Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201400118

# Towards Better BBB Passage Prediction Using an Extensive and Curated Data Set

Yoan Brito-Sánchez,[a, b] Yovani Marrero-Ponce,*[b, c, d] Stephen J. Barigye,[b, e] Iván Yaber-Goenaga,[c] Carlos Morell Pérez,[f] Huong Le-Thi-Thu,[g] and Artem Cherkasov[a]

**Abstract**: In the present report, the challenging task of drug delivery across the blood-brain barrier (BBB) is addressed via a computational approach. The BBB passage was modeled using classification and regression schemes on a novel extensive and curated data set (the largest to the best of our knowledge) in terms of log *BB*. Prior to the model development, steps of data analysis that comprise chemical data curation, structural, cutoff and cluster analysis (CA) were conducted. Linear Discriminant Analysis (LDA) and Multiple Linear Regression (MLR) were used to fit classification and correlation functions. The best LDA-based model showed overall accuracies over 85% and 83% for the training and test sets, respectively. Also a MLR-based model with acceptable explanation of more than 69% of the variance in the experimental log *BB* was developed. A brief and general interpretation of proposed models allowed the estimation on how 'near' our computational approach is to the factors that determine the passage of molecules through the BBB. In a final effort some popular and powerful Machine Learning methods were considered. Comparable or similar performance was observed respect to the simpler linear techniques. Most of the compounds with anomalous behavior were put aside into a set denoted as controversial set and discussion regarding to these compounds is provided. Finally, our results were compared with methodologies previously reported in the literature showing comparable to better results. The results could represent useful tools available and reproducible by all scientific community in the early stages of neuropharmaceutical drug discovery/development projects.

**Keywords:** Linear discriminant analysis · Multiple linear regression · P-glycoprotein · Quantitative structure pharmacokinetic (property) relationship · Blood–brain barrier · BBB endpoint · Dragon descriptor

## 1 Introduction

In early stages of drug development, knowledge on the ability of a compound to penetrate the blood–brain barrier (BBB) plays a key role.[1] The BBB is a complex physical and biochemical interface consisting of endothelial cells of the brain capillaries.[2–3] Its purpose is to maintain the homeostasis of the central nervous system (CNS) by separating the brain from the blood stream[4] and represents a major challenge to the treatment of most brain disorders.[5–6] The level of BBB penetration must be known not only for drugs targeting the CNS, but also in those ones in which low penetration is desirable to minimize the undesired CNS side effects.[7]

[a] *Y. Brito-Sánchez, A. Cherkasov*
*Vancouver Prostate Centre, University of British Columbia*
*Vancouver, British Columbia, V6H 3Z6, Canada*

[b] *Y. Brito-Sánchez, Y. Marrero-Ponce, S. J. Barigye*
*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research, International Network (CAMD-BIR International Network), Los Laureles L76MD, Nuevo Bosque, 130015, Cartagena de Indias, Bolivar, Colombia.*
**e-mail: ymarrero77@yahoo.es*
*Homepage: Homepage: http://www.uv.es/yoma/*
*Homepage: http://sites.google.com/site/ymponce/home*

[c] *Y. Marrero-Ponce, I. Yaber-Goenaga*
*Grupo de Investigación en Estudios Químicos y Biológicos, Facultad de Ciencias Básicas, Universidad Tecnológica de Bolívar*
*Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 vía Turbaco, 130010, Cartagena de Indias, Bolívar, Colombia*

[d] *Y. Marrero-Ponce*
*Facultad de Química Farmacéutica, Universidad de Cartagena*
*Cartagena de Indias, Bolívar, Colombia*

[e] *S. J. Barigye*
*Department of Chemistry, Federal University of Lavras*
*P.O. Box 3037, 37200-000, Lavras, MG, Brazil*

[f] *C. Morell Pérez*
*Center of Studies on Informatics, Universidad "Marta Abreu" de Las Villas*
*Santa Clara, 54830, Villa Clara, Cuba*

[g] *H. Le-Thi-Thu*
*School of Medicine and Pharmacy, Vietnam National University*
*Hanoi (VNU) 144 Xuan Thuy, CauGiay, Hanoi, Vietnam*

*Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201400118.*

Brain penetration is commonly assessed by two experimental approaches, namely equilibrium distribution between brain and blood and BBB permeability.[8] The former determines the total extent of brain distribution (quantified as log *BB*)[9] and despite all its limitations as a sole indicator of brain exposure,[10] is the most commonly used.[7,11–12] The latter is often expressed as permeability-surface area product (quantified as log *PS*).[13] Lately, another quantitatively meaningful measurement of brain exposure, expressed as steady-state unbound brain-to-plasma concentration ratio ($K_{p,uu,brain}$) have been proposed.[14] This parameter can be more likely linked to the compounds CNS activity because it give indications of free, unbound drug, that is responsible for the pharmacological effect. Alternatively the log*BB* essentially represents the inert partitioning into brain lipid matter. However, although log *PS* and $K_{p,uu,brain}$ has been accepted as important parameters in drug discovery, the scarcity of publically available data has limited their viability in modeling studies of BBB penetration.[9,15–16]

A poor pharmacokinetics profile, has been recognized as one of the leading causes of failure of a drug candidate in late-stage discovery/development,[17] with a recent shift in the thinking toward toxicity and efficacy as the major causes of attrition. Thus acquiring valid information on molecules' BBB permeation, toxicity and efficacy in the early stages of drug discovery is a subject of great scientific and economic value. In this sense, in silico prediction methods have gained popularity as they are cheaper and less time consuming,[7,18] allowing the screening of candidates BBB profile, even before synthesizing the molecule and resources invested in testing.[19] However, modeling BBB passage is a challenging task in drug design.

On one hand, finding quality (following a uniform standard protocol for experimental determination of the brain/plasma ratio) and quantity log *BB* data is very difficult. On the other hand, there are other factors like passive diffusion characteristics, active efflux and influx transporters, metabolism and relative drug binding affinity differences between the plasma proteins and brain tissue that may influence the log *BB* value.[9–10] Hence establishing a useful relationship between the molecular structure and the measured blood brain partitioning is a really difficult task.[7] Another important issue of data quality that inherently affects the performance of models is the step of chemical data curation and preparation prior to model development and validation.[20–21] Unfortunately, although there are compelling reasons to believe that chemical data curation should be given a lot of attention, it is also obvious that for the most part the basic steps to curate a dataset of compounds have been either considered trivial or ignored.[22]

Despite all the limiting factors, many efforts have been devoted into in silico models for BBB passage prediction using different sets of descriptors and modeling techniques.[23–26] Nevertheless, some of them share the same major drawbacks – small number of compounds are used to train the models and lacking external validation to prove their real predictive power.[27–28] As a consequence, it has been shown that these models are not suitable for high-throughput screening (HTS) of new chemical entities as they do not generalize outside the chemical space used to set up the models.[11] The largest publicly available data set of log *BB* values, which contains 362 compounds has been recently published,[12] but most data sets that have been used to build models for BBB penetration so far are much smaller.[23,29–33]

In the recent years, a frequent problem is that although a number of models reported in the literature give reasonably good performance on BBB passage prediction, details like, chemical structures in any chemical format, properties, descriptors used to encoded chemical information or software used at each stage of the workflow are often not available.[34–36] Consequently, these models cannot easily be tested or extended, and adherence to OECD principles remains unclear.[37] Altogether, these findings clearly show that there is still need for further research on BBB passage prediction.

Bearing in mind all mentioned above and in order to overcome the actual unsatisfactory situation, the present manuscript tackles five main objectives: 1) compiling the largest (to our knowledge) dataset with quantitatively measured log *BB* using data from all previous publications, 2) performing steps of chemical data curation, brief property and structural characterization, threshold and cluster analysis, 3) attempting to evaluate the performance of Dragon descriptors on their ability to be used to classify the compounds into BBB + and BBB − based on a threshold value and further to predict log *BB* values, using Linear Discriminant Analysis (LDA), Multiple Linear Regression (MLR), and other nonlinear machine learning techniques, respectively, 4) performing a consistent comparison between our models and those previously reported in the literature, and 5) describing all the workflow in a transparent manner that the report results could be easily reproduced, tested or extended by other researchers.

## 2 Materials and Methods

### 2.1 Data Compilation and Chemical Curation

After an extensive literature search, we have compiled the largest (to our knowledge) dataset with quantitatively measured log *BB,* in which some compounds were subjected to the QSAR study for the first time. The log *BB* is defined as the ratio of the steady-state total concentration of a compound in the brain to that in the blood.[9] It can be experimentally determined either by in vivo or in vitro methods. The in vivo methods involve the measurement of drug concentrations in brain and blood and provide the most reliable reference information for testing and validating other models.[9] The in vitro determinations have been used over the years to estimate in vivo BBB penetration as accurately as possible. They comprise cell based systems like *Madin-*

  

*Darby Canine Kidney* (*MDCK*), cell line or non-cell based systems e.g., *Parallel Artificial Permeability Assay* (*PAMPA*) and several reviews have summarized the state of the art of these systems.[9,38–39] Quantitative log *BB* values were collected from original experimental articles and earlier modeling works, the latter being rechecked from the original sources wherever possible. For the vast majority of compounds, the log *BB* values have been measured in vivo, for the most part in rats, but the dataset also includes 58 organic volatile compounds for which the log *BB* values have been determined in vitro.[33] We therefore combine all sets of distribution ratios, but do not average them. The final log *BB* values were selected on the basis of their uniformity with respect to experimental determinations.

Initially, the molecules were drawn and saved as MDL mol files using the ChemDraw software package.[40] The hydrogen atoms were added to the structures using Open Babel software.[41] Chemical data set curation was performed on the original data set. The initial step comprise tools available for dataset curation included in ChemAxon[42] and TOMOCOMD-CARDD (QuBiLS Suite).[43] The most important steps included the removal of inorganic and organo-metallic compounds, mixtures and curation of tautomeric forms. Also organic salts (salts with $Na^+$, $K^+$, $Ca^{2+}$) were converted to their corresponding neutral forms, and only one compound was retained in case of isomerism (any pair of enantiomers or diastereoisomers were recognized as duplicates). Additionally, at the end of the process manual data set curation was performed on the *original data set* as well. At this step each structure was visualized and manually inspected to detect structures that for some reasons escaped the automatic curation steps described above.

### 2.2 Dragon Descriptors Computation

Molecular descriptors (MDs) were calculated using the Dragon software.[44] This software computes descriptors based on 2D or 3D molecular structures and have been successfully applied for BBB passage prediction[32,45] and other research areas.[46–49] The MD were generated from the 2D structures in the appropriate .mol hydrogen added input format. The calculation procedures for these MDs are reported in reference.[50] The calculated MDs were filtered to exclude those ones with zero variance and low occurrence (MDs represented by less than 24% of compounds). Also, MDs with correlation coefficient ($x/x$) of 1.0 were eliminated. They were tested, on their quality of being able to classify the compounds into BBB+ and BBB− based on a threshold value and further to quantitatively predict the measured log *BB* values.

### 2.3 Statistical Analysis: Data Processing and Modeling

#### 2.3.1 Data Set Splitting

Clustering algorithms (CAs) are simple and useful data mining tools to explore relationships that exist among objects (or variables) and allocate to the same classes the similar ones, on the basis of predefined similarity (or dissimilarity) measures.[51–52] First *k*-nearest neighbors cluster analysis (*k*-NNCA), also known as hierarchical agglomerative clustering, was performed by using Complete Linkage and the Euclidean distance as amalgamation rule and proximity function, respectively, to have preliminary insight on the "possible" number of clusters that naturally exist in the examined data, to be later used in the *k*-Means Cluster Analysis (*k*-MCAs).

To evaluate the statistical quality of data partitions in the clusters a standard analysis of variance (ANOVA) for each dimension (variable) was performed. The values of the standard deviation (SS) between and within clusters, of the respective Fisher's ratio and their *p* level of significance, were examined.[53–54] The *training/prediction set* (*TS/PS*) splitting is based on the *k*-MCAs for each class (BBB+ or BBB−) and from each cluster of compounds approximately 20% (~ 20%) for the PS is randomly selected. Statistical analysis was carried out with STATISTICA package.[55]

#### 2.3.2 Qualitative Approach Using LDA

To obtain the binary predictions with QSAR models developed using real log *BB* values for the *modeling set*, we followed the criterion that compounds with experimental log $BB < 0$ were classified as relatively poor penetrators of the BBB (i.e., BBB−), while compounds with log $BB \geq 0$ were classified as relatively good penetrators of the BBB (i.e., BBB+). The dependent variable was then assigned a value of 1 or −1 when the compounds had log *BB* greater than or lower than the threshold, respectively. Statistical analysis was carried out with STATISTICA package.[55] The LDA was used to find the classifier functions.[56] The *forward stepwise* and *best subset* methods were employed for the attribute selection. The tolerance parameter was set to 0.01. By using the models, one compound can be classified as either active, if $\Delta P\% > 0$, being $\Delta P\% = [P\ (\text{Active}) - P\ (\text{Inactive})] \times 100$, or inactive otherwise. *P* (active) and *P* (inactive) are the probabilities with which the equations classify a compound as active and inactive, respectively. The quality of the models was determined according to Wilks' $\lambda$, the square of the Mahalanobis distance $D^2$, Fisher ratio (*F*), significance level (*p*) and the percentage of good classification (accuracy, *Q*). Therefore, parameters like sensitivity 'hit rate' (*SE*), specificity (*SP*), false positive rate ($fp_{rate}$) (also called false alarm rate) and Matthews' correlation coefficient (MCC) were taken into account.[57] Also the principle of parsimony (Occam's razor) was considered, in that models with high statistical significance but having as few parameters as possible were preferred. However, the main criterion

to select the best model is based on the prediction statistics for a PS that were never used in the process of model development.[22]

#### 2.3.3 Quantitative Approach Using MLR

In this study, one of our aims is to evaluate the predictive capacity of the DRAGON indices of log *BB* of the *modeling set*. In this report, we use MLR analysis coupled with the Genetic Algorithm (MLR-GA), using MobyDigs software.[58] This method is a variable selection strategy which imitates the "survival for the fittest" principle in the search for models that best explain a determined response variable.[59] Each chromosome is an *n*-dimensional binary vector in which each gene (position) is made to correspond to a variable, assigned 1 if present in the model and 0 otherwise. From an initial population of chromosomes (models), new ones are generated according a defined optimization function of fitness and using operations typical of the natural selection process such as: mutation, crossing-over, reproduction and tabu. The key benefit of the GA is the reduction in time required to arrive at an optimal solution.[60] As can be noted, computations with Dragon software yield high MDs dimensional space, justifying the need for data reduction. Accordingly, tabu list was used as preliminary screening of the original values to exclude variables with high correlation coefficients (x/x). The MDs with zero variance were also eliminated. The population size was set at 100 and the reproduction/mutation trade-off (T) at 0.70.

