

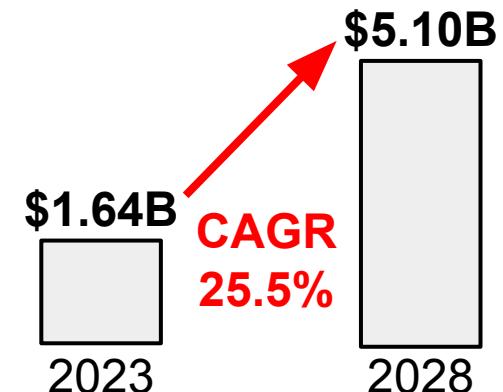
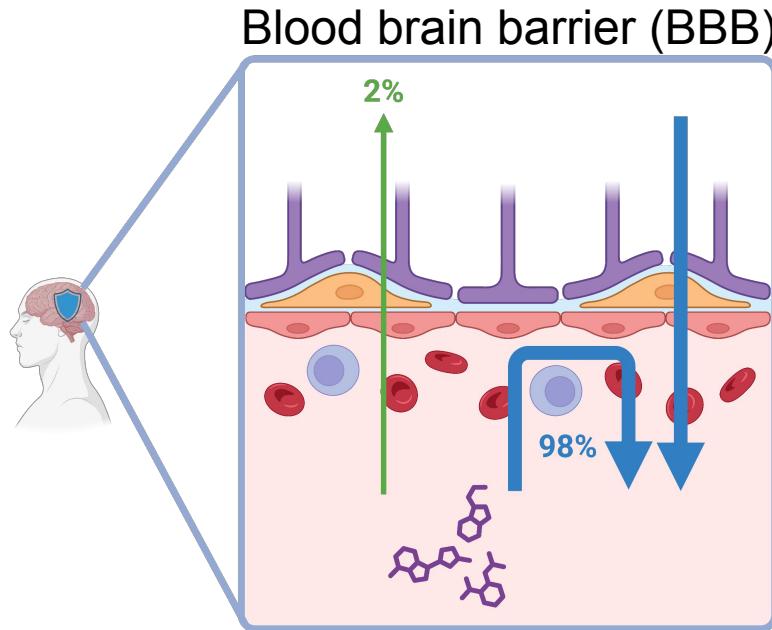
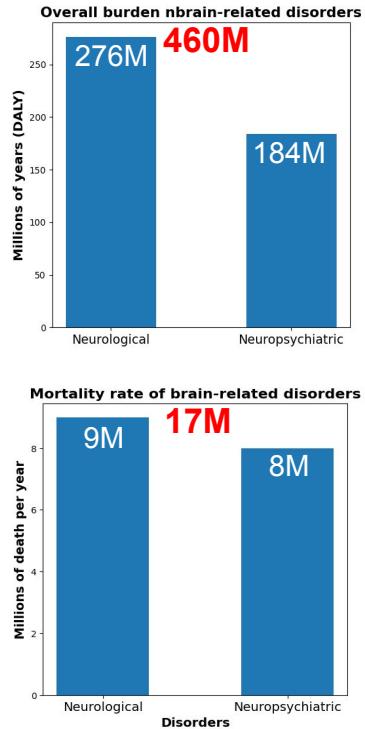
Harnessing Machine Learning Insights for Precision Prediction of Chemical Blood-Brain Barrier Permeability

4/11/2024 BME 790L Final project presentation

Junqi Lu, Jessop Oliver
Department of Biomedical Engineering

Mentor: Dr. Michael D Lynch
Department of Biomedical Engineering

BBB drug delivery challenge sparks huge market potential



Álvaro Nuno, P., Suarez, J. & Michel Le, B. Sizing the brain. Deloitte Insights (2023).

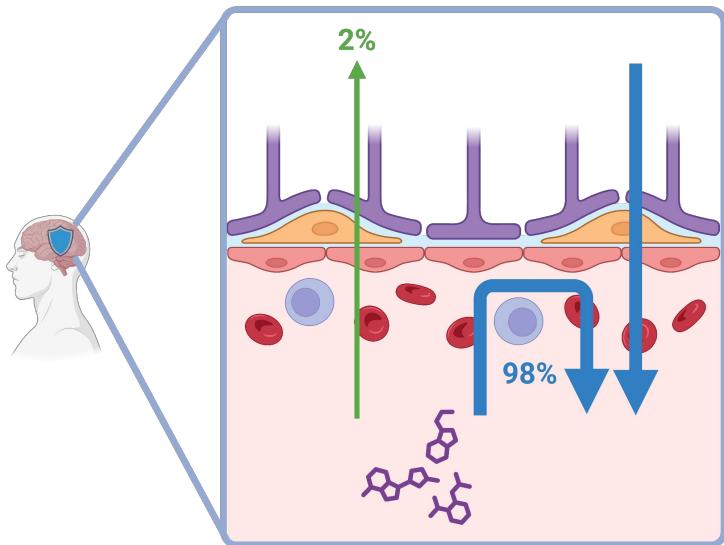
Wu, D. et al. The blood-brain barrier: structure, regulation, and drug delivery. *Signal Transduction and Targeted Therapy* 8, (2023).

Fayzullin, A., Bakulina, A., Mikaelyan, K., Shekhter, A. & Guller, A. Implantable Drug Delivery Systems and Foreign Body Reaction: Traversing the Current Clinical Landscape. *Bioengineering (Basel)* 8 (2021).

<https://doi.org/10.3390/bioengineering8120205>

The Insight Partners. Blood Brain Barrier Technologies Market Global Analysis by 2028. The Insight Partners <https://www.theinsightpartners.com/reports/blood-brain-barrier-technologies-market> (2021).

Graphical abstract



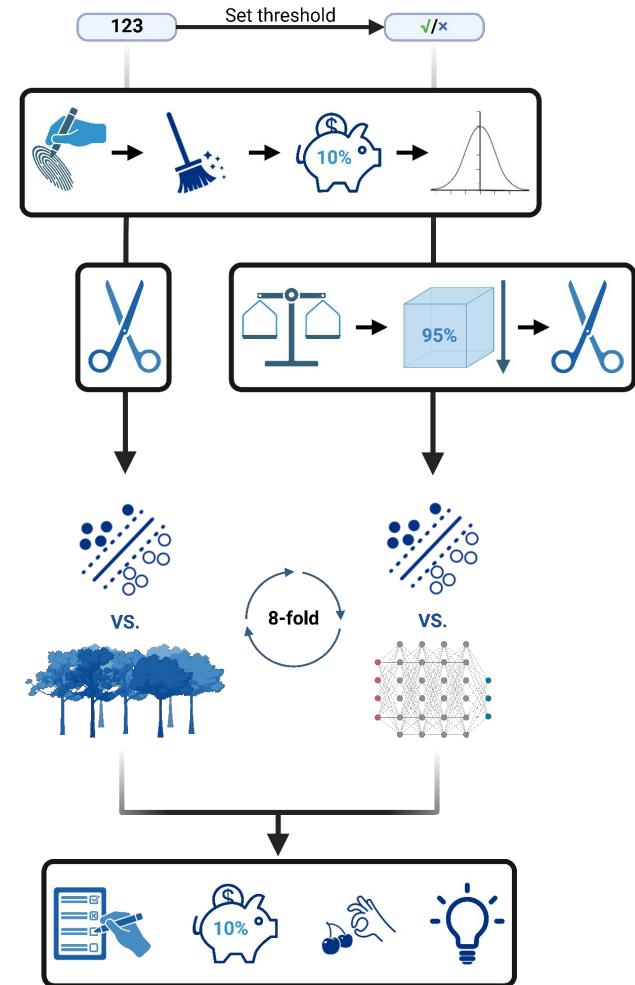
Raw data

Data preprocessing

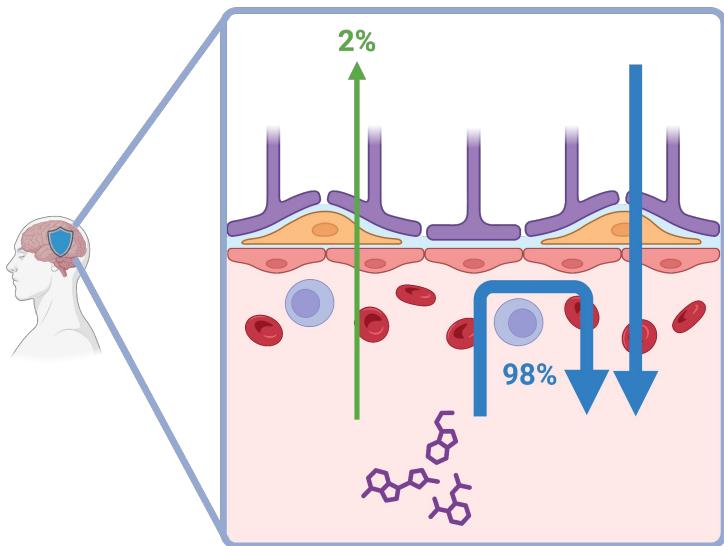
Data processing

Model training & comparisons
(baseline vs. complex)

