Machine Learning for Molecular Property Prediction

Nga Ngo*, Kha Dinh Luong†, and Ambuj Singh†
*Department of Chemical Engineering, UCSB, Santa Barbara, CA 93106, U.S.A.
†Department of Computer Science, UCSB, Santa Barbara, CA 93106, U.S.A.

DYNAMO Dynamic Networks: Analysis and Modeling

CSEP CNSI

Framework for Applying ML in Cheminformatics

- To develop efficient predictive modeling methods for molecular property prediction with applications in drug discovery.
- To translate molecules into mathematical representations and apply different ML algorithms to recognize patterns between those representations and their properties.
- The goal of the study is to reduce the need to synthesize a large number compounds during the drug development process.
- The field is relatively new and there is still a need for good benchmarks and comparisons between techniques. This project attempts to **reproduce and benchmark different methods across different datasets.**

Datasets

Dataset	Prediction Tasks	Tasks	Compounds	Properties	Type
MUV	Virtual screening	17	93,127	Biophysics	classification
HIV	Ability to inhibit HIV replication	1	41,913	Biophysics	classification
BBBP	Permeability	1	2,053	Physiology	classification
Tox21	Toxicity measurements	12	8,014	Physiology	classification
SIDER	Marketed drugs and adverse drug reactions (ADR)	27	1,427	Physiology	classification
QM8	Electronic spectra and excited state energy	12	21,786	Quantum Mechanics	regression
ESOL	Water solubility	1	1,128	Physical Chemistry	regression
LIPO	Membrane permeability and solubility	1	4,200	Physical Chemistry	regression

Translation of SMILES to Fingerprints

num	name	p_np	smiles	fingerprints
1	Propanolol	1	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2	Terbutylchlorambucil	1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3	40730	1	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
4	24	1	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
5	cloxacillin	1	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, ...
...	...	...	...	...
2049	licostinel	1	C1=C(Cl)C(=C(C2=C1NC(=O)C(N2)=O)[N+](=O)[O-])Cl	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2050	ademetionine(adenosyl-methionine)	1	[C@H]3([N]2C1=C(C(=NC=N1)N)N=C2)[C@@H]([C@@H](...	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2051	mesocarb	1	[O+]1=N[N](C=C1[N-]C(NC2=CC=CC=C2)=O)C(CC3=CC=...	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2052	tofisoline	1	C1=C(OC)C(=CC2=C1C(=[N+](C(=C2CC)C)[NH-])C3=CC...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2053	azidamfenicol	1	[N+](=NCC(=O)N[C@@H]([C@H](O)C1=CC=C([N+]([O-]...	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...

Create fingerprints using SMILES

Extracted from the BBBP dataset.

Append fingerprints column to dataset

num	name	p_np	smiles	fingerprints
1	Propanolol	1	[Cl].CC(C)NCC(O)COc1cccc2ccccc12	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2	Terbutylchlorambucil	1	C(=O)(OC(C)(C)C)CCCc1ccc(cc1)N(CCCl)CCCl	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3	40730	1	c12c3c(N4CCN(C)CC4)c(F)cc1c(c(C(O)=O)cn2C(C)CO...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
4	24	1	C1CCN(CC1)Cc1cccc(c1)OCCCNC(=O)C	[0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
5	cloxacillin	1	Cc1onc(c2ccccc2Cl)c1C(=O)N[C@H]3[C@H]4SC(C)(C)...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, ...
...	...	...	...	...
2049	licostinel	1	C1=C(Cl)C(=C(C2=C1NC(=O)C(N2)=O)[N+](=O)[O-])Cl	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2050	ademetionine(adenosyl-methionine)	1	[C@H]3([N]2C1=C(C(=NC=N1)N)N=C2)[C@@H]([C@@H](...	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2051	mesocarb	1	[O+]1=N[N](C=C1[N-]C(NC2=CC=CC=C2)=O)C(CC3=CC=...	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2052	tofisoline	1	C1=C(OC)C(=CC2=C1C(=[N+](C(=C2CC)C)[NH-])C3=CC...	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2053	azidamfenicol	1	[N+](=NCC(=O)N[C@@H]([C@H](O)C1=CC=C([N+]([O-]...	[0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...

