2020

Benchmarking Small-Dataset Structure-Activity-Relationship Models for Prediction of Wnt Signaling Inhibition

Mahtab Kokabi
University of Massachusetts Amherst

Matthew Donnelly
University of Massachusetts Amherst

Guangyu Xu
University of Massachusetts Amherst

Recommended Citation

Kokabi, Mahtab; Donnelly, Matthew; and Xu, Guangyu, "Benchmarking Small-Dataset Structure-Activity-Relationship Models for Prediction of Wnt Signaling Inhibition" (2020). *IEEE Access*. 1200.
https://doi.org/10.1109/ACCESS.2020.3046190

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 19, 2020, accepted December 16, 2020, date of publication December 21, 2020,
date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3046190

Benchmarking Small-Dataset Structure-Activity-Relationship Models for Prediction of Wnt Signaling Inhibition

MAHTAB KOKABI, MATTHEW DONNELLY, AND GUANGYU XU, (Member, IEEE)
Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA

Corresponding author: Guangyu Xu (guangyux@umass.edu)

This work was supported by the National Science Foundation under Grant ECCS-1835268.

ABSTRACT 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.

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

I. INTRODUCTION

Drug development often involves extensive investment and time effort on experimental screening of drug candidates. 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 [1]–[5]. To date, computational methods based on three-dimensional quantitative structure-activity relationship (3D QSAR) analysis, high-throughput imaging (HTI), and pharmacophore modeling [5], [6]–[10] have succeeded in predicting the effectiveness of drug compounds towards prevalent human diseases (e.g., cancer [10]). Nonetheless, these high-performance methods often require user intervention steps on molecular/ligand alignment [5], [8], [9] or high-resolution images that are not available for all drug compounds [7]. To this end, 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. 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 [1]–[3], [6].

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,

The associate editor coordinating the review of this manuscript and approving it for publication was Henry Hess.

functional groups, and connectivity pathways [3], [11]. 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) [6], [12]. The resulting QSAR models have succeeded in predicting effective drugs of psychological disorders [13], protein-ligand binding affinities [14], and mTOR kinase inhibitors [15].

Nonetheless, current 2D QSAR analysis often relies on training machine learning algorithms with a large-sized drug activity dataset (size $>$ 1000) [6], [16], which requires significant time and effort on both benchwork and statistical analysis. For this reason, developing new drugs can cost hundreds of millions of U.S. dollars [17] and can take over a decade to transition to a marketable state [18]. 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 [19], [20].

Wnt signaling pathways are essential in cell biology and the development of therapeutics for highly prevalent diseases such as cancer, Schizophrenia, and kidney damage [21]–[25]. 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 [21]. 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 [21], [26]–[29]. For instance, screening assays based on live cell imaging have been used to identify Wnt/β-catenin inhibitors [30]. 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 [21].

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.

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 2D QSAR models for drug activity prediction.

II. METHODS

A. DATASETS

To the best of our knowledge, there has been a total of 70 drug compounds available in literature with experimentally validated effectiveness for internalizing Frizzled receptor proteins, and thus inhibiting Wnt signaling in human U2OS cells [21], [27]–[29]. Specifically, all these 70 compounds have been classified as active [inactive] compounds if they were 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. As a result, 29 compounds were experimentally tested to be active and 41 were tested to be inactive, suggesting a mild class imbalance between active and inactive compounds. It is noted that 65 of these compounds (except 5 inactive compounds) are derivatives of niclosamide [26], suggesting high structural similarities among these 70 compounds.

In this work, we chose to build 2D QSAR models from the aforementioned 70 compounds for prediction of Wnt signaling inhibition. We chose this dataset because these 70 compounds are all experimentally validated with the same biological assay (i.e., the internalization of Frizzled receptor proteins in U2OS cells) and form a mild class imbalance. It is noted that assays targeted at the dynamics of other Wnt-signaling-inhibition related proteins are also available in the ChEMBL database [31]–[36]. 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 [36]). On the other hand, we found that the size of our dataset, 70, is on par with other small-dataset QSAR studies (e.g., 16 in [19], and 48 in [20]); we thus believe this dataset has a sufficient number of data to build good-performing QSAR models.

