Logistic regression and random forest unveil key molecular descriptors of druglikeness

Liza T. Billones*, Nadia B. Morales, Junie B. Billones

*Department of Physical Sciences and Mathematics, College of Arts and Sciences
University of the Philippines Manila, Padre Faura, Ermita, Manila 1000 Philippines*

**ltbillones@up.edu.ph*

(Received February 13, 2021; accepted August 15, 2021; published online September, 8, 2021)

Abstract

The identification of molecular descriptors that embody the chemical information for druglikeness will be a step forward in data-driven drug discovery and development endeavor. In this study, over 4000 Dragon-type molecular properties were generated for approximately 2000 known drugs and 2000 surrogate nondrugs. Logistic Regression (LogR) and Random Forest (RF) techniques were carried out to unveil the crucial molecular descriptors that can adequately classify a compound as drug or nondrug. Ten one-variable LogR models each demonstrated at least 70% prediction accuracy. A two-variable model consisting of *HVcpx* and *MDDD* correctly classified 85% of the test compounds. The best LogR model with 89.0% prediction accuracy identified five most influential descriptors for druglikeness: an information index *HVcpx*, topological index *MDDD*, a ring descriptor *NNRS*, *X2A* or average connectivity index of order 2, and walk and path count *SRW05*. The best RF model involving 10 only weakly correlated descriptors was found to be 92.5% accurate and at par with the RF and LogR models that consisted of over 200 variables. The model featured: molecular weight, *MW*; average molecular weight, *AMW*; rotatable bond fraction, *RBF*; percentage carbon, *C%*; maximal electrotopological negative variation, *MAXDN*; all-path Wiener index, *Wap*; structural information content index, neighborhood symmetry of 1 order, *SIC1*; number of nitrogen atoms, *nN*; 2D Petitjean shape index, *PJI2*; and self-returning walk count of order 5, *SRW05*. Many of these descriptors have straightforward chemical interpretability and future applicability as druglikeness filters in virtual high throughput drug discovery.

Keywords: Dragon molecular descriptors, Logistic Regression, Random Forest, druglikeness, high throughput, virtual drug discovery

Area of Interest: *In silico* drug discovery

1. Introduction

The use of computer-aided drug discovery techniques [1–2] and data analytics [3–5] that allows virtual screening of large databases, structure elaboration, and *in silico* drug development have accelerated the discovery of new therapeutic agents. In the traditional approach, most bioactive molecules have been discarded at the candidate drug development stage due to problems primarily associated with bioavailability, poor pharmacokinetic profiles, or issues with safety and efficacy [6–8]. Typically, this step is reached at only when substantial amount of resources have already been expended in discovering lead compounds and their derivatives. To avoid this undesirable scenario, additional screens [9–11] that distinguish potentially drug from nondrug compounds have been introduced at the hit-lead discovery stage [12–13] so that only those bioactive hits with druglike characteristics are forwarded for further development.

Several efforts have been exerted to find the chemical features that a compound need to possess for it to be considered a good lead-like or druglike compound [9–11,14–15]. A notable development in this area of medicinal chemistry is the groundbreaking paper of Lipinski on Rule of Five (ROF), which characterizes most orally bioavailable drug candidates [9, 16]. The original ROF specifies critical range of values for four simple physicochemical parameters (*i.e.* molecular weight, $MW \leq 500$; octanol-water partition coefficient, $\log P \leq 5$; hydrogen-bond donors, $HBD \leq 5$; hydrogen-bond acceptors; and $HBA \leq 10$) associated with 90% of orally active drugs that have reached the phase II clinical stage. Subsequently, Veber and co-workers established that the optimum number of rotatable bonds (NROT) is 7 and that NROT must not exceed 10 for a compound to display good oral bioavailability [10]. Clark and Pickett also stressed the importance of polar surface area (PSA) in the evaluation of druglikeness and proposed that PSA should not exceed $140 \text{ \AA}^2$ [11]. Moreover, in ‘fragment-based’ discovery [17–19], the hits identified generally obey a ‘Rule of Three’ (molecular weight, $MW \leq 300$; number of hydrogen bond donors, $HBD \leq 3$; number of hydrogen bond acceptors, $HBA \leq 3$; partition coefficient, $ClogP \leq 3$; number of rotatable bonds, $NROT \leq 3$; and polar surface area, $PSA \leq 60$) [20].

Furthermore, a more ambitious attempt to rapidly distinguish druglike compounds was made by Kubinyi and Sadowski [21] who developed a scoring scheme involving atomic physicochemical parameters [22] that can be used to classify molecules as drug or nondrug. A similar study aimed to distinguish between drugs and nondrugs by predicting the activity spectra for substances (PASS) [23] utilized substructure descriptors called “multilevel neighborhoods of atoms” based on the 2D representation of the molecules [24]. A more recent account also used limited number of molecular descriptors to classify compounds based on assigned probabilities [25].

Aside from the physicochemical properties associated with oral bioavailability [9–11], and the limited structure-derived descriptors used in the pioneering works in this area [21–25], the number of additional chemical descriptors that have been developed has grown [26–28] and the computational tools used to calculate them have become widely available [29]. An enormous number of physicochemical, electrotopological and graph theory-derived molecular descriptors [26–28] are still largely unexplored in virtual high throughput screening of compounds in drug discovery and development. These molecular properties can be quickly calculated from the composition and structure of a compound [30]. A cheminformatics software Dragon® [31], for example, is capable of generating over 4000 molecular descriptors per molecule. The descriptors can be classified as 0D (constitutional properties), 1D (*e.g.* functional groups, atom

centered fragments, information, and other molecular properties), 2D (e.g. topological, molecular walk counts, Burden eigenvalues, eigenvalue based indices, topological charge indices, connectivity, edge adjacency, and 2D autocorrelation), and 3D (charge, Randic molecular profiles, geometry, RDF, 3D-MoRSE, WHIM, and GETAWAY) [32]. We have demonstrated the utility of these molecular descriptors in predicting the activity of different sets of compounds against their corresponding molecular targets [33–35].

The identification of key molecular descriptors that confer druglikeness has been a long-standing goal in drug discovery. ROF, for instance, identified properties that promote oral bioavailability, which is being linked to druglike character of a compound. However, passing the ROF does not guarantee druglikeness [36]. A quick survey of the DrugBank (www.drugbank.ca) will show that TB drugs, antibiotics, and many other drug classes are not in agreement with ROF. The limited scope of ROF [9], the scanty descriptors used in earlier studies to classify molecules as drug or nondrug [21–25], and the scarcity of studies that exploit the intrinsic capacity of diverse molecular descriptors in discriminating drugs from nondrugs have prompted us to cast a wider net in hopes of finding alternative predictors of druglikeness.

