ADMET Property Prediction With Oloren ChemEngine

Andrew Li, David Huang

Oloren AI

Abstract

We present a set of model architectures built using the Oloren ChemEngine library for the Therapeutics Data Commons (TDC) benchmarks. We evaluate the models on ADMET benchmarks, representing various small molecule drug properties encompassing absorption, distribution, metabolism, excretion, and toxicity. The models predict values for these properties given the chemical structure of the molecule. The presented models draw from a wide set of state-of-art strategies, such as custom model gradient boosting, graph neural networks, molecular fingerprinting, and OlorenVec, a supervised learned molecular representation. We show that these models are superior to current leaderboard model approaches.

Introduction to OCE

Oloren ChemEngine[1] (OCE) is an open-source Python software package developed by Oloren AI for developing and using molecular property predictors. OCE implements current state-of-art methods such as message passing neural networks, graph neural networks, and molecular fingerprints and representations. It additionally allows for the creation of custom gradient boosting models consisting of user-defined submodels for each learner stage.

The TDC[2] ADMET benchmarks consist of 22 datasets composed of drug chemical structures and the property value for their respective ADMET property. Dataset sizes range from hundreds to ~15,000 molecules. We present the models which achieve a rank of first-place in 13 of these tasks, and top 3 in 5 tasks.

Model Architectures

We construct and test 8 unique models. This set of models consists of 1 graph neural network, 1 Extreme Gradient Boosting model (XGBoost), and 6 custom gradient boosting models composed of various submodels.

The models are represented in the table below:

Model Name	Model Description
BaseBoosting hMCiM8io	Random Forest (RDKit2DNormalized), Random Forest (Morgan Fingerprint), Random Forest (OlorenVec), SPGNN
BaseBoosting ADkCCrwJ	Random Forest (RDKit2DNormalized), Random Forest (Morgan Fingerprint), Random Forest (OlorenVec)
BaseBoosting OQRAYLPP	SPGNN, Random Forest (OlorenVec), BaseTorchGeometricModel (GINModel)
BaseBoosting xXOn1QFI	Random Forest (Morgan Fingerprint), Random Forest(Morgan Chiral Counts), Random Forest (Morgan Feature Counts), Random Forest (RDKit2DNormalized), Random Forest (OlorenVec)
BaseBoosting 1zpI0dIb	Random Forest (Morgan Fingerprint), Random Forest (Morgan Chiral Counts), Random Forest (Morgan Feature Counts), Random Forest (RDKit2DNormalized), Random Forest (OlorenVec)
BaseBoosting ZSCWPcLT	Random Forest (Morgan Fingerprint) Random Forest (Morgan Chiral Counts) Random Forest (Morgan Feature Counts) Random Forest (RDKit2DNormalized) Random Forest (OlorenVec)
BaseBoosting ktIIq91G	Random Forest (Morgan Fingerprint), Random Forest (Morgan Feature Counts), Random Forest (Morgan Chiral Counts), Random Forest (RDKit2DNormalized), Random Forest (Mordred Descriptor), Random Forest (OlorenVec)
ZWK_XGBoostModel u3zq9AAV	XGBoost Model with RandomizedSearchCV hyperparameter tuning.
RFStacker jTNhN7U7	Random Forest (RDKit2DNormalized) Random Forest (OlorenVec) Random Forest (Lipinski Descriptor) Random Forest (Morgan Feature Counts) SPGNN

Model names are defined as OCE's uniquely generated name for a given model. BaseBoosting models are custom-made gradient boosting models. SPGNN represents the graph neural network implemented in Hu et al. (2020)[3]. Certain gradient boosting models have identical model and feature architectures but differ in terms of the parameterization of the constituent submodels,

which can be observed in the source code. Random forest submodels are defined by the molecular fingerprint or representation that they use. BaseTorchGeometricModel is an implementation of the graph neural network in Brossard et al. (2021).

Results

Task	Model Name	Metric	Score	Stdev.	Leaderboard Rank
HIA	RFStacker jTNhN7U7	ROC-AUC	0.988	0.002	1
Clearance Microsome	RFStacker jTNhN7U7	Spearman	0.625	0.002	1
Pgp	BaseBoosting xXOn1QFI	ROC-AUC	0.946	0.001	1
BBB	BaseBoosting ADkCCrwJ	ROC-AUC	0.923	0.002	1
AMES	BaseBoosting 1zpI0dIb	ROC-AUC	0.865	0.002	1
DILI	BaseBoosting ADkCCrwJ	ROC-AUC	0.926	0.005	1
CYP3A4 Substrate	BaseBoosting ADkCCrwJ	ROC-AUC	0.679	0.033	1
Caco2	ZWK_XGBoost Model u3zq9AAV	Mean Absolute Error	0.289	0.011	1
CYP2C9 Inhibition	BaseBoosting 1zpI0dIb	Average Precision	0.791	0.005	1
CYP3A4 Inhibition	BaseBoosting ktIIq91G	Average Precision	0.882	0.001	1
CYP2D6 Substrate	BaseBoosting OQRAYLPP	Average Precision	0.711	0.006	1
Half Life	BaseBoosting ktIIq91G	Spearman	0.416	0.009	1
Clearance Hepatocyte	BaseBoosting xXOn1QFI	Spearman	0.491	0.006	1
Bioavailability_	BaseBoosting	ROC-AUC	0.736	0.004	2

Ma	1zpI0dIb				
hERG	BaseBoosting ADkCCrwJ	ROC-AUC	0.871	0.003	2
CYP2C9 Substrate	BaseBoosting ZSCWPcLT	Average Precision	0.419	0.007	2
PPBR	BaseBoosting ktIIq91G	Mean Absolute Error	8.503	0.023	2
LD50	BaseBoosting ktIIq91G	Mean Absolute Error	0.613	0.001	3

Source Code

OCE is open-source and accessible at https://github.com/Oloren-AI/olorenchemengine. Code to define, train, and use these models is available publicly at https://github.com/Oloren-AI/OCE-TDC.

References

[1] Oloren AI, Oloren ChemEngine, (2022), Github repository,

https://github.com/Oloren-AI/olorenchemengine

[2] Huang, Kexin, et al. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. arXiv, 28 Aug. 2021. arXiv.org, http://arxiv.org/abs/2102.09548.

[3] Hu, Weihua, et al. Strategies for Pre-Training Graph Neural Networks. arXiv, 18 Feb. 2020. arXiv.org, https://doi.org/10.48550/arXiv.1905.12265.

[4] Brossard, Rémy, et al. Graph Convolutions That Can Finally Model Local Structure. arXiv, 3 June 2021. arXiv.org, https://doi.org/10.48550/arXiv.2011.15069.