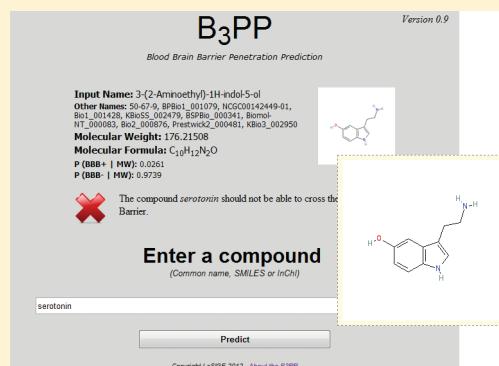


# A Bayesian Approach to *in Silico* Blood-Brain Barrier Penetration Modeling

Ines Filipa Martins,<sup>†</sup> Ana L. Teixeira,<sup>†,¶</sup> Luis Pinheiro,<sup>†</sup> and Andre O. Falcao<sup>\*,‡</sup><sup>†</sup>LaSIGE, Faculty of Sciences, University of Lisbon, Lisbon, Portugal<sup>‡</sup>Department of Informatics, Faculty of Sciences, University of Lisbon, Lisbon, Portugal<sup>¶</sup>CQB - Centro de Quimica e Bioquimica, Faculty of Sciences, University of Lisbon, Lisbon, Portugal

## S Supporting Information

**ABSTRACT:** The human blood-brain barrier (BBB) is a membrane that protects the central nervous system (CNS) by restricting the passage of solutes. The development of any new drug must take into account its existence whether for designing new molecules that target components of the CNS or, on the other hand, to find new substances that should not penetrate the barrier. Several studies in the literature have attempted to predict BBB penetration, so far with limited success and few, if any, application to real world drug discovery and development programs. Part of the reason is due to the fact that only about 2% of small molecules can cross the BBB, and the available data sets are not representative of that reality, being generally biased with an over-representation of molecules that show an ability to permeate the BBB (BBB positives). To circumvent this limitation, the current study aims to devise and use a new approach based on Bayesian statistics, coupled with state-of-the-art machine learning methods to produce a robust model capable of being applied in real-world drug research scenarios. The data set used, gathered from the literature, totals 1970 curated molecules, one of the largest for similar studies. Random Forests and Support Vector Machines were tested in various configurations against several chemical descriptor set combinations. Models were tested in a 5-fold cross-validation process, and the best one tested over an independent validation set. The best fitted model produced an overall accuracy of 95%, with a mean square contingency coefficient ( $\phi$ ) of 0.74, and showing an overall capacity for predicting BBB positives of 83% and 96% for determining BBB negatives. This model was adapted into a Web based tool made available for the whole community at <http://b3pp.lasige.di.fc.ul.pt>.



## ■ INTRODUCTION

The Blood-Brain Barrier (BBB) is a membrane that separates circulating blood and the brain extracellular fluid. Some of the main functions of this barrier comprise the protection of the brain from foreign substances in the blood that may injure it, protection against hormones and neurotransmitters in the rest of the body, and maintenance of a constant environment for the brain.<sup>1</sup> Therefore, the BBB has special features that make it almost impenetrable to most drugs. It has a selective permeability and the molecules that are generally able to cross have been found to be difficult to identify. The features of the BBB represent a problem in CNS drug development, and most pharmaceutical companies do not have a BBB drug targeting development program.<sup>2,3</sup> BBB penetration is one of the key factors that are taken into account in chemical toxicological studies and in drug design.<sup>4</sup> Furthermore direct measurement of BBB penetration is possible but experiments are very expensive and time-consuming<sup>5</sup> and constitute a time and financial hindrance when a large number of compounds are examined. Pardridge<sup>2</sup> discusses the complexity of the process of BBB penetration, and how crucial its understanding is for treatment of several CNS disorders and even some viral infections like AIDS, where the virus lodges itself in brain

tissues, where available antiviral drugs show minimal BBB penetration.

Although *in vitro* models and results from high throughput screening for BBB are becoming available,<sup>6,7</sup> Cucullo et al.<sup>8</sup> have argued that *in vitro* models have been unable to fully reproduce the BBB characteristics *in vivo*, impairing several drug development programs. On the other hand, Goodwin and Clark<sup>9</sup> contend that *in silico* models have so far also failed to be useful in real world scenarios, partly due to the quality and quantity of data available, as the molecules for which BBB penetration data are available have limited chemical representation. Ekins and Tropsha<sup>10</sup> further argue on the scarcity of available data, and the limitations of the chemical families represented, and discuss that not better statistical models are needed but better and more reliable data. The molecular factors that influence the ability of a compound to cross the BBB have been described by Banks<sup>11</sup> by the analysis of available data on BBB characteristics and molecular penetration data. Other studies (e.g., Ooms et al.<sup>12</sup>) used statistical models to infer molecular characteristics that influence BBB penetration.

Received: March 8, 2012

Published: May 21, 2012

**Table 1. Prediction Accuracies and Mean Square Contingency Coefficient for BBB Positives and BBB Negatives (p/n) Compounds from Different Studies Reported in the Literature (Adapted from Li et al.<sup>14</sup> with New Data)**

year	study	desc <sup>c</sup>	methods <sup>a</sup>	no. molecules (p/n)		model validation results			$\phi$
				training	test	sens	spec	acc	
2000	Crivori et al. <sup>17</sup>	72	PLS	46/64	49/71	90	65	74.8	-
2000	Cruciani et al. <sup>18</sup>	-	PLS	46/64	35	-	-	>75	-
2002	Doniger et al. <sup>13</sup>	9	SVM	154/120	25/25	82.7	80.2	81.5	-
2004	Adenot et al. <sup>16</sup>	67	PLS	1336/360	20/62	90	92	91	-
2005	Li et al. <sup>14</sup>	199	SVM	276/139	-	88.6	75.0	83.7	0.645
2007	Zhao et al. <sup>5</sup>	19	R, PLS	832/261	451/49	98.2	87.8	97.2	-
2007	Zhao et al. <sup>5</sup>	19	R, PLS	1283/310	267/130	80.1	63.1	74.6	-
2011	Guerra et al. <sup>24</sup>	-	NN	96/12 <sup>d</sup>	74/4 <sup>d</sup>	85.0	25.0	82.0	-
2008	Zhang et al. <sup>4</sup>	832 <sup>b</sup>	k-NN, SVM	124/20	99 + 267	-	-	82.5–100	-
2008	Kortagere et al. <sup>25</sup>	<100	SVM	186/165	-	84	79	82	0.635
2009	Wang et al. <sup>26</sup>	55	kohNN	1283/310	266/130	94.7	54.6	81.1	-

<sup>a</sup>SVM (support vector machine), PLS (partial least-squares), k-NN (k-nearest neighbors), R (recursive partitioning models), NN - Neural Networks, kohNN - Kohonen Neural Networks. <sup>b</sup>Descriptors are divided into three classes and are tested separately. <sup>c</sup>Number of descriptors. *sens* - sensitivity, *spec* - specificity, *acc* - accuracy,  $\phi$  - mean square contingency coefficient. <sup>d</sup>Number of instances assuming a *logBB* threshold of -0.9.

Automated prediction of drug molecules' BBB penetration would be a useful tool to assist the experimental drug discovery process,<sup>13</sup> decreasing the time of the initial stages and therefore the time required for a drug to reach the market. *In silico* approaches are a valid alternative and have been introduced as prescreening tools for large chemical databases to reduce the cost and enhance the speed of BBB permeability analysis.<sup>4,14</sup> Automated prediction of physiological characteristics of molecules is a complex task as minimal changes in molecular structure can have a big impact on a potential drug pharmacological properties. On the other hand, the large number of potential molecules and drugs does not allow systematic *in vitro* or *in vivo* testing. Computational models can therefore be a viable alternative to high throughput screening, and as computer power becomes more affordable, more data and more complex and accurate models may be produced.