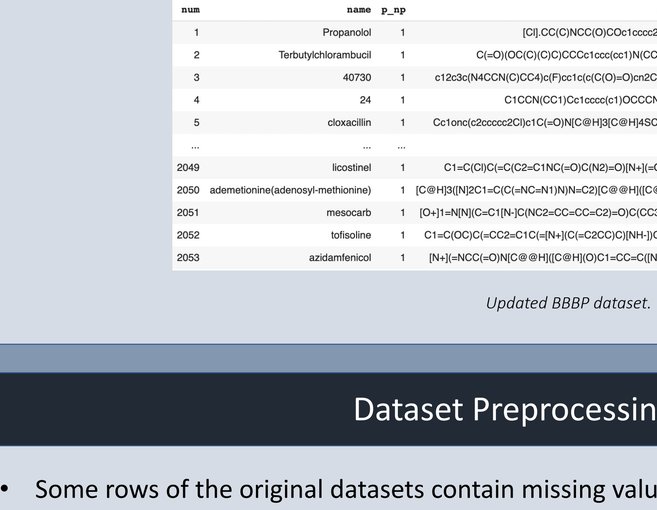
Updated BBBP dataset.

Dataset Preprocessing

- Some rows of the original datasets contain missing values.
- The datasets are modified to exclude the missing values before being passed into the ML models.

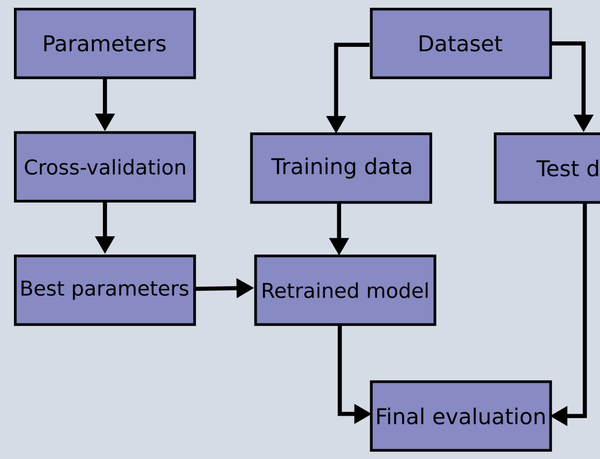
Exhaustive Search for Optimized Parameters

- Hyperparameter tuning to determine the optimal values for the ML models.
- The parameters are optimized by cross-validated grid-search over a parameter grid.



- Grid search with a “fit” and “score” method.
- Grid search loops through the predefined hyperparameters and fits the model on the training set.
- In the end, the best combination is retained.

Preventative Measure Against Overfitting



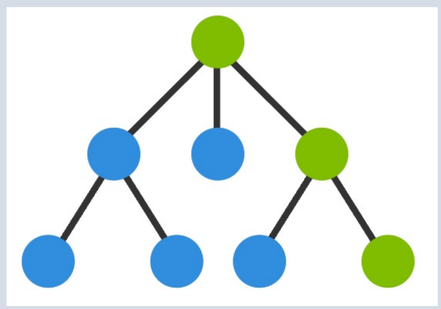
- Overfitting occurs when the model is trained so much to the existing data that it loses its ability to generalize cannot perform accurately against unseen data.
- Cross-validation can be used to assess the performance of the model with an unknown dataset.
- Cross-validation partitions the data into multiple folds, reserves one fold and trains the model on the remaining folds, then tests the model on the holdout fold.
- The process is repeated for each fold in the dataset.