As an offshoot of our previous work that employed Discriminant Analysis (DA) [37], this work also examined the two sets of compounds, approved drugs from DrugBank and surrogate nondrugs from the Enamine REAL database of synthetic compounds. The 3D structure of each compound was optimized and the Dragon-type molecular descriptors for each one were calculated. Likewise, the molecular descriptors served as the independent variables while the compound class (i.e. drug/nondrug) was the dependent variable. But, since DA assumes homogeneity of covariances matrices [38] that can however be relaxed on large data sets such as the one used in this study, we thought of considering other classification techniques that would not have such data requisite. Accordingly, we employed Logistic Regression and Random Forest methods in this follow-up study.

2. Methodology

Data collection was done using a personal computer running on Microsoft Windows 7 Professional 64-bit Operating System using a 3.50-GHz Intel® Core™ i7-4770K processor with 8.00-GB random access memory (RAM); and the data management, modeling, and analysis were handled using a machine with MacOS Catalina operating system, 3.1-GHz Dual-Core Intel Core i7 processor, and 16-GB RAM.

The structures of approved drugs were retrieved from the DrugBank database (<http://drugbank.ca>). The 3D structures of the compounds were generated using the CHARMM force field in BIOVIA Discovery Studio (DS) (<http://3dsbiovia.com>) and each structure was saved as structure-data file (.sdf). With the use of Dragon 6 software (<https://chm.kode-solutions.net>) over 4000 descriptors were calculated per molecule and the output was organized and cleaned in MS Excel. In the spreadsheet, rows and columns that have at least one *NAN* (*Not A Number*) entries and those with invariant values were removed; duplicate compounds, large compounds (MW > 1000), small compounds (MW < 100), and those with metal ions were likewise deleted. Some 25 compounds that were not successfully optimized in DS were also earlier excluded from the data set. Consequently, 1788 drug compounds were left for the analysis.

The structures of nondrug compounds were retrieved from Enamine REAL Collection database (<https://enamine.net>). Since the Enamine REAL Compounds database downloaded in 2017 is a huge data set of 22 million compounds, a probability sampling technique was applied

to derive a representative sample of nondrug compounds. With the observation that the classes of compounds in this collection were entered in no particular order, albeit a group of variants were entered successively, it was assumed that the list of compounds consist of chunks or clusters similar in terms of general attributes. Each cluster would be a mini-representation of the whole set of compounds. To assemble a good representation of the nondrugs, a three-stage cluster sampling was done. On the first stage, the whole collection of compounds was divided into chunks or clusters of 1 million compounds. These clusters were saved as different files, then a cluster or file (11th) was selected randomly (random.org) from among the 22 files. On the second stage, file 11 was divided into 4000 chunks or clusters consisting of 250 compounds each, and again saved as different files. From these 4000 clusters, 20 clusters were randomly selected and combined as one dataset of 5000 nondrug compounds; and then 3D structures were generated, molecular descriptors calculated, and data organized and cleaned like the drugs data set. On the third stage, 1787 compounds were randomly selected to obtain a nondrugs group of comparable size with the drugs group. Thus, there were a total of 3575 compounds in the working data set on 244 variables. It should be noted, however, that Enamine REAL Compounds database, the nondrugs data source, comprises synthetic compounds that already passed the ROF and Veber filters. But since a data source for nondrug compounds upon which to compare properties of drugs has not yet been established to date, compounds from the database were used as surrogate nondrugs. The choice of Enamine REAL database for this purpose was purely arbitrary, like all the previous studies that attempt to predict or quantify the druglikeness of a molecule wherein their choices had been a matter of preference. And as there is no existing criteria to evaluate a database of compounds qualifying it for a negative set, the only criterion applied in differentiating nondrugs from drugs is the absence of overlap between the two sets of compounds such that a compound not found in the drugs group was labeled nondrug.

A 70-30 split was applied to the data set of 3575 compounds that subsequently generated a training set of 2503 compounds (1252 drugs, 1251 nondrugs) and a test set of 1072 compounds (536 drugs, 536 nondrugs). Data modeling and analyses were done with SPSS Statistics 20.0.0 and RapidMiner Studio 9.7.001 (<https://rapidminer.com>) software. Logistic Regression and Random Forest modeling were carried out on the training set with the dependent variable (*i.e.* drug-nondrug classification) and the 244 independent variables (*i.e.* molecular descriptors).

Logistic Regression [39-40] determines a logistic or logit function (log-odds) to model a categorical dependent variable. This has been applied in various disciplines including mostly medical fields [41–42], social sciences [43], engineering [44], marketing [45], and drug discovery [46–47]. Logistic modeling is a suitable statistical and machine learning technique for classification giving the means for the identification of the determining predictors.

Random Forest (RF) is a machine learning algorithm that is used for classification and regression. As a classifier, it uses an ensemble of decision trees that output the mode of the results (classification) from the trees. A decision tree (DT) is a flowchart resembling a tree showing the process of arriving at an output of the target variable through the values of the features. The chart starts at a root node, then by the branches it reaches the internal nodes, up till the leaf nodes. The root node corresponds to the feature that has the most influence upon the target output, the internal nodes are the other features relevant for the target variable, the branches hold the feature values, and the leaf nodes carry the target output. The features generally characterize the instances in relation to the target variable. A trained DT classifies new instances by tracing their paths from the root node to a leaf node, and this process resembles following a set of decision rules. Thus, decision tree algorithm creates a model that predicts the value for the target variable gathering simple decisions from the data features. The predictive performance of a random forest depends on both the strengths of the individual trees

and on the correlation between the trees. Thus, aggregating the results of a multitude of trees trained by modified DT learning algorithm, taking only a random sample of features at each node, results in a robust RF model. In effect, RF combine all the information found in the forest making it more effective in predicting outcomes compared to single decision trees [48]. RF is now one of the most popular machine learning techniques and has been applied in data mining for computational biology [49–50] and medicine [51–52].