We then 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). We next randomly selected 80 % of these 70 compounds to form a training dataset (56 compounds; 31 inactive, 25 active) to develop 2D QSAR models and perform cross-validation to statistically analyze their performance; we used the remaining 20 % of these 70 compounds as an external validation dataset (14 compounds; 10 inactive, 4 active) to examine if these QSAR models can predict the activity of compounds that were not used in model training [37]. We noted that the class imbalance of the randomly selected training dataset (25/56 = 44 % active samples) is on par with the class imbalance in the overall 70-compound dataset (31/70 = 41 % active samples) [38].

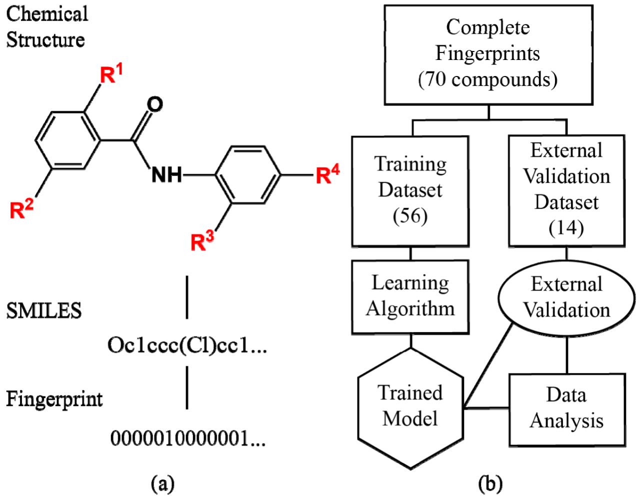


FIGURE 1. Schematic diagrams on benchmarking small-dataset QSAR models. R^1–R^4 represent the structural features in the compound.

B. FINGERPRINT REPRESENTATION

To train 2D QSAR models, 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) [39]–[41]. 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. 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, O=C(Nc1ccc([N+] (=O) [O-])cc1Cl)c1 cc(Cl)ccc1O; 2) converting this SMILES notation in the OpenBabel GUI to a hexadecimal vector 4a5124612940006 04091001f7aebecf6; and 3) converting the hexadecimal vector to a binary one, 01001010010100010 01001000110 000100101001010000000000000000011000000 10000001001 000100000000000011111011111010111010111 110110011110110. 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 2C).

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 [42], [43]. Specifically, FP2 (default length: 1024) is a path-based fingerprint which recognizes the rings and linear substructures in drug molecules [39]. FP3 (default length: 64) and FP4 (default length: 512) are substructure-based fingerprints to mark sub-structural patterns by SMILES arbitrary target specification (SMARTS) [39]. MACCS (default length: 256) is a substructure-key based fingerprint using 166 structural keys to characterize SMARTS patterns [39], [44]. ECFP4 and ECFP6 (no default lengths) are circular fingerprints stemming from the Morgan algorithm [39], [45] and are explicitly designed to capture molecular features related to molecular activity.

C. 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. Specifically, 1) QSVM is a binary classifier to define an optimal hyperplane that maximally separates two classes of high-dimensional data [46]; 2) fine tree algorithm (abbreviated as Fine) uses up to 100 decision rules (i.e. decision tree) for precise classification of the data [47]; 3) bagged tree algorithm (abbreviated as Bagged) first forms several subsets of data that are randomly sampled from the entire training dataset with replacement [48]. 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 [49]; 4) 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 [50], [51]. 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.

D. 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:

1) FOLDING NUMBER K

During the training of QSAR models, we applied the k-fold cross-validation procedure [52], which splits the training dataset into k sub-groups, iteratively selects one sub-group to validate the model trained by the remaining (k-1) sub-groups, and evaluates the model performance by the collective results. 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 [52], [53]. We then chose one preferred k value for the rest of our analysis based on the overall performance of these 72 models.

2) FINGERPRINT LENGTH

With the chosen k value, we next 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 [40], [54]. 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 [55].

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 = (TP + TN) / (TP + TN + FP + FN); sensitivity = TP / (TP + FN); and specificity = TN / (TN + FP); AUC was defined as the integrated area underneath the receiver operating characteristic curve (i.e., sensitivity versus 1-specificity).

