October 2021

BENCHMARKING SMALL-DATASET STRUCTURE-ACTIVITY-RELATIONSHIP MODELS FOR PREDICTION OF WNT SIGNALING INHIBITION

Mahtab Kokabi
University of Massachusetts Amherst

BENCHMARKING SMALL-DATASET STRUCTURE-ACTIVITY-RELATIONSHIP MODELS FOR PREDICTION OF WNT SIGNALING INHIBITION

A Thesis Presented

by

MAHTAB KOKABI

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

Master of Science in Electrical and Computer Engineering

September 2021

BENCHMARKING SMALL-DATASET STRUCTURE-ACTIVITY-RELATIONSHIP MODELS FOR PREDICTION OF WNT SIGNALING INHIBITION

A Master Thesis Presented

by

MAHTAB KOKABI

Approved as to style and content by:

_____
Guangyu Xu, Chair

_____
Weibo Gong, Member

_____
Lixin Gao, Member

_____
Christopher Hollot, Department Head
Electrical and Computer Engineering

DEDICATION

To my beloved family - My husband for his endless support, my parents for all the love and encouragement towards my life, and my brother and sister who have always been the closest friend to me.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Professor Guangyu Xu for his outstanding support and mentorship through my M.Sc. studies. This thesis would have not been possible without his advice, knowledge, and encouragement. I would also like to thank my thesis committee: Prof. Gong, and Prof. Gao for their insightful comments and feedback regarding this Thesis.

I want to thank my labmate, Matthew Donnelly for all his help, support, and knowledge toward this research. This thesis would not have been possible without his unselfish help in my research project.

I want to thank my husband, who I have shared the ups and downs of my life with, and who helped me throughout all the challenges I had throughout these years.

Above all, I would like to thank my parents for their support and encouragement from thousand miles away. Every step that I took in my life and academic journey comes from the confidence, strength, and love they gave me; for that, I am so grateful.

ABSTRACT

BENCHMARKING SMALL-DATASET STRUCTURE-ACTIVITY-RELATIONSHIP MODELS FOR PREDICTION OF WNT SIGNALING INHIBITION

SEPTEMBER 2021

MAHTAB KOKABI

B.Sc., AMIRKABIR UNIVERSITY

M.S.E.C.E, UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Guangyu Xu

Quantitative structure-activity relationship (QSAR) models based on machine learning algorithms are powerful tools to expedite drug discovery processes and therapeutics development. Given the cost in acquiring large-sized training datasets, it is useful to examine if QSAR analysis can reasonably predict drug activity with only a small-sized dataset (size < 100) and benchmark these small-dataset QSAR models in application-specific studies. To this end, here we present a systematic benchmarking study on small-dataset QSAR models built for prediction of effective Wnt signaling inhibitors, which are essential to therapeutics development in prevalent human diseases (e.g., cancer). Specifically, we examined a total of 72 two-dimensional (2D) QSAR models based on 4

best-performing algorithms, 6 commonly used molecular fingerprints, and 3 typical fingerprint lengths. We trained these models using a training dataset (56 compounds), benchmarked their performance on 4 figures-of-merit (FOMs), and examined their prediction accuracy using an external validation dataset (14 compounds). Our data show that the model performance is maximized when: 1) molecular fingerprints are selected to provide sufficient, unique, and not overly detailed representations of the chemical structures of drug compounds; 2) algorithms are selected to reduce the number of false predictions due to class imbalance in the dataset; and 3) models are selected to reach balanced performance on all 4 FOMs. These results may provide general guidelines in developing high-performance small-dataset QSAR models for drug activity prediction.

Keywords: Bioactivity prediction, drug discovery, machine learning, molecular fingerprint, quantitative structure-activity relationship, Wnt signaling.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1. Motivation

Drug development often involves extensive investment and time effort on experimental screening of drug candidates. On average, getting a potential drug candidate from laboratory to the pharmacy takes about 14 years, costs more than one billion dollars, and has a low success rate [1, 2]. To reduce the resource demand in such drug screening processes, predictive models based on advanced computational methods have been developed to help screen possible drug compounds with high cost-effectiveness [3-7]. To date, computational methods based on three-dimensional quantitative structure-activity relationship (3D QSAR) analysis, high-throughput imaging (HTI), and pharmacophore modeling [7-12], have succeeded in predicting the effectiveness of drug compounds towards prevalent human diseases (e.g., cancer [10]).

1.1.1. Three-dimensional quantitative structure-activity relationship

Three-dimensional quantitative structure-activity relationship (3D QSAR) analysis is a methodology with major applications in computer-assisted molecular design (CAMD) [13]. This technique has served as a valuable predictive tool in the design of pharmaceuticals and agrochemicals [14]. In general, 3D QSAR techniques require an appropriate spatial superimposition and three-dimensional structures of the molecules with known activities [11]. Various procedures have been pursued for this crucial step, many of them being entirely manual or at least requiring user intervention [11]. In addition to the

subjective nature of a user-supervised alignment, the enormous time effort makes such an approach inappropriate for large screening scenarios [11].

1.1.2. High-throughput imaging

High-throughput imaging (HTI), also known as high-content screening (HCS), captures the morphological features of the cell and its organelles by microscopy, which has yielded variety of biological discoveries [9, 15, 16]. In this method, a set of features including shape, spatial metrics, intensity, and patterning of fluorescently labeled markers, are used to describe chemical compounds. These features can be considered as an image-based compound fingerprint [9]. Predicting the activities of compounds using this computational method often requires high resolutions images that are not available for all drug compounds. The applications of this method are limited due to high computational cost and advanced hardware requirement for processing the high-resolution images [9, 16].

1.1.3. Pharmacophore modeling

Pharmacophore modeling has become one of the major tools in drug discovery field. Specifically, a pharmacophore model can be defined based on two approaches: ligand-based and structured-based pharmacophore modeling. Ligand-based methods extract the chemical features from 3D structures of a set of known ligands. The main challenge faced in this method is the modeling of ligand flexibility, which is used to represent the internal degrees of freedom of the ligands [10]. On the other hand, structured-based method works directly with the 3D structure of a macromolecular target, which requires efficient molecular alignment algorithms and highly accurate model optimization [10]. Similar to 3D QSAR technique, pharmacophore modeling needs efficient molecular

and ligand alignment, which is considered as the main difficulty in pharmacophore modeling [10].

To sum up, these high-performance methods often require user intervention steps on molecular/ligand alignment [7], [10, 11] or high-resolution images that are not available for all drug compounds [9]. For these reasons, two-dimensional (2D) QSAR analysis has emerged as a viable alternative method to build predictive models from the widely available chemical structures of drug candidates, which can perform well with no user intervention steps.

1.2. Two-dimensional (2D) QSAR

2D QSAR methods have been applied to the development of relationship between properties of chemical substances and their biological activities to obtain a reliable statistical model for activity prediction of new chemical entities [3]. This analysis correlates the structural details of drug molecules to their effectiveness in biological assays that correspond to specific diseases and builds models that can predict the bioactivity or physiochemical properties of unknown drug compounds [3-5], [8]. This method can reduce the costly failures of drug candidates by identifying promising lead compounds and reducing the number of costly experiments [8].

In 2D QSAR studies, the features of each drug molecule are often coded by a 2D molecular fingerprint, resulting in a numerical vector to describe the presence or absence of substructures in the molecule such as chemical bonds, functional groups, and connectivity pathways [5, 17]. The vectors from drug molecules with known effectiveness to one targeted biological assay (active vs. inactive) will be used to build predictive QSAR

models based on machine learning algorithms such as support vector machines (SVM), decision trees, k-nearest neighbors (KNN), and artificial neural network (ANN) [8, 18]. The resulting QSAR models have succeeded in predicting effective drugs of psychological disorders [19], protein-ligand binding affinities [20], and mTOR kinase inhibitors [21].

Nonetheless, current 2D QSAR analysis often relies on training machine learning algorithms with a large-sized drug activity dataset (size > 1000) [8, 22], which requires significant time and effort on both benchwork and statistical analysis. Given the cost in acquiring these large-sized datasets, it will be useful to examine if 2D QSAR analysis can result in reasonable prediction of drug activity with only a small-sized dataset (size < 100), and moreover benchmark these small-dataset QSAR models in application-specific studies. Such small-dataset QSAR analysis will be especially beneficial at early stages of drug development, when the activity data from potential drug candidates remain limited [23, 24].

1.3. Wnt Signaling

Wnt signaling pathways are essential in cell biology and the development of therapeutics for highly prevalent diseases such as cancer, Schizophrenia, and kidney damage [25-29]. Some of these diseases (e.g., lung cancer) are associated with altered function/levels of proteins in specific Wnt/β-catenin pathways (one type of Wnt signaling pathway), which lead to elevated gene expression that influences cell proliferation and survival [25]. For this reason, inhibition of Wnt/β-catenin signaling by small molecule modulators (e.g., Niclosamide) is being considered and developed as a candidate cancer treatment [25],[30-33]. For instance, screening assays based on live cell imaging have been

used to identify Wnt/β-catenin inhibitors [34]. These inhibitors induce the internalization of Frizzled receptor proteins (i.e. moving from cell membrane to cell cytoplasm) in human U2OS cells; such internalized receptors cannot be activated by extracellular Wnt proteins (secreted from other cells), effectively inhibiting the strength of Wnt signaling [25].