DT learning uses some criteria, such as *Information Gain* and *Gini Index*, that determine the best features on which to build a tree and cutoff values to split the nodes. Information gain is the reduction in entropy after splitting a node. Entropy is a measure of disorder of the target outputs in the nodes, in which higher values mean greater disorder (impure nodes) and lower values indicate purer nodes. Gini Index is a measure of inequality between the distributions of the target outputs. Like Entropy, lower values of Gini Index indicate purer nodes, and thus the feature that splits a node resulting in a smaller Gini Index or lower Entropy is chosen. *Information Gain Ratio* is another criterion used in the selection of features taking into account the number and size of branches thereby reducing the bias towards those features with multiple values. And being a classifier, *Accuracy*, is also used to evaluate the predictive performance of a DT model. Hyperparameters are also considered for speed and quality of the DT learning process, such as the number of predictors to be sampled (m_{try}) and considered to split a node, and the number of trees (n_{tree}) to be grown in the forest.

3. Results and Discussion

There are several approaches in determining the druglikeness of a compound. The most popular is the Lipinski rule of five (ROF) [9] that is based on certain structural features of the molecule (e.g. count of H-bond acceptors and H-bond donors) along with molecular weight and log P. While ROF and its successors [10,11] have the advantage of having features that require only simple counting and measurement, the rule does not necessarily point to druglikeness for such features were examined only for their association with oral bioavailability. Moreover, rule-based methods essentially use equal weights for all the parameters and entail sharp boundaries that lead to different classifications for highly similar molecules. These apparent limitations are overcome with the use of prediction models that do not only identify the crucial features but also provides more thorough quantitative appraisal for how druglike a compound is. For instance, the quantitative estimate of druglikeness (QED) developed recently by Bickerton [53] exhibits a continuous spectrum from the most druglike compounds to the least druglike ones. Unlike the rule-based classifiers that are prone to misclassification due to abrupt cut-offs, the structurally similar compounds will have comparable QED scores and consequently will have most likely the same predicted drug classification. Notwithstanding this appealing refinement, QED has also been defined mostly by only the traditional descriptors in ROF and its extensions, and need to determine for each class of therapeutics the adjustable weighting factors for the features. The proliferation of molecular descriptors in recent years provides great advantage of being able to consider more features that will be utilized in determining druglikeness. On top of that, the DrugBank database contains not only oral drugs but also those that are administered through other routes, paving the way for the discovery of descriptors that dispense druglikeness intrinsic to a molecule, and not just oral bioavailability.

3.1 Multiple Logistic Regression

The dependent variable in this study is called ‘Compound Class’ with categories ‘Drug’ and ‘Nondrug’ and hence the binary logistic model was used with the Dragon-type molecular descriptors as predictors. Logistic Regression was first implemented on the original set of 244 variables using the default setting of the software. At this instance, 13 descriptors (*nBO*, *nH*, *nHet*, *nCIC*, *Rbrid*, *Qindex*, *BB1*, *SNar*, *Psi_e_A*, *MPC01*, *X0*, *X1*, and *TIC0*) were excluded for having strong collinearity with the other variables in the set exceeding a correlation coefficient of 0.98. With 231 predictors, Model 1 (Table 1) was constructed with a prediction accuracy of 93.00%, that is, in the test set of 1072 compounds it correctly classified 997 and misclassified only 75 compounds. For the class recall of the model, the sensitivity was 91.60% (out of the 536 drugs it misclassified only 45 as nondrugs) and specificity was 94.40% (out of the 536 nondrugs it misclassified only 30 as drugs). As regards precision, the positive prediction value was 94.24% (only 30 out of 521 predicted drugs are actually nondrugs) and the negative prediction value was 91.83% (only 45 of the 551 predicted nondrugs are actually drugs).

Table 1. Confusion Matrix of Model 1 (LogR Model on Compound Class with 231 Dragon-type Descriptors)

accuracy: 93.00%			
	true Drug	true Nondrug	class precision
pred. Drug	491	30	94.24%
pred. Nondrug	45	506	91.83%
class recall	91.60%	94.40%	

One of the most popular evaluation tools for classification methods is the receiver operator characteristic curve or ROC curve. It is the true positive rate or TPR (sensitivity) as a function of false positive rate or FPR also called fall-out (*i.e.* 1-specificity). ROC is also known as relative operating characteristic curve as it depicts relative trade-offs between two operating characteristics, TPR and FPR, as the threshold values are varied. It displays the summary of all the corresponding confusion matrices produced by a classification method at different discrimination threshold settings [54–55] so that a point in the curve represents particular instance of a confusion matrix. Perfect classification occurs at point (0,1) where the prediction method yields 100% for both sensitivity and specificity. The diagonal line from point (0,0) to point (1,1) is so-called line of no discrimination that is much the same as random guesses. Curves above the diagonal line represent good classification and the closer an ROC curve is to the top left corner the better, such curves have higher values for both sensitivity and specificity. Compound classification by Model 1 is highly satisfactory with the optimal threshold of the model presenting high values for both sensitivity and specificity (Figure 1).

Figure 1. ROC curve of Model 1 (LogR model on Compound Class with 231 Dragon-type descriptors)

The area under the curve (AUC) of the ROC curve (or AUC-ROC) is another performance evaluation tool that indicates how much a model is able to separate between the categories of the dependent variable. When AUC equals 1, it means that the method perfectly distinguishes between the two classes so that an excellent model would have an AUC close to 1. An AUC near 0.5 means that the model is not able to differentiate the classes and a value near 0 implies that the classification method reverses its prediction [56]. Model 1 has an AUC of 0.976 conveying that it can almost perfectly distinguish drugs from nondrugs. Model 1 with 231 descriptors on 2503 compounds in the training set satisfies the ‘rule of ten’ (*i.e.* 10 compounds per descriptor) in Logistic Regression [57].

Nevertheless, the major goal of this work was to identify the key structure-derived Dragon-type molecular descriptors that confer druglikeness. Thus, a feature selection process is necessary to generate a model with a good prediction performance while constituting only the most important chemical features. With the 231 descriptors, 1-variable logit modeling was implemented using each of those descriptors. Ten of the resulting single-predictor logit models exhibited at least 70% prediction accuracy (Figure 2). The featured descriptors were *HVcpx*, *IDE*, *Yindex*, *DECC*, *ICR*, *AECC*, *NNRS*, *MSD*, *X5A*, and *MAXDP*. The two-tailed t-tests done on these variables at 5% level of significance showed significantly different means between drugs and nondrugs. As the best among them, *HVcpx* singlehandedly distinguishes drugs from nondrugs with 75% accuracy. This result had been consistent with our previous study that employed Discriminant Analysis where the best 2-variable discriminant function also included the descriptor *HVcpx*, which alone correctly classified 72% of the compounds [37].