For each family, the best ten, nine and eight variable models for log *BB* were constructed, using as optimization function the statistical parameter $Q_{loo}^2$ ("leave one out" cross-validation). Later, the best variables, for each family, were grouped together into a single set and ten, nine and eight variable models, developed. The model performance was evaluated by the following statistical parameters: the coefficient of determination ($R^2$), the adjusted ($R^2$), the standard deviation (s), and Fisher-ratio's *p*-level (*p*(F)). From the population of generated models, the "best" 10 in each case were retained for validation using the techniques "bootstrapping" ($Q^2_{boot}$) and "scrambling" (a($R^2$), a($Q^2$)). In addition the standard error of cross validation (*SECV*) was taken into account. Thus, using a multi-criteria perspective only those models that pass both internal and external statistics filters were retained for the final selection. In this step, the prediction statistics for the test set were the leading criteria at time of the final decision.

#### 2.3.4 Applicability Domain Analysis

The applicability domain (AD) of a QSPR model must be defined if the model is to be used for screening new compounds. In this report, the William plot was used to verify the AD. This plot reveals the leverage values versus standardized residual and permit the graphical detection of both the response outliers (Y outliers) and the structurally influential compounds (X outliers).

#### 2.3.5 Non-Linear Machine Learning Methods

Additionally in the present report more rigorous non-linear classification and regression methods have been considered. Four algorithms were applied: Logistic regression (LR)[61] and support vector machines using Gaussian Kernel (C-SVM)[62–63] were explored for classification purposes while epsilon- Support Vector Regression (ε-SVR)[64] and Gaussian process (GP)[65] were considered for regression. Their behavior in the prediction of BBB passage is reported. The models were developed using Waikato Environment for Knowledge Analysis (WEKA), version 3.6[66]

## 3 Results and Discussion

### 3.1 Data Analysis

To date many efforts have been devoted into computational approaches to answer the question of rapidly and effective methods for drug delivery across the BBB.[9,23] However, the scarcity of publicly available data without giving serious attention to the importance of chemical data curation inherently affects the quality of models.[22] In an effort to improve the quality of the *original data set* detailed steps of automatic and manual data set curation were conducted in the present report. After finishing all steps of data set preparation the *curated dataset* was denoted as BM581 (denoting the number of compounds utilized throughout this study) and is provided in the Excel format in Table S1 of the Supporting Information (SI), along with chemical formulas in smiles code format, log *BB* values and references. By far to our knowledge, this is the largest set in terms of log *BB* values reported so far. Therefore BM581 can be a useful tool for the scientific community or during early stages of neuropharmaceutical drug discovery projects.

#### 3.1.2 Threshold Analysis

To know if a compound will be able to cross the BBB or not is a subject of great interested in neuropharmaceutical research. However, establishing the threshold value at which a compound is defined as a good or poor penetrator (BBB+ or BBB−) of the BBB is a controversial theme,[9] because it is generally hard to assign a standard threshold value usable in all cases. In this report, in an effort to overcome this barrier, the effect of choosing this point at different values was studied. Statistical parameters like the 'hit rate' and $fp_{rate}$ were check for each classification model.[57] Also dataset balancing was taking into account,[22] trying to select the cut-off value that provide a well-balanced dataset, the lowest $fp_{rate}$, but without discarding the balance between sensitivity and specificity. Accordingly and following this multi-criteria workflow, in our case the best cut off was

**Table 1.** Main results for the analysis of threshold value.

| Cut-Off | BBB+ [a] | BBB− [a] | $Q_T$ [b] | $fp_{rate}$ [a] | $Se$ [b] |
|---|---|---|---|---|---|
| −0.40 | 73.00 | 27.00 | 78.23 | 23.87 | 78.98 |
| −0.30 | 70.00 | 30.00 | 76.19 | 24.12 | 76.32 |
| −0.20 | 65.00 | 35.00 | 76.36 | 24.00 | 76.55 |
| −0.10 | 60.00 | 40.00 | 79.25 | 21.67 | 79.89 |
| **0.00** | **51.00** | **49.00** | **80.00** | **20.64** | **80.46** |
| 0.10 | 45.00 | 55.00 | 78.23 | 22.01 | 78.52 |
| 0.20 | 41.00 | 59.00 | 76.87 | 22.83 | 76.45 |
| 0.30 | 36.00 | 64.00 | 75.34 | 23.20 | 72.77 |
| 0.40 | 29.00 | 61.00 | 75.00 | 22.30 | 68.42 |

[a] Percentage of compound by each class. [b] All values are expressed as percentage (%).

0.00. Interestingly this point is one of the most widely employed in the literature in the field of BBB passage prediction.[9] The main results at this stage are shown in Table 1, details in Table S2 of the Supporting Information.

### 3.1.3 Data Set Characterization

BBB penetration is mandatory for CNS drugs, while must be restricted for many of the non-CNS drugs to avoid undesirable side-effects so a clear understanding of structural differences between good and poor penetrators of the BBB may assist both research areas. Many properties directly related to the molecular structure were computed with Dragon software and the distribution of various types of them in both series (BBB+ and BBB−) is described below. Here, all the properties were within the 95% percentile property range.

*Atom Count*. Figure 1 illustrates the distribution of all atoms, non-including hydrogens (nSK). The major difference was in the slope of the curves and the locations of the maxima. The distribution indicated that a total of 5–20 and 20–25 non-hydrogen atoms may be the best region for BBB+ and BBB− compounds, respectively. Figure 2 illustrates the distribution of nitrogen atoms. The distribution indicated that compounds that cross the BBB tend to have zero to two nitrogen atoms, while BBB− compounds vary between two and four nitrogen atoms reaching a maxima of six atoms. Finally, Figure 3 shows the distribution of the number of oxygen atoms. Clearly, zero to one oxygen atoms is the best range for compounds that cross the BBB. By contrast two to three oxygen atoms may restrict the passage of compounds through the BBB.

*H-Bond Acceptors and Donors*. Figure 4A) and 4B) show the distribution of hydrogen bond acceptors and donors, respectively, as calculated by Dragon. According to the mo-

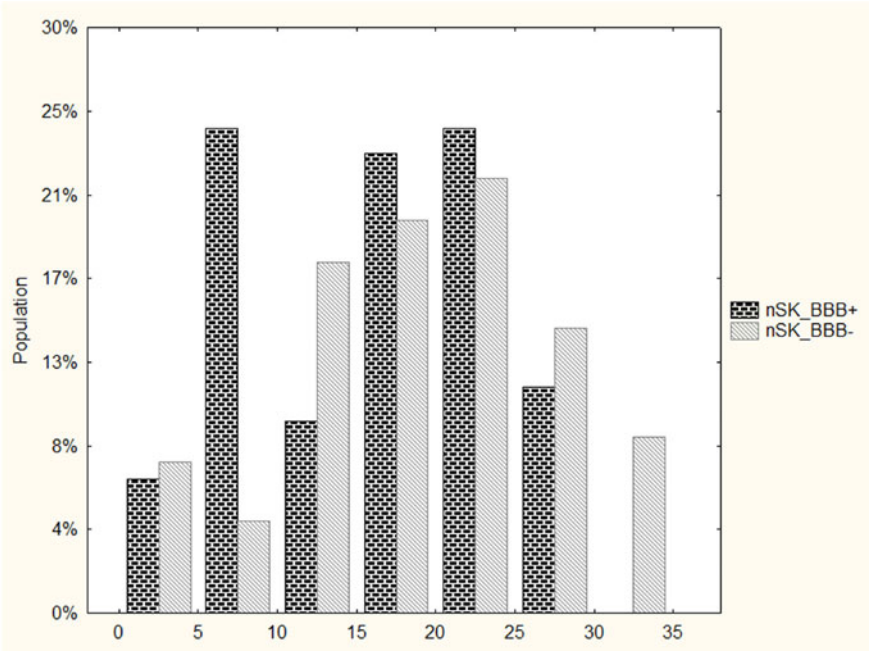


**Figure 1.** Distributions of the total number of atoms, non-including hydrogen atoms (nSK) in the BBB+ and BBB− sets.

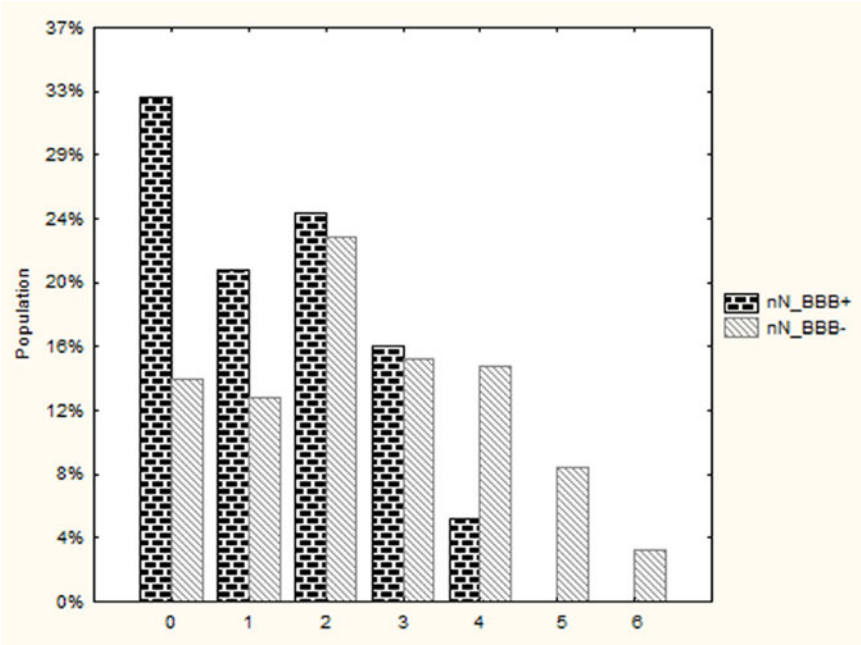


**Figure 2.** Distributions of the number of nitrogen atoms (nN) in the BBB+ and BBB− sets.

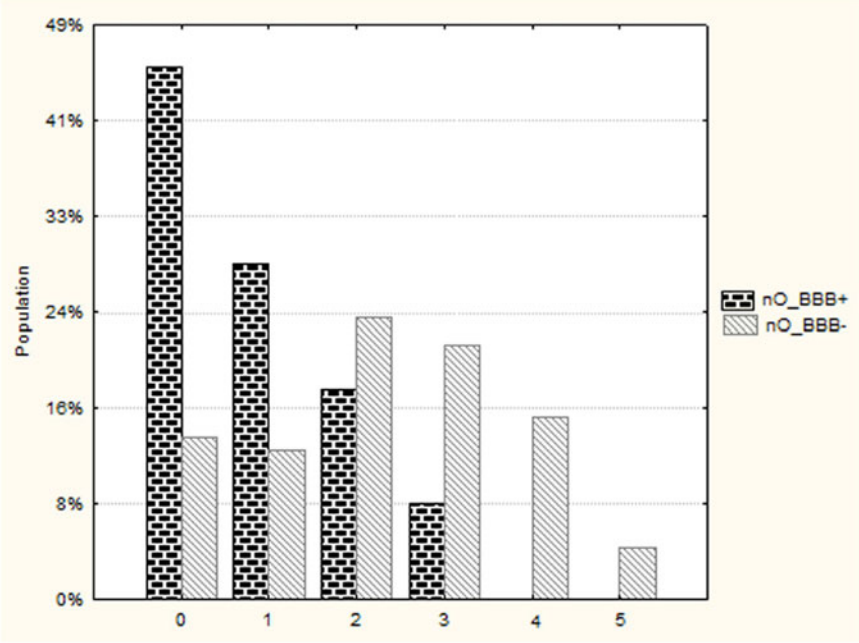


**Figure 3.** Distributions of the number of oxygen atoms (nO) in the BBB+ and BBB− sets.

lecular property calculator, the number of H-Bond Acceptors (nHAc) is the number of heteroatoms (oxygen, nitrogen,) with one or more lone pairs, excluding atoms with positive formal charges in heterocyclic rings or higher oxidation states. Similarly, the number of H-Bond Donors (nHDon) is the number of heteroatoms (oxygen, nitrogen) with one or more attached hydrogen atoms. The distribution differed in terms of not only the percentage of occurrence for different values but also the locations of the maximum. According to the molecular property calculator, the nHAc peak was at three for compounds that cross the BBB, while BBB− compounds showed the maximal population peak at five being almost equally populated. For nHDon, the best ranges are zero to one and one to two, for BBB+ and BBB− compounds, respectively.

