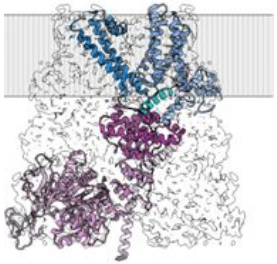
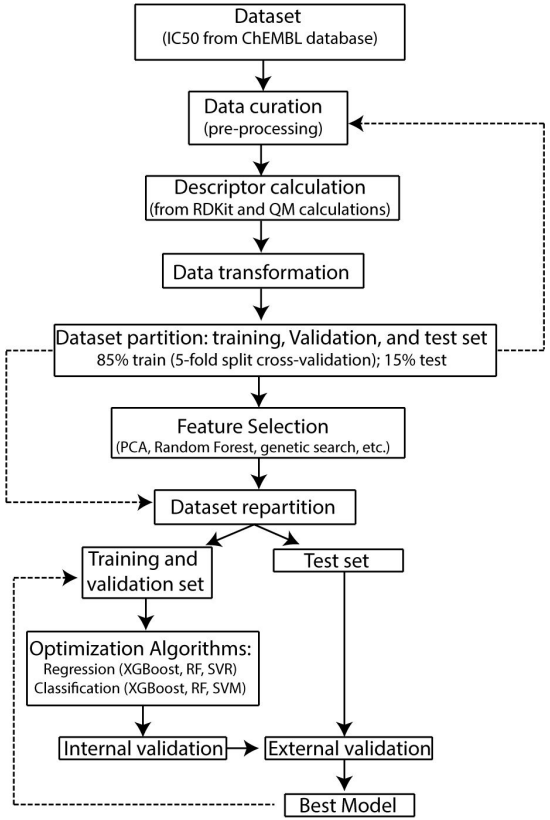
Predicting Drug Potency for the TRPM8 Cold Receptor:

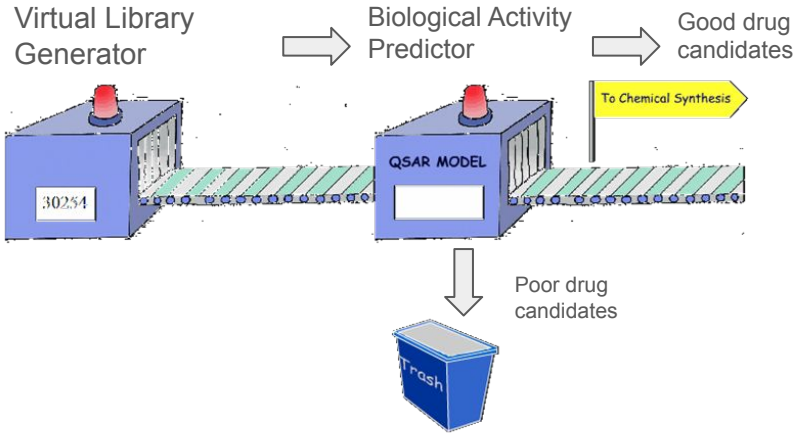
A Data Science Approach

Adedolapo Ojoawo, Carmen Al Masri, Jaehyoun Seiler, Jessica Pan



QSAR(Quantitative structure-activity relationship) Workflow for targeting TRPM8





Adapted from Palchevskyi et al. Commun. Biol 2023

Target: Transient receptor potential M8 protein (CHEMBL1667665) - ChEMBL

Method

Preprocessing: 1168 (ChEMBL) → 654 compounds

- Removed data with empty values, salt ions, small fragments, duplicates. Ensured correct SMILE representation. Standardized IC50.

Descriptor calculation

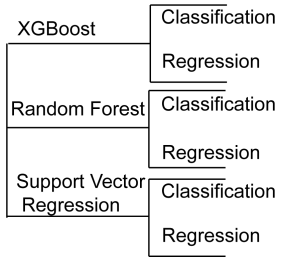
- 3D properties, 2D topological descriptors, quantum mechanical properties, an constitutional properties (number of atoms, type, size etc)

