Use of LogBB data to model and predict small molecule Blood Brain Barrier Penetrance

Caiden Lukan, Da Shi, Ph.D., Sarangan Ravichandran, Ph.D., and Amanda Paulson, Ph.D.
[1]Butler University, [2]ATOM Consortium
2021 ATOM Consortium Summer Internship Program
West Dundee, IL

Abstract

Modelling blood brain barrier (BBB) drug penetrance is essential for the discovery of small molecules, especially since cancer drugs typically do not cross the blood brain barrier. Additionally, an associated issue is the expense that comes along with testing *in vitro*, which is usually necessary for understanding which drugs can cross the BBB. The literature that has already been established contains data with LogBB values (the log of the ratio of concentrations of a drug in the brain to the body at steady state) as well as a categorical value of penetrance. While the data from the literature are used en masse to predict whether or not a molecule crosses the BBB, there are often conflicting values between sources of how a molecule was judged to cross the BBB. The data gathered from the literature were further curated to ensure BBB penetrance was assessed in a consistent manner. After this data curation, a regression and classification model were made. The best models made were neural networks with a validation regression R^2 score of 0.61 and a classification ROC AUC score of 0.94. Additionally, feature importance was analyzed for the best models. Features of a molecule that indicate polarity are a helpful measurement in determining the permeability of that molecule. Using artificial intelligence and machine learning techniques, models with high accuracy can be made and used for prediction of BBB penetrance in the future.

Introduction

- The Blood Brain Barrier (BBB) is a barrier between the Central Nervous System and the rest of the body that prevents molecules from causing CNS damage.
- The metric that is used as a parameter to measure whether or not a molecule passes through the BBB is LogBB.

$$\log(BB) = \log\left(\frac{[Brain]}{[Blood]}\right)$$

- LogBB us best used to measure concentrations of a drug at equilibrium.
- LogBB < 0 means that a molecule does not pass the BBB.
- LogBB ≥ 0 means that a molecule passes the BBB.

Materials and Methods

- Seven journal articles were used as sources of LogBB data.
- The assays that are used to calculate logBB are Parallel Artificial Membrane Permeability Assay (PAMPA) and Immobilized Artificial Membrane technique.
- The data from the journal article was the curated in order to improve the accuracy of the models.
- From the curated data, a Regression model and Classification model were made.
- The Regression model is useful for predicting a LogBB value.
- The Classification model is useful to predict whether or not a molecule will pass through the BBB.
- Molecules were then featurized and trained with an XGboost, Neural Network, and Random Forest model.
- Parameters for the models were analyzed and improved to improve the score of the valid set of data.
- Finally, results are analyzed and the best model can be used as a standard for predicting LogBB values or whether or not a molecule will pass the BBB.

Data Curation

- Collected 2590 compounds from 7 different sources.
- Converted Compounds to base_rdkit_smiles to look for duplicate compounds.
- Removed duplicate compounds to have a final total of 1181 compounds.
- Prior to moving to the modelling step, LogBB data was analyzed by a histogram **(Figure 1)** and Penetration data was analyzed with a bar graph **(Figure 2)** to ensure a proper balance of data.
- The data of the compounds structure was also analyzed with a heat map **(Figure 3)** and tanimoto distance plot **(Figure 4)** to ensure a proper diversification of SMILES strings.

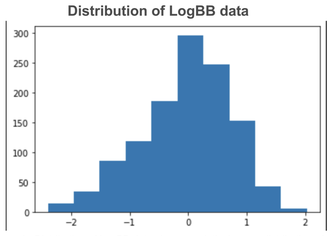

Figure 1. Distribution of LogBB data shows a relatively even distribution of values. The minimum value is -2.4 and the maximum value is 2.03.

