]. In the case of cancer cells, the over functionality of P-glycoprotein may prevent anticancer drugs' activity, bringing tumor cells a kind of resistancy. This phenomenon is known as multidrug resistance [28].

Another aspect of the functionality of P-glycoprotein is that it prevents permeability of drugs from certain barriers like BBB, as a defense mechanism for toxic chemicals targeting central nervous system [29]. Therefore, the pharmacodynamics of P-glycoprotein is very essential for drug discovery process. To be able to predict P-gp substrates and inhibitors, makes a great advance for a drug to continue its development and to be served into the market. That is why we also examine Pgp inhibitors and substances in our study.

4.2. Chemical Descriptors

In chemistry, it is an important matter to represent molecules by a a set of mathematical characteristics. These representation should reflect natural formation of atoms forming the molecule, including neighborhood-induced properties and relative arrangement in space. Chemical descriptor is defined as a mathematical method that makes use of chemical information in the symbolic representation of molecule into a numeric value, or an experimental value related to that molecule [30].

Molecule descriptors include values obtained from both experimental measurements and theoretical calculations. Descriptors from experimental measurements include values such as Log P, molar refractivity.

Theoretically calculated descriptors comprise 0D,1D,2D, 3D, 4D-descriptors. 0D-descriptors often refer to descriptors obtained from solely chemical formula of the molecule which does not carry much information about atom connections [31]. It includes counts of atoms, bonds and their constitution, summation or average of atomic properties, molecular weight [32]. 1D-descriptors are derived from structural fragments and functional groups included in the molecule. Fragment count, hydrogen bond donors and acceptors, molecular fingerprints (ie. Extended connectivity fingerprints, Morgan Fingerprints, MACCS, SMARTS) are type of 1D-descriptors [32].

2D-descriptors reflects topological structure of the molecule. It makes use of 2 dimensional molecular graph of the molecule which encodes atoms and their connectivity. Some examples of 2D-descriptors are Balaban J, Randic, kappa. 3D-descriptors are derived from 3 dimensional structure of molecules in which every atom is linked to a set of 3 dimensional coordinates, determining its own position in space. 4D-descriptors comprise 3D coordinates and additional information related to molecule interactions and electron distribution [31].

In coming sections, we will be explaining the set of molecular descriptors that we

choose to study on predicting ADMET properties.

4.2.1. Molecular Weight

Molecular weight is a commonly used 0D-descriptor that is used for indicating general size of the molecule. It is important for predicting whether a specific molecule can penetrate into certain types of barriers in the human body because nearly 100% of large molecules cannot get passage from highly selective barriers. For example, in the literature, Levin et al. has suggested that molecular weight threshold should be 400 for BBB penetration [33].

Although molecular weight performs well for discriminating large molecules, it is a high degeneracy descriptor, meaning its value can be same for different molecules. It is not very contributive to the uniqueness of the feature representation of molecule, but a good measure for filtering when combined with other descriptors. In their 'rule of 5', Lipinski et al. suggested that chemicals with molecular weight more than 500 are not likely to be absorbed by therapeutic target and to penetrate into certain living barriers in the body [34].

4.2.2. The Water Octanal Partition Coefficient

The water octanal partition coefficient is the proportion of dissolution of a molecule in organic solvent and water. The logarithm of water-octanol partition coefficient is defined as:

$$\log P = \log \frac{Co}{Cw} \tag{4.2}$$

A positive value of logP indicates that the molecule is more lipophilic, meaning mostly dissolved in lipids, lipid liking. A negative value for logP indicates that the molecule is more hydrophilic, meaning mostly dissolved in water, water liking. 0 value of logP means that the substance is dissolved in lipid and water equally [35].

The water octanal Partition coefficient is an important criterion for understanding physical nature of the molecule in different environments. It provides a foresight about dissolvement of molecule in given body part, indicating whether it is absorbed by the tissues or dispersed by the water.

A drug candidate molecule must have an optimal hydrophilicity and lipophilicity. Because very high hydrophilicity means that the molecule can be driven away by the water and cannot be absorbed enough by the tissues whereas very high lipophilicity means that the molecule can be trapped into the tissues, cannot be excreted, leaving a high chance of toxicity formulation in the body. Therefore, in drug discovery, candidate drug molecules are screened according to their logP values, to understand ADMET related behaviour of that molecule in the body. In their 'rule of 5', Lipinski et al. suggested that chemicals with calculated logp value greater than 5, have poor absorbtion and permeation characteristics [34]. In another study, Chen et al. have suggested that molecules having lipophilicity of logp greater than or equal to 3, show strong correlation with severe DILI [36].

4.2.3. Topological Polar Surface Area

Topological polar surface area is defined as the surface area that belongs to polar atoms like hydrogens, nitrogens, oxygens. It is mostly used as an indicator for molecular transport through biological barriers in the body. Therefore, for blood brain barrier penetration, TSPA value is a good feature to determine whether a certain molecule can penetrate into central nervous system.

In the literature, there have been experimental findings about correlation between penetration into living membranes and TPSA value. Palm et al. found out using Caco-2 cells that molecules with 60 $Å^2$ or less TPSA can penetrate into Caco-2 cells, on the other hand molecules with 140 $Å^2$ or more TSPA cannot [37]. For drugs targeted to central nervous system, it has been shown that only molecules with less than 60-70 $Å^2$ can be transported through the BBB [38]. Veber et al. conducted a study

on rats and concluded that TPSA of 140 $Å^2$ or less tends to have higher probability of bioavailability, reducing the TPSA also improves the permeability of the drug [39]. Hence, it will be a distinctive feature while prediciting the chemicals' ADMET related properties.

4.2.4. Number of Hydrogen Bond Donors and Acceptors

The hydrogen bond is an electrostatic attraction between a hydrogen bond donor and hydrogen bond acceptor in the same molecule or between molecules. Common hydrogen bond acceptors are Nsp^3 - amines, Nsp^2 - imines, pyridines, Osp^3 - alcohols, ethers, water, Nsp - nitriles, Osp^2 - amides, ureas, esters, ketones, Florine. Common hydrogen bond donors are water, alcohols, phenols, carboxylic acid, organophosphoric acid, mmonium cations, imidazoles [40].

Hydrogen bonds are very important for many biological functioning. For example, water shows unusual properties, like its unusual density in different molecular states. HB is also important in the protein structure and DNA base pairing, deciding the possible bond formation between drug candidate and therapeutic target.

Some studies showed that a difference created by only a hydrogen bond can alter the potency of a drug candidate molecule [41]. Lipinski et al. in their 'rule of 5', have showed that molecules having more than 5 hydrogen bond donors and 10 hydrogen bond acceptor are more likely to have poor absorbtion and permeation characteristics [34]. Therefore, we added HBD and HBA to our feature set.

4.2.5. Number of Rotatable Bonds

Rotatable bond is defined as any single bond attached to a non-terminal, non-hydrogen atom that is not ring. The number of rotatable bonds in a molecule affects the binding potency and bioavailability because while binding to another molecule, more rigid molecules tend to create less entropy loss than flexible molecules that have

rotatable bonds. Veber et al. suggested that as the number of rotatable bonds in a molecule increases, permeation rate decreases. Molecules with less than or equal to 10 rotatble bonds tend to have higher bioavailability [39]. We added number of rotatable bonds into our feature set due to its reflective characteristics of ADMET properties.

4.2.6. Molar Refractivity

According to Lorentz-Lorentz formula, molar refractivity is defined as:

$$[R] = \frac{n^2 - 1}{n^2 + 2} * \frac{M}{\sigma} \tag{4.3}$$

where M is molecular weight, n is refraction index and σ is the density related to light used to measure refractivity index.

Molar refractivity can be taken as the real volume of the molecule, as well as the London dispersive forces effective on drug-target interactions. [42] It reflects polarizability of the molecule and it is a characteristic for the molecule. In their work of characterization of known drug databases, Ghose et al. found out that the range for molar refractivity of known drugs is between 40 and 130 with average 97, [43] demonstrating molar refractivity is a characterizing feature for ADMET prediction of drug-like molecules.

4.2.7. Balaban's J Value

Balaban's J is a topological index value proposed by Alexandru Balaban in 1982 [44]. It is defined as:

$$J = \frac{B}{C+1} \sum_b (\sigma_i \sigma_j)_b^{\frac{-1}{2}} \tag{4.4}$$

where σ_i and σ_j are the vertex distance degree of adjacent atoms, summing over all the molecular bonds b, C is the number of rings and B is the number of bonds in the

molecule.

Topological indices are one of commonly used methods for converting molecular characteristics into a numerical value. Difference between the Balaban's J index and the other topological indices is that it calculates distance sum of each vertex in molecular graph by averaging the distance by the number of edges, rather than just adding the distances. By averaging the sum, Balaban's J index has the lower degeneracy compared to other topological indexes. For example, 5 different C_6 isomers yield different Balaban's J values. It is a very well discriminating single topological index, that's why we preferred to add it in our feature set.

4.2.8. Number of Valence Electrons

In an atom, valence electron is the farthest electron from the nucleus and it is more likely to have a bond formation or ionization. It is with the highest principle quantum number in the atom. In a covalent bond, one valence electron coming from both atoms form a pair. We added number of valence electrons in our feature set because probability of drug to bind to the therapeutic target is associated with number of valence electrons in the molecule.

4.2.9. Number of Aromatic Rings

Aromaticity of a molecule indicates that the molecule is cyclic, flat and the ring has resonance bonds which contribute to the stability of molecule. If an aromatic ring contains Nitrogen, it is called heteroaromatic rings. The term number of aromatic rings refer to both heteroaromatic rings and benzenoid aromatic rings.

On a study conducted on number of aromatic rings and ADMET related properties, Ritchie et al. suggested that for oral administration, the less number of aromatic rings decreases the risk of attrition rate of related drug candidate. Molecules having more than 3 aromatic rings are likely to fail in laboratory experiments [45]. Due

to high probability of increasing the total success of the molecule representation, we added number of aromatic rings to our feature set.

4.2.10. Atom Numbers

In our feature set, we added total number of atoms, number of H, B, C, N, O, S, P, F, Cl, Br and I atoms. Number of all atoms gives information about the molecular size whereas number of atom types reflects certain chemical interactions and bondings. These native features are suitable for predictions which are constructing molecule-property relations.

4.2.11. Acidic and Basic Group Counts

In drug discovery, acid-base properties of molecules are important because different pH values alter the charge state of molecules, affecting ADMET properties of that molecule. Acid-base properties of molecules are also determinants of ionisation constants, pK_a values. Ionization constant is measured by taking all functional groups into account and plays crucial role on solubility and lipophilicity [46]. That is why number of acidic and basic groups is another two descriptors in our feature set.

4.2.12. Eccentric Connectivity Index

A molecular graph is an important entity which reflects the topological structure of molecule, representing atoms and their connections to other atoms. Main concern in molecular graph theory is to derive structural characterization and ordering, eventually leading to decisions on what to put first in the sequence. Using these graph related properties, many mathematical formulas for topological indices were derived.

Eccentric connectivity index is another topological index that is used for a fast and cost effective drug design monitoring systems in pharmaceutical industry. It uses

eccentricity and valency of each atom in a molecular graph, and defined as:

$$\xi^c = \sum_{i=1}^{n} E(i)V(i) \tag{4.5}$$

where $E(i)$ is the eccentricity and $V(i)$ is the degree of vertex.

In their study, Sharma et al. [47] examined the correlations between ECI and properties ranging from physical and biological, gained correlation coeffients between 95% to 99%, much higher than Wiener index' results. They also concluded that for biological properties ECI has high correlation ability. Therefore, we added eccentric connectivity index to our feature set as another molecular graph descriptors.

4.2.13. Molecular Diameter and Radius

Molecular size plays a key role for the permeability, and ADMET properties. Among other size related descriptors like, atom number, surface area, volume etc., molecular diameter and radius are also used for representing molecular size in general.

In the literature, there are some work explaining some limits to diameter and radius for a molecule to permeate certain barriers. In their work on Caco-2 cells, Knipp et al. suggested that permeability decreases when the molecular radius increases [48]. Similar conclusions were also made by Hou et al. concluding that Caco-2 layer permeability rate tends to decrease for the molecules with larger radius than the ones with smaller radius [49]. With the molecular weight, we added molecular diameter and radius descriptors to our feature set, for a better representation of molecular size.

4.2.14. Petitjean Index

Petitjean index is a topological index proposed by Michel Petitjean in 1992 [50]. Petitjean defined the eccentricity of an atom as the longest path between that atom

and any other atom in the molecule, and the radius (R) is the lowest eccentricity in the molecule whereas the diameter (D) is longest. It is defined as:

$$I_{PJ} = \frac{D - R}{R} \tag{4.6}$$

Petitjean index is a good way of geometrical shapes and graph of molecules. While avoiding large complexities, index catches shape characteristics. Therefore we added petitjean index in our feature set.

4.2.15. Atomic and Bond Polarizability

Molecular polarizability is an important feature characterizing molecular behavior when it is interacted with an external sourced electric field. This behavior helps to model molecule's biological activities and internal properties [**?**]. In the literature, Breindl et al. investigated the correlation between a set of descriptors and indicated that logP values of molecules are affected by there descriptors and one of them was polarizability. Moreover, they displayed a linear dependency between polarizability and logP [51]. We added polarizability descriptors in our feature set to strengthen the ability to represent characteristics about solubility.

4.3. Chemical Fingerprints

Chemical fingerprints are commonly used for drug discovery virtual screening application, ADMET prediction, database searching and analysis. They represent the molecules in a high-dimensional vector space, each element of the vectors corresponds to the 2D, 3D properties of that molecule. Fingerprints are encoding of molecules, dividing the molecule into a large number of fragments. Presence or absence of every fragment corresponds to a bit in the fingerprint representation.