1.4. Research objective

Given the clinical significance of Wnt signaling in a variety of diseases and the progress made from screening assays, here we examine if small-dataset QSAR models could facilitate and expedite the process of identifying small molecule inhibitors. If successful, such predictive models and experimental QSAR studies can serve as complementary techniques in screening drug candidates for Wnt/β-catenin signaling inhibition and ultimately add to therapeutics development. To quantify the performance in our analysis, we benchmark 72 QSAR models based on: 1) 4 machine learning algorithms including quadratic support vector machine (QSVM), fine tree, random undersampling (RUS) boosted tree, and bagged tree; 2) 6 molecular fingerprints including fingerprint 2, 3, 4 (FP2, FP3, FP4), molecular access system fingerprint (MACCS), and extended-connectivity fingerprint 4 and 6 (ECFP4 and ECFP6) with three fingerprint lengths for each; and 3) a training dataset of 56 compounds and an external validation dataset of 14 compounds, both of which were experimentally tested in U2OS cells. We evaluate these models using 5- and 10-fold cross-validation and compare 4 figures of- merit (FOMs) in QSAR analysis including accuracy, area under curve (AUC), sensitivity, and specificity.

CHAPTER 2

BACKGROUND AND LITERATURE SURVEY

2.1. Related work

Studying bioactive molecular compounds is necessary for the process of drug development. Biological activity of these compounds needs to be predicted to determine the drug-target ability. As development and production of drugs require substantial amount of time, it is important to predict bioactive molecules with models having high predictive performance [22]. Computational methods based on machine learning approaches have shown great prospect in predicting the molecular activities and became a pivot tool for many projects in bioinformatics, and health informatics.

Many studies have been done within the literature to assess the performance of computational methods in prediction of molecular of activity. Simm et al. [9] demonstrated a computational method to predict the activities of compounds in hundreds of biological assays from a single image-based screen of half a million compounds. First, they extracted features from image-based cellular assay. The resulting set of features, which include not only shape and spatial metrics but also the intensity and patterning of fluorescently labeled markers. These features can be used to describe chemical compounds and can be considered as an image-based compound fingerprint. Second, they implemented Deep Neural Network (DNN), concretely feedforward artificial neural networks, to train a model to predict the bioactivity of new compounds. In the first layer of the network (the input layer), the neurons obtain an input vector which is the image-based fingerprint. The intermediate layers (hidden layers) comprise the hidden neurons that have weighted

connections to the neurons of the previous level layer and can be considered as abstract features. The last layer (the output layer) supplies the predictions of the model. They achieved up to 90% accuracy [9]. This study suggested that features from high-content screens or high- throughput imaging are a rich source of information that can be used to predict the molecular activity and replace customized biological assays.

In contrast, there are several studies have been done in the field of computational modeling based on deriving the features from chemical structures of compounds to predict their activity in assays. As an example, Myint et al. [8] have reported a novel 2D fingerprint-based artificial neural network QSAR (FANN-QSAR) method in order to effectively predict biological activities of structurally diverse chemical ligands. In this study, they used three types of molecular fingerprints namely fingerprint 2 (FP2), molecular access fingerprint (MACCS), and extended connectivity fingerprint 6 (ECFP6) (see section 3.2.) to train artificial neural network. In this study, a feed-forward neural network method was implemented using MATLAB R2007b Neural Network Toolbox. The number of hidden layer neurons was varied between 100 and 1000. They achieved an average mean square error of 75% [8]. This study demonstrated that combination of molecular fingerprints from chemical structures and ANN can lead to a reliable and robust method and can be a useful tool in computer-aided drug discovery research.

In addition, the resulting QSAR models have succeeded in predicting effective drugs of psychological disorders [19], protein-ligand binding affinities [20], and mTOR kinase inhibitors [21]. The predictive QSAR models based on machine learning algorithms such as support vector machine (SVM), decision tree, k-nearest neighbors (KNN) have shown promising ability in predicting biological activities of compounds. For example, Darnag et

al. [35] developed quantitative relationships to predict anti-HIV activity based on SVM algorithm. Many different techniques for QSAR modeling have been found useful for the establishment of the relationships between molecular structures and anti-HIV activity [36-38]. Most of these QSAR models have developed neural networks algorithms to predict anti-HIV activity. However, these neural systems have some problems inherent to its architecture such as over training, overfitting and network optimization [35]. In addition, neural networks algorithms require large dataset in order to train the model. Thus, it will be useful to employ other machine learning algorithms to evaluate the model. In this study [35] Support vector machines were applied to build up the QSAR model for predicting the anti-HIV-1 activity of 82 compounds, based on features calculated from molecular structure. In this work, they validated the models based on three different algorithms namely SVM, ANN, and multi linear regression (MLR) with the accuracy of 96%, 90%, and 80% respectively. The results showed that the SVM technique was able to establish a satisfactory relationship between the molecular descriptors and the anti-HIV-1 activity [35]. It has been shown in this study that SVM give a superior performance comparing to ANN, and MLR.

There are several studies offers the latest insights and approaches at targeting the Wnt/β-catenin pathway in various cancer diseases such as colorectal cancer, melanoma, leukemia, breast and lung cancers [39]. Biochemical and genetic data support the idea that inhibition of Wnt/β-catenin signaling is beneficial in cancer Therapeutics [39]. Binding of Wnt proteins to Frizzled receptors results in β-catenin being released into the cytoplasm, where its concentration increases and influences many cellular processes, including development and differentiation [39]. There are numerous quantitative studies performed

on other assays such as three-channel glucocorticoid receptor (GCR), inhibition of HIV integrase in a cell based, and inhibition of human thrombin to predict the bioactivity of molecules by implementing machine learning algorithms [9, 40, 41]. However, few quantitative studies have been performed on Wnt signaling to predict the bioactivities to the best of our knowledge. As an example Chen et al. [42] applied machine learning algorithms on Wnt/β-catenin, carbohydrate metabolism, and PI3K-Akt signaling pathway-related genes to classify tumors and normal samples. In this study, four machine learning methods namely support-vector machines, random forest, decision tree, and k-nearest neighbor algorithms was used to assess the accuracy of the mentioned genes in predicting colorectal cancer. Their results showed areas under the curve exceeding 95.00% for cancer outcomes [42].

CHAPTER 3

METHODS

3.1. Dataset

ChEMBL is an open large-scale bioactivity database containing 2-D structures, calculated properties such as Molecular Weight, and abstracted bioactivities of each compound. This database contains more than 1.6 million distinct compound structures with 14 million activity values experimentally tested from 1.2 million assays. These assays are mapped to around 11000 targets, including 9052 proteins [43]. In this study 70 drug compounds obtained from ChEMBL database. To the best of our knowledge, all these compounds are available in literature with experimentally validated effectiveness for internalizing Frizzled receptor proteins, and thus inhibiting Wnt signaling in human U2OS cells [44-47].

There are two main reasons for selecting this dataset. Firstly, because these 70 compounds are all experimentally validated with the same biological assay (i.e., the internalization of Frizzled receptor proteins in U2OS cells). It is noted that assays targeted at the dynamics of other Wnt signaling- inhibition related proteins are also available in the ChEMBL database (Table 1) [43, 48-52]. However, these assays have yet to experimentally test a sufficient number of active or inactive compounds, therefore making the QSAR modeling challenging (e.g., 3 active compounds in [52]). Secondly, we found that the size of our dataset is on par with other small-dataset QSAR studies (e.g., 16 in [53], and 48 in [54]); we thus believe this dataset has a sufficient number of data to build good-performing QSAR models.

Table 1. Summary of Assays for Wnt Signaling Inhibition

Assay ChEMBL ID	Cell Type	Target ChEMBL ID	Number of Compounds
CHEMBL2354178 [44], CHEMBL3606285 [45],CHEMBL3994283 [46], CHEMBL4276087 [47] Inhibition of Wnt/beta-catenin in human U2OS cells assessed as internalization of frizzled-GFP at 12.5 uM after 6 hrs by confocal microscopic analysis	U2OS	CHEMBL612545 CHEMBL2346493 Frizzled-1 (SINGLE PROTEIN)	29 active 41 inactive
CHEMBL1072837 [48] Inhibition of human recombinant SFRP1 expressed in human U2OS cells assessed as increase in Wnt signaling after 16 to 18 hrs by luciferase reporter gene assay	U2OS	CHEMBL5517 Secreted frizzled-related protein 1 (SINGLE PROTEIN)	9 inactive
CHEMBL3788739 [49] Inhibition of human TERT-regulated Wnt/beta-catenin signaling in human MGC803 cells assessed as decrease in cyclin D1 mRNA expression at 40 umol/L measured at 48 hrs by RT-PCR method	MGC-803	CHEMBL2916 Telomerase reverse transcriptase (SINGLE PROTEIN)	1 active
CHEMBL3101279 [50] Inhibition of CK2-mediated Wnt signaling in human MCF7 cells assessed as reduction of beta-catenin level at 1 uM after 15 mins by Western blot analysis	MCF7	CHEMBL2095191 Casein kinase II (PROTEIN COMPLEX GROUP)	1 active
CHEMBL1251260 [51] Inhibition of GSK-3-beta-mediated Wnt signaling in human ST14A cells assessed as increase in accumulation of beta-casein around nucleus after 6 hrs by microscopic analysis	ST14A	CHEMBL262 Glycogen synthase kinase-3 beta (SINGLE PROTEIN)	2 active
CHEMBL1251261 [51] Inhibition of GSK-3-beta-mediated Wnt signaling in human ReNcell VM cells assessed as increase in accumulation of beta-casein at <3 uM after 2 hrs by ELISA relative to control	ReNcell VM	CHEMBL262 Glycogen synthase kinase-3 beta (SINGLE PROTEIN)	2 inactive
CHEMBL4008143 [52] Agonist activity at PPARgamma in human HT-29 cells harboring APC mutant assessed as inhibition of Wnt/beta-catenin signaling pathway by measuring decrease in c-Myc level at 10 uM treated for 24 hrs by Western blot method	HT-29	CHEMBL235 Peroxisome proliferator-activated receptor gamma (SINGLE PROTEIN)	3 active