Figure 2. Prediction accuracy of 10 topmost one-variable LogR models on Drug Class by Dragon-type descriptor

However, with only 75% prediction accuracy, compounds classification was not well resolved yet. Instinctively, the top ten descriptors were put together in a single 10-predictor logit model (Model 2). As a matter of course, this resulted in a raised prediction accuracy of 87.5% and AUC of 0.94 (Table 2). Model 2 adequately separates drugs from nondrugs with high values for both sensitivity and specificity and low classification errors such as false negative rate and false positive rate (Figure 3). The ROC curve likewise presents Model 2 with optimal point on the top left corner near point (0,1) indicating satisfactory classification performance.

Table 2. Prediction Performance of LogR Models on Drug Class

Model	LogR Equation	P-value	Nagelkerke R Square (%)	Prediction Accuracy* (%)	Sensitivity (%)	Specificity (%)	AUC
1	$P = -4.858 MW + 11.611 AMW - \dots - 53.34I$ $SpPosA_A - 1775^a$	-	-	93.0	91.6	94.4	0.98
2	$P = -19.003 IDE - 12.468 HVcpx - 64.410 XSA - 2.771 NNRS - 0.243 MAXDP - 0.385 DECC + 4.047 Yindex + 4.784 ICR + 0.810 AECC + 6.152 MSD + 53.654^b$	<0.001	56.9	87.5	80.0	95.0	0.94
3	$P = -0.012 MW + 0.812 AMW + 3.966 RBF - 0.675 nN - 0.061$	<0.001	52.6	82.3	74.2	90.3	0.88

	$C\% - 2.848 PJI2 + 0.000 Wap + 0.022 MAXDN - 0.496 SRW05 - 5.444 SIC1 + 6.220^c$						
4	$P = -11.87 HVcpx + 0.41 MDDD - 3.42 NNRS + 8.98 X2A - 0.57 SRW05 + 31.09$	<0.001	62.9	89.0	83.2	94.8	0.95

P = probability that a compound is a drug, * test set ($N = 1072$), ^awith 231 predictors (P-value and r^2 calculation failed due to multicollinearity), ^bwith 10 topmost predictors, ^cwith 10 weakly correlated predictors.

Figure 3. Predictive performance of Model 2 (LogR model on drug classification with 10 topmost descriptors)

Heat map displays the test set: 1–536 drugs and 537–1072 surrogate nondrugs.

Nonetheless, the correlation matrix was re-examined for these 10 predictors in Model 2 and it revealed 6 highly correlated ($r > 0.9$) descriptors indicating the presence of multicollinearity and accordingly violating the assumption of independence of variables for the logistic model [58]. To resolve this, the default value set for the correlation coefficient was changed to 0.4 so that variables which are at least moderately correlated are excluded from the analysis [59]. Then another Logistic Regression was carried out on the original set of 244 predictors, and this new setting left only 13 variables for the model. The accuracy of the resulting model was already lower by about 4 percentage points than the previous 10-variable model. For straightforward comparison with the previous 10-variable model, the number of variables in this model was further reduced to 10 using the forward stepping method of feature selection [37]. With this technique, 1-variable logit models were generated in the first round. The descriptor that gave the highest prediction accuracy was carried over to the second round involving 2-variable logit models, that is, the best predictor in round 1 was paired with each and every predictor left in the set. Then the best pair of predictors in round 2 was forwarded to the third round to generate 3-variable models. This procedure was iterated until the best 10-predictor logit model or Model 3 was obtained consisting of only weakly correlated descriptors namely, *MAXDN*, *nN*, *Wap*, *MW*, *AMW*, *SRW05*, *SIC1*, *PJI2*, *RBF*, and *C%*, according to the

order of which they entered into the model. Compared to Model 2 with also 10 predictors but some with strong correlation, this Model 3 had lower values for both predictive accuracy (82.3% vs 87.5%) and AUC (0.88 vs 0.94).

It would seem that some of the descriptors that were excluded due to the stringent requirement on multicollinearity contain information valuable to explain druglikeness. Hence, the correlation coefficient was reset at default value so that only variables that have very strong correlation would be eliminated in the process. Then another feature selection by forward stepping method was done on the 244 variables. *HVcpx*, the best single variable classifier (Figure 2), was paired with every variable in the set creating 230 two-variable logit models. The best model in this round consisted of the combination of *HVcpx* and *MDDD* that displayed a prediction accuracy of 85.4% and AUC of 0.91. This duo is exactly the same pair that constituted the best model in a similar study that employed Linear Discriminant Analysis (LDA) [50]. The predictive performance of this 2-variable model was even higher than Model 3 (Accuracy: 85.4% vs 82.3%, AUC: 0.91 vs 0.88) that included 10 weakly correlated descriptors. The inclusion of *NNRS* as a third variable enhanced the performance of the model (86.7% accuracy), and with two more variables [*X2A* and *SRW05*] into the model the prediction accuracy furthermore increased to 89.0% (Figure 4). Every predictor added to the model increased the confidence of correctly classifying compounds and decreased the probability of misclassification.

Figure 4. Predictive performance of a LogR model on drug classification with sequential inclusion of descriptors *HVcpx*, *MDDD*, *NNRS*, *X2A*, and *SRW05*. Heat map displays the test set: 1–536 drugs and 537–1072 surrogate nondrugs.

The resulting 5-predictor logit model (Model 4) correctly classified 446 of the 536 drugs (*i.e.* sensitivity was 83.2%) and 508 of the 536 nondrugs (*i.e.* specificity was 94.8%). Its AUC (Figure 5) further indicated very satisfactory performance correctly classifying 95% of compounds. The addition of the next most influential variable, *ONOV*, contributed only additional 0.66 percentage point to the prediction accuracy thus further addition of descriptors

into the model was no longer pursued. The Nagelkerke R square shows that Model 4 can better predict drug class compared to the two ten-variable models. Except for *HVcpx* and *MDDD* which were moderately correlated ($r = 0.78$), the rest of the pairs of descriptors in Model 4 were not strongly correlated ($r < 0.4$) [59].

Figure 5. Predictive performance of Model 4 (LogR Model on drug classification with descriptors *HVcpx*, *MDDD*, *NNRS*, *X2A*, and *SRW05*)

Heat map displays the test set: 1–536 drugs and 537–1072 surrogate nondrugs.