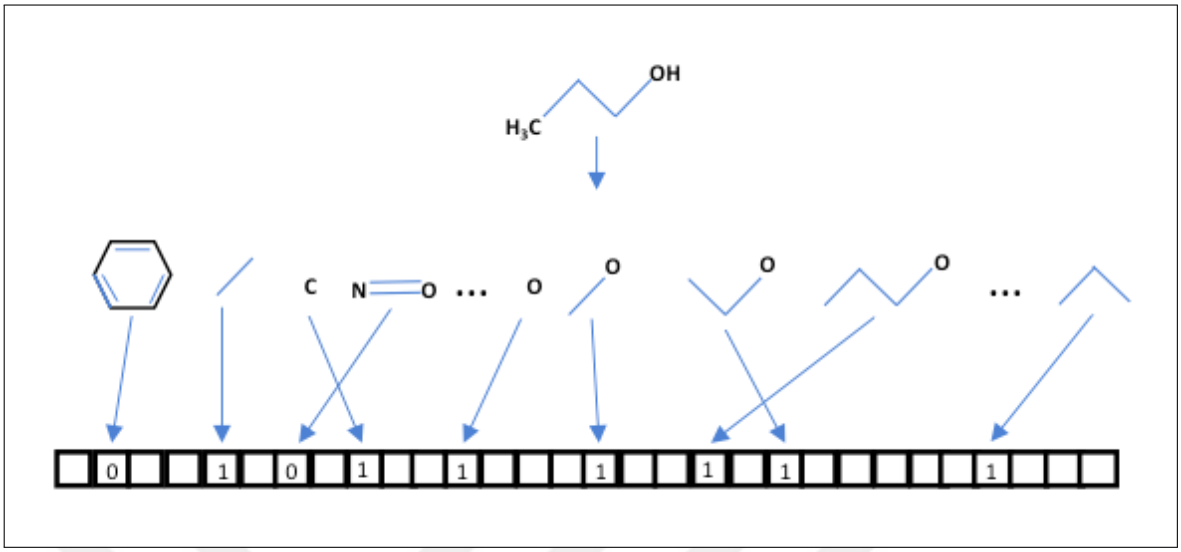

Figure 4.1. A Molecular Substructure Fingerprint with predefined SMARTS

In SMARTS representation, a predefined substructure pattern list constitutes each bit of the fingerprint, if particular substructure is in the molecule, than corresponding bit in the fingerprint is set to 1, otherwise 0.

Another fingerprinting approach is based on hashing.(Figure 4.2) Sometimes, number of all possible fragments, generated by obtaining all atom pairs, augmented atoms, all sequences of atoms in the sequence between 2-7, is huge. That huge number of elements in the fingerprint creates computational difficulties, taking into the sparseness and the size of the fingerprints. With hashing, fingerprint size is shrieked down to a fixed number of bits. (i.e 2048 bits)

Now, we will explain topological and Morgan fingerprints for ADMET related property prediction that we used in this work.

4.3.1. Topological Fingerprints

For topological fingerprints, we used RDKit's fingerprint representation. It is a fingerprint based on hashing on substructures, like the one illustrated in figure 4.2. It is yet another implementation of Daylight-like fingerprint [52]. Subgraphs in Daylight fingerprints includes;

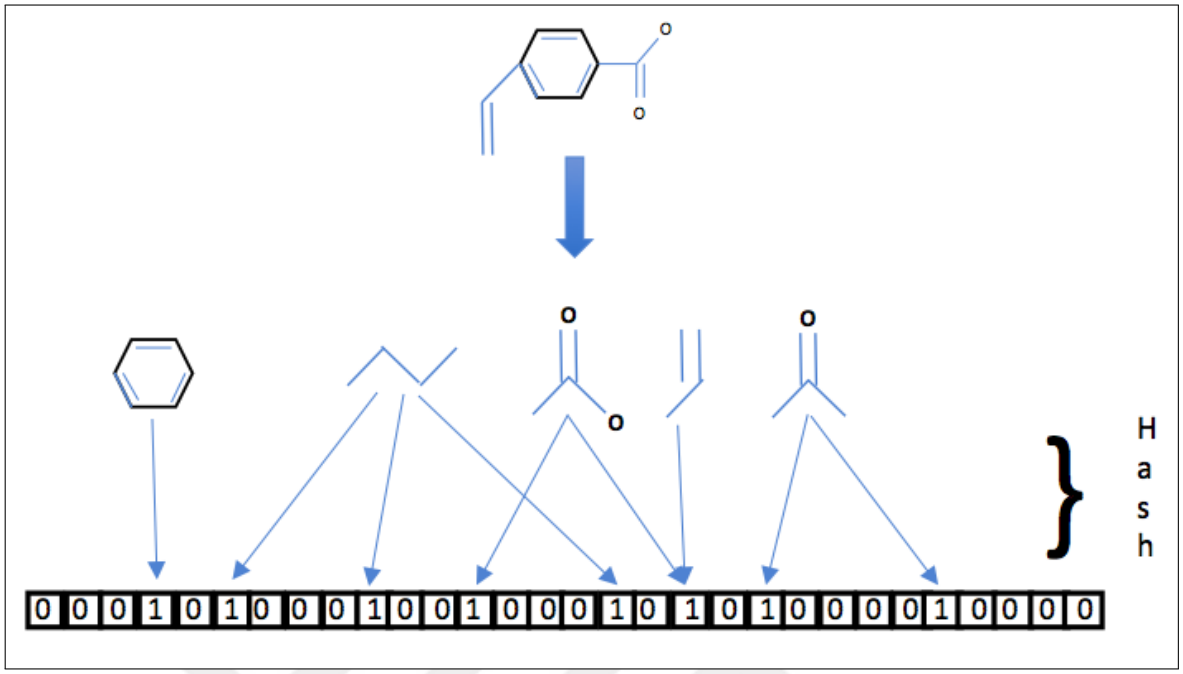

Figure 4.2. Hashed Fingerprints

- a pattern for each atom
- a pattern for each atom and its neighbor (1-bond)
- a pattern for each group of atoms and bonds connected by paths up to 2 bonds, 3 bonds, 4 bonds, ... (number of bonds can be up to 31)

For example "OC=CN" molecule includes:

- 0-bond paths: O, C, N
- 1-bond paths: OC, C=C, CN
- 2-bond paths: OC=N, C=CN
- 3-bond paths:OC=CN

The topological fingerprinting algorithm of RDKit does the following [53]:

- For all subgraphs in the molecule, generate a hash to generate a raw bit ID (Hashing the subgraph includes hashing *individual bonds* based on atom types (include atomic numbers modded by 128 and aromacity), *the degrees of two atoms* in the bond and *the bond type*)

- Seed random number generator with hash
- Generate random numbers between 0 and fingerprint size and set corresponding bit.

4.3.2. Morgan Fingerprints

Morgan fingerprint is a circular kind of fingerprint built on Morgan algorithm [54]. Morgan algorithm is based on the idea of descriminating isomorphism. The algorithm works as follows: On each atom, an integer number is assigned and for each integer an identifier is encoded. Until each atom gets a different (unique) identifier, algorithm iteratively calculates new identifiers using the the identifier encoded in the previous step. Intermediate identifiers are discarded.

Morgan fingerprints [55], also known as extended connectivity fingerprints, are calculated similar to Morgan algorithm except iteration steps take time for a pre-specified number of times. Instead of discarding intermediate calculations, algorithm collects them into the set that actually produce fingerprint itself. After iterations fingerprint is hashed into a fixed size of lenght using the set consisting of precollected identifiers. These identifiers can be interpreted as 'on' bits of a large fingerprint which is to be hashed into a fixed size.

Morgan fingerprints are useful for representing molecules substructural patterns however sometimes bit collisions occur, allowing more than one structure to affect same bit. Still, it is a very powerful technique for molecular level prediction which we will demonstrate the results of morgan fingerprints using different machine learning algorithms in the next chapter.

5. EXPERIMENTS AND RESULTS

In this chapter, we will be explaining our datasets and molecule representation choices for each run on each dataset. Then, we will explain parameter tuning for each dataset, for each classification method. Lastly, we will show results for each dataset for different representation and classification methods.

5.1. Datasets

We examined 9 datasets, related to a specific ADMET related property that were explained in the section 4.1 . We retrieved the datasets from the work of Schyman et al [6]. from the literature.

- Chemical mutagenicity (AMES): For chemical mutagenicity, the AMES test results of 6512 molecules were used. 3503 of them is AMES-positive which means the molecule is mutegenic and 3009 of them is AMES-negative which means the molecule is not mutegenic.

- Blood Brain Barrier Penetration: For BBB permeability, log BB values of molecules were filtered in a way that the ones with lower than -0.3 of $logBB$ was classified as BBB-nonpermeable, the ones with greater than 0.3 of $logBB$ was classified as BBB-permeable. There are 353 compounds in the dataset, and 197 of them is BBB-permeable (positive), 156 of them is BBB-nonpermeable (negative).

- Cytotoxicity (HepG2): For cytotoxicity prediction, the toxicity recordings against HepG2 cells were used. There are 6097 molecules in the dataset, which were classified as cytotoxic (positive) if the molecule with an IC50 of 10μM or less, otherwise is not cytotoxic (negative). 1970 of the are positive and 4127 of the are negative.

- Drug-Induced Liver Injury: In drug-induced liver injury dataset, there are 1427 molecules. 780 of them are positive which means that the molecule is associated with a high risk of liver injury, 647 of them are negative which means there is no risk of liver injury related to that drug.

- hERG Blockers: In human ether a-go-go related gene blockers dataset, we have 685 molecules. hERG-blockers are labeled as positives and there are 282 of them, whereas hERG non-blockers are labeled as negatives and there are 403 of them.

- Human Liver Microsomal Stability:In human liver microsomal stability dataset contains 3219 compounds. Molecules with having reported half life greater than 30 minutes was classified as stable (positive), whereas the ones having half life less than 30 minutes was classified as non-stable (negative). 2,047 positive and 1,166 negative samples are included in the dataset.

- Mitochondrial Membrane Potential Distruption: MMP distruption dataset consists of 6261 compounds gained by using measurements MMP distruption of HepG2 cells. 913 compounds were found to decrease MMP and labeled as positive, 5395 compounds were found to have no effect on MMP and labeled as negatives.

- Permeability Glycoprotein Inhibitors: Pgp inhibitors dataset includes 1,319 inhibitors and 937 non-inhibitors. Pgp inhibitors and non-inhibitors are labeled as positives and negatives, respectively.

- Permeability Glycoprotein Substrates:The Pgp substrate dataset includes measurements for 422 substrates and 400 non-substrates.Pgp substrates and non-substrates are labeled as positives and negatives, respectively.

5.2. Molecule Representation

In our work, for each of the 9 datasets, we used 4 different types of molecular representations, namely topological fingerprints, morgan fingerprints, SmilesVec representation and a feature vector consisting of a custom set of molecular descriptors. In this section, we will be explaining these representations.

5.2.1. Fingerprinting

We gave the detailed information about molecular fingerprints in the previous section. Here, a brief implementation detail is revealed.

- Topological Fingerprints: To construct topological fingerprinting, we used RDKit's Chem library. Our topological fingerprint size is 2048 bits with default set of parameters.

- Morgan Fingerprints For morgan fingerprints, we also used RDKit with fingerprint size of 2048 bits with default set of parameters.

5.2.2. SmilesVec Representation

Another approach that we used for molecular representations is proposed by Öztürk et al [56]. In SmilecVec representation, a molecular SMILES is divided into a set of molecular strings, using sliding window method with substring size of 8. For each substring of molecule, a Word2Vec [57] vector is constructed. A complete representation of that molecule is generated by adding all the Word2Vec vectors of 'n' words in the molecule. It makes use of Word2Vec model's ability of reflecting complex structures and advantage of vector learning that can detect some similarities in the ordering of substrings that frequently occur.

For calculating SmilesVec of each molecule, we used Öztürk et al.'s open source

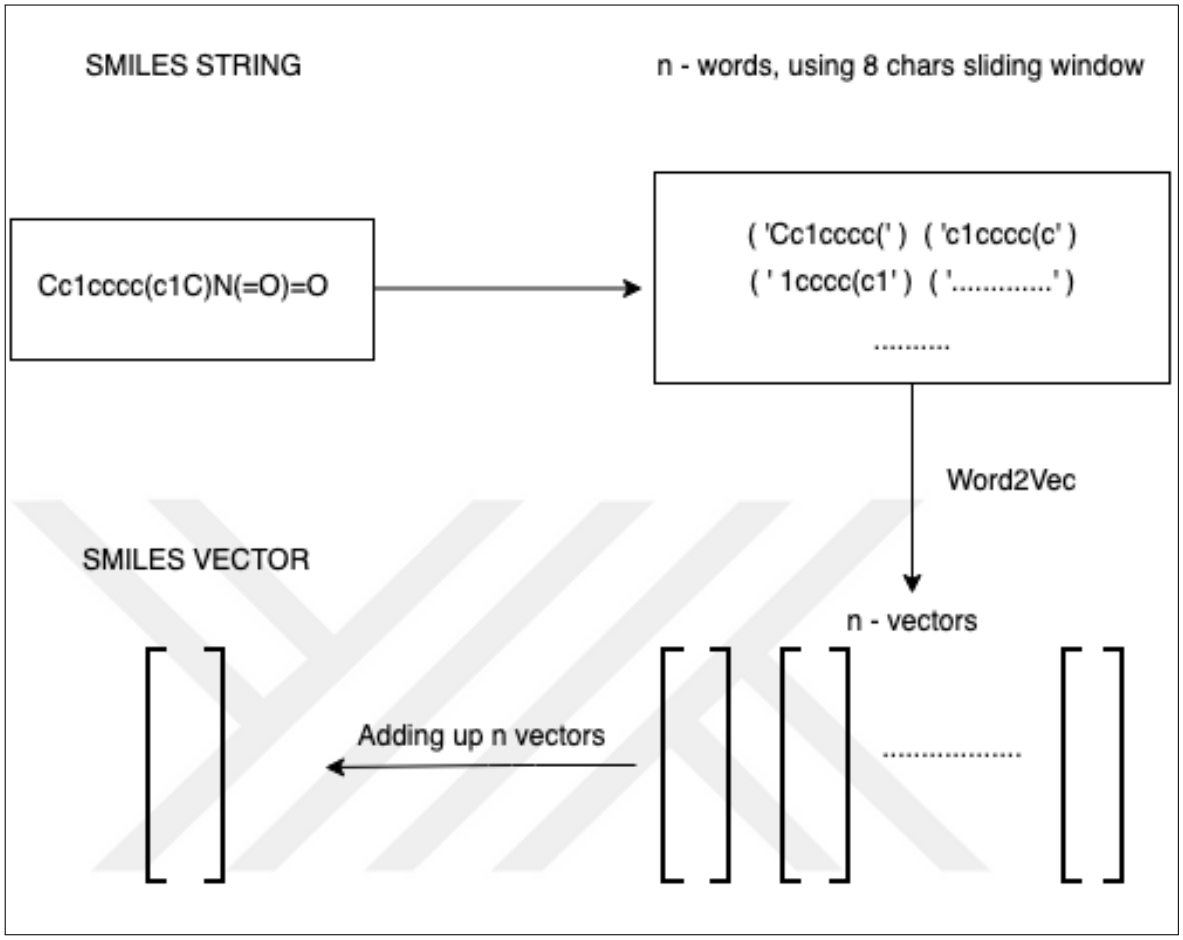

Figure 5.1. SmilesVec Representation

implementation along with the word level embeddings trained on 1.7 million samples on CHEMBL23 database. In SmilesVec representation, every molecule is represented as a 100-dimensional vector.

