Predictive modeling of potential substrates and inhibitors of BCRP in the blood-brain barrier

Kendra Schorr[1], Da Shi, Ph.D.[2], Sarangan Ravichandran, Ph.D.[2], and Amanda Paulson, Ph.D.[2]
[1]Butler University, Indianapolis, IN [2]ATOM Consortium, San Francisco, CA
2021 ATOM Consortium Summer Internship Program

Abstract

Poly(ADP-ribose) Polymerase Inhibitors (PARPi) are a new class of chemotherapy drugs used to treat cancer, yet are unable to treat brain cancer. The Breast Cancer Resistance Protein (BCRP) is a multi-drug resistance efflux transporter in the blood-brain barrier that could be responsible for conferring resistance in the brain to newly-discovered PARP inhibitors. Our goal is to make the drug discovery process for PARPi more efficient by using machine learning models to screen drug molecules as possible BCRP substrates or inhibitors. To do this, data was sourced from ChEMBL, DTC, Excape, and journal articles. The data was curated into regression or classification datasets, and then further subdivided into transport or inhibition data. The regression model for transport data was not trained because it did not have enough data values to build a model accurately. The data was split using a scaffold split and featurized using either ECFP, or computed descriptors (RDKit, Mordred, MOE). Models were trained using random forest, neural networks including graph convolutional neural networks, and XGBoost. The classification models for inhibition and transportation had best ROC AUC scores of 0.9634 and 0.9209, respectively. The regression model for inhibition had a best R^2 score of 0.4024. Future directions include further analyzing the regression data to determine why its model trained poorly and finding more data to work with.

Introduction

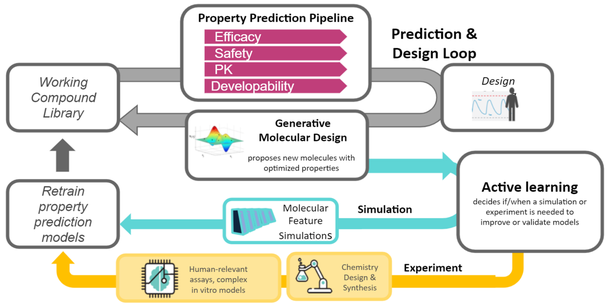

Figure 1: **ATOM Platform**
The prediction and design loop is utilized to build predictive models for the activity or inhibition of BCRP based on molecules. This project falls within the Efficacy portion of the Property Prediction Pipeline.

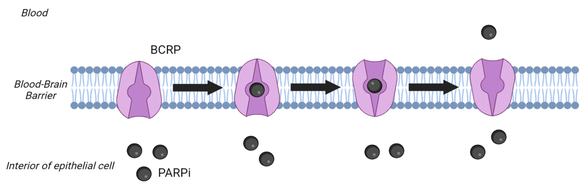

Figure 2: **BCRP Mechanism of Action**
The BRCP efflux pump transports drugs, such as PARP inhibitors, out of the brain, thereby conferring drug resistance. This project builds machine learning models in order to better understand the molecules that BCRP transports and is inhibited by, so that we can apply that to future potential PARP inhibitors.

Hypothesis

Predictive modeling of BCRP will help stream-line the process of creating potential PARP inhibitors for the treatment of brain cancer.

Methods

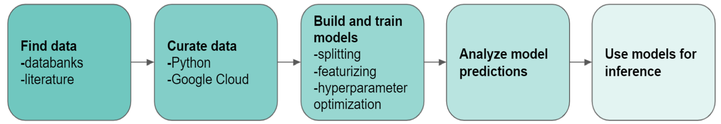

Figure 3: **Diagram of Project Workflow**
There were 5 main steps to this project, and they are outlined in this diagram.

Results

Source	Number of Values
Chembl	4035
DTC	1938
Excape	26514
ADMET	2800
Human	4789
Sel	4576
Total	44,652

Figure 4: **Starting Data**
Data was found from 3 different databases (ChEMBL, DTC, Excape) and 3 different journal articles (ADMET, Human, Sel).

Final Data Frame	Number of Values
BCRP_reg_tr	157
BCRP_reg_in	785
BCRP_cla_tr	766
BCRP_cla_in	814
Total	2,522

Figure 5: **Results from Data Curation**
The data was curated and concatenated into 4 different dataframes. The data was separated into classification or regression data and then transportation or inhibition data. The BCRP_reg_tr dataframe was not analyzed because it did not have enough values.

