RAIRO-Oper. Res. 55 (2021) 2769–2783
https://doi.org/10.1051/ro/2021134

RAIRO Operations Research
www.rairo-ro.org

DETERMINING THE BEST SET OF MOLECULAR DESCRIPTORS FOR A TOXICITY CLASSIFICATION PROBLEM

Badri Toppur[1,*] and K.J. Jaims[2]

Abstract. The safety norms for drug design are very strict with at least three stages of trials. One test, early on in the trials, is about the cardiotoxicity of the molecules, that is, whether the compound blocks any heart channel. Chemical libraries contain millions of compounds. Accurate *a priori* and *in silico* classification of non-blocking molecules, can reduce the screening for an effective drug, by half. The compound has to be checked for other risk factors alongside its therapeutic effect; these tests can also be done using a computer. Actual screening in a research laboratory is very expensive and time consuming. To enable the computer modelling, the molecules are provided in Simplified Molecular Input Line Entry (SMILE) format. In this study, they have been decoded using the chem-informatics development kit written in the Java language. The kit is accessed in the R statistical software environment through the *rJava* package, that is further wrapped in the *rcdk* package. The strings representing the molecular structure, are parsed by the *rcdk* functions, to provide structure-activity descriptors, that are known, to be good predictors of biological activity. These descriptors along with the known blocking behaviour of the molecule, constitute the input to the Decision Tree, Random Forest, Gradient Boosting, Support-Vector-Machine, Logistic Regression, and Artificial Neural Network algorithms. This paper reports the results of the data analysis project with shareware tools, to determine the best subset of molecular descriptors, from the large set that is available.

Mathematics Subject Classification. 62P10, 92-10.

Received February 22, 2021. Accepted August 14, 2021.

1. Introduction

There is another side to the Indian government response to handling the SARS-CoV2 pandemic, as the COVID-19 virus is also known. This response is from the molecular biologists and pharmacologists, who are designing a suitable antivirus to it. An invitation in the mail to participate in a drug design hackathon (DDH) was intriguing, and we would like to share what we have gathered, from the statements made by the biochemists, who have been working on the problem [2]. In the learning resources, are mentioned Indian herbal treatments such Ashwagandha (*Withania Somnifera*) and also a Chinese herbal remedy based on the Empress tree (*Paulownia Tomentosa*). Numerous synthesized compounds, that have been used effectively, for Middle Eastern respiratory syndrome (MERS) and the earlier version of SARS, have also been mentioned.

Keywords. Data mining, Bayesian classification problem, random forest, gradient boosting, biochemistry.

[1] Rajalakshmi School of Business, Chennai, India.
[2] DC School of Management and Technology, Kochi, India.
*Corresponding author: badri.toppur@rsb.edu.in; badri.toppur@gmail.com

The Protein Data Base (PDB) is a database repository for all the proteins, and researchers are attempting to identify precisely, the structure of the novel parts of the virus that has stormed the planet. Many parts of the virus, such as the spike and membrane appear to be identified, but there are missing elements that have not been crystallized and sequenced. It is common knowledge, that a protein is made up a sequence of amino acids, and there are 20 of these amino acids. The arrangement of amino acids, or signature is unique, and the challenge is to identify the sequence, so that one may inhibit the influence of the virus on human receptors. The identification of the protein folding, is complicated by mutations, that gives rise to different variants. Various companies in the field such as Centre for Development of Advanced Computing, India (CDAC), Schrödinger, ChemAxon, Optibrium, BioSolveIT, Molinspiration, and Cresset Software have provided tools for these *in silico* analysis and synthesis efforts. Problem statements and datasets were created, with expectations of the three dimensional structural model, and other information about the target molecule. Testing and search for the hit molecule is about using the computer to ascertain which drug molecule will inhibit the virus from entering into the human cell. Molecular Docking (MD) is a method used to try out a large number of small molecules from a database, and see which ones are suitable for a vaccine. In the case the structure of the target molecule is not clear, other approaches have to be used.

This paper is not targeted at the registered medical practitioner interested in drug discovery and design, which is a highly specialized and niche area. It is for the attention of the statistics and the machine learning community, because it reveals the complexities of a real world classification exercise, with a large dataset, 4.6 GB in size, involving many variables. A search in the Kaggle repository for similar datasets, brought up only two. This study, also emphasizes the use of easily available software, which democratizes the culture of statistical thinking and experiential learning considerably.

A second motivation for this work, is that a large segment of the educated population in India, is still unsure about the efficacy of the vaccination programme, after receiving information about fatalities soon after administration of the vaccine; only a small fraction of the Indian population has been vaccinated at present. A basic understanding of the one protocol of drug design discussed in the paper, will help clarify the safety pitch of the manufacturers to all citizens, that all manufactured drugs are screened scientifically during the clinical trials, to rule out allergic reactions.

2. Literature review

Notable researchers in the biophysics and biochemistry field are Paul and Gautham [17,18], Vengadesan and Gautham [23], and Wang [24]. A recent, but somewhat technical discussion of the viral pharmacokinetics is given by Hirano and Murakami [10]. Simplified Molecular Input Line Entry System is a notation or nomenclature that allows a user to represent a chemical structure in a way that can be parsed by the computer. This is compact way of representing the molecular structure of a drug molecule. A simple exposition of this notation is presented by Anderson *et al.* [3]. More detailed information about this notation is presented by Hunter *et al.* [11], Weininger [25] and Weininger *et al.* [26]. Roy, an expert in QSAR models, clarifies that toxic activity and therapeutic activity are the two sides of the same coin, in these classification studies. He also emphasizes, that in the end of the data analysis, a mechanistic interpretation of the final set of descriptors should be attempted [16]. Smith and Toppur, have investigated the geometrical structure of the Collagen protein, which is the *connective network* for transmitting forces in human and animal tissues, in the context of shortest interconnecting networks [20]. Collagen owes its unique properties not only to its chemical composition, but also to the physical arrangement of its individual molecules. The basic molecular polypeptide chain forms a left-handed helix, and three such helices are wrapped around each other to form a right-handed super-helix [12]. In retrospect, that was also a structure-activity study, which used the three-dimensional coordinates of the atoms. Atomic coordinates, have not been used so far in this study. Such an analytical approach combined with statistical methods, are sure to be useful with other molecules besides Collagen.