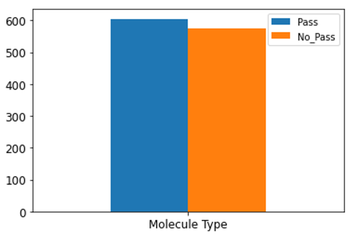

Figure 2. A graphical representation of the molecule that pass (605) and do not pass (576) the blood brain barrier

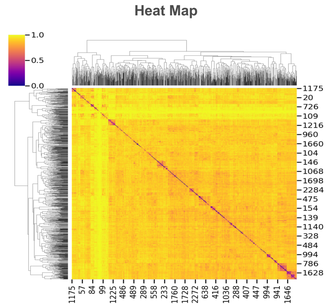

Figure 3. This Heat Map illustrates the similarities between the compounds. Note, 1 represents different and 0 represents similar. There are very few spots of purple, which means that there is a wide variety of compounds

Data Curation

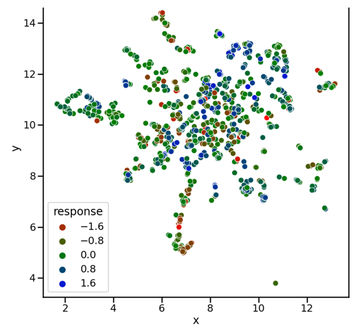

Figure 4. Distribution of LogBB based on tanimoto distance represents diversity of compounds

Modelling

- The features that were used were: MOE, Mordred, Rdkit, ECFP, Graph Convolutional Network.
- The models that were done were: Neural Network, Random Forest, and XGboost.
- The hyperparameters **(Figures 5 and 6)** are based on previous hyperparameter optimization trials.

Classification	NN	RF	xgboost
Rdkit	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [16,32,64,128,256] rfd_choice = [8,32,64,128,256] rff_choice = [4,8,16,32,48,64]	xgbG_choice = [0.05,0.1,0.15,0.2] xgbL_choice = [.4,.35,.3,.25,.2]
ECFP	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [128,192,256] rfd_choice = [32,64,128,256] rff_choice = [4,8,16,32,48,64]	xgbG_choice = [0.1,0.2,0.3] xgbL_choice = [.5,.3,.2,.1]
moe	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [128,192,256] rfd_choice = [32,64,128,256] rff_choice = [4,8,16,32,48,64]	xgbG_choice = [0.1,0.2,0.3] xgbL_choice = [.5,.3,.2,.1]
Mordred	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [128,192,256] rfd_choice = [32,64,128,256] rff_choice = [4,8,16,32,48,64]	xgbG_choice = [0.1,0.2,0.3] xgbL_choice = [.5,.3,.2,.1]
GCN	learning_rate_choice = [0.1,0.01, 0.001, 0.0001,0.00001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250,"300,"350","400"]		

Figure 5. Parameters chosen for Hyperparameter Optimization of Classification Model

Regression	NN	RF	xgboost
Rdkit	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [32,64,128,256] rfd_choice = [64,96,128,192,256] rff_choice = [4,8,16,32,64,128]	xgbG_choice = [0.0,1.0,15,0.2,0.25] xgbL_choice = [.3,.25,.2,.15,.1]
ECFP	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [32,64,128,256] rfd_choice = [32,64,128,256] rff_choice = [64,96,128,192,256]	xgbG_choice = [0.0,1.0,2.0,3] xgbL_choice = [.1,.15,.2,.25,.3]
moe	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [32,64,128,256] rfd_choice = [64,96,128,192,256] rff_choice = [4,8,16,32,64,128]	xgbG_choice = [0.0,1.0,15,0.2,0.25] xgbL_choice = [.3,.25,.2,.15,.1]
Mordred	learning_rate_choice = [0.2,0.1,0.01, 0.001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]	rfe_choice = [32,64,128,256] rfd_choice = [64,96,128,192,256] rff_choice = [4,8,16,32,64,128]	xgbG_choice = [0.0,1.0,15,0.2,0.25] xgbL_choice = [.3,.25,.2,.15,.1]
GCN	learning_rate_choice = [0.2,0.1,0.01, 0.001,0.0001] layer_sizes_choice = ["500,250","400,200","300,150","200,100","100,50"] dropouts_choice = ["0.1,0.1","0.2,0.2","0.3,0.3"] max_epochs_choice = ["100","150","200,"250","300"]		

