

# Finding Needles in a Haystack: Determining Key Molecular Descriptors Associated with the Blood-brain Barrier Entry of Chemical Compounds using Machine Learning

Subhabrata Majumdar,<sup>\*,[a]</sup> Subhash C. Basak,<sup>[b]</sup> Claudiu N. Lungu,<sup>[c]</sup> Mircea V. Diudea,<sup>[c]</sup> and Gregory D. Grunwald

**Abstract:** In this paper we used two sets of calculated molecular descriptors to predict blood-brain barrier (BBB) entry of a collection of 415 chemicals. First of 579 descriptors were calculated by Schrodinger and TopoCluj software. Polly and Triplet software were used to calculate the second set of 198 descriptors. Random forest modelling and a two-deep, repeated external validation method was used for QSAR formulation. Results show that both sets of descriptors individually and their combination give models

of reasonable prediction accuracy. We also uncover the effectiveness of a variable selection approach, by showing that for one of our descriptor sets, the top 5% predictors in terms of random forest variable importance are able to provide a better performing model than the model with all predictors. The top influential descriptors indicate important aspects of molecular structural features that govern BBB entry of chemicals.

**Keywords:** blood-brain barrier · molecular descriptors · variable selection · machine learning · random forest · two-deep cross validation · quantitative structure-activity relationship (QSAR)

## 1 Introduction

There is a lot of interest in the pharmacological, drug discovery, and computational chemistry literature in the prediction of blood-brain barrier (BBB) entry of molecules.<sup>[1–7]</sup> The BBB is an important barrier comprising endothelial cells and tight junctions between them which allow selective entry of solutes from the blood to the brain. The prospect of many promising psychoactive drug candidates will be diminished if they cannot cross the BBB to a desirable extent. Also, the drug designer would like to assure that drugs whose target is not the central nervous system do not cross the BBB to minimize undesirable side effects. Methods of rational drug design may produce many promising new drug entities, but they will not emerge as psychoactive drugs if they do not possess desirable BBB penetration profiles.

Various methods have been used before to assess the BBB permeability of chemicals. One quantifier of BBB entry is  $P_{app}$  which is the ratio of the concentration of a chemical in the brain to its concentration in the blood. Li et al.<sup>[7]</sup> developed computational models to predict this quantity for a set of 415 chemicals using molecular descriptors. But these days we have the availability of expanded set of descriptors, more sophisticated modeling methods as well as more computational power. Therefore, it was of interest to develop Quantitative Structure-Activity Relationship (QSAR)-s for the BBB entry of the 415-chemical set using molecular descriptors calculated by POLLY, APProbe, Triplet, and TopoCluj software and applying quantitative modeling

methods. For predictive modelling in the QSAR context, there is ample evidence that 'black box' type machine learning methods tend to perform better than conventional statistical techniques like logistic regression.<sup>[8,9]</sup> Among such methods, we decided to use random forest as our method of choice because of its effectiveness in the QSAR context when a large number of descriptors are involved.<sup>[9]</sup>

The present paper is organized as follows. In section 2 we give all details of our QSAR implementation: the data, descriptors, methods and validation techniques. Section 3 gives the results on our analysis, on which we elaborate

[a] S. Majumdar  
University of Florida Informatics Institute, 432 Newell Dr, CISE Bldg  
E251, Gainesville FL 32611, USA  
and  
Currently at AT&T Labs Research  
E-mail: subho@research.att.com

[b] S. C. Basak  
Department of Chemistry and Biochemistry, University of Minnesota,  
246 Chemistry Building, 1039 University Drive, Duluth MN  
55812, USA

[c] C. N. Lungu, M. V. Diudea  
Department of Chemistry, Babes-Bolyai University, Strada Arany  
János 11, Cluj-Napoca 400028, Romania

[d] G. D. Grunwald  
Natural Resources Research Institute, University of Minnesota, 5013  
Miller Trunk Highway, Duluth MN 55811, USA.

Supporting information for this article is available on the WWW  
under <https://doi.org/10.1002/minf.201800164>

upon in Section 4. We end the paper with concluding remarks in Section 5.

## 2 Methods and Materials

### 2.1 Data

The data (Supplementary File S1) on the BBB activity of 415 chemical compounds is taken from [7] the response variable is binary, and indicates whether a chemical is able to penetrate the BBB. We removed 3 compounds because of some descriptors not being calculated for them. Among the remaining chemicals, there are 275 compounds that are able to go through the BBB and 137 that are not able to. In the original paper, [7] we used several statistical and machine learning methods to model this binary BBB activity based on 199 molecular descriptors, and applied a feature selection method, Recursive Feature Elimination, [10] to determine influential features.