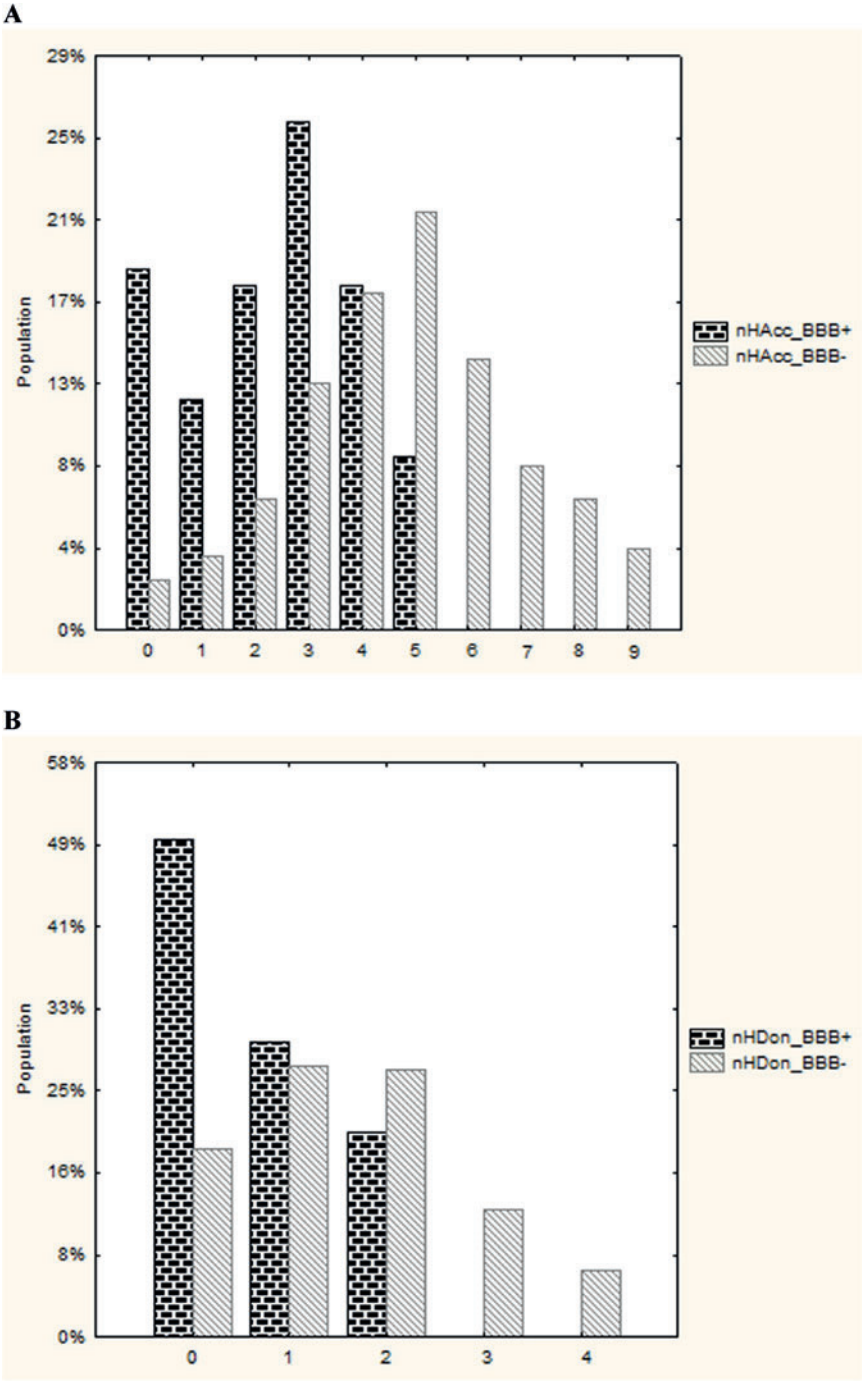


**Figure 4.** Distributions of the number of hydrogen bond acceptors (nHAc) (A) and the number of hydrogen bond donors (nHDon) (B) in the BBB+ and BBB− sets.

*Number of Aromatic Rings and Rotatable Bonds.* The distribution in counting the total number of aromatic rings (nBz) and rotatable bonds (nRB) was approximately identical for good and poor penetrators of the BBB (Figure 5 and 6, respectively). According to Figure 5, the number of aromatic rings in both series showed the maximum at two being the BBB+ set almost doubly populated. In the case of the number of rotatable bonds (Figure 6), the total number of them should not be more than six to facilitate the passage of compounds through the BBB and between two and forth for compounds with restrict access to pass the BBB.

*Molecular Weight.* Some properties directly related to molecular size are very useful during lead selection and lead optimization at early stages of drug discovery. Among them, molecular weight (MW) is commonly used. The distribution of MW in both series is shown in Figure 7. It indi-

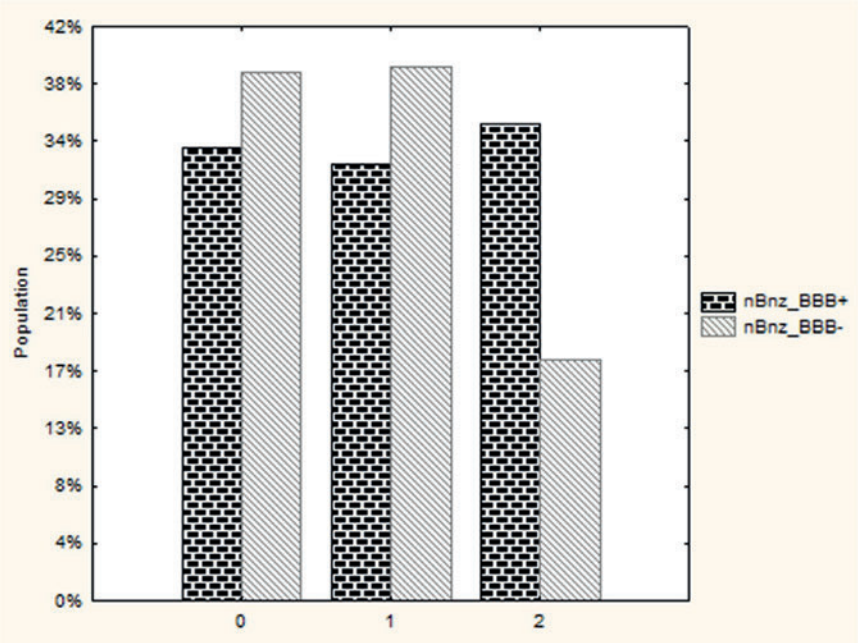


**Figure 5.** Number of aromatic rings in the BBB+ and BBB− sets.

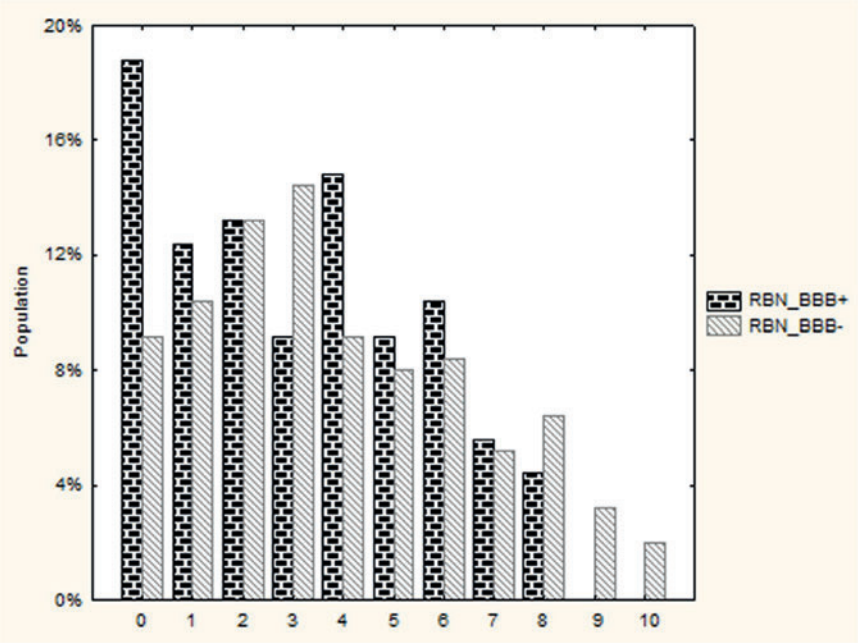


**Figure 6.** Number of rotatable bonds in the BBB+ and BBB− sets.

cates that the range of 250–300 was the best MW region for BBB+ compounds, though the maximal population peak for BBB− compounds is around 350.

*Topological Polar Surface Area.* The overall distributions of topological polar surface area using nitrogen and oxygen polar contributions (TPSA NO) differed not only in the location of the most populated bin, but also in the relative population of them. This property showed noticeable difference between BBB+ and BBB− sets (Figure 8). A small TPSA NO of 0–30 was the best range for BBB+ compounds, while values over 70 were preferential for BBB− compounds.

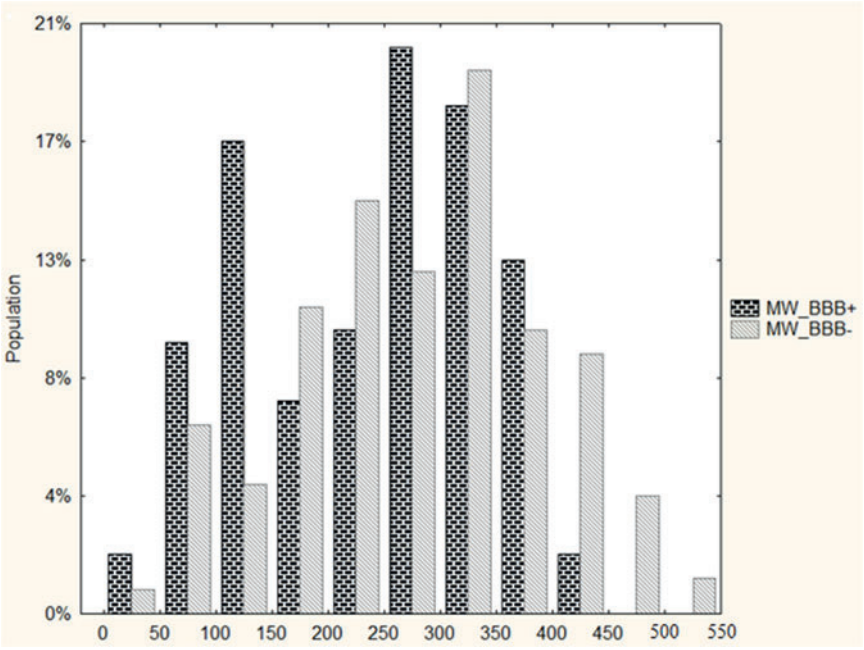


**Figure 7.** Distribution of molecular weight in the BBB+ and BBB− sets.

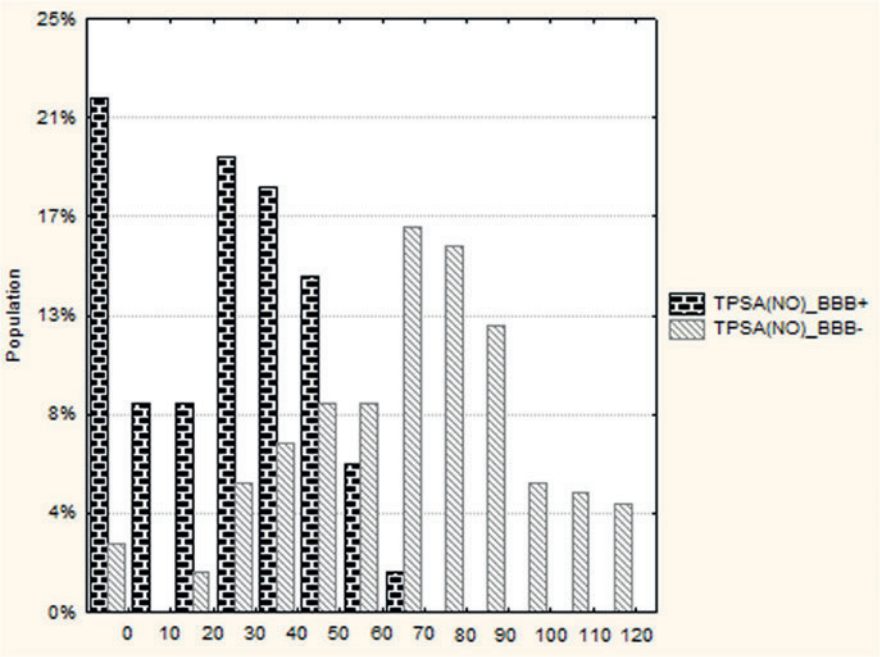


**Figure 8.** Distributions of topological polar surface areas in the non-CNS and CNS drugs.in the BBB+ and BBB− sets.

*Octanol-Water Partition Coefficient.* The distributions of log P values for BBB+ and BBB− compounds are shown in Figure 9. Log P distributions showed that the largest population for good penetrators of the BBB was from two to three good while 1.0 to 2.5 is the most populated range for poor penetrators of the BBB.

*Brief Conclusion of Multiple Properties Analysis.* For some of the properties studied before variation among their distributions between good and poor penetrators of the BBB can be noticed but any of them alone can discriminate very well between both series. TPSA NO was among the most discriminatory properties in differentiating BBB+ compounds from BBB− compounds while log P otherwise.

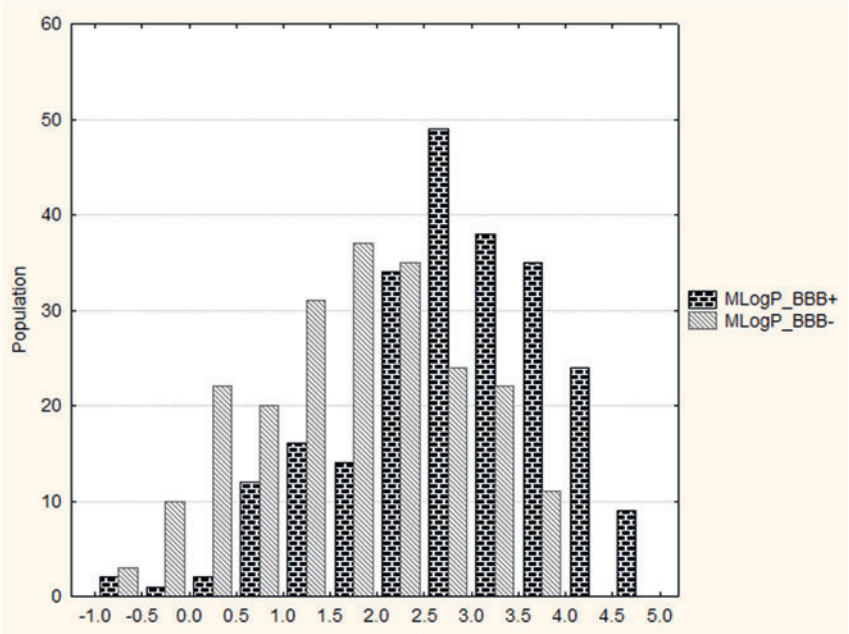


**Figure 9.** Distribution of Moriguchi Octanol-Water Partition Coefficient (*M*log*P*) values in the BBB+ and BBB− sets.

