American Journal of Biochemistry and Biotechnology

Original Research Paper

3D-QSAR and SVM Prediction of BRAF-V600E and HIV Integrase Inhibitors: A Comparative Study and Characterization of Performance with a New Expected Prediction Performance Metric

[1,2]Leonard Wesley, [3]Saihitha Veerapaneni, [3]Rachana Desai, [3]Francisco McGee, [3]Namrata Joglekar, [4]Sheela Rao and [5]Zeeshan Kamal

[1]MedLex Inc. www.medlex.com
[2]Department of Computer Science, College of Science, San Jose State University, One Washington Square, San Jose, CA 95192, USA
[3]Graduate Student, General Engineering, College of Engineering, San Jose State University One, Washington Square San Jose, CA 95192, USA
[4]Former Graduate Student, Biomedical, Department of Chemical and Materials Eng., College of Engineering, San Jose State Univ. One Washington Square San Jose, CA 95192, USA
[5]Nanosyn Inc., 3100 Central Expressway, Santa Clara, CA 95051, USA

Article history
Received: 19-09-2016
Revised: 09-10-2016
Accepted: 15-10-2016

Corresponding Author:
Leonard Wesley
Department of Computer Science, College of Science
San Jose State University
San Jose, CA 95192, USA
Email: Leonard.Wesley@sjsu.edu
URL: www.cs.sjsu.edu/~wesley

Abstract: The results of directly comparing the prediction accuracy of optimized 3D Quantitative Structure-Activity Relationship (3D-QSAR) models and linear Support Vector Machine (SVM) classifiers to identify small molecule inhibitors of the BRAF-V600E and HIV Integrase targets are reported. Performance comparisons were carried out using 303 compounds (68 active) against BRAF-V600E and 204 compounds (159 active) against HIV Integrase. A SVM prediction accuracy of 95% (BRAF-V600E) and 100% (HIV Integrase) and 3D-QSAR prediction accuracy of 76% (BRAFV600E) and 82% (HIV Integrase) was observed. To help explain the better performance of SVM in the comparison reported here and to help assess the degree to which a SVM or 3D-QSAR model is likely to perform best for other targetligands of interest a new EPP (Expected Predictive Performance) metric is introduced. How EPP can be used to help predict future performance of SVM and 3D-QSAR models by quantifying the degree of similarity between candidate compounds and training data is also demonstrated. Results show that the EPP metric is capable of predicting future prediction accuracy of SVM and 3D-QSAr models within 7% of actual performance.

Keywords: 3D-QSAR, SVM, BRAF, HIV Integrase, Machine Learning

Introduction

The drug discovery and development process is highly inefficient, risky and complex (DiMasi *et al*., 2015; Lamberti and Getz, 2015). Approximately 95% of candidate drug compounds fail to make it to market for a variety of complex reasons that range from imperfect science through business and economic related forces (Torfinn, 2014). Despite the use of High Throughput Screening (HTS) to help address efficiency concerns, drug failure rates and inefficiencies remain unacceptably high (Torfinn, 2014). Virtual screening attempts to improve HTS efficiency by using Machine Learning (ML) computational methods (Shoichet, 2004; Sengupta and Bandyopadhyay, 2012). The ML methods used in HTS carry out virtual screening by training supervised classifiers or regression-based methods to predict affinity and activity interactions between targets and candidate compounds. Then a subset of the candidate compounds that are predicted to be active and at times a smaller subset of compounds predicted to be not active, have bioassay tests conducted to confirm or refute predictions. This ML-based virtual screening process helps improve

efficiencies by significantly reducing the time and number of resource intensive bioassay tests.

Selecting an appropriate ML method remains a non-trivial and critical task because many complex aspects must be considered (Murphy, 2011). These range from the particular HTS questions to be answered, nature and amount of available data through the relevant physiochemical properties (i.e., descriptors) of candidate compounds, computational resources and the desired accuracy of classification results (Liu *et al*., 2014). Although there are many different types of ML-based classifiers and regression algorithms from which to select, typical practice is to analyze and compare the results from several distinct algorithms, or combine several methods into a single multiple classifier system (Wozniak *et al*., 2014). In either case, evaluating the credibility of several prediction results remain integral to identifying the most appropriate ML algoritm to use for the HTS task of interest.

Two among many well known ML algorithms that are used to help with drug discovery and development tasks include 3D Quantitative Structure-Activity Relationship (3D-QSAR) models and linear Support Vector Machine (SVM) models (Nantasenamat *et al*., 2010; Verma *et al*., 2010; Vapnik, 1999). Li *et al*. (2012), conducted a 500-compound comparative QSAR and SVM-regression study on estrogenic activities of persistent organic pollutants. Pourbasheer *et al*. (2015), developed QSAR and SVM models for predicting the activity of CK2 inhibitors. Darnag *et al*. (2010), used SVMs to build QSAR relationships between anti-HIV activity and four molecular descriptors of 82 TIBO derivatives. Yao *et al*. (2004), used SVMs to develop QSAR models that correlate molecular structures to their toxicity and bioactivities. Vasanthanathan *et al*. (2009), conducted comparative classification accuracy of binary quantitative structure activity relationship, SVM, random forest, kappa Nearest Neighbor (kNN) and decision tree methods to predict activity against cytochrome P450 targets.