### 2.2 Molecular Descriptors

We use two collections of molecular descriptors for our QSAR model building exercise. The first set of descriptors, calculated and utilized by the Cluj team of Diudea and collaborators [11–14] were calculated by the softwares Schrodinger [15] and TopoCluj. [16] Detailed references about these descriptors are given in the Supplementary Table S2. We include 579 such descriptors in our analysis.

The second set of descriptors, used frequently by Basak et al., [17–20] is calculated using the softwares POLLY [21] and Triplet. [22] We use 98 and 100 descriptors calculated by these softwares, respectively. These descriptors are generally referred to as topological indices (TIs), since they are derived from graph theoretical methods. TIs include both topostructural (TS) and topochemical (TC) subclasses. The former encode information strictly on molecular connectivity. The latter include chemical features in addition to topological information. These chemical features include atom and bond type. Table 1 provides a list of the TIs from the Basak lab used in this study, along with brief descriptions.

### 2.3 Modelling Technique

We use Random Forest (RF) to obtain our QSAR models. First introduced by [23] this method has become one of the main pillars of predictive modelling in the past two decades. Technically, RF is an ensemble of multiple decision trees. When training each of these trees, a different random subset of the full feature-set is considered for each split in that tree. To obtain predictions for a set of values of the full feature set, one prediction is obtained from each of the

decision trees trained. For a regression problem, these tree-level features are averaged to obtain the final prediction. On the other hand, in classification problems, the aggregate predicted class is that predicted by a majority of the trees. RF models have previously been used in QSAR modelling in diverse scenarios. [24–26] For BBB modelling, the efficacy of machine learning models has been demonstrated previously by [7,27] For this reason, we choose random forests to obtain our QSAR predictive models.

Variable importance in RF models is calculated by the total amount of Gini impurity reduction caused by trees containing a specific variable. [23,28] We use this as our measure of variable importance in the results section.

While there are a host of machine learning methods are available and widely used for predictive modelling, the measures of variable importance used for them are different and often empirical. [29] Because of the high degree of collinearity exhibited by QSAR descriptors, [9,30,31] it is likely that multiple methods would lead to very different sets of important features being selected for each method. This makes the discussion on outputs and possible mechanistic interpretations (as given in Section 3.1) difficult. To circumvent this confusion, we focus on RF as our chosen method of modelling.

### 2.4 Validation

#### 2.4.1 Beware of (Single-split) External Validation

We use a multi-split cross-validation method to evaluate our QSAR models. For a small set of chemical compounds, external validation based on a single train-test split is unstable, and a leave-one-out (LOO) approach is preferable for model assessment. [32,33] Furthermore, since the training set is smaller than the full dataset, it gives a biased estimate of the unknown standard error. [34] In our situation, even though we have a considerably large set of compounds, the model evaluation metrics vary considerably across different random splits of the data (see Supplementary file S4). To mitigate this, we use a multi-split external validation to evaluate the performance of our models. This simply means repeating an external validation method 100 times over different random train-test splits of the data, and taking the average of any metrics obtained over all such splits. Also known as Monte-Carlo Cross Validation (MCCV), [35] this validation technique provides stable estimates of evaluation metrics, [33] and provided that the number of splits considered is large, gives unbiased estimate of the generalization error, i.e.  $q^2$ . [34]

*Note:* There is some debate in the QSAR community regarding the 'best' method to perform cross-validation. As pointed out by one of our referees, this depends on the specific goal of validation exercise. When the aim is feature selection, Shao [36] pointed out that LOO-CV is prone to overfitting, and leave-many-out CV needs to be used. On

Table 1. Symbols and definitions of topological indices (TIs).

