PURDUE UNIVERSITY

ATOM

Accelerating Therapeutics for Opportunities in Medicine (ATOM)

Overview

In our research project we utilized DeepChem and AMPL to analyze hERG data. We constructed several types of single and multi-task models to then compare which more accurately predicts the data. We split the data by assay type, then train multi-task models to predict on other molecules.

Introduction

What is hERG data

- We are studying Human ether-a-go-go-related gene (hERG) data to find an accurate model that can predict on other hERG molecules as some assays have minimal data and are difficult to predict.

The motivation for modeling the single and multitask models

- Our hypothesis is that a multi-task prediction model will perform significantly better than a single task model of open sourced medical data.
- The multi-task model is assumed to make a prediction model more efficient by leveraging other assays to get higher coefficient of determination, R2, values.
- There are many obstacles in the way of drug discovery, including long lead times, skyrocketing expenses, and startlingly high failure rates. With only a 12% clinical success rate and an average development period of 6 years, traditional approaches are in need of a paradigm shift.

ATOM is an open public-private partnership for accelerating drug discovery

Goals
- Accelerate the drug discovery process
- Improve success rate in translation to patients

Approach
- Computation-driven drug design, supported and validated by targeted experiments
- Data-sharing to build models using everyone's data

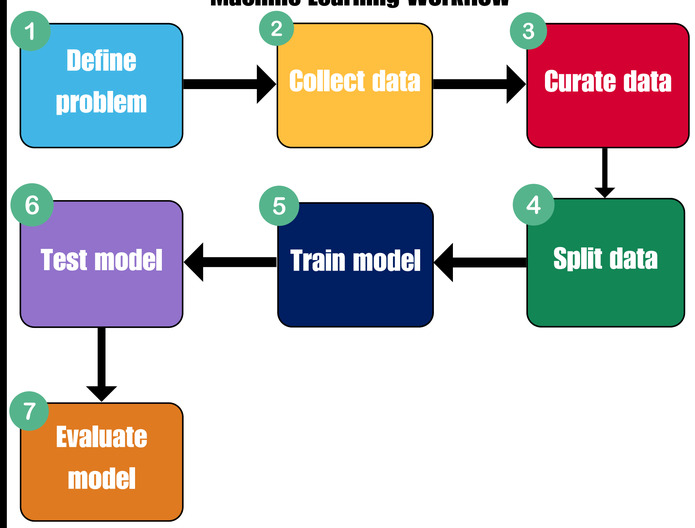
Product
- An open-source platform for active-learning based molecular design and optimization

Status
- Shared collaboration space at Mission Bay, SF
- R&D started February 2018

Example member contributions

Research Methodology

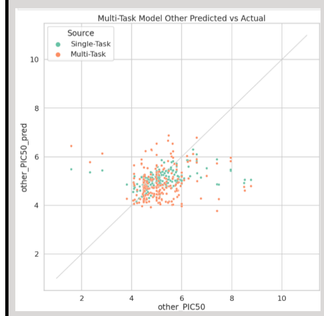
Machine Learning Workflow

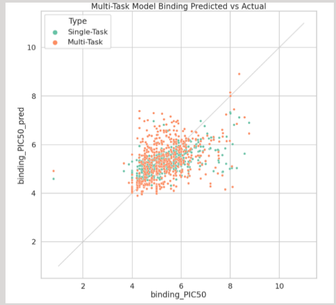


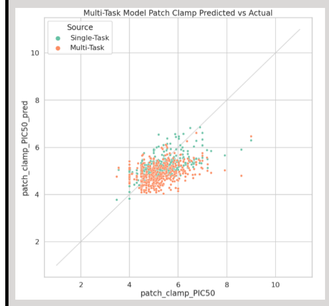
Research Methodology

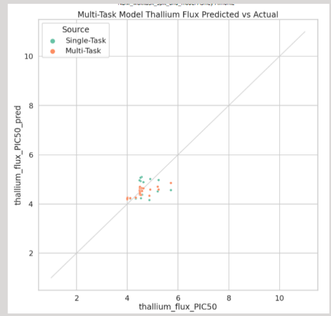
- Curation and featurization of data
- Split data using random or scaffold splitting
- Train and test various ML models:
  - Random Forest
  - Nueral Network
  - Single-Task vs. Multi-Task
- Tools: AMPL, DeepChem, rdkit, Ipython, uuid, and scipy

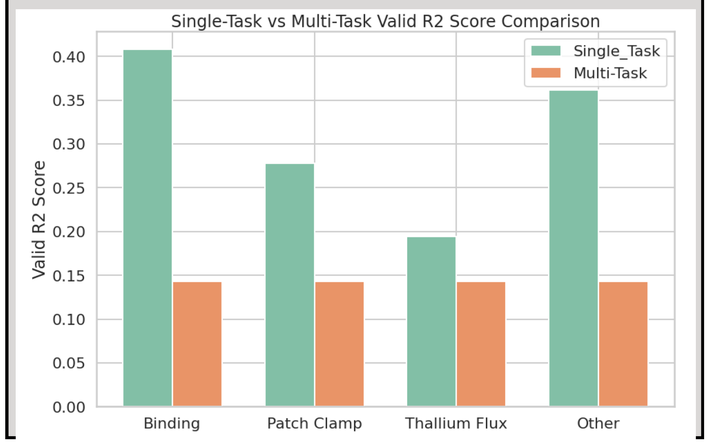
Results











Conclusion

- In conclusion, our goal at ATOM is to accelerate and streamline the process of drug development. This is important because of current challenges in drug discovery, such as lengthy development times, high costs, and low success rates, which necessitates predictive modeling as an effective, data-driven approach.
- Contrary to the initial hypothesis, hERG findings revealed the single-task model had a higher coefficient of determination (R2), indicating better performance.
- While multi-task modeling remained relatively constant, single-task modeling outperformed multi-task in Binding, Patch Clamp, Thallium Flux, and Other assays.

Future Steps

- Based on our findings, the single task model outperforms the multitask model, as indicated by its higher R2 score.
- Moving forward, our next course of action would involve expanding our ML models with additional data to validate these results.
- Additionally, we plan to investigate other molecule targets using these new sources of data.

Acknowlegements

Special thank you to our corporate partner mentor, Hiranmayi Ranganathan, our TA Fenil Gala, and Nicholas Lenfestey from the Data Mine Staff.

References

1. Minnich, Amanda J., et al. "AMPL: a data-driven modeling pipeline for drug discovery." Journal of chemical information and modeling 60.4 (2020): 1955-1968.
2. https://github.com/ATOMScience-org/AMPL