5.2.3. Descriptor Vector Representation

Here is the custom feature set which was explained in detail in section 4.2.:

$F_c = \{MW, LogP, TPSA, HBD, HBA, ROTB, MolMR, BalabanJ, N_{ValanceElectron}, AROM, N_{Atom}, N_{Acid}, N_{Base}, ECIndex, Diameter, Radius, PetitjeanI, Apol, Bpol, N_H, N_B, N_C, N_N, N_O, N_S, N_P, N_F, N_{Cl}, N_{Br}, N_I\}$

For each molecule in datasets, we calculated descriptor set above using Mordred

[58] molecular descriptor calculator. Then we normalized the values in the feature vector using min-max scaler. In descriptor vector representation, every molecule is represented as a 30-dimensional vector.

5.3. Parameter Tuning

Datasets for each ADMET properties are divided into train and test set by 70% and 30% accordingly. (Figure 5.2) To preserve the homogenicity of datasets, we split train and test sets in such a way that both of them contains an equal rate of positive and negative number of samples.

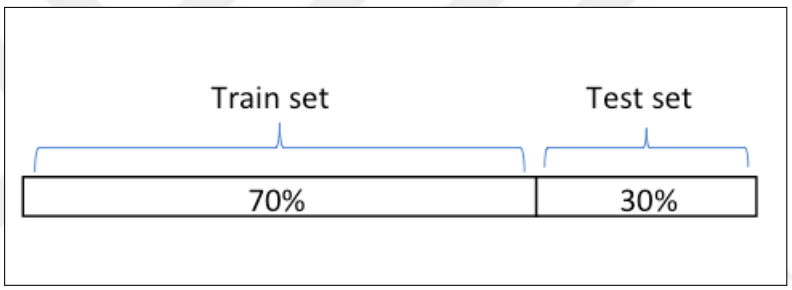

Figure 5.2. Data split for train and test

For parameter tuning, we trained a set of parameters for each classifier on train sets. By 10-fold cross validation, for each set of parameters, mean accuracy of 10 runs was taken into consideration. Then for each classification method, the most successful set of parameters were chosen for the testing. (Figure 5.3)

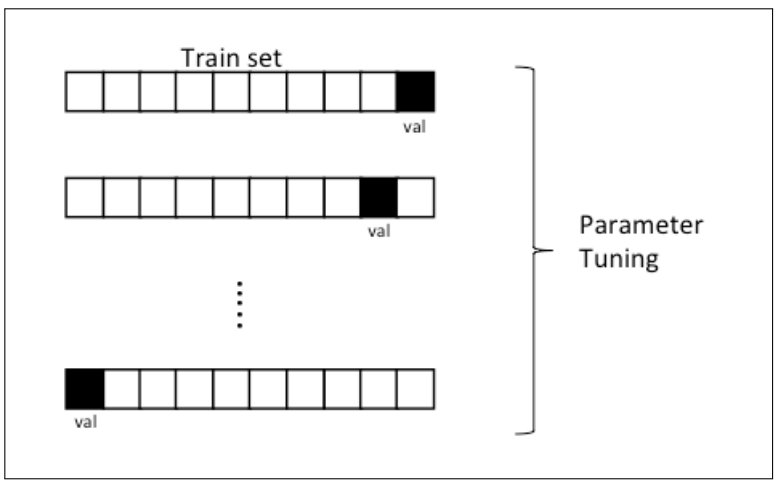

Figure 5.3. Parameter tuning on training set, by 10 fold cross validation.

- For kNN, we examined k values from 3 to 10.

- For SVM, rbf, linear and polynomial kernels were examined. For rbf kernel, we examined C and gamma values ranging from 0.001 to 10, increasing 10 times for the next value. For polynomial kernel, degree parameter values from 3 to 5.
- For RF, our criterion was either gini or entropy. Other parameters were number of estimators, in range [15, 20, 25, 30, 35]; minimum sample leaf, in range [1, 2, 3, 4]; minimum sample split, in range [3, 4, 5, 6].

5.4. Results

In testing phase, we constructed our models for each classifier by using the tuned parameters. To be able to get more realistic test results, we divided our training set into 10 folds, and for each run in testing phase, we took out one fold out of the train data. Then we tested our models using our test set (previously partitioned, 30% of the dataset) 10 times (Figure 5.4). Mean values and standard deviations of performance measures for all runs are reported. F- measure is calculated using the mean precision and recall values obtained at the end of all 10 runs.

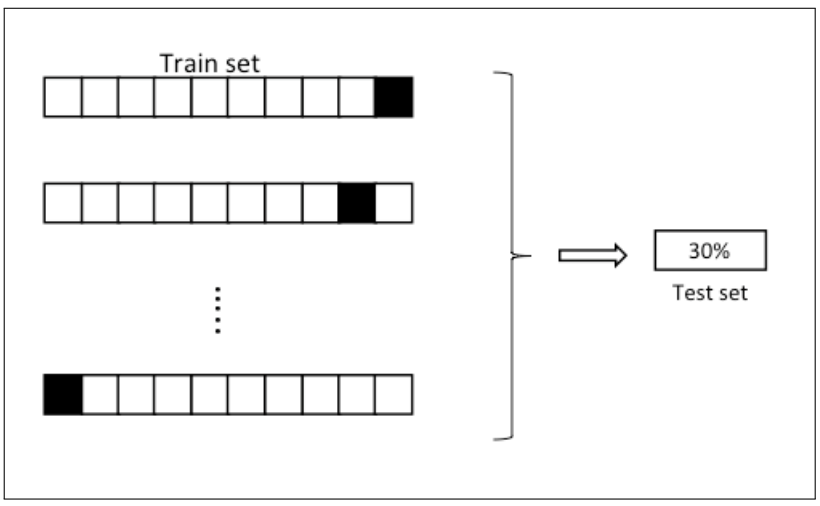

Figure 5.4. Experiments on test set 10 times, by using 90% of the training set each time.

5.4.1. Performance Measures

To evaluate performance of our model, we used following metrics:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

$$precision = \frac{TP}{TP + FP} \tag{5.2}$$

$$recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$fScore = \frac{2 * (Precision) * (Recall)}{Precision + Recall} \tag{5.4}$$

where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively.

5.4.2. Chemical mutagenicity Results

In table 5.1 you can see the hyperparameter tuning results for chemical mutagenicity dataset. For each data representation a set of hyperparameters are decided. In table 5.2, the evaluation metric scores of models constructed by using hyperparameters are displayed.

Table 5.1. Hyperparameters for AMES

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	3	rbf	10	0.001	gini	20	1	6
Topological FP	3	rbf	10	0.001	gini	25	1	5
SmilesVec	4	rbf	10	0.1	entropy	30	1	4
DescVec	10	rbf	10	0.1	gini	20	1	6

The highest scores are achieved using morgan fingerprints and SVM classifier. F-measure turns out to be 1 for our testing set. RF gives very close scores to SVM, 0.99. kNN gives 81% accuracy with 0.74 F-score.

Topological fingerprints are also a good representative of molecules for chemical mutagenicity dataset. In the case of topological fingerprints, SVM fits best with 99% accuracy. RF scores follows the SVM, with 91% accuracy with 0.88 F-score. kNN gives 77% accuracy with 0.71 F-score which is pretty comparable with kNN results of morgan fingerprints.

Smiles vector representation of molecules gives its highest accuracy with SVM which is 76% accuracy with 0.70 F-score. RF gives close results to SVM, 72% accuracy with 0.65 F-score. kNN also gives very similar results to SVM, 0.67 F-score with 72% accuracy.

In descriptor vector representation, best scores are provided by RF with 72% accuracy and 0.67 F-measure score. Next best scores belongs to SVM with same F-measure and 65% accuracy. kNN gives close scores, same accuracy as SVM, with 0.58 F-measure.

Table 5.2. Ames mutagenicity Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.8111 (0.0032)	Topological FP	kNN	Accuracy	0.7653 (0.0056)
		Precision	0.7936 (0.0062)			Precision	0.7756 (0.006)
		Recall	0.7084 (0.0058)			Recall	0.6654 (0.0076)
		F-measure	0.7422			F-measure	0.7162
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	0.9984 (0.0004)
		Precision	1.0 (0)			Precision	0.9991 (0.0006)
		Recall	1.0 (0)			Recall	0.9984 (0.0011)
		F-measure	1.0			F-measure	0.9992
	RF	Accuracy	0.9851 (0.0058)		RF	Accuracy	0.9085 (0.0108)
		Precision	0.9962 (0.0023)			Precision	0.9051 (0.0112)
		Recall	0.9935 (0.0039)			Recall	0.8519 (0.0161)
		F-measure	0.9953			F-measure	0.8777
SmilesVec	kNN	Accuracy	0.7226 (0.0041)	DescriptorVec	kNN	Accuracy	0.6482 (0.0026)
		Precision	0.7438 (0.0097)			Precision	0.6502 (0.0035)
		Recall	0.6154 (0.0062)			Recall	0.5236 (0.0002)
		F-measure	0.6735			F-measure	0.5800
	SVM	Accuracy	0.7571 (0.0050)		SVM	Accuracy	0.6469 (0.0015)
		Precision	0.7540 (0.0054)			Precision	0.7109 (0.0003)
		Recall	0.6489 (0.0060)			Recall	0.5301 (0.0018)
		F-measure	0.6975			F-measure	0.6073
	RF	Accuracy	0.7222 (0.0071)		RF	Accuracy	0.7271 (0.0039)
		Precision	0.7051 (0.0105)			Precision	0.7104 (0.0066)
		Recall	0.6030 (0.0086)			Recall	0.6087 (0.0051)
		F-measure	0.6500			F-measure	0.6556

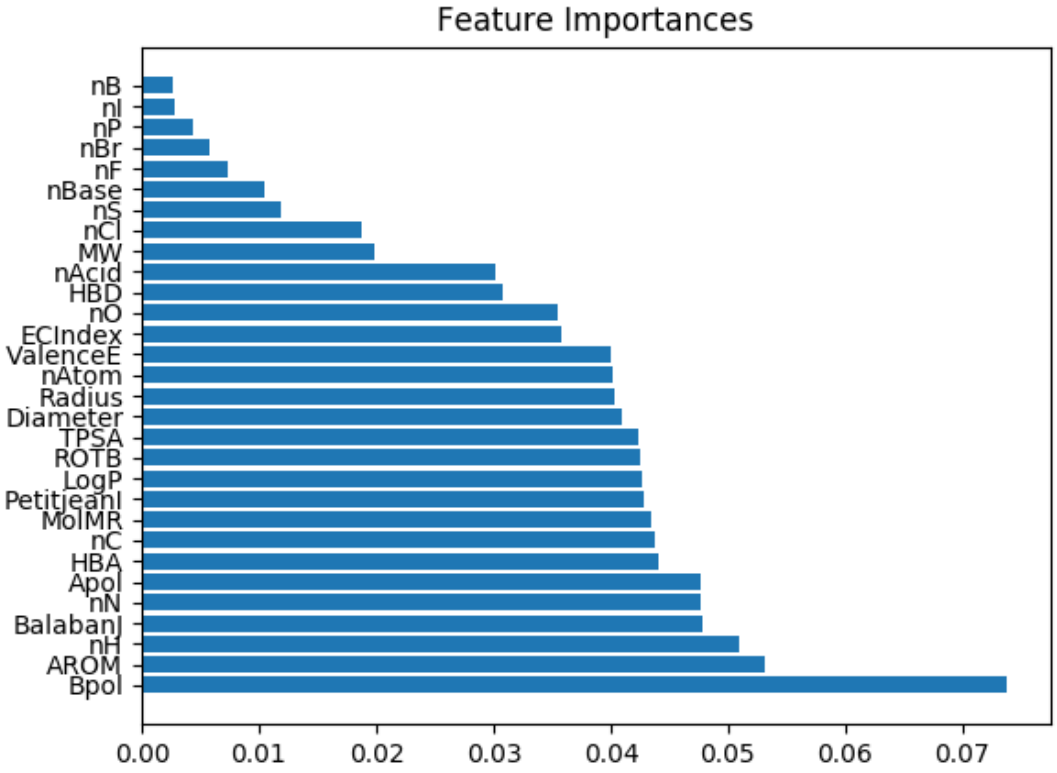

Figure 5.5. Feature importances of custom descriptor set for AMES dataset

In figure 5.5, RF feature importances of descriptor set for AMES dataset are plotted. Here, most important feature for AMES data set is displayed to be bond polarization with 7% importance. (Here importance is calculated by adding up the effectiveness of that feature decreasing impurity, meaning the possibility to classify it with wrong class.) We see a block of features that contributes to results almost equally; namely, AROM, nH, BalabanJ, nN, Apol, HBA, nC, MolMR, PetitjeanI, LogP, ROTB, TPSA, Diameter, Radium, nAtom, ValenceE.

5.4.3. Blood Brain Barrier Results

In table 5.3, hyperparameter tuning results are displayed. With these paramaters, we construct the models, and the scores of these models are displayed in Table 5.4.

Table 5.3. Hyperparameters for BBB

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	4	linear	-	-	gini	25	1	4
Topological FP	3	linear	-	-	gini	25	1	3
SmilesVec	9	rbf	10	0.01	entropy	25	1	6
DescVec	8	rbf	10	0.1	gini	30	1	4

For BBB dataset, we see the highest scores belong to SVM with morgan fingerprints yielding 100% accuracy. With morgan fingerprints, RF gives the next highest scores with 92% accuracy and 0.98 F-score. kNN gives 61%, the lowest accuracy with morgan fingerprints, with 0.94 F-score.

Second highest scores for BBB dataset belongs to descriptor vector representation. RF gives 88%, the highest accuracy with 0.88 F-measure. Next highest scores are achieved by kNN, 81% accuracy with 0.84 F-measure. Lowest accuracy score is by SVM with 77% accuracy with 0.91 F-measure.

Topological fingerprints gives the next highest results. Maximum accuracy for TPFP is yielded by SVM, in which accuracy is 86% with 0.93 F-measure. RF follows SVM, yielding 81% accuracy with 0.89 F-measure. The lowest results of TPFP are yielded by kNN. The accuracy is 0.67 and F-measure is 0.76.

For BBB dataset lowest overall scores are achieved by smiles vector representation. The classifier success is SVM,kNN , RF from highest to lowest. Accuracy values are 0.75, 0.75, 0,70 whereas F-mesures are 0.81, 0.85, 0.83.