Evaluation,
conclusions &
inspirations



Graphical abstract



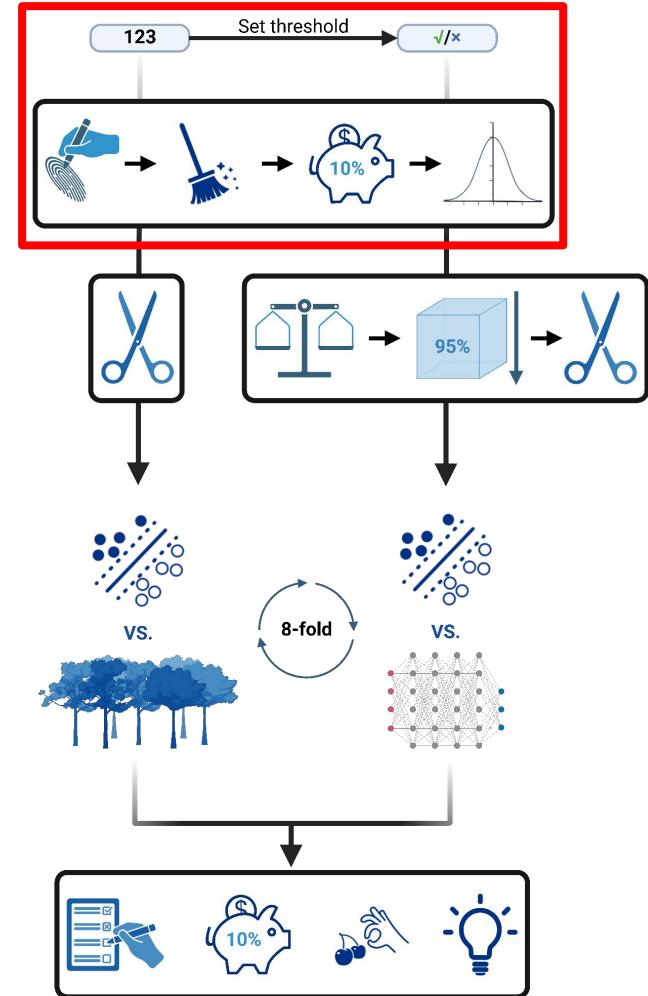
Raw data

Data preprocessing

Data processing

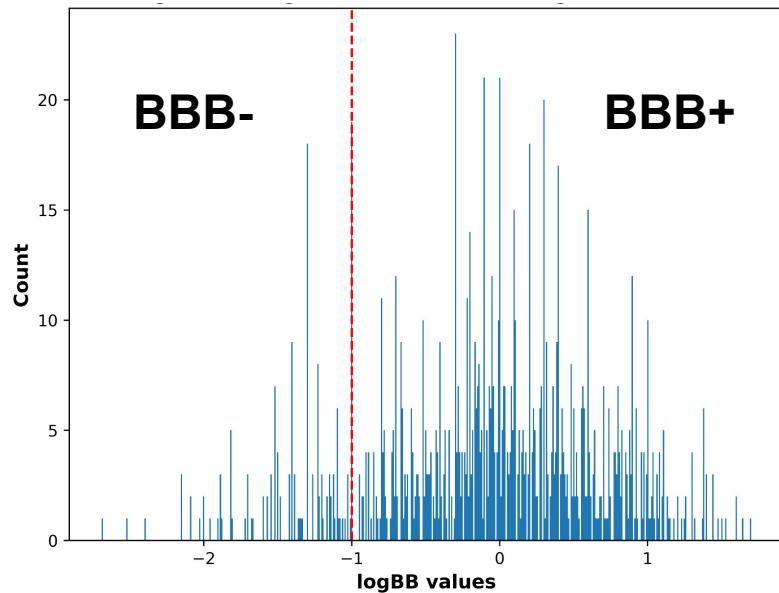
Model training & comparisons
(baseline vs. complex)

Evaluation,
conclusions &
inspirations

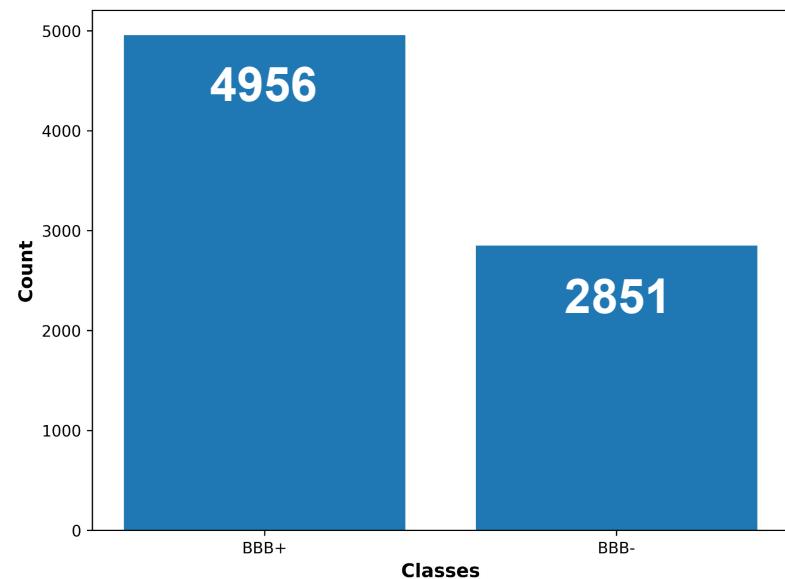


Original datasets have regression & classification data

Regression: 1058



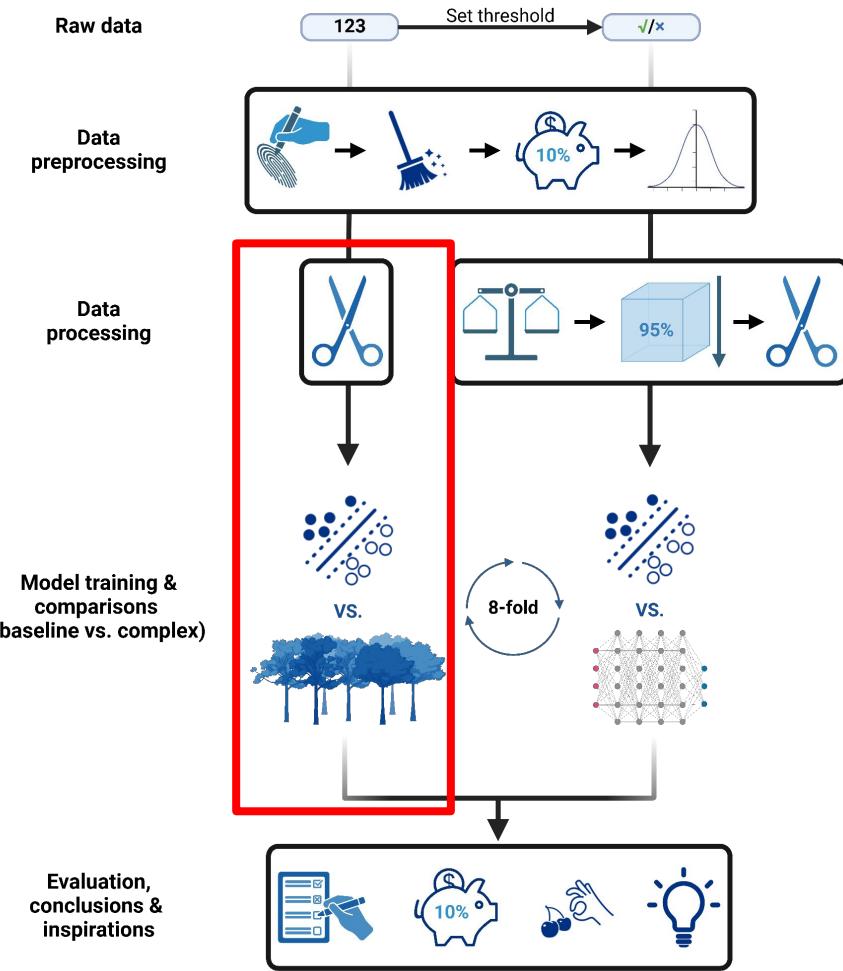
Classification: 7807



Common pre-processings on both datasets

1. Feature expansion
 - a. RDKit descriptors
 - b. Morgan fingerprints
 - c. MACCS keys
2. Feature selection
 - a. Get rid of 0 variance features
3. Hold out 10% original data
4. Center & standardization
 - a. Data processing transformers specific to regression / classification problem

logBB regression prediction



Data processing & baseline & complex regressors training

- 80-20 train test split
- Transformer shared by 2 models
- Cross-validation by 8 folds
- Scoring: MAE, MSE, R2
- Refit by R2

Baseline: SVR

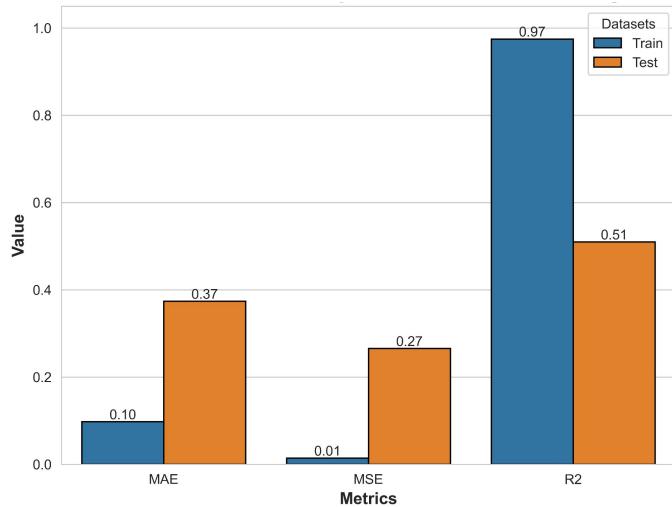
- 'kernel': ['rbf']
- 'gamma': ['auto', 3]
- 'C': [1, 5]