With respect to the work reported here, the importance of the above and similar related work is that the results of the comparative studies suggest that SVM classifiers tend to perform better than many other supervised ML methods, including QSAR models, across a variety of target-ligand combinations. However, it is important to note that some of the comparative work was conducted using only statistical validation metrics such as R^2, Q^2, RMSE and so forth. Other comparisons were made in which one model used different training and test datasets than the other. In contrast, all of the comparative work reported by Vasanthanathan *et al*. (2009), involved using isoforms of the cytochrome P450 family of targets. Evaluation of model performance using identical targets, training and test sets for each model has very rarely been conducted or reported. The work reported here helps address this gap.

Another remaining technical gap is being able to determine if a 3D-QSAR or a SVM model is likely to perform better or worse on targets and ligands that are different from that investigated in previous work. Furthermore, if 3D-QSAR performs better than SVM or vice versa, being able to explain the observed difference in performance is equally important.

To begin addressing these technical gaps, the focus of the work reported here is three fold: (1) Report the results of directly comparing the accuracy of 3D-QSAR and SVM models to predict the activity of small molecule compounds against the BRAF-V600E and HIV Integrase targets; (2) develop a new quantitative and comparative measure called the Expected Predictive Performance (EPP) of 3D-QSAR and SVM models given a set of unclassified compounds and a trained classifier; and (3) present a method by which the *EPP* metric can be used to:

- Help explain why one model performed better than the other
- Quantify the degree to which 3D-QSAR or SVM is likely to perform better on targets and ligands that are of current interest and yet to be classified
- Quantify the number of samples, descriptors and required similarity between unknowns and a training set to improve predictive performance

The reasons for using the BRAF-V600E and HIV Integrase protein for this investigation are that significant previous work has identified many inhibitors of these targets, however, there remains much interest in identifying additional inhibitory small molecule compounds (e.g., Wainber *et al*., 2012; Prahallad *et al*., 2012).

The *EPP* measure is intended to be a common metric that can be used to compare current and expected prediction performance across many different classification and regression models. A requirement of the *EPP* measure is that Vapnik's theoretical notion of a model's *capacity* can be quantified and used to discriminate between class instances (Vapnik, 1999). *EPP* combines a model's capacity with a measure of the similarity between the unknown to be classified and the examples used to train the model. The idea is that the observed and expected prediction accuracy of a model is based, in part, on and proportional to its theoretical capacity to classify and the degree to which an unknown to be classified is similar to the examples used to train the classification model.

For the small molecule compounds used here, similarity is based on the physiochemical properties of a model's training data and the same physiochemical

properties of an unknown candidate drug compound. The greater the similarity between an unknown and training examples, the greater the observed and expected prediction accuracy.

Materials and Methods

The compound data sets and details needed to reproduce the results reported here can be found in a supplement document at (Wesley, 2016).

The comparison of 3D-QSAR and SVM classifiers to predict the inhibitory activity of small molecule compounds was carried out on two different targets: BRAF-V600E and HIV Integrase. 3D-QSAR modeling and prediction was conducted in a Windows 7 environment using the Molecular Operating Environment (MOE, 2015) version 2014.0901 software product from the Chemical Computing Group. The 3D-QSAR classification involved thresholding the regression predicted IC50 value at 1.4 μM to discern active (≤1.4 μM) from non-active (>1.4 μM) compounds. SVM modeling and prediction was conducted in a Windows 10 environment using scikit's sklearn SVM classifier version 0.17.1 modules that were imported into a 64-bit Enthought Canopy environment version 1.6.2.3262 running Python 2.7.3.

Two different approaches were used to perform each comparison: "Best Possible Model" and "Constrained to MOE Descriptors." In the "Best Possible Model" approach, 3D-QSAR and SVM models were optimized using a number and type of descriptors that produced the best possible prediction accuracy. Each model need not use the same type and number of descriptors as the other. In the "Constrained To MOE Descriptors" approach, the optimal SVM classifier was constrained to use all or a subset of the descriptors available and used by MOE. In this approach, there was no need to re-build 3D-QSAR models because they were optimized in the "Best Possible Model" approach.

In the "Best Possible Model" approach, 303 small molecule compounds were selected and used to build and test 3D-QSAR and SVM classifiers for the BRAF-V600E target. Of the 303 compounds, 243 were used as a training set and 60 compounds were used as a test set. Of the 243 compounds in the training set, 48 are active, per PubChem data base bio assay results and 195 are not active (PubChem). Of the 60 test set compounds, 20 are active and 40 not active. The 20 active compounds were specifically chose to be analogs of Vemurafenib® to facilitate a more direct comparison with 3D-QSAR predictions.

For the HIV Integrase target, 204 small molecule compounds were selected and used to build and test 3D-QSAR and SVM classifiers. Of the 204 compounds, 163 were used as a training set and 41 compounds were used as a test set. Of the 163 training set compounds, 130 are active and 33 are not active. Of the 41 test set compounds, 29 compounds are active and 12 are inactive.

Comparisons were carried out with no restrictions on the descriptors used to build each model. The approach involved using the best model building method and practices to develop the best possible model before making predictions. In this approach, the PaDEL-descriptor software (Yap, 2011) was used to generate descriptors for SVM classifiers and MOE descriptors were used for the 3D-QSAR models.

In the "Constrained to MOE Descriptors" approach, the only difference from the "Best Possible Model" approach is that the descriptors used by SVM were constrained to the same 92 descriptors used by MOE. Table 1 provides a summary matrix of the experimental approaches, seven (7) prediction accuracy experiments and number of descriptors used to produce the results reported here.