Table 5.4. BBB Penetration Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.6085 (0.0180)	Topological FP	kNN	Accuracy	0.6670 (0.0184)
		Precision	0.9673 (0.0099)			Precision	0.7776 (0.0169)
		Recall	0.9140 (0.0267)			Recall	0.7493 (0.0170)
		F-measure	0.9399			F-measure	0.7632
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	0.8735 (0.0250)
		Precision	1.0 (0)			Precision	0.9245 (0.0183)
		Recall	1.0 (0)			Recall	0.9273 (0.0185)
		F-measure	1.0			F-measure	0.9259
	RF	Accuracy	0.9160 (0.0577)		RF	Accuracy	0.8223 (0.0181)
		Precision	0.9795 (0.0204)			Precision	0.8939 (0.0237)
		Recall	0.9789 (0.0216)			Recall	0.8912 (0.0202)
		F-measure	0.9792			F-measure	0.8925
SmilesVec	kNN	Accuracy	0.7472 (0.0205)	DescriptorVec	kNN	Accuracy	0.8122 (0.0165)
		Precision	0.8122 (0.0285)			Precision	0.8367 (0.0329)
		Recall	0.8111 (0.0256)			Recall	0.8500 (0.0245)
		F-measure	0.8116			F-measure	0.8433
	SVM	Accuracy	0.7519 (0.0119)		SVM	Accuracy	0.7679 (0.0638)
		Precision	0.8612 (0.0237)			Precision	0.9163 (0.0460)
		Recall	0.8472 (0.0202)			Recall	0.9036 (0.0333)
		F-measure	0.8541			F-measure	0.9099
	RF	Accuracy	0.6962 (0.0245)		RF	Accuracy	0.8821 (0.0135)
		Precision	0.8510 (0.0242)			Precision	0.8735 (0.0152)
		Recall	0.8150 (0.0235)			Recall	0.8912 (0.0117)
		F-measure	0.8326			F-measure	0.8823

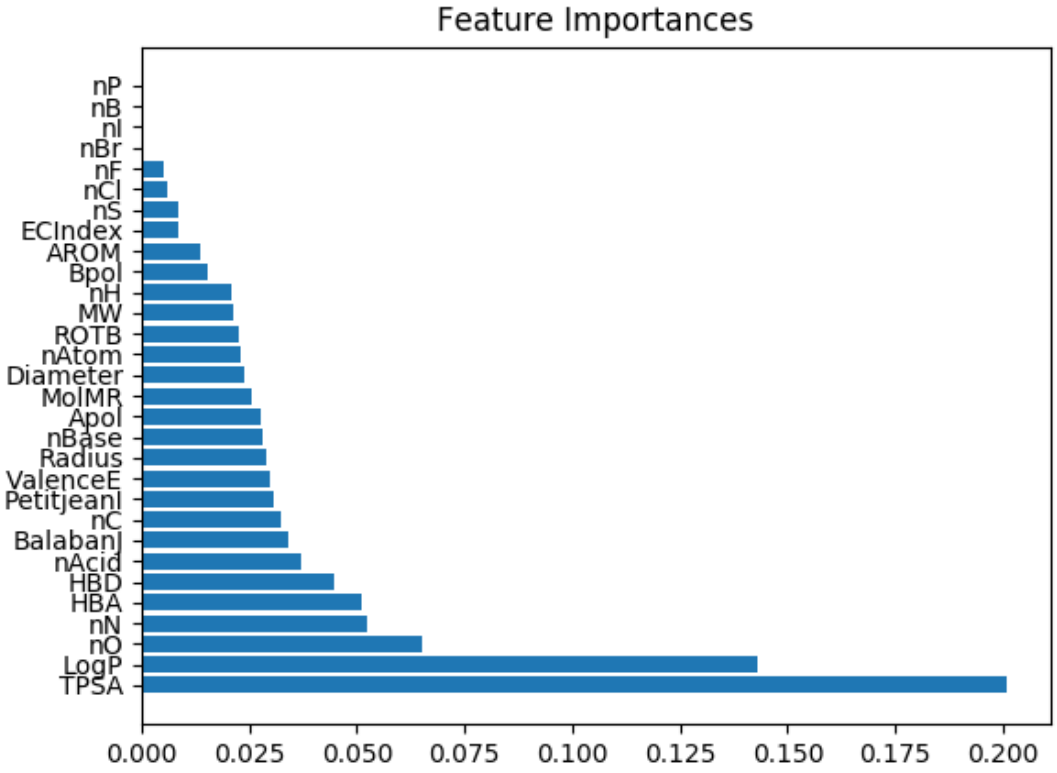

Figure 5.6. Feature importances of custom descriptor set for BBB dataset

In Figure 5.6, RF feature importances of descriptor set for BBB permeability dataset are plotted. In this case, we see a clear importance of topological polar surface area with 0.2 feature importance and water-octanol partition coefficient with 0.14 feature importance for BBB permeability classification.

In the literature, there have been experimental findings about correlation between penetration into living membranes and TPSA value [37] [39]. Waterbeemd et al. showed that for drugs targeted to central nervous system, only molecules with less than 60-70 $Å^2$ can be transported through the BBB [38]. For LogP, Lipinski et al. suggested that chemicals with calculated LogP value greater than 5, have poor absorption and permeation characteristics [34]. These findings in literature are correlated with what we found about BBB permeability.

Table 5.5. Results for morgan fingerprints and updated morgan fingerprints (2 descriptors) added with Random Forest classifier

	Morgan FP	Morgan FP + [TPSA, LogP]
Accuracy	0.92	0.95
Precision	0.98	0.94
Recall	0.98	0.95
F-Measure	0.98	0.95

In table 5.5, the results of updated morgan FP and morgan FP results are displayed. TPSA and LogP descriptors are added to morgan FP. The model accuracy increased by 3%.

5.4.4. Cytotoxicity Results

In table 5.6, parameter tuning results of cytotoxicity dataset is displayed. In table 5.7, results of the models created with hyperparameters are displayed.

Table 5.6. Hyperparameters for Cytotoxicity

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	3	rbf	1	0.001	gini	25	1	6
Topological FP	3	rbf	10	0.001	gini	20	1	5
SmilesVec	3	rbf	10	0.1	gini	30	1	5
DescVec	4	rbf	1	0.1	entropy	30	1	4

For cytotoxicity dataset, both of the fingerprint representations work very good for SVM yielding 100% accuracy and F-measure. RF also gives very good results with morgan fingerprints. For topological fingerprints, RF gives 86% accuracy with 0.78 F-measure. kNN is a little bit behind RF with 80% accuracy with 0.76 F-measure. kNN's results using morgan fingerprints are similar to kNN with TPFP results, 82% accuracy with 0.77 F-measure.

Table 5.7. Cytotoxicity Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.8192 (0.0038)	Topological FP	kNN	Accuracy	0.7971 (0.0032)
		Precision	0.6830 (0.0061)			Precision	0.6983 (0.0058)
		Recall	0.8234 (0.0029)			Recall	0.8233 (0.0028)
		F-measure	0.7667			F-measure	0.7557
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	1.0 (0)
		Precision	1.0 (0)			Precision	1.0 (0)
		Recall	1.0 (0)			Recall	1.0 (0)
		F-measure	1.0			F-measure	1.0
	RF	Accuracy	0.9942 (0.0029)		RF	Accuracy	0.8609 (0.0075)
		Precision	0.9960 (0.0037)			Precision	0.7188 (0.0148)
		Recall	0.9976 (0.0023)			Recall	0.8467 (0.0071)
		F-measure	0.9970			F-measure	0.7775
SmilesVec	kNN	Accuracy	0.7701 (0.0049)	DescriptorVec	kNN	Accuracy	0.6638 (0.0029)
		Precision	0.6538 (0.0095)			Precision	0.4856 (0.0072)
		Recall	0.7993 (0.0046)			Recall	0.7112 (0.0027)
		F-measure	0.7193			F-measure	0.5757
	SVM	Accuracy	0.7700 (0.0036)		SVM	Accuracy	0.6211 (0)
		Precision	0.5623 (0.0098)			Precision	0
		Recall	0.7706 (0.0035)			Recall	0.6211 (0)
		F-measure	0.6502			F-measure	0
	RF	Accuracy	0.7444 (0.0046)		RF	Accuracy	0.7310 (0.0035)
		Precision	0.4371 (0.0154)			Precision	0.4243 (0.0093)
		Recall	0.7308 (0.0049)			Recall	0.7233 (0.0027)
		F-measure	0.5470			F-measure	0.5348

Results of smiles vector representation and descriptor vector representation is 20-25% behind of FPs. SVM and kNN give same accuracy value 0.77 using smiles vector representation with 0.72 and 0.65 F-measure respectively. RF is worse than first two, with 74% accuracy and 0.55 F-measure. For descriptor set representation, RF gives highest accuracy, 73%, but low F-measure, 0.54. Similarly, kNN yields 66% accuracy wiht 0.58 F-measure. Oddly, here SVM gives 0 precision and therefore 0 F-measure. And accuracy value is 0.62.

We think the reason why SVM gives 0 precision is because cytotoxicity dataset consists of 1970 positive and 4127 negative samples. Since the dataset has unbalanced number of positive and negative classes, descriptor set representation cannot learn very well for positive class and classify every positive test instance as negative, resulting 0 precision.

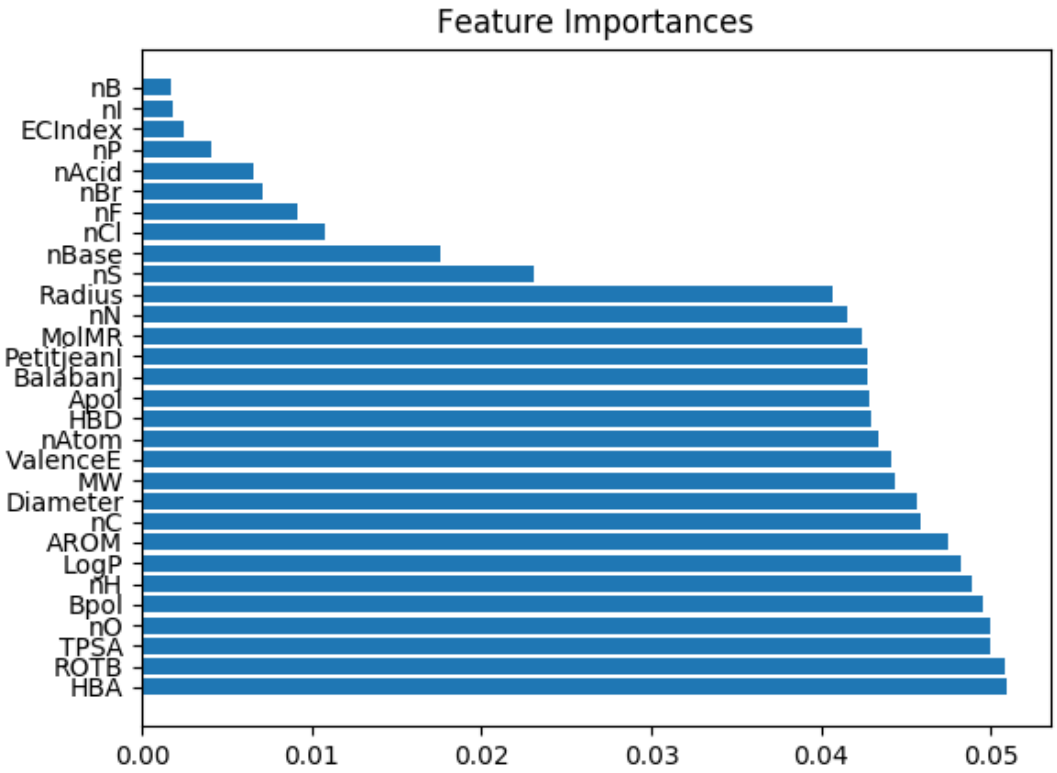

Figure 5.7. Feature importances of custom descriptor set for Cytotoxicity dataset

In figure 5.7, feature importances of RF are plotted. For cytotoxicity dataset,

hydrogen bond acceptor has largest feature importance value. Furthermore, number of hydrogen and nitrogen atoms also have strong effect on the classification result. In a cytotoxicity profiling study, it has been suggested that hydrogen bond acceptor affects solvents in such a way that it becomes toxic [59] which correlates with our findings.

5.4.5. Drug Induced Liver Injury Results

In table 5.8, hyperparameter tuning results are displayed. In table 5.9, results of the models constructed with hyperparameters are shown.

Table 5.8. Hyperparameters for DILI

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	6	rbf	10	0.001	gini	30	1	5
Topological FP	5	linear	-	-	entropy	20	1	5
SmilesVec	5	rbf	1	0.1	entropy	30	1	6
DescVec	7	rbf	10	0.1	entropy	30	1	3

For DILI dataset, SVM classifier with morgan and topological fingerprints give the best results with 100% accuracy. RF using morgan fingerprints gives similar results with 98% accuracy and 0.98 F-measure. RF using topological fingerprints yields 80% accuracy and 0.71 F-measure. kNN classifier however for both fingerprint representations gives the lowest scores, with 0.68 F-measure for morgan fingerprints and 0.53 F-measure for topological fingerprints.

Next highest result is achieved by descriptor vectors using RF classifier. In that case F-measure is 0.59 with 0.68 % accuracy results. kNN gives next highest results for descriptor vector representation wiht 60% accuracy and 0.50 F-measure. SVM in this case barely classified the samples, yielding 53% accuracy and 0.42 F-measure. Smiles vector representation's best scores are by SVM classifier with 0.57 F-measure and 67% accuracy. RF and kNN follows SVM with both 0.51 F-measure and 62% accuracy.