Topostructural (TS)	
	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
	Mean information index for the magnitude of distance
$W$	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
$P$	Degree complexity
$H^V$	Graph vertex complexity
$H^D$	Graph distance complexity
	Information content of the distance matrix partitioned by frequency of occurrences of distance $h$
$M_1$	A Zagreb group parameter = sum of square of degree over all vertices
$M_2$	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order $h = 0-6$
${}^h\chi_C$	Cluster connectivity index of order $h = 3-6$
${}^h\chi_{PC}$	Path-cluster connectivity index of order $h = 4-6$
${}^h\chi_{Ch}$	Chain connectivity index of order $h = 3-6$
$P_h$	Number of paths of length $h = 0-10$
$DN^2S_y$	Triplet index from distance matrix, square of graph order, and distance sum; operation $y = 1-5$
$DN^21_y$	Triplet index from distance matrix, square of graph order, and number 1; operation $y = 1-5$
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation $y = 1-5$
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation $y = 1-5$
$ASN_y$	Triplet index from adjacency matrix, distance sum, and graph order; operation $y = 1-5$
$DSN_y$	Triplet index from distance matrix, distance sum, and graph order; operation $y = 1-5$
$DN^2N_y$	Triplet index from distance matrix, square of graph order, and graph order; operation $y = 1-5$
$ANS_y$	Triplet index from adjacency matrix, graph order, and distance sum; operation $y = 1-5$
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation $y = 1-5$
$ANN_y$	Triplet index from adjacency matrix, graph order, and graph order again; operation $y = 1-5$
$ASV_y$	Triplet index from adjacency matrix, distance sum, and vertex degree; operation $y = 1-5$
$DSV_y$	Triplet index from distance matrix, distance sum, and vertex degree; operation $y = 1-5$
$ANV_y$	Triplet index from adjacency matrix, graph order, and vertex degree; operation $y = 1-5$
Topochemical (TC)	
$O$	Order of neighborhood when $IC_r$ reaches its maximum value for the hydrogen-filled graph
$O_{orb}$	Order of neighborhood when $IC_r$ reaches its maximum value for the hydrogen-suppressed graph
$I_{ORB}$	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
$IC_r$	Mean information content or complexity of a graph based on the $r^{th}$ ( $r = 0-6$ ) order neighborhood of vertices in a hydrogen-filled graph
$SIC_r$	Structural information content for $r^{th}$ ( $r = 0-6$ ) order neighborhood of vertices in a hydrogen-filled graph
$CIC_r$	Complementary information content for $r^{th}$ ( $r = 0-6$ ) order neighborhood of vertices in a hydrogen-filled graph
${}^h\chi^b$	Bond path connectivity index of order $h = 0-6$
	Bond cluster connectivity index of order $h = 3-6$
	Bond chain connectivity index of order $h = 3-6$
	Bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	Valence path connectivity index of order $h = 0-6$
	Valence cluster connectivity index of order $h = 3-6$
	Valence chain connectivity index of order $h = 3-6$
	Valence path-cluster connectivity index of order $h = 4-6$
$AZV_y$	Triplet index from adjacency matrix, atomic number, and vertex degree; operation $y = 1-5$
$AZS_y$	Triplet index from adjacency matrix, atomic number, and distance sum; operation $y = 1-5$
$ASZ_y$	Triplet index from adjacency matrix, distance sum, and atomic number; operation $y = 1-5$
$AZN_y$	Triplet index from adjacency matrix, atomic number, and graph order; operation $y = 1-5$
$ANZ_y$	Triplet index from adjacency matrix, graph order, and atomic number; operation $y = 1-5$
$DSZ_y$	Triplet index from distance matrix, distance sum, and atomic number; operation $y = 1-5$
$DN^2Z_y$	Triplet index from distance matrix, square of graph order, and atomic number; operation $y = 1-5$

the other hand, when the assessment of the predictive capability of a QSAR model is desired, LOO CV-based  $q^2$  estimates have very little bias.<sup>[34]</sup> However, this is true only if hyper-parameter tuning and model selection steps are based *only on the training data of that LOO-CV layer*. Otherwise that amounts to model selection, not model

assessment, making the LOO-CV procedure biased again.<sup>[37,38]</sup> This distinction between model selection (cf. internal validation) and model assessment (cf. external validation) is sometimes misunderstood in the QSAR community. A proper two-step CV (to be discussed below) contains both: *the outer layer performs model assessment*

using repetitive predictions on test sets, which is preferable to a single train-test split as it is a more unbiased estimator of the true prediction error,<sup>[34,39]</sup> and the inner layer performs model selection that may involve selecting tuning parameters, and/or variable selection using statistical model selection criteria depending on the modelling technique used.

#### 2.4.2 Two-deep Cross Validation

In this study, we aim to compare predictions of random forest models based on the top  $\theta\%$  of predictors in terms of variable importance across a range of values for  $\theta$ . Since the variable importance depends on the trained model, it changes based on training samples it is trained on, i.e. across different train-test splits of the data. For this reason, it is imperative that a separate all-descriptor model (referred to as full model from now on) is trained for each train-test split, and then perform retraining based on the top predictors from the trained model for that specific split. Even though training the full model based on all available samples might intuitively seem the correct thing to do, this naïve approach would use information from holdout compounds in the training step, thus synthetically inflating all model metrics. The cross-validated  $q^2$ : obtained from such practices has been coined as naïve  $q^2$ .<sup>[40]</sup>


Since training a random forest model involves using cross-validation as well, we essentially perform cross-validation twice in our procedure: first time to train the full model inside each train-test split, and then to obtain model evaluation metrics. The steps of this two-stage cross validation procedure was coined as *Two-deep Cross Validation* by John Tukey.<sup>[37]</sup> In a QSAR scenario, two-layered cross-validation schemes have been used by.<sup>[8,39–43]</sup>

The steps for performing a two-deep CV in our context are as given below:

- Split the data into training and test sets randomly.
- Using samples in the training set, train a random forest model that uses all predictors.
- Obtain variable importance for all descriptors. Select the top  $\theta\%$  descriptors, use them to train a new RF model, and obtain validation metrics by evaluating the new model on the test compounds. Repeat for all values of  $\theta$  that are being considered.
- Repeat steps (a)–(c) for the requisite number of train-test splits.