It suggest us the imperative need of employing modeling techniques based on a multivariable approach for discriminating between both series considering the complexity of the actual property (ability of compound to cross the BBB).

#### *3.1.4 Cluster Analysis*

In order to prove the structural diversity of the BM581 dataset (*curated dataset*), hierarchical agglomerative clustering was performed, for both BBB+ and BBB− series respectively.[53–54] As part of the *data fitting* process and before defining the modeling set several compounds with anomalous Euclidean distances with respect to the whole series (BBB− and BBB+) (the vast majority of them structurally extreme substances) were removed and are discussed in more detail in Section 3.4. The resulting dendrograms are depicted in Figure 10A) and B), using the Euclidean distance (*X*-axis) and the complete linkage (*Y*-axis). As can be seen, in both cases the dendrogram shows a clear and consistent tree structure. Also there are a great number of different structural patterns, which demonstrate the BM581 data set's molecular diversity. A cut-off of approximately 25% of maximum agglomerative distance was used as guide for the selection of an initial k value for performing *k*-MCAs. The main idea of *k*-MCAs consists in making a partition of either BBB+ or BBB− series into several statistically representative classes of compounds. Hence, this procedure allows a rational choice of compounds for the TS and PS considering the whole "experimental universe" of BM581.

A *k*-MCA was made first with BBB+ compounds and, afterwards, with BBB− ones. Several compounds were excluded from further analysis in the process of defining the optimum number of cluster. They were identified as singleton points (structural outliers), belonging to no cluster or forming clusters of five or less compounds. Also more reasons that could explain their anomalous behavior are given in Section 3.4. Finally, the first *k*-MCA (*k*-MCA I) partitioned the BBB+ set into 11 clusters and a second one (*k*-MCA II) split the BBB− set in 9 clusters. All variables that were used showed *p*-levels $<0.005$ for the Fisher test, more details about ANOVA results are depicted in the Supporting Information as Table S3. In both series, the selection of the TS and PS was performed by taking, in a random way, approximately 20% of compounds belonging to each cluster for the PS (details are in the Supporting Information Table S4 and Table S5). At the end of the process the *modeling set* (see Supporting Information Table S6) contains 497 unique compounds in which 381 of them form the TS and the remaining ones the PS. It very interesting to notice that for the BBB− all in vitro data belong to cluster seven while for the BBB+ over 72% correspond to cluster two. This result demonstrated that the performed cluster analysis was not only able to distinguish the optimum number of cluster based on chemical similarities but also captured biological trends in the proposed modeling set.

### 3.2 Qualitative Approach Using LDA

After performing a representative selection of *TS* and *PS*, *LDA* was used to fit discriminant functions that permit the classification of compounds as either BBB+ or BBB− using a cut-off value of 0.0 for the brain exposure classification.

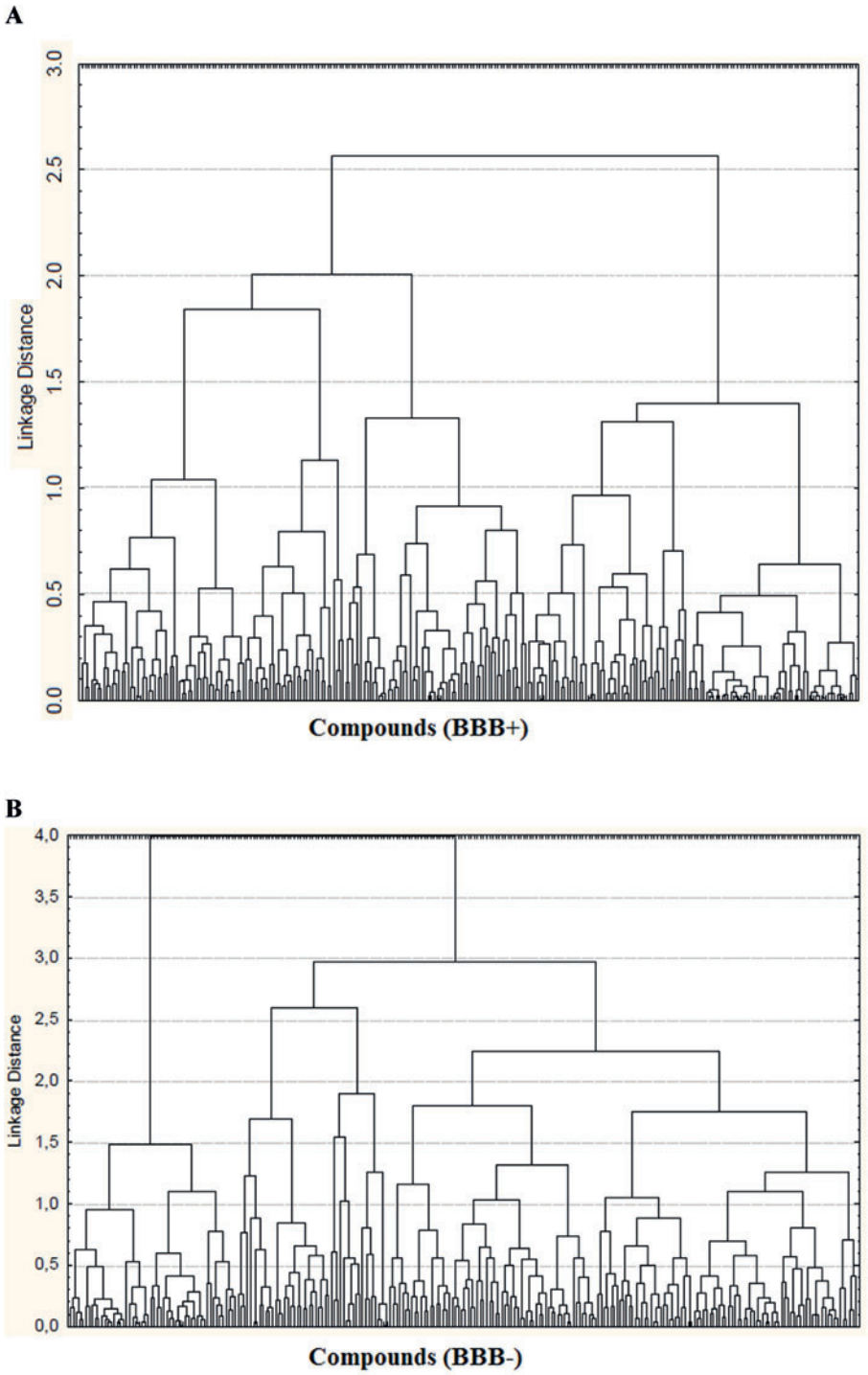


**Figure 10.** Dendograms for agglomerative hierarchical cluster analysis using the set of BBB+ and BBB−, A) *k*-NNCA I and B) *k*-NNCA II, respectively.

The *LDA* has become an important tool successfully applied in the field of BBB as well as others areas of drug design and property estimation.[67–69] In this sense, its application in the context of BBB passage prediction when it is not always necessary to predict an exact value, understand the probability that a compound will have passage to the brain or not can be very helpful.

During the process of fitting the best classification functions some compounds were identified as outliers and excluded before selecting the best model. Some examples

  

and details about possible reasons for their anomalous behavior can be found in Table 4. The best model employing six variables is given below together with its statistical parameters for the *TS*:

$$Class = 0.313 - 0.050 \times \mathrm{TPSA(NO)} - 0.568 \times \mathrm{BLTF96} + 0.588 \times \mathrm{C\text{-}008} - 1.470 \times \mathrm{nRCOOH} - 1.243 \times \mathrm{C\text{-}042} - 0.026 \times \mathrm{DELS} \quad (1)$$

$$N = 369 \;\; \lambda = 0.55 \;\; D^2 = 3.17 \;\; F(6,476) = 68{,}859 \;\; p < 0.001$$

In addition, for the *LDA*-based QSPkR model using the TS, we show in Table 2 most of the parameters commonly used to evaluate the performance of classification models. In the present report, we have selected overall accuracy ($Q_T$) and other statistical parameters like: MCC, SE, SP and $fp_{rate}$. However, for details about each method and others that are currently used, as well as the advantages and disadvantages of each approach see this reference.[57] The fitted Model 1 showed to be statistically significant at *p*-level $<0.0001$. Also, if we consider that the model has been trained using one of the largest sets reported so far and that prediction accuracies in the field of BBB are around 80%,[67,70] it becomes more clear the appropriateness and well balanced *Q* of 86.32% and 83.80% for BBB+ and BBB− compounds, respectively, in the TS. Additionally, for both BBB+ and BBB− compounds conforming the TS the posterior probabilities $\Delta P$% calculated from the Mahalanobis distance using Equation 1 are shown in Supporting Information as Table S7. Besides, in Figure 11 a plot of the $\Delta P$% (see Section 2.3.2) can be observed, based on classification obtained by Equation 1, for each compound in the TS.

However, although the statistical parameters for the *TS* provide some assessment of the goodness of fit of the model, the only way to prove its real predictive power is making predictions for a set of compounds that was never

**Table 2.** Prediction performances for linear and non-linear classification BBB–QSAR models.

| Sets | *N* | *MCC* | $Q_T$ [a] | $fp_{rate}$ [a] | $S_e$ [a] | $S_p$ [a] |
|---|---|---|---|---|---|---|
| ADL [b] | 381 | 0.70 | 85.09 | 16.20 | 86.32 | 84.97 |
| ADL [c] | 116 | 0.67 | 83.33 | 16.07 | 82.76 | 84.21 |
| LR [d] | 381 | 0.57 | 78.74 | 15.57 | 72.52 | 80.98 |
| LR [c] | 116 | 0.70 | 85.34 | 13.55 | 84.21 | 85.71 |
| C-SVM [d] | 381 | 0.57 | 78.74 | 17.08 | 74.17 | 79.88 |
| C-SVM [c] | 116 | 0.67 | 83.62 | 15.25 | 82.45 | 83.92 |

[a] All values are expressed as percentage (%). [b] Training set, [c] test Set, [d] 10-fold cross validation.

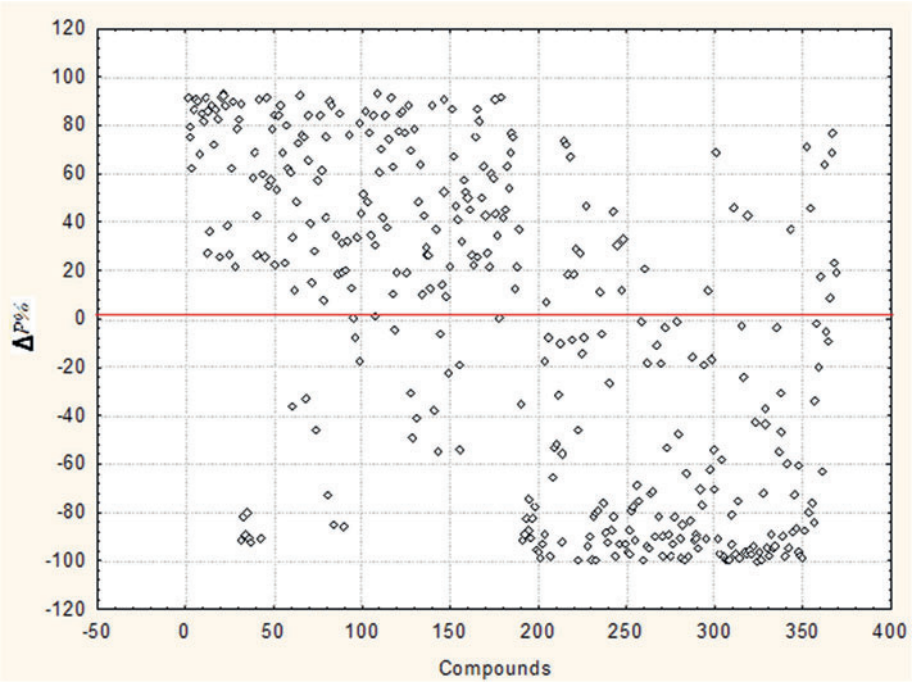


**Figure 11.** Plot of the predicted $\Delta P$% from Equation 1 for each compound in the training set. Compounds 1–190 are good penetrator (BBB+) of the BBB and chemicals 191–369 are poor penetrators (BBB−).

used in the process of defining classification functions.[22,71] Accordingly, Equation 1 was tested for its ability to predict the corresponding BBB class for a PS representative of whole "experimental universe" of BM581. As in the case of TS the accuracy and other statistical parameters were used to assess the performance and real predictive power our model in the PS, the results are depicted in Table 2. Also, for both BBB+ and BBB− sets conforming the PS the posterior probabilities $\Delta P\%$ using Equation **1** are shown in Supporting Information as Table S8. As can be seen Equation **1** correctly classified more than 83% of compounds in the PS. Therefore, since both training and prediction sets produced nearly identical statistical parameters the model is not likely to be over-fitted, which reinforces it the usefulness in the early stages of neuropharmaceutical drug discovery projects.