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 [56].

III. RESULTS AND DISCUSSION

A. 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 1). 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.

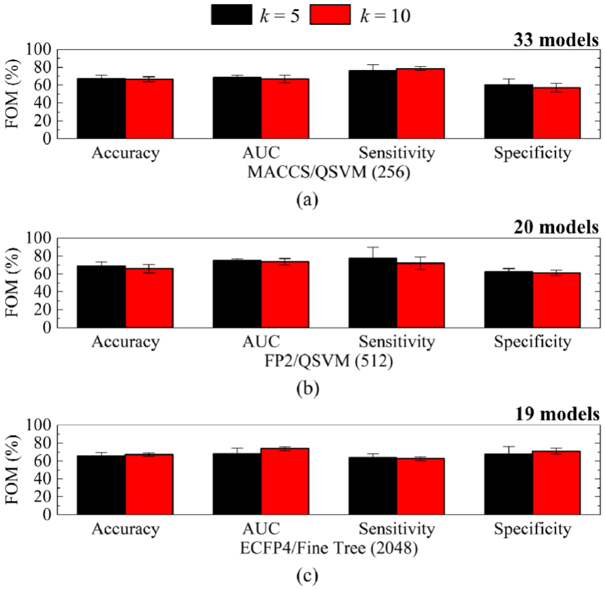


FIGURE 2. Effect of k values on FOMs in representative models. a) One model with no preferred k value. b) One model in which $k = 5$ is preferred. c) One model in which $k = 10$ is preferred. In a) – c), each model is noted as fingerprint/ algorithm (fingerprint length).

TABLE 1. 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

B. FINGERPRINTS

1) FINGERPRINT LENGTH

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 2). 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 a preferred length.

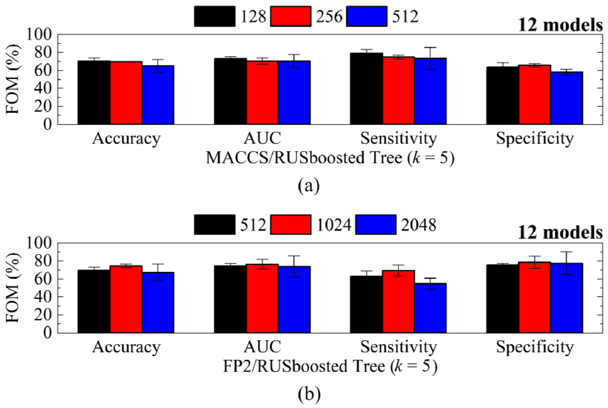


FIGURE 3. Effect of fingerprint lengths on FOMs in representative models. a) One model with no preferred length: b) One model with one preferred length: In a) and b), each model is noted as fingerprint/algorithm.

TABLE 2. 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

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 II.A) [35], [57], [58]. 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 [40], [58].

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 [46]. For ECFP4, we again chose to analyze