Complex: RF

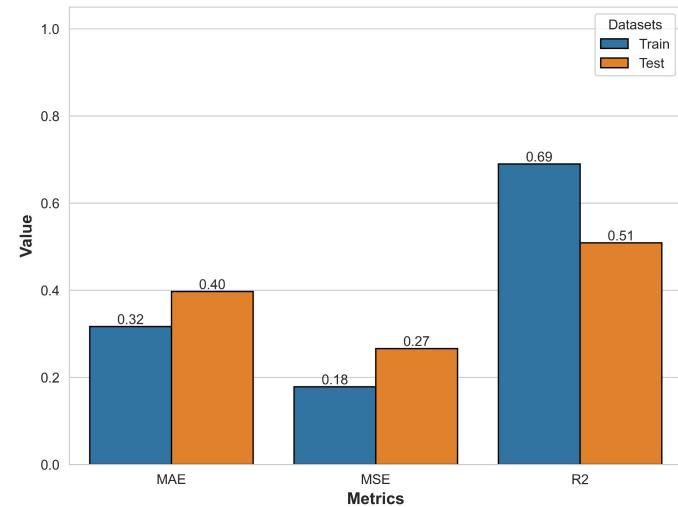
- 'n_estimators': [500, 1000]
- 'max_depth': [5, 20]
- 'min_samples_split': [2, 50]
- 'min_samples_leaf': [5, 50]
- 'criterion': ['squared_error', 'absolute_error']
- 'max_samples':[0.2, 0.85],
- 'max_features': ['sqrt']