#### 3.2.1 Interpretation LDA-Based QSPkR Model

After choosing the final classification model an attempt into a brief and general interpretation in structural terms of the obtained model was made. The first MD under analysis was TPSA(NO) which is among molecular properties and is calculated according to the model proposed by Ertl.[72] It encodes information related to contributions from polar fragments containing nitrogen and oxygen to polar surface area.[50] This MD seems to have negative contribution to compounds brain exposure. This result is in agreement with other researchers who found that PSA with its related hydrogen-bond donor and acceptor ability descriptors were one of the most important factors determining the passage of molecules through the BBB.[7,17,73–74]

BLTF96, is calculated according to the model proposed by Verhaar based on the Moriguchi Log *P* (MLOGP) and has been observed in the literature that MDs related to Log *P* are directly linked to key properties like lipophilicity, determining the passage of molecules through cell membranes and also a very important property for their passage through the BBB.[1,17,45,73–75]

Some atom-centered fragments descriptors like C-008 (CHR2X) and C-042 (X− –CH··X) also take part in the model.[76] In the case of the former R represents any group linked through carbon and X represents any electronegative atom (O,N,S, halogens), while in the latter X represents any electronegative atom (O, N,S, halogens); − − an aromatic bond as in benzene or delocalized bonds such as the N–O bond in a nitro group; and "··" represents aromatic single bonds as the C–N bond in pyrrole. Note that previous studies have shown that some fragments descriptors were necessary for the modeling of the BBB passage.[19,30,32]

The effect of the addition of a functional group or its replacement by another on the ability of compounds to cross or not the BBB can be a promising alternative to carry out in the assessing of candidate drug molecules. The inclusion of nRCOOH (number of aliphatic carboxylic acids),[50] in the model maybe indicates that $-CO_2H$ group acts to prevent brain penetration. It would simply be due to the intrinsic hydrogen bonding and polarity properties of neutral acids increasing the ability of binding to albumin present in plasma and blood. This result is in agreement with previous experiments that achieved an improvement in the model when a variable indicator of the presence of carboxylic acid or MDs related to functional group counts were included.[19,73,77]

The DELS corresponds to molecular electrotopological variation and it is among the topological descriptors, encoding information related to both partial charges of atoms and their topological position relative to the whole molecule.[78] This *MD* is related to the relative availability of electrons accessible to intermolecular interactions. Hence compounds with high electronic density will increase binding affinity (e.g., to albumin present in plasma and blood), decreasing their ability to cross the BBB and thus the extent of brain exposure. It is worthwhile point out that previous derived models have proven that DELS is a significant descriptor for modeling drug transport properties such as human intestinal permeation.[79]

### 3.3 Quantitative Approach Using MLR

MLR-GA analysis was used to select the best subset of descriptors and to develop the linear models on the training set. It is noteworthy that some compounds were identified as outliers because of their particular structural features poorly represented in the training set, which could affect the variable selection for a better modeling of those compounds (*X* outliers) or the experimental uncertainties (Y outliers). Generally, if the residual value is larger than +/−3.0 s (where, s mean standard deviation of model), the sample can be considered as a response outlier (Y outlier), which could be associated with errors in the experimental values. Subsequently, all of them were removed before selecting the best model and some of them are listed in Table 3. Finally, the best model involved ten *MDs* and is shown below:

$$\begin{aligned}\log BB &= 0.1380\ (0.0683) - 0.0300\ (0.0048)\\ &\times \mathrm{T(N\cdot\cdot Br)} - 0.4467\ (0.0513) \times \mathrm{BELm3} - 0.7043\ (0.1023)\\ &\times \mathrm{nRCOOH} - 0.3809\ (0.0742) \times \mathrm{nOHp} + 0.1054\ (0.0367)\\ &\times \mathrm{nArX} + 0.1378\ (0.0241) \times \mathrm{nHDon} - 0.0516\ (0.0160)\\ &\times \mathrm{H\text{-}051} - 0.0280\ (0.0075) \times \mathrm{H\text{-}052} - 0.3002\ (0.0505)\\ &\times \mathrm{N\text{-}074} - 0.0167(0.0010) \times \mathrm{TPSA(NO)}\end{aligned} \quad (2)$$

$N = 357 \quad R^{2=}0.693 \quad Q_{loo}{}^2 = 0.670 \quad s = 0.373 \quad F = 78.06$

$p < 0.001$

The high F ratio of 78.06 indicates that Equation 2 does good prediction of the log *BB* values. Equation 2 has an adjusted $R^2$ value of 0.684, which indicates very good agree-

**Table 3.** Statistical parameters for the best MLR-based BBB–QSAR model in the training and test sets.

| Sets | *N* | $r^2$ | *SDEC* | $Q_{loo}^2$ | $Q^2_{boot}$ | $Q^2_{ext}$ | *SDEP* | *F* |
|---|---|---|---|---|---|---|---|---|
| Training set | 381 | 0.693 | 0.373 | 0.671 | 0.658 | – | – | 78.06 |
| Test set | 116 | – | – | – | – | 0.664 | 0.386 | – |

ment between the correlation and the variation in the data. The cross-validated correlation coefficient $Q_{loo}^2 = 0.670$ illustrates the performance of the model by focusing on the sensitivity of the model to the elimination of any one data points. Also the model was validated by Y randomization test achieving a($R^2$) = 0.009, which indicate that the good results of the original model are not due to *chance correlation* or structural dependency of the training set. In addition, the statistical parameters applying the AD to the TS and PS (shown in Table 6) satisfy the generally accepted condition, see this reference for details[32] and thus demonstrate the predictive power of the present model (see Table 3). Also the experimental and predicted log *BB* values from Equation **2** for the training and test sets are shown in Supporting Information as Table S9; and the plotted values are shown in Figure 12. The mean absolute error (*MAE*) for the training and test set is 0.10 and 0.30, respectively, resulting in a *MAE* of 0.14 for the entire dataset. Thus, the model can be regarded as an optimal regression equation and these values are acceptable if the uncertainty that accompanies the experimental determination of log *BB* is taken into consideration.

#### 3.3.1 Interpretation MLR-Based QSPkR Model

As it can be observed, in the regression model, most of the included variables are very close to the structural features that influence on the log *BB* values or the passage of compounds through the BBB. The vast majority of them tend to impact negatively log *BB* values so for a given compound experimental log *BB* should tend to decrease with the rise of the value of this *MDs*. Among them some like the nRCOOH (number of aliphatic carboxylic acids) and nOHp (number of primary alcohols) are related with the number of functional groups. These are logical results be-

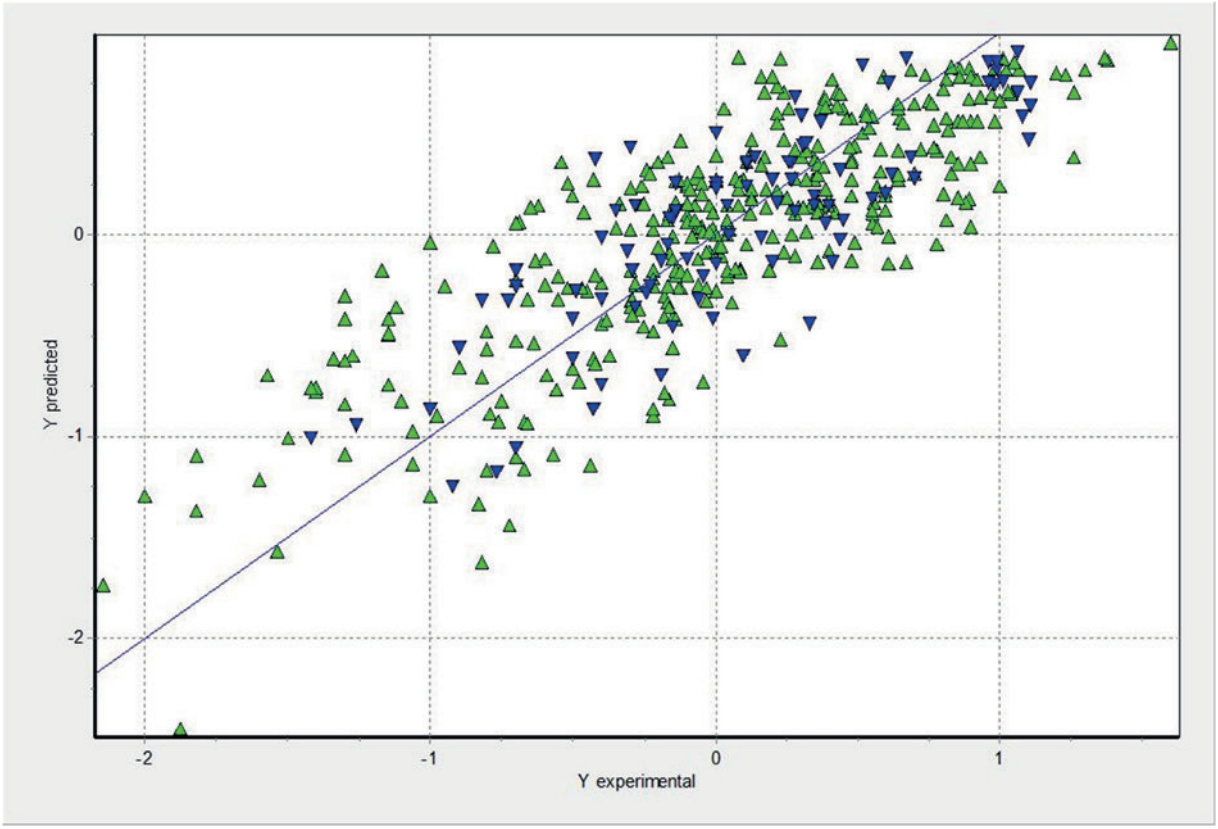


**Figure 12.** The observed and predicted log *BB* values (Equation 2) for training and test sets.

cause the presence of such groups could be related with the possibility of molecule ionization and obtain a charge decreasing the ability of compounds to cross the BBB. Also nArX (number of halogens on aromatic ring) is among *MDs* related to the number of functional groups but it impact positively log *BB* values. This result is in agreement with the fact that increasing lipophilicity (in this case increasing the number of halogens in aromatic rings) is rather close with a better ability of compounds to cross biological membranes. Noteworthy that previous reports also recognizes the inclusion of descriptors related to functional groups counts as way to explain the passage of compounds through the BBB.[19,73,77]

Other descriptors such as H-051 contains information about hydrogen (H) attached to alpha-carbon (C attached through a double (=), triple (#) and single (– –) bond with halogens –C=X, –C#X, –C– –X , respectively; while H-052 means H attached to C (sp3) with one halogen atom attached to next C. As well N-074 contains information related to any group linked through carbon by double or triple bond to nitrogen atoms. Those *MDs* are related with the structural features of compounds and experimental log *BB* values negatively depend on these descriptors. This fact can be due those fragments are related with the ability of compounds to donate H-atoms or hydrogen bond to heteroatoms in the molecules. That's in line with the fact that the increasing of the number of heteroatoms and the hydrogen bound to heteroatoms in the molecules decrease the permeability across the biological membrane. This fits or become in agreement with other researchers who found that the inclusion of some fragments descriptors achieved an improvement in the model.[30,32]

There are two variables in the model due their nature is not evident to explained in an understandable way how they can be related with the ability of compounds to penetrate the BBB. Those are among Burden eigenvalue and topological descriptors. The former represent by T(N··Br), which is related to the sum of topological distances between N··Br, while the latter corresponds to BELm3 lowest eigenvalue n. 3 of Burden matrix /weighted by atomic masses. Both descriptors also impact negatively the passage of compounds through the BBB. Although the inclusion of those MDs is not evident to explain must be remember that the present model have been constructed on the largest available data set in terms of log *BB* and is very challenging try to explain all general trends.

Finally, in our model TPSA(NO) and nHDon (number of donor atoms for H-bonds (N and O)) are variables that contributes negatively to the log *BB* of the compounds. This fits with the fact that the ability of forming H-bonds act preventing a high brain/blood partitioning.[68]

#### 3.3.2 Measures of Quality from Differences Between Published log BB values

In this contribution also and additional attempt was made to estimate the experimental uncertainty associate with log *BB* for compounds with more than one experimental value reported. Following the approach proposed by Kramer et al.[80] the standard deviation of the measurements ($\sigma E$) and the mean unsigned error (*MUE*) was determined. After applying those equations to all compounds with at least two independently measured values a $MUE = 0.35$ log *BB* units and $\sigma E = 0.39$ log *BB* units were derived as error estimates for individual published log *BB* values. These results can help future modeling works in setting the maximum performance achievable using BM581.

### 3.4 Non-Linear Machine Learning Models

The results for each ML technique used to develop various models for classification and regression purposes are given. Here only the best model for each ML approach is displayed. Meanwhile that most of the compounds with anomalous behavior (outliers detected in previous linear models) are included in the present models. The performance of the models is discussed below.

#### 3.4.1 Classification Models

*C-SVM:* A gaussian kernel implemented in WEKA was explored. For each configuration, different parameters were examined by "*trial and error*" strategy ranking the results of performance. The configuration that yields highest ranking was selected. For the gaussian kernel we ran the experiments with gamma values from 0.0 to 10 and ranked the obtained results. Finally, gamma = 0.01 yielded the highest ranking with C value of 1.0. In this model, an average of 78.74% of the chemicals were correctly classified in the 10-fold cross validation with $fp_{rate}$ of 2.16 . Interestingly in the *PS*, 83.62% of the compounds were correctly predicted.

*LR:* In the initial step the ridge parameter was varied between 0.0 and 0.15 without a significant different in the models performance. Finally the configuration that used ridge parameter of 0.1 was selected. That yields, an average of 78.74% of the chemicals correctly classified and $fp_{rate}$ of 2.18 during 10-fold cross validation. Using the *PS*, a better degree of classification was achieved (85.34%) and $fp_{rate}$ of 1.47.

#### 3.4.2 Regression Models

*ε-SVR:* Several kernels implemented in WEKA were attempted. For each one, some parameters were analyzed following "*trial and error*" strategy. Finally the results were ranked choosing the highest ranked configuration. We found the gaussian kernel produced the best results. Finally the highest ranking was achieved, at epsilon and gamma values of

1.0 and 0.001. The MAE of 0.40 and the correlation coefficient value of 0.66 for the PS is quite similar to the results using the linear methods.

*GP:* Another model is developed by Gaussian process. The best result is achieved again using a gaussian kernel. After all exploratory analysis gamma value of 1.0 was found the optimum. Again the performance was almost the same in comparison to previous methods achieving correlation coefficient value of 0.66 and MAE of 0.43 for the PS.

#### 3.4.3 Brief Conclusion Machine Learning Methods

In this section, making use of several ML techniques in addition to LDA and MLR, we have demonstrated that even using more powerful methods still been a challenge to model this very difficult and attractive biological endpoint. However must be highlighted that the performance of more complex methods appear quite similar to previous simpler linear models even using several strategies and combinations of 14 *MDs* . Additionally should be considered that the actual non-linear models have been build considering all the outliers previously removed from linear models.

### 3.5 Compounds Removed in Each Step and Reasons for Their Exclusion

Due the importance of data and model fitting,[22] in the present report this process received special attention comprising some compounds that were set aside at each step. The *first step* allows us to identify those compounds with anomalous behavior in the whole series (BBB+ and BBB−) during the *data fitting* procedure and they were excluded before defining the *modeling set*, training and test models. The *second step* comprises of the exclusion of compounds causing large errors during training and testing the models so they were identified as possible *outliers* in the process of fitting Equation 1 and Equation 2. All those compounds were set aside into an additional set denoted as *controversial set*, which is available in the Supporting Information Table S10.

