Polaris ASAP Discovery x OpenADMET challenge Report

For this submission, we train a multi-head MLP to predict the log-transformed labels from a latent representation. We benchmarked multiple different small molecule representations and submitted the predictions made on the representation with the best performance in a cross-validation setting.

Data Preparation:

We start by running our internal SMILES standardization pipeline on the input SMILES. Prediction labels are also log-transformed to simulate the setting on which we will be evaluated.

Latent Representation:

We compute multiple latent representation for the SMILES using different models:

- `GraphMiniMol`
- `ChemBERTa`
- `MoleculeDescriptors`
- `AtomPairFingerprint`
- `AvalonFingerprint`
- `MolFormer`

GraphMiniMol seemed to be the best performing representation.

Prediction Model:

Rather than training a prediction model for each label, we train a single multi-head MLP to leverage a shared representation layer and learn from multiple labels at once. All labels are scaled together and the model is trained to minimize L1 loss.
We also tried single task MLPs as well as XGB, RandomForest and Linear Regression classifiers, but the multihead-MLP seemed to be the best performing model in our crossvalidation setting. We did not tweak many parameters of the MLP or any of the other models, mainly wanting to look at the impact of various different latent representation.

Cross-validation:

Cross validation was run on scaffold splits, to assure better generalization of the chosen model.

Test predictions were then re-transfomed to remove the log-transform (apart from LogD).