All of the 70 compounds were experimentally classified as active or inactive in which 29 compounds were tested to be active, and 41 compounds to be inactive. Generally, Active Compound means a compound that specifically inhibits, stimulates or alters the

production or activity of a Target. In our dataset, a compound has been classified as active [inactive] compounds if it was able [unable] to induce the internalization of Frizzled receptor proteins, according to the cell imaging data from before and after applying the compound to the cell culture.

To generate 2D QSAR models, we represented the chemical structures of these 70 compounds listed in the ChEMBL database in a simplified molecular-input line-entry system (SMILES) notation. Each compound was labeled as 0 (inactive) or 1 (active) by its effectiveness on Wnt signaling inhibition, which was tested in the assay of internalizing Frizzled receptor proteins (Fig. 1). In order to evaluate 2D QSAR methods, dataset was split into two nonoverlapping subsets: a training set and a test set. To do this, we shuffled the data and selected 80% of these 70 compounds for training process (56 compounds; 31 inactive, 25 active) to develop 2D QSAR models and perform cross-validation to statistically analyze their performance. Then we selected the remaining 20 % of these 70 compounds (14 compounds; 10 inactive, 4 active) as an external validation dataset to examine if these QSAR models can predict the activity of compounds that were not used in the training model [55].

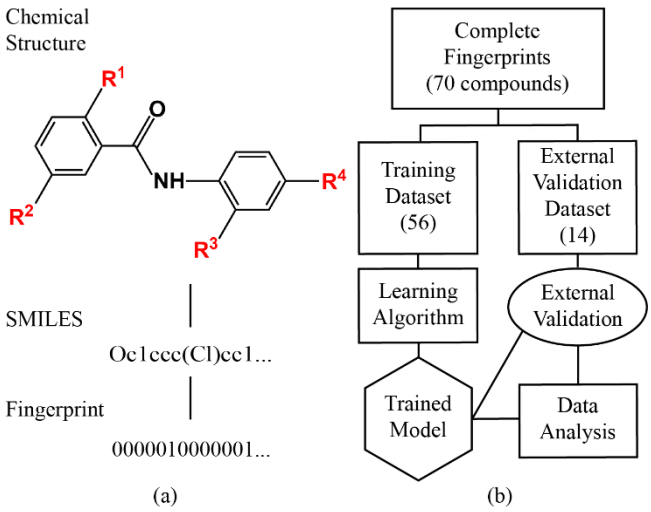

Figure 1. Schematic diagrams on benchmarking small-dataset QSAR models. $R^1 - R^4$represent the structural features in the compound.

3.2. Fingerprint Representation

Each molecule is represented by a list of features, i.e. "descriptors" in QSAR nomenclature. In this study, we used SMILES strings as the textual representation of molecules. SMILES is a linear notation for representing molecular structures. For SMILES to be processed by machine learning models, they need to be transformed into numeric representations. This numeric representation is called fingerprint. Each of these fingerprint representations is a binary bit vector with a defined length; each bit or group of bits represents the presence or absence of structural features in the compound. In order to generate the fingerprint, we used OpenBabel graphical user interface (GUI) to convert the SMILES notation of the compounds in the training dataset to 2D molecular fingerprint representations (Fig. 1) [56-58]. For instance, the niclosamide compound is represented by MACCS fingerprint with a length of 128 in the following steps: 1) finding the SMILES notation of niclosamide in the ChEMBL database, ODC(Nc1ccc([NC] (DO) [O-])cc1Cl)c1 cc(Cl)ccc1O; 2) converting this SMILES notation in the OpenBabel GUI to a hexadecimal vector 4a5124612940006 04091001f7aebecf6; and 3) In the last step, we should convert

the hexadecimal vector to a binary one, 010010100101000100100100011000010010100101000000000000000011000000100000010010001000000000001111101111010111010111111010110011110110. The resulting binary vector and the effectiveness of niclosamide for Wnt signaling inhibition (active) will then be used to train QSAR models using MATLAB Classification Learner application (see Section 3.3).

Molecular descriptors can be divided into four broader categories: constitutional (1D), topological (2D), geometrical (3D), and physico-chemical properties-based (4D) descriptors. By far, the most commonly used structural representation for comparing molecules is the use of two-dimensional (2D) molecular fingerprints [59]. In this work, we chose to benchmark QSAR models using 3 linear 2D fingerprints (FP2, FP3, and FP4) and 3 nonlinear 2D fingerprints (MACCS, ECFP4, and ECFP6). These fingerprints are computationally effective and have been broadly used in drug activity prediction based on solubility, permeability, and protein-ligand interactions [60, 61].

3.2.1. Linear 2D Fingerprint

FP2, FP3, and FP4 are subset of linear Fingerprints in which all linear paths containing up to seven bonds are mapped to a structure. Then, a hashing operation is performed on a string-based description of each linear fragment to produce an integer bit address [57]. FP2 (default length: 1024) is a path-based fingerprint which recognizes the rings and linear substructures in drug molecules of lengths 1 to 7 (excluding the 1-atom substructures C and N) and then maps them onto a bit-string of length 1024 using a hash function [56]. FP3 (default length: 64) and FP4 (default length: 512) are substructure-based fingerprints to mark sub-structural patterns by SMILES arbitrary target specification (SMARTS). In addition, FP4 consists of a set of SMART patterns corresponding to functional groups [56].

3.2.2. Nonlinear 2D Fingerprint

MACCS (default length: 256) is substructural-key based fingerprint using 166 structural keys to characterize SMARTS patterns where each specific bit position represents the presence (1) or absence (0) of predefined functional groups, substructure motifs, or fragments [59], [56], [62]. ECFPs are a novel class of 2D circular fingerprints used for molecular characterization. These circular fingerprints have many useful qualities, including: (i) being fast to calculate; (ii) representing a very large number of different features; and (iii) not relying on predefined features; thus, they can represent novel structural variation [63]. Specifically, ECFP4 and ECFP6 (no default lengths) are circular fingerprints stemming from the Morgan algorithm [56], [64] and are explicitly designed to capture molecular features related to molecular activity.

3.3. Algorithms

Using the fingerprint representations of 56 compounds in the training dataset with known activity for Wnt signaling inhibition, we developed predictive QSAR models based on four machine learning algorithms: QSVM, fine tree, bagged tree, and RUSboosted tree. We selected these algorithms in our benchmarking study since their resulting QSAR models showed the highest accuracy and AUC values among 25 available algorithms in MATLAB Classification Learner application.

3.3.1. QSVM

Support Vector Machine (SVM) was first proposed by Vapnik [65]. It is based on finding an optimal hyper-plane which separates the data into two classes with the largest margin. Some applications of the SVM are: Histogram-based Image Classification [66], Spam Categorization [67], Face Membership Authentication [68], and data analysis and

classification [68]. Quadratic Support Vector Machine (QSVM) is a quadratic kernel-free non-linear support vector machine [69] which is a binary classifier to define an optimal hyperplane that maximally separates two classes of high-dimensional data [65].

3.3.2. Fine tree

Fine tree is subset of decision tree algorithms. The goal of these algorithm is to create a model that predicts the value of a target variable. The decision tree algorithm uses a tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree. Fine tree algorithm (abbreviated as Fine) uses up to 100 decision rules (i.e. decision tree) for precise classification of the data [70].

3.3.3. Bagged tree

Bagged tree algorithm (abbreviated as Bagged) is subset of decision tree algorithms. This algorithm first forms several subsets of data that are randomly sampled from the entire training dataset with replacement [71]. Each subset of data will be used to train a decision-tree based sub-model. This algorithm finally makes a robust classification of an unknown data by either voting or averaging the prediction results of this data from all sub-models [72].

3.3.4. RUSboosted tree

RUSboosted tree algorithm (abbreviated as RUSboosted) iteratively trains a series of decision-tree based sub-models, each of which is based on a subset of data formed by randomly under-sampling the majority class of the training dataset to alleviate the class imbalance [73, 74]. During the iteration, each data used for internal validation will increase its weight if it was incorrectly classified during the previous iteration, so that it is likely to

be correctly classified in the current iteration. For this reason, the decision tree upon the completion of the iteration is a weighted vote from all involved sub-models and will be used to classify unknown data.