Table 5.9. Drug Induced Liver Injury Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.7790 (0.0136)	Topological FP	kNN	Accuracy	0.6389 (0.0099)
		Precision	0.7523 (0.0175)			Precision	0.6035 (0.0112)
		Recall	0.6279 (0.0163)			Recall	0.4730 (0.0099)
		F-measure	0.6845			F-measure	0.5303
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	1.0 (0)
		Precision	1.0 (0)			Precision	1.0 (0)
		Recall	1.0 (0)			Recall	1.0 (0)
		F-measure	1.0			F-measure	1.0
	RF	Accuracy	0.9828 (0.0179)		RF	Accuracy	0.8004 (0.0209)
		Precision	0.9831 (0.0073)			Precision	0.7789 (0.0274)
		Recall	0.9670 (0.0141)			Recall	0.6586 (0.0298)
		F-measure	0.9750			F-measure	0.7137
SmilesVec	kNN	Accuracy	0.6233 (0.0150)	DescriptorVec	kNN	Accuracy	0.60 (0.0102)
		Precision	0.5989 (0.0199)			Precision	0.60 (0.0096)
		Recall	0.4570 (0.0149)			Recall	0.4293 (0.0115)
		F-measure	0.5184			F-measure	0.5005
	SVM	Accuracy	0.6654 (0.0151)		SVM	Accuracy	0.5295 (0.0754)
		Precision	0.6620 (0.0196)			Precision	0.4505 (0.2150)
		Recall	0.4998 (0.0170)			Recall	0.3962 (0.0304)
		F-measure	0.5685			F-measure	0.4216
	RF	Accuracy	0.6169 (0.0135)		RF	Accuracy	0.6773 (0.0189)
		Precision	0.6079 (0.0174)			Precision	0.6835 (0.0226)
		Recall	0.4484 (0.0136)			Recall	0.5136 (0.0224)
		F-measure	0.5161			F-measure	0.5866

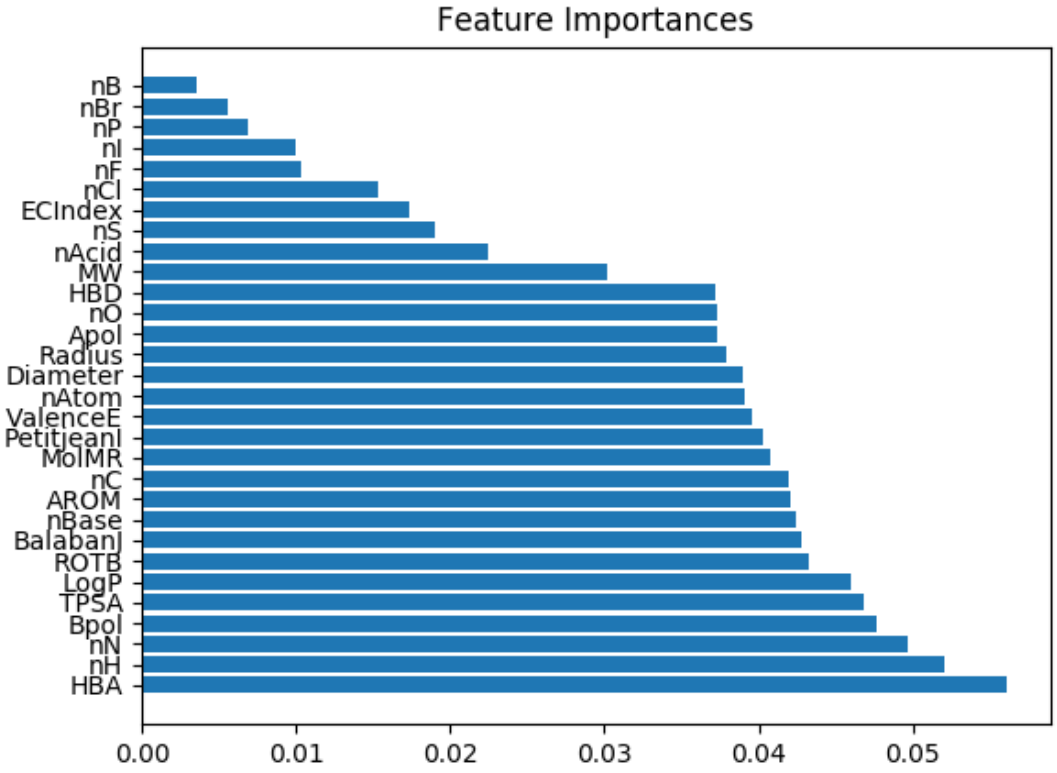

Figure 5.8. Feature importances of custom descriptor set for DILI dataset

In figure 5.8, random forest feature importances are plotted for DILI dataset. Hydrogen bond acceptor descriptor contributes to the classification results of random forest classifier the most. Number of hdrogen and nitrogen atoms, bond polarization, topological polar surface area and water-octanol partition coefficient are also contributive to the results.

5.4.6. hERG Blockers Results

In table 5.10 parameter tuning results of hERG dataset is displayed. In table 5.11 scores of classifiers modelled on hyperparameters are explained.

Table 5.10. Hyperparameters for hERG

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	3	rbf	10	0.001	entropy	20	1	7
Topological FP	5	linear	-	-	entropy	30	1	6
SmilesVec	4	rbf	1	0.1	entropy	25	1	5
DescVec	10	rbf	10	0.1	entropy	30	1	3

For hERG database, SVM classifier using morgan fingerprints yields 99% accuracy and 0.99 F-measure. Random forest using morgan fingerprints gives pretty high results with 92% accuracy and 0.88 F-measure. kNN gives lowest scores with 76% accuracy and 0.58 F-measure.

Topological fingerprints yield next highest scores. Random forest classifier gives the highest scores among SVM and kNN classifiers. It gives 88% accuracy and 0.81 F-score. Next highest scores are achieved by SVM with 86% accuracy and 0.79 F-score. Last scores are with kNN, yielding 81% accuracy and 0.73 F-measure.

For this dataset smiles vector and descriptor vector representations give very close scores. Best scores of descriptor vector representation ia achieved by random forest classifier. It yields 82% accuracy with 0.75 F-measure. kNN's scores follow RF, producing 73% accuracy and 0.67 F-measure. Lastly SVM yields 70% accuracy and 0.61 F-measure.

Smiles vector representation's best scores are achieved by SVM classifier. It predicts test samples with 82% accuracy and 0.74 F-measure. RF and kNN for this case, yield similar results; accuracy scores are 0.77 and 0.76 respectively. F-measure scores are 0.74 and 0.66 respectively.

Table 5.11. Human Ether a-go-go Related Gene Blockers Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.7618 (0.0055)	Topological FP	kNN	Accuracy	0.8050 (0.0086)
		Precision	0.4946 (0.0121)			Precision	0.6957 (0.0153)
		Recall	0.6953 (0.0049)			Recall	0.7747 (0.0079)
		F-measure	0.5780			F-measure	0.7330
	SVM	Accuracy	0.9995 (0.0015)		SVM	Accuracy	0.8648 (0.0139)
		Precision	0.9989 (0.0032)			Precision	0.7630 (0.0187)
		Recall	0.9991 (0.0027)			Recall	0.8238 (0.0126)
		F-measure	0.9990			F-measure	0.7923
	RF	Accuracy	0.9166 (0.0203)		RF	Accuracy	0.8759 (0.0134)
		Precision	0.8609 (0.0452)			Precision	0.7902 (0.0233)
		Recall	0.8909 (0.0317)			Recall	0.8406 (0.0150)
		F-measure	0.8756			F-measure	0.8146
SmilesVec	kNN	Accuracy	0.7603 (0.0110)	DescriptorVec	kNN	Accuracy	0.7307 (0.0129)
		Precision	0.7130 (0.0195)			Precision	0.6272 (0.0266)
		Recall	0.7647 (0.0126)			Recall	0.7191 (0.0145)
		F-measure	0.7380			F-measure	0.6700
	SVM	Accuracy	0.8161 (0.0055)		SVM	Accuracy	0.6980 (0.0027)
		Precision	0.6967 (0.0141)			Precision	0.5641 (0.0032)
		Recall	0.7790 (0.0071)			Recall	0.6845 (0.0009)
		F-measure	0.7356			F-measure	0.6185
	RF	Accuracy	0.7713 (0.0131)		RF	Accuracy	0.8231 (0.0077)
		Precision	0.5815 (0.0234)			Precision	0.7141 (0.0109)
		Recall	0.7222 (0.0116)			Recall	0.7886 (0.0053)
		F-measure	0.6643			F-measure	0.7495

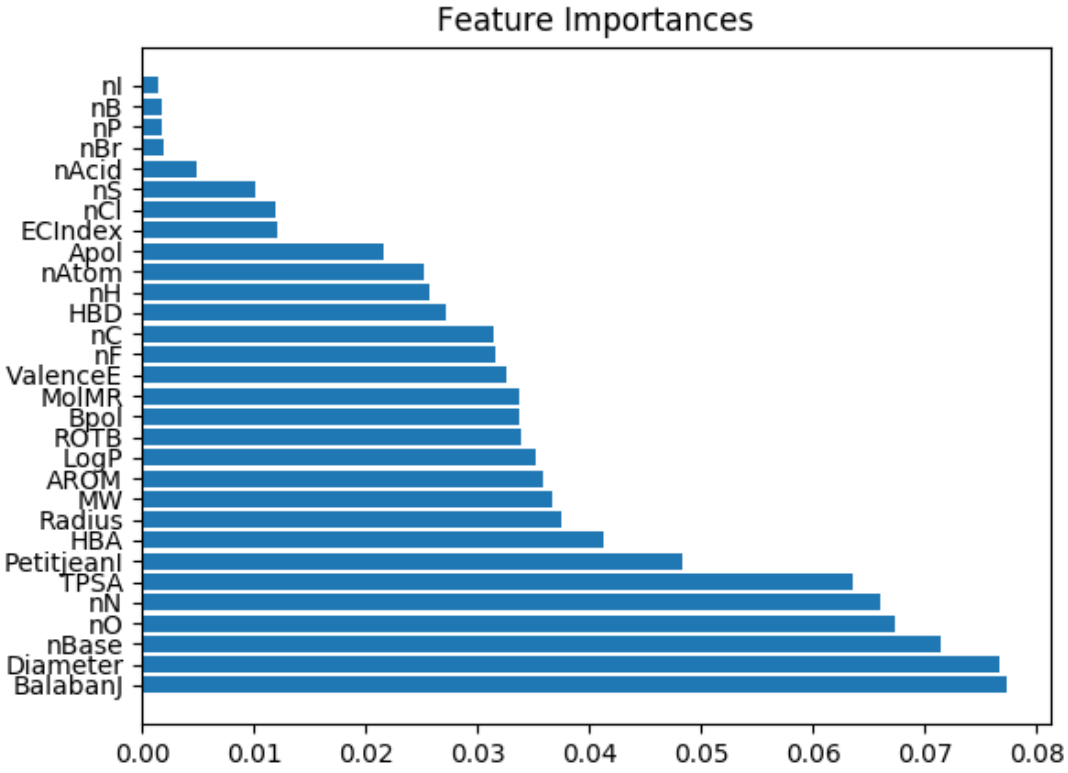

Figure 5.9. Feature importances of custom descriptor set for hERG blockers dataset

In figure 5.9, feature importances of random forest classifier using descriptor vector are plotted. For human ether a-go-go related gene blockers dataset, most contributive feature turns out to be Balaban's J value. Other important features from most effective to the least are diameter, number of basic group counts, number of oxgyen and nitrogen and topological polar surface area.

RF classifier gives 92 percent accuracy using morgan fingerprints. Here we see a room for a possible improvement by adding more feature(s) feature to morgan fingerprint representation. In table 5.12, you can see the scores of morgan fingerprint and updated morgan fingerprint ([morgan fp, BalabanJ]). Here adding balaban's J index to the feature set made 3 percent improvement in accuracy, 7 percent improvement in precision, 5 percent improvement in recall and 5 percent improvement in F-measure.

Table 5.12. Results of Random Forest classifier with morgan fingerprints and updated morgan fingerprints (Balaban's J index descriptor added)

	Morgan FP	Morgan FP + [BalabanJ]
Accuracy	0.92	0.95
Precision	0.86	0.93
Recall	0.89	0.94
F-Measure	0.88	0.93

5.4.7. Human Liver Microsomal Stability Results

In 5.13, you can see the parameter tuning results of HLM dataset and in table 5.14 the scores of classifiers modelled with hyperparameters are displayed.

Table 5.13. Hyperparameters for HLM

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	9	rbf	1	0.001	entropy	30	1	7
Topological FP	3	linear	-	-	entropy	20	1	5
SmilesVec	3	rbf	10	0.1	gini	30	1	6
DescVec	5	rbf	1	0.1	gini	30	1	6

As in all of the cases, for HLM dataset, morgan fingerprints also gives the best scores. SVM classifier using morgan dataset yields a 100 percent accuracy and F-score. Random forest's scores follow SVM closely, with 0,97 accuracy score and 0.97 F-measure. kNN yields a lower accuracy score compared to other two classifiers. Its accuracy score is 0.78 with 0.78 F-measure.

Topological fingerprints are also good representative power of molecules in HLM dataset. SVM classifier with topological fingerprints achieves 100% accuracy. RF gives 88 % percent accuracy and 0.86 F-measure whereas kNN yields 78% accuracy with 0.82 F-measure.

Table 5.14. Human Liver Microsomal Stability Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.7846 (0.0061)	Topological FP	kNN	Accuracy	0.7817 (0.0075)
		Precision	0.7064 (0.0151)			Precision	0.7517 (0.0201)
		Recall	0.8792 (0.0054)			Recall	0.8934 (0.0072)
		F-measure	0.7834			F-measure	0.8165
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	1.0 (0)
		Precision	1.0 (0)			Precision	1.0 (0)
		Recall	1.0 (0)			Recall	1.0 (0)
		F-measure	1.0			F-measure	1.0
	RF	Accuracy	0.9679 (0.0081)		RF	Accuracy	0.8774 (0.0134)
		Precision	0.9619 (0.0146)			Precision	0.7981 (0.0260)
		Recall	0.9852 (0.0056)			Recall	0.9218 (0.0097)
		F-measure	0.9734			F-measure	0.8555
SmilesVec	kNN	Accuracy	0.7326 (0.0086)	DescriptorVec	kNN	Accuracy	0.6466 (0.0068)
		Precision	0.7128 (0.0199)			Precision	0.5649 (0.0095)
		Recall	0.8712 (0.0075)			Recall	0.8033 (0.0042)
		F-measure	0.60			F-measure	0.6633
	SVM	Accuracy	0.7740 (0.0054)		SVM	Accuracy	0.7240 (0)
		Precision	0.6879 (0.0208)			Precision	0
		Recall	0.8715 (0.0069)			Recall	0.7240 (0)
		F-measure	0.7689			F-measure	0
	RF	Accuracy	0.7691 (0.0060)		RF	Accuracy	0.7709 (0.0044)
		Precision	0.5826 (0.0332)			Precision	0.6260 (0.0249)
		Recall	0.8409 (0.0093)			Recall	0.8529 (0.0071)
		F-measure	0.6884			F-measure	0.7221

In figure 5.10, feature importances of random forest classifier using descriptor vector representation are plotted. Number of rotatable bonds descriptor contributes to the results of HLM dataset the most. Topological polar surface area, water-octanol partition coefficient and number of aromatic rings take great portion of contribution to classify test samples correctly.

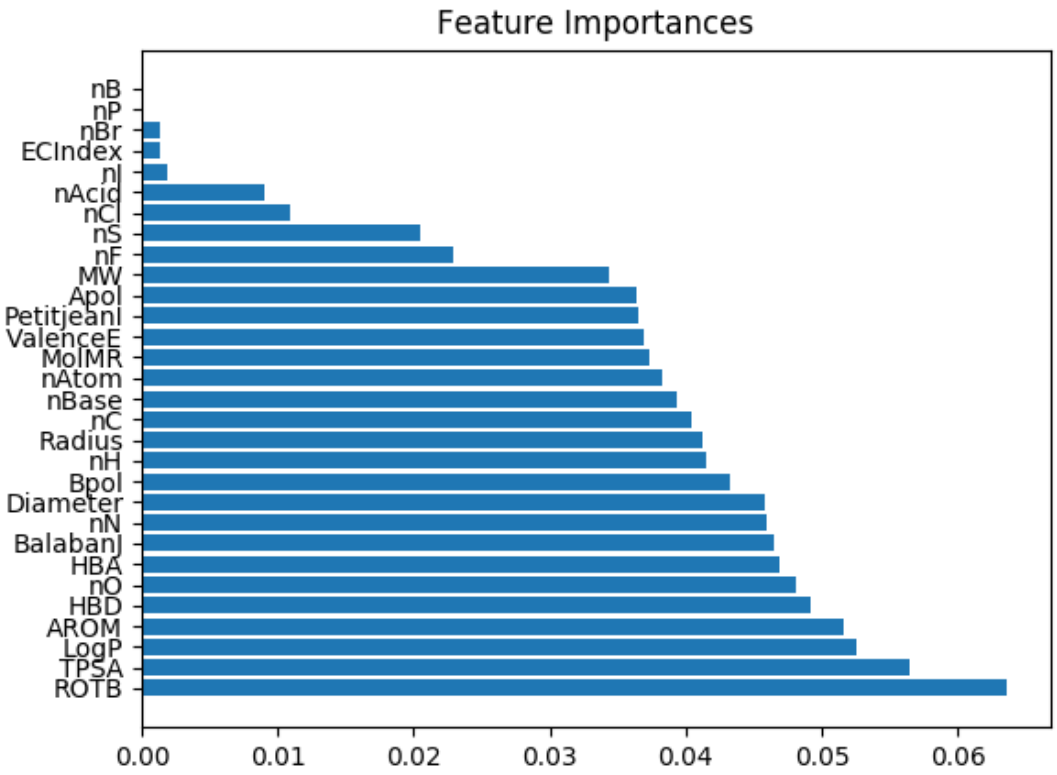

Figure 5.10. Feature importances of custom descriptor set for HLM stability dataset

5.4.8. Mitochondrial Membrane Potential Distruption Results

In table 5.15, hyperparameters for HLM dataset are displayed. In table 5.16 results from different classifiers using different vector representations are shown.