SVR likely to be overfitting

Baseline: SVR

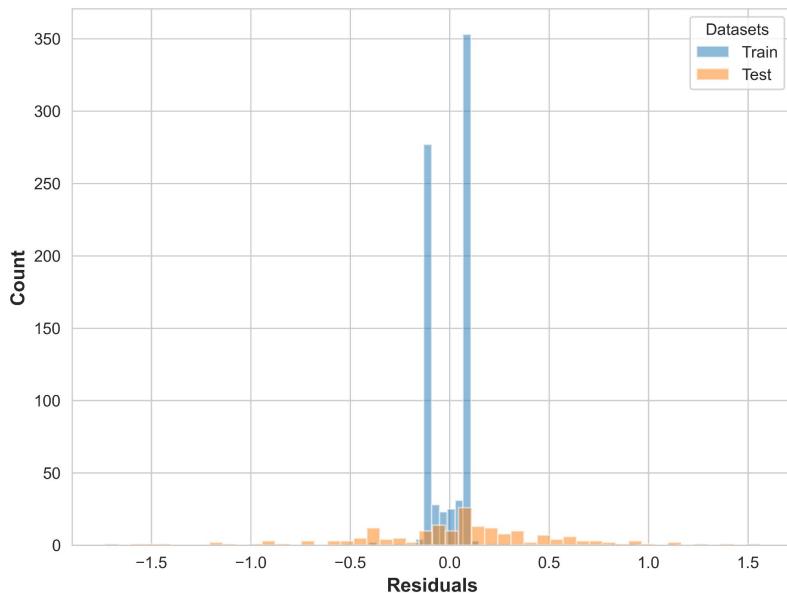


Complex: RF

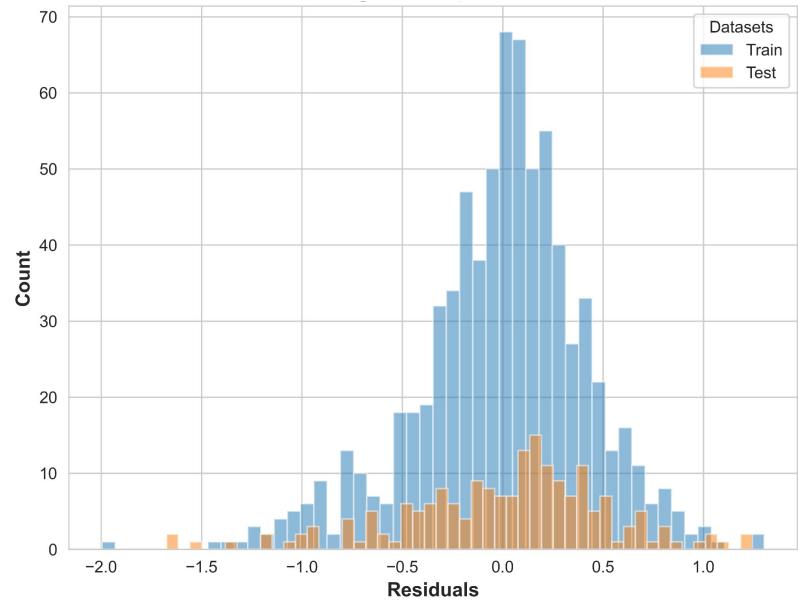


SVR likely to be overfitting

Baseline: SVR

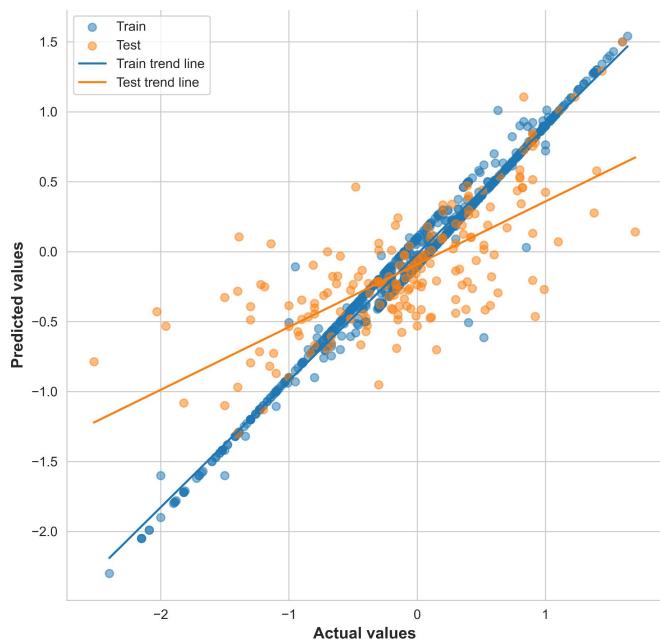


Complex: RF

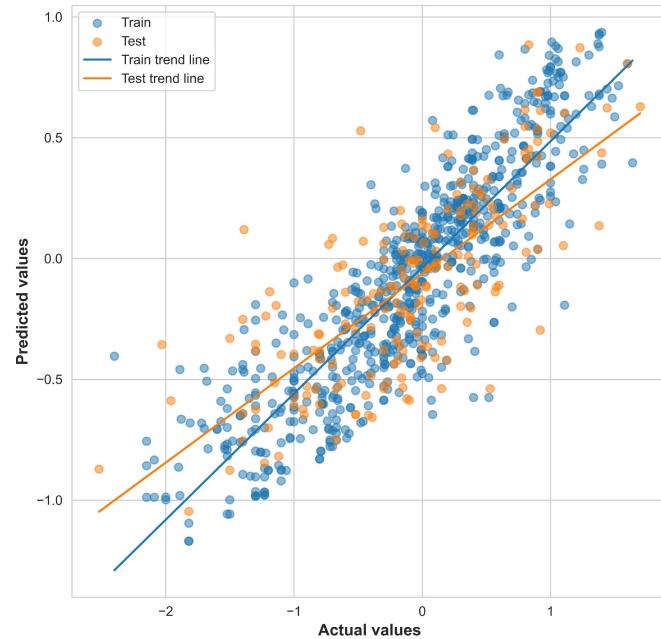


SVR likely to be overfitting

Baseline: SVR

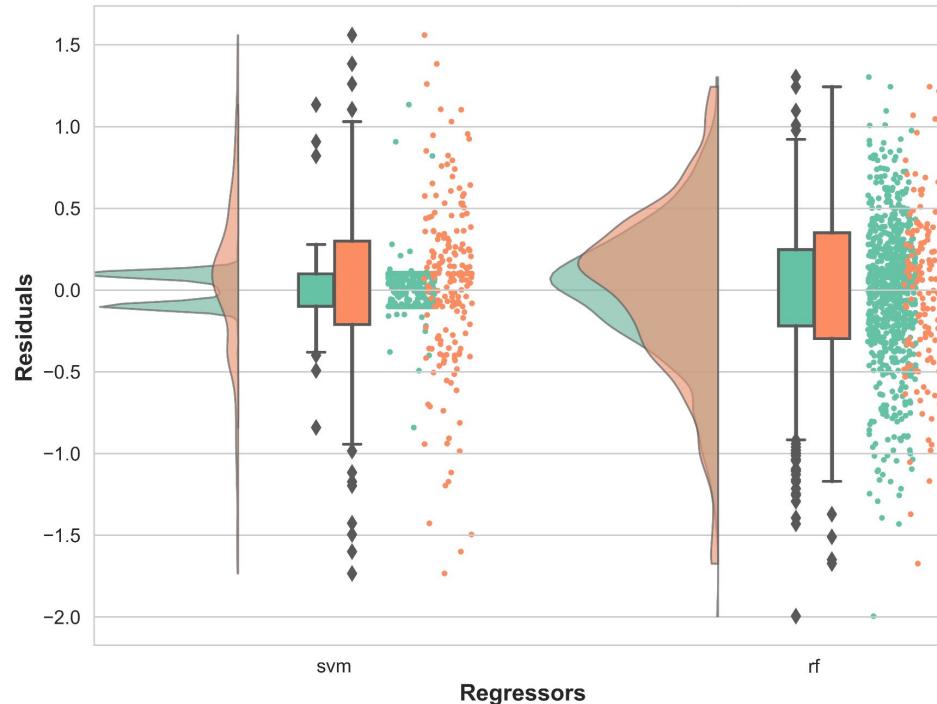


Complex: RF



SVR likely to be overfitting

Baseline: SVR



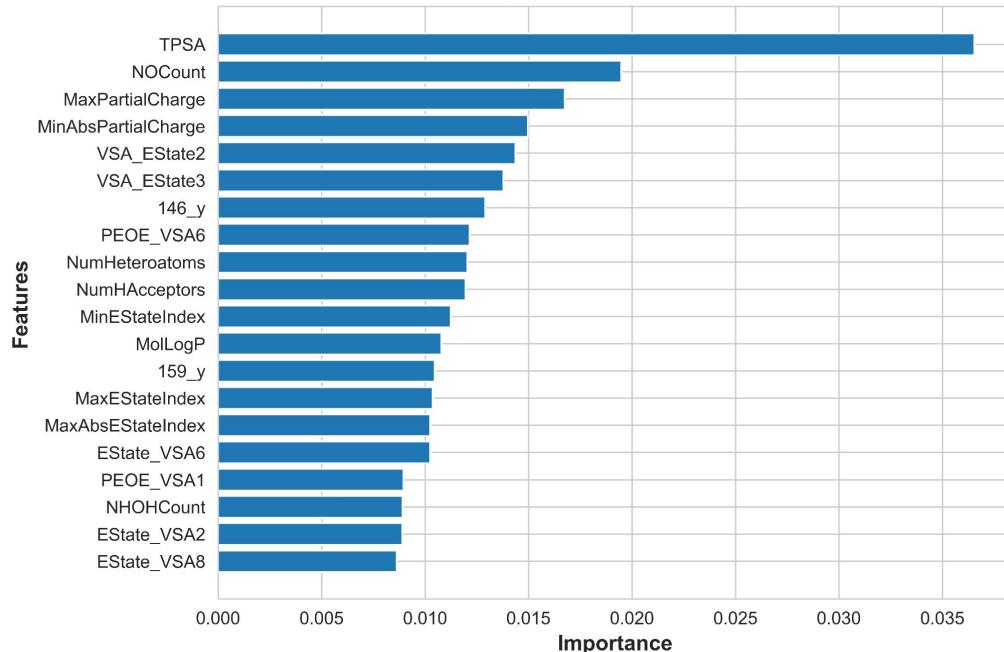
Complex: RF

Datasets
train
test

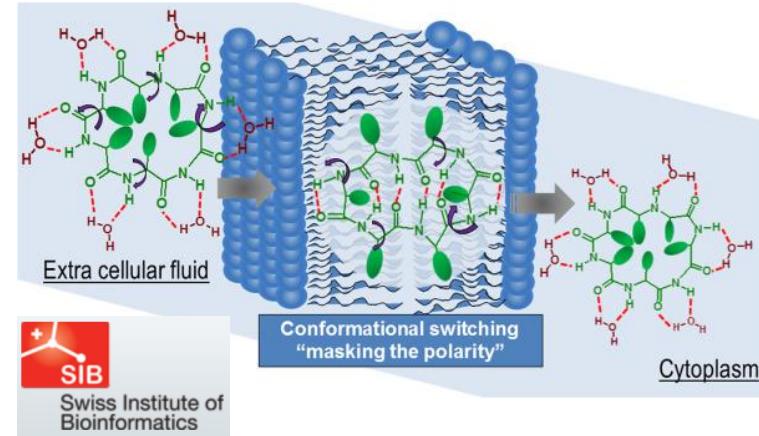
T-test showed

- Between 2 train sets: p-value = 0.32
- Between 2 test sets: p-value = 0.06

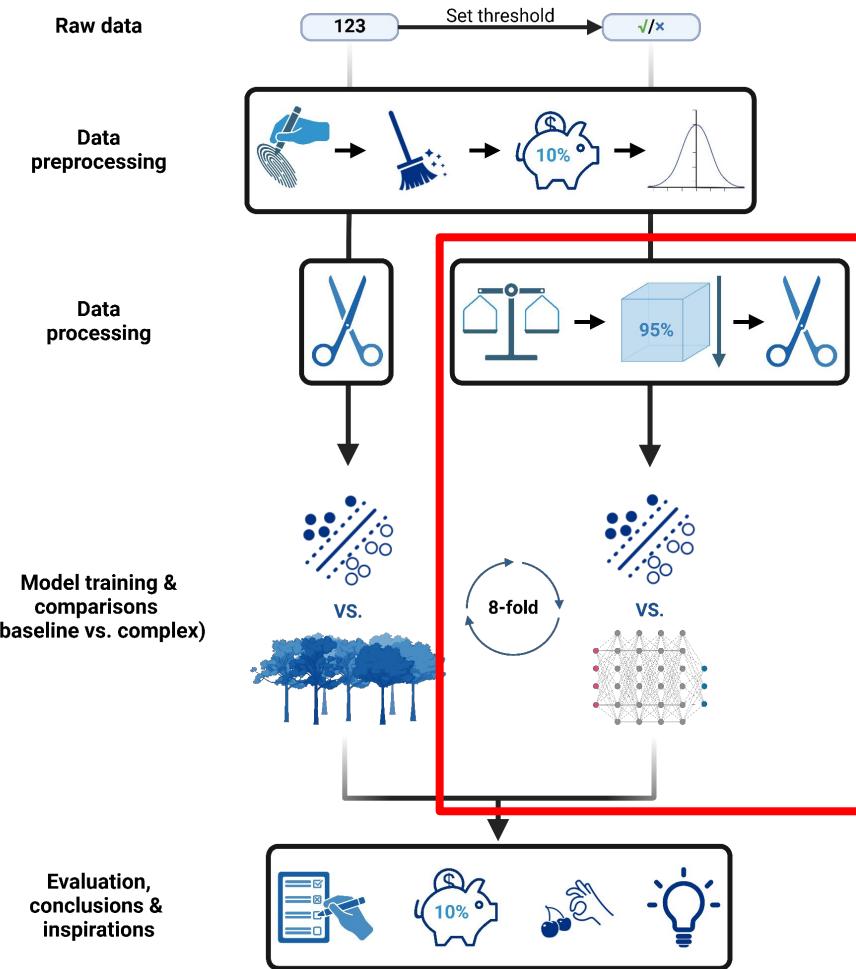
RF shows important features related to polarity & charges



Topological polar surface area (TPSA): measure of molecule's polar surface area

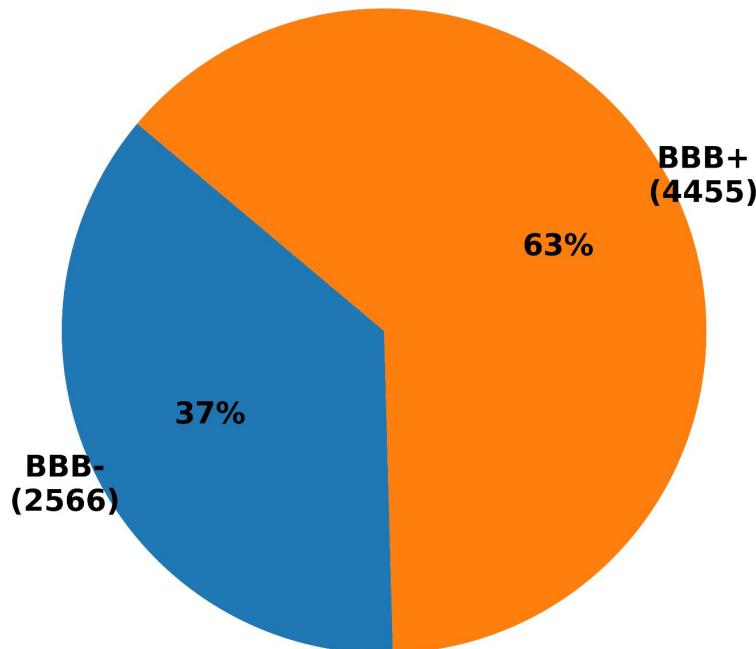


BBB+/BBB- classification prediction



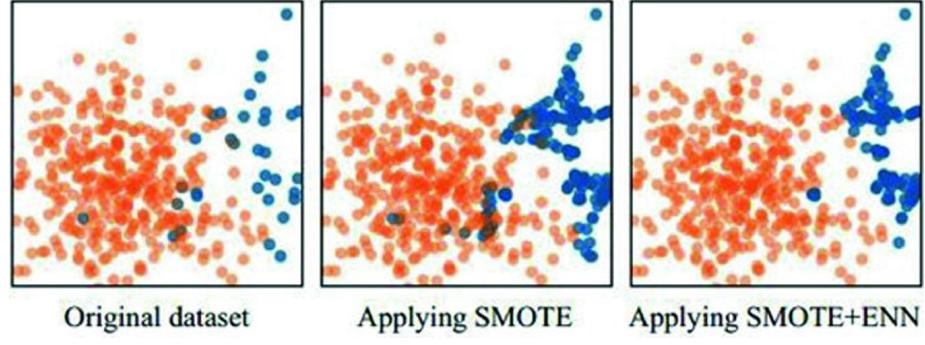
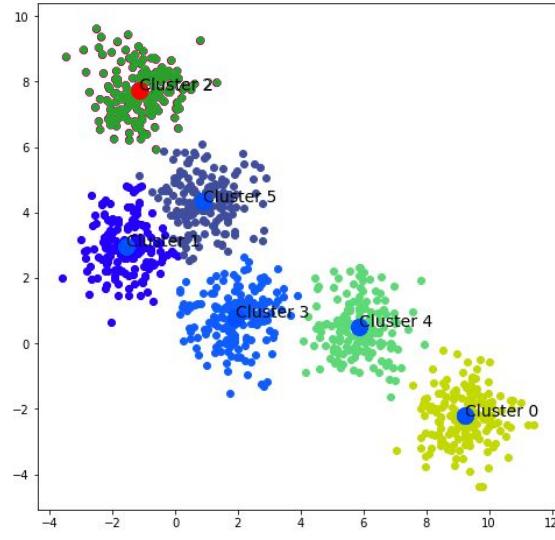
Continue dataset preprocessing for classification problem