Literature on computational methods for BBB penetration prediction is now quite extensive, despite the relative lack of empirical evidence on molecular BBB penetration properties. Some studies (e.g., Ajay et al.<sup>15</sup> Adenot and Lahana<sup>16</sup>) have circumvented this data limitation by focusing on predicting CNS activity instead. Focusing on BBB penetration prediction, Crivori et al.<sup>17</sup> used a method to transform molecular 3D fields into descriptors that are fitted into a partial least-squares model trained over 110 compounds and their model validated with 85 molecules. The same descriptor generating approach was followed by Cruciani et al.<sup>18</sup> for discriminating molecules BBB penetration characteristics. Doniger et al.<sup>13</sup> have gathered an important molecular database with BBB penetration data for 325 compounds that has been used for comparison by several other studies. These authors compared Support Vector Machines (SVM) with Neural Networks over a 9 descriptor base, concluding that SVM show significant better results than Neural Networks. Iyer et al.<sup>19</sup> used membrane interaction quantitative structure–activity relationship (QSAR) analysis to derive multivariate linear models to predict the blood-brain partition (*logBB*) including membrane-solute descriptors. Hou and Xu<sup>20</sup> also tried to predict the *logBB* gathered from *in vivo* data, using molecular structural descriptors and testing a variety of multilinear models with commonly used descriptors. A different effort using a variety of 4D molecular similarity measures and cluster analysis was used to predict the blood-brain partition over a data set of 150 molecules,<sup>21</sup> suggesting

that consensus models that include different models for different chemical families may be able to perform better than simple models fitted to undifferentiated data sets. Li et al.<sup>14</sup> have assessed different molecular descriptor sets, concluding it to be a fundamental aspect for BBB modeling. These authors have tested a variety of machine learning methods over a database of 415 molecules and 199 descriptors, having selected only 37 using recursive feature elimination. Of all machine learning methods tested, SVM proved to provide the best results. Gerebtzoff and Seelig<sup>22</sup> have estimated the molecular cross sectional area for BBB penetration classification, over a data set 122 known drugs, finding significant linear correlations. Svetnik et al.<sup>23</sup> used random forests (RF) for a variety of QSAR classification problems including BBB penetration prediction over Doniger's<sup>13</sup> database of 325 molecules using 9 chemical and physical molecular properties, concluding that RF outperformed other available methods. SVM were also tested over several combinations of different descriptor sets over 366 molecules for *logBB* prediction, producing models with good prediction accuracy.<sup>4</sup> An indirect way for BBB classification was tried by Guerra et al.<sup>24</sup> These authors estimated *logBB* using a neural network and then defined several *logBB* thresholds with which to make a classification. Also, a novel descriptor set generated by the method of Shape Signatures was tested with linear regression and SVMs, the latter proving to be superior to the former on a variety of published data sets.<sup>25</sup> A large database mostly derived from Zhao et al.<sup>5</sup> was used by Wang et al.,<sup>26</sup> which tested SVM and Kohonen Neural Networks over a 2D property autocorrelation descriptors. A different approach based on genetic algorithms was also used for feature selection, coupled with linear regression models for predicting blood-brain partition in a database of 193 compounds and 217 molecular descriptors,<sup>27</sup> and the resulting model used only 5 descriptors for prediction. Zhao et al.<sup>5</sup> used binomial partial least-squares for classifying an extensive data set with 1593 molecules and 18 descriptors. More recently, Muehlbacher et al.<sup>28</sup> also used random forests for *logBB* prediction using a data set of 362 molecules and 4 descriptors. A selection of classification studies for BBB penetration prediction adapted from Li et al.<sup>14</sup> is presented in Table 1.

The current work aims at providing a fresh approach to virtual screening molecules regarding their BBB penetration.

First and foremost, it was clear from the beginning that there was an important aspect never considered by previous studies, namely the inclusion in the models of the likelihood of any organic molecule to cross the BBB. That is, if a molecule is picked up randomly from an organic chemistry database, how likely is it that it can cross the BBB? Recent studies<sup>2,29,30</sup> point to values of near 2% of small molecules that may be able to cross the BBB. This means that in a data set of randomly sampled 2000 molecules, the BBB-positives should be around 40 molecules while the number of BBB-negatives, totalling around 1960. Models fitted from adequately sampled data will have no problems in adapting themselves to reality, while, on the other hand, if the ratio positives/negatives diverges markedly from what is expected in a real world scenario, then the training set is defined as biased. In the available studies presented in the literature, to the present knowledge, this aspect has never been adequately addressed, and the number of BBB-positives has generally been consistently larger than the number of BBB-negatives, as can be perceived in Table 1, thus the resulting models will reflect this same bias and will show a level of inadequacy when dealing with real world scenarios where the *a priori* probabilities of BBB penetration are expected to hold. Traditional straightforward application of machine learning methods is inadequate to deal with data set bias as respective to the available knowledge of reality, and, in this study, we present a Bayesian approach that aims at improving the model fitting process by *unbiasing* the data set, making it more likely to approach reality, and also modifying the testing procedures so that the model can be adequately assessed. This *unbiasing* consists in differentially sampling the available data for building training and testing data sets that approximate what is expected for any molecule for which there is no previous knowledge.

Also, aiming for a robust and reliable model, it was found necessary to aim for the largest data set possible, and so an extensive data collection from published sources was amassed and verified for consistency. The model development followed the guidelines suggested by Tropsha,<sup>31</sup> where the available data were validated and curated, separate cross-validation testing was performed, and the final models tested with an independent validation set. To assess for spurious relationships, the best resulting model was fitted with the same data where the predicting class was y-randomized.

One final relevant aspect is that the produced model has been made available in a publicly accessible and free Web-tool both for research usage and scrutiny.

## BAYESIAN LEARNING

Bayes theorem<sup>32</sup> states that the *a posteriori* probability of a given hypothesis  $h$  verifies on the occurrence of an observation  $D$  can be written as

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (1)$$

Thus  $P(h|D)$  is a function of the probability of  $D$  holding if  $h$  is true, the probability of the hypothesis  $h$  occurring and the probability of the instance  $D$  to be observed.

To apply this theorem to the current context of molecular BBB penetration prediction, it is necessary to instantiate these variables more concretely. Let  $BBB_+$  be the set of all molecules that are able to penetrate the blood-brain barrier and, respectively,  $BBB_-$  the set of molecules that do not cross it.

Also, for a given molecule  $m$ , consider  $P(h_+|m)$  as the probability that molecule  $m$  crosses the BBB (the *a posteriori* probability), where  $h_+$  is the hypothesis of  $m$  belonging to  $BBB_+$ , thus

$$P(h_+|m) = \frac{P(mlh_+)P(h_+)}{P(m)} \quad (2)$$

and correspondingly

$$P(h_-|m) = \frac{P(mlh_-)P(h_-)}{P(m)} \quad (3)$$

According to eqs 2 and 3, the *a posteriori* probability of a model predicts whether a given molecule ( $m$ ) crosses the BBB ( $P(h_+|m)$ ) is a function of i) the *likelihood*, that is the capability of the model to predict accurately whether a molecule is able to cross the BBB ( $P(mlh_+)$ ); ii) the *a priori* probability of a molecule belonging to  $BBB_+$  ( $P(h_+)$ ); and iii) the probability that a given molecule appears ( $P(m)$ ). The latter, for all practical purposes, can be considered as a constant. Classification in a given class is given by selecting the class with the *maximum a posteriori* probability ( $h_{MAP}$ )

$$h_{MAP} \equiv \operatorname{argmax}_{h \in \{h_+, h_-\}} P(h|m) \quad (4)$$

Most models that aim at predicting BBB penetration, simplify the basic Bayes theorem by assuming that  $P(h_+) = P(h_-)$ , which eliminates these terms when comparing probabilities and leaving only the respective likelihoods ( $P(mlh_+)$  and  $P(mlh_-)$ ) which then are then equated to the *a posteriori* probabilities.