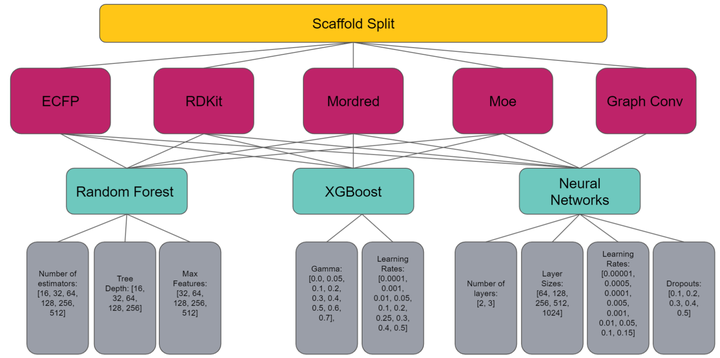

Figure 6: **Model Training Diagram**
As the figure shows, each model was split with scaffold split, then featurized with either ECFP, RDKit, Mordred, Graph Conv, or Moe. Then, the individual models were trained with Random Forest, XGBoost, or Neural Networks. Grid Search was used for hyperparameter optimization, and the options used are shown in the gray boxes.

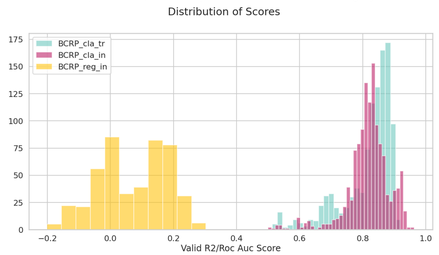

Figure 7: **Distribution of R^2 and ROC AUC Scores**
The R^2 scores from the regression model training were significantly worse than the Roc Auc scores from classification model training. We went back to the curated data to analyze what was causing this issue.

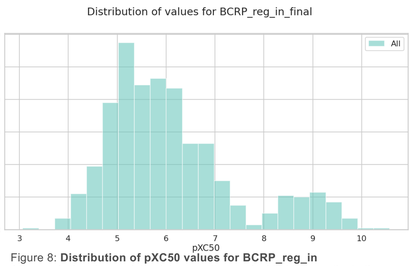

Figure 8: **Distribution of pXC50 values for BCRP_reg_in**
After analyzing the model training results, we decided to investigate the BCRP_reg_in data more. The bimodal distribution of the pXC50 values was thought to be the cause of the poor R^2 values. To do this, we subset the regression/inhibition data into pXC50 values <7.5 and >7.5.

Results

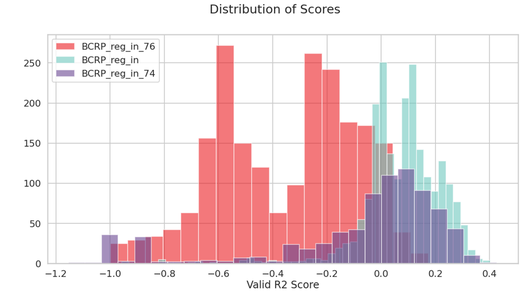

Figure 9: **Distribution of R^2 Scores for Original and Subset Data Frames**
After we subset the data, we trained models using both new dataframes. The data with pXC50 values < 7.5 was named "BCRP_reg_in_74, and the data with pXC50 values > 7.5 was named "BCRP_reg_in_76. As the graph shows, there was no improvement in the R^2 scores for the subset data frames relative to the original data frame. Therefore, the cause for the poor regression results could not be determined.

Conclusions

	Classification		Regression
	Inhibition	Transportation	Inhibition
Featurizer	RDKit	Mordred	Mordred
Model	Neural Networks	Neural Networks	Neural Networks
Roc Auc Score/R^2 score	0.9634	0.9209	0.4024

Figure 10: **Best Models Table**
This table shows the best models and their R^2 and Roc Auc results from model training. The classification models trained significantly better than the regression model.

Future Directions

- Study the regression data more
  - What is the cause of the bimodal distribution values?
  - Why did the models train so poorly, if it was not for the bimodal distribution?
- Find more data
  - Having more data would hopefully train more accurate models
  - More data would allow for the regression/transport data to be trained

References

Bioactivities for ABCG2. Drug Target Commons. Accessed June 9, 2021. https://drugtargetcommons.fimm.fi/bioactivities?id=DTCT0027880&category=Target&name=ABCG2
Excape Chemogenomics Database. Excape. Accessed June 9, 2021. https://solr.ideaconsult.net/search/excape/
Jiang D, Lei T, Wang Z, Shen C, Cao D, Hou T. ADMET evaluation in drug discovery. 20. Prediction of breast cancer resistance protein inhibition through machine learning. *J Cheminform*. 2020;12(1):16. Published 2020 Mar 5. doi:10.1186/s13321-020-00421-y
Sedykh A, Fourches D, Duan J, et al. Human intestinal transporter database: QSAR modeling and virtual profiling of drug uptake, efflux and interactions. *Pharm Res*. 2013;30(4):996-1007. doi:10.1007/s11095-012-0935-x
Shaikh N, Sharma M, Garg P. Selective Fusion of Heterogeneous Classifiers for Predicting Substrates of Membrane Transporters. *J Chem Inf Model*. 2017;57(3):594-607. doi:10.1021/acs.jcim.6b00508
Target Report Card. ChEMBL. 2018. Accessed June 9, 2021. https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL5393/
What to Know about PARP Inhibitors. Medical News Today. Accessed July 6, 2021. https://www.medicalnewstoday.com/articles/parp-inhibitor