Predicting a Compounds Blood-Brain-Barrier Permeability with Lantern Pharma's AI and ML Platform, RADR®

Rick Fontenot[1*], Umesh Kathad[1*], Joseph McDermott[1*], Drew Sturtevant[1], Panna Sharma[1], Peter Carr[1]

[1] Lantern Pharma Inc., Dallas, TX, United States; [*] These authors contributed equally to these works

Abstract:

This report presents methodologies and a series of machine learning generated models to predict the blood-brain-barrier (BBB) permeability for a collection of drugs in the Therapeutic Data Commons (TDC) BBB, Martins et al. benchmark group challenge. The semi-permeable BBB prevents the delivery of most drug compounds to the central nervous system (CNS) , limiting the efficacy of treatments in CNS disorders. The ability to predict which drugs are likely to pass through the BBB aids in identifying candidate treatments for disorders of the CNS. In this work, SMILES (Simplified Molecular Input Line Entry System) strings for drugs were transformed into molecular fingerprints and descriptor features along with already known important factors such as molecular weight and surface area for use in machine learning models with high accuracy in identifying a drug's ability to penetrate the BBB. After generating thousands of candidate features that could potentially be useful in predictions, different subsets of the most important features are selected for training models with logistic regression, random forest, and deep neural network methods as well as an ensemble of these base learner models. At the time of this report, all five models presented rank in the top 6 of the TDC leaderboard with area under the receiver operating characteristic (AU-ROC) ranging from 0.903 to 0.915 and accuracies ranging from 88% to 90%.

Introduction:

The blood-brain-barrier (BBB) is a highly selective brain barrier that presents challenges in the effective delivery of therapeutic compounds for the treatment of brain and central nervous system (CNS) disorders. The blood-brain-barrier (BBB) is a highly selective brain barrier that presents challenges in the effective delivery of therapeutic compounds for the treatment of brain and central nervous system (CNS) disorders. Rather than a single entity, the BBB is the combined function of properties of endothelial cells and cell adhesions that together limit the permeability of a vessel. [1] Molecules with molecular mass greater than 500 daltons do not cross the BBB, and approximately only 2% of molecules with mass under this threshold have been shown to cross the BBB. [2]

Prior work utilizing machine learning models to predict BBB permeability were created by generating molecular fingerprints as features using software packages such as PaDEL-Descriptor software [3] or DeepChem and RDKit [4] and produced results with AUC-ROC ranging from 0.849 [3] to 0.905 [4] and accuracy of 0.798 [3]. While molecular fingerprints as machine learning features have shown this predictive success, previous research also suggests that BBB permeability decreases as the molecule's surface area increases [1]. BBB permeability also decreases as hydrogen bonds are added to the structure, whether it is a hydrogen bond donor or hydrogen bond acceptor [1]. Incorporating these features for drugs in addition to the fingerprints previously shown to be valuable may provide for higher prediction accuracy.

Although BBB permeability prediction with machine learning and deep learning methods based on molecular fingerprints have demonstrated accuracy beyond short rule based criteria, expanding the number of descriptors available as candidate features may improve accuracy. The workflow approach here expands the candidate features based on insights from literature reviews, then reduces the final amount of features used in training models based on importance in feature selection algorithms. In addition to better model performance on test data, these models use fewer features and are more robust and interpretable. Interpretation of key features can improve the understanding of how specific chemical structures impact BBB permeability, and how molecule design can be improved to enhance permeability.

Methods:

Here, Simplified Molecular Input Line Entry System (SMILES) structures were used to translate a chemical's three-dimensional structure into a string of symbols that can be processed by computer software programs. The “Rdkit” python library was used to convert the SMILES drug structures into numerical features such as Morgan fingerprints , rdk fingerprints, MACCS fingerprint, and descriptors including 2D and 3D Autocorrelations as well as 3D Getaway and WHIM descriptors. [5] These binary and non-binary numerical features served as a proxy for different atomic properties including element connectivity, chemical features, bond type, atomic mass, and electrotopological state. Well established atomic rules (e.g. lipinski rules, ghose filter, veber filter etc.) and their attributes were also generated as features. These processes generated 4,552 candidate features for each drug (Table 1), which were subsequently fed into feature reduction methods prior to machine learning model generation.