The dataset below in Table 1 shows an extract from a large dataset of molecules compiled by Sharath Kumar *et al.* [14]. The concern is about how they affect the human heart which is described by the dichotomous variable

TABLE 1. Extract from dataset displaying Molecular Notation and binary classification.

ID	SMILE	Class
1	[11CH3]Oc1ccc2cccc(N3CCN(CCCCN4N=CC(=O)N(C)C4=O)CC3)c2c1	Blocker
2	[2H]C([2H])([2H])Oc1cc(ncc1C#N)C(O)CN2CCN(C[C@H](O)c3ccc4C(=O)OCc4c3C)CC2	Blocker
3	[2H]C([2H])(O)CN1CCN(CC1)c2cnc3cc(cc(NCc4cccc(c4)[N+](=O)[O−])c3c2)C(F)(F)F	Blocker
4	[2H]C(Nc1cc(cc2ncc(cc12)N3CCN(C)CC3)C(F)(F)F)c4cccc(c4)[N+](=O)[O−]	Blocker
5	[2H]C(Nc1cc(cc2ncc(cc12)N3CCN(CC([2H])([2H])O)CC3)C(F)(F)F)c4cccc(c4)[N+](=O)[O−]	Blocker
6	[2H]C(Nc1cc(cc2ncc(cc12)N3CCN(CC3)C(=O)C)C(F)(F)F)c4cccc(c4)[N+](=O)[O−]	Non-Blocker
7	[C@@](c1c(F)cc(F)cc1)([C@H](N2Cc3c(nc(−c4cncnc4)s3)CC2)C)(O)Cn5ncnc5	Non-Blocker
8	[C@]123c4c5c(O)ccc4CC(N(CC3)CC=C)[C@]2(O)CCC([C@H]1O5)=N/N=C(/[C@H]6O7)CC[C@@]8(O)C(N(CC9)CC=C)Cc1ccc(O)c7c1[C@@]689	Blocker
9	[Cl−]	Blocker
10	[O−][N+](=Nc1ccccc1)c2ccccc2	Non-Blocker
11	[O−][N+](=O)C(Br)Br	Non-Blocker
12	[O−][N+](=O)c(c1)ccc(c12)nc(cc2)N3CCNCC3	Blocker
13	[O−][N+](=O)c1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	Blocker
14	[O−][N+](=O)c1c2ccccc2cc3ccccc13	Blocker
15	[O−][N+](=O)c1cc(c2ccccc2c1)[N+](=O)[O−]	Blocker

Blocker/Non-blocker. The researchers used Random Forest, Multilayer Perceptron and Sequential Minimal Optimization techniques. Random Forest models, were found to be most robust.

Various classification methods, have been explained in the SPSS guide by Elliott [7]. The mathematics underneath the method is explained, in the textbooks by James [13], and Kumar [15]. There are various diagnostic tests for determining the goodness of the model, such as precision, recall/sensitivity, specificity, Omnibus test, Wald test, Hosmer–Lemeshow test, Classification Plot, Receiver Operating Curve (ROC), and Area Under Curve (AUC). The precision of a model is the ratio of the number of true positives to the total number of predicted positives. The recall of a model is another name for the true positive rate. The specificity refers to the true negative rate. Approaches to Pattern Recognition (PR) in electrical and computer engineering, using structural, syntactic, and neural networks have been systematized as far back as 1992, by Schalkoff [19]. A recent book on deep learning, as the field of neural networks is referred to these days, and one specific to the R statistical environment, is by Chollet and Allaire [6].

Quantitative Structure-Activity Relationship (QSAR) models attempt to relate the structure of the molecule with its chemical and biological activity. A comprehensive and recent survey of QSAR approaches is presented by Abdel-Illah *et al.* [1]. They differentiate between structure based and ligand based approaches to drug design. QSAR is a type of ligand based drug design, in which the structure of the target molecule is not available. QSAR models are divided further into two dimensional and three dimensional models. We have seen in the literature, that some physical property such as boiling point of a compound can be predicted very precisely by a neural network that has been trained on inputs obtained from the structure of the molecule of the compound [8].

We next provide some context about the computer packages used in this paper; the Java computer language was developed at Oracle Corporation, in California. The Java software package CDK, stands for Cheminformatics Developer Kit and can be used for analysing molecular information. The OpenScience Project, can be used

for free, to analyse molecular information [22]. A derived wrapper package rcdk by Rajarshi Guha, makes the functionality of the CDK accessible, to beginner and intermediate users of the R language [9]. Although advanced, licensed, and commercial systems are being used at the pharmaceutical companies, these shareware packages can be used by small researchers, who are interested in the search for the present day holy grail – an effective antivirus for COVID-19. PaDEL is another popular software for generating the set of molecular descriptors using the CDK. Dong *et al.* have developed the Rcpi package that also provides similar descriptors [5]. Their paper mentions 48 descriptor types, yielding 288 molecular descriptors. The rcdk package by Guha, describe 55 molecular descriptors across 5 descriptor categories, which are enumerated in the next section. This is the package that we have used for obtaining the input variables.

The *Rattle* package in R from Togaware Pty, offers many classification tools [27]. We have tried out all the classification techniques, available in *Rattle*, namely, decision trees, Random Forest, Gradient Boosting, Logistic Regression, followed by Neural Network with one hidden layer. The package gives the results of most of the statistical tests such as sensitivity and specificity, Wald's test, ROC curve, and area under the curve, to determine the significance of the predictor variables, and many charts that can help one determine the robustness of the model. The package exhibits artificial intelligence; the variable selection rules for example, are built into the various classification algorithms. However, if the end user so wishes, he may also specify the variables to be used in a model.