SVM classifier using morgan fingerprint representation gives highest accuracy and F-measure for MMP dataset. Random forest yields 99% accuracy with 0.99 F-measure which are also very high scores. kNN on the other hand, gives a considerably high accuracy (92%) but very low precision (0.50) and therefore F-score (0.66).

Table 5.15. Hyperparameters for MMP

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	3	rbf	1	0.001	gini	25	1	4
Topological FP	4	linear	-	-	gini	25	1	3
SmilesVec	8	rbf	10	0.1	entropy	30	1	3
DescVec	8	rbf	1	0.1	entropy	25	1	3

SVM classifier also yields 100% accuracy using topological fingerprints. Like kNN classifier using morgan fingerprints, using topological fingerprints SVM and kNN gives a considerably good accuracy value but low precision and F-measure. RF gives 95% accuracy with 0.73 F-measure whereas kNN yields 87% accuracy and 0.68 F-score.

Next highest results belong to descriptor vector representation. RF yields 92% accuracy and SVM yields 90% accuracy in the case of descriptor vectors. kNN scores follow RF and SVM with 84% accuracy. However, low precision and F-measure problem can also be seen in this case. F-measures are 0.64, 0, 0.67 for kNN, SVM and RF respectively. Because SVM classifier cannot classify any test samples that are member of positive class, 0 precision and F-measure is obtained.

Smiles vector representation scores highest accuracy in RF with 91%. SVM yields 89% and kNN yields 85% accuracy. F-measure values are 0.71, 0.74 and 0.46 for kNN, SVM and RF respectively.

In the case of MMP disruption dataset, we clearly see the effects of imbalanced number of samples in positive and negative classes. In this dataset, number of samples in positive class is 913 and number of samples in negative class is 5395. Therefore, classifiers learn negative class members well but due to low number of samples in training set, positive class members cannot be classified correctly.

Table 5.16. Mitochondrial Membrane Potential Distruption Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.9160 (0.0036)	Topological FP	kNN	Accuracy	0.8736 (0.0031)
		Precision	0.5043 (0.0218)			Precision	0.5314 (0.0165)
		Recall	0.9456 (0.0023)			Recall	0.9457 (0.0017)
		F-measure	0.6578			F-measure	0.6804
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	1.0 (0)
		Precision	1.0 (0)			Precision	1.0 (0)
		Recall	1.0 (0)			Recall	1.0 (0)
		F-measure	1.0			F-measure	1.0
	RF	Accuracy	0.9965 (0.0022)		RF	Accuracy	0.9521 (0.0033)
		Precision	0.9805 (0.0171)			Precision	0.5881 (0.0221)
		Recall	0.9978 (0.0019)			Recall	0.9557 (0.0023)
		F-measure	0.9891			F-measure	0.7282
SmilesVec	kNN	Accuracy	0.8485 (0.0028)	DescriptorVec	kNN	Accuracy	0.8425 (0.0034)
		Precision	0.5676 (0.0165)			Precision	0.4892 (0.0119)
		Recall	0.9480 (0.0017)			Recall	0.9393 (0.0013)
		F-measure	0.7100			F-measure	0.6433
	SVM	Accuracy	0.8903 (0.0039)		SVM	Accuracy	0.8996 (0)
		Precision	0.6032 (0.0192)			Precision	0
		Recall	0.9542 (0.0021)			Recall	0.8996 (0)
		F-measure	0.7392			F-measure	0
	RF	Accuracy	0.9122 (0.0043)		RF	Accuracy	0.9159 (0.0020)
		Precision	0.3081 (0.0474)			Precision	0.5222 (0.0143)
		Recall	0.9269 (0.0045)			Recall	0.9473 (0.0014)
		F-measure	0.4625			F-measure	0.6732

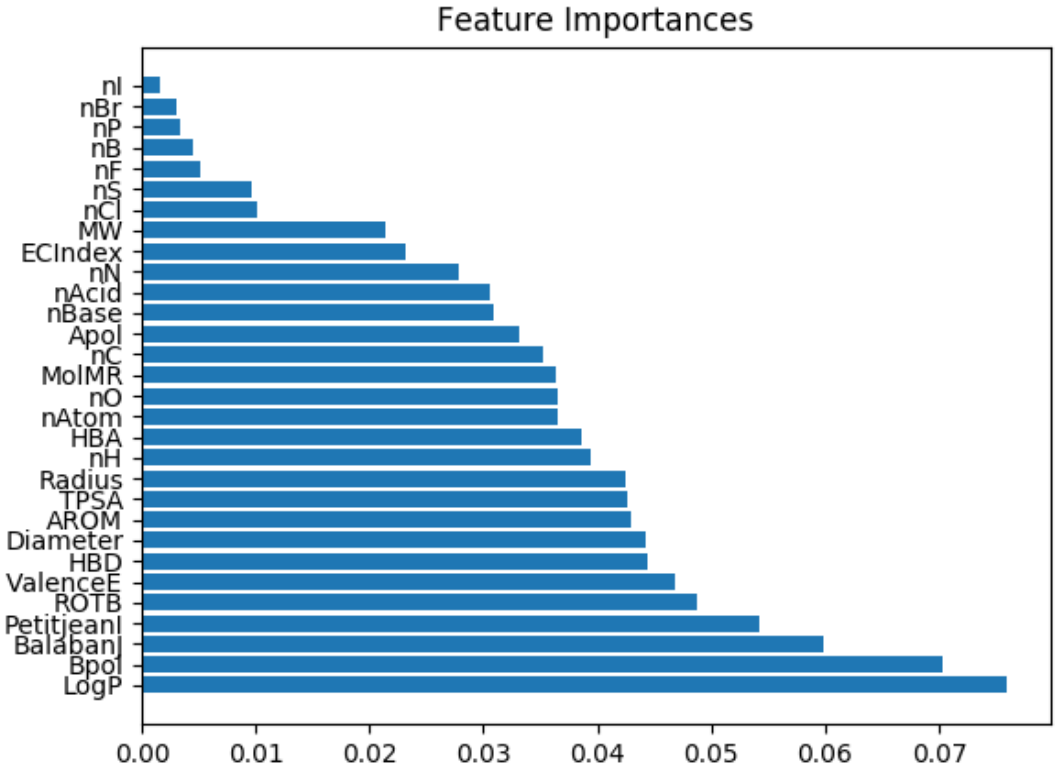

Figure 5.11. Feature importances of custom descriptor set for MMP distruption dataset

In figure 5.11, feature importances of random forest classifier using descriptor vector representation are plotted. Water-octanol partition coefficient turns out to be most important feature for MMP distruption dataset. Bond polarizability and Balaban's J index are also contributive to correctly classify test samples.

5.4.9. Permeability Glycoprotein Inhibitors Results

In table 5.17, parameter tuning results of PgP inhibitors results are shown. In table 5.18, scores of classifers built with hyperparamters are displayed.

Table 5.17. Hyperparameters for Pgp inhibitors

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	3	rbf	10	0.001	gini	20	1	6
Topological FP	3	linear	-	-	gini	20	1	5
SmilesVec	3	rbf	1	0.1	gini	30	1	5
DescVec	10	rbf	10	0.1	gini	30	1	4

Morgan fingerprint yields 100% accuracy with using SVM classifier. Random forest gives 96% accuracy with 0.98 F-measure. kNN is behind SVM and RF with 82% accuracy and 0.79 F-score. Topological fingerprints also yield 100% accuracy with SVM classifier. Next highest results are achieved by RF which makes predictions with 87% accuracy and 0.91 F-measure. kNN with topological fingerprints gives similar results to kNN with morgan fingerprints. Prediction accuracy is 78% and F-measure is 0.83.

Descriptor vector representation on the other hand gives a slightly higher results than smiles vector representation. RF works better than kNN and SVM. It yields 84% prediction accuracy and 0.88 F-measure. kNN and SVM yield almost the same results. Prediction is done with 77% accuracy and 0.79 - 0.84 F-measure.

kNN with smiles vector representation yields same accuracy result with descriptor vector and with 0.85 F-measure. SVM gives best scores with smiles vector representation, 82% accuracy and 0.88 F-measure. RF's results are very close to SVM, prediciting with 79% accuracy and 0.84 F-measure.

Table 5.18. Permeability Glycoprotein Inhibitors Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.8184 (0.0060)	Topological FP	kNN	Accuracy	0.7768 (0.0070)
		Precision	0.8037 (0.0064)			Precision	0.8600 (0.0081)
		Recall	0.7821 (0.0065)			Recall	0.8033 (0.0097)
		F-measure	0.7928			F-measure	0.8307
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	1.0 (0)
		Precision	1.0 (0)			Precision	1.0 (0)
		Recall	1.0 (0)			Recall	1.0 (0)
		F-measure	1.0			F-measure	1.0
	RF	Accuracy	0.9554 (0.0100)		RF	Accuracy	0.8651 (0.0210)
		Precision	0.9832 (0.0099)			Precision	0.9227 (0.0146)
		Recall	0.9790 (0.0123)			Recall	0.8966 (0.0198)
		F-measure	0.9811			F-measure	0.9094
SmilesVec	kNN	Accuracy	0.7705 (0.0065)	DescriptorVec	kNN	Accuracy	0.7720 (0.0073)
		Precision	0.8842 (0.0079)			Precision	0.8120 (0.0067)
		Recall	0.8224 (0.0103)			Recall	0.7646 (0.0080)
		F-measure	0.8522			F-measure	0.7876
	SVM	Accuracy	0.8182 (0.0051)		SVM	Accuracy	0.7690 (0.0032)
		Precision	0.8984 (0.0048)			Precision	0.8672 (0.0039)
		Recall	0.8571 (0.0052)			Recall	0.8055 (0.0031)
		F-measure	0.8773			F-measure	0.8352
	RF	Accuracy	0.7935 (0.0102)		RF	Accuracy	0.8415 (0.0053)
		Precision	0.8651 (0.0083)			Precision	0.8989 (0.0043)
		Recall	0.8157 (0.0107)			Recall	0.8657 (0.0056)
		F-measure	0.8396			F-measure	0.8820

In figure 5.12, feature importances of random forest classifier on Pgp inhibitors are plotted. Here, most contributive descriptor is diameter. Other important descriptors are Balaban's J index and topological polar surface area. All top three descriptors correlate with each other because they are representatives of 3-dimensional molecular structure. We also see that the Petitjean index as a very contributive descriptor which is described as a measure molecular volume.

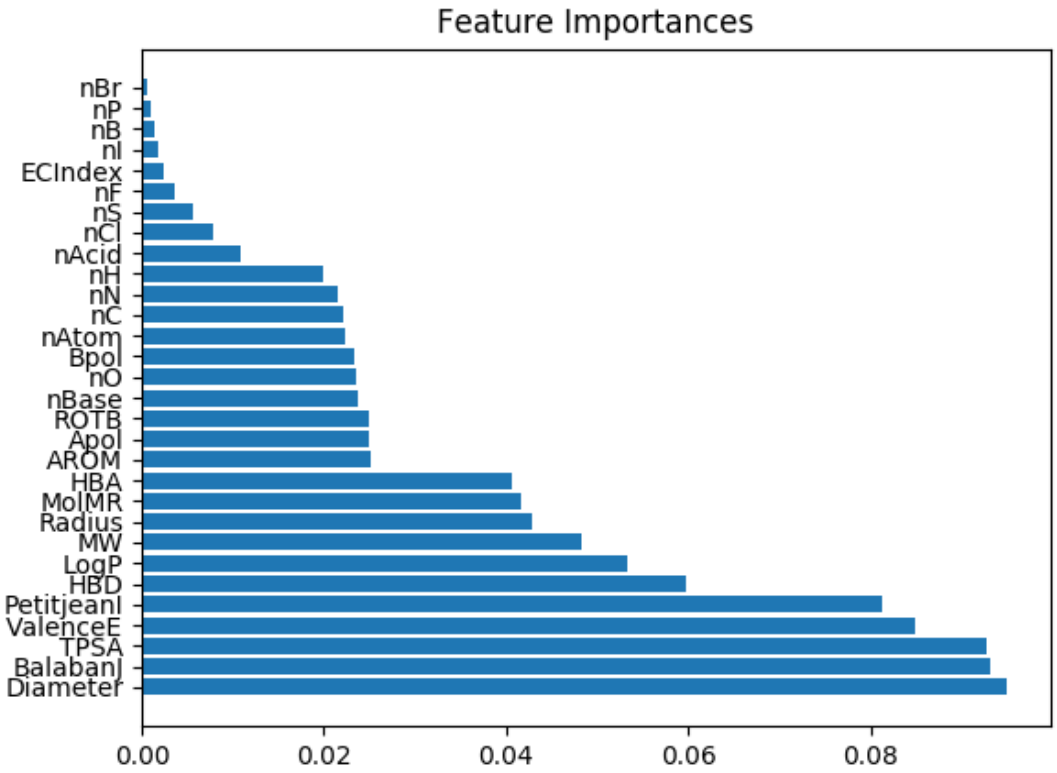

Figure 5.12. Feature importances of custom descriptor set for Pgp inhibitors dataset

5.4.10. Permeability Glycoprotein Substrates Results

In table 5.19, results of parameter tuning for Pgp substrates are displayed. In table 5.20 prediction scores of classifiers modelled with hyperparameters are shown.