Composition of 2 categories in regression dataset before balancing



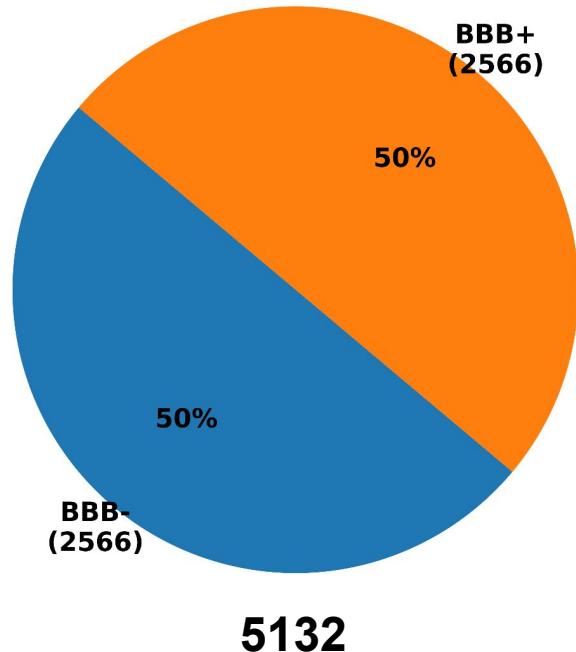
Balancing classification dataset by 2 methods

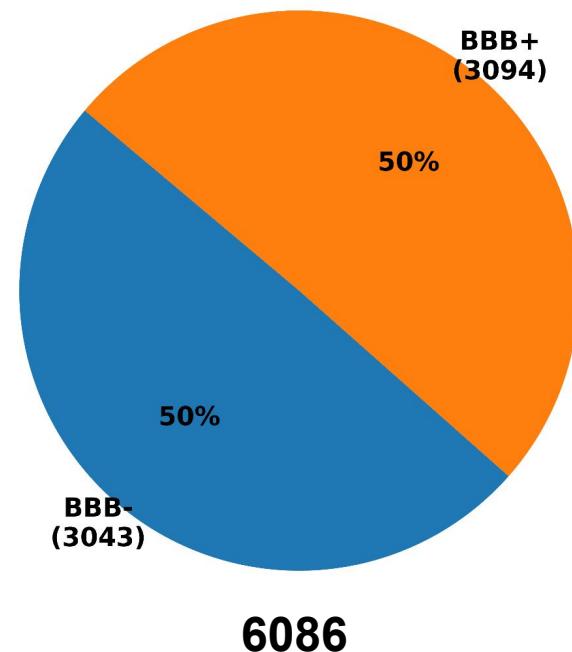
- **ClusterCentroids**
 - ↓ bigger category
- **SMOTEENN (SMOTE+ENN)**
 - ↑ smaller category
 - Cleaning



Balancing classification dataset by 2 methods

- ClusterCentroids
- SMOTEENN (SMOTE+ENN)





Data processing & baseline regressors training

- PCA to cover 95% variance
- 80-20 stratified train test split
- Transformer shared by 2 models
- Repeated stratified cross validation by 8 folds & 2 repeats
- Scoring: recall, precision, F1, accuracy, balanced accuracy, AUROC
- Refit by AUROC

Baseline: SVC on ClusterCentroids

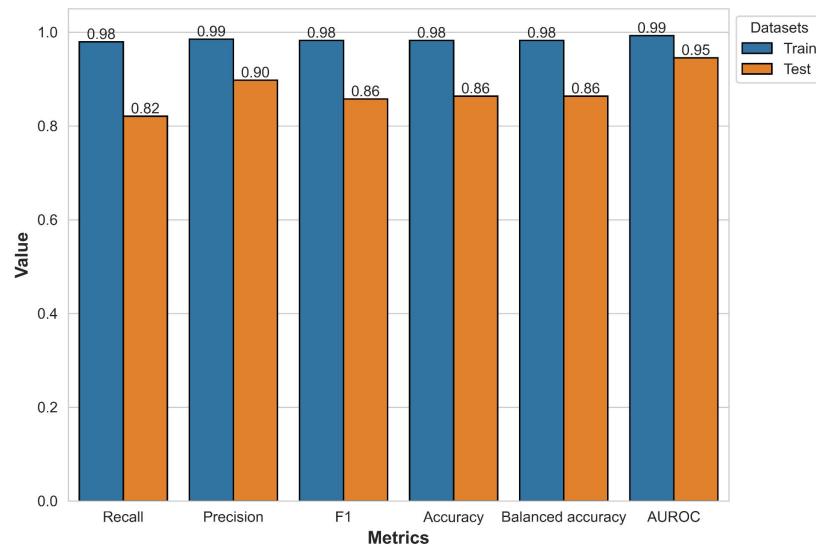
- 'kernel': ['rbf']
- 'gamma': ['auto', 3]
- 'C': [1, 5]

Baseline: SVC on SMOTEENN

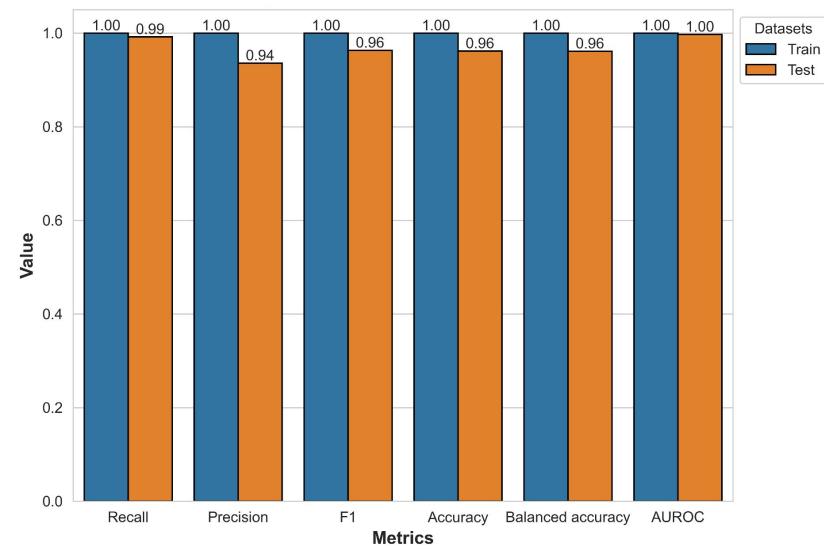
- 'kernel': ['rbf']
- 'gamma': ['auto', 3]
- 'C': [1, 5]