As we can expected the statistics for these set were not good when Equations 1 and 2 were test on the whole set. However, after trying to find additional explanations (not only statistical reasons) to the anomalous behavior of some compounds, it seem really interesting, that the vast majority of them are related to active transporters, metabolic factors, carrier mediated transport (CMT) (either influx or efflux) and so forth, details in Table 4. These findings are in agreement with those ones of related studies that excluded compounds either because they were *outliers* or had structural features related to these processes, so they become as *outliers* to the passive diffusion algorithm generally assumed by the model.[9]

### 3.6 Comparison with Other Published Models

#### 3.6.1 Classification Models for BBB

Equation 1 is considered to be the best model created in the present report. To compare their performance respect to other classification models published, overall accuracies from different studies reported in the literature are provided in Table 5. However, direct comparison of our results to those previously reported should be done with caution. First, it must be kept in mind that classification approaches to predict BBB passage reported differ on the number and sets of compounds used to train and test the models, MDs used to encode chemical information, classification methods, parameters used as performance criteria and methods for generating *TS* and *PS*. In this sense, to make our comparison with previously approaches in a more rational way, emphasis is placed on those studies when a reasonable number of compounds is used and for the vast majority of them a threshold value based on real log *BB* values is used as the main criteria for defining classification groups or afford interpretation of predictions made with QSPkR models.

The statistical parameters of the LDA based BBB−model generated in our study for our extended *modeling set* appear statistically equivalent to or better than those reported in the literature (see Table 5). Of course, as previously mentioned the direct comparison is difficult, since different models were generated for different sets of compounds but we stress that our model have been trained and test using sets of compounds by far larger than any of the available datasets studied previously (see Table S1 at Supporting Information). Among the other approaches that at least used a reasonable number of compounds, only the best model reported in ref.[19] achieved better results. However, should be noted that those are statistics for a smaller training set (132 compounds); besides, their *PS* (15 compounds) is way smaller than our PS set (116 compounds). Also the model was obtained employing a boosting method and 3D Dragon MDs. By contrast our model is based on a simpler approach that derives a global model using only *LDA* and 0D-2D Dragon *MDs*. All these results are summarized in Table 4, where a comparison among different computational schemes can be performed.

Finally, our model was compared with the best models that only employ one kind of indices by each 0D-2D *MDs* implemented in the Dragon software. Their development was under the same conditions involved in the development of our best LDA-based QSPkR. The five obtained models together with their statistical parameters are shown in Table 6. However, a more detailed discussion on these models is out of the scope of the present report.

#### 3.6.2 Quantitative Models for log BB Prediction

Considering that the experimental log *BB* measurement error is around 0.3 log units and taking into account the

  

**Table 4.** Examples of omitted compounds and reasons for their exclusion.

| No | Name | Comment | Reference (when available) |
|---|---|---|---|
| 1 | Dipyridamole [a] | [d] P-gpefflux | [88–89] |
| 2 | Sparfloxacin [a] | [d] P-gpefflux | [90] |
| 3 | ZiPr-DOPA(P)2 [a] | Prodrug of levodopa a well-known compound transported throughCMT | [91] |
| 4 | Ivermectin 1a [a] | [d] P-gpefflux | [89] |
| 5 | Flunitrazepam [a] | [d] P-gpefflux | [12] |
| 6 | Fluvoxamine [a] | Possible OCT* influx | [36] |
| 7 | Mefloquine [a] | Identified as outlier by Cabrera, M.A et al. | [92] |
| 8 | Diltiazem [a] | [d] P-gp efflux<br>Identified as controversial compound by Broccatelli, F. et al. | [70, 89] |
| 9 | Rapacuronium [a] | Identified as outlier by Garg, P. and Verma, J. | [7] |
| 10 | Cyclosporine A [a] | [d] P-gp efflux | [89, 93–94] |
| 11 | Norcuron [a] | Identified as outlier by Garg, P. and Verma, J. | [7] |
| 12 | Lovastatin [a] | [d] P-gpefflux | [89] |
| 13 | Warfarin [a] | Undergo active transport mechanisms or ionization under physiological conditions that complicates membrane passage<br>Identified as outlier by Garberg, P. et al. | [94–95] |
| 14 | Fluphenazine [a] | [d] P-gp efflux<br>Identified as outlier by Platts, J. A. et al.<br>Abraham, M. H.et al. agree with previously reports that exclude this compound possibly because of active transport or metabolism. | [33, 77, 96] |
| 15 | Morpholinyloxaunomycin [a] | Identified as singleton and possible structural outlier in this work | |
| 16 | 4-Fluoropaclitaxel [a] | possible substrate of P-gp<br>Identified as singleton and possible structural outlier in this work | [89] |
| 17 | Cyanidin 3-O-glucoside [a] | Identified as singleton and possible structural outlier in this work | |
| 18 | Ondansetron [b] | [d] P-gp efflux<br>Identified as outlier in this work. | [89] |
| 19 | Phenserine [b] | Identified as outlier by Platts, J. A. et al.<br>Abraham, M. H. et al. agree with previously reports that exclude this compound possibly because of active transport or metabolism.<br>Identified as outlier in this work. | [33, 77] |
| 20 | Morphine [b] | [d] P-gp efflux<br>Identified as outlier by Cabrera, M.A et al.<br>Identified as outlier in this work. | [12, 92, 94] |
| 21 | Paraxanthine [b] | Identified as a problematic comp by Abraham, M. H. et al.<br>Identified as outlier in this work. | [33] |
| 22 | Bunitrolol [b] | [d] P-gp efflux<br>Identified as outlier in this work. | [12] |
| 23 | Gentisicacid [b,c] | Identified as outlier by Urbano-Cuadrado, M. et al.<br>Identified as outlier in this work. | [97] |
| 24 | Metoclopramide [b] | [d] P-gp efflux<br>Identified as a notable outlier by Di, L. et al.<br>Identified as outlier in this work. | [96] |
| 25 | Oxazepam [b] | Identified as outlier by Rose, K. et al.<br>Identified as outlier in this work. | [98] |
| 26 | Mesoridazine [b,c] | Identified as marked outlier by Abraham, M. H. et al.<br>Identified as outlier in this work. | [33] |
| 27 | Bisphenol A [b] | Identified as a problematic comp by Abraham, M. H. et al.<br>Identified as outlier in this work. | [33] |
| 28 | Pentobarbital [b] | Identified as a problematic comp by Narayanan, R. and Gunturi, S. B.<br>Identified as outlier by Rose, K. et al.<br>Identified as outlier in this work. | [83, 98] |
| 29 | Phenothiazine [b,c] | [d] P-gp efflux<br>Identified as outlier by Fan, Y. et al.<br>Identified as outlier in this work. | [34] |
| 30 | Nevirapine [b] | Transported throughCMT<br>Identified as outlier in this work. | [93] |
| 31 | YG20 [b] | possible substrate of P-gp<br>Identified as strong outlier by Kaznessis, Y. N. et al.<br>Identified as outlier in this work. | [32, 77, 99] |
| 32 | Brezal [c] | Identified as experimental or response outlier in this work | |

**Table 4.** (Continued)

| No | Name | Comment | Reference (when available) |
|---|---|---|---|
| 33 | Carnitine [c] | Identified as experimental or response outlier in this work | |
| 34 | TZ-15 [c] | Identified as response outlier in this work | |
| 35 | Flurbiprofen [c] | Identified as response outlier in this work | |
| 36 | DPDPE [c] | Identified as response outlier in this work | |
| 37 | YG15 [c] | Identified as response outlier in this work | |
| 38 | Citalopram [c] | Identified as structural or X outlier in this work | |
| 39 | Baclofen [c] | Identified as structural outlier in this work | |
| 40 | MGS0028 [c] | Identified as structural outlier in this work | |
| 41 | Pfizer Compound 3 [c] | Identified as structural outlier in this work | |
| 42 | Sulfasalazine [c] | Identified as structural outlier in this work | |

[a] Compound excluded in the process of data fitting. [b] Compound excluded in the process of LDA model fitting. [c] Compound excluded in the process of MLR model fitting. *P-glycoprotein. [d] OCT (a transporter that facilitates transport across the apical membrane of an epithelial cell).[93]

**Table 5.** Summary of classification models and data sets used for performance comparison. PLS-DA: Principal Least Squares -Discriminant Analysis. RF: Random forest.

| Study | *N* | $Q_T$ (%) | *Q* (%) | | Method | Program (or descriptor type) |
|---|---|---|---|---|---|---|
| | | | BBB+ | BBB− | | |
| Crivori et al.[100] | 110 [a,c] | – | – | – | PCA<br>PLS-DA | VolSurf |
| | 120 [b,c] | 71.70 | 90.00 | 65.00 | | |
| Feher et al.[81] | 61 [a] | – | – | – | MLR | Chemical & physical properties |
| | 39 [b] | 74.35 | 73.00[1] | 78.00 [e] | | |
| | | 82.05 | 85.00[2] | 67.00 [f] | | |
| Ooms et al.[101] | 83 [a] | 81.48 | 79.00[3] | 85.00 [g] | PCA | VolSurf |
| | | 74.07 | 65.00[4] | 90.00 [h] | PLS-DA | |
| | – | – | – | – | | |
| Hutter[102] | 90 [a] | – | – | – | MLR | VAMP |
| | 60 [b] | 76.66 | 71.00 | 81.00 | | |
| Narayanan and Gunturi[83] | 88 [a] | – | – | – | MLR | 'Bio-suite'<br>(in-house software) |
| | 13 [b] | – | – | – | | |
| | 15 [b] | – | – | – | | |
| | 92 [b] | 80.43 | 80.00 | 81.00 | | |
| Li et al.[67] | 415 [a,d] | 71.00 | 83.90 | 46.40 | LR | chemical & physical properties, topological, and quantum chemical |
| | | 71.20 | 78.20 | 58.30 | LDA | |
| | | 74.30 | 80.30 | 62.80 | CT | |
| | | 77.10 | 85.50 | 61.40 | k-NN | |
| | | 76.50 | 84.30 | 62.10 | PNN | |
| | | 83.70 | 88.60 | 75.00 | SVM | |
| Deconinck et al.[19] | 132 [a] | 80.60 | – | – | CT-1 | Dragon |
| | | 83.60 | – | – | CT-2 [l] | |
| | | 91.10 | – | – | Bootstrap | |
| | 15 [b] | – | – | – | - | |
| Guerra et al.[18] | 30 [a] | 83.00 | – | – | ANN | CODE |
| | 74 [b] | 73.00 | – | – | | |
| Shen et al. [16] | 151 [a] | – | – | – | MLR | Dragon |
| | 28 [b] | – | – | – | | |
| | 91 [b] | 82.42 | 77.50 | 86.27 | | |
| Zhang et al.[32] | 144 [a] | – | – | – | | |
| | 99 [b] | 82.50 | – | – | kNN | Molecular Operating Environment method (MOE) |
| | | 86.50 | – | – | Consensus kNN,SVM | Dragon - MOE -MolConnZ |
| | 267 [b] | 59.00 | – | – | kNN | MOE |
| | | 80.90 | – | – | Consensus kNN,SVM | Dragon - MOE -MolConnZ |

**Table 5.** (Continued)

| Study | *N* | $Q_T$ (%) | *Q* (%) | | Method | Program (or descriptor type) |
|---|---|---|---|---|---|---|
| Kortagere et al.[74] | 78 [a] | 88.00 | 83.00 | 95.00 | MLR | MOE and Shape Signatures method |
| | 181 [b] | 66.00 | 47.00 | 86.00 | | |
| | 269 [b] | 71.00 | 67.00 | 75.00 | | |
| | 61 [b] | 93.00 | 92.00 | 94.00 | | |
| | 376 [b] | 60.00 | 45.00 | 89.00 | | |
| | 351 [b] | 71.00 | 66.00 | 78.00 | | |
| | 389 [b,k] | 77.00 | 45.00 | 88.00 | | |
| | 351 [b,i] | 82.00 | 84.00 | 79.00 | SVM | 2D (shape + charges) Shape Signatures |
| | | 80.00 | 80.00 | 79.00 | | MOE |
| | 376 [b,i] | 80.00 | 89.00 | 62.00 | | 2D (shape + charges) Shape Signatures |
| | | 76.00 | 89.00 | 51.00 | | MOE |
| | 351 [b,j] | 77.00 | | | | 1D Shape Signatures (shape only) |
| | | 83.00 | | | | 2D Shape Signatures (shape + charges) |
| | 376 [b,j] | 73.00 | | | | 1D Shape Signatures (shape only) |
| | | 80.00 | | | | 2D Shape Signatures (shape + charges) |
| | 389 [b,k] | 66.00 | 56.00 | 69.00 | Consensus Model based on MLR, Rule based and SVM models | MOE and Shape Signatures method |
| Vilar et al.[68] | **307** [a] | 79.47 | 78.20 | 80.20 | LDA | MOE method |
| | | 80.13 | 80.40 | 78.40 | | |
| Muehlbacher et al.[12] | **202** [a] | 88.20 | – | – | RF | MOE, ACD/Labs method and size intensive descriptors |
| **This work** | **381** [a] | **85.09** | **86.32** | **83.80** | **LDA** | **Dragon** |
| | **116** [b] | **83.33** | **83.92** | **82.75** | | |

[a] Training set. [b] Test set. [c] Counting isomers. [d] According to Li et al.[67] the total of 415 agents have known BB ratios, but experimental values are not available, also the accuracy of each method is taken from the average accuracy of a 5-fold cross validation by using the whole dataset and different statistical learning methods; LR (logistic regression), LDA (linear discriminate analysis), C4.5 DT (C4.5 decision tree), *k*-NN (*k* nearest neighbor), PNN (probabilistic neural network), SVM (support vector machine), RFE (recursive feature elimination). [e] Employing cut-off value of −0.5 and [f] employing cutoff value of −1.0. [g] Employing cut-off zero and [h] cut-off value of −0.5. [i] According to Kortagere et al.[74] Leave-20%-out testing. [j]According to Kortagere et al.[74] 10-fold cross validations. [k] SCUT database of FDA approved drugs with BBB permeation classified based on known therapeutic use (knowledge based method). [l] the best model.