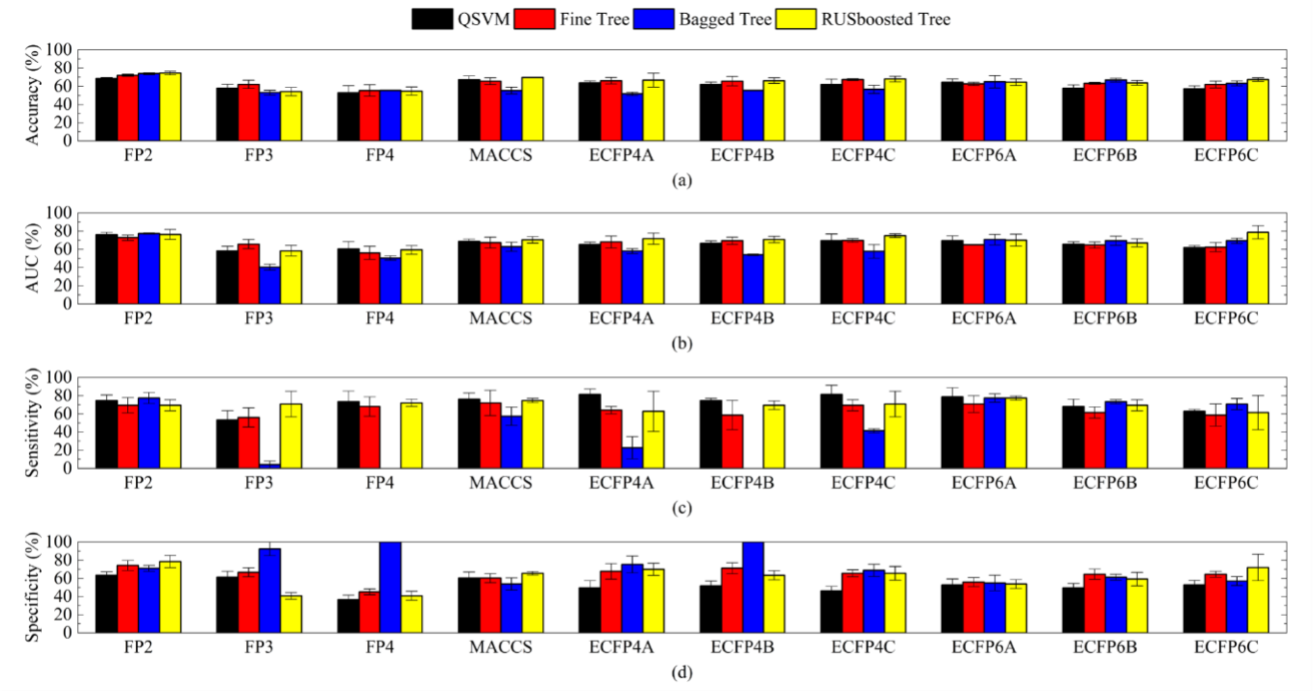


FIGURE 4. FOMs values across 40 models with $k = 5$ and the chosen lengths for each fingerprint.

all 3 lengths (2048, 4096, and 8192 labeled as ECFP4A, ECFP4B, and ECFP4C, respectively).

2) 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 [3], [59]. 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 [39], [45], [58]. In contrast, FP3, FP4, and MACCS fingerprints each yields over 20 identical vectors among the training dataset, suggesting that they are less unique in representing compound structures [58].

C. 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 3), with the results described as follows:

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 [60].

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 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.

TABLE 3. 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

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) [60], [61], 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 [60] 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) [48], [60], [62], [63].

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.

D. 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 4) [4]. 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.

TABLE 4. 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

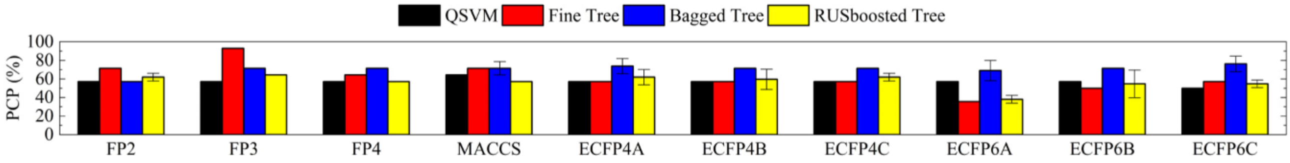


FIGURE 5. PCP values across 40 models with $k = 5$ and the chosen lengths for each fingerprint.

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 [56]. 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 [64].

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.

E. 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. They built predictive models of the drug activity (in the respective assay) by machine learning algorithms [65]. 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 by neural network algorithms [66]. 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 %. 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.

IV. CONCLUSION

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: 1) 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); 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 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).

ACKNOWLEDGMENT

(Mahtab Kokabi and Matthew Donnelly contributed equally to this work.)

REFERENCES

[1] H. M. Patel, M. N. Noolvi, P. Sharma, V. Jaiswal, S. Bansal, S. Lohan, S. S. Kumar, V. Abbot, and S. Dhiman, "Quantitative structure-activity relationship (QSAR) studies as strategic approach in drug discovery," *Medicinal Chem. Res.*, vol. 23, no. 12, pp. 4991–5007, 2014.