Feature reduction: 1008 → 564 features

- Factor Analysis
- PCA
- Genetic Algorithm
- Random Forest elimination

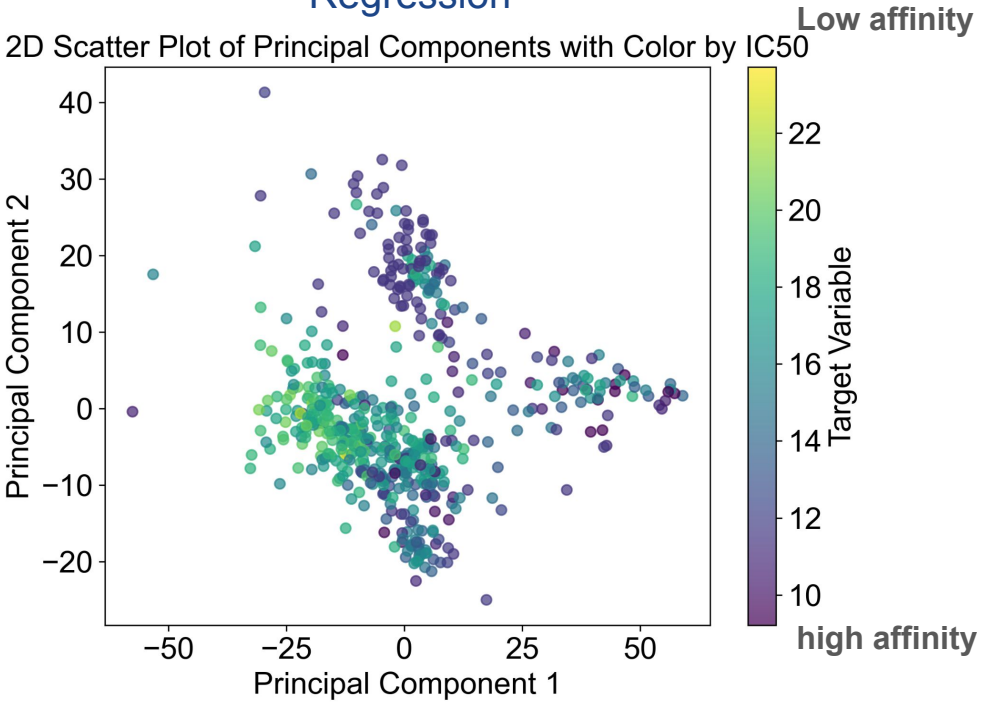
- 6 quantum -, 23 physicochemical -, 62 topological -, 917 3D -descriptors/features selected.
- Removed features with low variance/high correlation, then by Factor Analysis, PCA, Genetic Algorithm, or Random Forest elimination

