ALY 6080: Integrated Experiential Learning

Annotated Bibliography I

Submitted to

Bryce Allen

Submitted by

Md Tajrianul Islam

Date: Sep 30, 2020

Annotated Bibliography I

I. Summary

Discovery of new compounds to treat human disease is a multifaceted process involving the selection of chemicals with favorable pharmacological properties: a high potency to the desired target, elimination or minimization of safety liabilities, and a favorable pharmacokinetic (PK) profile. To address this challenge, the drug discoverer has a wealth of choices, with total "drug-like" chemical matter estimated between 1022 and 1060 unique molecules. However, evaluating the desirability of these molecules with respect to potency, pharmacokinetics, and safety liabilities is a time-consuming and expensive process. Many of these molecules require de novo synthesis, which is a rate-limiting step. Furthermore, evaluation of pharmacological properties both in vitro and especially in vivo is prohibitively expensive given the universe of possible choices. To aid in this design challenge, the field of computer-aided drug design has evolved to rapidly predict the properties of pharmacological matter in silico. These techniques generally fall into two categories: (1) structure-based drug design, which relies on knowledge of the target structure (i.e., docking, molecular dynamics, free energy perturbation), and (2) ligand-based drug design, which uses known properties of molecules to develop models of quantitative structure−activity relationships (QSAR). In this paper, we introduce a new small molecule property prediction pipeline, AMPL. This software was developed through the Accelerating Therapeutics Opportunities in Medicine (ATOM) Consortium as the ATOM Modeling PipeLine.

AMPL includes several modules to curate data into machine learning-ready data sets. Functions are provided to represent small molecules with canonicalized SMILES strings using RDKit9 and the MolVS package,10 by default stripping salts and preserving isomeric forms. Data curation procedures are provided with AMPL as Jupyter notebooks,11 which can be used as examples for curating new data sets. Procedures allow for averaging response values for compounds with replicate measurements and filtering com- pounds with high variability in their measured response values. AMPL also provides functions to assess the structural diversity of the data set, using either Tanimoto distances between fingerprints or Euclidean distances between descriptor feature vectors. AMPL supports two kinds of interaction with external featurizers: a dynamic mode in which features are computed on-the-fly and a persistent mode whereby features are read from precomputed tables and matched by compound ID or SMILES string. AMPL includes a train/ tune/ predict framework to create high-quality models. This framework supports a variety of model types from two main libraries: scikit-learn and DeepChem. Currently, specific input parameters are supported for the following:

- Random forest models from scikit-learn
- XGBoost models
- Fully connected neural network models
- Graph convolution neural network models

A module is available to support distributed hyperparameter search for HPC clusters. This module currently supports linear grid, logistic grid, and random hyperparameter searches, as well as iteration over user- specified values.

II. References

1. https://pubs.acs.org/action/showCitFormats?doi=10.1021/acs.jcim.9b01053&ref=pd