#### 2.4.3 Metrics

We use the following performance metrics to compare predictive model outputs using the above validation technique.

-  **Area Under Curve (AUC):** This is defined as the area covered under a Receiver Operating Characteristic (ROC) curve

that plots the precision and recall values obtained from setting different thresholds to a set of predicted probabilities obtained from a classification model. The maximum value of AUC is 1, denoting perfect separation of the two classes of the response variable. Thus, a larger value of AUC indicates a better predictive model.

b) **Top 20% lift:** The lift metric evaluates performance of classification models at either extreme of the AUC, i.e. the accuracy of predictions with predicted probabilities close to 0 or 1. Specifically, the top  $x\%$  lift denotes the percentage of positive samples captured by the top  $x\%$  of predicted probabilities. This is useful in a QSAR situation where limited resources are available for further screening of chemicals, and the QSAR model is used to prioritize among a large number of sample compounds. We consider a hypothetical situation where only 20% of our samples can be further screened, i.e.  $x = 20$ . Thus, in order to ensure that further screening is highly likely to produce a positive compound, we would like our predictive model to have a high top 20% lift value.

### 3 Results

We present and discuss the outputs from our QSAR analysis in this section. In the first subsection, we give implementation details for our methodology. The following two subsections are concerned with the results obtained from our analysis, which can be classified into two parts. Firstly we aim to find out the important variables in our developed QSAR models, and compare these variables across the two different predictor sets as well as the combined set of predictors. After this we use subsets of the top predictors for each methods (like top 5%, 10%, 15%, ...), build new models with these variables, and compare their performance with the full models using different validation metrics. We used the statistical software R v3.3.2<sup>[44]</sup> to do all our data analyses.

#### 3.1 Details of Modelling

We used an ensemble of 500 trees in each of the RF models, and use the default setting in R to set the minimum node size in each tree to 1, which means that each tree is grown to the maximum possible depth. For model assessment, in the outer loop of our two-deep repeated external CV procedure, we use a 75:25 train-test split of the full set of samples, and 100 such splits are considered. In the inner loop of model training, we use a bootstrapped sample of the same size as training sample for model evaluation in order to utilize the most amount of data to train the decision trees.

### 3.2 Top Descriptors in QSAR Models

Table 2 lists the top 10 descriptors in the full models built on the Basak, Diudea and combined set of descriptors, respectively.

For the indices calculated using the POLLY<sup>[21]</sup> and Triplet<sup>[22]</sup> software by Basak laboratory, two classes of indices, viz. information theoretic neighborhood complexity indices and Triplet descriptors, emerged as the most influential in predicting BBB entry of chemicals. The IC-indices, developed by,<sup>[45]</sup> are related to the overall heterogeneity of atomic neighborhoods in the molecular structure. The ANZ4, AZN4, ANZ5, and DN2 N3 are triplet descriptors developed by.<sup>[46]</sup> These are local vertex invariants (LOVI's) which encode information for the presence of multiple bonds and/or heteroatoms in the molecular architecture. Consequently, these LOVIs may represent polarity/ polarizability in the molecules. Among previous studies,<sup>[7]</sup> found that polarity of molecules play important role in the prediction of BBB entry of molecules, and<sup>[47]</sup> reported that it is inversely correlated with the BBB permeability.

Regarding the influential indices from the Diudea lab of descriptors, the topological indices Sum.of.topological.distances.between.O..O, E.state.topological.parameter and Sum.of.topological.distances.between.N..O are descriptors developed at Topo Group Cluj, Romania, and others.<sup>[14,48,49,50]</sup> These are based on topological distance and detour, then on molecular graph fragmentation and collection of this information as fragmental property indices. Such indices express the presence of heteroatoms, by atomic radii and

Sanderson electronegativities, then converted in topological partial charges on the heavy atoms in the molecule. ALOGP3 is a quantifier of hydrophobicity and so may aid in passage of chemicals through biological membranes.<sup>[7]</sup> Both polar surface area (PSA) and hydrophobicity calculated by the ClogP program have been found to be influential in the prediction of blood-brain barrier entry of chemicals.<sup>[51,52]</sup>

### 3.3 Validation

We use a repeated external validation, also known as monte-carlo cross validation<sup>[35]</sup> in the literature, to evaluate the predictive performance of our QSAR models. To this end, we calculate each of the six metrics considered for the full models for each descriptor set, as well as random forest models built from the top 5%, 10%, ..., 90%, 95% of descriptors as per variable importance. Notice that in a two-deep validation setup, this means that for each train-test split, we take the descriptors that have the highest 5%, 10% etc. variable importances in the model obtained *using the training data of that specific split*. We summarize our results in Table 3 and Figures 1.

