Frederick National Laboratory for Cancer Research (FNLCR) Project Description

Drug Discovery with AMPL: Combination of Disparate Chemoinformatics Datasets for Chemical Property and Activity Prediction *(Estimated time: ~ 16 weeks; Suggestions: teaming or pairing of students is encouraged)*

The pharmaceutical industry has invested heavily in the development of large chemical libraries with millions of compounds in order to search for new drugs. Chemical compounds are run through high-throughput biological assays in order to identify compounds with activities towards molecular targets for diseases. This has resulted in large chemoinformatics datasets which include chemical structure and associated biological responses from high throughput screening. High throughput screens have been useful to not only identify compounds for further development into a drug, but as data sources for predictive machine learning models.

The field of Quantitative Structure Activity Relationships (QSAR) leverages the massive investment in these datasets to fit machine learning models that predict the properties and biological activity of new compounds before they are synthesized. The ATOM Consortium (Accelerating Therapeutics for Opportunities in Medicine) is dedicated to using machine learning algorithms to accelerate the drug discovery process. ATOM has developed a machine learning platform called the ATOM Modeling Pipeline (AMPL) which is used to curate chemoinformatics datasets and create machine learning models.

Public chemoinformatics datasets come from many sources including scientific literature and patents. Disparate sources and conditions make curation and modeling of these datasets difficult. In this project, students will curate and model new public chemoinformatics datasets from multiple sources. Students will learn to work with scientific data including chemical and biological databases. Ultimately the models developed will be applied to drug discovery projects to increase the chances of finding effective treatments for diseases.

Benefits to the community

- AMPL scripts, data and models from this project will be shared with the community

Deliverables

- Collected chemoinformatics datasets of priority to ATOM drug discovery projects
- Characterization and curation of associated chemoinformatics datasets
- Scripts for exploratory data analysis (EDA), machine learning model fitting and hyperparameter optimization
- Evaluation of accuracy and domain of applicability of machine learning models

Supporting documents

- ATOM
- Accelerating Therapeutics for Opportunities in Medicine: A Paradigm Shift in Drug Discovery

- Chemoinformatics dataset sources:
    - PubChem: https://pubchem.ncbi.nlm.nih.gov/
    - NCATS: https://ncats.nih.gov/expertise/covid19-open-data-portal
    - CheMBL: https://www.ebi.ac.uk/chembl/