3D-QSAR Models

The practice of building QSAR models for drug discovery and development involves, in part, first identifying a chemical compound that is known to be active via bioassay results and IC50 values (Madhavan, 2012). For the BRAF-V600E target, Vemurafenib® (also known as PLX4032) was selected as a canonical active target inhibitor compound (Bollag *et al*., 2012). Then a data set consisting of Vemurafenib® analogs and related compounds along with respective IC50 values was created. The PLS (Partial Least Squares) method was then used to build a regression model to predict an IC50 value that indicates an active compound. The predicted IC50 value was then used to filter candidate compounds to identify the most appropriate ones to use and build pharmacophore models used by the MOE 3D-QSAR classifier. For the BRAF-V600E target, two 3D-QSAR models were built, one manually and a second using the MOE auto-build QSAR option. The manual-built method resulted in using five optimal descriptors and the auto-QSAR method used nine optimal descriptors shown in Table 2. The manual approach considered the allosteric method of inhibition by Vemurafenib® (i.e., high-level mechanism-based domain knowledge used for descriptor selection) that was not available using MOE's auto-QSAR method (PDB-5HES, 2016). The 3D-QSAR model validation results and BRAF-V600E inhibitor prediction accuracies are reported in Table 3. Only auto- QSAR was used to build a model to predict HIV Integrase inhibitors. Table 4 summarizes the 3D-QSAR model validation and prediction results for HIV Integrase inhibitors.

Table 1. Summary of the seven (7) prediction accuracy experiments conducted, number of descriptors used in each experiment and the experimental approach. SVD (Singular Value Decomposition) was used by to reduce the dimension to the indicated number of descriptors

Experiment	Target →	BRAFV600E	HIV Integrase
Approach ↓	Model ↓	(303 Compounds: 243 train, 60 test)	(204 Compounds: 163 train, 41 test)
Best Possible	3D-QSAR	Manual: 5 descriptors	Auto QSAR: 5 descriptors
Model		Auto QSAR: 9 descriptors	(SVD) 18 descriptors
		SVM (SVD) 16 descriptors	
Constrained to MOE descriptors	SVM	(SVD) 7 descriptors	(SVD) 71 descriptors

Table 2. (a) Manually identified optimal descriptors, (b) auto-QSAR identified optimal descriptors for predicting BRAF-V600E inhibitors

Descriptor	Description of the descriptor
(a)	
a_acc	Number of hydrogen bond acceptor atoms
a_don	Number of hydrogen bond donor atoms.
a_aro	Number of aromatic atoms
b_ar	Number of aromatic bonds
slogp	Log octanol/water partition coefficient.
(b)	
PEOE_VSA +3	Sum of v_i such that q_i is in the range (0.15, 0.20)
PEOE_VSA +5	Sum of v_i such that q_i is in the range (0.25, 0.30)
PEOE_VSA -4	Sum of v_i such that q_i is in the range (-0.25, -0.20)
SlogP_VSA 2	Sum of v_i such that L_i is in (-0.2, 0)
SlogP_VSA 5	Sum of v_i such that L_i is in (0.15, 0.20)
SlogP_VSA 7	Sum of v_i such that L_i is in (0.25, 0.30)
SMR_VSA 3	Sum of v_i such that R_i is in (0.35, 0.39)
SMR_VSA 4	Sum of v_i such that R_i is in (0.39, 0.44)
SMR_VSA 6	Sum of v_i such that R_i is in (0.485, 0.56)

Table 3. 3D-QSAR validation metrics and prediction accuracy between auto-QSAR and manual-QSAR models for BRAF-V600E inhibitors. Accuracy = (# Correct Prediction/Total # Predictions)×100%

Validation tests		Auto QSAR	Manual QSAR
CORRELATION COEFFICIENT ($R \geq 0.8$ is good)		$R = 0.8$	$R = 0.6$
COEFFICIENT OF DETERMINATION ($R^2 \geq 0.6$ is good)		$R^2 = 0.71$	$R^2 = 0.51$
CROSS-VALIDATED R^2 ($Q^2 > 0.5$)		$R^2 = 0.68$	$R^2 = 0.46$
$R^2_{ADJ} \geq 0.7$		$R^2_{ADJ} = 0.79$	$R^2_{ADJ} = 0.71$
R^2-$Q^2 < 0.3$		0.03	0.03
R^2_{ADJ}-$R^{2<0.3}$		0.08	0.07
$R^2_{PRED} > 0.6$		0.82	0.65
Y-RANDOMIZATION (Low R^2 is better)		0.002	0.004
BOOTSTRAPPING	R^2	0.73	0.57
	Q^2	0.70	0.54
	R^2_{PRED}	0.72	0.55
RMSE		0.89	0.70
ACCURACY		76%	75%

Table 4. Validation and prediction accuracy of the MOE's Auto-QSAR built 3D-QSAR model for HIV Integrase inhibitors. Accuracy = (# Correct Prediction/Total # Predictions)×100%

Validation tests	Auto-QSAR on HIV Integrase inhibitors
Correlation Coefficient (R)	0.89
Coefficient of determination (R^2)	0.72
Cross-validated R^2 (Q^2)	0.67
R^2-Q^2	0.05
R^2_{adj}& Q^2	0.82 and 0.80
R^2_{adj}- R^2	0.2
R^2_{Pred}	0.81
Y-randomization	0.04
Bootstrapping	$R^2 = 0.73$, $Q^2 = 0.70$ &$R^2_{pred} = 0.74$
RMSE	0.68
Accuracy	82.03%

Table 5. Summary of prediction accuracy of 3D-QSAR and SVM models for BRAF-V600E and HIV Integrase targets