The two sets of descriptors behave very differently in the prediction curves. For the Diudea set, the top 5% descriptors are extremely predictive, and the variable selection approach actually manages to give a better-performing model than the full model according to all metrics except sensitivity. For sensitivity the performance of plateaus after 5% top descriptors are used. On the other hand, models corresponding the Basak set of descriptors

**Table 2.** Most important descriptors for the three predictor sets.

Basak lab Variable	Importance	Variable	Importance
IC1	2.23	IC4	1.32
IC2	1.90	SIC3	1.27
ANZ4	1.58	AZN4	1.22
SIC1	1.56	ANZ5	1.21
DN2 N3	1.54	SIC2	1.20
Diudea lab Variable	Importance	Variable	Importance
PSA	2.51	PEOE1	1.83
Sum.of.topological.distances.between.O..O	2.35	E.state	1.66
E.state.topological.parameter	2.20	Superpendentic	1.49
ALOGP3	2.06	Topological.charge.index.of.order.5	1.31
Sum.of.topological.distances.between.N..O	1.99	PEOE12	1.28
Combined Variable	Importance	Variable	Importance
Sum.of.topological.distances.between.O..O	2.30	E.state	1.31
E.state.topological.parameter	2.22	Sum.of.topological.distances.between.N..O	1.21
PSA	1.56	Molecular.electrotopological.variation	1.12
Superpendentic	1.44	PEOE1	0.99
ALOGP3	1.37	PEOE12	0.87



**Table 3.** Average and standard deviation (in brackets) of model metrics for the three descriptor sets.

Metrics	Descriptor set		
	Basak lab	Diudea lab	Combined
AUC	0.81 (0.05)	0.80 (0.05)	0.82 (0.05)
20 % Lift	4.70 (0.26)	4.65 (0.30)	4.77 (0.24)
Sensitivity	0.86 (0.05)	0.87 (0.04)	0.87 (0.05)
Specificity	0.56 (0.09)	0.53 (0.09)	0.55 (0.08)
Overall accuracy	0.75 (0.05)	0.75 (0.04)	0.76 (0.04)
MCC	0.44 (0.10)	0.43 (0.09)	0.44 (0.09)

perform more or less similarly for values of  $\theta$  larger than 25 across the metrics. On AUC, lift, sensitivity and MCC, the combined set of descriptors constantly perform better than both the individual descriptor sets for all values of  $\theta$ . This improvement is slightly better for all values of  $\theta$  when the Diudea set is considered, and for  $\theta > 25$  when the Basak set is compared with. AUC performance of the Diudea and combined sets are similar across  $\theta$ , while top 20% lifts of the combined set are slightly better. For specificity and accuracy, the performance comparisons are interesting. The combined descriptor set gives overall best performance when top 5% descriptors are used. But it deteriorates for higher values of  $\theta$ , and becomes worse than the individual predictor sets.

## 4 Discussion

The main objective of this paper was to use computed molecular descriptors in the prediction of BBB entry of a diverse set of 415 chemicals, and find out influential descriptors with potential mechanistic interpretations. To this end we used two sets of molecular descriptors, viz., set 1 of 198 descriptors calculated by POLLY and Triplet software routinely used by the Basak group at the University of Minnesota, and set 2 of 579 indices calculated by Diudea's group using TopoCluj and Schrodinger software. Results of our analyses using six model evaluation metrics in Table 3 show that the two sets of descriptors give similar results, 0.813 and 0.818, respectively, for set 1 and 2. When the combined set of 777 descriptors the AUC was 0.82 which indicates that the increase in the number of descriptors did not make any significant improvement in model quality. The strong mutual intercorrelation of many descriptors that is common in QSAR problems<sup>[9,31]</sup> may explain such results.

If we look at the most influential 10 descriptors (Table 2), the indices selected from the 198 POLLY and Triplet descriptor set are the information theoretic indices and triplet descriptors.<sup>[45]</sup> The former group quantify neighborhood complexity of atomic neighborhoods in the molecule, while the triplet indices characterize the heterogeneity of atoms in the molecular graph. It is noteworthy that in previous research, principal component analysis

(PCA) for a diverse collection of 3,692 chemicals showed that the information theoretic indices are strongly correlated with a distinct principal component (PC) indicating that such indices quantify some aspects of molecular structure not represented by other indices.<sup>[30]</sup>

High dimensional QSAR spaces described by computed molecular descriptors and machine learning methods in model building have been explored by various authors.<sup>[8,12,53,54]</sup> For the models developed from the set 2 of 578 descriptors calculated by TopoCluj and Schrodinger, PSA, sum of topological distances between O..O, E state topological parameter, ALOGP, and topological charge index of order 5 emerge as important variables. Others have found that hydrophobicity descriptors like CLogP value is important for BBB prediction.<sup>[55]</sup> PSA may be related to the extent of molecular polarity which is related to the hydrophobicity of the chemical. E state topological parameter and topological charge index quality the electronic nature of the solute.

