Sina Khosravi

Department of Medicinal Chemistry, School of Pharmacy, Mashhad University of Medical Sciences, Mashhad, Iran

GitHub Gmail : Khosravi.sina2001@gmail.com

Initially, I analyzed the data using the DataWarrior software to evaluate the scaffold diversity and the distribution between the training and test sets. The analysis revealed a good overlap between the test and train sets, confirming that the applicable domain was satisfied. Upon further inspection of the data for each protein individually, I encountered issues with NaN values and duplicates. I removed the NaN entries and handled duplicates by calculating the mean for pairs and the median for groups of more than two, while attempting to eliminate the more outlier-prone layers.

Next, I proceeded to calculate various descriptors. For this, I utilized the OCHEM web platform, the scikit-fingerprints package, and embedding features from Chemprop. During data preparation and prior to model training, I removed features that were constant (i.e., all values identical), as well as those with variance below 0.01 or correlation exceeding 95%. Generally, I employed a consensus model approach, averaging the results of individual models without applying any weights. Below, I outline the specific methods used for each protein.

SARS-CoV-2 Mpro

1. **XGBoost Model**: Utilized all extracted features, followed by feature selection using SHAP values, retaining the top 3.5% of features. This reduced the feature count from 5,440 to 191. Cross-validation (5-fold) score: **0.366**.
2. **LightBoost Model**: Followed a similar approach with all extracted features, selecting the top 3% of SHAP values, reducing features from 5,540 to 164. 5-fold CV score: **0.358**.
3. **CatBoost Model**: Instead of using all features, an optimal combination was identified using a genetic algorithm (-map4-mold2-Estate). Subsequently, the top 20% of SHAP values were selected, reducing features from 1,376 to 275. 5-fold CV score: **0.401**.
4. **SVR Model**: Employed a combination of descriptors (PubChem, ECFP, MACCS) without additional feature selection, training on 1,109 features. 5-fold CV score: **0.423**.

MERS-CoV Mpro

1. **XGBoost Model**: Used all extracted features, followed by feature selection with SHAP, retaining the top 4.5% of features, reducing the count from 5,599 to 252. 5-fold CV score: **0.4622**.
2. **LightBoost Model**: Applied the same method as above, selecting the top 2.5% of SHAP values, reducing features from 5,599 to 140. 5-fold CV score: **0.4655**.
3. **CatBoost Model**: Identified an optimal feature combination using a genetic algorithm (map4-mold2-RDKit-FFn(Chemprop)). The top 20% of SHAP values were selected, reducing features from 1,672 to 335. 5-fold CV score: **0.441**.
4. **SVR Model**: Used a combination of descriptors (PubChem, ECFP, MACCS) without feature selection, training on 1,186 features. 5-fold CV score: **0.4562**.
5. **XGBoost with Clustering**: First, reduced data dimensionality to two using PubChem descriptors and PCA. Then, clustered the train and test data using the nearest neighbor algorithm. For each cluster, the best features were selected using SHAP, and the model was trained accordingly. Results:
   - **Cluster 0**: Train = 285, Test = 133; 5-fold CV = **0.44**.
   - **Cluster 1**: Train = 562, Test = 164; 5-fold CV = **0.46**.