POLARIS ADMET 2025 CHALLENGE

REPORT: DrAshokAndFriends-SASTRAUniversity

1. Data collection and preprocessing: Datasets from diverse sources (incl. Polaris Train dataset, ChEMBL API, OChem, Deepchem, ADMET-AI, TDC, biogen and DrugBank) were used to prepare consolidated datasets for each ADMET endpoint of interest, namely MDR1-MDCKII permeation rate, KSOL, MLM, LogD and HLM.
2. Feature space from SMILES: Using a library of antivirals, 1826 features were computed using Mordred, and filtered for variance. Consensus with the 199 features used in Chemprop yielded a final set of 55 features. Both the raw SMILES as well as the feature spaces were used as inputs to ML models.
3. Model training: Two types of Graph neural networks (Attentive GNNs and GCNNs) were explored in addition to three types of Boosting models (XGBoost, CatBoost, and LightGBM), for each problem. Following a train-test split of 80:20, these five models were subjected to k-fold cross-validation and their hyperparameters were optimized. Then the models were evaluated on the hold-out test set. The best performing for each problem was identified and used to rebuild the model with the optimal hyperparameters on the full dataset.
4. The above procedure yielded XGBoost as the best performing model for four of the ADMET endpoints of interest, namely MDR1-MDCKII, KSOL, LogD, and HLM. The attentive GNN was the best model for predicting MLM. <u>These models were then applied on the Polaris test set to generate the test predictions for submission.</u>
5. Size of the consolidated datasets are provided below: MLM: 2082, HLM: 8107, LogD: 17356, MDR1-MDCKII: 1001, KSOL: 24740.
6. Model performance (in terms of adjusted R^2) is summarized in the below table:

Models	KSOL	LogD	HLM	MLM	MDR1-MDCK permeation
XGBoost	0.9729	0.9384	0.5698	0.6252	0.4427
CatBoost	0.9166	0.9009	0.5304	0.6349	0.4215
LightGBM			0.2	0.6236	0.0095
GNN	0.8968	0.8739	0.3584	0.7410	0.2796
GCNN	-2.1688	0.8679	0.1885	0.5173	0.1550