3.4. Model Assessment

To benchmark our models, we first studied the dependence of their FOMs on the cross-validation folding number k and the fingerprint length, respectively. We then benchmarked the FOM values of these models using the preferred k value and fingerprint lengths, followed by evaluating their capability to predict the activity of the 14 compounds in the external validation dataset. To evaluate the statistical significance in our results, 1) all these models were trained for 3 independent times to obtain the mean values and the standard deviation of all 4 FOMs; 2) selected models (see details below) were then applied to the external validation dataset to obtain the mean values and the standard deviation of correct predictions. All QSAR models were trained and validated using the MATLAB Classification Learner application, detailed as follows:

3.4.1. Folding Number K

During the training of QSAR models, we applied the k-fold cross-validation procedure [75]. In k-fold cross validation (k-cv), the data set is divided into k folds, a classifier is learned using k-1 folds, and an error value is calculated by testing the classifier in the remaining fold. Finally, the k-cv estimation of the error is the average value of the errors committed in each fold. Thus, the k-cv error estimator depends on two factors: the training set and the partition into folds [76]. Specifically, we compared the 4 FOM values in 72 QSAR models with both 5- and 10-fold cross validation, which are commonly used in training machine learning models and they are less biased [75, 76]. We then chose one

preferred k value for the rest of our analysis based on the overall performance of these 72 models.

3.4.2. Fingerprint Length

Molecular fingerprints are often very different in length and complexity, ranging from 2D/simple representations of relevant structural features to 3D/complicated pharmacophore arrangements. Thus, many types of fingerprints have been generated with different settings (generation method, length, size of patterns, and number of bits activated by each pattern, etc.) and are further deployed as descriptors for predictive modeling to estimate biological activities [59]. Therefore, one of the most important part of QSAR modeling is to find the best fingerprint setting in terms of length, simplicity, size, and uniqueness [57].

In this study, we evaluated the FOM values in 24 models (based on 6 fingerprints by 4 algorithms) with 3 different fingerprint lengths, aiming to balance simplicity, resolution, and uniqueness of the fingerprint representations [57, 77]. For FP2, FP3, FP4, and MACCS, we trained our models using: 1) half the default length, 2) the default length, and 3) double the default length, and chose one preferred length for each fingerprint that yielded higher FOM values than the other two lengths (see details below). For ECFP4, we chose lengths of 2048, 4096, and 8192, whereas for ECFP6, we chose lengths of 1024, 2048, and 4096 in our analysis, because ECFP6 has no default length reported and its performance was suggested to likely improve when the length increases [78].

3.4.3. Model FOMs

We next benchmarked the 4 FOM values of FP2, FP3, FP4, and MACCS models with their chosen fingerprint lengths and those of ECFP4 and ECFP6 models with all three

lengths (a total of 40 models based on 4 algorithms by 10 lengths, evaluated at the chosen k value). For each model, we analyzed its confusion matrix results in the MATLAB classification learner toolbox to obtain the values of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Here TPs [FPs] refer to the number of correct [incorrect] predictions of active compounds, whereas TNs [FNs] refer to the number of correct [incorrect] predictions of inactive compounds. We then obtained the four FOM values as accuracy, sensitivity, specificity, and area under curve (AUC).

3.4.3.1. Accuracy

The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified.

3.4.3.2. Sensitivity

Sensitivity measures the proportion of positives that are correctly identified. Sensitivity is also termed as Recall. This implies that there will be another proportion of actual positive cases, which would get predicted incorrectly as negative (and, thus, could also be termed

as the false negative). This can also be represented in the form of a false negative rate. The sum of sensitivity and false negative rate would be 1.

$$Sensitivity = \frac{TP}{(TP\ +\ FN)}$$

3.4.3.3. Specificity

Specificity is defined as the proportion of actual negatives, which got predicted as the negative (or true negative). This implies that there will be another proportion of actual negative, which got predicted as positive and could be termed as false positives. This proportion could also be called a false positive rate. The sum of specificity and false positive rate would always be 1.

$$Specificity = \frac{TN}{(TN\ +\ FP)}$$

3.4.3.4. Area Under Curve (AUC)

AUC was defined as the integrated area underneath the receiver operating characteristic curve (i.e., sensitivity versus 1-specificity). The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.

To evaluate if these models can well predict the activity of unknown compounds, we benchmarked their percentage of correct predictions (PCP) out of the 14 compounds in the external validation dataset that were not used in model training [79].

CHAPTER 4

RESULTS AND DISCUSSION

4.1. Folding Number K

We first studied the effect of k values (5 and 10) on FOMs in 72 QSAR models based on 4 algorithms by 6 fingerprints by 3 fingerprint lengths (see representative cases in Fig. 2 and Table 2).

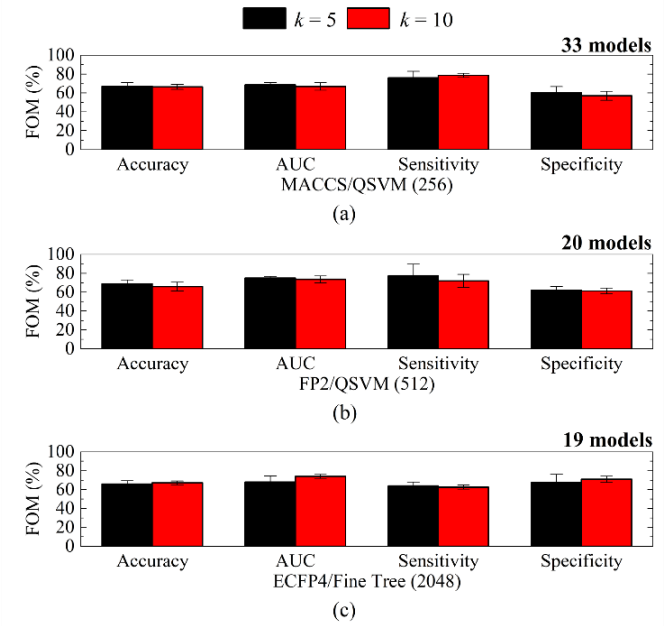

Figure 2. Effect of k values on FOMs in representative models. a) One model with no preferred k value. b) One model in which k = 5 is preferred. c) One model in which k = 10 is preferred. In a) – c), each model is noted as fingerprint/ algorithm.

If one model shows less than 5 % difference in all 4 FOMs between two k values, or if one model shows that k = 5 and k = 10 yields more than 5 % improvement in different FOMs, we will view this model as one that has no preferred k value. If one model shows

more than 5 % improvement in 1-4 FOMs at one k value (either 5 or 10), we will select this k value as the preferred k value for that model. According to these definitions, our data show that: 1) half of the models (33/72, in Fig. 2a) have no preferred k value; and 2) about one quarter of the models (20/72 in Fig. 2b, 19/72 in Fig. 2c) have a preferred k value (either 5 or 10). This result shows that overall k = 5 and k = 10 yield comparable performance among these 72 models. We therefore chose k = 5 for the following analysis.

Table 2. Effect of k values on FOMs in representative models.

	Accuracy (%)	AUC (%)	Sensitivity (%)	Specificity (%)
MACCS/QSVM (k = 5)	67.23 ± 4.10	68.67 ± 2.52	76.00 ± 6.93	60.21 ± 6.71
MACCS/QSVM (k = 10)	66.67 ± 2.69	67.00 ± 4.00	78.67 ± 2.31	56.99 ± 4.93
FP2/QSVM (k = 5)	69.03 ± 4.10	75.00 ± 1.73	77.33 ± 12.22	62.36 ± 3.72
FP2/QSVM (k = 10)	66.07 ± 4.70	73.67 ± 3.51	72.00 ± 6.93	61.29 ± 3.22
ECFP4/Fine (k = 5)	66.07 ± 3.55	68.00 ± 6.56	64.00 ± 4.00	67.74 ± 8.53
ECFP4/Fine (k = 10)	67.27 ± 2.02	74.00 ± 2.00	62.67 ± 2.31	70.97 ± 3.23

4.2. Fingerprints

Using k = 5, we next evaluated the effect of fingerprint lengths (3 lengths per fingerprint) on FOMs in 24 models based on 4 algorithms by 6 fingerprints (see representative cases in Fig. 3 and Table 3). Our data show that 16/24 models have at least one FOM where one length yields more than 5% improvement over the other two lengths.

If one model shows that different lengths yield more than 5%improvement in different FOMs, or if one model shows less than 5 % difference in all 4 FOMs among all 3 lengths, we will view this model as one that has no preferred length. If one model shows more than 5 % improvement in 1 to 3 FOMs at one length, we will select this length as the preferred length for that model (note: no model has one preferred length that yields more than 5 % improvement in 4 FOMs). According to these definitions, our data show that 50 % of the models (12/24, Fig. 3a) had no preferred length and 50 % of the models (12/24, Fig. 3b) had preferred length.

Table 3. Effect of fingerprint lengths on FOMs in representative models (k = 5).