Table 5.19. Hyperparameters for Pgp substrates

	kNN	**SVM**			**RF**			
	k	kernel	C	gamma	criterion	n_{est}	min_{leaf}	min_{split}
Morgan FP	7	rbf	10	0.001	entropy	30	1	4
Topological FP	3	linear	-	-	gini	30	1	6
SmilesVec	4	rbf	10	0.1	gini	20	1	5
DescVec	3	rbf	1	0.001	gini	20	1	3

For pgp substrates dataset, prediction accuracy of SVM classifier with morgan fingerprints is 100%. Random forest also yields pretty good accuracy that is 92 % and F-measure which is 0.94. kNN on the other hand, gives lower scores compared to SVM and RF. Prediction accuracy is 69% and F-measure is 0.66.

Topological fingerprints are also very successful with SVM classifiers, yielding a 100 % accuracy. Whereas RF gives 82% accuracy with 0.79 F-measure. Similar to morgan fingerprints, kNN gives lower results compared to SVM and RF classifiers. Prediction accuracy of kNN is 70% and F-measure is 0.70.

RF with descriptor vector representation gives better scores than all of the classifiers using SMILES vector representation however SVM with descriptor vectors yield very low results 50% accuracy and 0.31 F-measure. kNN yields 0.66 F-measure and prediciton accuracy is 64%.

Smiles vector representation gives around 0.62 accuracy for all of the classifiers. Maximum F-measure is achieved by kNN classifier, with 0.64 score. For this dataset, RF and kNN classifiers gives approximately same scores and SVM yields same accuracy scores whereas it is 10 percent behind in F-measure value.

Table 5.20. Permeability Glycoprotein Substrates Results (with standard deviations in parenthesis)

Morgan FP	kNN	Accuracy	0.6915 (0.0103)	Topological FP	kNN	Accuracy	0.6988 (0.0122)
		Precision	0.6746 (0.0697)			Precision	0.7299 (0.0175)
		Recall	0.6499 (0.0275)			Recall	0.6720 (0.0162)
		F-measure	0.6620			F-measure	0.6998
	SVM	Accuracy	1.0 (0)		SVM	Accuracy	1.0 (0)
		Precision	1.0 (0)			Precision	1.0 (0)
		Recall	1.0 (0)			Recall	1.0 (0)
		F-measure	1.0			F-measure	1.0
	RF	Accuracy	0.9240 (0.0176)		RF	Accuracy	0.8163 (0.0213)
		Precision	0.9604 (0.0163)			Precision	0.7985 (0.0292)
		Recall	0.9493 (0.0205)			Recall	0.7771 (0.0254)
		F-measure	0.9448			F-measure	0.7877
SmilesVec	kNN	Accuracy	0.6236 (0.0119)	DescriptorVec	kNN	Accuracy	0.6431 (0.0139)
		Precision	0.6948 (0.0461)			Precision	0.7067 (0.0226)
		Recall	0.5964 (0.0103)			Recall	0.6182 (0.0193)
		F-measure	0.6418			F-measure	0.6595
	SVM	Accuracy	0.6268 (0.0132)		SVM	Accuracy	0.50 (0.0447)
		Precision	0.4612 (0.0318)			Precision	0.50 (0.0050)
		Recall	0.5617 (0.0115)			Recall	0.2276 (0.0227)
		F-measure	0.5065			F-measure	0.3128
	RF	Accuracy	0.6350 (0.0213)		RF	Accuracy	0.6980 (0.0134)
		Precision	0.5910 (0.0276)			Precision	0.6843 (0.0221)
		Recall	0.5842 (0.0196)			Recall	0.6544 (0.0145)
		F-measure	0.5876			F-measure	0.6690

In figure 5.13, feature importances of random forest classifier on Pgp substates dataset are plotted. Here we can see the diameter descriptor is the most effective feature in our vector representation. Next two most contributive descriptors are Petitjean index and Balaban's J index as in the case of Pgp inhibitors. Results describe us the importance of 3-dimensional molecular structure in prediciton of P-glycoprotein inhibitors and substrates.

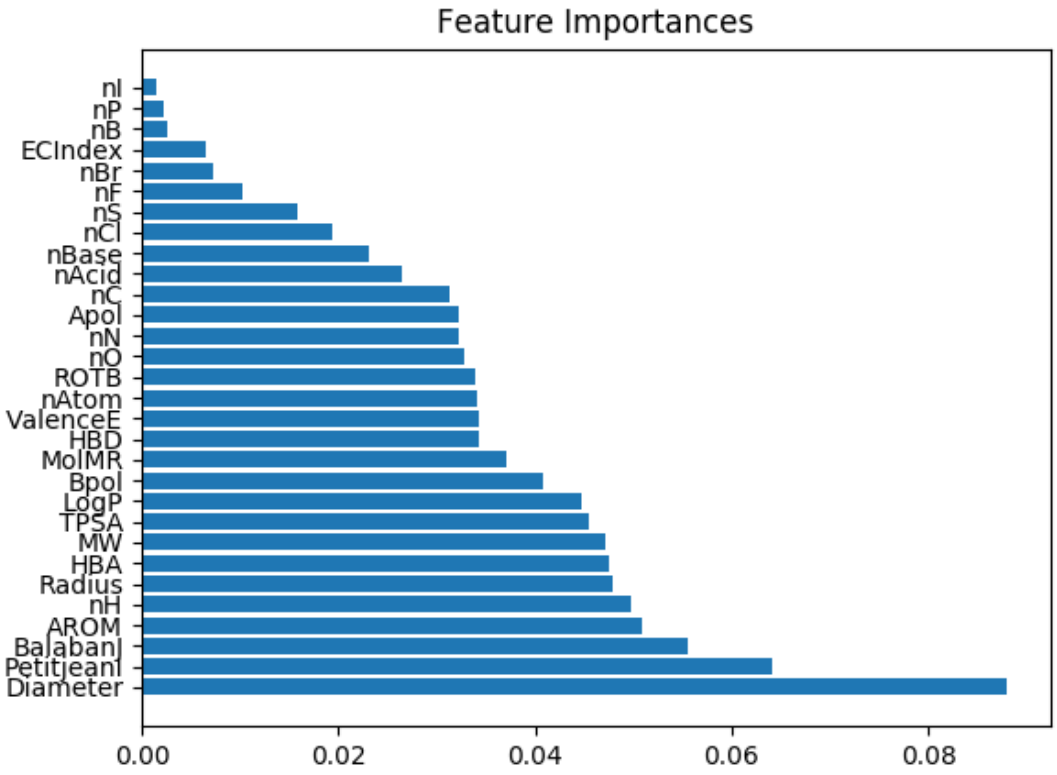

Figure 5.13. Feature importances of custom descriptor set for Pgp substrates dataset

Random forest using morgan fingerprints produces 92% accuracy and 0.93 F-measure. Since there is a room for a possible increase in the accuracy and F-measure score, we added diameter descriptor to morgan fingerprints. In table 5.21, you can see the scores of random forest classifier with updated morgan fingerprints.

Table 5.21. Results of Random Forest classifier with morgan fingerprints and updated morgan fingerprints (diameter descriptor is added)

	Morgan FP	Morgan FP + [diameter]
Accuracy	0.92	0.94
Precision	0.96	0.97
Recall	0.95	0.96
F-Measure	0.94	0.96

Here, adding diameter descriptor to the feature vector increased accuracy and F-measure by 2 percent and precision and recall by 1 percent which is very pleasing by adding just one numeric value increased our results.

6. DISCUSSION

In this study, we examined morgan fingerprint, topological fingerprint, Smiles vector and descriptor vector representations of molecules for new test samples predictions using k-nearest neighbor, support vector machine and random forest classifiers. We compare our random forest results with Schyman et al.'s VNN ADMET prediction tool [6] in table 6.1.

Table 6.1. Comparison of VNN and Random Forest classifier using morgan fingerprints (AMES, Cytotoxicity, DILI, HLM, MMP, Pgp inhibitors) or updated morgan fingerprint (BBB, hERG, Pgp substrates) representation

	VNN			Our results		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
AMES	0.82	0.86	0.75	0.99	0.99	0.99
BBB	0.90	0.94	0.86	0.95	0.95	0.95
Cytotoxicity	0.84	0.88	0.76	0.99	0.99	0.99
DILI	0.71	0.70	0.73	0.98	0.97	0.98
hERG	0.84	0.84	0.83	0.95	0.94	0.95
HLM	0.81	0.72	0.87	0.97	0.99	0.94
MMP	0.89	0.64	0.94	0.99	0.99	0.99
Pgp_inhibitors	0.85	0.91	0.73	0.96	0.98	0.97
Pgp_substrates	0.79	0.80	0.79	0.94	0.96	0.94

For our approach, random forest classifier makes predicitions with 94% and higher accuracy scores. In VNN ADMET predictor, they used an ECFP4 (diameter 4) it is a similar circular fingerprint to morgan fingerprints. Our representation is Rdkit's morgan with radius 2 fingerprints that are comparable with ECFP4. For all datasets, our approach yielded higher scores than VNN predictor.

For large datasets, having more than a thousand elements, random forest and support vector machine classifier yield prediction accuracy 97% and higher. For datasets

Table 6.2. Top 3 most important features for each dataset

Dataset	RF top 3 most important descriptors
AMES	Bpol (0.07), AROM (0.05), nH(0.05)
BBB	TPSA (0.20), LogP (0.14), nO (0.06)
Cytotoxicity	HBA (0.05), ROTB (0.05), TPSA (0.05)
DILI	HBA (0.05), nH (0.05), nN (00.05)
hERG	BalJ (0.08), diameter (0.08), nBase (0.07)
HLM	ROTB (0.06), TPSA (0.06), LogP (0.05)
MMP	LogP (0.08), Bpol (0.07), BalJ (0.06)
Pgp inhibitors	Diameter (0.10), BalJ (0.09), TPSA(0.09)
Pgp substrates	Diameter (0.09), PetitjeanI (0.06), BalJ (0.06)

having number of elements less than a thousand, SVM still yields prediction accuracy scores 98% accuracy and higher but random forest's accuracy score for these datasets are around 90%. With feature importance, we can see the most contributive molecular descriptors for each dataset and we improved our RF results for the datasets having less than a thousand elements. For each dataset, most important descriptors are shown in table 6.2.

Our results showed that morgan fingerprints that are a type of circular fingerprints, are very strong representations of molecular structure and it shows its highest scores using SVM classifier. Morgan fingerprints yield a 100% accuracy with SVM classifier on datasets that have more than approximately overall 700-750 samples since we witness only on hERG with 685 samples and BBB with 353 samples SVM do not give a 100% accuracy. Random forest also classifies test samples with very high accuracy with morgan fingerprints. Only for 3 datasets, we see an opportunity to make an improvement on RF classification success . These are BBB, hERG and Pgp substrates datasets.

Topological fingerprints are also good representations of molecules, in some cases they yield the same high results as morgan fingerprints with SVM classifier. How-

ever for RF, topological fingerprints are good feature vectors mostly for large datasets that have more than 6 thousands of elements but exceptions occur.For that reason we could not make any generalizations for topological fingerprints. In the case for large datasets, RF yields more than 90% accuracy. But for a small dataset it also yield 89% accuracy which proves that the success of topological fingerprints does not solely depend on dataset size.

Smiles vector representation on the other hand, prediction accuracy is mostly around 70%, there are only 2 exceptions, for MMP it gives 91% accuracy and for Pgp inhibitors it gives 80%. It is a light weight representation with a vector having only 100 dimensions. Once calculated and stored, it is fast to use for classification.

Descriptor vector representation's scores are greater than or equal to the scores of Smiles vector representation. Descriptor vector yields its highest scores using random forest classifier. SVM using descriptor vector is affected by the distribution of number of samples belonging to positive and negative classes. For HLM and MMP datasets, because the number of samples in each class is unbalanced, precision turns to be 0. However, there is an important aspect of descriptor vector representation. It provides us insight about the correlations between ADMET property and molecular descriptors. Using these insights, we had chance to improve our molecule representations.

As far as the classifiers are concerned, kNN is the fastest to train, tune the parameters and test, but its prediction accuracy is the lowest among others. It is also behind the VNN ADMET prediction tool [6].SVM fits best with fingerprints and Smiles vector representation however, training and parameter tuning phase is very long for polynomial and linear kernels, for especially fingerprint representations because of high dimensionality of fingerprints. RF also fits very good and its results are comparable with results of SVM. It is also faster than SVM but not as much as kNN.

All in all, for all cases, fingerprinting algorithms work better than descriptor vec-

tor representation. But training and parameter tuning phase with fingerprints takes time. Our parameter tuning, train and test phase for all datasets took approximately 18 hours with fingerprints whereas it took approximately 2,5 hours with descriptor vector representation. In addition, although fingerprints give best prediction scores, it is still a black box for understanding physicochemical properties of molecules and relations to ADMET properties. Descriptor vector representation on the other hand, even though it yields lower prediciton scores, allows us to understand molecular level properties and their effects in different environments of living bodies. This understanding may lead faster and more efficient drug discovery processes.

7. CONCLUSION AND FUTURE WORK

In this work, we examine 9 ADMET properties and relations of molecular descriptors with them. In table 6.2, most important top 3 molecular descriptors for each dataset are displayed. We see that hydrogen bonding acceptors and water-octanol partition coefficient (LogP) have a great significance about ADMET properties, as Lipinski suggested in his rule of 5 [34]. Topological polar surface area is also an important descriptor for living barrier penetration, hence we see TPSA as one of the most contributive descriptors. Topological indices namely Balaban's J index and Petitjean index which reflect the graph structure of molecules are also seen in the table 6.2. Similarly 2D diameter descriptor that is also constructed on molecular graphs has an important effect on predictions. Number of hydrogen, oxygen and nitrogen atoms are also contribute to prediction greatly on some datasets.

First future work to accomplish is to enhance the dataset size for training and test. For especially SVM classifier, datasets having unbalanced number of positive and negative classes face very low precision and F-score, since the classifier does not learn the characteristics of the class having less number of samples. We will collect new samples for each dataset and try to improve the prediciton accuracy rate for each class. Another future work is to select most relevant features according to feature importances and add new features similar to most relevant feature set that may be more contributive to the prediction. We will also gather the source code implemented in this work and serve as a publicly available tool for extracting relations between molecular descriptors and ADMET properties, and also for ADMET prediction.

REFERENCES

1. "Understanding the Drug Discovery Process", https://www.compoundchem.com/2016/01/16/dr discovery/.

2. Lipinski, C., F. Lombardo, B. Dominy and P. Feeney, "Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings", *Advanced Drug Delivery Reviews*, Vol. 64, pp. 4–17, 12 2012.

3. Norinder, U. and C. A. S. Bergström, "Prediction of ADMET Properties", *ChemMedChem*, Vol. 1, No. 9, pp. 920–937, September, 2006.

4. van de Waterbeemd, H. and E. Gifford, "ADMET in silico modelling: towards prediction paradise?", *Nature Reviews Drug Discovery*, Vol. 2, No. 3, pp. 192–204, 2003, `https://doi.org/10.1038/nrd1032`.