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

[3] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, and A. Tropsha, "QSAR modeling: Where have you been? Where are you going to," *J. medicinal Chem.*, vol. 57, no. 12, pp. 4977–5010, 2014.

[4] 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.

[5] K. Roy, S. Kar, and R. N. Das, "Other related techniques," in *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Amsterdam, The Netherlands: Elsevier, 2015, p. 357.

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

[7] J. Simm, G. Klambauer, A. Arany, M. Steijaert, J. K. Wegner, E. Gustin, and V. Chupakhin, "Repurposing high-throughput image assays enables biological activity prediction for drug discovery," *Cell Chem. Biol.*, vol. 25, no. 5, pp. 611–618, 2018.

[8] S.-Y. Yang, "Pharmacophore modeling and applications in drug discovery: Challenges and recent advances," *Drug Discovery Today*, vol. 15, nos. 11–12, pp. 444–450, 2010.

[9] A. Hillebrecht and G. Klebe, "Use of 3D QSAR models for database screening: A feasibility study," *J. Chem. Inf. Model.*, vol. 48, no. 2, pp. 384–396, 2008.

[10] G. Mustata, "Discovery of novel Myc? Max heterodimer disruptors with a three-dimensional pharmacophore model," *J. Medicinal Chem.* vol. 52, no. 5, pp. 1247–1250, 2009.

[11] J. Dong, D. S. Cao, H. Y. Miao, S. Liu, B. C. Deng, Y. H. Yun, N. N. Wang, A. P. Lu, W. B. Zeng, and A. F. Chen, "ChemDes: An integrated Web-based platform for molecular descriptor and fingerprint computation," *J. Cheminformatics*, vol. 7, no. 1, p. 60, 2015.

[12] 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.

[13] K. Zhao and H.-C. So, "Drug repositioning for schizophrenia and depression/anxiety disorders: A machine learning approach leveraging expression data," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 1304–1315, Jul. 2018.

[14] 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 Trans. Comput. Biol. Bioinf.*, vol. 12, no. 2, pp. 335–347, 2014.

[15] C. Kumari, M. Abulaish, and N. Subbarao, "Exploring molecular descriptors and fingerprints to predict mTOR kinase inhibitors using machine learning techniques," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 6, 2020, doi: 10.1109/TCBB.2020.2964203.

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

[17] O. J. Wouters, M. McKee, and J. Luyten, "Estimated research and development investment needed to bring a new medicine to market, 2009-2018," *Jama*, vol. 323, no. 9, pp. 844–853, 2020.

[18] V. A. Nyigo and H. Malebo, "Drug discovery and developments in developing countries: Bottlenecks and way forward," *Tanzania J. Health Res.*, vol. 7, no. 3, pp. 154–158, 2005.

[19] I. V. Tetko, A. I. Luik, and G. I. Poda, "Applications of neural networks in structure-activity relationships of a small number of molecules," *J. Medicinal Chem.*, vol. 36, no. 7, pp. 811–814, 1993.

[20] Y. Hao, "Prediction on the mutagenicity of nitroaromatic compounds using quantum chemistry descriptors based QSAR and machine learning derived classification methods," *Ecotoxicology Environ. Safety* vol. 186, Dec. 2019, Art. no. 109822.

[21] 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 Chem. Lett.*, vol. 23, no. 7, pp. 2187–2191, 2013.

[22] R. T. Moon, A. D. Kohn, G. V. De Ferrari, and A. Kaykas, "WNT and β-catenin signalling: Diseases and therapies," *Nature Rev. Genet.*, vol. 5, no. 9, pp. 691–701, 2004.

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

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

[25] 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.

[26] 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 Lett.*, vol. 349, no. 1, pp. 8–14, 2014.

[27] R. A. Mook, Jr., J. Wang, X. R. Ren, M. Chen, I. Spasojevic, L. S. Barak, H. K. Lyerly, and W. Chen, "Structure-activity studies of Wnt/β-catenin inhibition in the Niclosamide chemotype: Identification of derivatives with improved drug exposure," *Bioorganic Medicinal Chem.*, vol. 23, no. 17, pp. 5829–5838, 2015.