Classification and Regression Algorithms

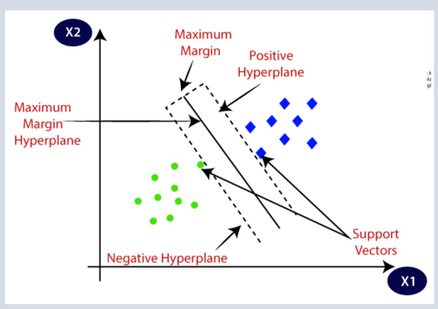
Classification
DecisionTreeClassifier
SVC
RandomForestClassifier

scikit learn

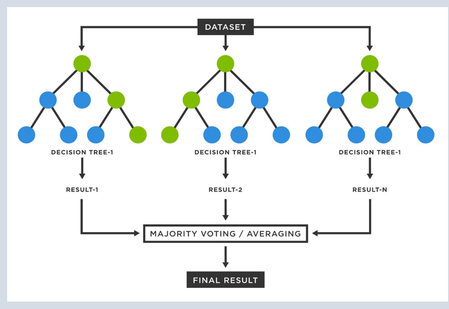
Regression
DecisionTreeRegressor
SVR
RandomForestRegressor



Decision Tree



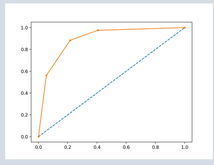
SVM



Random Forest

Performance Evaluation Metrics

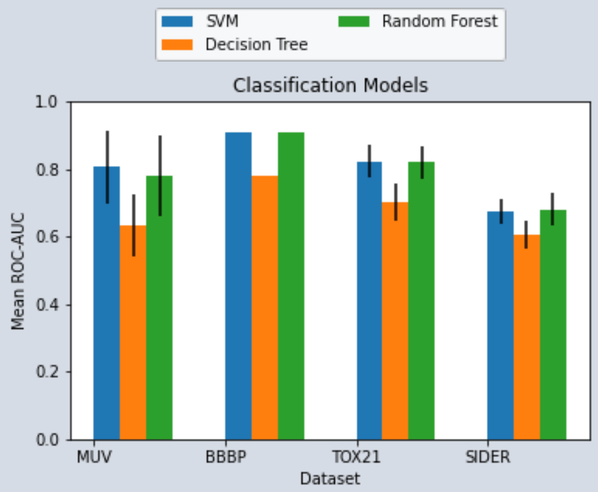
Metric for Classification Models:
Area under the curve of receiver operating characteristic curve (ROC-AUC)



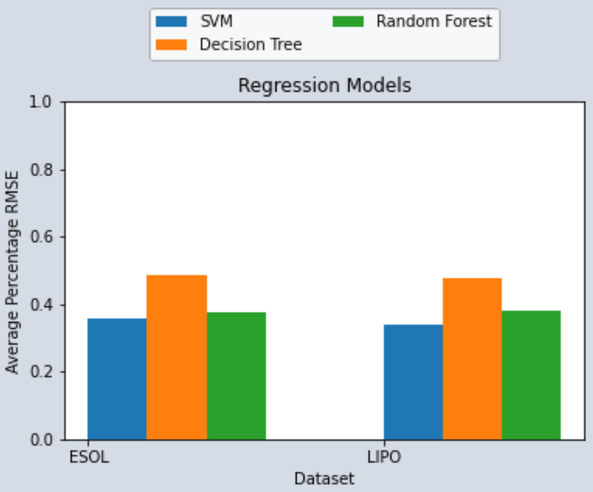
Metric for Regression Models
Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Predictive Performance Comparison of Machine Learning Models



Predictive performance of machine-learning approaches. ROC-AUC is the metric used. The higher the y-axis, the better the model performs.



Regression errors of machine-learning approaches. RMSE is the metric used. The lower the y-axis, the better the model performs.

What’s Next?

- If we can continue to optimize the performance of these models, we will be able to profile molecules and efficiently obtain their properties in the future without having to physically synthesize them in a laboratory.
- This work provides a good benchmark for future work.

Acknowledgements

This work was supported in part by **California NanoSystem Institute, Center for Science and Engineering Partnerships, and EUREKA.**

References

J. Shen *et al.*, Drug Discovery Today: Technologies, 32-33, 29–36

M. Withnall *et al.*, Journal of Cheminformatics 12(1)

Z. Wu *et al.*, Chemical Science, 9(2), 513–530