	BRAFV600E inhibitors		HIV Integrase inhibitors	
	SVM	QSAR	SVM	QSAR
Number of compounds in test set	60	60	41	41
Number of descriptors	16	9	18	5
Prediction accuracy	95%	76%	100%	82.03%

SVM Models

The same compounds used for building 3D-QSAR prediction models for the BRAFV600E and HIV Integrase targets were used for building SVM models for both the "Best Possible Model" and "Constrained to MOE Descriptors" approaches. 2D .sdf (structure definition files) for the candidate compounds were obtained from the PubChem database and then converted to 3D .sdf files using OpenBabel version 2.3.1 (O'Boyle *et al*., 2011). Descriptors were generated from the 3D .sdf files using PaDEL-descriptor software (Yap, 2011). For the "Best Possible Model" approach, PaDEL-descriptor generated 2,070 2D and 3D descriptors for 303 BRAF-V600E target compounds and 204 HIV Integrase compounds. Singular Value Decomposition (SVD) was used to reduce the number of descriptors to the numbers shown in Table 1 (Wall *et al*., 2003). The SVM training and testing sets consist of the minimum number of descriptors that can achieve the highest prediction accuracy. This was achieved by a gridding process and decreasing the number of descriptors, prioritized by SVD, for each gridding iteration. Table 5 summarizes prediction accuracy and comparison of the 3D-QSAR and SVM models for the BRAF-V600E and HIV Integrase targets.

Expected Prediction Performance

The *EPP* measure can help to answer questions such as: (1) "Why has an optimized SVM model performed better than an optimized 3D-QSAR model?"; (2) "What is the likelihood that optimized SVM and 3D-QSAR models will accurately predict potential inhibitors of interest in the future?"; and (3) "How many training samples and what properties must unknown ligands/compounds have in order to achieve optimal performance?"

Gunawardana and Shani (2009), conducted work to identify evaluation metrics that can be used to assess the appropriateness of a machine learning-based recommender system. Han *et al*. (2008), developed a means to evaluate decision tree models that can be used as a virtual screening technique as well as a complement to traditional approaches for hits selection. Reich and Barai (1999), proposed a systematic evaluation procedure for machine learning. An intent of this and related reported work is to improve research and practice in the use of machine learning algorithms in engineering applications. Zadrozny (2004), formalized the sample selection bias problem in machine learning algorithms and presented a bias correction method that is particularly useful for classifier evaluation under sample selection bias.

To date, previous work to develop model prediction performance metrics have not captured the measure of observed or predicted performance that is characterized by the *EPP* measure. Here, our *EPP* measure of 3D-QSAR and SVM models is based not only on the available training and test data, but also on the characteristics (e.g., physiochemical descriptors) of the unknown compounds that the models are intended to classify.

Many classifiers have, as part of their calculus, the notion of capacity that is related to the complexity, flexibility and power of a set of classifier functions F intended to correctly classify example data (Vapnik, 1999). The notion of capacity can be quantified by a measure called the *VC dimension*, named after Vapnik and Chervonenkis (Vapnik, 1999).

VC dimension for indicator functions is defined to be the largest number h of points (i.e., largest number of vectors v_1, v_2, …, v_h) in all 2^h combinations of points/vectors that can be *shattered* (i.e., separated into two classes, e.g., class 0 or class 1) by all members $f(x,\alpha)\in F$, where x is a data point/vector and α, ($\alpha\in\Lambda\equiv$ the set of admissible parameters), is a parameter that specifies the function (e.g., if f happens to be a linear function, then α would represent the slope and y-intercept parameters of f).

Where l is the size of the training set (i.e., number of training samples), $\nu(\alpha)$ is the frequency of training errors on the training set (e.g., 1-cross validation score from a SVM classifier) and Λ is the set of admissible parameters that specify a f, then with probability 1-η (i.e., η is the likelihood that a $f_i \in F$ will misclassify a single example (v_i,y)) the upper bound of $p(\alpha)$ (i.e., defined as the probability of error on the test set) is a function of just l, h, $\nu(\alpha)$, η).

VC dimension of real valued functions (e.g., regression algorithms), let $A \leq Q(v, \alpha) \leq B$ and $\alpha \in F$, where F is a set of functions bounded by constants A and B and where A can be $-\infty$ and B can be ∞. Let β be an indicator of the level for the function $Q(z,\alpha)$ that shows for which $\boldsymbol{v}$ the function $Q(v, \alpha)$ exceeds β and for which it does not (e.g., β might be 1.4 μM which was the IC50 threshold level used by the 3D-QSAR models described

here to distinguish active from non-active compounds). The function Q(v, α) can be described by the set of all its indicators.

Let us consider along with the set of real functions Q(v, α), α ∈ F, the set of indicators:

$$I(v,\alpha,\beta)=\theta\{Q(v,\alpha)-\beta\},\alpha\in F,\beta\in(A,B) \tag{1}$$

where, $\theta(v)$ is the step function:

$$\theta(v)=(0 if\ v<0,1 if\ v\geq 0) \tag{2}$$

In statistical learning theory, a simplified probabilistic estimate of an upper bound on the prediction error rate on a test data set can be defined under two regimes (Vapnik, 1999). One regime is when the error rate on the test data is large, i.e., ≥50%. In this regime, the upper bound can be defined as:

$$\Pr ob(test\ error)\leq training\ error + \sqrt{\frac{VC_{Dim}\cdot\left(\log\left(\frac{2N}{VC_{Dim}}\right)+1\right)-\log\left(\frac{\eta}{4}\right)}{N}} \tag{3}$$