3. Methodology

Just like there are various atomic descriptors, such as mass and charge, to tell apart different elements in the periodic table, the differences in the structure of a molecule is captured by the molecular descriptors. Common sense descriptors are counts of atoms, bonds, acids, bases, and about the lengths of chains and rings. The less obvious descriptors, are about attributes such as charge, polar surface area, anisotropy, and immiscibility. Anisotropy or *Aelotropy* is about possessing different physical properties in different directions; for example, certain crystals have a different refractive index, in different directions [12]. These descriptors may be used to relate the structure of the molecule to the biological activity of the molecule. First, we must look at the available descriptor categories. According to one popular taxonomy there are five descriptor classes:

(1) Hybrid – 2 descriptors.
(2) Constitutional – 16 molecular descriptors.
(3) Topological – 25 molecular descriptors.
(4) Electronic – 7 molecular descriptors.
(5) Geometrical – 5 molecular descriptors.

Adding across these five descriptor types, there are 55 molecular descriptors. However they are not mutually exclusive categories, and one or two descriptors appear in more than one category. The dataset in question, has 9204 molecules in SMILE format. Though we have taken a black-box approach and simply used the classification tools in the *Rattle* data-miner, a few words about the internal mechanism of these tools, is essential, so that one may know how they are different from each other. Naïve Bayesian classification is traditionally, the simplest approach to classification, based on branching at some critical value of a variable. With multiple input variables, the branching decisions becomes increasingly complex. This methodology is generalized to Random Forest which creates numerous decision trees to determine the best one. Random Forest is an ensemble of unpruned decision trees, that are robust to variance and bias. Random forests are often used when we have large training datasets and particularly a very large number of input variables [27]. In Gradient Boosting, a weight is associated with each observation in the dataset. A series of models are built and the weights are increased or boosted, if a model incorrectly classifies an observation. Support Vector Machine models (SVM) attempt to identify a separating hyperplane between the two classes of points reminiscent of discriminant analysis. Logistic regression works by generalizing the multiple regression function, with a sigmoid function. The logistics function is typically defined

in the form of a probability:

$$p = \frac{e^{(b_0+b_1X_1+b_2X_2+b_3X_3+\ldots+b_nX_n)}}{\left(1+e^{(b_0+b_1X_1+b_2X_2+b_3X_3+\ldots+b_nX_n)}\right)}.$$

This can be written as:

$$p = \frac{e^z}{(1+e^z)}.$$

After finding the odds ratio and taking the natural logarithm of both sides, we get the Generalized Linear Model (GLM) which resembles the multiple regression function. After the estimation of the beta parameters, one can calculate the class probability, for any set of values. From the probability, the odds, and furthermore, the Odds Ratio (OR) is calculated, that measures the importance of a predictor variable on the response. The relative importance of predictor variables is also ranked [15]. Finally neural networks, based on a connectionist architecture and view of learning patterns, adjust the values of a large network of weights, over many iterations, to correctly identify the output categories, when presented with an input.

Out of the 9204 molecules, 7630 molecules (83%) were classified as blockers, and 1574 molecules (17%) were classified as Non-blockers. Forty-nine unique descriptors, gave a total of 303 related molecular descriptors in the dataset. Only 190 were usable in the classification exercise, others being constant or "NA" or missing for the molecules. This then was the unbalanced dataset to use in the classification algorithms, with many of the molecules being of the blocker type. More weightage is to be given to the non-blocker data, since they constituted only 17% of the entire dataset, otherwise any hyperplane of discrimination, would be overfitted to the more prevalent class. A balanced dataset taking equal number of molecules from both classes had 3148 observations. All the experiments were performed on a Lenovo desktop computer running Windows 10, with an i3 dual core Intel chip, and 16 GB RAM.

4. Results

Though the variable selection was automatically done by the *Rattle* software, some exploratory data analysis was done to create box and whisker plots, and also histograms including probability density functions across the two classes. A few examples for selected descriptors are given in Appendix B. Such charts are useful to conjecture, which descriptors are going to be significant in the classification. If the boxplot for a class is higher or lower, or wider than for the other class, it is highly probable that the descriptor will play an important role in the model. Similarly if the overlap in the histograms for the two classes is almost perfect for some variables, then those descriptors are likely to play a minor role in the model. For any descriptor, a statistical two sample t-test gives the p-value for which the hypothesized mean difference for the two categories, is significantly non-zero. These t-tests can be used with the boxplots and other charts, to build up, like in step-wise regression, the number of input variables in the classification function.

4.1. Models with complete dataset

The complete dataset as mentioned earlier, has 9204 observations, and 192 variables. In our preliminary analysis, we used the entire dataset, and followed this up with the same battery of tests, on the balanced dataset.

4.1.1. Decision tree and random forest

The decision tree tool took 9.33 s, and uses only four variables, namely ALogP, khs.aaCH, MDEN.22, and XLogP. A definition for each descriptor is given in Appendix A. The classification error is very low at 2.2% for blocker class, and quite high at 84.5% for the non-blocker class. If we only want to eliminate the blocker class, this is good, but it is by no means a robust classification model. The random forest method took 3.50 min. The runtime is comparatively long, because it creates 500 decision trees, sampling 13 variables from the set of 190 input variables. The error for the Blocker class is 1.2%, and the error for the Non-Blocker class is reduced

Table 2. Classification matrix for the Test data – Logistic Regression.

	Predicted		
Actual	Blocker	Non-Blocker	Error
Blocker	78.4%	3.9%	4.7%
Non-Blocker	12.2%	5.5%	69%

Table 3. Classification matrix for the Test data – Neural Network.

	Predicted		
Actual	Blocker	Non-Blocker	Error
Blocker	72.6%	9.7%	11.8%
Non-Blocker	11.6%	6.2%	65.3%

to 57.6%. The top ten descriptors from this execution were, XLogP, ALogP, ALogp2, WTPT.5, MDEC.33, MDEC.13, ATSc3, Fsp3, tpsaEfficiency, and MDEO.11.