## 5 Conclusion

In conclusion, the two sets of descriptors of 198 and 579 calculated for the set of 415 BBB data led to the development of good quality predictive models. The individual sets were independently as good as one another in the formulation of QSARs for the BBB data. Furthermore, through AUC and lift analysis of several sized subsets of the full feature sets, we underline the utility of sparse models in a QSAR context.

We demonstrate through our analysis the importance of parsimonious models, selected by using the top few important descriptors from an all-variable model, in QSAR model building in the context of BBB entry of chemicals. Considering the importance and practical relevance of QSAR problems, further studies with other data sets are needed to understand the utility of these mathematical molecular descriptors in assessing BBB entry, as well as other problem scenarios.

The use of reduced variable models represents its own set of challenges and extra cautions to be undertaken. According to the OECD principles for the validation of QSAR models,<sup>[56]</sup> any QSAR should be associated with a defined domain of applicability (AD) before deployment. When using a reduced variable model one needs to keep this mind that the AD for the full model and reduced model might not be the same. AD comparisons between these two approaches across different datasets and modelling methods are warranted to investigate this. We aim to address such questions in a analytically rigorous manner through our future research.

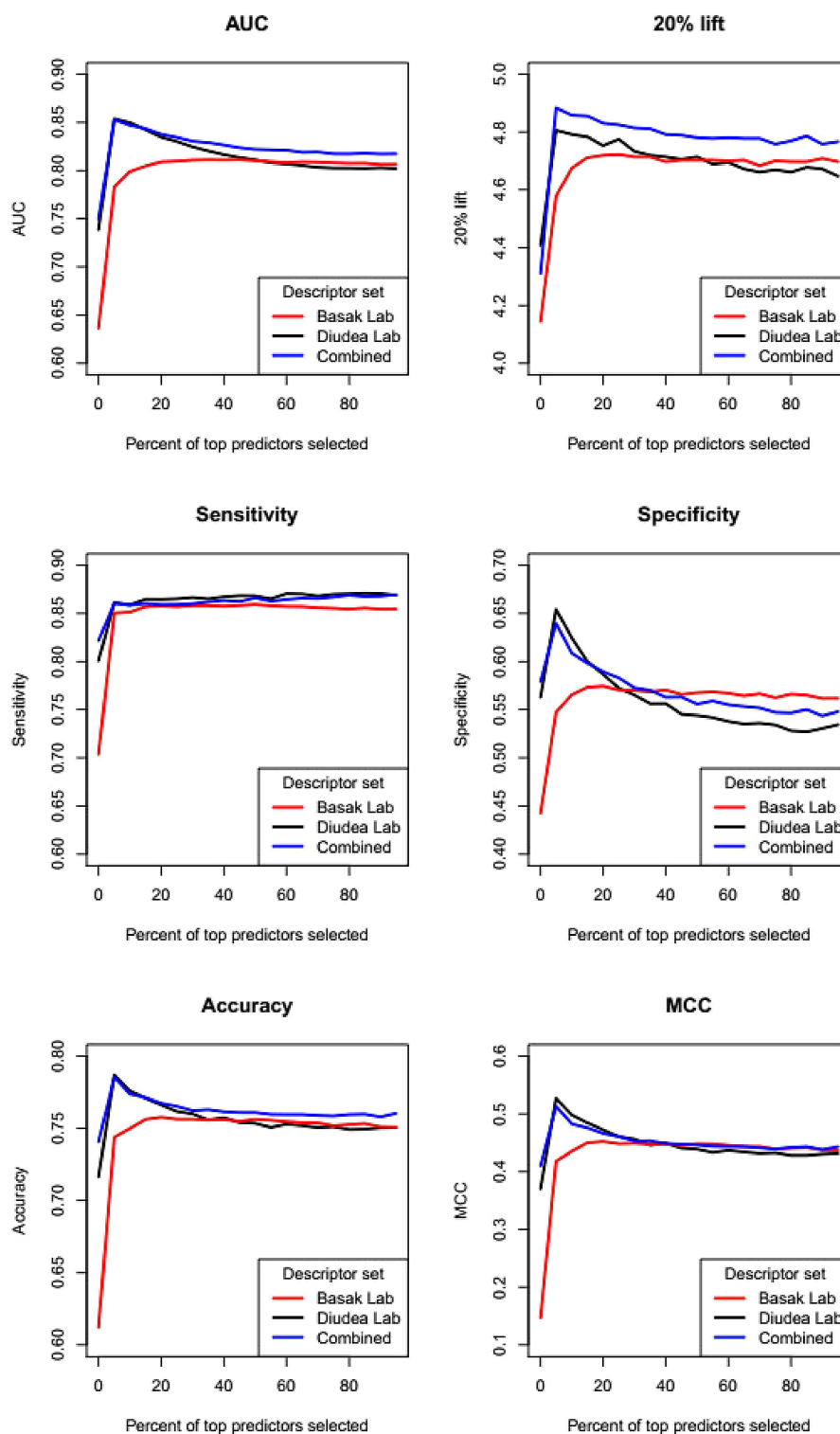


Figure 1. Validation metrics for reduced models with top  $\theta\%$  important descriptors.

## Conflict of Interest

We confirm that there is no conflict of interest on the content of this paper.

## Supplementary material

Supplementary file S1 contain all datasets used in our analysis. Files S2 and S3 contain information on the set of descriptors calculated by the Diudea lab. Supplementary file S4 contains outputs for all 6 metrics corresponding to the different train-test splits, and different values of top  $\theta$  % predictors. All code and outputs of the paper are available in <https://github.com/shubhobm/Blood-brain>.

## Acknowledgements

The authors dedicate this paper to Professor Paola Gramatica for her valuable contributions to the field of QSAR. The research of SM is supported by Prof. George Michailidis during his time at University of Florida Informatics Institute. We also thank the two anonymous referees for their thoughtful comments, which led to significant improvements in the paper.