ClusterCentroids gave more reliable SVC

Baseline: SVC on ClusterCentroids

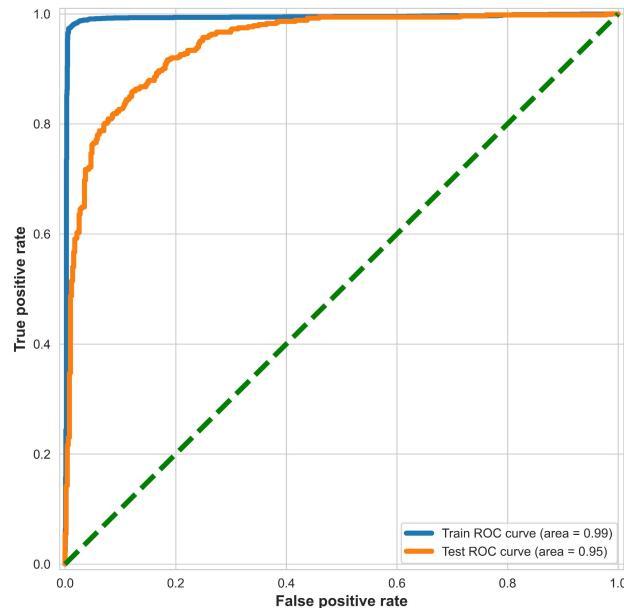


Baseline: SVC on SMOTEENN

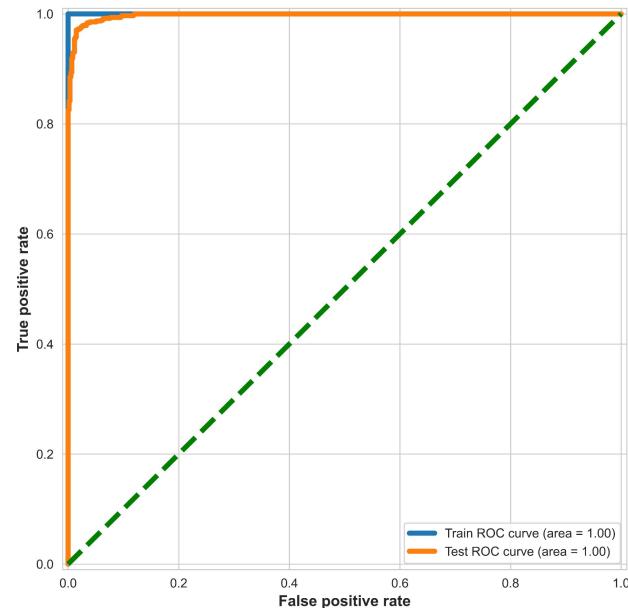


ClusterCentroids gave more reliable SVC

Baseline: SVC on ClusterCentroids

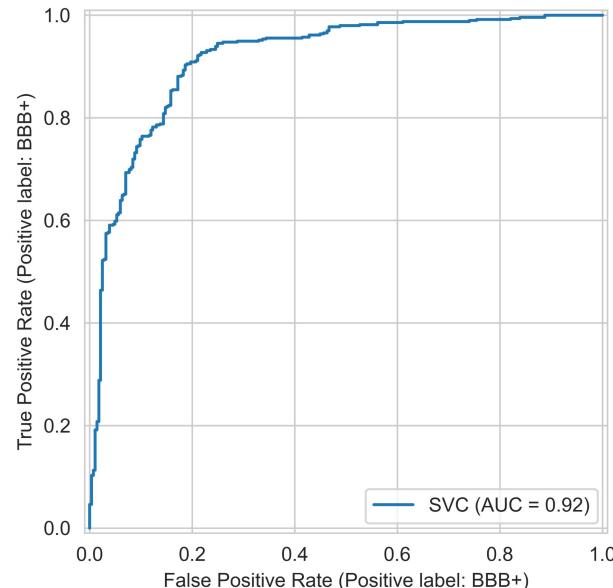


Baseline: SVC on SMOTEENN

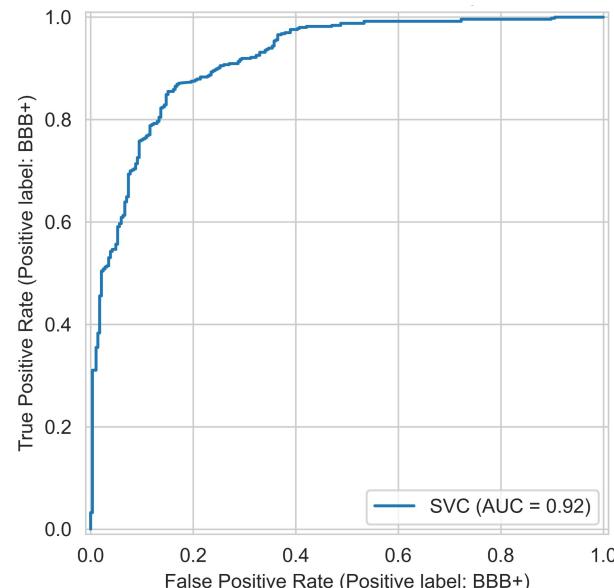


ClusterCentroids gave more reliable SVC performance on holdout

Baseline: SVC on ClusterCentroids



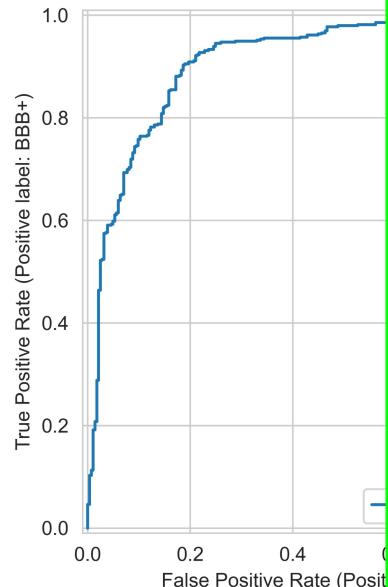
Baseline: SVC on SMOTEENN



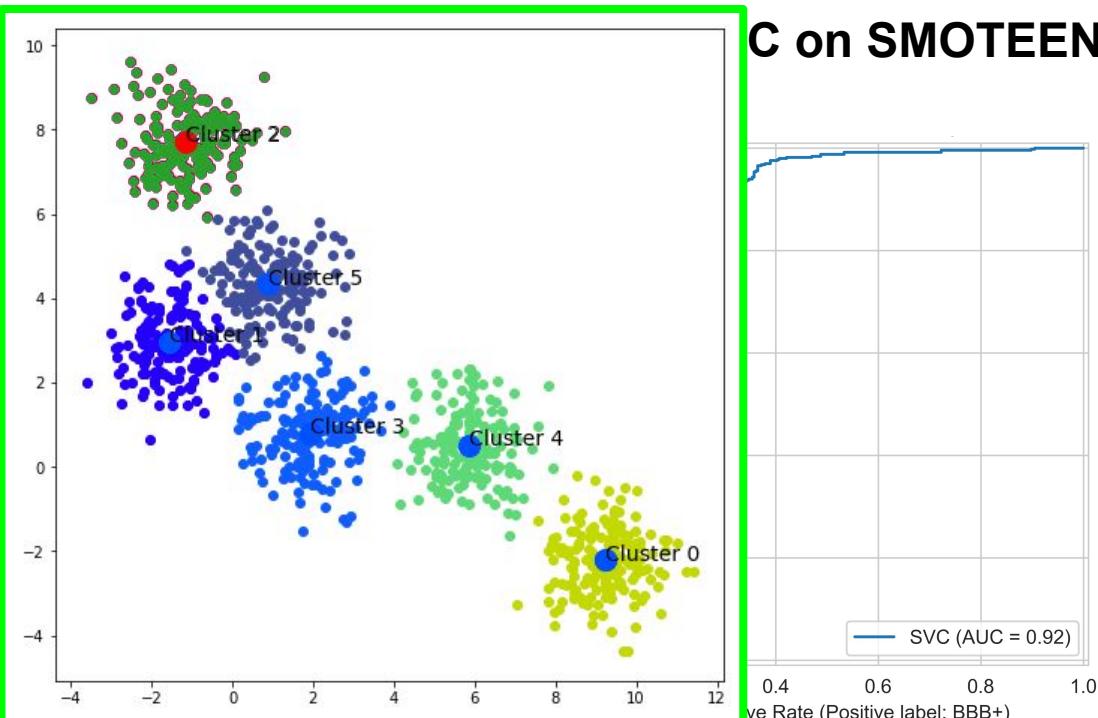
DeLong's test gives p-value = 0.944

ClusterCentroids gave more reliable SVC performance

Baseline: SVC on Clu



C on SMOTEENN



DeLong's test gives p-value = 0.944

Data processing & regressors training

- ClusterCentroids to undersample
- PCA to cover 95% variance
- 80-20 stratified train test split
- Transformer shared by 2 models
- Repeated stratified cross validation by 8 folds & 2 repeats
- Scoring: accuracy, balanced accuracy, AUROC
- Refit by AUROC

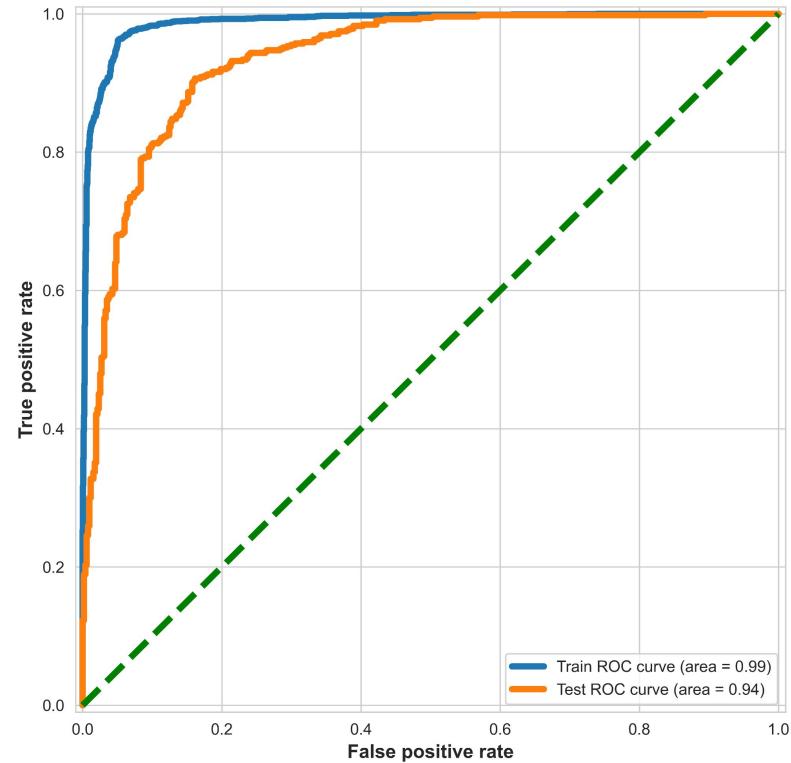
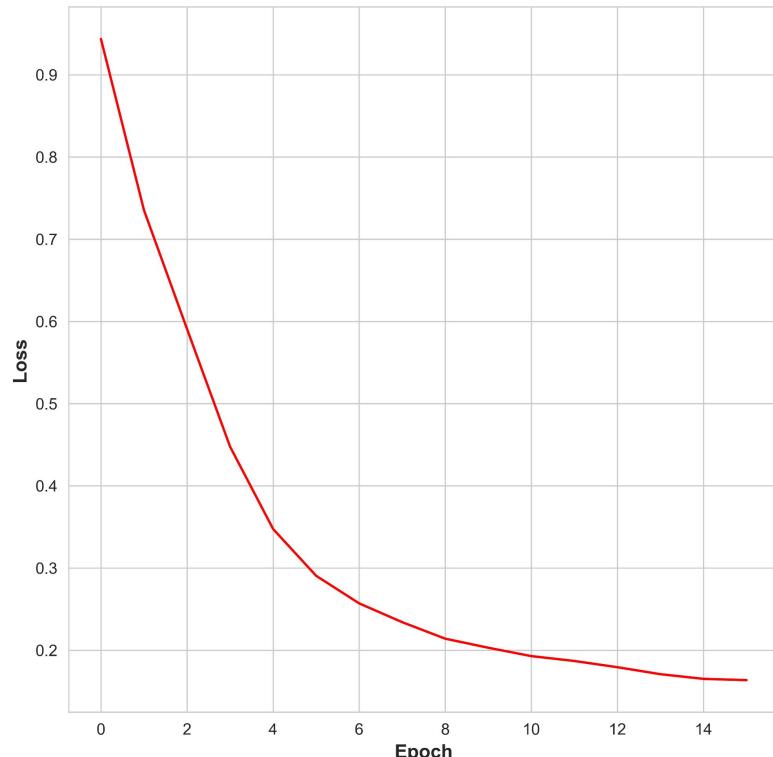
Baseline: SVC on ClusterCentroids