**Table 6.** Prediction performances for the best LDA-QSAR models employing one kind of indices by each 0D-2D MDs in the training set.

| Family | Index | *MCC* | $Q_T$ [a] | $fp_{rate}$ [a] | *Se* [a] | *Sp* [a] |
|---|---|---|---|---|---|---|
| 0D | Constitutional Descriptors | 0.56 | 78.05 | 23.46 | 79.47 | 78.24 |
| | | 0.53 | 76.32 | 23.21 | 75.86 | 77.19 |
| 1D | Atom-centered fragments | 0.53 | 76.69 | 28.49 | 81.58 | 75.24 |
| | | 0.52 | 75.43 | 35.71 | 86.21 | 71.43 |
| 2D | Topological descriptors | 0.55 | 77.24 | 21.23 | 75.79 | 79.19 |
| | | 0.55 | 77.19 | 17.86 | 72.41 | 80.77 |

[a] All values are express as percentage (%).

uncertainty in available log *BB* values, it can be concluded that, the predictive model for BBB penetration (Equation 2) performs reasonably well. As previously mentioned the direct comparison is difficult, since different models were generated for different sets of compounds and modeling techniques. A summary of some models of blood–brain distribution that used reasonably large data sets are shown in Table 7. For the vast majority of previous reports smaller training were used to develop their models,[30,75,81–85] and most of them do not demonstrate their real predictive power using a test set.[28,73,83,86–87] As shown in Table 7, our prediction model is as good as others models that employ complex machine learning techniques. However, our model is much simpler than most models and has similar predictive ability. From all the data set analyzed, the training set of our model is the largest available in relation to the rest of the models. Also, our model showed an adequate level of mean absolute error (*MAE* of 0.31) for the test being a good confirmation of the predictive quality of the model. Therefore although some recent reports achieved comparable results using acceptable number of compounds, must be notice that some of them share the major drawback related to the unavailability of data set used to generate the model. Hence though the models are available the results

  

**Table 7.** Summary of correlation models and data sets used for performance comparison. PLS: Partialleast-squares, PCR: principle component regression, VSMP: Variable Selection and Modeling method, *SE*: Standard error, MCCV: Monte Carlo cross-validation, GAVS: Genetic Algorithm Based Variable Selection, BRT: Boosted regression trees, MARS: Multivariate adaptive regression splines, GP-Nest:Gaussian Process Nested Sampling, PLSR: Partial least squares regression, SVM: support vector machine, NLSMP: Nonlinear least-squares minimization procedure.

| Study | *N* | $r^2$ | $q^2$ | *RMSE* | Method |
|---|---|---|---|---|---|
| Luco[85] | 58 [a] | 0.92 | 0.87 | 0.40 | PLS |
| | 12 [b] | 0.92 | | 0.54 | |
| | 25 | 0.79 | | 0.79 | |
| Feher et al.[81] | 61 [a] | 0.85 | 0.83 | 0.42 | PCR |
| | 12 [b] | 0.97 | | 0.24 | |
| | 25 [b] | 0.76 | | 0.52 | |
| Hou and Xu[82] | 57 [a] | 0.93 | 0.89 | 0.35 | MLR |
| | 12 [b] | 0.94 | | 0.31 | |
| | 23 [b] | 0.80 | | 0.52 | |
| Stanton[86] | 47 | 0.78 | 0.77 | | PLS |
| Cabrera[92] | 114 [a] | 0.84 | 0.65 | 0.43 | MLR |
| | 28 [b] | | | 0.33 (*MAE*) | |
| Narayanan and Gunturi[83] | 88 [a] | 0.86 | 0.85 | 0.39 (*SE*) | VSMP |
| Abraham[33] | 302 [a] | 0.75 | | 0.30 (s) | MLR |
| Wichmann[28] | 103 [a] | 0.71 | 0.68 | 0.40 | MLR |
| Konovalov[87] | 291 [a] | 0.75 | 0.73 | 0.30 (s) | *k*NN-MLR |
| Obrezanova[75] | 85 [a] | 0.59 [a] | | 0.52 | PLS |
| | 21 [b] | 0.73 [b] | | 0.40 | |
| | | 0.61 [a] | | 0.50 | GP-Basic |
| | | 0.74 [b] | | 0.39 | |
| | | 0.61 [a] | | 0.50 | GP-FVS |
| | | 0.74 [b] | | 0.39 | |
| | | 0.66 [a] | | 0.47 | GP-Opt |
| | | 0.77 [b] | | 0.36 | |
| | | 0.69 [a] | | 0.44 | GP-Nest |
| | | 0.81 [b] | | 0.34 | |
| Deconinck et al.[45] | 183 [a] | 0.82 [a] | | 0.34 | BRT |
| | 61 [b] | 0.71 [b] | | 0.53 | |
| | | 0.88 [a] | | 0.32 | MLR |
| | | 0.72 [b] | | 0.48 | |
| | | 0.90 [a] | | 0.26 | MLR–BRT |
| | | 0.90 [b] | | 0.46 | |
| | | 0.82 [a] | | 0.39 | PLS |
| | | 0.80 [b] | | 0.41 | |
| | | 0.83 [a] | | 0.37 | PLS–BRT |
| | | 0.80 [b] | | 0.41 | |
| Konovalov[73] | 289 [a] | 0.57 | | 0.39 (SE) | MCCV& MLR |
| Shen et al.[16] | 151 [a] | 0.85 | 0.82 | | GAVS&Dragon |
| | 28 [b] | 0.84 | | | |
| Fu et al.[84] | 86 [a] | 0.74 | 0.71 | 0.37 (s) | MRL |
| | 25 [b] | | | 0.53 | |
| Zhang[32] | 144 [a] | 0.92 | | 0.18 | kNN-Dragon |
| | | 0.86 | | 0.27 | SVM-Dragon |
| | | 0.75 | | 0.31 | kNN-MOE |
| | | 0.82 | | 0.24 | SVM-MOE |
| | | 0.95 | | 0.15 | kNN-MolConnZ |
| | | 0.87 | | 0.25 | SVM-MolConnZ |
| Kortagere[74] | 78 [a] | 0.70 | | | MRL-MOE |
| | 100 [b] | 0.65 | | | |
| Deconinck[103] | 224 [a] | 0.85 [a] | | 0.52 | BRT |
| | 75 [b] | 0.54 [b] | | 0.68 | |
| | | 0.88 [a] | | 0.41 | MARS |
| | | 0.24 [b] | | 1.09 | |
| | | 0.67 [a] | | 0.62 | Stepwise-MRL |
| | | 0.51 [b] | | 0.71 | |
| | | 0.68 [a] | | 0.62 | Stepwise MLR–BRT |

**Table 7.** (Continued)

| Study | *N* | $r^2$ | $q^2$ | *RMSE* | Method |
|---|---|---|---|---|---|
| | | 0.52 [b] | | 0.71 | |
| | | 0.71 [a] | | 0.60 | TMARS |
| | | 0.61 [b] | | 0.64 | |
| | | 0.64 [a] | | 0.65 | PLS |
| | | 0.57 [b] | | | |
| | | 0.87 [a] | | 0.32 | PLS–BRT |
| | | 0.58 [b] | | 0.65 | |
| | | 0.64 [a] | | 0.65 | PLS–MARS |
| | | 0.61 [b] | | 0.64 | |
| Obrezanova[104] | 106 [a] | 0.79 | | 0.32 | GP-Nest |
| | 23 [b] | 0.72 | | 0.38 | |
| | 22 [b] | 0.66 | | 0.49 | |
| | 143 [c] | 0.27 | | 0.49 | |
| | 205 [a+] | 0.80 | | 0.29 | GP-2D |
| | 44 [b] | 0.73 | | 0.33 | |
| | 43 [b] | 0.67 | | 0.35 | |
| Fan[34] | 193 [e] | 0.74 | 0.72 | | (GA)-MLR |
| | 81 [f] | 0.67 | | 0.60 | |
| | 160 [b] | 0.65 | | 0.32 | |
| Zhang[30] | 57 [a] | 0.92 | 0.79 | 0.28 | PLSR |
| | 13 [b] | 0.66 | | 0.78 | |
| Chen[35] | 432 [e] | 0.94[e] | 0.5 | 0.18 | RF |
| | 73 [f] | 0.57[f] | | 0.58 | |
| | | 0.83[e] | 0.6 | 0.3 | SVM |
| | | 0.55[f] | | 0.59 | |
| Lanevskij[36] | 329 [e] | 0.75 | | 0.38 | NLSMP |
| | 141 [b] | 0.74 | | 0.39 | |
| | 30 [f] | 0.72 | | 0.43 | |
| | 21 [f] | 0.71 | | 0.35 | |
| Muehlbacher[12] | 362 [a] | 0.59 | | | MLR |
| Yan[105] | 198 [a] | 0.90 [a] | | 0.62(s) | MLR |
| | 122 [b] | 0.89 [b] | | 0.58(s) | |
| | | 0.90 [a] | | 0.61(s) | SVM |
| | | 0.89 [b] | | 0.56 (s) | |
| | | 0.90 [a] | | 0.63(s) | ANN |
| | | 0.90 [b] | | 0.58 (s) | |
| This work | 369 [a] | 0.69 | 0.67 | 0.10 (*MAE*) | MLR |
| | 116 [b] | | | 0.31 (*MAE*) | |

[a] Training set. [b] Test set. [c] Accordingly to Ref Abraham 143 set. [a+] Acordingly to ref combined original training set and the Abraham 143 set. [e] In house training set (not available). [f] In house test set (not available).

cannot be easily reproduced or extended. Also some the essential structural factors that are considered to affect the passage through the BBB by several authors have been expressed by our model. Altogether confirms the well position of the present model as more conveniently filter for BBB passage prediction.

## 4 Conclusions

In summary, this report constitutes an effort to address the best practice of QSAR modeling in the field of BBB passage prediction. We have compiled the largest publicly available dataset of diverse organic molecules with experimentally available log *BB* values and details steps of compilation, steps of chemical curation, threshold analysis, characterization and splitting were also conducted. Also, new validated reproducible and extendable models for classification and regression purposes using simple linear techniques were developed making also an effort to provide brief interpretation of the proposed models. Additionally, serious discussion regarding the anomalous behavior of several compounds is provided.

Finally, an extensive but not direct comparison with previously approaches show that the present models appear statistically equivalent to better than those reported in the literature. There is no doubt about the need for new reliable in silico tools to predict BBB passage thus seems very interesting in futures works to explore others kinds of *MDs* to get new prediction tools. Also the extension of the proposed data set with lately reports and the compilation of new ones related to carrier mediated transport, active

  

transporters and metabolic factors may be a good strategy to cover a more extensive range of drugs entry and efflux mechanisms. These strategies could open new doors resulting in very powerful prediction tools really close to the real life scenario of the transport through the BBB.

## Acknowledgement

Yoan Brito-Sánchez acknowledges the Emerging Leaders in the Americas Program (ELAP) for a fellowship at *Vancouver Prostate Centre, University of British Columbia* (2014). Yovani Marrero-Ponce thanks to the program 'International Professor' for a fellowship to work at *Cartagena University* in 2013–2014. Le-Thi-Thu, H. acknowledges the *National Vietnam National University,* Hanoi for the support provided to the researches. Stephen J. Barigye acknowledges financial support from *CNPq*.