## References

- [1] T. Hou, J. Wang, W. Zhang, *Curr. Med. Chem.* **2006**, *13*, 2653–2667.
- [2] M. Hammarlund-Udenaes, M. Fridén, S. Syvänen, A. Gupta, *Pharm. Res.* **2008**, *25*, 1737.
- [3] X. Liu, C. Chen, B. J. Smith, *J. Pharmacol. Exp. Ther.* **2008**, *325*, 349–356.
- [4] A. R. Mehdi pour, M. Hamidi, *Drug Discovery Today* **2009**, *14*, 1030–1036.
- [5] R. Cecchelli, V. Berezowski, S. Lundquist, *Nat. Rev. Drug Discovery* **2007**, *6*, 650–661.
- [6] S. Aday, R. Cecchelli, D. Hallier-Vanuxeem, *Trends Biotechnol.* **2016**, *34*, 382–393.
- [7] H. Li, C. W. Yap, C. Y. Ung, *J. Chem. Inf. Model.* **2005**, *45*, 1376–1384.
- [8] S. Majumdar, S. C. Basak, C. N. Lungu, *SAR QSAR Environ. Res.* **2018**, *2018*, 579–590.
- [9] O. T. Devinyak, R. B. Lesyk, *Curr. Comput.-Aided Drug Des.* **2016**, *12*, 265–271.
- [10] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Mach. Learning* **2012**, *46*, 389–422.
- [11] C. N. Lungu, *Studia UBB Chemia* **2018**, *63*, 177–188.
- [12] C. Lungu, S. Ersali, B. Szefer, *Studia UBB Chemia* **2017**, *62*, 197–204.
- [13] C. N. Lungu, I. Bratanovici, G. Mirabela, *Curr. Med. Chem.* **2018**, in press.
- [14] M. V. Diudea, G. Katona, I. Lukovits, N. Trinajstić, *Croat. Chem. Acta* **1998**, *71*, 459–471.
- [15] *Small-Molecule Drug Discovery Suite*, **2009**, Schrödinger, LLC, New York, NY.
- [16] O. Ursu, M. V. Diudea, *TOPOCLUJ software program*, Cluj, Romania: Babes-Bolyai University, **2005**.
- [17] S. C. Basak, B. D. Gute, G. D. Grunwald, in *Topological Indices and Related Descriptors in QSAR and QSPR* (Eds: J. Devillers, A. T. Balaban), Amsterdam, The Netherlands, **1999**, pp. 675–696.
- [18] S. C. Basak, D. Mills, B. D. Gute, D. M. Hawkins, in *Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens* (Ed: R. Benigni), Boca Raton, FL, **2007**, pp. 215–242.
- [19] S. C. Basak, R. Natarajan, D. Mills, *J. Chem. Inf. Model.* **2006**, *46*, 65–77.
- [20] S. Majumdar, S. C. Basak, G. D. Grunwald, *Curr. Comput.-Aided Drug Des.* **2013**, *9*, 463–471.
- [21] S. C. Basak, D. K. Harriss, V. R. Magnuson, *POLLY v2.3*, Copyright of the University of Minnesota, **1988**.
- [22] S. Basak, G. Grunwald, A. Balaban, *TRIPLET*, Copyright of the Regents of the University of Minnesota, **1993**.
- [23] L. Breiman, *Mach. Learning* **2001**, *45*, 5–32.
- [24] V. E. Kuz'min, P. G. Polishchuk, A. G. Artemenko, S. A. Andronati, *Mol. Inf.* **2011**, *30*, 593–603.
- [25] P. G. Polishchuk, E. N. Muratov, A. G. Artemenko, *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.
- [26] V. Svetnik, A. Liaw, C. Tong, *J. Chem. Inf. Model.* **2008**, *43*, 1947–1958.
- [27] S. Kortagere, D. Chekmarev, W. J. Welsh, S. Ekins, *Pharm. Res.* **2008**, *25*, 1836–1845.
- [28] J. Friedman, *Ann. Statist.* **2001**, *29*, 1189–1232.
- [29] H. Liu, H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, New York, NY, **1998**.