## ESTIMATING LIKELIHOODS

As defined,  $P(mlh_+)$  and  $P(mlh_-)$  represent the likelihood that a model classifies correctly a given molecule. This relates directly to the ability of the model to behave in a real world situation. An empirical approximation of these model components are the true positive ratio (*sensitivity*) and the true negative ratio (*specificity*) of the application of the model to an independent set of instances

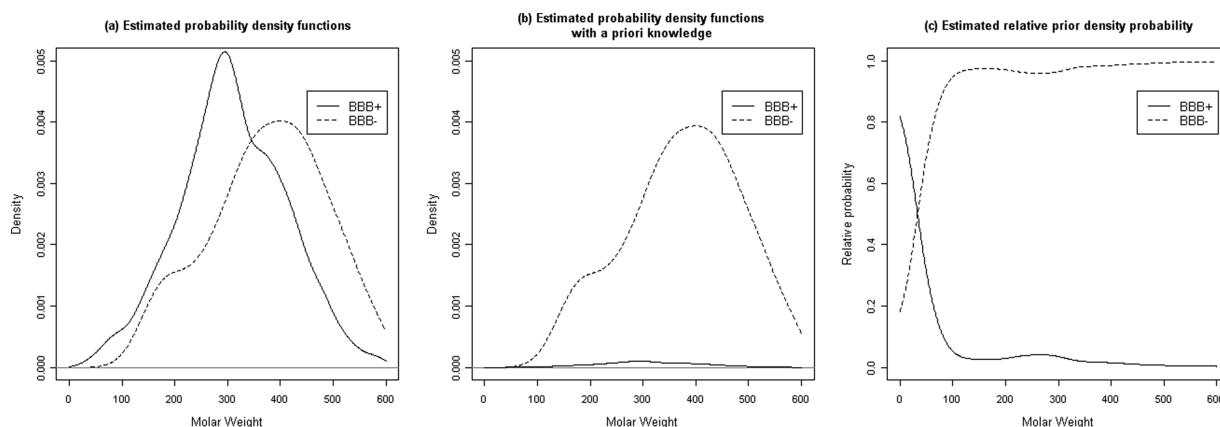
$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (6)$$

Where  $TP$  (true positives) is the total number of instances correctly classified as positive, that is molecules that belong to  $BBB_+$ ;  $TN$  (true negatives), the number of molecules that were ascribed by the model as belonging to  $BBB_-$ ;  $FP$  are the false positives, the number of molecules that were misidentified by the model as belonging to  $BBB_+$  while they are known to belong to  $BBB_-$ ;  $FN$  the false negatives, the number of molecules that indeed cross the BBB but were classified as belonging to  $BBB_-$  by the model. Thus sensitivity measures, for all the molecules that are known to be positive in a validation set, the fraction that has been correctly classified. Similarly specificity reports the fraction of known negatives that have been correctly classified.

## ESTIMATING A PRIORI PROBABILITIES

As referred, for the current problem, the assumption that  $P(h_+) = P(h_-)$  is not tenable. It is known that the probability of a



**Figure 1.** (a) Absolute probability density functions for  $BBB_+$  and  $BBB_-$  according to molecular weight; (b) absolute probability density functions scaled with *a priori* knowledge; (c) estimated relative  $BBB_+$  and  $BBB_-$  class probabilities.

given random molecule to cross the BBB is much lower than the opposite. Recent estimates point to values of 2%.<sup>2,30</sup> This fact presents difficulties not only in the assessment and evaluation of models as well as in the model building process itself. This is mainly because, for a Bayesian approach, it would be required that the probabilities of occurrence of  $BBB_+$  molecules be 49 times smaller than the fraction of  $BBB_-$  compounds.

Yet, the mentioned estimate of 2% is apparently not uniform, depending on the molecular weight (MW) of each molecule, with several studies empirically pointing to the fact that penetration rate decreases with an increase of MW.<sup>20,25</sup> Smaller molecules are in fact known to cross the BBB easier than larger ones.<sup>2,11</sup> With the current assembled data set this dependency of the penetration probability to the molecules' MW is clearly verified. Figure 1a represents the empirical probability density functions (PDFs) of  $BBB_+$  and  $BBB_-$  molecules empirically calculated with a kernel based function, over the MW domain up to 600 Da. This figure however does not take into account the known *a priori* knowledge between BBB penetration classes. Therefore, if it is assumed that the whole of  $BBB_-$  molecules correspond to 98% of all molecules (and assuming that this ratio is applied throughout the MW domain) this figure can be more adequately represented as Figure 1b, where the PDFs for both classes were adequately rescaled. However to actually have a clear view of the relative probabilities of each class across the MW range, it is possible to depict (Figure 1c) the relative importance of each class assuming for each MW that each molecule either permeates the BBB or not (as  $P(h_+|m_{MW}) + P(h_-|m_{MW}) = 1$ ). This figure is important as it shows how the *a priori* probabilities of a given compound change with the MW. Thus in fact  $P(h_+)$  is a function of MW, and the whole procedure needed to produce Figure 1c can be resumed in the following expressions

$$P(h_+) = P(h_+|m_{MW}) = \frac{P(m_{MW}|h_+)P(h_+)}{P(m_{MW})} \quad (7)$$

and

$$P(h_-) = P(h_-|m_{MW}) = \frac{P(m_{MW}|h_-)P(h_-)}{P(m_{MW})} \quad (8)$$

Accordingly, the *a priori*  $P(h_+)$  probability is in fact the probability that a compound with a given MW can cross the BBB ( $P(h_+|m_{MW})$ ). This can be calculated using Bayes theorem

with the probability distribution function of all compounds that penetrate the BBB ( $P(m_{MW}|h_+)$ ), the *a priori* knowledge of how the probability that an unknown compound may cross the BBB ( $P(h_+)$ ) and the probability of occurrence of a given MW ( $P(m_{MW})$ ). These latter values are identical for  $BBB_-$  and  $BBB_+$  molecules and thus can be discarded.

## ■ BLOOD-BRAIN BARRIER PENETRATION MODELING

Building a machine learning model that takes into account Bayesian logic is counterintuitive if the available data set for training and testing appears to have a strong bias when compared to what is actually expected in the real world chemical space, as is the present case. An unbiased data set should approximate the population statistics and could be produced by randomly sampling from a pool of molecules of unknown BBB penetration. The current data set, with a stronger component of  $BBB_+$  molecules, is clearly biased. The importance of an unbiased sample cannot be underestimated in prediction modeling. If the training data set does not represent the population, the fitted model also will not be adequate and will be biased in the same direction of the training set.

To solve this problem, an approach was followed that aimed to produce an unbiased training data set from the currently biased data set. The basic approach followed for model fitting used a differential sampling according to the *a priori* probabilities of each molecule to belong to  $BBB_+$  or  $BBB_-$ :

- 1 From the full data set  $S$  randomly select a training set  $T$  of arbitrary size;
- 2 Use the molecules of training set  $T$  to compute the prior probability density function (PDF) according to molecular weight ( $P(m_{MW}|h_+)$  and  $P(m_{MW}|h_-)$ ) where  $m \in T$ ;
- 3 Using the molecules in  $T$ , proceed iteratively until a user-specified number of instances ( $N$ ) for a resampled new training set ( $T'$ ) is selected:
  - a From  $T$  select randomly with replacement one instance and check its observed class  $\theta$ :  $\theta = (+)$  or  $\theta = (-)$  for  $BBB_+$  or  $BBB_-$  respectively;
  - b According to the MW of the compound get its appropriate likelihood ( $P(m_{MW}|h_\theta)$ ) from the generated prior PDF;
  - c Calculate  $P(h_\theta|P(m_{MW}))$  according to eqs 7 and 8;



- d Generate a random number  $r$  and if  $r \leq P(h_+|m_{MW})$ , update set  $T'$  adding the instance selected in (a).

This procedure will produce a training set that will mirror more adequately the real world molecule space, as each molecule will be selected according to its a priori probability. The above method is further able to produce arbitrarily large training sets as each molecule may be selected more than once because the sampling process is constructed with replacement. One inevitable consequence of this differential selection procedure is that if the size ( $N$ ) of the resampled training set ( $T'$ ) is not large enough, some molecules may never be selected for model building.