Figure 6. Parameters chosen for Hyperparameter Optimization of Regression Model

Modelling

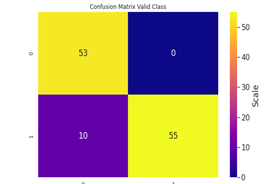

Figure 7. Confusion Matrix Valid. AUC_ROC = 0.94

Figure 8. Confusion Matrix Test. AUC_ROC = 0.83

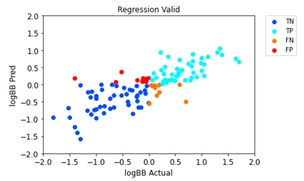

Figure 9. Regression Valid. r2 = 0.61. ROC_AUC based on threshold of 0 = .83

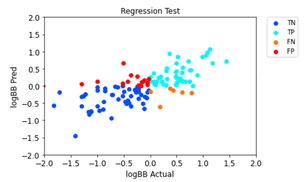

Figure 10. Regression Test. r2 = 0.43. ROC_AUC based on threshold of 0 = .82

Classification	NN	RF	xgboost
Rdkit	learning_rate_choice = .01 layer_sizes_choice =750,500 dropouts_choice = .3,.3 max_epochs_choice = 300 SCORE = .943	rfe_choice =64 rfd_choice = 64 rff_choice =8 SCORE = .91	xgbG_choice = .1 xgbL_choice = .25 SCORE = .89
ECFP	learning_rate_choice = .01 layer_sizes_choice = 400,200 dropouts_choice = .3,.3 max_epochs_choice = 300 SCORE = .903	rfe_choice = 128 rfd_choice = 128 rff_choice = 8 SCORE = .90	xgbG_choice = .1 xgbL_choice = .5 SCORE = .84
moe	learning_rate_choice = .001 layer_sizes_choice = 200,100 dropouts_choice = .2,.2 max_epochs_choice = 250 SCORE = .935	rfe_choice = 1218 rfd_choice = 32 rff_choice =8 SCORE = .91	xgbG_choice = .1 xgbL_choice = .2 SCORE = .928
Mordred	learning_rate_choice = .001 layer_sizes_choice = 300,150 dropouts_choice = .3,.3 max_epochs_choice = 100 SCORE = .915	rfe_choice = 128 rfd_choice = 64 rff_choice = 64 SCORE = .924	xgbG_choice = .2 xgbL_choice = .3 SCORE = .927
GCN	learning_rate_choice = .001 layer_sizes_choice = 100,50 dropouts_choice = .1,.1 max_epochs_choice =400 SCORE = .944		

Figure 11. Table of best results for Classification models

Regression	NN	RF	xgboost
Rdkit	learning_rate_choice = .1 layer_sizes_choice = 100,50 dropouts_choice = .2,.2 max_epochs_choice =300 SCORE = .612	rfe_choice = 256 rfd_choice = 64 rff_choice =16 SCORE = .56	xgbG_choice = 0 xgbL_choice = .3 SCORE = .54
ECFP	learning_rate_choice = .01 layer_sizes_choice = 300.150 dropouts_choice = .3,.3 max_epochs_choice = 300 SCORE = .49	rfe_choice = 256 rfd_choice = 64 rff_choice = 96 SCORE = .41	xgbG_choice = 0 xgbL_choice = .2 SCORE = .43
moe	learning_rate_choice = .01 layer_sizes_choice = 100,50 dropouts_choice = .1,.1 max_epochs_choice = 300 SCORE = .616	rfe_choice = 32 rfd_choice = 192 rff_choice = 4 SCORE = .523	xgbG_choice = .15 xgbL_choice = .2 SCORE = .53
Mordred	learning_rate_choice = .001 layer_sizes_choice =500,250 dropouts_choice = .2,.2 max_epochs_choice = 100 SCORE = .612	rfe_choice = 32 rfd_choice = 64 rff_choice = 64 SCORE = .605	xgbG_choice = 0 xgbL_choice = .1 SCORE = .57
GCN	learning_rate_choice = .0001 layer_sizes_choice = 100,50 dropouts_choice = .1,.1 max_epochs_choice = 300 SCORE = .54		

Figure 12. Table of best results for Regression models

Data Exploration

- Feature importance of the classification and regression model is important in understanding what part of the molecule has the greatest impact on the outcome.
- The way that feature importance can be accomplished is via shuffling the data. “If we, with our shuffle, break a strong relationship will compromise what our model has learned during training, resulting in higher errors (high error = high importance)”.

X_1	X_2	X_...	y
X_1a	X_2a	:	y_1
X_1b	X_2b	:	y_2
X_1c	X_2c	:	y_3
:	:	:	:
X_1z	X_2z	X_....	y_n

X_1	X_2	X_...	y
X_1f	X_2a	:	y_1?
X_1p	X_2b	:	y_2?
X_1r	X_2c	:	y_3?
:	:	:	:
X_1b	X_2z	X_....	y_n?