Feature Generation Method	**# of Features**	**Values**
Rdk fingerprints	2048	Binary
Morgan fingerprints	2048	Binary
MACCS fingerprints	167	Binary
2D Autocorrelation descriptors	192	Continuous
3D Autocorrelation descriptors	80	Continuous
Rules/filters (6) and it’s attributes (11)	17	Mix
Getaway 3D descriptors	273	Continuous
WHIM 3D descriptors	114	Continuous
Total	4939	Mix

Table 1. Details of feature generation methods

Following the initial candidate feature generation, feature reduction was performed to reduce feature multicollinearity prior to feature selection and modeling. Tanimoto similarity was used to screen binary features such as fingerprints and pearson’s R was used to screen numeric features. For both metrics one feature was dropped in each pair of features with a correlation higher than 95%.

Next chi-square tests were performed for all individual fingerprint features to determine whether permeability is dependent on the fingerprint. Fingerprints with a *p*-value less than 0.05 were considered as significant in the first phase. In the second phase, the ratio of permeable and non-permeable samples were calculated for drug samples containing each fingerprint. Fingerprints with a permeability of 50% or lower were categorized as significantly negatively associated fingerprints, and those with 80% or greater permeability were considered significantly positively associated fingerprints. Based on these results newly engineered features were created totaling the count of negatively and positively associated fingerprints for each drug sample.

Following feature generation, reduction and engineering steps described above, features were transformed using kernel principal component analysis.

Due to the class imbalance between the majority of drug samples being permeable (75.3%) and a minority of samples being non-permeable (24.7%), Synthetic Minority Oversampling Technique (SMOTE) [6] was performed on the training set of data utilizing a k-nearest neighbor algorithm to create synthetic data for the minority class until the sample counts were balanced between permeable and non-permeable observations. This augmentation of the training data set served as the input to feature selection and model training phases of the workflow. No augmented samples were generated for the validation or test sets of data.

Candidate features were reduced using logistic regression with the least absolute shrinkage and selection operator penalty to shrink the least important feature's coefficients to zero thus eliminating from features selected for use in model training. Ten-fold cross-validations were performed on a search range for the L1 regularization parameter value, which provided the highest average accuracy across the folds. The L1 regularization value identified as optimal in the search was used to train a final feature selection model that eliminated features with a coefficient of zero and ranked remaining features in order of importance by the absolute value of their coefficient. Feature selection lists were generated for the original un-transformed features as well as the kernel PCA transformed features to evaluate the use of each feature set on each predictive modeling type.

Four base learner methods of modeling were utilized to generate diverse methods of predictions that serve as inputs to an ensemble meta-learner in a subsequent phase. The first base learner method used was a logistic regression. Previous implementation during feature selection utilized L1 regularization to reduce features. The base learner model included a further search using ten-fold cross validation to determine the optimal L2 regularization parameter values. This model used the kernel pca transformed features without oversampling augmentation.

The second base learner method used was a Deep Neural Network. The design of the neural network was generated using a search on the optimal architecture for a range of two to five fully connected dense hidden layers. Each hidden dense layer included L2 regularization and was followed by a dropout layer. The search additionally included a range of neurons used in each hidden layer, a selection of optimizers, and parameters for learning rate reductions over the course of training. Each iteration of the architecture search included early stopping criteria using a holdout subset from the training data to stop model training at the epoch which represented the highest area under the receiver operating characteristic curve (AUC-ROC) on the holdout samples. A final model was constructed with the 4 hidden layers consisting of [67, 70, 84, 88] neurons with [relu, tanh, relu, relu] activation functions,, and dropout rates of [0.029, 0.122, 0.039, 0.162] and L2 regularization rate of 1.6e-8. The model used the RMSprop optimizer with an initial learning rate of 1e-3 which is reduced by a factor of 0.5 after 3 epochs without improvement in loss. identified in the search and was fully trained up to the early stopping criteria. This model included kernel pca transformed features without augmentation at the feature selection stage but with augmentation during training.