5. Walker, T., C. M. Grulke, D. Pozefsky and A. Tropsha, "Chembench: a cheminformatics workbench", *Bioinformatics (Oxford, England)*, Vol. 26, No. 23, pp. 3000–3001, Dec 2010, `https://www.ncbi.nlm.nih.gov/pubmed/20889496`, 20889496[pmid].

6. Schyman, P., R. Liu, V. Desai and A. Wallqvist, "vNN Web Server for ADMET Predictions", *Frontiers in pharmacology*, Vol. 8, pp. 889–889, Dec 2017, `https://www.ncbi.nlm.nih.gov/pubmed/29255418`, 29255418[pmid].

7. Cheng, F., W. Li, Y. Zhou, S. Jie, Z. Wu, G. Liu, P. Lee and Y. Tang, "admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties", *Journal of chemical information and modeling*, Vol. 52, 10 2012.

8. Yang, H., C. Lou, L. Sun, J. Li, Y. Cai, Z. Wang, W. Li, G. Liu and Y. Tang, "admetSAR 2.0: web-service for prediction and optimization of chemical AD-

MET properties", *Bioinformatics*, Vol. 35, No. 6, pp. 1067–1069, 08 2018, https://doi.org/10.1093/bioinformatics/bty707.

9. Daina, A., O. Michielin and V. Zoete, "SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules", *Scientific reports*, Vol. 7, pp. 42717–42717, Mar 2017, https://www.ncbi.nlm.nih.gov/pubmed/28256516, 28256516[pmid].

10. Sushko, I., S. Novotarskyi, R. KÃ¶rner, A. K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V. V. Prokopenko, V. Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, I. I. Baskin, V. A. Palyulin, E. V. Radchenko, W. J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de Sousa, Q.-Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko and I. V. Tetko, "Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information", *Journal of computer-aided molecular design*, Vol. 25, No. 6, pp. 533–554, Jun 2011, https://www.ncbi.nlm.nih.gov/pubmed/21660515, 21660515[pmid].

11. Miteva, M. A., S. Violas, M. Montes, D. Gomez, P. Tuffery and B. O. Villoutreix, "FAF-Drugs: free ADME/tox filtering of compound collections", *Nucleic acids research*, Vol. 34, No. Web Server issue, pp. W738–W744, Jul 2006, https://www.ncbi.nlm.nih.gov/pubmed/16845110, 16845110[pmid].

12. Sun, L. Z., Z. L. Ji, X. Chen, J. F. Wang and Y. Z. Chen, "ADME-AP: a database of ADME associated proteins", *Bioinformatics*, Vol. 18, No. 12, pp. 1699–1700, 12 2002, https://doi.org/10.1093/bioinformatics/18.12.1699.

13. Grzegorzewski, J., J. Brandhorst, D. Eleftheriadou, K. Green and M. König, "PK-DB: PharmacoKinetics DataBase for Individualized and Stratified Computational Modeling", *bioRxiv*, 2019, https://www.biorxiv.org/content/early/2019/09/09/760884.

14. Maunz, A., M. GÃŒtlein, M. Rautenberg, D. Vorgrimmler, D. Gebele and C. Helma, “lazar: a modular predictive toxicology framework”, *Frontiers in pharmacology*, Vol. 4, pp. 38–38, Apr 2013, https://www.ncbi.nlm.nih.gov/pubmed/23761761, 23761761[pmid].

15. Tom Mitchell, M. H., *Machine Learning*, 1997.

16. Cortes, C. and V. Vapnik, “Support-Vector Networks”, *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995, https://doi.org/10.1023/A:1022627411411.

17. Lam, L. and S. Y. Suen, “Application of majority voting to pattern recognition: an analysis of its behavior and performance”, *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, Vol. 27, No. 5, pp. 553–568, Sep. 1997.

18. Breiman, L., “Random Forests”, *Machine Learning*, Vol. 45, No. 1, pp. 5–32, Oct 2001, https://doi.org/10.1023/A:1010933404324.

19. Fawagreh, K., M. M. Gaber and E. Elyan, “Random forests: from early developments to recent advancements”, *Systems Science & Control Engineering*, Vol. 2, No. 1, pp. 602–609, 2014, https://doi.org/10.1080/21642583.2014.956265.

20. Ames, B. N., F. D. Lee and W. E. Durston, “An improved bacterial test system for the detection and classification of mutagens and carcinogens”, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 70, No. 3, pp. 782–786, Mar 1973, https://www.ncbi.nlm.nih.gov/pubmed/4577135, 4577135[pmid].

21. N. J., L., Rönnbäck and E. Hansson, “Astrocyte-endothelial interactions at the blood-brain barrier”, *Natural Reviews Neuroscience*, Vol. 7, pp. 41–53, 2006.

22. Pardridge, W. M., “Blood–brain barrier delivery”, *Drug Discovery Today*, Vol. 12, No. 1, pp. 54 – 61, 2007,

http://www.sciencedirect.com/science/article/pii/S1359644606004363.

23. Muehlbacher M, L. K. K. J., Spitzer GM, "Qualitative prediction of blood-brain barrier permeability on a large and refined dataset.", *Journal of computer-aided molecular design*, Vol. 25(12), p. 1095–1106, 2011.

24. Chen, M., A. Suzuki, J. Borlak, R. J. Andrade and M. I. Lucena, "Drug-induced liver injury: Interactions between drug properties and host factors", *Journal of Hepatology*, Vol. 63, No. 2, pp. 503 – 514, 2015, http://www.sciencedirect.com/science/article/pii/S0168827815002998.

25. Lamothe, S. M., J. Guo, W. Li, T. Yang and S. Zhang, "The Human Ether-a-go-go-related Gene (hERG) Potassium Channel Represents an Unusual Target for Protease-mediated Damage", *The Journal of biological chemistry*, Vol. 291, No. 39, pp. 20387–20401, Sep 2016, https://www.ncbi.nlm.nih.gov/pubmed/27502273, 27502273[pmid].

26. Lee, H.-M., M.-S. Yu, S. R. Kazmi, S. Y. Oh, K.-H. Rhee, M.-A. Bae, B. H. Lee, D.-S. Shin, K.-S. Oh, H. Ceong, D. Lee and D. Na, "Computational determination of hERG-related cardiotoxicity of drug candidates", *BMC bioinformatics*, Vol. 20, No. Suppl 10, pp. 250–250, May 2019, https://www.ncbi.nlm.nih.gov/pubmed/31138104, 31138104[pmid].

27. Finch, A. and P. Pillans, "P-glycoprotein and its role in drug-drug interactions", *Australian Prescriber*, Vol. 37, No. 4, pp. 137–139, 2014, https://app.dimensions.ai/details/publication/pub.1068690922 and https://cdn0.scrvt.com/08ab3606b0b7a8ea53fd0b40b1c44f86/955390525ab19130/d3a6

28. Borst, P. and R. O. Elferink, "Mammalian ABC Transporters in Health and Disease", *Annual Review of Biochemistry*, Vol. 71, No. 1, pp. 537–592, 2002, https://doi.org/10.1146/annurev.biochem.71.102301.093055, pMID: 12045106.

29. König, J., F. Müller and M. F. Fromm, "Transporters and Drug-Drug Interactions: Important Determinants of Drug Disposition and Effects", *Pharmacological Reviews*, Vol. 65, No. 3, pp. 944–966, 2013, http://pharmrev.aspetjournals.org/content/65/3/944.

30. Prof., D. R. T. D. V. C., *Molecular Descriptors for Chemoinformatics*, Vol. 41, Wiley[U+2010]VCH Verlag GmbH Co. KGaA, 7 2009.

31. Mauri, A., V. Consonni and R. Todeschini, *Molecular Descriptors*, 01 2017.

32. Roy, K., *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*, 01 2017.

33. Levin, V. A., "Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability", *Journal of Medicinal Chemistry*, Vol. 23, No. 6, pp. 682–684, 1980, https://doi.org/10.1021/jm00180a022.

34. Lipinski, C. A., F. Lombardo, B. W. Dominy and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1", *Advanced Drug Delivery Reviews*, Vol. 46, No. 1, pp. 3 – 26, 2001, http://www.sciencedirect.com/science/article/pii/S0169409X00001290, special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998.

35. Bhal, S. K., *LogP—Making Sense of the Value*, Application note, Advanced Chemistry Development, Inc.

36. Chen, M., J. Borlak and W. Tong, "High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury", *Hepatology*, Vol. 58, No. 1, pp. 388–396, 2013,

https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.26208.

37. Palm, K., K. Luthman, A.-L. Unge, G. Strandlund and P. Artursson, "Correlation of Drug Absorption with Molecular Surface Properties", *Journal of Pharmaceutical Sciences*, Vol. 85, No. 1, pp. 32 – 39, 1996, http://www.sciencedirect.com/science/article/pii/S0022354915499782.

38. van de Waterbeemd, H., G. Camenisch, G. Folkers, J. R. Chretien and O. A. Raevsky, "Estimation of Blood-Brain Barrier Crossing of Drugs Using Molecular Size and Shape, and H-Bonding Descriptors", *Journal of Drug Targeting*, Vol. 6, No. 2, pp. 151–165, 1998, https://doi.org/10.3109/10611869808997889, pMID: 9886238.

39. Veber, D. F., S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, "Molecular Properties That Influence the Oral Bioavailability of Drug Candidates", *Journal of Medicinal Chemistry*, Vol. 45, No. 12, pp. 2615–2623, Jun 2002, https://doi.org/10.1021/jm020017n.

40. Laurence, C. and M. Berthelot, "Observations on the strength of hydrogen bonding", *Perspectives in Drug Discovery and Design*, Vol. 18, No. 1, pp. 39–60, Jun 2000, https://doi.org/10.1023/A:1008743229409.

41. Hamaguchi, W., N. Masuda, S. Miyamoto, Y. Shiina, S. Kikuchi, T. Mihara, H. Moriguchi, H. Fushiki, Y. Murakami, Y. Amano, K. Honbou and K. Hattori, "Synthesis, SAR study, and biological evaluation of novel quinoline derivatives as phosphodiesterase 10A inhibitors with reduced CYP3A4 inhibition", *Bioorganic Medicinal Chemistry*, Vol. 23, No. 2, pp. 297 – 313, 2015, http://www.sciencedirect.com/science/article/pii/S0968089614008347.

42. Padron, J., R. Carrasco-Velar and R. Pellón, "Molecular descriptor based on a Molar Refractivity partition using Randic-type graphtheoretical invariant", *J. Pharm. Pharmaceut. Sci.*, Vol. 5, pp. 267–274, 10 2002.

43. Ghose, A. K., V. N. Viswanadhan and J. J. Wendoloski, "A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases", *Journal of Combinatorial Chemistry*, Vol. 1, No. 1, pp. 55–68, Jan 1999, https://doi.org/10.1021/cc9800071.

44. Balaban, A. T., "Highly discriminating distance-based topological index", *Chemical Physics Letters*, Vol. 89, No. 5, pp. 399 – 404, 1982, http://www.sciencedirect.com/science/article/pii/0009261482800092.

45. Ritchie, T. J. and S. J. Macdonald, "The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design?", *Drug Discovery Today*, Vol. 14, No. 21, pp. 1011 – 1020, 2009, http://www.sciencedirect.com/science/article/pii/S1359644609002785.

46. Manallack, D. T., R. J. Prankerd, E. Yuriev, T. I. Oprea and D. K. Chalmers, "The significance of acid/base properties in drug discovery", *Chemical Society reviews*, Vol. 42, No. 2, pp. 485–496, Jan 2013, https://www.ncbi.nlm.nih.gov/pubmed/23099561, 23099561[pmid].

47. Sharma, V., R. Goswami and A. K. Madan, "Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies", *Journal of Chemical Information and Computer Sciences*, Vol. 37, No. 2, pp. 273–282, Mar 1997, https://doi.org/10.1021/ci960049h.

48. Knipp, G. T., N. F. H. Ho, C. L. Barsuhn and R. T. Borchardt, "Paracellular Diffusion in Caco-2 Cell Monolayers: Effect of Perturbation on the Transport of Hydrophilic Compounds That Vary in Charge and Size", *Journal of Pharmaceutical Sciences*, Vol. 86, No. 10, pp. 1105–1110, Oct 1997, https://doi.org/10.1021/js9700309.

49. Hou, T. J., W. Zhang, K. Xia, X. B. Qiao and X. J. Xu, "ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties", *Journal of Chemical Information and Computer Sciences*, Vol. 44, No. 5, pp. 1585–1600, Sep 2004, https://doi.org/10.1021/ci049884m.

50. Petitjean, M., "Applications of the radius-diameter diagram to the classification of topological and geometrical shapes of chemical compounds", *Journal of Chemical Information and Computer Sciences*, Vol. 32, No. 4, pp. 331–337, 1992, https://pubs.acs.org/doi/abs/10.1021/ci00008a012.

51. Breindl, A., B. Beck, T. Clark and R. C. Glen, "Prediction of the n-Octanol/Water Partition Coefficient, logP, Using a Combination of Semiempirical MO-Calculations and a Neural Network", *Molecular modeling annual*, Vol. 3, No. 3, pp. 142–155, Mar 1997, https://doi.org/10.1007/s008940050027.

52. "Daylight fingerprint", https://www.daylight.com/meetings/summerschool01/course/ba

53. "Topological Fingerprints in RDKit", https://www.rdkit.org/UGM/2012/Landrum$_R DKit_U G$

54. Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.", *Journal of Chemical Information and Modeling*, Vol. 5, No. 2, pp. 107–113, 1965, https://app.dimensions.ai/details/publication/pub.1055225052.

55. Rogers, D. and M. Hahn, "Extended-Connectivity Fingerprints", *Journal of Chemical Information and Modeling*, Vol. 50, No. 5, pp. 742–754, May 2010, https://doi.org/10.1021/ci100050t.

56. Ã[U+0096]ztÃ¼rk, H., E. Ozkirimli and A. Ã[U+0096]zgÃ¼r, "A novel methodology on distributed representations of proteins using their interacting ligands", *Bioinformatics*, Vol. 34, No. 13, pp. i295–i303, 06 2018, https://doi.org/10.1093/bioinformatics/bty287.

57. Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger (Editors), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, Curran Associates, Inc., 2013.

58. Moriwaki, H., Y.-S. Tian, N. Kawashita and T. Takagi, “Mordred: a molecular descriptor calculator”, *Journal of Cheminformatics*, Vol. 10, No. 1, p. 4, 2018, https://doi.org/10.1186/s13321-018-0258-y.

59. MacÃ¡rio, I. P. E., H. Oliveira, A. C. Menezes, S. P. M. Ventura, J. L. Pereira, A. M. M. GonÃ§alves, J. A. P. Coutinho and F. J. M. GonÃ§alves, “Cytotoxicity profiling of deep eutectic solvents to human skin cells”, *Scientific Reports*, Vol. 9, No. 1, p. 3932, 2019, https://doi.org/10.1038/s41598-019-39910-y.