## ■ MODEL EVALUATION

**Estimation of Classification Quality.** To evaluate the results of a classification model, several commonly used statistics are extant, where *accuracy* is the most used, indicating the overall performance of the model

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

This statistic however is quite unsuitable for unbalanced data sets as it will account similarly the contribution of each class. For instance in a classification problem with 2 classes A and B, where A contributes to 90% of the data, a trivial model that classifies everything as A will reach an accuracy of 90%. A more robust statistic is the *mean square contingency coefficient*<sup>33</sup> (also known as Matthews correlation coefficient), or  $\phi$

$$\phi = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (10)$$

This coefficient is more robust than simple accuracy to unbalanced data sets. The previous example will produce a  $\phi$  of 0.0, which shows the actual inability of the model for classification, indistinct from random guessing.

Other statistics include sensitivity and specificity (eqs 5 and 6) as well as *precision*

$$\text{precision} = \frac{TP}{TP + FP} \quad (11)$$

Sensitivity will assess the model's ability to identify the positive instances, while specificity measures the correctly identified negatives. Precision, on the other hand, is a measure of the quality of positive identification, that is for each positive identified, how many negatives were wrongly classified as positives.

A good model has to score high in all described coefficients. The  $\phi$  statistic provides a good overall metric, but to correctly identify a problem with the model it is necessary to check sensitivity, specificity, and precision.

**Model Assessment.** A classification model when applied to a given random molecule will provide an estimate of whether it belongs to  $BBB_+$  or to  $BBB_-$ , however, similarly to model fitting, if the validation data set does not reflect the a priori knowledge, as is the present case, the statistics for a model may become biased. As an illustrative example suppose that for a given model the following were verified to hold on an independent validation set of 400 molecules evenly distributed among both classes:  $TP = 180$ ,  $TN = 160$ ,  $FP = 40$ , and  $FN = 20$ . Applying the above-defined statistics, the accuracy of the

model will be 0.85, with a  $\phi$  score of 0.70, sensitivity of 0.90, and specificity reaching 0.80, while precision will reach a value of 0.82, showing apparently a very convincing model. However, if this same classifier is applied to a real population where it is known that only a fraction of the molecules cross the BBB, the basic model statistics must be scaled according to the a priori knowledge to assess meaningful results. One way is to scale the validation set by attributing to each molecule a representation adequate to its class, which actually is  $P(h_+)$ . Thus if  $V_+$  is the number of  $BBB_+$  in the validation set, an estimate of the real representativity of  $BBB_+$  compounds in the validation set ( $R_+$ ) can be estimated as

$$R_+ = V_+ P(h_+) \quad (12)$$

and similarly for  $BBB_-$

$$R_- = V_- P(h_-) \quad (13)$$

Expanding on the previous example and, as before, assuming that  $P(h_+) = 0.02$  and  $P(h_-) = 0.98$ , and calculating  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  in the same way - as the sum of the priors of the instances of the respective classes - the very same model will yield the same specificity and sensitivity, but the overall accuracy will decrease to 0.80, the  $\phi$  coefficient falls to 0.24, and precision will now equal 0.08. These latter values imply a model with difficult acceptance. As an example, for every molecule correctly predicted as positive, over 12 molecules will be incorrectly identified as positives. This is what is to be expected if the model is applied in real world conditions, where the above assumption is assumed to hold, and is therefore the more adequate way to assess the model quality. However, this simplified approach must be refined if the priors are not constant for all classes, as is the present case. Thus, as the priors have been found to depend on each molecule molecular weight, expressions 12 and 13 can be rewritten as

$$R_+ = \sum_{m \in BBB_+} P(h_+|m_{MW}) \quad (14)$$

and respectively

$$R_- = \sum_{m \in BBB_-} P(h_-|m_{MW}) \quad (15)$$

Which define the equivalent representativity of  $BBB_+$  and  $BBB_-$  according to each molecule MW as defined by eqs 7 and 8. To assess a model when validating it with a new set of molecules the a posteriori probabilities  $P(h_+|m_{MW})$  for each molecule must be computed beforehand according to the known class of each molecule and process the respective sums for calculating  $TP$ ,  $TN$ ,  $FP$ , and  $FN$ . For instance, if a known positive molecule with  $P(h_+|m_{MW}) = 0.035$  is classified correctly as  $BBB_+$  it will add 0.035 to  $TP$ . On the other hand if it is classified as a  $BBB_-$ , the same amount will be added to  $FN$ , thus accounting for the relative rarity of  $BBB_+$  compounds. Similarly a  $BBB_-$  molecule with a  $P(h_-|m_{MW}) = 0.910$  if classified as  $BBB_+$  will increase  $FP$  with 0.910, therefore weighting considerably the effect of a wrong attribution. Due to the Bayesian approach followed, each molecule in the testing set will then account for its actual representativity in the BBB chemical space as determined by its prior rather than its raw classification value.

## MACHINE LEARNING METHODS FOR BBB PENETRATION PREDICTION

For the present study, two machine-learning models were selected, namely support vector machines (SVM) and random forests (RF). These models have been proven to be reliable, capable of finding robust solutions even when the number of variables is large, and able to adapt to a wide range of situations. Both random forests and support vector machines are among the most used algorithms in recent virtual screening and QSAR studies, and both have been described in the literature for the prediction of BBB penetration.

### SUPPORT VECTOR MACHINES

Support Vector Machines are now an ubiquitous machine learning algorithm used for a variety of problems. Burbidge et al.<sup>34</sup> published one of the first studies that featured SVM tested in SAR problems, and this methodology proved superior to other machine learning tools, either in results and computational efficiency. Other works<sup>4,13,14,25</sup> have described the application of SVM to the issue of BBB penetration prediction modeling, similarly with consistent good results.

Some of its unique characteristics include the capability of handling a very large number of descriptor variables with minimal overfitting (as it is often the problem with other methodologies like neural networks). Also, differently from other methodologies based on heuristic optimization methods, SVM are based on the solution of a convex quadratic programming problem, for which it is guaranteed to reach a minimum solution, which is deemed to be unique. The foundation of SVM is the discovery of instances in the data (the support vectors) which are able to maximize the separation between classes (when in a classification problem) according to a mathematical transformation of the variable space through a kernel function applied to the support vectors. Kernel functions are usually linear, polynomial, radial, or sigmoid, and generally machine learning libraries provide implementations to all.

### RANDOM FORESTS

Differently from SVM, Random Forests are an *ensemble classification method*. Ensemble methods are based on the iterative application of a simple classification algorithm over a randomly defined subset of the data and use a *consensus* voting procedure for determining the outcome of its application. Random Forests use as a basic classification algorithm, simple decision trees fitted to a randomly sampled set of instances and variables. Similarly to SVM, RF have been used in BBB penetration modeling (e.g., Zhang et al.,<sup>4</sup> Svetnik et al.,<sup>23</sup> and Muehlbacher et al.<sup>28</sup>).

The basic process of random forest building can be summarily described in the following sequence of steps. The process is repeated once for every iteration ( $i = 1..N$ ), according to a value specified by the user ( $N$ ). One iteration will produce a simple decision tree from a set of variables and instances. For each fitted tree a distinct set of variables and instances is used. From the training data set, a bootstrapping procedure is run, a set of instances ( $\gamma_i$ ) is randomly selected with replacement, with size equal to the training set. Also a small subset of independent variables are randomly selected from all the available variables ( $\Delta_i$ ). Then a decision tree model  $DT_i = f(\gamma_i, \Delta_i)$  is fitted to  $\gamma_i$  and  $\Delta_i$ . The set of all decision tree models ( $DT_i$ , where  $i = 1..N$ ) is a random forest. Using it for prediction implies running all trees with a new data set and producing a consensus result

from the classification outcome of the individual decision trees. Random forests allow natively for an out-of-bag validation, that is, each tree is validated with the instances that were not selected for its training, and global consensus statistic can be produced. Nevertheless, out-of-bag validation was not used due to the differential sampling applied, which increases the likelihood of instances being repeated to appear in both the training and out-of-the-bag validation sets. Therefore the more stringent 5-fold validation is used, which also permits a more objective comparison to SVM results.