## References

[1] A. R. Katritzky, M. Kuanar, S. Slavov, D. A. Dobchev, D. C. Fara, M. Karelson, W. E. Acree JR, V. P. Solov'ev, A. Varnek, *Bioorg. Med. Chem.* **2006**, *14* 4888 – 4917.
[2] D. Pan, M. Iyer, J. Liu, Y. Li, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2083 – 2098.
[3] N. Strazielle, J. F. Ghersi-Egea, *Mol. Pharmaceutics* **2013**, *10*, 1473 – 1491.
[4] R. Harati, H. Benech, A. S. Villégier, A. Mabondzo, *Mol. Pharmaceutics* **2013**, *10*, 1566 – 1580.
[5] N. J. Abbott, A. A. Patabendige, D. E. Dolman, S. R. Yusof, D. J. Begley, *Neurobiol. Discov.* **2010**, *37*, 13 – 25.
[6] A. Lindqvist, J. Rip, P. J. Gaillard, S. Björkman, M. Hammarlund-Udenaes, *Mol. Pharmaceutics* **2013**, *10*, 1533 – 1541.
[7] P. Garg, J. Verma, *J. Chem. Inf. Model.* **2006**, *46*, 289 – 297.
[8] X. Liu, M. Tu, R. S. Kelly, C. Chen, B. J. Smith, *Drug Metab. Dispos.* **2004**, *32*, 132 – 139.
[9] J. Mensch, J. Oyarzabal, C. Mackie, P. Augustijns, *J. Pharm. Sci.* **2009**, *98*, 4429 – 4468
[10] J. T. Goodwin, D. E. Clark, *J. Pharmacol. Exp. Ther.* **2005**, *315*, 477 – 483.
[11] Y. H. Zhao, M. H. Abraham, A. Ibrahim, P. V. Fish, S. Cole, M. L. Lewis, M. J. de Groot, D. P. Reynolds, *J. Chem. Inf. Model.* **2007**, *47*, 170 – 175.
[12] M. Muehlbacher, G. M. Spitzer, K. R. Liedl, J. Kornhuber, *J. Comput. Aided. Mol. Des.* **2011**, *25*, 1095 – 1106.
[13] W. M. Pardridge, *Drug Discov. Today* **2004**, *9*, 392 – 393.
[14] M. Fridén, S. Winiwarter, G. Jerndal, O. Bengtsson, H. Wan, U. Bredberg, M. Hammarlund-Udenaes, M. Antonsson, *J. Med. Chem.* **2009**, *52*, 6233 – 6243.
[15] K. Lanevskij, P. Japertas, R. Didziapetris, A. Petrauskas, *J. Pharm. Sci.* **2009**, *98*, 122 – 134.
[16] J. Shen, Y. Du, Y. Zhao, G. Liu, Y. Tang, *QSAR Comb. Sci.* **2008**, *27*, 704 – 717.
[17] B. Hemmateenejad, R. Miri, M. A. Safarpour, A. R. Mehdipour, *J. Comput. Chem.* **2006**, *27*, 1125 – 1135.
[18] A. Guerra, J. A. Paez, N. E. Campillo, *QSAR Comb. Sci.* **2008**, *27*, 586 – 594.
[19] E. Deconinck, M. H. Zhang, D. Coomans, Y. Vander Heyden, *J. Chem. Inf. Model.* **2006**, *46*, 1410 – 1419.
[20] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **2013**, 4977 – 5010.
[21] A. J. Williams, S. Ekins, V. Tkachenko, *Drug Discov. Today* **2012**, *17*, 685 – 701.
[22] A. Tropsha, *Mol. Inf.* **2010**, *29*, 476 – 488.
[23] U. Norinder, M. Haeberlein, *Adv. Drug Deliv. Rev.* **2002**, *54*, 291 – 313.
[24] A. R. Mehdipour, M. Hamidi, *Drug Discov Today* **2009**, *14*, 1030 – 1036.
[25] D. E. Clark, *Drug Discov. Today* **2003**, *8*, 927 – 933.
[26] M. Hammarlund-Udenaes, U. Bredberg, M. Friden, *Curr. Top. Med. Chem* **2009**, *9*, 148 – 162.
[27] J. H. A. Al-Fahemi, D. L. Cooper, N. L. Allan, *J. Mol. Graph. Model.* **2007** *26*, 607 – 612.
[28] K. Wichmann, M. Diedenhofen, A. Klamt, *J. Chem. Inf. Model.* **2007**, *47*, 228 – 233.
[29] M. H. Abraham, A. Hersey, in *Comprehensive Medicinal Chemistry II*, Vol. 5 (Eds: J. B. Taylor, D. J. Triggle), Elsevier, Oxford, **2006**, pp 745 – 766.
[30] Y. H. Zhang, Z. N. Xia, L. T. Qin, S. S. Liu, *J. Mol. Graph. Model.* **2010**, *29*, 214 – 220.
[31] H. Dureja, A. K. Madan, *Int. J. Pharm.* **2006**, *323* 27 – 33.
[32] L. Zhang, H. Zhu, T. I. Oprea, A. Golbraikh, A. Tropsha, *Pharm. Res.* **2008**, *25*, 1902 – 1914.
[33] M. H. Abraham, A. Ibrahim, Y. H. Zhao, W. E. Acree Jr, *J. Pharm. Sci* **2006**, *95*, 2091 – 2100.
[34] Y. Fan, R. Unwalla, R. A. Denny, L. Di, E. H. Kerns, D. J. Diller, C. Humblet, *J. Chem. Inf. Model.* **2010**, *50*, 1123 – 1133.
[35] H. Chen, S. Winiwarter, M. Fridén, M. Antonsson, O. Engkvista, *J. Mol. Graph. Model.* **2011**, *29*, 985 – 995.
[36] K. Lanevskij, J. Dapkunas, L. Juska, P. Japertas, R. Didziapetris, *J. Pharm. Sci.* **2011**, *100*, 2147 – 2160.
[37] in *37th Joint Meeting of the Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology*, Paris, 17–19 November **2004**.
[38] R. Cecchelli, V. Berezowski, S. Lundquist, M. Culot, M. Renftel, M. P. Dehouck, L. Fenart, *Nat. Rev.* **2007**, *6*, 650 – 661.
[39] A. Reichel, D. J. Begley, N. J. Abbott, *Biol. Res. Protoc.* **2003**, *89*, 307 – 324.
[40] *ChemDraw*, Version 7.0.1 ed., CambridgeSoft Co, Cambridge, **2002**.
[41] *OpenBabel*, Version 2.3.0 ed., **2010**.
[42] *JChem*, Version 6.1.2 ed., **2013**.
[43] Y. Marrero-Ponce, C. R. García Jacas, J. R. Valdés Martini, *TOMOCOMD-CARDD* software (TOpological MOlecular COMputational Design – Computer-Aided Rational Drug Design), (www.tomocomd.com), Santa Clara, Villa Clara, Cuba, **2002–2014**. The QUBILs' Framework (v1.0) allows easy calculation of algebraic forms-based molecular descriptors. Three modules are included, a) *QuBiLs-MAS*, b) *QuBiLs-MIDAS* and c) *QuBiLs-POMAS*. They are based on the application of mathematical N-linear transformations using 2–4 *n*-tuple matrix representations. A professional version can be obtained upon request to Y. Marrero-Ponce: ymarrero77@yahoo.es.
[44] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, *MATCH Commun. Math. Comput. Chem.* **2006**, *56*, 237 – 248.
[45] E. Deconinck, M. H. Zhang, D. Coomans, Y. Vander Heyden, *J. Chemom.* **2007**, *21*, 280 – 291.
[46] A. Borota, M. Mracec, A. Gruia, R. Rad-Curpăn, L. Ostopovici-Halip, M. Mracec, *Eur. J. Med. Chem.* **2011**, *46*, 877 – 884.

  

[47] C. Tebby, E. Mombelli, P. Pandard, A. R. R. Péry, *Sci. Total Environ.* **2011**, *409*, 3334 – 3343.
[48] M. P. González, P. L. Suárez, Y. Fall, G. Gómez, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 5165 – 5169.
[49] G. M. Casañola-Martín, Y. Marrero-Ponce, M. T. H. Khan, A. Ather, K. M. Khan, F. Torrens, R. Rotondo, *Eur. J. Med. Chem.* **2007**, *42*, 1370 – 1381.
[50] R. Todeschini, V. Consonni, in *Handbook of Molecular Descriptors,* Vol. 11, 1st ed. (Eds: R. Mannhold, H. Kubinyi, H. Timmerman), Wiley-VCH, Weinheim, Germany, **2000**, p. 667.
[51] R. D. Brown, Y. C. Martin, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572 – 584.
[52] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1992**, *32* 644 – 649.
[53] R. A. Johnson, D. W. Wichern, in *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, **1988**.
[54] J. W. Mc Farland, D. J. Gans, in *Chemometric Methods in Molecular Design* (Ed: H. van de Waterbeemd), VCH, Weinheim, Germany, **1995**, pp. 295 – 307.
[55] *STATISTICA*, Version 6.0 ed., StatSoft Inc, Tulsa, OK, **2001**.
[56] H. van de Waterbeemd, in *Chemometric Methods in Molecular Design* (Ed: H. van de Waterbeemd), VCH Publishers, Weinheim, Germany, **1995**, pp. 265 – 288.
[57] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, H. Nielsen, *Bioinformatics* **2000**, *16*, 412 – 424.
[58] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, 1.0 ed., Milano, **2005**.
[59] A. J. Hopfinger, H. C. Patel, in *Genetic Algorithms in Molecular Modeling* (Ed: J. Devillers), Academic Press, London, **1996**, pp. 131 – 157.
[60] M. Pavan, V. Consonni, P. Gramatica, in *Partial Order in Environmental Sciences and Chemistry* (Eds: R. Brüggeman, L. Carlsen), Springer, Berlin, **2006**, pp. 181 – 217.
[61] S. Le Cessie, J. C. van Houwelingen, *Applied Statistics* **1992**, *41*, 191 – 201.
[62] B. E. Boser, I. M. Guyon, V. N. Vapnik, in *Proc. 5th Ann. ACM Workshop on Computational Learning Theory*, **1992**.
[63] C. Cortes, V. N. Vapnik, *Machine Learning* **1995**, *20*, 273 – 297.
[64] V. Vapnik, in *Statistical Learning Theory*, Wiley, New York, **1998**.
[65] C. E. Rasmussen, C. K. I. Williams, in *Gaussian Processes for Machine Learning*, Springer, Cambridge, **2006**.
[66] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, *SIGKDD Explorations* **2009**, *11*.
[67] H. Li, C. Yap, C. Ung, Y. Xue, Z. Cao, Y. Chen, *J. Chem. Inf. Model.* **2005**, *45*, 1376 – 1384.
[68] S. Vilar, M. Chakrabarti, S. Costanzi, *J. Mol. Graph. Model.* **2010**, *28*, 899 – 903.
[69] Y. Brito-Sánchez, J. A. Castillo-Garit, H. Le-Thi-Thu, Y. González-Madariaga, F. Torrens, Y. Marrero-Ponce, J. E. Rodríguez-Borges, *SAR QSAR Environ. Res.* **2013**, *24*, 235 – 251.
[70] F. Broccatelli, C. A. Larregieu, G. Cruciani, T. I. Oprea, L. Z. Benet, *Adv.Drug Deliv. Rev.* **2012**, *64*, 95 – 109.
[71] A. Golbraikh, A. Tropsha, *J. Mol. Graph Model.* **2002**, *20*, 269 – 276.
[72] P. Ertl, in *Polar Surface Area, in Molecular Drug Properties*, Vol. 7 (Ed: R. Mannhold), Wiley-VCH, Weinheim, Germany, **2008**.
[73] D. A. Konovalov, N. Sim, E. Deconink, Y. V. Heyden, D. Coomans, *J. Chem. Inf. Model.* **2008**, *48*, 370 – 383.
[74] S. Kortagere, D. Chekmarev, W. J. Welsh, S. Ekins, *Pharm. Res.* **2008** *25*.
[75] O. Obrezanova, G. Csányi, J. M. R. Gola, M. D. Segall, *J. Chem. Inf. Model.* **2007**, *47*, 1847 – 1857.
[76] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Comb. Chem.* **1999**, *1*, 55 – 68.
[77] J. A. Platts, M. H. Abraham, Y. H. Zhao, A. Hersey, L. Ijaz, D. Butina, *Eur. J. Med. Chem.* **2001**, *36* 719 – 730.
[78] P. Gramatica, M. Corradi, V. Consonni, *Chemosphere* **2000**, *41*, 763 – 777.
[79] H. Pham-The, I. González-Alvarez, M. Bermejo, V. Mangas Sanjuan, I. Centelles, T. M. Garrigues, M. A. Cabrera-Pérez, *Mol. Inf.* **2011**, *30*, 376 – 385.
[80] C. Kramer, T. Kalliokoski, P. Gedeck, A. Vulpetti, *J. Med. Chem* **2012**, *55*, 5165 – 5173.
[81] M. Feher, E. Sourial, J. M. Schmidt, *Int. J. Pharm.* **2000**, *201*, 239 – 247.
[82] T. Hou, X. Xu, *J. Mol. Model.* **2002**, *8*, 337 – 349.
[83] R. Narayanan, S. B. Gunturi, *Bioorg. Med. Chem.* **2005**, *13*, 3017 – 3028.
[84] X.-C. Fu, G.-P. Wang, H.-L. Shan, Wen-Quan Liang, J.-Q. Gao, *Eur. J. Pharm. Biopharm.* **2008**, *70*, 462 – 466.
[85] J. M. Luco, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 396 – 404.
[86] D. T. Stanton, B. E. Mattioni, J. J. Knittel, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1010 – 1023.
[87] D. A. Konovalov, D. Coomans, E. Deconinck, Y. V. Heyden, *J. Chem. Inf. Model.* **2007**, *47*, 1648 – 1656.
[88] Q. Wang, J. D. Rager, K. Weinstein, P. S. Kardos, G. L. Dobson, J. Li, I. J. Hidalgo, *Int. J. Pharm.* **2005**, *288,* 349 – 359.
[89] M. Adenot, R. Lahana, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 239 – 248.
[90] E. C. M. de Lange, S. Marchandb, D. J. van den Berg, I. C. J. van der Sandt, A. G. de Boer, A. Delon, S. Bouquet, W. Couet, *Eur. J. Pharm. Sci.* **2000**, *12* 85 – 93.
[91] T. Hou, J. Wang, W. Zhang, X. Xu, *J. Chem. Inf. Model.* **2007**, *47*, 208 – 218.
[92] M. A. Cabrera, M. Bermejo, M. Pérez, R. Ramos, *J. Pharm. Sci.* **2004**, *93*, 1701 – 1717.
[93] H. H. Usansky, P. J. Sinko, *Pharm. Res.* **2003**, *20*.
[94] P. Garberg, M. Ball, N. Borg, R. Cecchelli, L. Fenart, R. D. Hurst, T. Lindmark, A. Mabondzo, J. E. Nilsson, T. J. Raub, D. Stanimirovic, T. Terasaki, J.-O. Öberg, T. Österberg, *Toxicol. In Vitro* **2005**, *19*, 299 – 334.
[95] C. Andres, M. C. Hutter, *QSAR Comb. Sci.* **2006**, *25*, 305 – 309.
[96] L. Di, E. H. Kerns, I. F. Bezar, S. L. Petusky, Y. Huang, *J. Pharm. Sci* **2009**, *98*, 1980 – 1991.
[97] M. Urbano-Cuadrado, I. Luque-Ruiz, M. A. Gómez-Nieto, *J. Comput. Chem.* **2007**, *28*, 1252 – 1260.
[98] K. Rose, L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput.Sci.* **2002**, *42*, 651 – 666.
[99] Y. N. Kaznessis, M. E. Snow, C. J. Blankley, *J. Comput-Aided Mol. Des.* **2001**, *15*, 697 – 708.
[100] P. Crivori, G. Cruciani, P. A. Carrupt, B. Testa, *J. Med. Chem.* **2000**., *43*, 2204 – 2216.
[101] F. Ooms, P. Weber, P. A. Carrupt, B. Testa, *Biochim. Biophys. Acta* **2002**, *1587*, 118 – 125.
[102] M. C. Hutter, *J. Comput. Aided. Mol. Des.* **2003**, *17*, 415 – 433.
[103] E. Deconinck, M. H. Zhang, F. Petitet, E. Dubus, I. Ijjaali, D. Coomans, Y. Vander Heyden, *Anal. Chim. Acta* **2008**, *609*, 13 – 23.
[104] O. Obrezanova, J. M. R. Gola, E. J. Champness, M. D. Segall, *J. Comput. Aided. Mol. Des.* **2008**, *22*, 431 – 440.
[105] A. Yana, H. Lianga, Y. Chonga, X. Niea, C. Yu, *SAR QSAR Environ. Res.* **2013**, *24*, 61 – 74.

Received: August 29, 2014
Accepted: January 20, 2015
Published online: May 7, 2015

  