The third base learner method used was a random forest. This base learner model included a search using ten-fold cross-validation to determine the optimal number of estimators, depth, and minimum samples per split and leaf to reduce the effects of overfitting. This model included original features without kernel pca transformation and used augmentation for both the feature selection and model training processes.

The fourth modeling method was a Support Vector Machine (SVM). This base learner model included a search using ten-fold cross-validation to determine the optimal kernel, degree, regularization strength, and kernel gamma coefficient which resulted in a final model with a linear kernel and regularization strength equal to 2. This model included kernel pca transformed features without augmentation at the feature selection stage but with augmentation during training.

After training each of the base learner models, the predicted probability of permeability was calculated on holdout validation samples that were not included in the model's training samples. These validation sample predictions were subsequently used to train an ensemble meta-learner.

The fifth modeling method was an ensemble method where validation sample predictions from the three base learner models were used as feature inputs to a logistic regression meta-learner ensemble model. All base models were evaluated as meta-learner inputs and permutations of the base-learner combinations. The combination of base-learners with the highest area under the receiver operating characteristic curve was selected as the final meta-learner. Pruning evaluation showed that the best ensemble results based on AU-ROC included the deep neural network, random forest, and SVM models while excluding the logistic regression which showed a strong correlation to the deep neural networks predictions and did not add further gains in ensembling.

Generating predictions on the test set for evaluation was completed in two phases. First, the probability of permeability for each sample was predicted for each base learner and its associated selected features. Second, the validation set base learner probabilities were used as inputs to train a logistic regression meta-learner ensemble model used to make the final predictions on the test set. BBB permeability classification labels were assigned using a threshold of greater than 0.5 predicted probability as the drug being permeable for all model types for the purposes of reporting accuracy, sensitivity and specificity scores.

Results:

The ensemble model performed best in terms of the AU-ROC and specificity by blending the diverse predictions of the deep neural network, random forest, and SVM models as base learners.

All four machine learning models generated have AUC-ROC scores ranging from 0.92-0.96, which placed them at the top of the BBB-Martins leaderboard as follows as of the date of submission.

Model	AU-ROC	Accuracy	Sensitivity	Specificity
Ensemble	0.915 ± 0.003	0.879 ± 0.007	0.915 ± 0.009	0.726 ± 0.033
Deep Neural Network (DNN)	0.912 ± 0.003	0.893 ± 0.009	0.947 ± 0.013	0.664 ± 0.033
Random Forest (RF)	0.907 ± 0.006	0.891 ± 0.007	0.955 ± 0.007	0.621 ± 0.037
SVM-linear	0.905 ± 0.008	0.898 ± 0.006	0.965 ± 0.008	0.615 ± 0.024
Logistic Regression	0.903 ± 0.002	0.896 ± 0.001	0.970 ± 0.003	0.582 ± 0.007

Table 2. Model BBB Prediction Performance

Full results including predictions on each drug are available on FigShare found at the links below:
Fontenot, Rick (2023): Lantern Pharma - TDC BBB-Martins Leaderboard. figshare. Collection.
https://doi.org/10.6084/m9.figshare.c.6491158.v1

Open source code with documentation and installation instructions to reproduce results is accessible at:
https://github.com/lanternpharma/tdc-bbb-martins

References:

1. Profaci CP, Munji RN, Pulido RS, Daneman R. The blood–brain barrier in health and disease: Important unanswered questions. J Exp Medicine. 2020;217:e20190062.
2. Pardridge WM. The blood-brain barrier: Bottleneck in brain drug development. Neurorx. 2005;2:3–14.
3. Liu L, Zhang L, Feng H, Li S, Liu M, Zhao J, et al. Prediction of the Blood–Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. Chem Res Toxicol. 2021;34:1456–67.
4. Tian H, Ketkar R, Tao P. Accurate ADMET Prediction with XGBoost. Arxiv. 2022. https://doi.org/10.48550/arxiv.2204.07532.
5. RDKit: Open-source cheminformatics. https://www.rdkit.org https://doi.org/10.5281/zenodo.7671152.
6. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57.