	Accuracy (%)	AUC (%)	Sensitivity (%)	Specificity (%)
MACCS/RUSboosted (128)	70.23 ± 3.69	73.00 ± 2.00	78.67 ± 4.62	63.44 ± 4.93
MACCS/RUSboosted (256)	69.60 ± 0.00	70.33 ± 3.51	74.67± 2.31	65.59 ± 1.86
MACCS/RUSboosted (512)	64.87 ± 7.23	70.33 ± 7.37	73.33 ± 12.22	58.06 ± 3.22
FP2/RUSboosted (512)	69.63 ± 3.55	74.33 ± 3.05	62.67 ± 6.11	75.27 ± 1.86
FP2/RUSboosted (1024)	74.40 ± 2.08	76.33 ± 5.51	69.33 ± 6.11	78.49 ± 6.72
FP2/RUSboosted (2048)	67.23 ± 9.18	73.67 ± 11.85	54.67 ± 6.11	77.42± 12.90

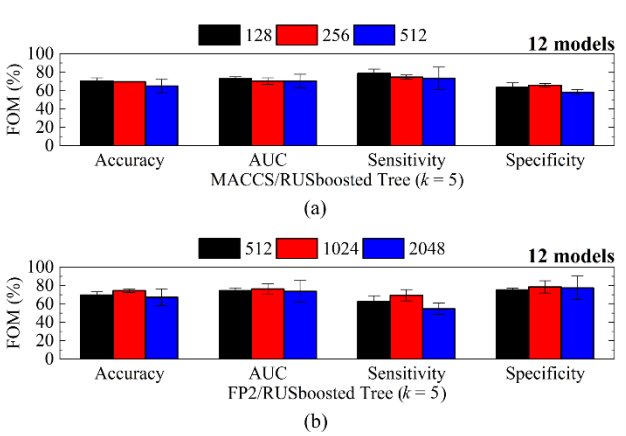

Figure 3. Effect of fingerprint lengths on FOMs in representative models. a) One model with no preferred length: b) One model with one preferred length: In a) and b), each model is noted as fingerprint/algorithm.

In FP2, FP3, FP4, and MACCS models (a total of 16 by 4 algorithms), we found that: 1) increasing the length from default values does not capture additional structural details of the compounds in their fingerprint representations (i.e., merely adding extra zeros to representation vectors). As a result, half of the models (9/16) do not have more than 5 % improvement in any FOM, whereas 2 of the 7 remaining models do not have their longest length as the preferred length; 2) decreasing the length from default values will make fingerprints lose their resolution and likely fail to capture structural details that are needed to differentiate highly similar compound structures (see Section 3.1) [51, 80, 81]. As a result, one quarter of the models (4/16) have more than 5 % degradation in 1 or 2 FOMs, whereas 50 % of the models (8/16) do not have their shortest length as the preferred length. In ECFP6 models (a total of 4 by 4 algorithms), we found that at the length of 2048 and/or 4096: 1) 1 model has more than 5 % improvement in 2 FOMs than those at the length of 1024; 2) 2 models have more than 5 % degradation in 1 or 2 FOMs than those at length 1024; and 3) one model shows less than 5 % difference in all 4 FOMs compared to those at the length of 1024. This result shows that the ECFP6 fingerprint does not always capture more structural details in our dataset at lengths longer than 1024 [57, 81]. Based on these

analyses, we chose the default lengths in FP2 (1024), FP3 (64), FP4 (512), and MACCS (256) models for the rest of our analysis because: 1) only half (9/16) of the models have a preferred length, 2) a longer length often adds no new structural information, and 3) a shorter length often results in a loss of structural details. For ECFP6, we chose to analyze all 3 lengths in the following (1024, 2048, and 4096 labeled as ECFP6A, ECFP6B, and ECFP6C, respectively) because there is no default length reported for this fingerprint [64]. For ECFP4, we again chose to analyze all 3 lengths (2048, 4096, and 8192 labeled as ECFP4A, ECFP4B, and ECFP4C, respectively).

4.3. Fingerprint Uniqueness

Due to the structure similarity of the compounds in our dataset, we also examined if these fingerprints at their chosen lengths can uniquely represent the compound structures. If not, there would be identical representation vectors representing both active and inactive compounds, which can result in misclassifications by the corresponding model [82, 83]. From this respective, our data show that FP2, ECFP4, and ECFP6 fingerprints each yield only 2 identical vectors across 56 compounds in the training dataset, suggesting that they can represent most compound structures in a unique vector [56, 64, 81]. In contrast, FP3, FP4, and MACCS fingerprints each yield over 20 identical vectors among the training dataset, suggesting that they are less unique in representing compound structures [81].

4.4. Model FOMs

Using k = 5 and the fingerprint lengths we chose, we next benchmarked the 4 FOM values in 40 models based on 4 algorithms by 10 fingerprints (ECFP4 and ECFP6 each with 3 lengths) (see Fig. 4 and Table 4), with the results described as follows:

4.4.1. Accuracy and AUC

Our accuracy and AUC data (Figs. 4a and 4b) show that: 1) all 40 models have more than 50 % accuracy with less than 10 % standard deviation; 2) except FP3/Bagged tree and FP4/Bagged tree models, all the other 38 models have more than 51 % AUC with less than 10 % standard deviation; 3) FP2/QSVM, MACCS/RUSboosted tree, ECFP6B/RUSboosted tree, and ECFP6C/RUSboosted tree (x/y: x: fingerprint, y: algorithm) models have more than 70 % accuracy and more than 75 % AUC, suggesting the promise of these 4 small-dataset models. Based on accuracy and AUC values, we found that FP2/QSVM, MACCS/RUSboosted tree, ECFP6B/RUSboosted tree, and ECFP6C/RUSboosted tree models performed the best, whereas FP3/Bagged tree and FP4/Bagged tree models performed the worst. The overall fair performance of the remaining 34 models (50-70%accuracy and AUC) can result from the small size of the training dataset and the challenge in classifying compounds with similar structures [84].

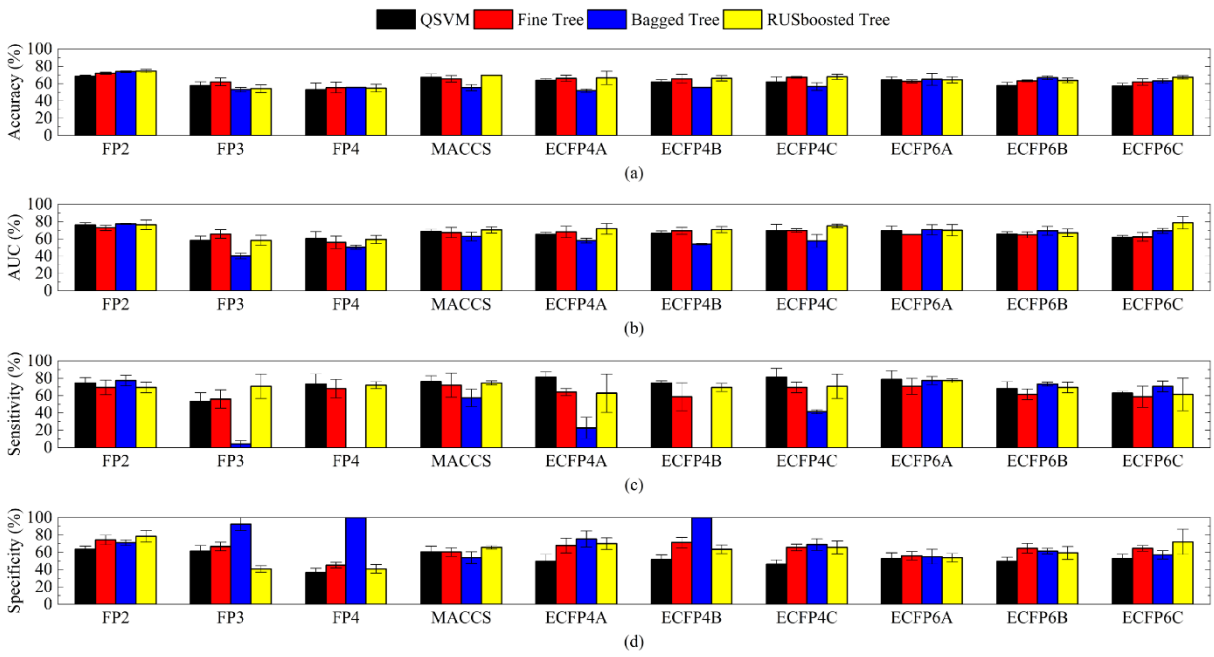

Figure 4. FOMs values across 40 models with k = 5 and the chosen lengths for each fingerprint.

4.4.2. Sensitivity and Specificity

Our sensitivity and specificity data (Figs. 4c and 4d) show that: 1) except for the FP3/Bagged tree, FP4/Bagged tree, ECFP4A/Bagged tree, ECFP4B/Bagged tree, and ECFP4C/Bagged tree models, the remaining 35 models have more than 50 % sensitivity; the majority of these models (21/35) show less than 10%standard deviation; 2) 15 models show > 10% standard deviation; 3) except for the FP4/QSVM model, the remaining 39 models have more than 40 % specificity; the majority of these models (39/40) show less than 10 % standard deviation; 4) 36 models have less than 40 % difference between their

Table 4. FOM values of best performing models for each fingerprint (k = 5).

