Development of Semi-Automated Workflow to Predict ADMET Endpoints via Support Vector Regression

congruence TX

Poster # :

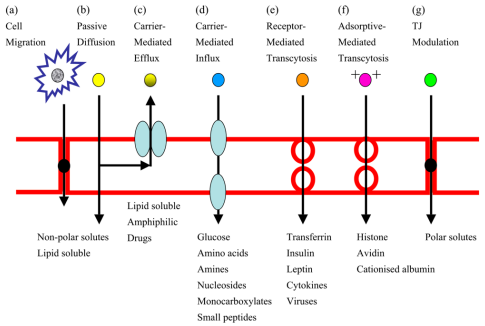
By: Behrgen Smith, Rocky Goldsmith and Maximilian Ebert

Abstract

The presence of the blood-brain barrier (BBB) presents a constant challenge in optimizing the delivery of novel therapeutics to the central nervous system (CNS); with many potential therapeutics never making it to clinical trials due to poor BBB permeability (BBBp). In this study, we aimed to develop a workflow to generate a Support Vector Regression (SVR) model to quantitatively predict LogBBBp using a curated dataset of 595 molecules, the largest dataset used to train a regression model to date. Key features of the workflow include filtering of molecules that are volatile or have unconstrained chiral centers, and procedural exploration of descriptor space via forward feature selection while keeping model complexity low. Both descriptor generation and chiral filtering were performed in MOE. The resulting SVR was able to predict LogBBBp to an R^2 of 0.54 on the test set and an R^2 of 0.667 on a smaller, external set . The results of this study show that the SVR performs relatively well compared to previous quantitative models trained on smaller datasets and offers a workflow to create similar models for other endpoints.

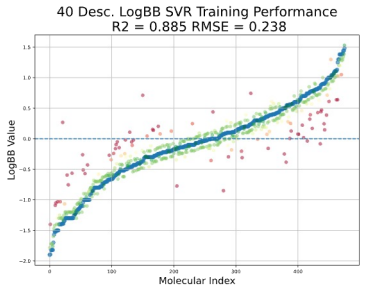
Background

BBBP is a critical parameter for the development of small molecule therapeutics that target diseases of the CNS. The development of therapeutics that target proteins in the brain is difficult because of the complexity of the BBB, as it is not only an anatomical barrier, but also consists of transport and metabolic barriers, such as paracellular permeability and the expression of multi-specific efflux transporters. There are a variety of pathways that a molecule can take to pass into the brain with varying efficacy, which results in a diverse set of molecules that are BBB penetrant.



Therapeutics that target the CNS have less than half the success rate than those that target more accessible areas of the body (i.e. cardiovascular), and with clinical experiments on BBBp being time consuming and expensive, there is a need for accurate models to predict the endpoint. Historically, most ML models that target BBBp fit into two categories: classification models trained on large datasets (>1000 molecules) with high accuracy, and regression models trained on small datasets (50-300 molecules) with average performance.

Training Performance

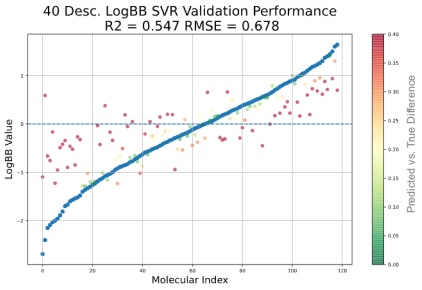


Training Set Notes

The training set was 80% of the dataset, selected by taking a diverse subset of the LogBBB endpoint.

Given the size of the training set and relatively few number of descriptors (<10% size of the training dataset), the model fit well. There is a clear fit with some outliers, but the plot doesn't immediately indicate overfitting.

Validation Performance

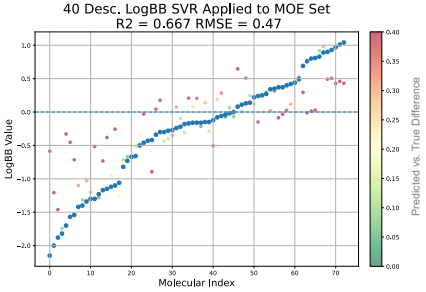


Validation Set Notes

The validation set was 20% of the dataset, selected by taking a diverse subset of the LogBBB endpoint.

