Prediction of P-gp Efflux within the Blood Brain Barrier Impacting Molecule Transport or Inhibition

Sarah Abu-Salih,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

Fishers, IN

Abstract

PARP-inhibitors are compounds that inhibit the enzyme Poly (ADP-ribose) polymerase and are increasingly becoming utilized in chemotherapeutic regimens in order to prevent the repair of damaged DNA. Specific to the ATOM Consortium, understanding Pgp-efflux is important for the development of PARP-inhibitors. P-glycoprotein, or P-gp, functions as a transmembrane efflux pump and plays a significant role in the uptake and efflux of a range of drugs. The consequence of Pgp-efflux is that it prohibits entry of anticancer drugs and prevents the ability of the compounds to reach their targets within the brain. When trying to identify what compounds are susceptible to Pgp-efflux, in-vitro models are time consuming and costly. Utilization of in-silico models reduces the time and cost associated with previous techniques. Modeling Pgp-efflux is crucial within the discovery of potential small molecules to be used in chemotherapeutic treatments. Published data on P-glycoprotein inhibitions and transport was acquired from public databases and used to build predictive models. Through employment of classification and regression models, we are able to predict which compounds are able to bypass P-gp and enter the brain in order to inhibit PARP. The highest performing models for, both, classification and regression models were random forest models. Implementation of machine learning techniques and artificial intelligence allow researchers the elimination of in-vitro models, and provides them with high-accuracy models that can predict whether a compound will be delivered to the brain, or, if it will be effluxed via Pgp.

Introduction

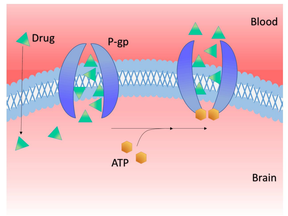
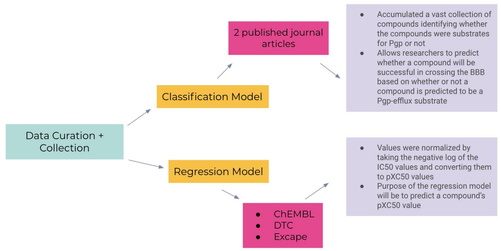


Figure 1. www.ScienceDirect.com

- P-glycoprotein is responsible for the uptake and efflux of many compounds.
- This efflux occurs when drugs are transported from the plasma to the blood brain barrier and are met with P-gp.
- A huge challenge for researchers today when creating new anti-therapeutic drugs is identifying whether the compound will be able to pass the BBB or not.
- This can be resolved through utilization of prediction models which help researchers better understand P-gp's behavior.

Materials and Methods



Data Curation

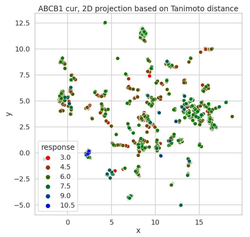


Figure 2. Distribution of compound diversity based on Tanimoto distance is represented by the above figure.

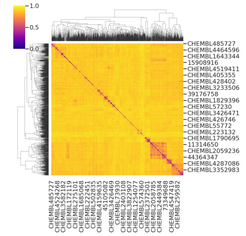


Figure 3. The Heat Map demonstrates the diversity within the compounds curated for the models. A value of 1 means the compounds are different, while a value of 0 means the compounds are similar. As the figure shows, the compounds are dominantly diverse, however, the purple square in the lower right hand corner represents a cluster of similar compounds.

Classification Data

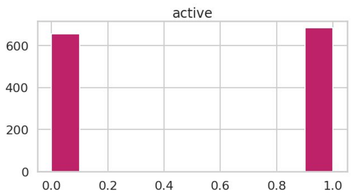


Figure 4. The data above represents data curated for Classification and Inhibition model. A value of 0 means the compound was not an inhibitor of Pgp, and a value of 1 means the compound was an inhibitor of Pgp. An even distribution of inhibitors vs non-inhibitors was utilized to create the Classification and Inhibition model.

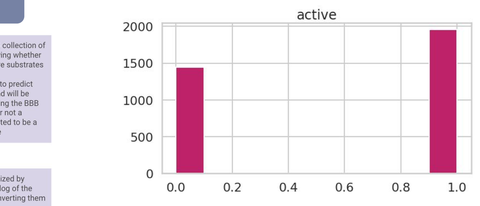


Figure 5. The data used for the Classification and Transportation model was also fairly equally distributed, A value of 0 meant the compound was a non-substrate of Pgp, while a value of 1 means the compound was a substrate for Pgp.

Regression Data

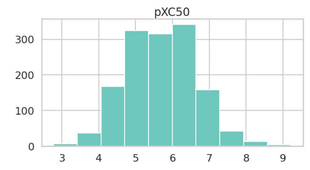


Figure 6. The figure above represents the pXC50 values for the Regression and Inhibition model.