### DATA SET AND DESCRIPTOR SETS

For the present work, we compiled a data set of 2053 molecules selected from a number of publications discussing BBB penetration.<sup>4,5,13,14</sup> Of these, only 1970 were used for modeling purposes as all compounds that exceeded a molecular weight of 600 Da were excluded. Most studies divide molecules as being able or not to cross the BBB, that is belonging to  $BBB_+$  or  $BBB_-$ , respectively. When the blood-brain penetration partition ( $\log BB$ ) is available, molecules were divided into  $BBB_+$  and  $BBB_-$  classes if  $\log BB \geq -1$  and  $\log BB < -1$ , respectively.<sup>4,5,14</sup>

In total there are 1570  $BBB_+$  and 483  $BBB_-$  molecules, which are given in Table S1 of the Supporting Information. The data set includes 312 molecules (172  $BBB_+$  and 140  $BBB_-$ ) from Doniger et al.,<sup>13</sup> 316 molecules (202  $BBB_+$  and 114  $BBB_-$ ) from Li et al.,<sup>14</sup> 100 molecules (91  $BBB_+$  and 9  $BBB_-$ ) from Zhang et al.,<sup>4</sup> and 1325 molecules (1105  $BBB_+$  and 220  $BBB_-$ ) from Zhao et al.<sup>5</sup>

The resulting data set was manually curated to remove generic name and SMILES duplicates and to treat some of the found ambiguities, e.g. Ribavirin, an antiviral drug with broad spectrum, which it is ineffective against viral encephalitis because it fails to cross the BBB,<sup>35</sup> though in another study<sup>14</sup> the same molecule is described as capable of entering the CNS. The final data set used for chemical descriptor generation includes a self-generated alphanumeric ID, the generic name as referred in the literature, the binary classification ( $p$  ( $BBB_+$ ) or  $n$  ( $BBB_-$ )), and the respective SMILES<sup>36</sup> representation.

The structural molecular information was obtained using the Chemical Identifier Resolver,<sup>37</sup> an online service, provided by the National Cancer Institute.<sup>38</sup> This service allows to convert a given structure identifier, e.g., molecule generic name, into another representation, for example as a SMILES string, that can later be read and processed by chemical processing software. However, some of the molecules were not recognized by the service, and, when available, the original SMILES provided by the source reference were used. Molecules where no recognizable name was found nor with a valid SMILES string were eliminated. Furthermore molecules where contradicting data were found were also eliminated. The full data set used for this work is made available in the Supporting Information.

### MOLECULAR DESCRIPTORS

Molecular descriptors have been used as input for statistical studies as a source to make assumptions in very different studies that treat the permeability of BBB to drug molecules.<sup>14</sup> This is a key issue in virtual screening studies, where it is fundamental to use relevant descriptors for prediction. Four different descriptor sets were tested in several combinations, in order to assess their relative importance.

- **A. Daylight Fingerprints** - The simple molecular fragments as calculated by the FP2 Daylight fingerprints<sup>39</sup> were used as descriptors. Daylight fingerprints work by decomposing molecules into atomic chains that may reach up to seven atoms. Each chain is then hashed using a special algorithm and coded as a binary string (a fingerprint) with 1021 positions or bits. As a descriptor set, each bit was assumed to correspond to a distinct chemical descriptor. This descriptor set was used as it is easily calculated and was conjectured that, as it is frequently used for assessing chemical similarity, whether this binary representation might be able to contribute to classification models.
- **B. Atomic and Ring Multiplicities** - This set of descriptors includes the number of occurrences of each atom according to element and number of bonds to non-hydrogen atoms. The bond order is not considered. Also within this descriptor set are the average molecular weight, the number of rings in a molecule, and the number of bonds in rings. This descriptor set totals 28 different parameters.
- **C. Simple Molecular Parameters** - In this set are included the number of single, double, triple, and aromatic bonds, as well as chemical descriptors usually found in BBB penetration prediction, namely the calculated logarithm of octanol/water partition (ClogP<sup>40</sup>), molecular radius, and topological polar surface area (TPSA).<sup>41</sup> All these parameters were computed directly with the *OpenBabel-PyBel*<sup>42</sup> libraries which in turn use *joelib2*,<sup>43</sup> for computation.
- **D. e-Dragon** - The Web application E-Dragon 1.0<sup>44,45</sup> was used to compute a set of chemical descriptors commonly used in QSAR analysis. These include constitutional descriptors, walk and path counts, information indices, edge adjacency indices, topological charge indices, randic molecular profiles, radial distribution function descriptors, weighted holistic invariant molecular (WHIM) indices, functional group counts, charge descriptors, topological descriptors, connectivity indices, 2D autocorrelations, eigenvalue based indices, geometrical descriptors, 3D MoRSE descriptors, GET-AWAY (GEometry, Topology, and Atom-Weights assembly) descriptors, and atom centered fragments as well as 31 other molecular properties.<sup>46</sup> The e-Dragon version used computes 1666 chemical descriptors.

All data sets included the molecular weight which was found fundamental for sampling data for the constitution of training sets.

## RESULTS AND DISCUSSION

**Data Processing.** From the whole data set of 1970 molecules a set of 120 randomly selected molecules was withdrawn for constituting an independent validation set to be used in the final phase, after all the model selection procedures.

The selected molecules were processed using the library *PyBel*<sup>42,43</sup> to compute the descriptors of sets A, B, and C, while parameter set D was computed using the e-Dragon Web application.<sup>44,45</sup> For the latter case, as there is a limitation on the number of molecules processed per batch, the whole data set was divided into small batches which were later reassembled. Molecules were submitted to the e-Dragon Web site in their respective SMILES representation. However, of the

remaining data set of 1970 molecules some molecules, exactly 100 could not be processed by e-Dragon. Thus the absolute results presented for this descriptor set, albeit similar, cannot be directly compared to the other results.

For inference and statistical processing, the R language and environment was used. Specifically library *randomForest*<sup>47</sup> was used for Random Forests while *e1071*<sup>48</sup> was used for SVM fitting, while this latter package is an interface to *libsvm*,<sup>49</sup> one of the most widely used SVM implementations.

The full statistical processing for model fitting validation and testing was run on a Dell server with an Intel Xenon processor with 4 cores running at 3.0 GHz, with 8 GB of RAM.

**Model Screening and Selection.** The model selection process aimed at analyzing the validation results of a set of variables not only identifying the best descriptor set, not only capable of reaching the best classification results but also verifying whether the molecular blood-brain barrier penetration is better classified with a single nonlinear model (SVM) or an ensemble method (RF). One further parameter was considered for analysis, the *training sample factor* (TSF). This parameter proved to be a necessity because the molecules' differential sampling process, according to their BBB penetration properties and MW, implies that not every molecule may be selected for training, and some might appear multiple times within each training set. One way to increase the probability of entering as many unique molecules into the model is by considering this factor as a model parameter, which may be different from the fixed  $TSF = 1.0$ , common in bootstrapping methods. A  $TSF = 1.0$  means that the actual training set ( $T'$ ) will have a size equal to the original molecules set selected for training ( $T$ ). Thus  $TSF$  values were selected from the set  $[0.5, 1.0, 2.0, 4.0]$ . Values above 4.0, for the available data, proved to be too computationally intensive for the software and hardware platforms available and were not tested.

Taking into account the above considerations, an exhaustive search procedure was performed where several combinations were tested and the results analyzed conjointly. The whole screening procedure was tested using 5-fold cross-validation. That is, the whole data set is divided randomly into 5 equally sized subsets, 4 of which are then used for training and the fifth used for validation. The process is repeated 4 times by changing the subset selected for validation. Both SVM and RF were tested with the following parameter set combinations: A, A+B, A+C, A+B+C, B+C, and D. As to assess the best configuration for RF, just the size of the forest was tested, and models were adjusted with 500, 1000, and 2000 trees. For SVM, all 4 kernel functions were tested (linear, radial, polynomial of degree 3, and sigmoidal). Results for the full cross-validations and the synthetic analysis table are presented in the Supporting Information.

**Random Forests.** For RF, out of the projected 72 combinations of descriptor sets, training sample factors and model size, 5 were not tested due to computational constraints. Therefore it was not possible to run any of the RF models with 2000 trees, neither the model with a train sample factor of 4.0 nor 1000 trees, for the same descriptor set. These models required more RAM beyond the test machine's capacity.