- 'kernel': ['rbf']
- 'gamma': ['auto', 3]
- 'C': [1, 5]

Complex: MLP on ClusterCentroids

- 'hidden_layer_sizes': [(5,), (10,), (10,10)],
- 'alpha': [1e-3, 1],
- 'activation': ['relu', 'tanh'],
- 'batch_size': [16, 64],
- 'learning_rate': ['constant', 'adaptive']

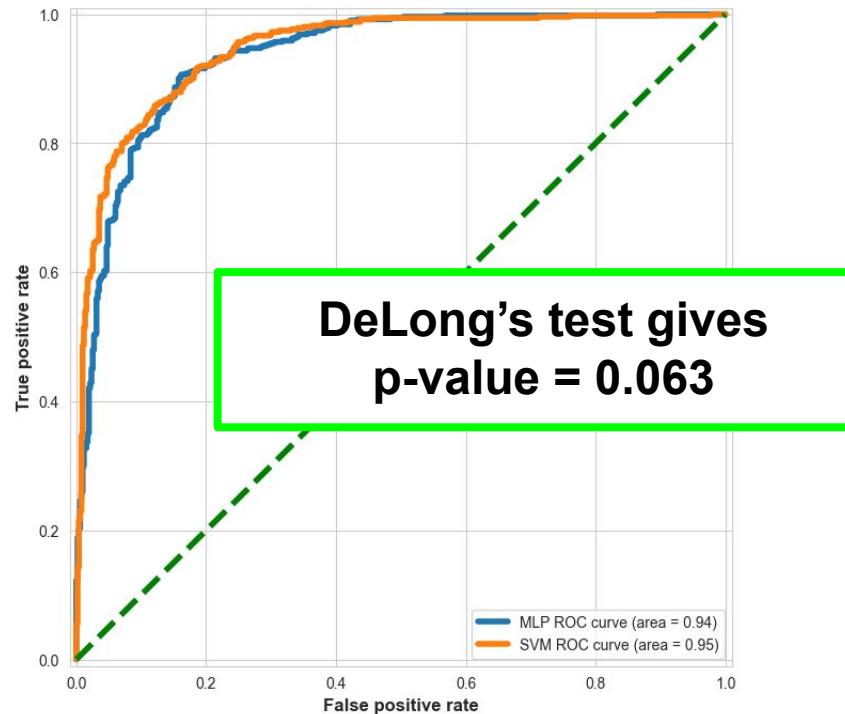
Best MLP classifier showed reasonable learning process



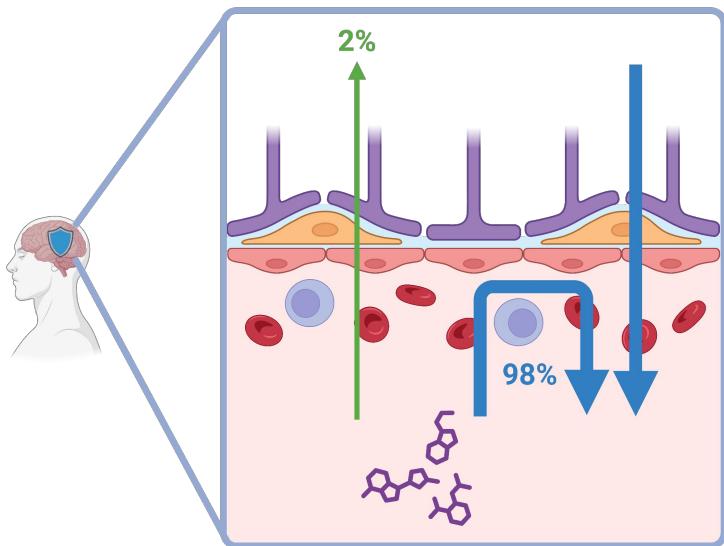
ClusterCentroids gave more reliable SVC

Baseline: SVC on ClusterCentroids

Complex: MLP on ClusterCentroids



Graphical abstract



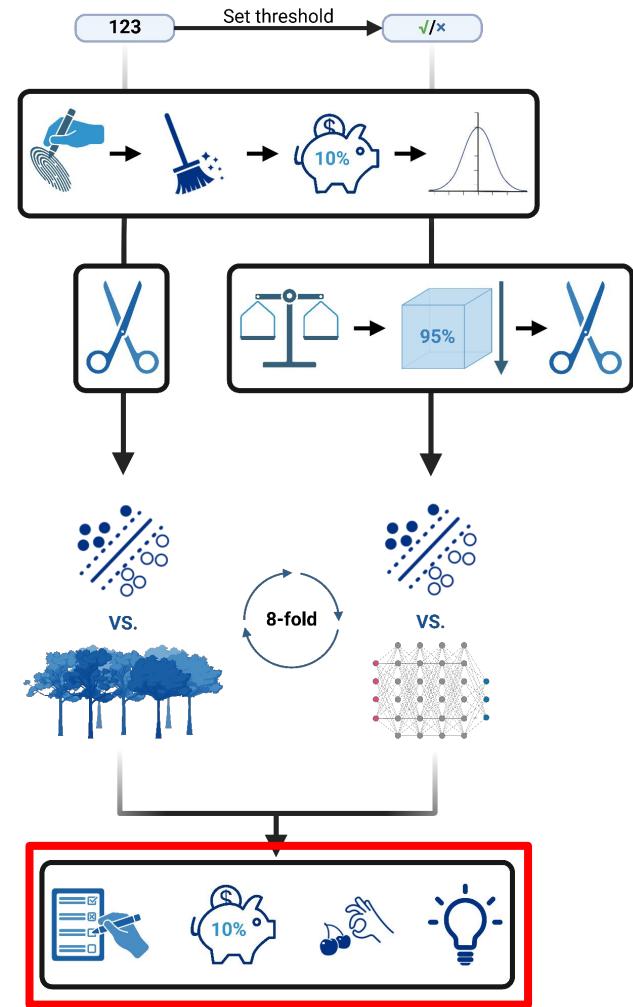
Raw data

Data preprocessing

Data processing

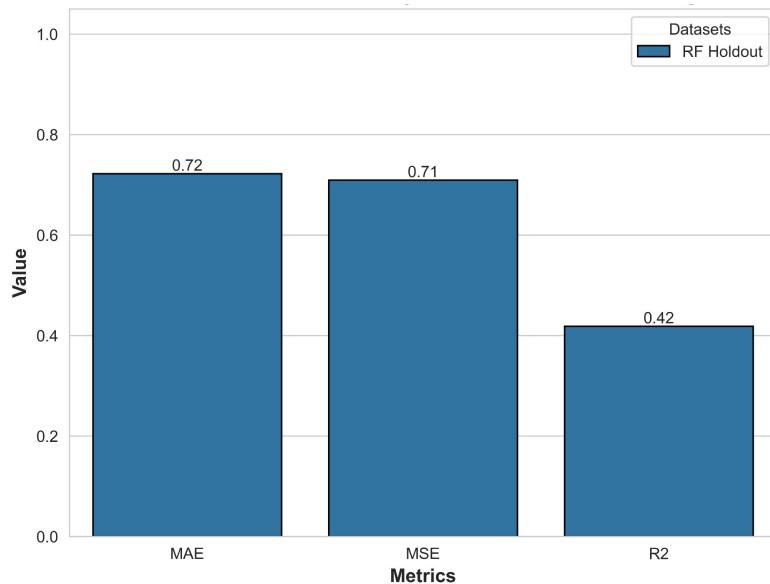
Model training & comparisons
(baseline vs. complex)

