POLARIS - OPEN ADMET - ASAP DISCOVERY 2025 CHALLENGE

REPORT: SystemsCBLab – SASTRA University **[apalania]**

1. Dataset preparation:
    a. External data sources were consulted for each problem, and the units were harmonized. Final dataset sizes along with their provenance, are shown for each endpoint in the table below. SMILES duplicated with same endpoint values were removed.

DATASET	**LogD**	**KSOL**	**HLM**	**MLM**	**MDR1-MDCKII**
Polaris Train	434	434	434	434	434
DrugBank	2579	2579	2579		
ADMET-AI	4200	9982	1314		
TDC	4200	9982	1102		
BIOGEN		2173	3088		
ChEMBL				2058	126
OCHEM	6353				851
TOTAL	**17,766**	**25,150**	**8,517**	**2,492**	**1,411**

    b. Generation of feature space for tabular models was based on a custom model of antivirals:
        i. **Dataset Compilation:** Collected and cleaned antibiotic datasets from PubChem, ChEMBL, and other sources.
        Descriptor Calculation: Used **Mordred** (1826 descriptors) and compared with **Chemprop (**"A Deep Learning Approach to Antibiotic Discovery") (199 descriptors); identified **55 common descriptors**. Variance threshold was applied.

    c. Final feature space for tabular models:
    SlogP_VSA2, EState_VSA8, SlogP_VSA6, SlogP_VSA3, PEOE_VSA4, EState_VSA5, VSA_EState3, SMR_VSA6, BertzCT, SlogP_VSA8, EState_VSA1, VSA_EState1, VSA_EState8, PEOE_VSA8, SlogP_VSA11, PEOE_VSA1, VSA_EState7, VSA_EState6, EState_VSA3, EState_VSA10, SMR_VSA8, PEOE_VSA13, PEOE_VSA6, EState_VSA9, SlogP_VSA5, SlogP_VSA9, PEOE_VSA5, EState_VSA7, PEOE_VSA11, SlogP_VSA7,

PEOE_VSA2, SMR_VSA1, PEOE_VSA3, SlogP_VSA1, SMR_VSA9, VSA_EState2, EState_VSA2, TPSA, SlogP_VSA10, PEOE_VSA10, SlogP_VSA4, SMR_VSA7, VSA_EState9, EState_VSA6, SMR_VSA4, LabuteASA, VSA_EState5, SMR_VSA2, PEOE_VSA12, SMR_VSA5, PEOE_VSA7, SMR_VSA3, VSA_EState4, PEOE_VSA9, EState_VSA4.

2. Models explored and their hyperparameters
    a. **XGBoost**: "n_estimators": [250,500, 700, 800, 900, 1000], "max_depth": [3, 5, 7, 9, 11], "learning_rate": [0.01, 0.1], "subsample": [0.2,0.4,0.6,0.8, 1.0]
    b. **Catboost**: "iterations": [100, 200, 300], "depth": [3, 5, 7], "learning_rate": [0.01, 0.1, 0.2]
    c. **RandomForest** : n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap
    d. **LightGBM** models
    e. **AttentiveGNN**:in_channels, hidden_channels, out_channels, edge_dim, num_layers, num_timesteps, dropout
    -Training hyperparameters -batch_size, learning_rate, weight_decay, epochs, gradient_accumulation_steps, loss_function, optimizer.
    f. **GCNN** :epochs, batch_size, lr, patience, weight_decay, factor, num_workers.

3. Model training:
    a. Train-test split - 80:20
    b. 5-fold Cross-validation in Train set to optimize hyperparameter values.
    c. Loss function: MAE
    d. Early stopping with validation dataset, eval_metric: mae
    e. Rebuilding model with full dataset.

XGBoost: optimized hyperparameters for each learning problem:

ENDPOINT	learning_rate	max_depth	n_estimators	subsample
LOGD	0.1	9	700	1.0
KSOL	0.1	11	1000	1.0
MDR1-MDCKII	0.1	7	500	0.8
HLM	0.1	7	1000	1
MLM	0.01	7	1000	0.8

4. Model evaluation. MAE values are emphasized.

Models	KSOL	LogD	HLM	MLM	MDR1-MDCK II permeation
XGBoost: R^2	0.9809	0.9232	0.5358	0.6240	0.4301
and **MAE**	**0.2060**	**0.0345**	**0.5578**	**1.3812**	**2.6249**
GNN: R^2	0.8968	0.8739	0.3584	0.7410	0.2796

Thus XGBoost was identified as optimal for KSOL, LogD, HLM, and MDR1-MDCKII endpoints, whereas the Attentive GNN model was identified as optimal for MLM endpoint.
The optimal models were applied on the Polaris Test set to generate the predictions.

TEAM: DrAshokAndFriends-SASTRAUniversity