	Accuracy (%)	AUC (%)	Sensitivity (%)	Specificity (%)
FP2 (1024)	RUSboosted 74.40 ± 2.08	Bagged 77.33 ± 0.58	Bagged 77.33 ± 6.11	RUSboosted 78.49 ± 6.71
FP3 (64)	Fine 62.03 ± 4.56	Fine 65.67 ± 4.93	RUSboosted 70.67 ± 14.05	Bagged 92.47 ± 7.45
FP4 (512)	Bagged 55.40 ± 0.00	QSVM 60.33 ± 8.14	QSVM 73.33 ± 11.55	Bagged 100.00 ± 0.00
MACCS (256)	RUSboosted 69.60 ± 0.00	RUSboosted 70.33 ± 3.51	QSVM 76.00 ± 6.93	RUSboosted 65.59 ± 1.86
ECFP4A (2048)	RUSboosted 66.53 ± 7.60	RUSboosted 71.66 ± 6.11	QSVM 81.33 ± 6.11	Bagged 75.26 ± 9.31
ECFP4B (4096)	RUSboosted 66.06 ± 3.05	RUSboosted 70.66 ± 3.51	QSVM 74.66 ± 2.30	Bagged 100.00 ± 0.00
ECFP4C (8192)	RUSboosted 67.86 ± 3.05	RUSboosted 75.00 ± 2.00	QSVM 81.33 ± 10.06	Bagged 68.81 ± 6.71
ECFP6A (1024)	Bagged 64.87 ± 6.78	Bagged 70.67 ± 5.86	QSVM 78.67 ± 10.07	Fine 55.91 ± 4.93
ECFP6B (2048)	Bagged 66.70 ± 2.08	Bagged 69.33 ± 5.13	Bagged 73.33 ± 2.31	Fine 64.52 ± 5.59
ECFP6C (4096)	RUSboosted 67.27 ± 2.02	RUSboosted 78.67 ± 7.23	Bagged 70.67 ± 6.11	RUSboosted 72.04 ± 14.55

sensitivity and specificity values; of the four exceptions, FP3/Bagged tree, FP4/Bagged, and ECFP4B/Bagged tree models showed low sensitivity (< 5 %) due to a large number of FNs, and high specificity (> 90 %) due to a small number of FPs.

The imbalance between sensitivity and specificity in FP3/Bagged tree, FP4/Bagged tree, and ECFP4B/Bagged tree models is likely due to a significant bias they develop to the majority class (inactive compounds) in our training dataset. This bias can result from the class imbalance in our training dataset (31 inactive versus 25 active) [84, 85], which can make these models form classification rules primarily on inactive compounds. This in turn would lead to 1) misclassifications of active compounds, thus increasing the number of FNs [84] and 2) overall a small number of true predictions, thus decreasing the number of FPs. Furthermore, such imbalance can be worsened by the way the bagged tree algorithm from sub-models based on randomly sampled subsets of the entire training dataset. Such sampling process may drop active compounds and result in subsets where inactive compounds are even more dominated (i.e., yielding a greater imbalance between inactive and active compounds) [71, 84, 86, 87].

Overall, our sensitivity and specificity data highlight the importance of benchmarking all 4 FOMs when evaluating the model performance. Accuracy and AUC alone may not fully capture the downside of the model performance, such as the imbalance between sensitivity and specificity trained from imbalanced training datasets.

4.4.3. Model Validation

To evaluate if the aforementioned 40 models can predict the activity of unknown compounds, we examined their PCP on 14 compounds (10 inactive versus 4 active) in the external validation dataset (see Fig. 5 and Table 5) [88]. Our data show that: 1) FP3/Fine

tree model performs the best with PCP = 92.86 %, whereas the FP2, FP4, MACCS, ECFP4, and ECFP6 models have their PCP up to 76.19 %; 2) PCP values across all 40 models have less than 15 % standard deviation.

These results suggest the promise of our small-dataset models in predicting Wnt inhibitors. For each of these models, we compared its PCP from the validation process (Fig. 5) with its accuracy value from the training process (Fig. 4a) to check if it is an overfitted model [79]. Our data show that: 1) PCP is more than 15 % lower than the accuracy in 1 FP2 model and 2 ECFP6A models; and 2) PCP is less than 15 % lower than the accuracy in all ECFP4C, ECFP6B, and ECFP6C models. For models listed in the first category, PCP is significantly lower than the accuracy, suggesting that these models likely overfitted compound structures (e.g., captured unnecessary structural details) in the training dataset [89].

Table 5. Models with the maximum PCP values in each fingerprint (k = 5).

Fingerprint	PCP (%)
FP2 (1024)	Fine; 71.43 ± 0.00
FP3 (64)	Fine; 92.86 ± 0.00
FP4 (512)	Bagged; 71.43 ± 0.00
MACCS (256)	Fine; 71.43 ± 0.00 Bagged; 71.43 ± 7.14
ECFP4A (2048)	Bagged; 73.8 ± 8.24
ECFP4B (4096)	Bagged; 71.42 ± 0.00
ECFP4C (8192)	Bagged; 71.42 ± 0.00
ECFP6A (1024)	Bagged; 69.04 ± 10.91
ECFP6B (2048)	Bagged; 71.43 ± 0.00
ECFP6C (4096)	Bagged; 76.19 ± 8.25

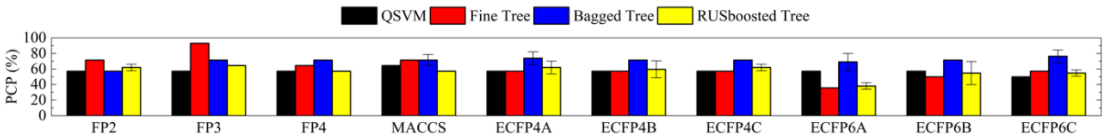

Figure 5. PCP values across 40 models with k = 5 and the chosen lengths for each fingerprint

Based on both PCP and 4 FOMs of these 40 models, we observe that ECFP4 and ECFP6 fingerprint at the longer lengths offers unique and sufficient representations of structural details with no overfitting. In contrast, FP3, FP4, and MACCS fingerprints also show no overfitting but fail to offer unique representations. FP2 fingerprint features high accuracy and AUC but also shows overfitting. These results suggest that fingerprints should be chosen to sufficiently, uniquely, but not overly represent structurally similar compounds in developing high performance small-dataset QSAR models.

4.5. Performance comparison

We finally remarked that the FOMs in our QSAR models are on par with other computational methods used for drug discovery. For instance, Mayr et al. have comprehensively studied *ca*. 500,000 drug compounds across more than 1000 assays from ChEMBL dataset. They built predictive models of the drug activity (in the respective assay) by machine learning algorithms namely support vector machine, random forest, k-nearest neighbor, naive bayes, and deep learning [90]. By averaging the AUC values of each model, they reported typical AUC values around 70 %.

As another example, Hofmarcher et al. have built predictive models from over 30000 compounds across 209 assays from Cell painting dataset by neural network algorithms [91]. In this study, they derived features from High-throughput fluorescence

microscopy imaging (HTI). By averaging the FOMs over all assays, they reported typical accuracy values around 77 %, AUC values around 70 %, sensitivity values around 50 %, and specificity values around 76 %. This results indicating that the cell morphology changes contain a large amount of information about compound activities in the field of drug discovery.

In comparison, our models typically obtained accuracy values around 65 %, AUC values around 70 %, sensitivity values around 70 %, and specificity values around 60 %. Nonetheless, we noted that computational methods on prediction of Wnt signaling inhibitors are still at their early stage of development at this moment. We expect that future efforts on this essential field of cell biology will allow more direct comparison with our QSAR models.

CHAPTER 5

CONCLUSION

Machine learning is currently one of the most important and rapidly evolving topics in computer-aided drug discovery. One of the primary application areas for machine learning in drug discovery is helping researchers understand and exploit relationships between chemical structures and their biological activities or QSAR. The general protocol for constructing QSAR models for drug discovery has been systematized and consists of several modular steps involving:

- First, Molecular Encoding, where the chemical features and properties are derived from chemical structures.
- Second, a feature selection step is performed where unsupervised learning techniques are used to identify the most relevant properties and reduce the dimensionality of the feature vector.
- Finally, in the learning phase, a supervised machine learning model is applied to discover the relationship between the input feature vectors and the biological responses.

Building an accurate QSAR model also requires careful consideration and selection of the QSAR datasets used for training and model validation This includes separation of training and test sets for initial model creation and the test sets for final model performance evaluation. The performances of the QSAR models are commonly evaluated by standard metrics such as sensitivity, specificity, precision and recall [6]. For unbalanced datasets,

area-under-curve (AUC) derived from receiver-operating-characteristics (ROC) curves can be used [6].

In this study, we present a systematic small-dataset QSAR study for prediction of effective Wnt signaling inhibitors that are essential to therapeutics development in prevalent human diseases. Specifically, we trained 72 QSAR models based on 4 algorithms, 6 fingerprints, and 3 fingerprint lengths using a training dataset (56 compounds), evaluated their performance on 4 FOMs, and examined their PCP using an external validation dataset (14 compounds). Our data show that the model performance is maximized when:

- Molecular fingerprints are selected to provide sufficient, unique, and not overly detailed representations of the compound structures (i.e. to avoid fingerprint lengths that lose fine structural features, identical representation vectors for multiple compounds, and overfitting);
- Algorithms are selected to reduce the number of false predictions due to class imbalance in the dataset; and
- Models are selected to reach balanced performance on all 4 FOMs.

