**Finding needles in a haystack: determining molecular descriptors associated with the blood-brain barrier entry of chemical compounds using machine learning**

S. Majumdar a, S.C. Basak\*b, C.N. Lungu, M.V. Diudea c and G.D. Grunwald d

aUniversity of Florida Informatics Institute, Gainesville, USA; bDepartment of Chemistry and Biochemistry,

University of Minnesota, Duluth MN, USA;

cDepartment of Chemistry, Babes-Bolyai University, Cluj-Napoca, Romania;

dNatural Resources Research Institute, University of Minnesota, Duluth, USA

**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 [1-7]. 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 are used to assess the BBB permeability of chemicals. One quantifier of BBB entry is BB, which is the ratio of the concentration of a chemical in the brain to its concentration in the blood. Li et al [7] developed computational models to predict BB of a set of 415 chemicals using molecular descriptors and a few statistical as well as machine learning methods. But these days we have the availability of expanded set of descriptors and more sophisticated modeling methods. Therefore, it was of interest to develop QSARs for the BBB entry of the 415-chemical set using molecular descriptors calculated by POLLY [ 8], APProbe, Triplet [9] and TopoCluj [10] , and Schrodinger [11] software and modeling methods, viz., RR, Machine learning (Subho to fill up)

**Method and material**

BB data: The data (Supplementary Table 1) on the penetration of the BBB is taken from Li et al. [7]. The data is binary, n indicating not penetrating and p indicates the chemical crosses the BBB. We use the Simplified Molecular Input Line Entry System (SMILES) notations of these compounds and feed them into several in-house software to compute hundreds of molecular descriptors (vide infra) and .

**Molecular Descriptors**

(To be modified)

For this study we use two collections of molecular descriptors. One set of descriptors, used frequently by the Cluj team of Diudea and collaborators, are calculated by the programs Schrodinger [11] and TopoCluj [10]. More detailed references about these descriptors are given in the Supplementary material. We include 579 such descriptors in our analysis.

The second set of molecular descriptors, used frequently by Basak et al, is calculated by the software POLLY [8] and Triplet [9]. We use 98 and 100 descriptors calculated by these software, respectively.

**Statistical and Machine learning methods**

In view of the stated external validation bias of the journal, we must say very explicitly what we are doing here and why along with two deep cross validation ideas and our refs on this topic.

(Subho: I was wondering whether in this section you would like to have some comments on external validation based on our CCADD paper and also the attached papers one of which cites our paper with Hawkins)

*Predictive models*

We use two statistical/machine-learning methods to build our QSAR models.

Random Forest (RF): This method trains multiple decision trees on a dataset, each based on a randomly selected subset of total features. The final prediction in a regression problem is taken as the average of individual predictions from all the trees, while in classification problem the final class prediction is done by majority voting. Previous examples of the use of RF models in QSAR include [**33-35].** Variable importance in RF models is calculated by the total amount of Gini impurity reduction caused by trees containing a specific variable (https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/). we use this as our measure of variable importance in the results section.

Lasso: to be filled up.

*Validation technique*

*Metrics*

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

1. *Area Under Curve (AUC)*: This is defined as the area covered under a Receiver Operating Characteric (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 indicate a better predictive model.
2. *Top 20% lift*: Suppose the response has two levels in a classification problem: 0 and 1, and the model outputs predicted probabilities of a sample belonging to class 1. Then the Top 20% lift denotes the ratio of class 1 samples which are in the samples in the top 20% predicted probabilities, divided by the ratio of class 1 samples in the overall population. For a bad model, this ratio will be closer to 1, since that means the top 20% of prediction scores cover about the same proportion of class 1 as a random guess from the full population. Whereas in a good model most of class 1 samples will have a high score, and thus the top 20% scores are more likely to cover those samples.

**Results**

The results obtained from our analysis 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 methods. After this we use subsets of the top predictors for each method (like top 10%, 20%), 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 [43] to do all of our data analyses.

*Top descriptors in QSAR models*

Tables 1-3 list the top 10 descriptors in the full models built on the Basak, Diudea and combined set of descriptors, respectively.

|  |  |  |  |
| --- | --- | --- | --- |
| *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 |
| DN2N3 | 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 |

(Please explain the significance of these findings)