The most influential descriptor *HVcpx* is an information index called graph vertex complexity index [60–61]. Model 4 implies that a small value for *HVcpx* renders a molecule to be more druglike or that a more complex molecule is more likely a drug. The second important descriptor *MDDD* (mean distance degree deviation or $\Delta\sigma$) is the sum of the differences between the vertex distance degree (σ_i) and the mean vertex distance degree ($\bar{\sigma}$), where σ_i is sum of the topological distances or distances of all the other atoms j to atom i (d_{ij}). Terminal vertices are observed to have high values of vertex distance degree compared to central vertices. The vertex distance degree is small if the vertex is near a branching site compared to a terminal vertex that is far away from it [28]. The model shows that large values of *MDDD* are associated with druglikeness, thus, a molecule with distant peripheral groups tend to be a drug. The third most influential variable, *NNRS*, is a ring descriptor that stands for normalized number of ring systems [27]. A smaller *NNRS* value favors druglikeness. The ring portions in the molecule provide structural rigidity, which can be useful for anchorage or precise disposition of functional groups during interaction with a biological target. However, the model conveys that smaller *NNRS* is needed for a molecule to be druglike. The fourth and fifth variables, *X2A* (average connectivity index of order 2) and *SRW05* (self-returning walk count of order 5), are connectivity and walk count indices, respectively [27]. Larger values of *X2A* and smaller values of *SRW05* increase the probability of druglikeness. In general, these descriptors encode the structural complexity of a molecule. Incidentally, the *MW* in known classifiers are strongly correlated with *MDDD* ($r = 0.89$) and *HVcpx* ($r = 0.71$). In addition, the normalized number of ring systems (*NNRS*), although absent in ROF, is related to the number of aromatic rings (AROM) in Bickerton's QED [53].

3.2 Random Forest

In this work, each tree was grown using 67% of the instances randomly chosen from the training set. The remaining 33% so-called out-of-bag (OOB) samples were utilized to calculate the unbiased prediction error (*OOBError*) that was used to evaluate the model accuracy during training. The behavior of the *OOBError* as a function of the number of trees, n_{tree} , (Figure 6) was observed for each of the four criteria of feature selection and model performance namely, *Gini Index*, *Information Gain*, *Gain Ratio*, and *Accuracy*. This in turn became the basis for the determination of the number of decision trees that should be grown in the forest. The *OOBError* settled at a constant minimum starting at around 90 trees for all the four criteria. In fact, there were no remarkable changes in the *OOBError* values when the number of trees was increased up to 1000 (not shown). Therefore, $n_{tree}=100$ is just fitting and so this value for the number of trees was used in subsequent analyses. Information Gain and Gini Index showed comparable *OOBErrors* that are remarkably lower than those of Gain Ratio and Accuracy. Accordingly, either Information Gain or Gini Index was used in the prediction and validation steps.

Figure 6. Out-of-bag error with Gini Index, Gain Ratio, Information Gain, and Accuracy by number of trees

Calculated on 826 OOB samples (33% of training dataset).

The suitable maximal depth of each tree was also determined by plotting the *OOBError* against the depth of each decision tree (figure not shown). The *OOBError* curve was stabilized starting at depth of 7 for the criterion Accuracy, and at 20 for the other criteria. Using the optimal parameters ($n_{tree} = 100$, depth of tree = 20) with Gini Index as the selection criterion and confidence vote as the voting strategy, an RF model involving the 244 features (Model 5) was applied to the test set. With prediction accuracy of 94.2% and AUC of 0.982 (Table 3), this model performed better than the 231-predictor logit model (Model 1). Based on the attribute weights shown by the model, the 10 most influential descriptors are simply constitutional indices: *N%* (percentage of N atoms), *AMW* or average molecular weight (MW/nAT , where *nAT* is the number of atoms), *MW* (molecular weight, the sum of atomic weights), *Si* (sum of first ionization potentials), *Mv* (mean atomic van der Waals volume), *Mp*

(mean atomic polarizability), *Se* (sum of atomic Sanderson electronegativities), *Mi* (mean first ionization potential), *Sp* (sum of atomic polarizabilities), and *nBT* (number of bonds)[29].

However, when RF was implemented on only the five predictors in Model 4, the accuracy of the corresponding model (Model 6) was decreased to 86.7%, which is 2.3 percentage points lower than that of Model 4, albeit AUC remained high at 0.940. This slightly inferior outcome was the apparent result of having very few variables to create a forest. With only subsets of already small set of variables on which to create trees, only very few tiny trees would have been built that resulted in an information deficient random forest. More features would have been needed in order to build a powerful random forest. As a matter of fact, when another RF was carried out on the 10 predictors in Model 2, the resulting model (Model 7) performed much better (Accuracy = 89.9% vs 87.5 and AUC = 0.96 vs 0.94). Like Model 5 that performed better than its corresponding logit Model 3, Model 7 as well performed better than its 10-variable logit Model 2 counterpart. Most interestingly, the random forest Model 8 propagated by the 10 weakly correlated predictors performed even much better (Figure 7) than the corresponding logit Model 3 (Accuracy = 92.5% vs 82.3% and AUC = 0.97 vs 0.88). This remarkable performance of Model 8 with 10 predictors was only 1 and 2 percentage points shy of the accuracy of logit Model 1 and random forest Model 5, respectively, that consisted of more than 200 predictors!

Table 3. Prediction Performance of RF Models on Drug Class

Model	Input Variables in Random Forest Model (in order of decreasing attribute weight)	Prediction Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
5	<i>N%</i> > <i>AMW</i> > <i>MW</i> > <i>Si</i> > <i>Mv</i> > <i>Mp</i> > <i>Se</i> > <i>Mi</i> > <i>Sp</i> > <i>nBT</i> >>> <i>IDMT</i> ^a	94.2	91.4	97.0	0.982
6	<i>MDDD</i> > <i>HVcpx</i> > <i>X2A</i> > <i>NNRS</i> > <i>SRW05</i>	86.7	83.0	90.3	0.941
7	<i>MSD</i> > <i>AECC</i> > <i>DECC</i> > <i>MAXDP</i> > <i>ICR</i> > <i>Yindex</i> > <i>X5A</i> > <i>NNRS</i> > <i>IDE</i> > <i>HVcpx</i> ^b	89.9	86.2	93.7	0.962
8	<i>MW</i> > <i>AMW</i> > <i>RBF</i> > <i>C%</i> > <i>MAXDN</i> > <i>Wap</i> > <i>SIC1</i> > <i>nN</i> > <i>PJI2</i> > <i>SRW05</i> ^c	92.5	90.1	95.0	0.970

^awith 231 descriptors, ^bwith 10 topmost descriptors, ^cwith weakly correlated descriptors ($r < 0.4$).