These results may provide general guidelines in developing high-performance small dataset 2D QSAR models for drug activity prediction. Moving forward, it will be useful to test if these guidelines would apply to QSAR studies based on other Wnt signaling related assays. To achieve this, we will need to expand the experimental data in those assays, which are often associated with other targeted proteins (e.g., Wnt-3a, kinases) or host cells (e.g., MCF7, ST14A).

BIBLIOGRAPHY

[1] A. Schuhmacher, O. Gassmann, and M. Hinder, "Changing R&D models in research-based pharmaceutical companies," *Journal of translational medicine,* vol. 14, no. 1, pp. 1-11, 2016.

[2] S. M. Paul *et al.*, "How to improve R&D productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery,* vol. 9, no. 3, pp. 203-214, 2010.

[3] H. M. Patel *et al.*, "Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery," *Medicinal chemistry research,* vol. 23, no. 12, pp. 4991-5007, 2014.

[4] D. A. Winkler, "The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery," *Briefings in bioinformatics,* vol. 3, no. 1, pp. 73-86, 2002.

[5] A. Cherkasov *et al.*, "QSAR modeling: where have you been? Where are you going to?," *Journal of medicinal chemistry,* vol. 57, no. 12, pp. 4977-5010, 2014.

[6] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," *Drug discovery today,* vol. 23, no. 8, pp. 1538-1546, 2018.

[7] K. Roy, S. Kar, and R. N. Das, *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press, 2015.

[8] K.-Z. Myint, L. Wang, Q. Tong, and X.-Q. Xie, "Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions," *Molecular pharmaceutics,* vol. 9, no. 10, pp. 2912-2923, 2012.

[9] J. Simm *et al.*, "Repurposing high-throughput image assays enables biological activity prediction for drug discovery," *Cell chemical biology,* vol. 25, no. 5, pp. 611-618. e3, 2018.

[10] S.-Y. Yang, "Pharmacophore modeling and applications in drug discovery: challenges and recent advances," *Drug discovery today,* vol. 15, no. 11-12, pp. 444-450, 2010.

[11] A. Hillebrecht and G. Klebe, "Use of 3D QSAR models for database screening: a feasibility study," *Journal of chemical information and modeling,* vol. 48, no. 2, pp. 384-396, 2008.

[12] G. Mustata *et al.*, "Discovery of novel Myc− Max heterodimer disruptors with a three-dimensional pharmacophore model," *Journal of medicinal chemistry,* vol. 52, no. 5, pp. 1247-1250, 2009.

[13] Y. C. Martin, "3D QSAR: current state, scope, and limitations," *3D QSAR in drug design,* pp. 3-23, 1998.

[14] T. Madhavan, "A review of 3D-QSAR in drug design," *Journal of the Chosun Natural Science,* vol. 5, no. 1, pp. 1-5, 2012.

[15] G. Pegoraro and T. Misteli, "High-throughput imaging for the discovery of cellular mechanisms of disease," *Trends in Genetics,* vol. 33, no. 9, pp. 604-615, 2017.

[16] L. Li, Q. Zhou, T. C. Voss, K. L. Quick, and D. V. LaBarbera, "High-throughput imaging: Focusing in on drug discovery in 3D," *Methods,* vol. 96, pp. 97-102, 2016.

[17] J. Dong *et al.*, "ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation," *Journal of cheminformatics,* vol. 7, no. 1, pp. 1-10, 2015.

[18] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: progress in machine intelligence for rational drug discovery," *Drug discovery today,* vol. 22, no. 11, pp. 1680-1685, 2017.

[19] K. Zhao and H.-C. So, "Drug repositioning for schizophrenia and depression/anxiety disorders: A machine learning approach leveraging expression data," *IEEE journal of biomedical and health informatics,* vol. 23, no. 3, pp. 1304-1315, 2018.

[20] H. M. Ashtawy and N. R. Mahapatra, "A comparative assessment of predictive accuracies of conventional and machine learning scoring functions for protein-ligand binding affinity prediction," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 12, no. 2, pp. 335-347, 2014.

[21] C. Kumari, M. Abulaish, and N. Subbarao, "Exploring molecular descriptors and fingerprints to predict mTOR kinase inhibitors using machine learning techniques," *IEEE/ACM transactions on computational biology and bioinformatics,* 2020.

[22] O. O. Petinrin and F. Saeed, "Stacked ensemble for bioactive molecule prediction," *IEEE Access,* vol. 7, pp. 153952-153957, 2019.

[23] I. V. Tetko, A. I. Luik, and G. I. Poda, "Applications of neural networks in structure-activity relationships of a small number of molecules," *Journal of medicinal chemistry,* vol. 36, no. 7, pp. 811-814, 1993.

[24] Y. Hao *et al.*, "Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods," *Ecotoxicology and environmental safety,* vol. 186, p. 109822, 2019.

[25] R. A. Mook Jr, M. Chen, J. Lu, L. S. Barak, H. K. Lyerly, and W. Chen, "Small molecule modulators of Wnt/β-catenin signaling," *Bioorganic & medicinal chemistry letters,* vol. 23, no. 7, pp. 2187-2191, 2013.

[26] R. T. Moon, A. D. Kohn, G. V. De Ferrari, and A. Kaykas, "WNT and β-catenin signalling: diseases and therapies," *Nature Reviews Genetics,* vol. 5, no. 9, pp. 691-701, 2004.

[27] N. Barker and H. Clevers, "Mining the Wnt pathway for cancer therapeutics," *Nature reviews Drug discovery,* vol. 5, no. 12, pp. 997-1014, 2006.

[28] H. Clevers, "Wnt/β-catenin signaling in development and disease," *Cell,* vol. 127, no. 3, pp. 469-480, 2006.

[29] G. S. Coombs, T. M. Covey, and D. M. Virshup, "Wnt signaling in development, disease and translational medicine," *Current drug targets,* vol. 9, no. 7, pp. 513-531, 2008.

[30] Y. Li, P.-K. Li, M. J. Roberts, R. C. Arend, R. S. Samant, and D. J. Buchsbaum, "Multi-targeted therapy of cancer by niclosamide: A new application for an old drug," *Cancer letters,* vol. 349, no. 1, pp. 8-14, 2014.

[31] R. A. Mook Jr *et al.*, "Structure–activity studies of Wnt/β-catenin inhibition in the Niclosamide chemotype: Identification of derivatives with improved drug exposure," *Bioorganic & medicinal chemistry,* vol. 23, no. 17, pp. 5829-5838, 2015.

[32] R. A. Mook Jr *et al.*, "Benzimidazole inhibitors from the Niclosamide chemotype inhibit Wnt/β-catenin signaling with selectivity over effects on ATP homeostasis," *Bioorganic & medicinal chemistry,* vol. 25, no. 6, pp. 1804-1816, 2017.

[33] J. Wang *et al.*, "Identification of DK419, a potent inhibitor of Wnt/β-catenin signaling and colorectal cancer growth," *Bioorganic & medicinal chemistry,* vol. 26, no. 20, pp. 5435-5442, 2018.

[34] M. Grimaldi, A. Boulahtouf, C. Prévostel, A. Thierry, P. Balaguer, and P. Blache, "A Cell Model Suitable for a High-Throughput Screening of Inhibitors of the Wnt/β-Catenin Pathway," *Frontiers in pharmacology,* vol. 9, p. 1160, 2018.

[35] R. Darnag, E. M. Mazouz, A. Schmitzer, D. Villemin, A. Jarid, and D. Cherqaoui, "Support vector machines: development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives," *European journal of medicinal chemistry,* vol. 45, no. 4, pp. 1590-1597, 2010.

[36] R. Darnag *et al.*, "Quantitative structure-activity relationship studies of TIBO derivatives using support vector machines," *SAR and QSAR in Environmental Research,* vol. 21, no. 3-4, pp. 231-246, 2010.

[37] A. S. Mandal and K. Roy, "Predictive QSAR modeling of HIV reverse transcriptase inhibitor TIBO derivatives," *European journal of medicinal chemistry,* vol. 44, no. 4, pp. 1509-1524, 2009.

[38] B. Hemmateenejad, S. M. H. Tabaei, and F. Namvaran, "Computer-aided design of potential anti-HIV-1 non-nucleoside reverse transcriptase inhibitors by contraction of β-ring in TIBO derivatives," *Journal of Molecular Structure: THEOCHEM,* vol. 732, no. 1-3, pp. 39-45, 2005.

[39] K. Dzobo, N. E. Thomford, and D. A. Senthebane, "Targeting the versatile Wnt/β-catenin pathway in cancer biology and therapeutics: from concept to actionable strategy," *Omics: a journal of integrative biology,* vol. 23, no. 11, pp. 517-538, 2019.

[40] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure–activity relationships," *Journal of chemical information and modeling,* vol. 55, no. 2, pp. 263-274, 2015.

[41] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme gradient boosting as a method for quantitative structure–activity relationships," *Journal of chemical information and modeling,* vol. 56, no. 12, pp. 2353-2360, 2016.