4.1.2. Gradient boosting & SVM

From the execution of the Gradient Boosting method on the full dataset, the important descriptors were, ALogP, XLogP, MDEO.11, MDEN.23, and WTPT.4. The Boosting method took 8.07 s. The results are inferior to the Random Forest method, with error for the Blocker class at 3.3% and error for the Non-Blocker class at 52.7%. The SVM method took 49.87 s. The error for the Blocker class was 0% but 98.4% for the Non-Blocker class, a highly polarized outcome.

4.1.3. Logistic regression

An execution, of the logistic regression method, took 5.16 min. It yielded a Chi-square p-value of 0.000 and a pseudo R-Square (optimistic) of 0.54497665. For the test set, the overall error was 16.1%, and the averaged class error was 36.85%. The results for this experiment are displayed in Table 2. We also ran the logistic regression without partitioning the dataset into training, validation and testing parts. The pseudo r^2 improved to 0.53 from 0.51, and the misclassification of non-blockers was reduced by about 2.4% from 69%.

4.1.4. Neural network

An artificial neural network with 10 neurons in a single hidden layer took 33.09 s, and the classification matrix was as in Table 3. Increasing the number of neurons did not improve the performance. Furthermore, in neural network implementations, one is unable to tell the most important variables.

Comparing all these methods with respect to classification error, the Random Forest, ensemble method emerges the best. However the almost 50% misclassification for Non-Blocker class indicates that the using the full dataset, with an under-represented class is not effective.

4.2. Models with balanced dataset

To counter the effects of imbalance, we next took equal number of molecules from each class in a dataset referred to in the following tables as the balanced dataset. Since we had 1574 non-blocker molecules, we took 1574 molecules from the blocker class, for a total of 3148 observations. All cited results are for the test data which constitutes 15% of the balanced dataset.

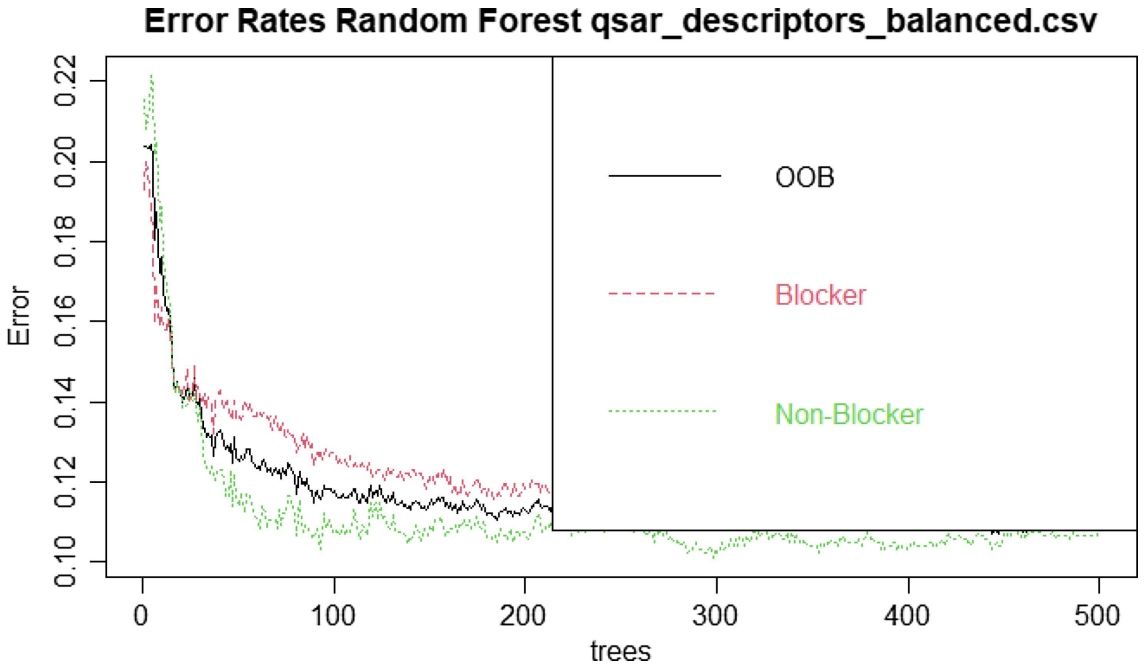


FIGURE 1. Error rate.

4.2.1. Decision tree and random forest

The decision tree method for the reduced dataset, took 2.29 s. The error for the Blocker class was 17.4% and for the Non-Blocker class, it was 19.0%. The eleven variables selected for splitting the tree were ALogP, C2SP2, ECCEN, FMF, MDEO.11, nAcid, nBase, nG, SC.5, SP.5, and VCH.7.

The Random Forest method took 37.49 s. The error for the Blocker class was 9.8% and for the Non-Blocker class was 6.4% with an overall error of only 8.1%. This is the best performance so far. Figure 1 shows how the error rate falls, as the number of decision trees calculated increases. An ROC chart plots the true positive rate against the false positive rate [27]. The Area under the curve (AUC) of 0.957 in the Receiver Operating Characteristics (ROC) curve, in Figure 2 shows that the accuracy of the model is very good. In this particular figure the false positive rate along the x-axis, is indicated as False Alarm Rate, and the true positive rate along the y-axis is indicated as Hit Rate.

The ten most important descriptors in descending order are ALogP, Alogp2, XLogP, tpsaEfficiency, MDEO.11, SC.5, nHBAcc, nAcid, TopoPSA, and WTPT.5. In fact, all the variables are ranked according to the mean decrease in the Gini index. Statistical t-tests for difference of two means resulted in very small p-values indicating that the means were significantly different for the two classes. One can see from the correlation clusters in Figure 3, that the partition coefficient variables (ALogP, Alogp2, and XLogP) that measure the hydrophobicity or immiscibility of organic compounds in water, are correlated at a low minimum distance (MD). The other seven variables, such as number of acids (nAcid) or number of Hydrogen bond acceptors (nHBAcc) are relatively independent.