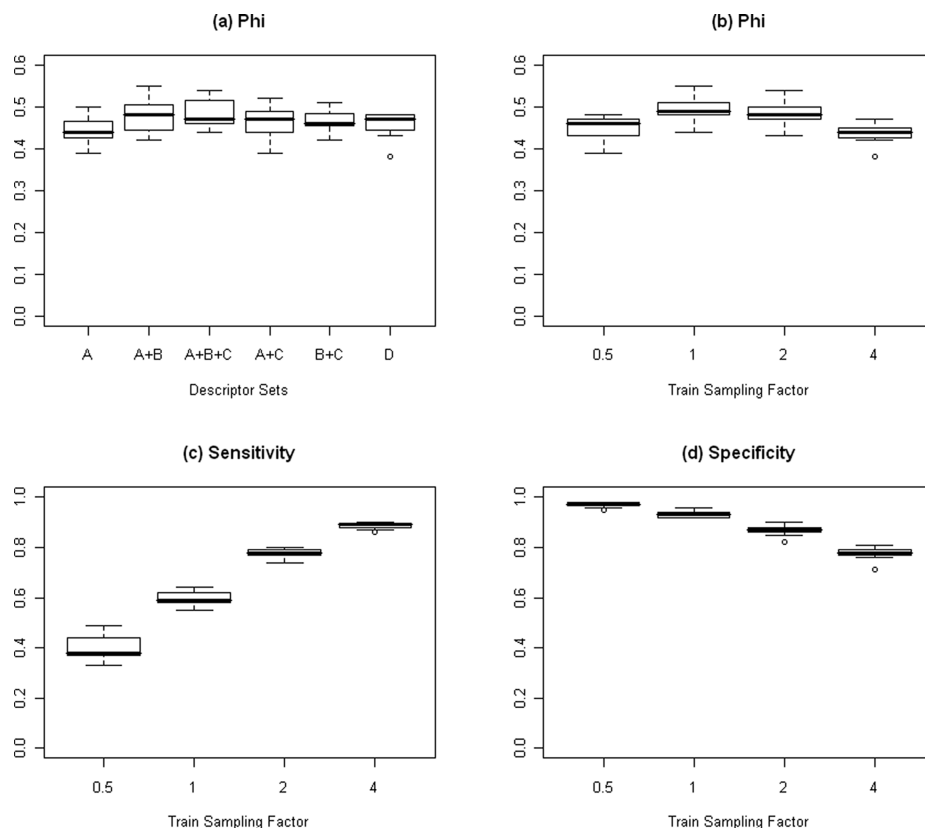
It was found that the number of trees in a random forest does not significantly impact the statistics results. On the other hand, the descriptor set and the TSF seem to be markedly important. It is clear that either descriptor set A+B or A+B+C produce the better overall results. Despite the fact that descriptor A+B is able to reach better  $\phi$  values, it is not as consistent as descriptor



Table 2. Classification Statistics for 5-Fold Cross-Validation Results of the Best 5 Random Forest Models<sup>a</sup>

descriptor set	N. trees	TSF	TP	TN	FP	FN	accuracy	$\phi$	sensitivity	specificity	precision
A+B	2000	1	24.6	370.8	15.5	17.7	0.922	0.554	0.581	0.960	0.613
A+B+C	1000	2	33.2	346.5	40.9	8.4	0.885	0.542	0.798	0.894	0.448
A+B	500	1	26.2	365.0	21.4	15.9	0.913	0.537	0.623	0.945	0.551
A+B+C	500	1	24.6	368.9	18.4	17.3	0.917	0.532	0.586	0.952	0.571
A+B+C	2000	2	33.1	343.7	43.7	8.4	0.878	0.528	0.797	0.887	0.431

<sup>a</sup>TSF - training sample factor; TP, true positives; TN, true negatives; FP - false positives; FN - false negatives.  $\phi$  - mean square contingency coefficient.



**Figure 2.** Random Forest 5-fold cross-validation results for all combinations of model size (number of trees), descriptor sets, and training sample factors. (a) Box plots of the mean square contingency coefficient ( $\phi$ ) for all 6 descriptor set combinations; (b)  $\phi$  ranges across the training sample factor (TSF); (c) and (d) effects of the TSF on sensitivity and specificity, respectively.

set A+B+C, for which no model was below 0.45. Differently from the descriptor set used for model fitting, the TSF appears as a particularly relevant one. Clearly both TSF values of 0.5 or 4.0 achieve subpar  $\phi$  results, and the differences between 1.0 and 2.0 are not significant. The reasons why the extremes produce poor overall results is due to the influence they show both on sensitivity and specificity. Lower TSF values emphasize the correct classification of  $BBB_-$  compounds (high specificity) while, on the other hand, reduce substantially the ability to correctly identify  $BBB_+$  molecules (low sensitivity). The opposite is verified for TSF = 4.0. This makes sense as the increase in the number of molecules on the training sets increase the a priori low representation of  $BBB_+$  compounds and thus makes the model more capable to correctly identify these molecules. On the other hand, as the number of  $BBB_-$  compounds increases at the expense of numerous repetitions of the same instance, the model ability for extrapolate outside the training set is reduced thus causing low specificity scores.

Analyzing the best absolute models as sorted by  $\phi$  (Table 2), it is perceptible that out of the 67 model combinations tested

and validated with 5-fold cross-validation, the 5 best models used parameter sets A+B or A+B+C. No trend was observed regarding the adequate number of trees required; however, as expected, the TSF ranged between 1.0 and 2.0 for all cases. The best model used descriptor sets A+B, using 2000 trees and a TSF of 1.0, reaching a  $\phi$  value of 0.55, but with a sensitivity value low, reaching 0.58. As mentioned, this statistic gives an estimate of the ratio of identification of  $BBB_+$  molecules, and this score is clearly low. The second best model, with only a slightly lower  $\phi$  score presents a sensitivity of 0.80 and specificity of 0.89. The precision of 0.45 is however a bit lower. This latter model nonetheless appears to be more adequate for usage in real-world applications due to its better sensitivity and adequate specificity (see Figure 2).

**Support Vector Machines.** SVM are generally faster to fit, although increasing the number of training samples places an important computational burden, and it was not possible to compute the larger instances of descriptor set D.

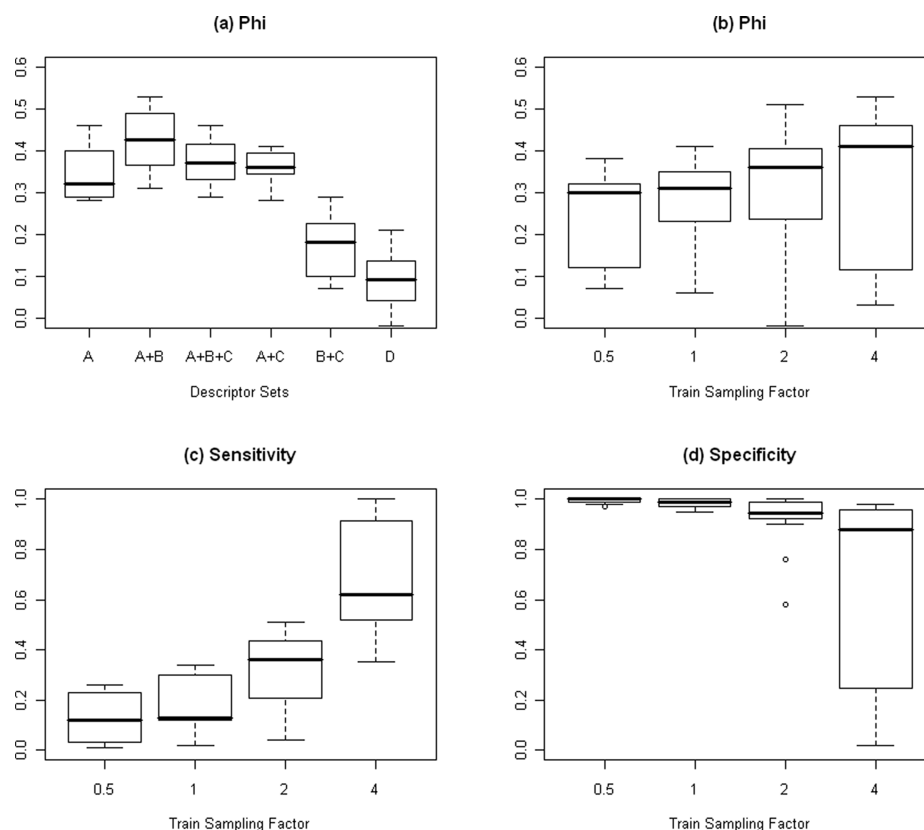
Support vector machine models were tested with the same descriptor sets and ranges of sampling factors, the sole