[42] P. Chen, P. Shi, G. Du, Z. Zhang, and L. Liu, "Wnt/β-Catenin, Carbohydrate Metabolism, and PI3K-Akt Signaling Pathway-Related Genes as Potential Cancer Predictors," *Journal of healthcare engineering,* vol. 2019, 2019.

[43] A. Gaulton *et al.*, "The ChEMBL database in 2017," vol. 45, no. D1, pp. D945-D954, 2017.

[44] R. A. Mook Jr *et al.*, "Small molecule modulators of Wnt/β-catenin signaling," vol. 23, no. 7, pp. 2187-2191, 2013.

[45] R. A. Mook Jr *et al.*, "Structure–activity studies of Wnt/β-catenin inhibition in the Niclosamide chemotype: Identification of derivatives with improved drug exposure," vol. 23, no. 17, pp. 5829-5838, 2015.

[46] R. A. Mook Jr *et al.*, "Benzimidazole inhibitors from the Niclosamide chemotype inhibit Wnt/β-catenin signaling with selectivity over effects on ATP homeostasis," vol. 25, no. 6, pp. 1804-1816, 2017.

[47] J. Wang *et al.*, "Identification of DK419, a potent inhibitor of Wnt/β-catenin signaling and colorectal cancer growth," vol. 26, no. 20, pp. 5435-5442, 2018.

[48] W. J. Moore *et al.*, "Modulation of Wnt signaling through inhibition of secreted frizzled-related protein I (sFRP-1) with N-substituted piperidinyl diphenylsulfonyl sulfonamides," vol. 52, no. 1, pp. 105-116, 2009.

[49] Y. Y. Chen, X. Q. Wu, W. J. Tang, J. B. Shi, J. Li, and X. H. J. E. j. o. m. c. Liu, "Novel dihydropyrazole-chromen: design and modulates hTERT inhibition proliferation of MGC-803," vol. 110, pp. 65-75, 2016.

[50] X. Cheng *et al.*, "7, 7′-diazaindirubin—a small molecule inhibitor of casein kinase 2 in vitro and in cells," vol. 22, no. 1, pp. 247-255, 2014.

[51] A.-C. Schmöle *et al.*, "Novel indolylmaleimide acts as GSK-3β inhibitor in human neural progenitor cells," vol. 18, no. 18, pp. 6785-6795, 2010.

[52] L. Piemontese *et al.*, "New diphenylmethane derivatives as peroxisome proliferator-activated receptor alpha/gamma dual agonists endowed with anti-proliferative effects and mitochondrial activity," vol. 127, pp. 379-397, 2017.

[53] I. V. Tetko, A. I. Luik, and G. I. J. J. o. m. c. Poda, "Applications of neural networks in structure-activity relationships of a small number of molecules," vol. 36, no. 7, pp. 811-814, 1993.

[54] Y. Hao *et al.*, "Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods," vol. 186, p. 109822, 2019.

[55] J. Bergstra and Y. J. J. o. m. l. r. Bengio, "Random search for hyper-parameter optimization," vol. 13, no. 2, 2012.

[56] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. J. J. o. c. Hutchison, "Open Babel: An open chemical toolbox," vol. 3, no. 1, pp. 1-14, 2011.

[57] J. Duan, S. L. Dixon, J. F. Lowrie, W. J. J. o. M. G. Sherman, and Modelling, "Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods," vol. 29, no. 2, pp. 157-170, 2010.

[58] M. Sastry, J. F. Lowrie, S. L. Dixon, W. J. J. o. c. i. Sherman, and modeling, "Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments," vol. 50, no. 5, pp. 771-784, 2010.

[59] S. Kausar and A. O. J. M. Falcao, "Analysis and comparison of vector space and metric space representations in QSAR modeling," vol. 24, no. 9, p. 1698, 2019.

[60] T. Chen, T. Wu, N. Li, Y. Jiang, H. Yin, and M. J. D. P.-A. I. J. o. P. S. Wu, "Simulation-based comparison of Biopharmaceutics Classification System and drug structure," vol. 75, no. 4, pp. 124-130, 2020.

[61] T. Braun *et al.*, "Quantitative time-resolved measurement of membrane protein–ligand interactions using microcantilever array sensors," vol. 4, no. 3, pp. 179-185, 2009.

[62] J. L. Durant, B. A. Leland, D. R. Henry, J. G. J. J. o. c. i. Nourse, and c. sciences, "Reoptimization of MDL keys for use in drug discovery," vol. 42, no. 6, pp. 1273-1280, 2002.

[63] A. U. J. D. d. t. Khan, "Descriptors and their selection methods in QSAR analysis: paradigm for drug design," vol. 21, no. 8, pp. 1291-1302, 2016.

[64] D. Rogers, M. J. J. o. c. i. Hahn, and modeling, "Extended-connectivity fingerprints," vol. 50, no. 5, pp. 742-754, 2010.

[65] C. Cortes and V. J. M. l. Vapnik, "Support-vector networks," vol. 20, no. 3, pp. 273-297, 1995.

[66] O. Chapelle, P. Haffner, and V. N. J. I. t. o. N. N. Vapnik, "Support vector machines for histogram-based image classification," vol. 10, no. 5, pp. 1055-1064, 1999.

[67] H. Drucker, D. Wu, and V. N. J. I. T. o. N. n. Vapnik, "Support vector machines for spam categorization," vol. 10, no. 5, pp. 1048-1054, 1999.

[68] S. Pang, D. Kim, and S. Y. J. I. t. o. N. n. Bang, "Face membership authentication using SVM classification tree generated by membership-based LLE data partition," vol. 16, no. 2, pp. 436-446, 2005.

[69] I. J. J. o. G. O. Dagher, "Quadratic kernel-free non-linear support vector machine," vol. 41, no. 1, pp. 15-30, 2008.

[70] A. T. Azar, S. M. J. N. C. El-Metwally, and Applications, "Decision tree classifiers for automated medical diagnosis," vol. 23, no. 7, pp. 2387-2403, 2013.

[71] M. J. W. A. o. S. Pal, Engineering and Technology, "Ensemble learning with decision tree for remote sensing classification," vol. 36, pp. 258-260, 2007.

[72] L. J. M. l. Breiman, "Bagging predictors," vol. 24, no. 2, pp. 123-140, 1996.

[73] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," in *2008 19th international conference on pattern recognition*, 2008: IEEE, pp. 1-4.

[74] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, A. J. I. T. o. S. Napolitano, Man,, C.-P. A. Systems, and Humans, "RUSBoost: A hybrid approach to alleviating class imbalance," vol. 40, no. 1, pp. 185-197, 2009.

[75] T.-T. J. P. R. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," vol. 48, no. 9, pp. 2839-2846, 2015.

[76] J. D. Rodriguez, A. Perez, J. A. J. I. t. o. p. a. Lozano, and m. intelligence, "Sensitivity analysis of k-fold cross validation in prediction error estimation," vol. 32, no. 3, pp. 569-575, 2009.

[77] J. Cai, J. Luo, S. Wang, and S. J. N. Yang, "Feature selection in machine learning: A new perspective," vol. 300, pp. 70-79, 2018.

[78] N. M. O'Boyle and R. A. J. J. o. c. Sayle, "Comparing structural fingerprints using a literature-based similarity benchmark," vol. 8, no. 1, pp. 1-14, 2016.

[79] S. Bleeker *et al.*, "External validation is necessary in prediction research:: A clinical example," vol. 56, no. 9, pp. 826-832, 2003.

[80] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," 1998.

[81] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. J. M. Pujadas, "Molecular fingerprint similarity search in virtual screening," vol. 71, pp. 58-63, 2015.

[82] A. Cherkasov *et al.*, "QSAR modeling: where have you been? Where are you going to?," vol. 57, no. 12, pp. 4977-5010, 2014.

[83] A. Bender *et al.*, "How similar are similarity searching methods? A principal component analysis of molecular descriptor space," vol. 49, no. 1, pp. 108-119, 2009.

[84] P. Banerjee, F. O. Dehnbostel, and R. J. F. i. c. Preissner, "Prediction is a balancing Act: importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets," vol. 6, p. 362, 2018.

[85] O. Soufan, W. Ba-Alawi, A. Magana-Mora, M. Essack, and V. B. J. S. r. Bajic, "DPubChem: a web tool for QSAR modeling and high-throughput virtual screening," vol. 8, no. 1, pp. 1-10, 2018.

[86] Z. Afzal *et al.*, "Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records," vol. 13, no. 1, pp. 1-11, 2013.

[87] H. He, E. A. J. I. T. o. k. Garcia, and d. engineering, "Learning from imbalanced data," vol. 21, no. 9, pp. 1263-1284, 2009.

[88] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. J. D. d. t. Altman, "Machine learning in chemoinformatics and drug discovery," vol. 23, no. 8, pp. 1538-1546, 2018.

[89] D. M. J. J. o. c. i. Hawkins and c. sciences, "The problem of overfitting," vol. 44, no. 1, pp. 1-12, 2004.

[90] A. Mayr *et al.*, "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL," vol. 9, no. 24, pp. 5441-5451, 2018.

[91] M. Hofmarcher, E. Rumetshofer, D.-A. Clevert, S. Hochreiter, G. n. J. J. o. c. i. Klambauer, and modeling, "Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks," vol. 59, no. 3, pp. 1163-1171, 2019.