4.2.2. Gradient boosting & SVM

The gradient boosting method takes only 2.10 s on this dataset. The error is 11.5% and 5.5% respectively, for the Blocker and Non-Blocker class. The overall error was 8.5% which is the best amongst all the methods used. Comparing this to the skewed and poor results obtained from applying Gradient Boosting to the full dataset implies that a balanced dataset is necessary for the method to work well. The five best descriptors are ALogP,

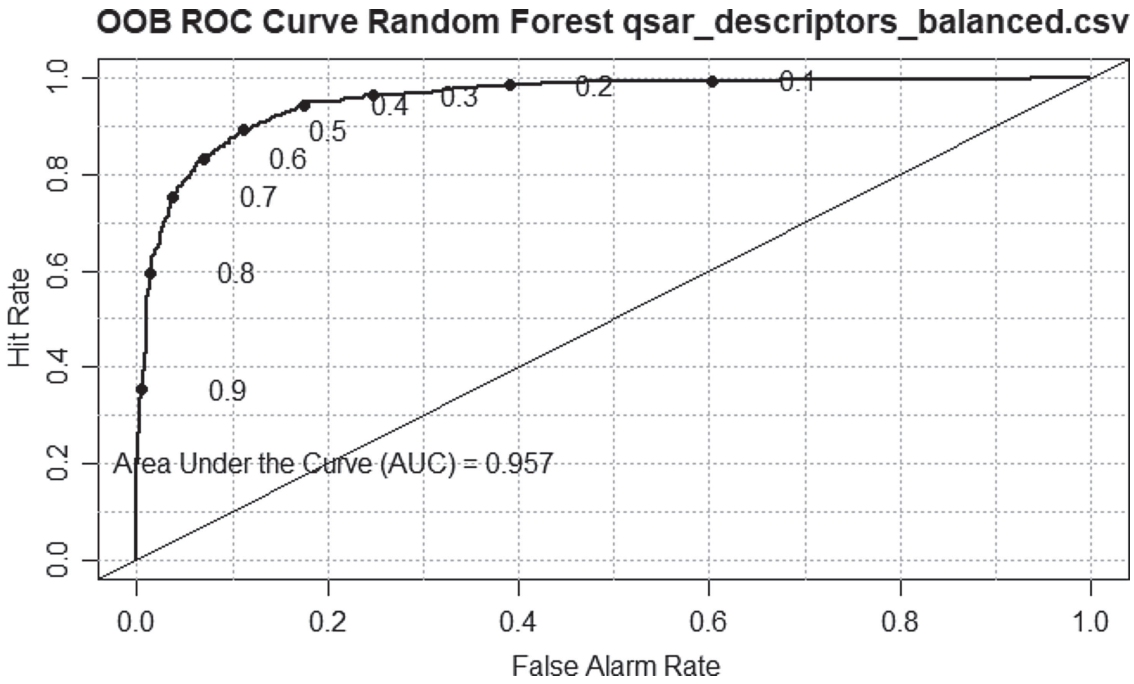


FIGURE 2. Receiver Operating Characteristic (ROC) curve.

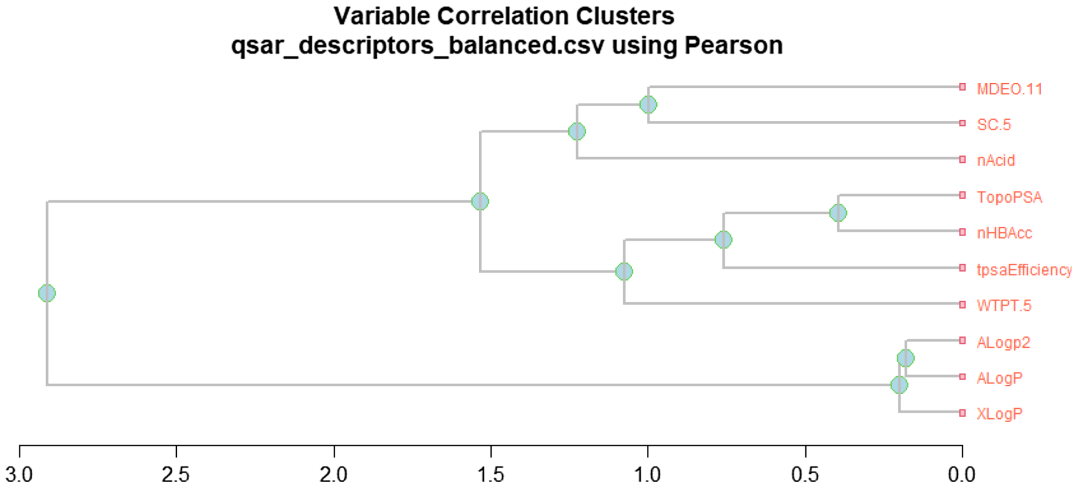


FIGURE 3. Variable correlation clusters.

MDEO.11, XLogP, nBase and SC.5. The classification matrix is displayed in Table 4. The SVM method takes 8.95 s. The classification error is high at 26.4% and 26.3% for the two classes. However it is equally accurate with both classes.

TABLE 4. Classification matrix for the Test set – Gradient Boosting.

	Predicted		
Actual	Blocker	Non-Blocker	Error
Blocker	44.2%	5.7%	11.5%
Non-Blocker	2.8%	47.3%	5.5%

4.2.3. Logistic regression

The logistics regression on this reduced dataset provided a much higher pseudo r^2 of about 0.62, up from 0.54 obtained for the full dataset. This does also suggest that imbalanced datasets detract significantly from the classification accuracy. However, from Table 4 for the test dataset, we can see that the classification error for the majority class, has gone up to 20.9% from 4.7%. The classification error for the minority class, is down from 69% to 26.3% which is 42.7% percentage points reduction in error. The overall error is 21%, and the averaged class error is 23.6%. Twenty four descriptors are shown to be significant at $\alpha = 0.001$. These are nAcid, aLogP, Alogp2, apol, nA, nG, nAromBond, ATSc1, ATSc3, ATSm4, ATsp2, nBase, C1SP2, C3SP3, SCH.3, VCH.4, VCH.7, VP.3, Kier3, khs.sCH3, khs.sBr, MDEN.33, WTPT.2, and WTPT.5. Out of all these descriptors, four descriptors are common with the set obtained from Random Forests. These are nAcid, ALogP, ALogp2, and WTPT.5, and they represent all three clusters of variables in Figure 3.

