BIOINF 595 Final Project A: Curation of Bioactivity Data

Bioactivity data is the basis for training and evaluating machine learning models for drug discovery. For this half of the final project, the goal is to identify, curate and train models for new bioactivity datasets.

Learning objectives

- Building on Labs 4-6, test your understanding of the process
- Develop skills to be able to work with new data that is relevant for your research project
- Contribute to improving drug discovery through making data more accessible

Steps

Select a dataset

The dataset should:

- involve either experimental, simulated, or QM calculations of the properties of molecules.
- be publicly available with a permissive license (or seek permission from creator)
- not already curated into an ML ready public repository (e.g. MaomLab ChemPile, TDCommons).
- Note: It is ok if that data is in ChEMBL, PubChem, but the dataset should be transformed in someway, e.g. curating different PubChem screens into a common dataset, or filtering for ChEMBL activities that interact with metal cofactors.

Here is a list of possible datasets BIOINF595w25 Bioactivity ML Datasets.

- Most of these datasets should satisfy the requirements, but still check that they do
- You can choose a dataset that is not on the list, confirm with Prof. O'Meara first and add it to the list

Once you have selected a dataset

- Add your name to the BIOINF595w25 Bioactivity ML Datasets spreadsheet to claim it.

Download the dataset and prepare it

Follow the instructions here for curating and uploading the dataset to HuggingFace

Make sure to

- Sanitize molecules e.g. using MolVS
- Split the data into test/train subsets if it is not already split (e.g. using splito)
- include detailed and useful metadata and README.md

Train and evaluate a baseline model

Use the data to train a baseline machine-learning classifier or generative model to predict the data

- You can use H2O AutoML or other ML focused predictor including chemprop, or Uni-Mol tools
- key decisions
    - molecule representation / features

    - training and evaluation objectives (e.g. loss function and interpretation of model quality)
    - model selection and hyperparameters
    - training strategy including test/train splitting
- Evaluate the trained model
    - Performance over the dataset and relevant data subsets
    - Example predictions and how they can be used interpreted
    - Comparison to published models for the same or related datasets

Report dataset and analysis

Create a technical report of the dataset and model training. It should clear and concise with relevant technical details (1-2 pages). Include the following sections

- Description of dataset
    - Broad problem/question the data seeks to inform
    - Data source / generation / curation strategy
    - Specifics about the curation and resulting dataset (subsets, number of samples, data columns)
- Description of model development
    - Justify key modeling decisions
    - Dataset specific challenges if any
- Results
    - Context of performance for related models
    - Overall performance and performance for relevant subsets
    - Example predictions and their interpretation
- Describe opportunities/limitations and potential use cases of the dataset
- A link to the HuggingFace dataset, which contains
    - Scripts to reproduce curation and model fitting, and analysis