Model training on reduced features

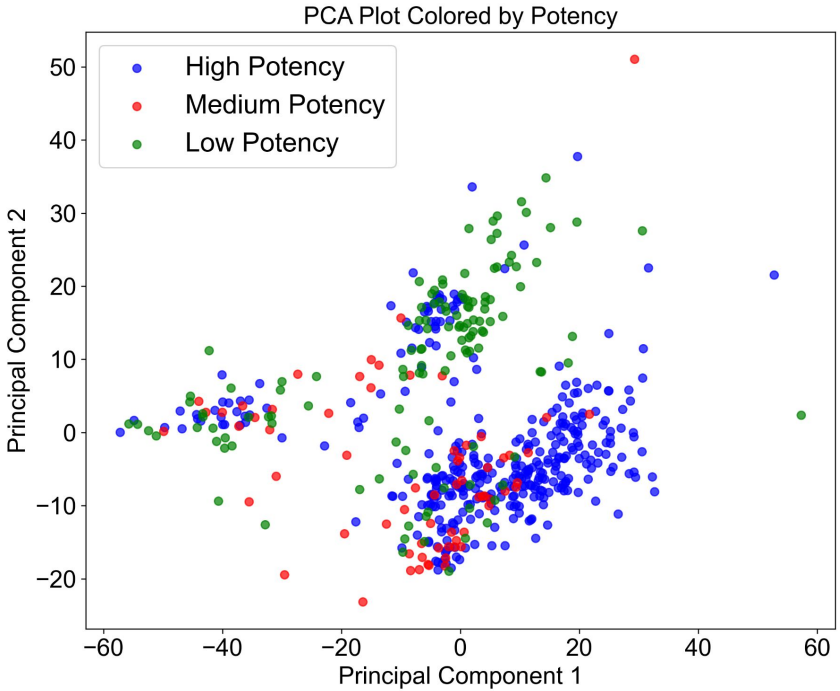


PCA-based feature reduction

Regression



Classification

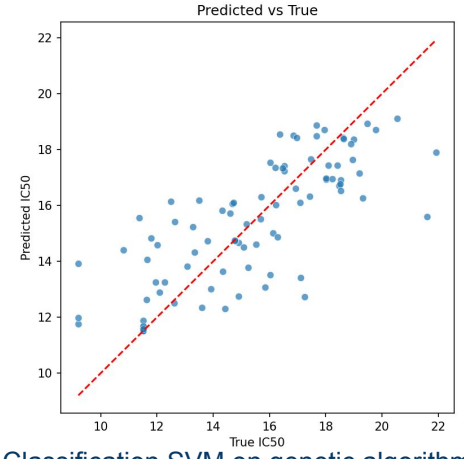


Model training on reduced features

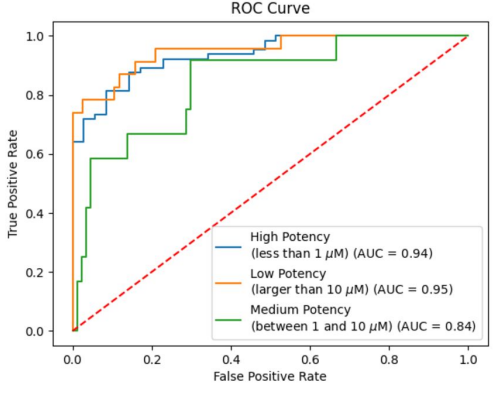
model_type	problem_type	input_type	mse	r2_score	pearson_corr
XGBoost	regression	PCA	3.2704973524449734	0.6420536843052471	0.8013067671152532
XGBoost	regression	Factor_Analysis	3.7684174475335417	0.5875577944326891	0.772378957618285
XGBoost	regression	correlation_variance_filter	3.412345623444292	0.6265288082634062	0.7967625992038058
XGBoost	regression	genetic_algorithm	10.871522837415622	-0.0898126817928817	0.0636711776215474
XGBoost	regression	random_forest_elimination	3.587119082748313	0.6074003672046381	0.7800653648363141
random_forest	regression	PCA	3.149953623164591	0.6552468409191063	0.8095983688675181
random_forest	regression	genetic_algorithm	10.926797065431074	-0.0953536308915889	0.0729185251657888
random_forest	regression	random_forest_elimination	3.167114897972596	0.6533685898679144	0.8086356378510251
random_forest	regression	Factor_Analysis	3.572296932225368	0.6090226079829083	0.7805909424400309
random_forest	regression	correlation_variance_filter	3.259094175886536	0.6433017284393474	0.8023156357013755
SVM	regression	random_forest_elimination	3.2313366798065664	0.646339704742037	0.8051846703835307
SVM	regression	PCA	3.5482581575654395	0.6116535811640684	0.7841156051821632
SVM	regression	Factor_Analysis	3.9124194415509614	0.5717971997412733	0.7656218872317448
SVM	regression	genetic_algorithm	11.561558285184107	-0.1589850869021825	0.0320107087875223
SVM	regression	correlation_variance_filter	3.745183812068768	0.5901006474970993	0.7690865408730042

model_type	problem_type	input_type	accuracy	precision	recall	f1_score
XGBoost	classification	random_forest_elimination	0.8015267175572519	0.774837858003788	0.8015267175572519	0.7823988768974293
XGBoost	classification	Factor_Analysis	0.7862595419847328	0.7336113952939324	0.7862595419847328	0.7505746162234712
XGBoost	classification	genetic_algorithm	0.8080808080808081	0.7890032162420222	0.8080808080808081	0.7966141182428475
XGBoost	classification	PCA	0.8091603053435115	0.7882316656112057	0.8091603053435115	0.7854483550250262
XGBoost	classification	correlation_variance_filter	0.7633587786259542	0.7136565724351984	0.7633587786259542	0.737670972219113
random_forest	classification	Factor_Analysis	0.7709923664122137	0.6870229007633588	0.7709923664122137	0.7239124876730341
random_forest	classification	random_forest_elimination	0.7938931297709924	0.765727963724147	0.7938931297709924	0.7750683528440687
random_forest	classification	genetic_algorithm	0.8080808080808081	0.8030519094690217	0.8080808080808081	0.8040239388708917
random_forest	classification	correlation_variance_filter	0.7557251908396947	0.6756633951290439	0.7557251908396947	0.7075576983170394
random_forest	classification	PCA	0.8015267175572519	0.7885712738887988	0.8015267175572519	0.7912313708722937
SVM	classification	correlation_variance_filter	0.7862595419847328	0.7787262312137279	0.7862595419847328	0.782235635244689
SVM	classification	PCA	0.732824427480916	0.7958726554574072	0.732824427480916	0.7560231756452505
SVM	classification	random_forest_elimination	0.7251908396946565	0.7580128964961358	0.7251908396946565	0.7398609988765293
SVM	classification	genetic_algorithm	0.8282828282828283	0.8275340820795367	0.8282828282828283	0.8277816928979721
SVM	classification	Factor_Analysis	0.7557251908396947	0.7452477174075737	0.7557251908396947	0.7491139585605234