- [30] S. C. Basak, V. R. Magnusson, G. J. Niemi, R. R. Regal, *Discrete Appl. Math.* **1988**, *19*, 17–44.
- [31] C. Yoo, M. Shahlaei, *Chem. Biol. Drug Des.* **2018**, *91*, 137–152.
- [32] D. Hawkins, S. Basak, D. Mills, *J. Chem. Inf. Comput. Sci.* **2003**, *3*, 579–586.
- [33] S. Majumdar, S. C. Basak, *Curr. Comput.-Aided Drug Des.* **2018**, *14*, 284–291.
- [34] A. M. Molinaro, R. Simon, R. M. Pfeiffer, *Bioinformatics* **2005**, *21*, 3301–307.
- [35] Q.-S. Xu, Y.-Z. Liang, *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1–11.
- [36] J. Shao, *J. Am. Stat. Assoc.*, **1993**, *88*, 486–494.
- [37] J. Tukey, in *The Handbook of Social Psychology*, 2 ed., Vol. 2 (Eds.: G. Lindzey, E. Aronson), Oxford, England, **1968**, pp. 147.
- [38] M. Stone, *J. R. Statist. Soc. B*, **1977**, *39*, 44–47.
- [39] D. Baumann, K. Baumann, *J. Cheminf.* **2014**, *6*, 1–19.
- [40] D. Hawkins, S. Basak, D. Mills, *Environ. Toxicol. Pharmacol.* **2004**, *16*, 37–44.
- [41] S. Majumdar, S. C. Basak, *Curr. Comput.-Aided Drug Des.* **2016**, *12*, 294–301.
- [42] P. Filzmoser, B. Liebmann, K. Varmuza, *J. Chemom.* **2009**, *23*, 160–171.
- [43] K. Roy, P. Ambure, *Chemom. Intell. Lab. Syst.* **2016**, *159*, 108–126.
- [44] R Core Team, *R: A Language and Environment for Statistical Computing version 3.1.1*, **2014**.
- [45] S. C. Basak, *Med. Sci. Res.* **1987**, *15*, 605–609.
- [46] A. T. Balaban, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 23–28.
- [47] F. Ooms, P. Weber, P. A. Carrupt, B. Testa, *Biochim. Biophys. Acta* **2002**, *1587*, 118–125.
- [48] L. Jäntschi, G. Katona, M. V. Diudea, *MATCH, Commun. Math. Comput. Chem.* **2000**, *41*, 473–488.
- [49] O. M. Minailiuc, G. Katona, M. V. Diudea, *Croat. Chem. Acta* **1998**, *71*, 473–488.
- [50] O. Ivanciuc, T. Ivanciuc, M. V. Diudea, *SAR QSAR Environ. Res.* **1997**, *7*, 63–87.



- [51] H. Pajouhesh, G. R. Lenz, *NeuroRx*. **2005**, 2, 541–553.
- [52] B. Hemmateenejad, R. Miri, M. A. Safarpour, A. R. Mehdipour, *J. Comput. Chem.* **2006**, 27, 1125–1135.
- [53] Z. Y. Algamil, M. H. Lee, A. M. Al-Fakih, M. Aziz, *J. Chemom.* **2015**, 29, 547–556.
- [54] M. A. Lill, *Drug Discovery Today* **2007**, 12, 1013–1017.
- [55] M. Iyer, R. Mishru, Y. Han, A. J. Hopfinger, *Pharm. Res.* **2002**, 19, 1611–1621.
- [56] *OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models*, available at: <https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>, accessed 11 March 2019.

Received: December 1, 2018

Accepted: April 11, 2019

Published online on ■■■, ■■■■

### Graphical Abstract

*S. Majumdar\*, S. C. Basak, C. N. Lungu,  
M. V. Diudea, G. D. Grunwald*

1 – 10

**Finding Needles in a Haystack: De-  
termining Key Molecular Descriptors  
Associated with the Blood-brain  
Barrier Entry of Chemical  
Compounds using Machine  
Learning**