Figure 7. Predictive performance of Model 8 (RF model on Drug Class with 10-weakly correlated predictors *MAXDN*, *nN*, *Wap*, *MW*, *AMW*, *SRW05*, *SIC1*, *PJI2*, *RBF*, and *C%*)

Heat map displays the test set: 1–536 drugs and 537–1072 surrogate nondrugs.

Each of the ten weakly correlated descriptors in Model 8 (Table 4) carried unique and vital information required in producing a diversified forest that yielded more accurate predictions. Half of the descriptors in Model 8 (and Model 3) were constitutional indices: *MW* (molecular weight), *AMW* (average molecular weight), *RBF* (rotatable bond fraction), *C%* (percentage Carbon), and *nN* (number of nitrogen atoms). On the other hand, *MAXDN* (maximal electrotopological negative variation), *Wap* (all-path Wiener index), and *PJI2* (2D Petitjean shape index) are topological indices; *SIC1* (structural information content index, neighborhood symmetry of 1 order) is an information index; and *SRW05* (self-returning walk count of order 5) is a walk and path count index [27]. The model correctly classified 90% of the drugs and 95% of the nondrugs and misclassified no greater than 10% of the drugs.

Table 4. Correlation of Descriptors in Model 8 (RF Model on Drug Class)

Attribu...	MW	AMW	RBF	nN	C%	PJI2	Wap	MAXDN	SRW05	SIC1
MW	1	0.006	0.065	0.266	0.127	0.113	0.360	0.251	0.113	-0.204
AMW	0.006	1	-0.017	-0.041	-0.328	-0.099	-0.022	0.253	-0.068	0.277
RBF	0.065	-0.017	1	-0.010	-0.313	0.152	-0.054	0.119	-0.183	-0.070
nN	0.266	-0.041	-0.010	1	-0.031	0.055	0.090	-0.016	0.307	0.273
C%	0.127	-0.328	-0.313	-0.031	1	0.085	0.040	-0.266	0.092	0.138
PJI2	0.113	-0.099	0.152	0.055	0.085	1	-0.033	0.065	0.055	0.002
Wap	0.360	-0.022	-0.054	0.090	0.040	-0.033	1	0.047	0.103	-0.089
MAXDN	0.251	0.253	0.119	-0.016	-0.266	0.065	0.047	1	-0.066	0.225
SRW05	0.113	-0.068	-0.183	0.307	0.092	0.055	0.103	-0.066	1	0.147
SIC1	-0.204	0.277	-0.070	0.273	0.138	0.002	-0.089	0.225	0.147	1

In this endeavor, the interpretation of the descriptors poses a challenge as most Dragon-type descriptors have extremely difficult chemical interpretations, and this makes their application in drug design quite daunting. Fortunately, for Model 5 and Model 8, the descriptors with the most information for druglikeness are largely constitutional indices that have straightforward chemical interpretations. Model 8 had a false negative rate of 9.89%, that is, it misclassified 53 out of 536 true drugs. Notably, a sizeable number (20 or 38%) of the misclassified drugs act on receptors associated with the central nervous system including the drugs for schizophrenia (Remoxipride, Sulpiride, and Olanzapine), and depression or anxiety (S-Adenosylmethionine, Fluvoxamine, Droperidol, Buspirone, Trazodone, and Flurazepam). The rest are antihypertensive drugs Isradipine, Nisoldipine, Prazosin, Brimonidine, Carteolol, Labetalol, Phentolamine, and Furosemide, and some commonly known drugs shown in Table 5. The ten descriptors of Model 8 (Table 3) are arranged in decreasing order of attribute weight. MW has the largest attribute weight with the drugs group having larger molecular weights. Of the 53 false negatives, 34 or 64% have molecular weights smaller than the lower bound of the 95% CI for the drugs group. Misclassifications result from having values that are not typical for the particular drug group, like values outside the 95% CI, on at least one of these 10 descriptors. The values on these features of a given molecule contribute to the probability that such would be classified as drug or nondrug. Although an optimal threshold value such as that determined by the Youden index [62] could afford an improved prediction performance, the cutoff value applied here was fixed at 0.5 simply due to its straightforwardness. Thus, the drugs with calculated probabilities below 0.5 were classified as nondrugs and comprise the false negatives (FN).

Table 5. Commonly Known Drugs Misclassified as Nondrugs by Model 8

Drug	Indication	Chemical Structure	Probability as Drug
Cimetidine	acid-reflux		0.41
Oseltamivir	influenza		0.40
Fluconazole	fungal infection		0.37
Chloroquine	viral infection		0.30
Linezolid	bacterial infection		0.04

In all of these, it should be noted that the nondrug compounds used in this study are rather surrogate nondrugs, which were taken from Enamine REAL Collection database that comply with ROF and Veber criteria for oral bioavailability. Consequently, the crucial descriptors unveiled in these study are those that distinguish approved drugs of all sorts (*i.e.* oral drugs and those that are administered through other routes) from nondrugs, which are represented by Enamine REAL compounds.

4. Conclusion

Logistic Regression and Random Forest models were generated using 244 variables from the original number of 4000 Dragon-type molecular descriptors derived from 3575 compounds. With the main objective of identifying alternative predictors for druglikeness, the models were trained on 70% of the samples (2503 compounds) and tested on the remaining portion of the samples (30% or 1072 compounds).

Based on the best LogR model in this study with prediction accuracy of 89.0%, the five most influential descriptors for druglikeness arranged in decreasing order of importance are: an information index *HVcpx* called graph vertex complexity index, topological index *MDDD* or mean distance degree deviation, a ring descriptor *NNRS* that stands for normalized number of ring systems, *X2A* or average connectivity index of order 2, and walk and path count *SRW05*. In general, a more structurally complex molecule (*i.e.* smaller *HVcpx*), with more distant peripheral groups (*i.e.* larger *MDDD*, *X2A*), fewer rings (*i.e.* smaller *NNRS*), and less branching around odd-membered rings (*i.e.* smaller *SRW05*) tends to be a drug. Interestingly, *HVcpx*, *MDDD*, and *NNRS* are correlated with some descriptors in well-known classifiers.