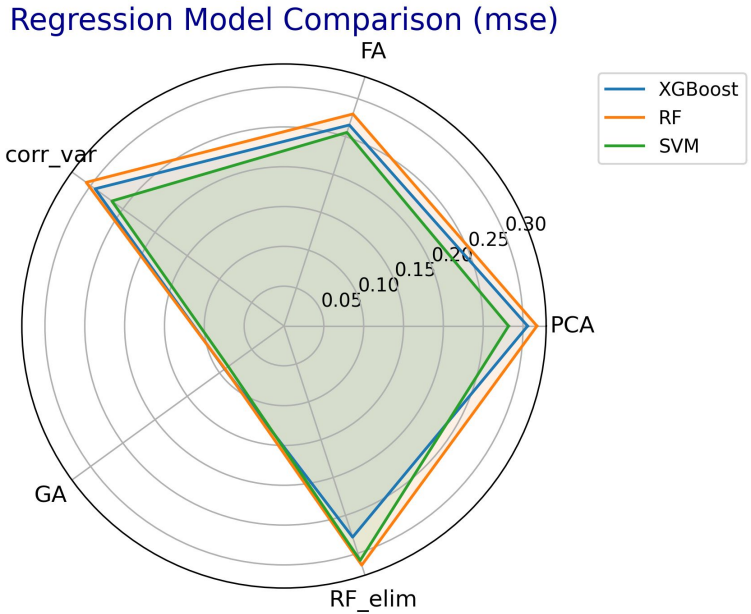
Regression Random Forest on PCA feature selection

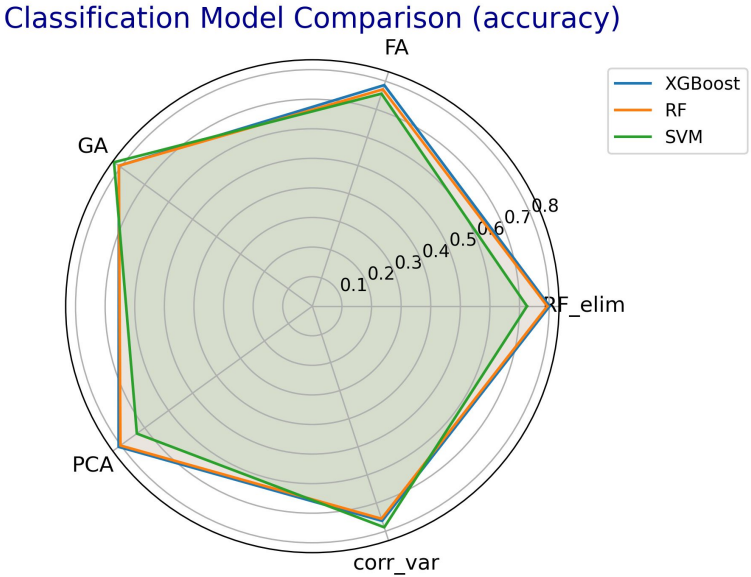


Classification SVM on genetic algorithm feature selection



Model comparison





Conclusion

- Using publicly available TRPM8 targeting drugs, features were calculated and reduced using multiple methods then modeled using XGBoost, Random Forest, Support Vector Machine
- It’s easier to predict high or low potency than medium potency
- For regression, Random Forest model performed best
- For classification, most methods performed well
- The models performed well with accuracies and correlations with experimental data compared to state-of-the art