ML4 Science 2023 Hackathon

Team: sMLe_pOdu

Sandeep Nagar, Litralson E R

Flow overview:

Literature review

Understanding dataset, plotting the dataset (properties) to visualize.

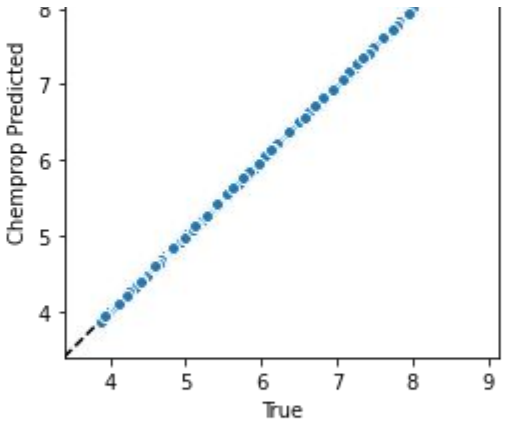
Dataset representation for ML methods

Understanding ML models and best fit to solve our problem.

Challenges and solution.

Methods:

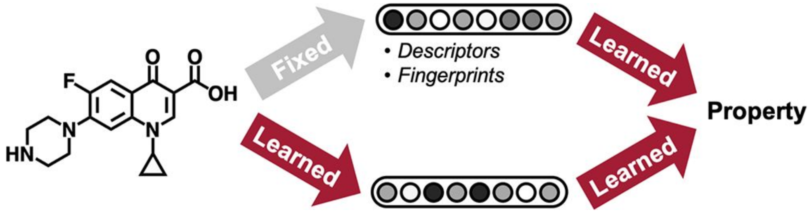
- Linear regression
- MLP
- MLT-BERT
- Chemprop: D-MPNN



Model overview:

- Message-Passing Neural Networks for molecular property prediction.
- Neural networks: applied to computed molecular fingerprints.
- Iteratively aggregate local chemical features to predict properties.
- From RDkit Additional Molecular Features.
- Predicted uncertainty distribution
- Metrics: rmse (default), mae, mse, r2, bounded_rmse, bounded_mae, bounded_mse.

Why Chemprop:



1. bond-level message passing
2. Hyperparameter Optimization
3. Ensembling
4. Better than 19 different baseline models

Challenges: overfitting, model size, small dataset

Solutions: dropout regularization, few layers, Cross-Validation
dataset split type (random & scaffold)