Evaluation,
conclusions &
inspirations

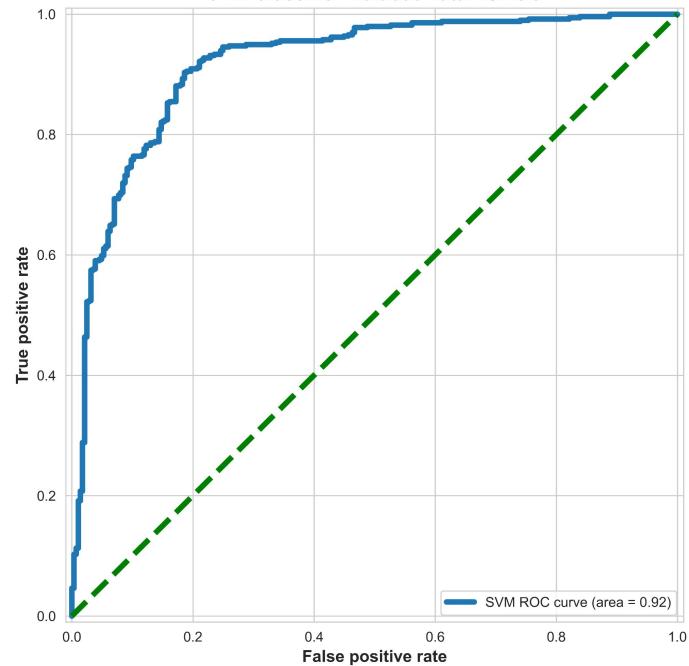


Sanity check: Holdout dataset

RF regressor

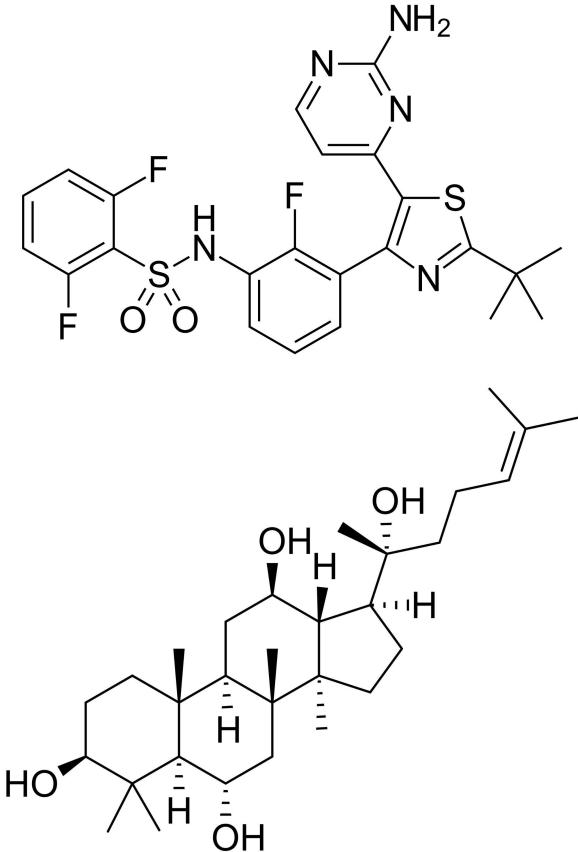
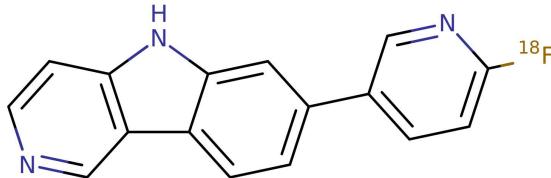


SVC



Sanity check: cherry picking for SVC

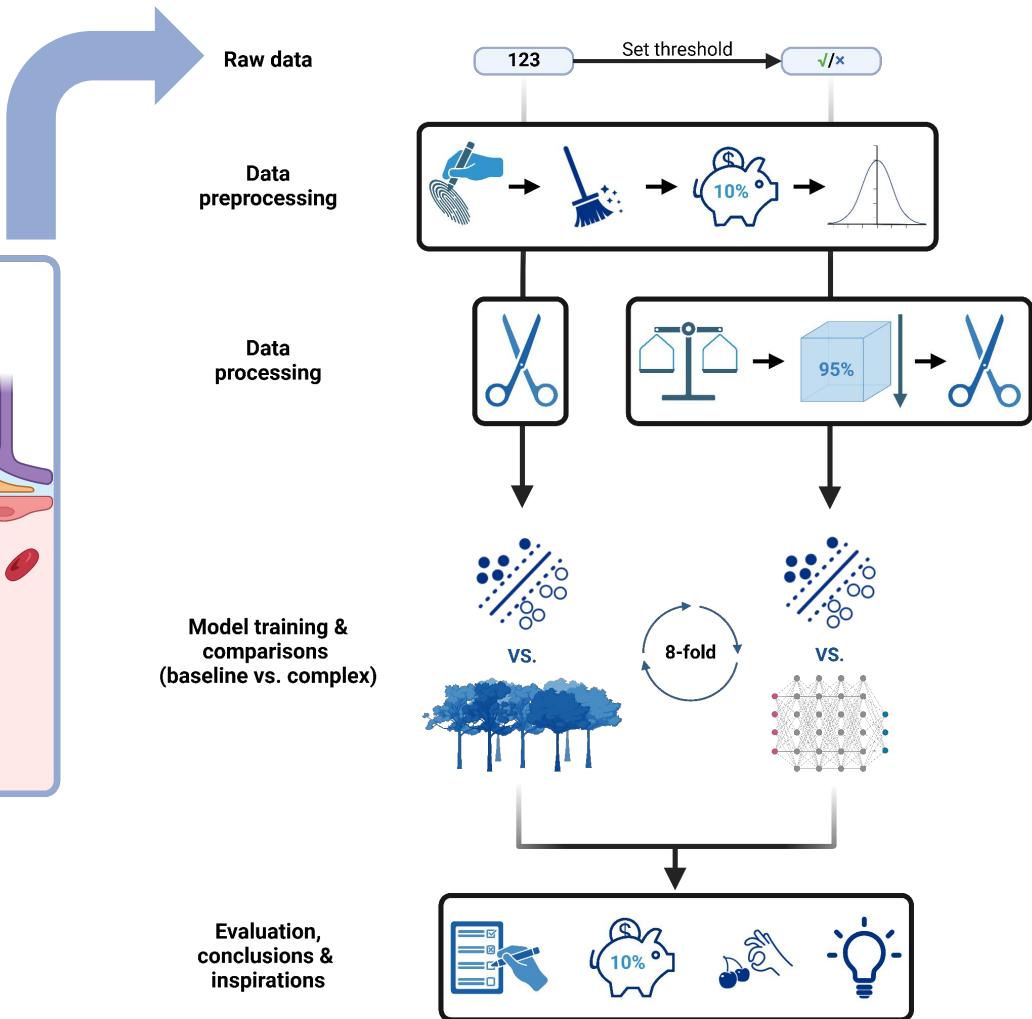
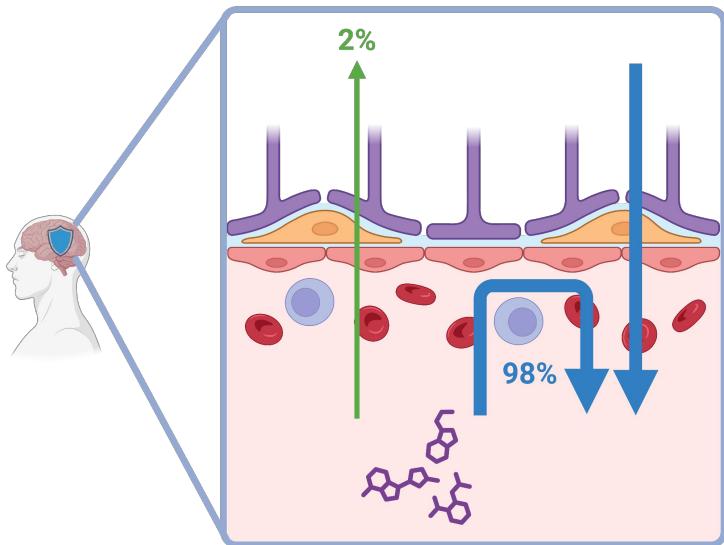
- BBB-
 - Dabrafenib: predicted to be BBB-
 - Protopanaxatriol: predicted to be BBB-
- BBB+
 - Flortaucipir F-18: predicted to be BBB-



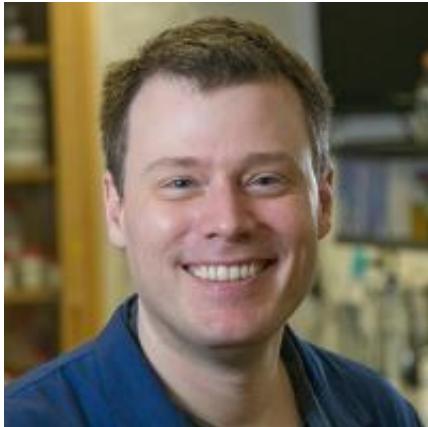
Conclusions

- Categorical dataset successfully balanced using multiple methods
- RF regressor & SVC performed & generalized better than other models
- Successful sanity check with orthogonal domain knowledge, holdout dataset verification, cherry picking (mostly)
- Complex models tend to overfit & require heavier regulations
- Polarity & charges play critical roles in determining ability to penetrate BBB
 - Agreed with commercial tool SwissADME
- Improvements & future directions
 - Correlation doesn't guarantee causation
 - Validations in vitro & in vivo

Graphical abstract



Thank you!



Duke
UNIVERSITY

BIOMEDICAL
ENGINEERING

Thank you!