Table 3. Classification Statistics for 5-Fold Cross-Validation Results of the Best 5 Support Vector Machines<sup>a</sup>

descriptor set	kernel function	TSF	TP	TN	FP	FN	accuracy	$\phi$	sensitivity	specificity	precision
A+B	radial	4	22.2	372.8	13.5	20.5	0.921	0.526	0.520	0.965	0.621
A+B	polynomial	2	16.0	381.4	4.8	26.3	0.928	0.508	0.378	0.988	0.769
A+B	sigmoid	4	21.5	371.9	14.5	21.0	0.917	0.505	0.506	0.963	0.597
A+B	polynomial	4	22.1	369.9	16.5	20.0	0.915	0.501	0.524	0.957	0.572
A+B	radial	2	14.9	381.5	4.8	27.7	0.924	0.481	0.349	0.988	0.755

<sup>a</sup>TSF - training sample factor; TP, true positives; TN, true negatives; FP - false positives; FN - false negatives.  $\phi$  - mean square contingency coefficient.



**Figure 3.** Support vector machines 5-fold cross-validation results for all combinations of kernel function, descriptor sets, and training sample factors. (a) Box plots of the mean square contingency coefficient ( $\phi$ ) for all 6 descriptor set combinations; (b)  $\phi$  ranges across the training sample factor (TSF); (c) and (d) effects of the TSF on sensitivity and specificity, respectively.

Table 4. Results of Validation of the Best Model Fitted with All the Available Data with an Independent Validation Set

	TP	TN	FP	FN	accuracy	$\phi$	sensitivity	specificity	precision
Y-randomization	17.6	6.9	1.7	0.2	0.676	-0.092	0.123	0.034	0.080
best model	23.6	1.0	0.5	2.4	0.947	0.737	0.826	0.712	0.800

difference was that instead of testing the importance of the model size (number of trees in a RF model), the four available kernel functions were tested (see Table 3).

SVM also showed no marked difference between algorithm configurations. On SVM, different kernel functions were tested and although radial kernels appeared to show slightly better results none showed to be markedly superior to the others. On the other hand, the  $\phi$  statistic proved to be sensitive to the different descriptor sets used. Yet, the same trend as verified for RF was verified with descriptor sets A+B+C and A+B getting the top scores. Here it is even more obvious the importance of including descriptor set A, as descriptor sets B+C and D show poor results. The TSF also shows its importance on the results of SVM models with a similar trend to RF. Increasing the TSF

increases the sensitivity and generically decreases the specificity, but this trend is less marked than for RF. It is also clear that SVM are better able to cope with a larger number of instances, as the best results are reached with TSF values of 4.0 (see Figure 3).

**Model Validation.** According to the above results, the selected model was Random Forests, fitted with descriptor set A+B+C, thus the 1021 Daylight fingerprints plus the atomic multiplicities and the simple molecular parameters. Random forests were fitted with 1000 trees using a TSF factor of 2.0.

To assess the model validity the same model was tested with the same validation set, yet the training set was y-Randomized,<sup>31</sup> by scrambling among training instances each actual class, and fit an equivalent model with the same

B<sub>3</sub>PP

Version 0.9

Blood Brain Barrier Penetration Prediction

**Input Name:** 3-(2-Aminoethyl)-1H-indol-5-ol

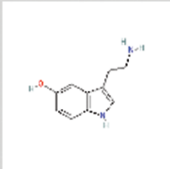
**Other Names:** 50-67-9, BPBio1\_001079, NCGC00142449-01, Bio1\_001428, KBioSS\_002479, BSPBio\_000341, Biomol-NT\_000083, Bio2\_000876, Prestwick2\_000481, KBio3\_002950


**Molecular Weight:** 176.21508

**Molecular Formula:** C<sub>10</sub>H<sub>12</sub>N<sub>2</sub>O

**P (BBB+ | MW):** 0.0261

**P (BBB- | MW):** 0.9739





The compound *serotonin* should not be able to cross the Blood Brain Barrier.

Enter a compound

(Common name, SMILES or InChI)

serotonin

Name

Predict

Copyright LaSIGE 2012 - About the B<sub>3</sub>PP

Figure 4. B3PP Web application example, showing the output for running a sample molecule (serotonin).

parameters. For the latter case, results were very poor (Table 4), which suggests that the model is actually finding meaningful relationships between the descriptor sets and the molecules BBB classes. Thus a final model was fitted with the same parameters and descriptor sets and tested with the independent validation set selected earlier. Results are indicative that the model is indeed capable of learning, and the results produced appear consistent with the values produced in the screening phase which is a further guarantee of the model reliability (Table 4).

**Web Tool.** With the selected model, a public and free Web tool (B3PP - Blood Brain Barrier Penetration Prediction) has been developed. The objective was to produce an application simple to use with easily readable results. When starting B3PP, the user can input any molecule using its common name, SMILES string, or InChI identifier. The common name is resolved using the Chemical Identifier Resolver<sup>37</sup> (CIR), directly called by the application. The output produced consists in the model output, some other common names, when available, and a graphical representation of the molecule, also generated by the CIR. Also provided, for completeness, is the a priori probabilities of crossing the BBB according to the molecular weight ( $P(\text{BBB}_+, m_{\text{MW}})$  and  $P(\text{BBB}_-, m_{\text{MW}})$ ). This application is able to predict the BBB penetration class for molecules with a MW below 600 Da.

B3PP was mainly developed in the PHP programming language. The application communicates with a Python script that uses PyBel for computing the required descriptors of the input molecule for model usage. These are then passed to a R script that loads the model and produces the model result, which is finally processed by the PHP front end for output to the user (see Figure 4).

B3PP is available at <http://b3pp.lasige.di.fc.ul.pt>

## CONCLUSIONS

In this study, we have gathered a large data set with known BBB penetration data and tested two state-of-the-art machine learning models, namely random forests and support vector machines, for compound classification. We have taken into account the a priori knowledge of BBB penetration, to derive a more robust model applicable to a real world prescreening situation. This requirement forced important changes in the processes of data selection for model fitting, model testing, and model validation. Furthermore several descriptor sets were tested as well as different combinations of the same.

The resulting model was achieved by a thorough process where different model conformations were tested and validated through a 5-fold cross-validation. With the current data the model that showed better performance used random forests with 1000 trees, and requiring as descriptors the 1021 Daylight fingerprints, simple molecular parameters and atom-bond multiplicities. It was actually quite surprising to find that, in particular, the Daylight fingerprints were fundamental for all the best models both for support vector machines and random forests. The best model conformation (a random forest) was then fitted to the whole available data and validated with an independent validation set separated from the beginning from any process of model screening. Results with this latter validation set suggest that the fitted model can be an important tool usable in a real-world scenario providing researchers with valuable insight in the early phases of drug development and general drug research. The model developed is further made available to the whole community as a public and free Web tool. A Web service capable of batch processing for large data sets is currently under development. By running the application

it is possible to verify the strengths and weaknesses of the proposed approach by testing known compounds.

Although it is believed that the present work is a valid contribution to the field of *in silico* prediction of BBB penetration, the resulting model could provide better results if the data available were not so biased, namely by including a larger number of BBB-negatives. This would allow a less stringent selection procedure for the BBB-positives and include a larger number of different molecules in each training iteration. In fact, what is now missing is not so much empirical evidence on molecules that can cross the BBB but rather a larger number of molecules for which it is known that are not able to cross it.

Finally, despite the fact that, due the methodology used, the achieved results are not directly comparable to other efforts described in the literature, it can be seen that end model is of comparable quality to the best published efforts. This suggests that even with limited data, strongly biased against what is known to be the reality of molecular BBB penetration, it is possible to infer good quality models that may be useful to the whole community, if proper measures are taken to guarantee appropriate model fitting and validation procedures.