where, *training error* is ν(α) for SVM and 1-R^2 for regression methods (Vapnik, 1999), VC_{Dim} is the VC-Dimension, N is the number of training samples and η is chosen based on the training set. For SVMs, an approximation of η is taken to be the normalized inverse of the perpendicular (⊥) distance from any support vector and a f_i that corresponds to the boundary yielding the max margin distance. For regression methods, an approximation of η is taken to be 1-max($\Delta R^2|X\text{-}x_i|$ for $1\leq i\leq |X|$). For SVMs, VC_{Dim} can be approximated as N +1 where N is the number of training samples. For real valued functions, VC_{Dim} can be estimated as p +1 where p is the dimension of the data (Akaike, 1974). If computational resources are available, VC_{Dim} can be estimated with greater accuracy if F is a member of the class of linear discriminate functions (Vapnik, 1999).

For the second regime, where the error rate on the test data is <50%, the upper bound can be defined as:

$$\Pr ob(test\ error)\leq training\ error + \frac{VC_{Dim}\cdot\left(\log\left(\frac{2N}{VC_{Dim}}\right)+1\right)-\log\left(\frac{\eta}{4}\right)}{N} \tag{4}$$

In the interest of brevity and without loss of generality, the discussion going forward will continue just with respect to Equation 3.

The above probabilistic estimate of the upper bound on the prediction error rate is based on the assumption that the test data set is selected, i.i.d., from the same distribution as the training data. However, the assumption that an unknown example, for which a prediction is made, is a sample from the same distribution as the training data will hold to varying degrees.

If a fully specified distribution of the training data is available, where parameters of the distribution do not need to be estimated, then the Anderson-Darling test is potentially applicable to assess the degree to which the unknown example is a member of the distribution (Anderson and Darling, 1954). When this requirement is not satisfied, an alternative is proposed here where a similarity measure between the unknown example to be predicted and the training data is computed and integrated into the calculation of the probabilistic estimate of an upper bound on the prediction error rate.

Where the average vector for class 0 data is defined as:

$$\overline{V_0}=\frac{1}{n}\left(v_0^1+v_0^2+\ldots+v_0^n\right) \tag{5}$$

and the average vector for class 1 data is defined as:

$$\overline{V_1}=\frac{1}{m}\left(v_1^1+v_1^2+\ldots+v_1^m\right) \tag{6}$$

and the max Cosine similarity between an unknown and yet to be classified vector u and $\overline{V_0}$ or between u and $\overline{V_1}$ is defined as:

$$MaxCosSim\left(u,\overline{V_0},\overline{V_1}\right) = \max\left(\frac{u\cdot\overline{V_0}}{\|u\|\|\overline{V_0}\|},\frac{u\cdot\overline{V_1}}{\|u\|\|\overline{V_1}\|}\right) \tag{7}$$

Then a proposed similarity measure that is maximal if and only if two vectors are identical and less than maximal otherwise can be defined as:

$$SimEC\left(u,V_0,V_1\right) = \frac{\left(\left(1-\min\left(\|u-\overline{V_0}\|,\|u-\overline{V_1}\|\right)\right)\right)+\left(MaxCosSim\left(u,\overline{V_0},\overline{V_1}\right)\right)}{2} \tag{8}$$

A measure that can accomplish this is a combination of the Euclidian distance and Cosine similarity measure *SimEC*.

Assuming the similarity measure *SimEC* is >0, The normalized *EPP* can now be defined as:

$$EPP = 1 - \left(training\ error + \left(\frac{\sqrt{\frac{VC_{Dim} \cdot \left(\log\left(\frac{2N}{VC_{Dim}} \right) + 1 \right) - \log\left(\frac{\eta}{4} \right)}{N}}}{\sqrt{\frac{VC_{Dim} \cdot \left(\log\left(\frac{2N}{VC_{Dim}} \right) + 1 \right) - \log\left(\frac{\eta}{4} \right)}{N \cdot SimEC}}} \right) \right) \tag{9}$$

EPP is intended to characterize a minimum expected prediction performance based on the training error, capacity of the model and the degree of similarity between an unknown to be classified and the model's training data. The idea is that if an unknown is very similar to the training data and the test data is assumed to be i.i.d. of the training data, then the upper bound on Prob(*test error*) is minimal and the expected prediction accuracy of the unknown is highest. As similarity between an unknown and training set decreases the upper bound on Prob(*test error*) increases and the expected prediction accuracy decreases. Intuitively, the higher the *EPP* value the higher the expected correct classification for a given unknown and its corresponding similarity measure.

The *EPP* measure can be used to characterize the expected predictive performance of a model given the model's training data and an unknown to be classified. It can also be used to characterize the relative expected predictive performance between an optimized SVM classifier and an optimized 3D-QSAR model by comparing their respective *EPP* measures.

The question, "Which trained and optimized SVM or trained and optimized 3D-QSAR classifier is more likely to correctly classify a given unknown example?" can be answered, in part, by comparing their respective *EPP* measures. The classifier with the higher *EPP* measure is more likely to correctly classify the unknown. The answer to the question, "Why is one classifier (SVM or 3D-QSAR) better than the other for a given unknown?" can be answered, in part, by examining several aspects such as the classifier's VC_{Dim} (i.e., capacity), validation error rates of the respective classifiers and similarity of the unknown example to the training set.

Figure 1 shows a comparison between the *EPP* measure as defined in Equation 9, the expected prediction accuracy and the actual prediction accuracy for the 3D-QSAR and SVM models used for predicting BRAF-V600E inhibitors. For space reasons, the same plot for HIV Inhibitors is not shown.