4.2.4. Neural network

The neural network with 10 neurons in the hidden layer, is trained in 1.63 s. The error is quite high at 53.2% and 33.5%.

5. CONCLUSIONS

We have tried the entire range of methods available in the *Rattle* data-miner. If we are concerned only with excluding the blocker type of molecule, in the interest of reducing the search space, then the balanced dataset is not a necessary condition, and the error percentage for the class, which constitutes 83% of the data, is less than 3%. If good classification performance is required for both classes, a balanced dataset is a must.

With the balanced dataset, the Random forest method shows accuracy over 91% for both classes. Gradient Boosting method has shown that classification accuracy can be as good as 88.5% for both classes with up to 94.5% accuracy for one class. The significant descriptors have been identified. These descriptors may be used to create a mechanistic model of the pharmacokinetic action of concern.

The neural network implementation did not show very good results. Neural networks with its many weights and parameters require many hours of fine tuning and the development of a tuned neural network for this dataset, in a dedicated deep learning framework such as *Keras*, is definitely a research direction to be taken. Even if classification is near perfect in such a neural network, a method to determine the best descriptors from the weights would be a scientific advance.

This study has shown that shareware software such as *rcdk* package and the *Rattle* data-mining tool, are competitive with the commercial grade software in this area of scientific research.

Appendix A. Molecular descriptors

Definitions of molecular descriptors

Descriptor name	Descriptor class	Definition
ALogP, ALogp2, AMR	Constitutional	Calculates atom additive logP and molar refractivity values as described by Ghose and Crippen
Apol	Electronic	Descriptor that calculates the sum of the atomic polarizabilities (including implicit hydrogens
nA, nR, nN, nD, nC, nF, nQ, nE, nG, nH, nI, nP, nL, nK, nM, nS, nT, nY, nV, nW	Protein, Constitutional	Returns the number of amino acids found in the system
naAromAtom	Constitutional	Descriptor based on the number of aromatic atoms of a molecule
nAromBond	Constitutional	Descriptor based on the number of aromatic bonds of a molecule
nAtom	Constitutional	Descriptor based on the number of atoms of a certain element type
ATSc1, ATSc2, ATSc3, ATSc4, ATSc5	Topological	The Moreau–Broto autocorrelation descriptors using partial charges
ATSm1, ATSm2, ATSm3, ATSm4, ATSm5	Topological	The Moreau–Broto autocorrelation descriptors using atomic weights
ATSp1, ATSp2, ATSp3, ATSp4, ATSp5	Topological	The Moreau–Broto autocorrelation descriptors using polarizability
BCUTw-1l, BCUTw-1h, BCUTc-1l, BCUTc-1h, BCUTp-1l, BCUTp-1h	Hybrid	Eigenvalue based descriptor noted for its utility in chemical diversity described by Pearlman *et al.*
bpol	Electronic	Descriptor that calculates the sum of the absolute value of the difference between the atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens)
nB	Constitutional	Descriptor based on the number of bonds of a certain bond order
PPSA-1 PPSA-2 PPSA-3 PNSA-1 PNSA-2 PNSA-3 DPSA-1 DPSA-2 DPSA-3 FPSA-1 FPSA-2 FPSA-3 FNSA-1 FNSA-2 FNSA-3 WPSA-1 WPSA-2 WPSA-3 WNSA-1 WNSA-2 WNSA-3 RPCG RNCG RPCS RNCS THSA TPSA RHSA RPSA	Electronic Geometrical	A variety of descriptors combining surface area and partial charge information
C1SP1 C2SP1 C1SP2 C2SP2 C3SP2 C1SP3 C2SP3 C3SP3 C4SP3	Topological	Characterizes the carbon connectivity in terms of hybridization
SCH-3 SCH-4 SCH-5 SCH-6 SCH-7 VCH-3 VCH-4 VCH-5 VCH-6 VCH-7	Topological	Evaluates the Kier & Hall Chi chain indices of orders 3, 4, 5 and 6
SC-3 SC-4 SC-5 SC-6 VC-3 VC-4 VC-5 VC-6	Topological	Evaluates the Kier & Hall Chi cluster indices of orders 3, 4, 5, 6 and 7
SPC-4 SPC-5 SPC-6 VPC-4 VPC-5 VPC-6	Topological	Evaluates the Kier & Hall Chi path cluster indices of orders 4, 5 and 6
SP-0 SP-1 SP-2 SP-3 SP-4 SP-5 SP-6 SP-7 VP-0 VP-1 VP-2 VP-3 VP-4 VP-5 VP-6 VP-7	Topological	Evaluates the Kier & Hall Chi path indices of orders 0, 1, 2, 3, 4, 5, 6 and 7

Definitions of molecular descriptors (continued)