[28] R. A. Mook, Jr., X. R. Ren, J. Wang, H. Piao, L. S. Barak, H. K. Lyerly, and W. Chen, "Benzimidazole inhibitors from the Niclosamide chemotype inhibit Wnt/β-catenin signaling with selectivity over effects on ATP homeostasis," *Bioorganic Medicinal Chem.*, vol. 25, no. 6, pp. 1804–1816, 2017.

[29] J. Wang, R. A. Mook, Jr, X. R. Ren, Q. Zhang, G. Jing, M. Lu, I. Spasojevic, H. K. Lyerly, D. Hsu, and W, Chen, "Identification of DK419, a potent inhibitor of Wnt/β-catenin signaling and colorectal cancer growth," *Bioorganic Medicinal Chem.*, vol. 26, no. 20, pp. 5435–5442, 2018.

[30] 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 Pharmacol.*, vol. 9, p. 1160, Oct. 2018.

[31] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, and M. Davies, "The ChEMBL database in 2017," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D945–D954, 2017.

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

[33] Y.-Y. Chen, "Novel dihydropyrazole-chromen: Design and modulates hTERT inhibition proliferation of MGC-803," *Eur. J. Medicinal Chem.* vol. 110, pp. 65–75, Mar. 2016.

[34] X. Cheng, "7, 7'-diazaindirubin—A small molecule inhibitor of casein kinase 2 *in vitro* and in cells," *Bioorganic Medicinal Chem.* vol. 22, no. 1, pp. 247–255, 2014.

[35] A.-C. Schmöle, "Novel indolylmaleimide acts as GSK-3β inhibitor in human neural progenitor cells," *Bioorganic Medicinal Chem.* vol. 18, no. 18, pp. 6785–6795, 2010.

[36] L. Piemontese, "New diphenylmethane derivatives as peroxisome proliferator-activated receptor alpha/gamma dual agonists endowed with anti-proliferative effects and mitochondrial activity," *Eur. J. Medicinal Chem.* vol. 127, pp. 379–397, Feb. 2017.

[37] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[38] R. Kurczab, S. Smusz, and A. J. Bojarski, "The influence of negative training set size on machine learning-based virtual screening," *J. Cheminformatics* vol. 6, no. 1, p. 32, 2014.

[39] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *J. Cheminformatics*, vol. 3, no. 1, p. 33, 2011.

[40] J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman, "Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods," *J. Mol. Graph. Model.*, vol. 29, no. 2, pp. 157–170, 2010.

[41] M. Sastry, J. F. Lowrie, S. L. Dixon, and W. Sherman, "Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 771–784, 2010.

[42] T. Chen, T. Wu, N. Li, Y. Jiang, H. Yin, and M. Wu, "Simulation-based comparison of Biopharmaceutics Classification System and drug structure," *Die Pharmazie, Int. J. Pharmaceutical Sci.*, vol. 75, no. 4, pp. 124–130, 2020.

[43] T. Braun, M. K. Ghatkesar, N. Backmann, W. Grange, P. Boulanger, L. Letellier, H. P. Lang, A. Bietsch, C. Gerber, and M. Hegner, "Quantitative time-resolved measurement of membrane protein-ligand interactions using microcantilever array sensors," *Nature Nanotechnol.*, vol. 4, no. 3, pp. 179–185, 2009.

[44] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 6, pp. 1273–1280, 2002.

[45] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010.

[46] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[47] A. T. Azar and S. M. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 2387–2403, Dec. 2013.

[48] M. Pal, "Ensemble learning with decision tree for remote sensing classification," *World Acad. Sci. Eng. Technol.*, vol. 36, pp. 258–260, Jan. 2007.

[49] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[50] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, ‘‘RUSBoost: Improving classification performance when training data is skewed,’’ in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.

[51] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, ‘‘RUSBoost: A hybrid approach to alleviating class imbalance,’’ *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

[52] T.-T. Wong, ‘‘Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation,’’ *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, 2015.

[53] J. D. Rodriguez, A. Perez, and J. A. Lozano, ‘‘Sensitivity analysis of k-fold cross validation in prediction error estimation,’’ *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 569–575, Mar. 2010.