On the other hand, the best RF model with prediction accuracy of 92.5% identifies the following 10 descriptors in order of decreasing attribute weights: molecular weight, *MW*; average molecular weight, *AMW*; rotatable bond fraction, *RBF*; percentage carbon, *C%*; maximal electrotopological negative variation, *MAXDN*; all-path Wiener index, *Wap*; structural information content index, neighborhood symmetry of 1 order, *SICI*; number of nitrogen atoms, *nN*; 2D Petitjean shape index, *PJI2*; and self-returning walk count of order 5, *SRW05*.

Having straightforward chemical interpretability, future endeavors in drug discovery can exploit the key predictors of druglikeness identified in this study serving as alternative filters in virtual high throughput screening of compounds. The new generation molecular descriptors identified here offer fresh and alternative leverage in quantifying druglikeness.

As classifiers, RF models performed better than their LogR counterparts when there were more predictors. Comparing LogR models, the one in which stronger collinearity was allowed came out best in terms of accuracy and parsimony. Among RF models, the best one was that grown on only weakly correlated predictors.

A follow up study involving comparison of relative performance of our models, the rule-based classifiers, and quantitative measures of druglikeness, is underway in our group.

Acknowledgement

This work was fully supported by the Office of the Vice President of Academic Affairs, University of the Philippines System through the Enhanced Creative Work and Research Grant (ECWRG 2016-1-006).

References

- [1] Kapetanovic, I. M. Computer-aided drug discovery and development (CADD): in silico-chemico-biological approach. *Chem Biol Interact.* **2008**, 171(2), 165–176. doi:10.1016/j.cbi.2006.12.006
- [2] Zhang, W. (Ed). *Computer-aided drug discovery*, (1st Ed), Humana Press. 2016.
- [3] Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo D. P.; *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nat Mater.* **2019**, 18(5), 435–441. doi:10.1038/s41563-019-0338-z
- [4] Rifaioğlu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; *et al.* Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform.* **2019**, 20(5), 1878–1912. doi:10.1093/bib/bby061
- [5] Zhao, L.; Ciallella, H. L.; Aleksunes, L. M.; Zhu, H. Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. *Drug Discov*

- Today*. **2020**, 25(9), 1624–1638. doi:10.1016/j.drudis.2020.07.005
- [6] Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discov*. **2015**, 14, 475–486. doi:10.1038/nrd4609
 - [7] Hay, M.; Thomas, D. W.; Craighead, J. L.; Economides, C.; Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **2014**, 32(1), 40–51. doi:10.1038/nbt.2786
 - [8] Bunnage, M. Getting pharmaceutical R&D back on target. *Nat. Chem. Biol.* **2011**, 7(6), 335–339. doi:10.1038/nchembio.581
 - [9] Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, 46, 3–26. doi:10.1016/s0169-409x(00)00129-0
 - [10] Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smit, B. R.; Ward, K. W.; *et al.* Molecular Properties That Influence the Oral Bioavailability of Drug Candidates *J. Med. Chem.* **2002**, 45(12), 2615–2623. doi:10.1021/jm020017n
 - [11] Clark, D. E.; Pickett, S. D. Computational methods for the prediction of ‘drug-likeness’. *Drug Discov Today*. **2000**, 5(2), 49–58. doi:10.1016/s1359-6446(99)01451-8
 - [12] Peach, M. L.; Tan, N.; Choyke, S. J.; Giubellino, A.; Athauda, G.; *et al.* Directed discovery of agents targeting the Met tyrosine kinase domain by virtual screening. *J. Med. Chem.* **2009**, 52(4), 943–951. doi:10.1021/jm800791f
 - [13] Luethi E.; Nguyen, K. T.; Burzle, M.; Blum, L. C.; Suzuki, Y.; *et al.* Identification of selective norbornane-type aspartate analogue inhibitors of the glutamate transporter 1 (GLT-1) from the chemical universe generated database (GDB). *J. Med. Chem.* **2010**, 53(19), 7236–7250. doi:10.1021/jm100959g
 - [14] Hann, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J Chem Inf Comput Sci.* **2001**, 41(3), 856–864. doi:10.1021/ci000403i
 - [15] Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J Chem Inf Comput Sci.* **2001**, 41(5), 1308–1315. doi:10.1021/ci010366a
 - [16] Lipinski, C. A. Chris Lipinski discusses life and chemistry after the Rule of Five. *Drug Discov Today*. **2003**, 8(1), 12–16. doi:10.1016/s1359-6446(02)02556-4
 - [17] Carr, R.; Jhoti, H. Structure-based screening of low-affinity compounds. *Drug Discov Today*. **2002**, 7(9), 522–527. doi:10.1016/s1359-6446(02)02245-6
 - [18] Erlanson, D. A.; Braisted, A. C.; Raphael, D. R.; Randal, M.; Stroud, R. M.; *et al.* Site-directed ligand discovery. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, 97(17), 9367–9372. doi: 10.1073/pnas.97.17.9367
 - [19] Vetter, D. Chemical microarrays, fragment diversity, label-free imaging by plasmon resonance - a chemical genomics approach. *J. Cell. Biochem.* **2002**, 39, 79–84. doi:10.1002/jcb.10408
 - [20] Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov Today*. **2003**, 8(19), 876–877. doi:10.1016/s1359-6446(03)02831-9
 - [21] Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, 41(18), 3325–3329. doi:10.1021/jm9706776
 - [22] Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J Chem Inf Comput Sci.* **1987**, 27(1), 21–35. doi:

- 10.1021/ci00053a005
- [23] Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; *et al.* Discriminating between drugs and nondrugs by prediction of activity spectra for substances (PASS). *J. Med. Chem.* **2001**, 44(15), 2432–2437. doi:10.1021/jm0010670
 - [24] Filimonov, D.; Poroikov, V.; Borodina, Y.; Glorizova, T. Chemical similarity assessment through multilevel neighborhood of atoms: definition and comparison with the other descriptors. *J Chem Inf Comput Sci.* **1999**, 39(4), 666–670. doi:10.1021/ci980335o
 - [25] García-Sosa, A. T.; Oja, M.; Heteñyi, C.; Maran, U. DrugLogit: Logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. *J Chem Inf Model.* **2012**, 52(8), 2165–2180. doi:10.1021/ci200587h
 - [26] Todeschini, R.; Consonni, V. In: Molecular Descriptors for Chemoinformatics, WILEY-VCH, Weinheim (Germany), 2009.
 - [27] Talete. List of Molecular Descriptors Calculated by Dragon, n.d. Available at: http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf
 - [28] Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
 - [29] Yap, C. W. Pa-DEL Descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem.* **2011**, 32(7), 1466–1474. doi: 10.1002/jcc.21707
 - [30] Balaban, A. T. Drug Design, Molecular Descriptors in. In *Encyclopedia of Complexity and Systems Science*, Meyers, R. Ed.; Springer, New York, 2014.
 - [31] Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; *et al.* Virtual computational chemistry laboratory--design and description. *J. Comput. Aid. Mol. Des.* **2005**, 19(6), 453–463. doi: 10.1007/s10822-005-8694-y
 - [32] Tetko IV. Computing chemistry on the web. *Drug Discov Today.* **2005**, 10(22), 1497–1500. doi:10.1016/S1359-6446(05)03584-1
 - [33] Reyes, A. M. M.; Billones, J. B. Quantitative structure-activity relationship study of curcumin analogues as anti-proliferative agents of human prostate cancer cell line (PC-3), *Kimika.* **2013**, 24(1), 8–17. doi:10.26534/kimika.v24i1.8-17
 - [34] Billones, L. T.; Billones, J. B. Multiple linear regression model of shiga toxin inhibitory activity of dihydroquinazoline derivatives of retro-2^{cycl}. *Philippine Science Letters.* **2013**, 6(2), 231–240.
 - [35] Billones, L. T.; Billones, J. B. A Univariate analysis of molecular properties and inhibitory activity of dihydrothiophenones against dihydroorotate dehydrogenase of malaria parasite. *J. Chem. Pharma. Res.* **2014**, 6(8), 209–217.
 - [36] Petit, J.; Meurice, N.; Kaiser, C.; Maggiora, G. Softening the Rule of Five -- where to draw the Line?. *Bioorg Med Chem.* **2012**, 20(18), 5343–5351. doi:10.1016/j.bmc.2011.11.064
 - [37] Billones, L.; Gonzaga, A. C.; Billones, J. B. Molecular descriptors for drugs: A discriminant analysis, *Phil J Health Res Devel.* **2019**, 23(4), 11–16.
 - [38] Tabachnick, B. J.; Fidell, L. S. Using Multivariate Statistics, Harper Collins College Publishers, New York. 1996.
 - [39] Cramer, J. S. (2002). *The origins of logistic (PDF)* (Technical report). 119. Tinbergen Institute. pp 167–178.
 - [40] Cox D. R. The regression analysis of binary sequences (with discussion). *J R Stat Soc B.* **1958**, 20(2), 215–242. doi:10.1111/j.2517-6161.1958.tb00292.x
 - [41] Freedman, D. A. Statistical Models: Theory and practice. Cambridge University Press, 2009, pp 128.
 - [42] Boyd, C. R.; Tolson, M. A.; Copes, W. S. Evaluating trauma care: The TRISS method. Trauma score and the injury severity score. *J Trauma.* **1987**, 27(4), 370–378.

- [43] Harrell, F.; E. Regression modeling strategies (2nd ed.). Springer-Verlag, 2001.
- [44] Palei, S. K.; Das, S. K. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: an approach. *Saf Sci.* **2009**, *47*, 88–96. doi:10.1016/j.ssci.2008.01.002
- [45] Berry, M. J. A. Data mining techniques for marketing, sales and customer support. Wiley. 1997, pp 10.
- [46] Ferreira, L. L. G.; Andricopulo, A. D. ADMET modeling approaches in drug discovery. *Drug Discov Today.* **2019**, *24*(5), 1157–1165. doi:10.1016/j.drudis.2019.03.015
- [47] Chen, C.; Wang, H. Logistic regression method for ligand discovery. *J Comput Biol.* **2020**, *27*(6), 934–940. doi:10.1089/cmb.2019.0232.
- [48] Breiman, L. Random Forests. In: *Machine learning*, Kluwer Academic Publishers, The Netherlands. 2001, *45*, pp 5–32
- [49] Boulesteix, A.; Janitza, S.; Kruppa, J.; König, I. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics, *WIREs Data Mining Knowl Discov.* **2012**, *2*(6), 493–507. doi:10.1002/widm.1072.
- [50] Hsueh, H. M.; Zhou, D. W.; Tsai, C. A. Random forests-based differential analysis of gene sets for gene expression data, *Gene.* **2013**, *518*(1), 179–186. doi: 10.1016/j.gene.2012.11.034
- [51] Lind, A. P.; Anderson, P. C. Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties. *PLoS ONE.* **2019**, *14*(7), e0219774. doi:10.1371/journal.pone.0219774.
- [52] Tetschke, F.; Schneider, U.; Schleussner, E.; Witte, O. W.; Hoyer, D. Assessment of fetal maturation age by heart rate variability measures using random forest methodology, *Comput Biol Med.* **2016**, *70*, 157–162. doi:10.1016/j.combiomed.2016.01.020.
- [53] Bickerton, G. R.; Paolinim, G. V.; Besnard, J.; Muresan, S, Hopkins, A. L. Quantifying the chemical beauty of drugs, *Nat Chem.* **2012**, *4*(2), 90–98. doi:10.1038/nchem.1243
- [54] Fawcett, T. An introduction to ROC analysis. *Pattern Recognit Lett.* **2006**, *27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010
- [55] Detector performance analysis using ROC curves - MATLAB & Simulink Example". Retrieved from www.mathworks.com on Aug 21, 2020.
- [56] Narkhede, S. Understanding AUC-ROC curve, 2018. Retrieved from <https://towardsdatascience.com>.
- [57] Moons, K. G. M. ; de Groot, J. A. H.; Bouwmeester, W.; Vergouwe, Y.; Mallett, S.; *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS Checklist. *PLoS Med.* **2014**, *11*(10), e1001744. doi: 10.1371/journal.pmed.1001744
- [58] Assumptions of logistic regression. Statistics Solutions. Retrieved from <http://www.statisticssolutions.com> on Aug 22, 2020.
- [59] Rumsey, D. J. How to interpret a correlation coefficient r, Dummies, (nd). Retrieved from <https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/>.
- [60] Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.* **1984**, *5*(6), 581–588. doi:1002/jcc.540050612
- [61] Shojaie, A.; Sedaghat, N. How Different Are Estimated Genetic Networks of Cancer Subtypes? In *Big and Complex Data Analysis*, Ahmed, S. E.; Ed., Springer International Publishing AG, Switzerland, 2017.
- [62] Youden, W. J. Index for rating diagnostic tests, *Cancer.* **1950**, *3*(1), 32–35. doi:10.1002/1097-0142(1950)3:1<32::aid-cnrcr2820030106>3.0.co;2-3