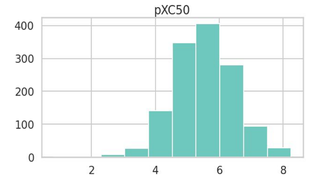


Figure 7. The figure above represents the pXC50 values curated for Regression and Transportation model.

Modeling

Regression Model Parameters

Feature	NN	Random Forest	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 = [4,8,16,32,48,64,128]	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 = [0.1,0.2,0.3] 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 = [64,96,128,256] 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 = [4,8,16,32,48,64,128]	xgbG_choice = [0.0,1.,15,0.2,0.25] xgbL_choice = [.3,.25,.2,.15,.1]
GCN	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"]		

Figure 8. Parameters chosen for hyperparameter optimization of Regression models.

Classification Model Parameters

Feature	NN	Random Forest	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 = [16,32,64,128,256] rfd_choice = [8,32,64,128,256] 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 = [16,32,64,128,256] rfd_choice = [8,32,64,128,256] 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 = [16,32,64,128,256] rfd_choice = [8,32,64,128,256] 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 = [0.2,0.1,0.01,0.001] 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", "350", "400"]		

Figure 9. Parameters chosen for hyperparameter optimization of Classification models.

Classification Inhibition Results

- **Best model performance for Classification + Inhibition:**
  - Feature: RDKit
  - Model: Random Forest
  - valid_roc_auc_score: 0.974

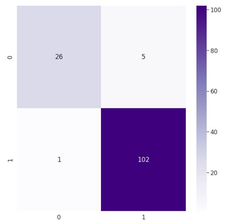


Figure 10. A confusion matrix for the best classification inhibition model (RDKit + Random Forest) based on valid set.

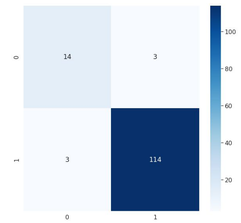


Figure 11. A confusion matrix for the best classification inhibition model (RDKit + Random Forest) based on test set.

Classification Transport Results

- **Best model performance for Classification + Transport:**
  - Feature: MOE
  - Model: Random Forest
  - valid_roc_auc_score: 0.863

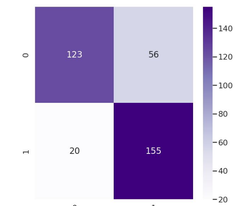


Figure 12. A confusion matrix for the best classification transport model (MOE + Random Forest) based on valid set.

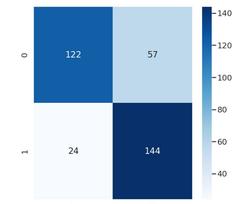
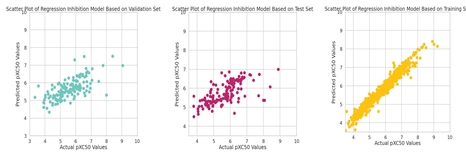


Figure 13. A confusion matrix for the best classification transport model (MOE + Random Forest) based on test set.

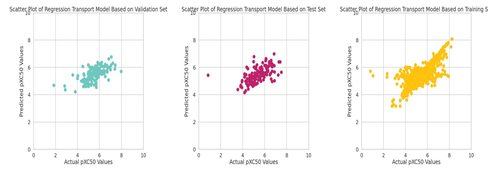
Regression Inhibition Results

- **Best model performance for Regression + Inhibition:**
  - Feature: RDKit
  - Model: Random Forest
  - valid_roc_auc_score: 0.448



Regression Transport Results

- **Best model performance for Regression + Transport:**
  - Feature: Mordred Filteredt
  - Model: Random Forest
  - valid_roc_auc_score: 0.385



Conclusion

- For all 4 of the different data frames, my highest scoring models were **Random Forest**
  - Classification inhibition data: **RDKit + Random Forest**
  - Classification transport data: **MOE + Random Forest**
  - Regression inhibition data: **RDKit + Random Forest**
  - Regression transport data: **Mordred Filtered + Random Forest**

Future Aims

- Improve the regression models
  - The scores for my regression models were significantly lower than the scores of my classification models
  - Find more data to work with in order to build more efficient regression models!
- Collect + curate more data for the classification models

References

- ChEMBL Database. EBI. https://www.ebi.ac.uk/chembl/g/#search_results/targets/query=abcb1. Accessed July 13, 2021.
- H2020 ExCAPE. ExCAPE-DB: ExCAPE chemogenomics database. https://solr.ideaconsult.net/search/excape/. Accessed July 13, 2021.
- 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 Mar 27;57(3):594-607. doi: 10.1021/acs.jcim.6b00508. Epub 2017 Mar 6. PMID: 28228010.

Acknowledgements

- Amanda Paulson, Ph.D.
- Susan Mertins, Ph.D
- Caleb Class, Ph.D
- Sarangan Ravichandran, Ph.D.
- Da Shi, Ph.D.
- Fellow ATOM summer interns

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.