The actual prediction accuracy is determined by carrying out predictions on a test set that is 20% the size of the original training set. All of the unknown examples in the test set are unique but have the same *SimEC* measure. That is, the descriptor values of compounds in the original test set were modified to achieve a specified degree of similarity/dissimilarity with the training set. Then predictions were carried out to achieve the actual prediction accuracy shown in the Fig. 1 and were observed to be within an average of 7% of the predicted accuracy.

The results of comparing the predictive accuracy between SVM and 3D-QSAR for BRAF-V600E and HIV-Integrase, discussed earlier, clearly indicates that the performance of SVM is better for the BRAF-V600E and HIV-targets than 3D-QSAR. With respect to just the BRAF-V600E target, Fig. 1 shows that it is possible for 3D-QSAR models to achieve comparable or better prediction results than SVM if:

- Candidate unknowns are more similar to the training data for the 3D-QSAR model
- The VC_{Dim} is decreased relative to the number of samples
- The cross validation error rate is reduced

Such plots can help answer questions such as, "Which classification models are most appropriate to use for the screening task of interest?" by first assessing the similarity of prospective unknown examples and training data. Then compare the respective *EPP* values. The classifier with the lower *EPP* value is more appropriate to use for the given unknown example and training set.

From Fig. 1, it can be seen that for the trained and optimized SVM and 3DQSARclassifiers used to make BRAF-V00E inhibitor prediction, the performance of the SVM model meets or exceeds the performance of the 3D-QSAR model using unknown examples that are between 20 to 30% more similar to the respective training set. Conversely, classifying unknown examples using 3D-QSAR would require they be 20 to 30% more similar to the 3D-QSAR training set to achieve comparable prediction accuracy of the SVM.

The question "How many samples should be used to carry out a classification task?" can be answered by using, again, the VC_{Dim} and following equation:

$$N = \Theta\left(\frac{VC_{Dim} + \ln\left(\frac{1}{\delta} \right)}{\varepsilon} \right) \tag{10}$$

where, δ is the desired minimum successful prediction probability and ε is the desired max training error (Vapnik, 1999). Answers to the question, "How many descriptors should be used to carry out a classification task?" can be approximated by solving Equation 10 for VC_{Dim} once a sample size, δ and ε have been chosen.

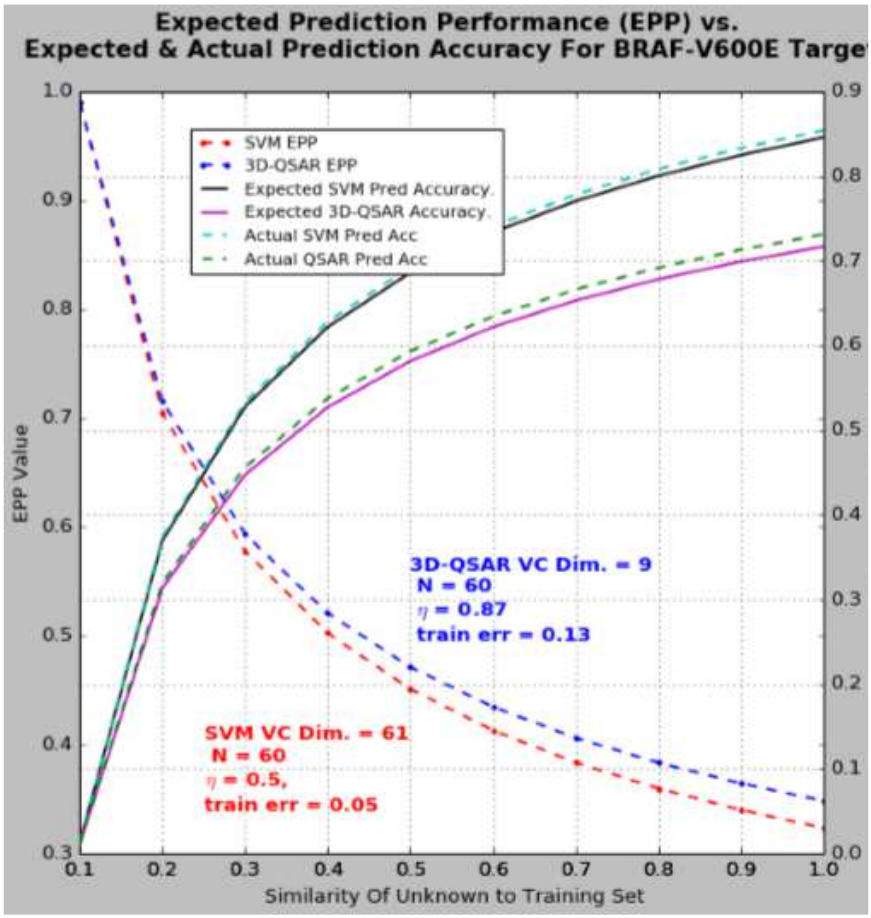

Fig. 1. Graph of the EPP metric (Shown as Upper Bound of Prob Prediction Error Rate) vs. similarity between unknown examples to be predicted and training data, versus expected and actual prediction accuracy

Answers to the questions like "Which descriptors should be identified and ordered in terms of importance to carry out a classification task?" can be answered by carrying out the desired dimension reduction step, such as SVD, PCA and so forth.