X_1	X_2	X_...	y
X_1a	X_2g	:	y_1?
X_1b	X_2e	:	y_2?
X_1c	X_2s	:	y_3?
:	:	:	:
X_1z	X_2q	X_....	y_n?

Figure 13. Steps on how feature shuffling is done. The first dataframe is unchanged. The second dataframe shows the first feature being shuffled while the second columns remains as it was in the initial dataframe and the result column changing. The third dataframe shows the second feature being shuffled while the first feature remains as it was in the initial dataframe and the result column changing. This process is repeated for as many features are in the featurizer.

Data Exploration

- Comparing the regression and classification model with the shuffling method **(Figure 13)**, the most important features were Qed and PEOE_VSA6 respectively.
- A change in PEOE (Partial Equalization of Orbital Electronegativities) would change the polarity of the molecule. This means that a change in polarity has an effect on the molecules ability to penetrate the BBB.
- A change in Qed (Quantitative estimation of drug likeness) is similar to changing parameters in Lipinski’s rule of 5. This means that a change in those properties has a great effect on the molecules ability to penetrate the BBB.

Conclusion

- Molecules that can cross the BBB have properties such as being non-polar based on a feature importance algorithm.
- Neural networks show the best results for all feature types.
- BBB penetrance modelling is important for future drug development regardless of whether or not the purpose of the drug is to cross.

Future Aims

- After completing hyperparameter optimization, the models chosen for both classification **(Figures 7 and 8)** and regression **(Figures 9 and 10)** are neural networks featurized by rdkit raw descriptors.
- The molecules that were predicted incorrectly in the classification model **(Figures 7 and 8)** had LogBB values close to 0.
- Best results for all models are listed in **Figures 11 and 12.**

References

- Radchenko EV, Dyabina AS, Palyulin VA. Towards Deep Neural Network Models for the Prediction of the Blood–Brain Barrier Permeability for Diverse Organic Compounds. *Molecules*. 2020; 25(24):5901. https://doi.org/10.3390/molecules25245901
- Shin HY, Lee S, Oh H, Yoo S, Kim W, Kang M. Development of blood brain barrier permeation prediction models for organic and inorganic biocidal active substances. *Chemosphere*. 2021; 277: 130330. https://doi.org/10.1016/j.chemosphere.2021.130330
- Vilar S, Chakrabarti M, Costanzi S. Prediction of passive blood-brain partitioning: Straightforward and effective classification models based on in silico derived physicochemical descriptors. *J. Mol. Graph.* 2010; 28(8): 899. https://doi.org/10.1016/j.jmgm.2010.03.010
- Zhang L, Zhu H, Oprea TI, Golbraikh A, Tropsha A. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm Res*. 2008;25(8):1902-1914. doi:10.1007/s11095-008-9609-0
- Plisson F, Piggott AM. Predicting Blood–Brain Barrier Permeability of Marine-Derived Kinase Inhibitors Using Ensemble Classifiers Reveals Potential Hits for Neurodegenerative Disorders. *Marine Drugs*. 2019; 17(2):81. https://doi.org/10.3390/md17020081
- Zhu, L., Zhao, J., Zhang, Y. *et al.* ADME properties evaluation in drug discovery: in silico prediction of blood–brain partitioning.*Mol Divers* 22, 979–990 (2018). https://doi.org/10.1007/s11030-018-9866-8
- Cerliani M. Feature Importance with Neural Network. Medium. https://towardsdatascience.com/feature-importance-with-neural-network-346eb6205743. Published January 21, 2020. Accessed July 13, 2021.

Acknowledgements

I would like to thank Amanda Paulson for directing me through this project and her excellent teaching skills. Next, I would like to thank Da Shi for helping me with resolving coding troubles that I encountered during the time he was with ATOM. Finally, I would like to thank Sarangan Ravichandran (Ravi) for answering additional questions of mine about docking and other steps of the process of modelling.

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

All animals used in this research project were cared for and used humanely according to the following policies: the *U.S. Public Health Service Policy on Humane Care and Use of Animals* (2000); the *Guide for the Care and Use of Laboratory Animals* (1996); and the *U.S. Government Principles for Utilization and Care of Vertebrate Animals Used in Testing, Research, and Training* (1985). All Frederick National Laboratory animal facilities and the animal program are accredited by the Association for Assessment and Accreditation of Laboratory Animal Care International.