Descriptor name	Descriptor class	Definition
ECCEN	Topological	A topological descriptor combining distance and adjacency information
fragC	Topological	Class that returns the complexity of a system. The complexity is defined as @cdk.cite{*Nilakantan06*}
GRAV-1 GRAV-2 GRAV-3 GRAVH-1 GRAVH-2 GRAVH-3 GRAV-4 GRAV-5 GRAV-6	Geometrical	Descriptor characterizing the mass distribution of the molecule
nHBAcc	Electronic	Descriptor that calculates the number of hydrogen bond acceptors
nHBDon	Electronic	Descriptor that calculates the number of hydrogen bond donors
Kier1 Kier2 Kier3	Topological	Descriptor that calculates Kier and Hall kappa molecular shape indices
khs.sLi khs.ssBe khs.ssssBe khs.ssBH khs.sssB khs.ssssB khs.sCH3 khs.dCH2 khs.ssCH2 khs.tCH khs.dsCH khs.aaCH khs.sssCH khs.ddC khs.tsC khs.dssC khs.aasC khs.aaaC khs.ssssC khs.sNH3 khs.sNH2 khs.ssNH2 khs.dNH khs.ssNH khs.aaNH khs.tN khs.sssNH khs.dsN khs.aaN khs.sssN khs.ddsN khs.aasN khs.ssssN khs.sOH khs.dO khs.ssO khs.aaO khs.sF khs.sSiH3 khs.ssSiH2 khs.sssSiH khs.ssssSi khs.sPH2 khs.ssPH khs.sssP khs.dsssP khs.ssssssP khs.sSH khs.dS khs.ssS khs.aaS khs.dssS khs.ddssS khs.sCl khs.sGeH3 khs.ssGeH2 khs.sssGeH khs.ssssGe khs.sAsH2 khs.ssAsH khs.sssAs khs.sssdAs khs.ssssssAs khs.sSeH khs.dSe khs.ssSe khs.aaSe khs.dssSe khs.ddssSe khs.sBr khs.sSnH3 khs.ssSnH2 khs.sssSnH khs.ssssSn khs.sI khs.sPbH3 khs.ssPbH2 khs.sssPbH khs.ssssPb	Topological	Counts the number of occurrences of the E-state fragments
nAtomLC	Constitutional	Returns the number of atoms in the longest chain
nAtomP	Constitutional	Returns the number of atoms in the longest pi chain
LOBMAX LOBMIN	Geometrical	Calculates the ratio of length to breadth
nAtomLAC	Constitutional	Returns the number of atoms in the longest aliphatic chain
MDEC-11 MDEC-12 MDEC-13 MDEC-14 MDEC-22 MDEC-23 MDEC-24 MDEC-33 MDEC-34 MDEC-44 MDEO-11 MDEO-12 MDEO-22 MDEN-11 MDEN-12 MDEN-13 MDEN-22 MDEN-23 MDEN-33	Topological	Evaluate molecular distance edge descriptors for C, N and O

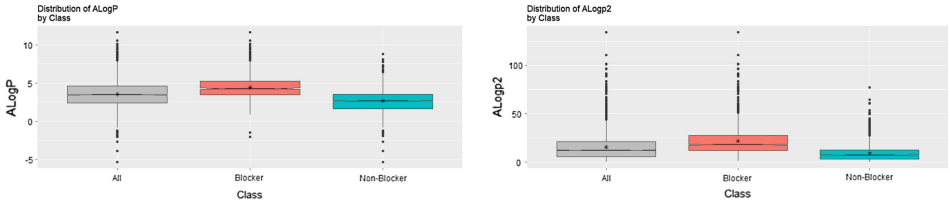
Definitions of molecular descriptors (continued)

Descriptor name	Descriptor class	Definition
MOMI-X MOMI-Y MOMI-Z MOMI-XY MOMI-XZ MOMI-YZ MOMI-R	Geometrical	Descriptor that calculates the principal moments of inertia and ratios of the principal moments. Also calculates the radius of gyration
PetitjeanNumber	Topological	Descriptor that calculates the Petitjean Number of a molecule
topoShape geomShape	Geometrical, Topological	The topological and geometric shape indices described by Petitjean and Bath *et al.* respectively. Both measure the anisotropy in a molecule
nRotB	Constitutional	Descriptor that calculates the number of nonrotatable bonds on a molecule
LipinskiFailures	Constitutional	This Class contains a method that returns the number of failures of Lipinski's Rule Of Five
TopoPSA	Topological, Electronic	Calculation of topological polar surface area based on fragment contributions
VAdjMat	Topological	Descriptor that calculates the vertex adjacency information of a molecule
Wlambda1.unity Wlambda2.unity Wlambda3.unity Wnu1.unity Wnu2.unity Wgamma1.unity Wgamma2.unity Wgamma3.unity Weta1.unity Weta2.unity Weta3.unity WT.unity WA.unity WV.unity WK.unity WG.unity WD.unity	Hybrid	Holistic descriptors described by Todeschini *et al.*
MW	Constitutional	Descriptor based on the weight of atoms of a certain element type. If no element is specified, the returned value is the Molecular Weight
WTPT-1 WTPT-2 WTPT-3 WTPT-4 WTPT-5	Topological	The weighted path (molecular ID) descriptors described by Randic. They characterize molecular branching
WPATH WPOL	Topological	This class calculates Wiener path number and Wiener polarity number
XLogP	Constitutional	Prediction of logP based on the atom-type method called XLogP
Zagreb	Topological	The sum of the squared atom degrees of all heavy atoms

Notes. Source: OCHEM (https://ochem.eu)

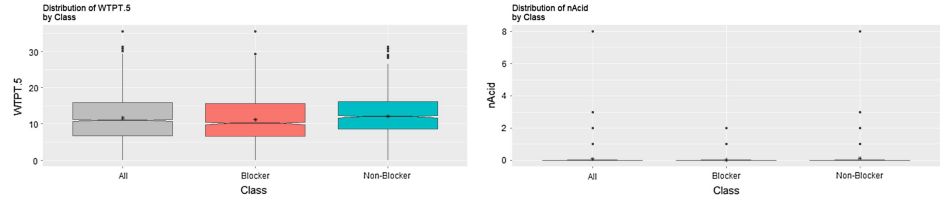
Appendix B. Exploratory data analysis

See Figures B.1–B.4.



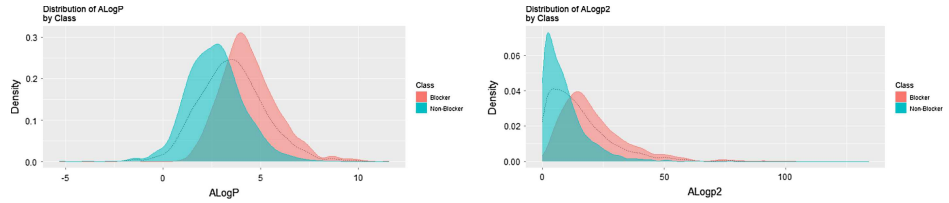
(A) (B)

Figure B.1. Boxplots for descriptors (a) ALogP and (b) ALogp2.