Using the EPP Metric

The *EPP* metric is intended to help explain why one of the 3D-QSAR or SVM models performs better than the other. It is also intended to help quantify the degree to which 3D-QSAR or SVM model is likely to perform better on targets and ligands that are of current interest. Finally, algebraic manipulations of the *EPP* metric can be used to quantify the number of samples, descriptors, similarity between unknown and training set in order to improve predictive performance.

This can be accomplishing by completing the following steps:

- Build optimized 3D-QSAR and SVM classifiers using best practices
- Calculate/estimate VC_{Dim} for each classifier.
- Generate an *EPP* Vs. similarity Vs. expected prediction accuracy plot for each classifier
- Complete predictions of desired compounds
- Compute *SimEC* measure for each desired compound
- Look up *EPP* values for corresponding *SimEC* value for a given compound. The classifier with a lower *EPP* value will be the classifier with a higher expected prediction accuracies. The EPP and corresponding expected prediction accuracy helps explain the relative performance of the classifiers
- Use the *EPP* vs. similarity vs. expected prediction accuracy plot to quantify the degree of change in similarity, *EPP*, or VC_{Dim} that is needed to improve a classifier's performance
- Use Equation 2 to determine the minimum number of samples required to achieve a desired prediction accuracy. Alternatively, solve for VC_{Dim} in Equation 10 to determine the change in capacity that is needed to achieve the desired prediction accuracy

Results and Discussion

About 303 potential BRAF-V600E inhibitors and 204 HIV Integrase inhibitors were collected and used to test the prediction accuracies of 3D-QSAR and SVM classifiers. Observed results indicate that the SVM classifier performed over 15% better than the 3D-QSAR classifier for both targets. A similarity measure *SimEC* and *EPP* measure was developed and used to explain and predict the difference in prediction accuracy of both models. Part of the explanation why the 3D-QSAR model did not perform as well as the SVM classifiers is that the 3D-QSAR model's "shatter capacity" was approximately 15% that of the SVM model's shatter capacity for the BRAF-V600E target. From Fig. 1, the EPP measure allows us to predict that the performance of the 3D-QSAR model might have been comparable to that of the SVM model if the unknown compounds predicted by the 3D-QSAR model were approximately $((3.6\text{-}3.3)/3.3) \times 100\% \approx 9\%$ more similar to the 3D-QSAR training data than that of the unknown compounds predicted by the SVM model. In other words, if the prediction accuracy of the 3D-QSAR model is expected to be comparable to the "optimal" SVM model, the unknowns predicted by the "optimal" 3D-QSAR model developed here need to be 10% more similar to the training data set than that of the unknowns predicted by the SVM model. Due to space constraints, a comparable Fig. 1 for the HIV-Integrase target was not generated and discussed here.

Conclusion

An optimized SVM classifier performs significantly better than an optimized 3D-QSAR model when predicting BRAF-V600E and HIV Integrase inhibitors.

The developed *SimEC* and *EPP* metrics appear to provide a means to explain, compare and predict performance with respect to prediction accuracy between SVM and 3D-QSAR models. These metrics can also be used to assess the likelihood of prediction accuracy of either or both models for yet to be classified compounds. The most appropriate ML-based classifier to use for HTS can thus be identified with greater fidelity by using the described *SimEC* and *EPP* metrics.

Future work to generalize the *EPP* and *SimEC* measures for a wider set of classification, regression and ML algorithms remains a topic area where even small advances are likely to yield significant benefits to the biopharmaceutical and larger ML communities.

Acknowledgment

The authors are thankful to Nathan Choo and Niharika Mandadi-Reddy for their insights and discussion about their previous work that involved using SVM classifiers to predict HIV Integrase inhibitors.

Funding Information

This work was supported, in part, by the SJSU Tower Foundation Account #034-1312- 0541

Author Contributions

Leonard Wesley: Performed BRAFV600E, HIV-Integrase, EPP work and preparing paper.

Saihitha Veerapaneni: Performed 3DQSAR BRAF-V600E and 3D-QSAR HIV Integrase work and preparing paper and supplement to paper.

Rachana Desai: Performed 3D-QSAR BRAF-V600E work and preparing supplement to paper.

Francisco McGee: Performed 3D-QSAR BRAF-V600E work.

Namrata Joglekar: Performed 3D-QSAR BRAF-V600E work.

Sheela Rao: Developed SVM classifier to help predict BRAF-V600E inhibitors.

Zeeshan Kamal: Provided information about HIV Integrase and related work on finding inhibitors.

Ethics

Authors declare no ethics violations.

Conflict of Interest

Authors declare no conflict of interests.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control, 19: 716-723. DOI: 10.1109/TAC.1974.1100705

Anderson, T.W. and D.A. Darling, 1954. A test of goodness-of-fit. J. Am. Stat. Assoc., 49: 765-769.

Bollag, G., J. Tsai, J. Zhang, C. Zhang and P. Ibrahim *et al*., 2012. Vemurafenib: The first drug approved for BRAF-mutant cancer. Nat. Rev. Drug Discovery, 11: 873-86. DOI: 10.1038/nrd3847

Darnag, R., E.L. Mostapha Mazouz, A. Schmitzer, D. Villemin and A. Jarid *et al*., 2010. Support vector machines: Development of QSAR models for predicting anti-HIV-1 activity of TIBO derivatives. Eur. J. Med. Chem., 4: 1590-1597. DOI: 10.1016/j.ejmech.2010.01.002

DiMasi, J.A., H.G. Grabowski and R.W. Hansen, 2015. The cost of drug development. N Engl. J. Med., 372: 1972-1972. DOI: 10.1056/NEJMc1504317

Gunawardana, A. and G. Shani, 2009. A survey of accuracy evaluation metrics of recommendation tasks. J. Machine Learn. Res., 10: 2935-2962.