While an R^2 of 0.547 isn't exceptional, inspection of the graph reveals a clear correlation, with the exception of a handful of outliers. The drop in performance when evaluated on the training set could be attributed to the diversity of the dataset, and not necessarily underfitting of the model.

External Set Performance



External Set Notes

The resulting model was evaluated on the blood_brain dataset provided in MOE. It consists of a diverse set of drug-like compounds as well as some volatiles.

The model performed well, and a clear correlation is observed. The higher performance on the external set than the validation could be attributed to its smaller size; however, the validation set may have had a particularly diverse set of molecules the model had not seen before.

Methodology

Dataset Curation

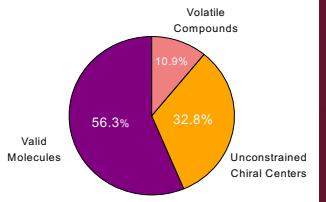
The dataset used for this experiment was obtained from the B3DB database, which had 1056 entries with continuous LogBB values. Molecules were filtered based on two criteria to increase the model's robustness:

Volatility:

- *It is common for LogBB datasets to have volatile compounds (i.e., inhalants), but these are not drug-like molecules and skew the model.*

Unconstrained Chiral Centers:

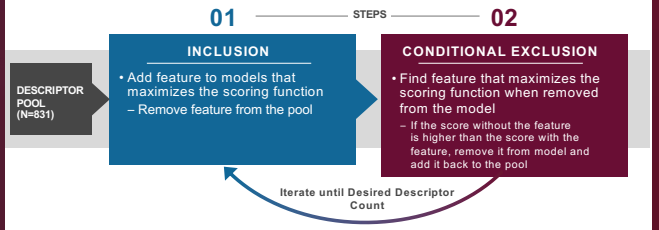
- *While chirality of molecules isn't necessarily a key consideration in predicting LogBB, unconstrained chiral centers can lead to inconsistent descriptor calculations and call the quality of data into question.*



Dataset Composition

Feature Selection

Given that MOE offers hundreds of molecular descriptors it is necessary to evaluate which combination of descriptors leads to the highest predictive power. However, due to the number of combinations of descriptors being a factorial function, performing an exhaustive test is not practical. Additionally, it is ideal to minimize the number of descriptors relative to the model performance, reducing the complexity of the model while increasing its interpretability.



Hyperparameter Optimization

Support Vector Regression Machines have 3 hyperparameters available for optimization : C, γ and ε. C controls the trade-off between the complexity of the model and the degree to which deviations larger than epsilon are tolerated. A large value of C results in a more complex model that fits the training data closely and can lead to overfitting, conversely a small value of C can lead to underfitting. γ determines the influence of a single training example on the model, specifically the shape of the kernel when transforming the input data into a higher dimension. High values of γ lead to a more complex decision boundary and can lead to overfitting, low values of γ lead a simpler boundary which can generalize well. The ε parameter controls the size of the region where errors are not penalized. Larger values of ε allow more datapoints to be in the insensitive zone and leads to a more generalized model that is less sensitive to noise in the training data.

The 3 hyperparameters were optimized by performing a grid search of the parameters with a linear space of 50 points between 0.1 and 50 for C and γ, and 25 points between 0.1 and 0.25 for ε. The optimal combination that maximized performance on the training set was 2.139, 0.071 and 1.641 for C, ε and γ.

References

Alsenan et al., 2020; Begley, 2004; Castillo-Garit et al., 2017; Ciura et al., 2020; Di et al., 2013; Gabathuler, 2010; Jiang et al., 2016; Kortagere et al., 2008; Kumar et al., 2022; Miao et al., 2019; Muehlbacher et al., 2011; Patel, 2020; Pedregosa et al., 2011; Radan et al., 2022; Radchenko et al., 2020; Shaker et al., 2021; Tong et al., 2022; Toropov et al., 2017; Wang et al., 2018; Yuan et al., 2018; Zhang, 2006

Congruence Therapeutics
congruencetx.com

Abbreviations: BBB, Blood Brain Barrier; BBBp, Blood Brain Barrier Permeability; ML, Machine Learning; SVR, Support Vector Regression