## ■ ASSOCIATED CONTENT

### ■ Supporting Information

Full data set used for this work (Table S1) and also the complete cross validation statistics for Tables 2 and 3 and Figures 2 and 3. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [afalcao@di.fc.ul.pt](mailto:afalcao@di.fc.ul.pt).

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors thank the Portuguese Fundacao para a Ciencia e a Tecnologia for the Multiannual Funding Programme of the LaSIGE laboratory and PhD grant SFRH/BD/64487/2009.

## ■ REFERENCES

- (1) *Introduction to the Blood-Brain Barrier: Methodology, biology and pathology*; Pardridge, W. M., Ed.; Cambridge University Press: 1998.
- (2) Pardridge, W. M. The Blood-Brain Barrier: Bottleneck in Brain Drug Development. *NeuroRx* **2005**, *2*, 3–14.
- (3) King, A. Breaking through the barrier. *Chem. World* **2011**, *8*, 36–39.
- (4) Zhang, L.; Zhu, H.; Oprea, T.; Golbraikh, A.; Tropsha, A. QSAR Modeling of the Blood-Brain Barrier Permeability for Diverse Organic Compounds. *Pharm. Res.* **2008**, *25*, 1902–1914.
- (5) Zhao, Y. H.; Abraham, M. H.; Ibrahim, A.; Fish, P. V.; Cole, S.; Lewis, M. L.; de Groot, M. J.; Reynolds, D. P. Predicting Penetration Across the Blood-Brain Barrier from Simple Descriptors and Fragmentation Schemes. *J. Chem. Inf. Model.* **2007**, *47*, 170–175.
- (6) Di, L.; Kerns, E. H.; Fan, K.; McConnell, O. J.; Carter, G. T. High throughput artificial membrane permeability assay for blood-brain barrier. *Eur. J. Med. Chem.* **2003**, *38*, 223–232.
- (7) Lu, J. A novel hypothesis of blood-brain barrier (BBB) development and in vitro BBB model: neural stem cell is the driver of BBB formation and maintenance. *J. Exp. Integr. Med.* **2012**, *2*, 39–43.
- (8) Cucullo, L.; Aumayr, B.; Rapp, E.; Janigro, D. Drug delivery and in vitro models of the blood-brain barrier. *Curr. Opin. Drug Discovery Dev.* **2005**, *8*, 89–99.
- (9) Goodwin, J. T.; Clark, D. E. In Silico Predictions of Blood-Brain Barrier Penetration: Considerations to “Keep in Mind”. *J. Pharmacol. Exp. Ther.* **2005**, *315*, 477–483.
- (10) Ekins, S.; Tropsha, A. A Turning Point For Blood-Brain Barrier Modeling. *Pharm. Res.* **2009**, *26*, 1283–1284.
- (11) Banks, W. A. Characteristics of compounds that cross the blood-brain barrier. *BMC Neurol.* **2009**, *9* (Suppl1), S3.
- (12) Ooms, F.; Weber, P.; Carrupt, P.-A.; Testa, B. A simple model to predict blood-brain barrier permeation from 3D molecular fields. *Biochim. Biophys. Acta* **2002**, *1587*, 118–125.
- (13) Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. *J. Comput. Biol.* **2000**, *9*, 849–864.
- (14) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Cao, Z. W.; Chen, Y. Z. Effect of Selection of Molecular Descriptors on the Prediction of Blood–Brain Barrier Penetrating and Nonpenetrating Agents by Statistical Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 1376–1384.
- (15) Ajay; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942–4951.
- (16) Adenot, M.; Lahana, R. Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 239–248.
- (17) Crivori, P.; Cruciani, G.; Carrupt, P.-A.; Testa, B. Predicting Blood-Brain Barrier Permeation from Three-Dimensional Molecular Structure. *J. Med. Chem.* **2000**, *43*, 2204–2216.
- (18) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11* (Supplement2), S29–S39.
- (19) Iyer, M.; Mishra, R.; Han, Y.; Hopfinger, A. J. Predicting Blood-Brain Barrier Partitioning of Organic Molecules Using Membrane-Interaction QSAR Analysis. *Pharm. Res.* **2002**, *19*, 1611–1621.
- (20) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137–2152.
- (21) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. J. Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D-Molecular Similarity Measures and Cluster Analysis. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2083–2098.
- (22) Gerebtzoff, G.; Seelig, A. In Silico Prediction of Blood-Brain Barrier Permeation Using the Calculated Molecular Cross-Sectional Area as Main Parameter. *J. Chem. Inf. Model.* **2006**, *46*, 2638–2650, PMID: 17125204.
- (23) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
- (24) Guerra, A.; Páez, J.; Campillo, N. Artificial Neural Networks in ADMET Modeling: Prediction of Blood–Brain Barrier Permeation. *QSAR Comb. Sci.* **2008**, *27*, S86–S94.
- (25) Kortagere, S.; Chekmarev, D.; Welsh, W.; Ekins, S. New Predictive Models for Blood-Brain Barrier Permeability of Drug-like Molecules. *Pharm. Res.* **2008**, *25*, 1836–1845.
- (26) Wang, Z.; Yan, A.; Yuan, Q. Classification of Blood-Brain Barrier Permeation by Kohonen’s Self-Organizing Neural Network (KohNN) and Support Vector Machine (SVM). *QSAR Comb. Sci.* **2009**, *28*, 989–994.
- (27) Fan, Y.; Unwalla, R.; Denny, R. A.; Di, L.; Kerns, E. H.; Diller, D. J.; Humblet, C. Insights for Predicting Blood-Brain Barrier Penetration of CNS Targeted Molecules Using QSPR Approaches. *J. Chem. Inf. Model.* **2010**, *50*, 1123–1133.
- (28) Muehlbacher, M.; Spitzer, G.; Liedl, K.; Kornhuber, J. Qualitative prediction of blood-brain barrier permeability on a large and refined dataset. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 1095–1106.
- (29) Pardridge, W. M. Blood-brain barrier delivery. *Drug Discovery Today* **2007**, *12*, 54–61.

- (30) Tsaion, K.; Bottlaender, M.; Mabondzo, A. ADDME - Avoiding Drug Development Mistakes Early: central nervous system drug discovery perspective. *BMC Neurol.* **2009**, *9* (Suppl 1), S1.
- (31) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476–488.
- (32) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*; Wiley-Interscience, 2000.
- (33) Baldi, P.; Brunak, S. *Bioinformatics - the machine learning approach*, 2nd ed.; MIT Press: 2001; pp I–XXI, 1–452.
- (34) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *BMC Neurol.* **2001**, *26*, 5–14.
- (35) Jeulina, H.; Venard, V.; Carapito, D.; Finance, C.; Kedzierewicz, F. Effective ribavirin concentration in mice brain using cyclodextrin as a drug carrier: Evaluation in a measles encephalitis model. *Antiviral Res.* **2009**, *81*, 261–266.
- (36) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- (37) Chemical Identifier Resolver beta 4. <http://cactus.nci.nih.gov/chemical/structure> (accessed May 18, 2011).
- (38) NCI/CADD Group Chemoinformatics Tools and User Services. <http://cactus.nci.nih.gov/> (accessed May 18, 2011).
- (39) Daylight Fingerprints - Screening and Similarity. <http://www.daylight.com/dayhtml/doc/theory/theory.finger.html> (accessed May 18, 2011).
- (40) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
- (41) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, 3714–3717.
- (42) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2008**, *2*.
- (43) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J. K.; Willighagen, E. L. The Blue Obelisk-Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991–998.
- (44) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453–463.
- (45) Tetko, I. V. Computing chemistry on the web. *Drug Discovery Today* **2005**, *10*, 1497–1500.
- (46) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: 2009.
- (47) randomForest: Breiman and Cutler's random forests for classification and regression. <http://cran.r-project.org/web/packages/randomForest/randomForest> (accessed April 10, 2011).
- (48) e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. <http://cran.r-project.org/web/packages/e1071/e1071.pdf> (accessed February 12, 2012).
- (49) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.