Han, L., Y. Wang and S.H. Bryant, 2008. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. BMC Bioinform., 9: 401-401. DOI: 10.1186/1471-2105-9-401

Lamberti, M.J. and K. Getz, 2015. Profiles of new approaches to improving the efficiency and performance of pharmaceutical drug development. Tufts Center for the Study of Drug Development.

Li, C.Y., Q.S. Li, L. Yan, X.G. Sun and R. Wei *et al*., 2012. Synthesis, biological evaluation and 3D-QSAR studies of novel 4,5-dihydro-1H-pyrazole niacinamide derivatives as BRAF inhibitors. Bioorganic Med. Chem., 20: 3746-3755. DOI: 10.1016/j.bmc.2012.04.047

Liu, Y., Y. Zhou, S. Wen and C. Tang, 2014. A strategy on selecting performance metrics for classifier evaluation. Int. J. Mobile Comput. Multimedia Commun., 6: 35-35. DOI: 10.4018/IJMCMC.2014100102

Madhavan, T., 2012. 3D-QSAR in drug design-a review. J. Chosun Nat. Sci., 5: 1-5. DOI: 10.2174/156802610790232260

MOE, 2015. MOE is the commercial product of the Chemical Computing Group.

Murphy, R., 2011. An active role for machine learning in drug development. Nat. Chem. Biol., 7: 327-330. DOI: 10.1038/nchembio.576

Nantasenamat, C., C. Isarankura-Na- Ayudhya and V. Prachayasittikul, 2010. Advances in computational methods to predict the biological activity of compounds. Expert Opin. Drug Discovery, 5: 633-54. DOI: 10.1517/17460441.2010.492827

O'Boyle, N.M., M. Banck, C.A. James, C. Morley and T. Vandermeersch *et al*., 2011. Open Babel: An open chemical toolbox. J. Cheminform., 3: 33-33. DOI: 10.1186/1758-2946-3-33

PDB-5HES, 2016 3D-Structure image of BRAFV600E with Vermufanib inhibitor.

Pourbasheer, E., R. Alizadeh and M.A. Ganjali, 2015. QSAR study of CK2 inhibitors by GA-MLR and GA-SVM methods. Arab. J. Chem. DOI: 10.1016/j.arabjc.2014.12.021

Prahallad, A., C. Sun, S. Huang, F. Di Nicolantonio and R. Salazar *et al*., 2012. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature, 483: 100-103. DOI: 10.1038/nature10868

PubChem, PubChem source information. The PubChem Project. National Center for Biotechnology Information, USA.

Reich, Y. and S.V. Barai, 1999. Evaluating machine learning models for engineering problems. Artificial Intell. Eng., 13: 257-272. DOI: 10.1016/S0954-1810(98)00021-1

Sengupta, S. and S. Bandyopadhyay, 2012. Application of support vector machines in virtual screening. Int. J. Computat. Biol., 1: 56-62.

Shoichet, B.K., 2004. Virtual screening of chemical libraries. Nature, 432: 862-865. DOI: 10.1038/nature03197

Torfinn, S., 2014. R&D Cost estimates: MSF response to tufts CSDD study on cost to develop a new drug. MSF USA.

Vapnik, V.N., 1999. The Nature of Statistical Learning Theory. 2nd Edn., Springer Science and Business Media, New York, ISBN-10: 0387987800, pp: 314.

Vasanthanathan, P., O. Taboureau, C. Oostenbrink, N.P.E. Vermeulen and L. Olsen *et al*., 2009. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. Drug Metabolism Dispos., 37: 658-664. DOI: 10.1124/dmd.108.023507

Verma, J., V.M. Khedkar and E.C. Coutinho, 2010. 3D-QSAR in drug design--a review. Curr. Top. Med. Chem., 10: 95-115. PMID: 19929826

Wall, M.E., A. Rechtsteiner and L.M. Rocha, 2003. Singular Value Decomposition and Principal Component Analysis. In: A Practical Approach to Microarray Data Analysis, Berrar, D.P., W. Dubitzky and M. Granzow (Eds.), Kluwer, Norwell, MA., pp: 91-109.

Wainber, M. A., T. Mesplède and P.K. Quashie, 2012. The development of novel HIV Integrase inhibitors and the problem of drug resistance. Curr. Opin. Virol., 2: 656-662. DOI: 10.1016/j.coviro.2012.08.007

Yap, C.W., 2011. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Computat. Chem., 32: 1466-1474. DOI: 10.1002/jcc.21707

Wesley, 2016. 3D-QSAR and SVM prediction of BRAF-V600E and HIV integrase inhibitors: A comparative study and characterization of performance with a new expected prediction performance metric supplement.

Wozniak, M., M. Grana and E. Corchado, 2014. A survey of multiple classifier systems as hybrid systems. Inform. Fus., 16: 3-17. DOI: 10.1016/j.inffus.2013.04.006

Yao, X.J., A. Panaye, J.P. Doucet, R.S. Zhang and H.F. Chen *et al*., 2004. Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks and multiple linear regression. J. Chem. Inf. Comput. Sci., 44: 1257-1266. DOI: 10.1021/ci049965i

Zadrozny, B., 2004. Learning and evaluating classifiers under sample selection bias. Proceedings of the 21th International Conference on Machine Learning, (CML' 04), ACM, New York, pp: 114-114. DOI: 10.1145/1015330.1015425