Safety, Reproducibility, Performance: Accelerating cancer drug discovery with ML and HPC technologies

Minnich, Amanda J. [1]

[1]Lawrence Livermore National Laboratory, ATOM Consortium

1. Introduction

The drug discovery process is costly, slow, and failure-prone. It takes an average of 5.5 years to get to the clinical testing stage, and in this time millions of molecules are tested, thousands are made, and most fail. The ATOM Consortium is working to transform the drug discovery process by utilizing machine learning to pretest many molecules *in silico* for both safety and efficacy, reducing the costly iterative experimental cycles that are traditionally needed. This consortium is comprised of LLNL, GlaxoSmithKline, NCI's Frederick National Laboratory for Cancer Research, and UCSF. Through ATOM's unique combination of partners, machine learning experts are able to use LLNL's supercomputers to develop models based on proprietary and public pharma data for over 2 million compounds. The goal of the consortium is to create a new paradigm of drug discovery that would drastically reduce the time from identified drug target to clinical candidate, and we intend to use oncology as the first exemplar of the platform.

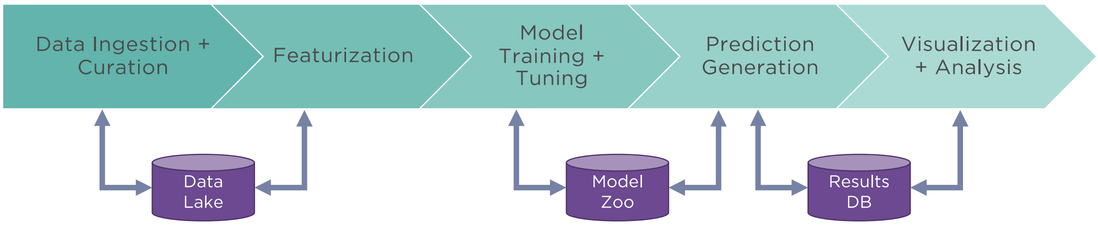

Figure 1 *Overall schema of our data-driven modelling pipeline*

To this end, we have created a computational framework (Figure 1) to build ML models that generate all key safety and pharmacokinetics parameters needed as input for Quantitative System Pharmacology and Toxicology models. Our end-to-end pipeline first ingests raw datasets, curates them, and stores the result in our data lake. Next it extracts features from these data and trains and saves the model to the model zoo. Our pipeline generates a variety of molecular features and both shallow and deep ML models. We are currently using a Python library Deepchem (Zhenqin Wu), which acts as a molecular data-specific wrapper for TensorFlow (Martín Abadi), to build our models. The HPC-specific module we have developed conducts efficient parallelized search of the model hyperparameter space and reports the best-performing hyperparameters for each of these feature/model combinations. We use NVIDIA Tesla P100 GPU nodes to speed up training of our neural net models, allowing us to conduct efficient hyperparameter searches over a large combinatorial parameter space.

Using this pipeline, we have created tens of thousands of deep learning and random forest models, which predict a variety of key safety and pharmacokinetic parameters with a high degree of accuracy. By leveraging the large number of nodes available on LLNL's supercomputers, we have been able to comprehensively explore a variety of feature types and model types. For neural nets, we have tested a wide range of model architectures, dropout schemes, and learning rates. We also optimized our random forest baseline models, exploring a variety of forest sizes and max depth thresholds. Through this extensive exploration, we have been able to improve the performance of our models significantly and generate neural net models that out-perform these random forest baseline models. We are now able to provide a robust survey of dataset-dependent performance for our curated pharma datasets.

To ensure complete traceability of results, we save the training, validation, and testing dataset version IDs, the Git hash of the code used to generate the model, and the OS- and library-related version information. We have set up a Docker/Kubernetes (Kelsey Hightower) infrastructure, so when a promising model has been identified, we can encapsulate the pipeline that created it, supporting both reproducibility and portability. Our system is designed to handle protected data and support incorporating proprietary models, which allows the framework to be run on real drug design tasks.

Using a containerized workflow also empowers rapid data science software development. Using Docker, we are able to stand up a variety of data services, such as MongoDB and MySQL servers, as well as an in-house datastore service that allows for the association of metadata with any type of file, including raw data files, serialized models, and stored training and testing data frames. Docker allows efficient updating of software packages while retaining containers with previous software versions, ensuring traceability and reproducibility of results. Kubernetes allows orchestration of compute resources, including GPU nodes, and access to data services, as well as the restriction of access to sensitive data. This infrastructure has been crucial in the building of our HPC-driven computational drug discovery pipeline.

The best-performing models (subset of classification results presented in Figure 2) are currently being integrated into an active learning pipeline to aid in de novo compound generation, as well as being sent back to consortium members to incorporate into their drug discovery efforts. The predicted PK parameters are also being used as input to our in-house PBPK models to predict *in-vivo* behavior. We are also currently preparing our code base and models for open source release. To make these models usable externally, we have built a module that can load in a model from our model zoo or disk and generate predictions for a list of compounds on-the-fly. If ground truth is known, a variety of performance metrics are generated and stored in our model performance tracker database, allowing for easy querying and comparison of model performance.

We are confident that our work building up internal infrastructure and software packages will help to transform cancer drug discovery from a time-consuming, sequential, and high-risk process into an approach that is rapid, integrated, and with better patient outcomes.

Disclaimer: The opinions or assertions contained herein are the private ones of the author/speaker and are not to be construed as official or reflecting the views of the Department of Defense, the Uniformed Services University of the Health Sciences or any other agency of the U.S. Government. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344.

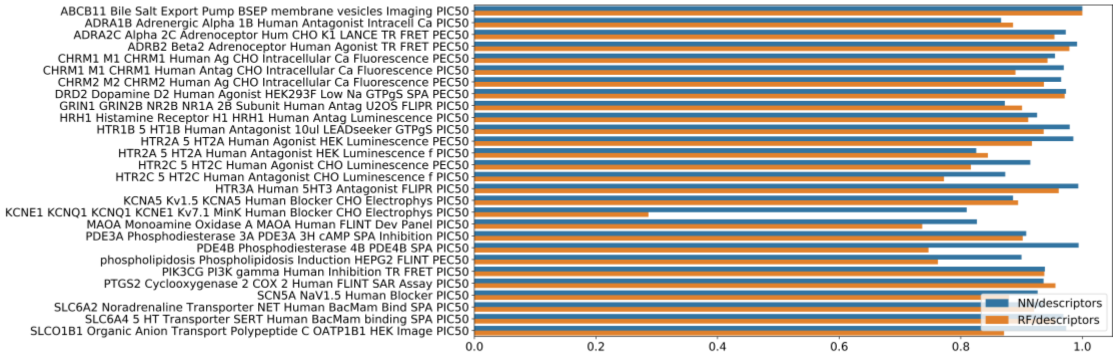

Figure 2. ROC-AUC scores for panel of safety target pIC50 and pEC50 classifications

2. References

Kelsey Hightower, Brendan Burns, and Joe Beda. *Kubernetes: Up and Running Dive into the Future of Infrastructure (1st ed.)*. O'Reilly Media, Inc., 2017.

Martín Abadi, et al. “TensorFlow: Large-scale machine learning on heterogeneous systems.” 2015. *http://tensorflow.org/.*

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. “MoleculeNet: A Benchmark for Molecular Machine Learning.” *CoRR* (2017).