[54] J. Cai, J. Luo, S. Wang, and S. Yang, ‘‘Feature selection in machine learning: A new perspective,’’ *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018.

[55] N. M. O’Boyle and R. A. Sayle, ‘‘Comparing structural fingerprints using a literature-based similarity benchmark,’’ *J. Cheminformatics*, vol. 8, no. 1, pp. 1–14, 2016.

[56] S. Bleeker, H. A. Moll, E. W. Steyerberg, A. R. T. Donders, G. Derksen-Lubsen, D. E. Grobbee, and K. G. M. Moons, ‘‘External validation is necessary in prediction research: A clinical example,’’ *J. Clin. Epidemiology*, vol. 56, no. 9, pp. 826–832, 2003.

[57] M. A. Hall and L. A. Smith, ‘‘Practical feature subset selection for machine learning,’’ in *Proc. 21st Australas. Comput. Sci. Conf. (ACSC)*, Perth, WA, Australia, Feb. 1998, pp. 181–191.

[58] 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,’’ *Methods*, vol. 71, pp. 58–63, Jan. 2015.

[59] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies, ‘‘How similar are similarity searching methods? A principal component analysis of molecular descriptor space,’’ *J. Chem. Inf. Model.*, vol. 49, no. 1, pp. 108–119, 2009.

[60] P. Banerjee, F. O. Dehnbostel, and R. Preissner, ‘‘Prediction is a balancing act: Importance of sampling methods to balance sensitivity and specificity of predictive models based on imbalanced chemical data sets,’’ *Frontiers Chem.*, vol. 6, p. 362, Aug. 2018.

[61] O. Soufan, W. Ba-Alawi, A. Magana-Mora, M. Essack, and V. B. Bajic, ‘‘DPubChem: A Web tool for QSAR modeling and high-throughput virtual screening,’’ *Sci. Rep.*, vol. 8, p. 9110, Jun. 2018.

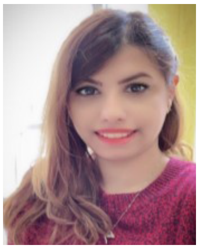
[62] Z. Afzal, ‘‘Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records,’’ *BMC Med. Informat. Decis. Making*, vol. 13, p. 30, Dec. 2013.

[63] H. He and E. A. Garcia, ‘‘Learning from imbalanced data,’’ *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[64] D. Hawkins, ‘‘The problem of overfitting,’’ *J. Chem. Inf. Comput. Sci.*, vol. 44, vol. 1, pp. 1–12, 2003.

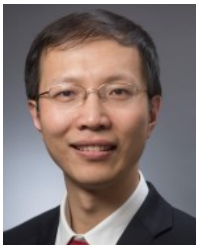
[65] A. Mayr, ‘‘Large-scale comparison of machine learning methods for drug target prediction on ChEMBL,’’ *Chem. Sci.* vol. 9, no. 24, pp. 5441–5451, 2018.

[66] M. Hofmarcher, E. Rumetshofer, D.-A. Clevert, S. Hochreiter, and G. Klambauer, ‘‘Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks,’’ *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1163–1171, 2019.



MAHTAB KOKABI received the B.S. degree in computer engineering from the Amirkabir University of Technology, Tehran, Iran, in 2017. She is currently pursuing the Master of Science degree in computer engineering with the University of Massachusetts, Amherst, MA, USA. Her current research interest includes developing machine learning models towards drug discovery and computational engineering science.

MATTHEW DONNELLY received the B.S. degree in electrical and computer engineering from the University of Massachusetts, Amherst, MA, USA in 2018, where he is currently pursuing the Ph.D. degree in electrical engineering. His current research interests include designing and fabricating bioelectronic devices towards drug delivery and health monitoring.



GUANGYU XU (Member, IEEE) received the B.S. and M.S. degrees in fundamental sciences and electrical engineering from Tsinghua University, Beijing, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from the University of California, Los Angeles, CA, USA, in 2011.

Since 2016, he has been an Assistant Professor of electrical and computer engineering with the University of Massachusetts, Amherst, MA, USA. His current research interests include building integrated neurointerfacing and biosensing tools, aiming to provide new capabilities for next-generation precision medicine.