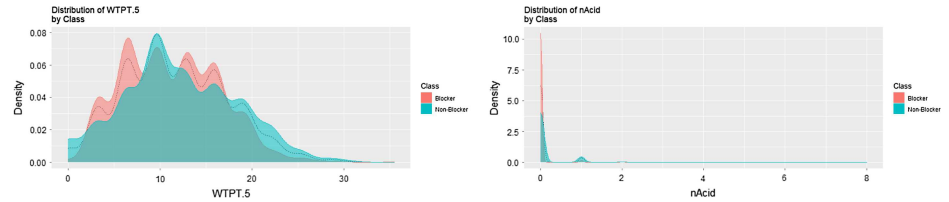
(A) (B)

Figure B.2. Boxplots for descriptors (a) WTPT.5 and (b) nAcid.



(A) (B)

Figure B.3. Histograms for descriptors (a) ALogP and (b) ALogp2.



(A) (B)

Figure B.4. Histograms for descriptors (a) WTPT.5 and (b) nAcid.

Acknowledgements. Kunal Roy, from Jadavpur University, piqued our interest in the classification problem, by sharing the dataset at the onset of the public health crisis. Rajarshi Guha, at the National Institutes of Health (NIH), Maryland, USA, was supportive, about the usage of the *rcdk* package.

References

[1] L. Abdel-Illah, E. Veljovic, L. Gurbeta and A. Badnjevic, Applications of QSAR study in drug design. *Int. J. Eng. Res. Technol.* **6** (2017) 582–587.

[2] P. Ambure, R. Balasaheb Aher, A. Gajewicz, T. Puzyn and K. Roy, "NanoBRIDGES" software: open access tools to perform QSAR and nano-QSAR modeling. *Chemom. Intell. Lab. Syst.* **147** (2015) 1–13.

[3] E. Anderson, G.D. Veith and D. Weininger, SMILES: a line notation and computerized interpreter for chemical structures. Report No. EPA/600/M-87/021. U.S. Environmental Protection Agency, Environmental Research Laboratory-Duluth, Duluth, MN 55804 (1987).

[4] M. Bruder and G. Polo and D.B.B. Trivella, Natural allosteric modulators and their biological targets: molecular signatures and mechanisms. *Nat. Prod. Rep. R Soc. Chem.* **37** (2020) 488–514.

[5] D.-S. Cao, N. Xiao, Q.-S. Xu and A.F. Chen, Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics* **31** (2015) 279–281.

[6] F. Chollet and J.J. Allaire, Deep Learning with R. Manning Publications Co. (2018)

[7] A.C. Elliott and W. Woodward, Statistical Analysis – Quick Reference Guide, With SPSS Examples. SAGE Publications, Inc. (2006).

[8] E.S. Goll and P.C. Jurs, Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model. *J. Chem. Inf. Comput. Sci.* **39** (1999) 974–983.

[9] R. Guha, Chemical informatics functionality in R. *J. Stat. Softw.* **18** (2007) 1–16.

[10] T. Hirano and M. Murakami, COVID-19: a new virus, but a familiar receptor and cytokine release syndrome. *Immunity* **52** (2020) 731–733.

[11] R.S. Hunter, F.D. Culver and A. Fitzgerald, SMILES user manual. A simplified molecular input line entry system. Includes extended SMILES for defining fragments. Review Draft, Internal Report, Montana State University, Institute for Biological and Chemical Process Control (IPA), Bozeman, MT (1987).

[12] A. Issacs and E.B. Uvarov, A Dictionary of Science. The English Language Book Society (1979).

[13] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, 1st edition. Springer (2013).

[14] L.S. Konda, S. Keerthi Praba and R. Kristam, hERG liability classification models using machine learning techniques. *Comput. Toxicol.* **12** (2019) 100089.

[15] D. Kumar, Business Analytics. John Wiley (2017).

[16] P.K. Ojha, S. Kar, J.G. Krishna, K. Roy and J. Leszczynski, Therapeutics for COVID-19: from computation to practices – where we are, where we are heading to. *Mol Divers* **25** (2021) 625–659.

[17] D.S. Paul and N. Gautham, MOLS 2.0: software package for peptide modeling and protein–ligand docking. *J. Mol. Model* **22** (2016) 239.

[18] D.S. Paul and N. Gautham, Protein-small molecule docking with receptor flexibility in iMOLSDOCK. *J. Comput.-Aided Mol. Design* **32** (2018) 889–900.

[19] R. Schalkoff, Pattern Recognition – Statistical, Structural and Neural Approaches. John Wiley & Sons, Inc. (1992).

[20] J.M. Smith and B. Toppur, Euclidean Steiner minimal trees, minimum energy configurations, and the embedding problem of weighted graphs in E^3. *Discrete Appl. Math.* **71** (1996) 187–215.

[21] M. Tardu, F. Rahim, H. Kavakli and M. Turkay, MILP-hyperbox classification for structure-based drug design in the discovery of small molecule inhibitors of SIRTUIN6. *RAIRO:OR* **50** (2016) 387–400.

[22] The OpenScience Project. https://cdk.github.io/cdk/2.3/docs/api/index.html?overview-summary.html.

[23] K. Vengadesan and N. Gautham, Enhanced sampling of the molecular potential energy surface using mutually orthogonal latin squares: application to peptide structures. *Biophys. J.* **84** (2003) 2897–906.

[24] J. Wang, Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. *J. Chem. Inf. Model.* **60** (2020) 3277–3286.

[25] D. Weininger, SMILES, a chemical language and information system. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28** (1988) 31–36.

[26] D. Weininger, A. Weininger and J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29** (1989) 97–101.

[27] G.J. Williams, Data mining with rattle and R: The art of excavating data for knowledge discovery. *Series Use R!* Springer (2011).