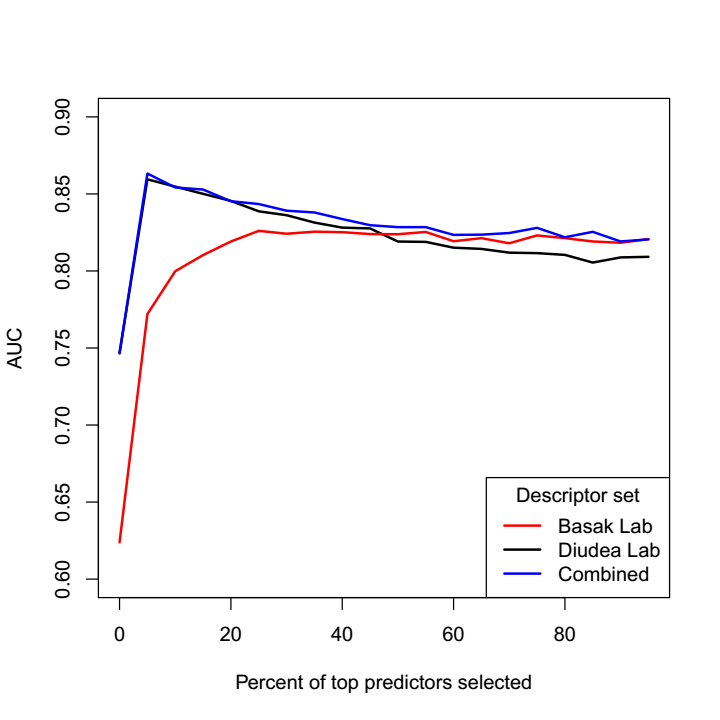
*Validation*

We use a repeated external validation, also known as monte-carlo cross validation in the literature, to evaluate the predictive performance of our QSAR models. To this end, we calculate each of the two metrics: AUC and top 20% lift 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 4 and Figures 1-3 (2 columns of Table 4 and Figs. 2-3 to be added).

|  |  |  |
| --- | --- | --- |
| Descriptor set | AUC | Top 20% lift |
| Basak lab | 0.813 |  |
| Diudea lab | 0.808 |  |
| Combined | 0.82 |  |

*Table 4: Full model metrics for the three descriptor sets*



*Figure 1: Area Under Curves for reduced models with top θ% important descriptors*

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. On the other hand, models corresponding the Basak set of descriptors perform more or less similarly for values of *θ* larger than 25. The combined set of descriptors constantly perform better than both the individual descriptor sets for all values of *θ*. This improvement is slightly better for all values of *θ* when the Diudea set is considered, and for *θ* > 25 when the Basak set is compared with.

**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. 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 AUC analyses in Table 4 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 may explain such results.

If we look at the most influential 10 descriptors (Table 1-3), the indices selected from the 198 POLLY and Triplet [9] descriptor set are the information theoretic [ Dis ref # 1] indices and the triplet descriptors. The former group quantify molecular neighborhood complexity while the triplet indices characterize the heterogeneity of atoms in the molecular graph. It is noteworthy that principal component analysis (PCA) for a diverse collection of 3,692 chemicals showed that the information theoretic indices are strongly correlated with a distinct PC indicating that such indices quantify some aspects of molecular structure not represented by other indices [ Dis Ref # 2)

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 [Dis ref # 3]. 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.

In conclusion, the two sets of descriptors of 198 and 579 calculated for the set of 415 BBB data led to the development of goof quality predictive models. The individual sets were independently as good as one another in the formulation of QSARs for the BBD data. Further studies with other data sets are needed to understand the utility of these mathematical molecular descriptors in assessing the blood-brain entry of chemicals.

.

**References**

[1] T. Houa, J. Wangb, W. Zhangc, W. Wang, Xiaojie Xu, Current Medicinal Chemistry, 2006, 13, 2653-2667

[2] M. Hammarlund-Udenaes, M. Fridén, S. Syvänen, A. Gupta, Pharm Res., 2008, 25: 1737. https://doi.org/10.1007/s11095-007-9502-2

[3] X. Liu, C. Chen, B. J. Smith, THE JOURNAL OF PHARMACOLOGY AND EXPERIMENTAL THERAPEUTICS, 2008, 325, JPET 325:349–356, 2008

[4] A. R. Mehdipour, M. Hamidi, , Drug Discovery Today Volume 14, Numbers 21/22 November 2009

[5] R. Cecchelli, V. Berezowski, S. Lundquist‡ M. Culot, M. Renftel‡, M. Dehouck, L. Fenart,

NATURE REVIEWS | DRUG DISCOVERY VOLUME 6 | AUGUST 2007 | 651

[6] S. Aday, R. Cecchelli, D. Hallier-Vanuxeem, M.P. Dehouck, L. Ferreira1, Trends in Biotechnology, 2016, Vol. 34, No. 5 , <http://dx.doi.org/10.1016/j.tibtech.2016.01.001>

[7] H. Li, C. W. Yap, C. Y. Ung, Y. Xue, Z. W. Cao, Y. Z. Chen, J. Chem. Inf. Model. 2005, 45, 1376-1384

[8] S.C. Basak, D.K. Harriss, and V.R. Magnuson, POLLY 2.3, University of Minnesota, MN, 1988.

[9] S. Basak, G. Grunwald, and A. Balaban, TRIPLET, University of Minnesota, MN, 1993.

[10] O. Ursu and M.V. Diudea, TOPOCLUJ, Babes-Bolyai University, Cluj, Romania, 2005.

[11] Small-Molecule Drug Discovery Suite 2009, Schrödinger LLC, New York, NY, 2009.

[Dis Ref # 1] Basak, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. Med. Sci. Res., 1987, 15, 605-609.

[Dis ref # 2] Basak, S.C.; Magnuson, V.R.; Niemi, G.J.; Regal, R.R. Determining structural similarity of chemicals using graph theoretic indices. Discrete Appl. Math., 1988, 19, 17-44.

[Dis ref # 3) Iyer, M.; Mishru, R.; Han, Y.; Hopfinger, A. J. Predicting bloodbrain barrier partitioning of organic molecules using membraneinteraction QSAR analysis. Pharm. Res. 2002, 19 (11), 1611-1621