Using Machine Learning Models to Accelerate Drug Discovery Process

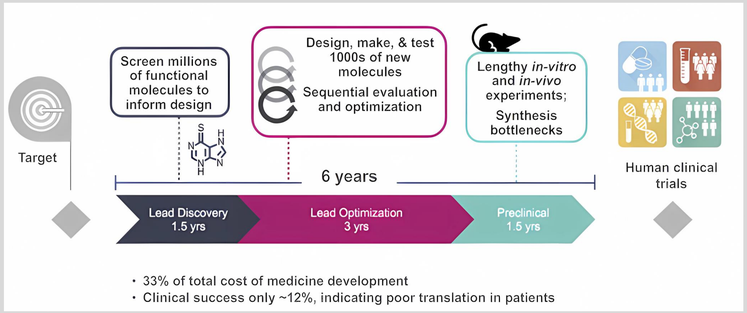
PURDUE UNIVERSITY | The Data Mine

Nihar Atri, Mariama Barry, Aryamaan Dhomne, Veerupaksh Singla, Phoebe Le (My Phuong Thi), Rohit Varanasy

ATOM

INTRODUCTION

ATOM is working to accelerate drug discovery, specifically molecule screening, by creating ML models for drug molecule inhibitory behavior, but their performance is yet to be evaluated.

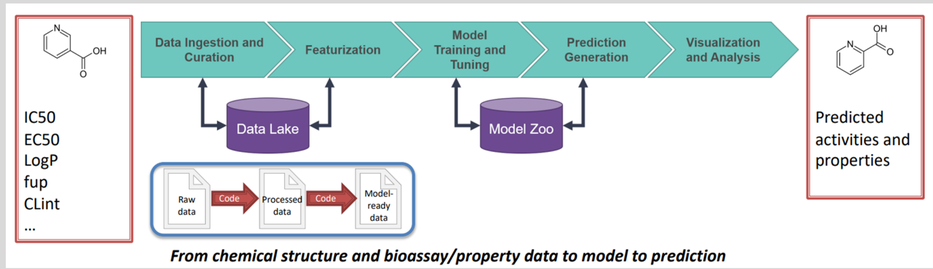


QUESTIONS

- How well do single task models predict inhibitory activity of drug molecules?
- How well do multi-task models to predict inhibitory activity of drug molecules?
- Do multi-task models perform better than single-task models?
- What descriptor performed the best?

METHODOLOGY

We first conducted exploratory data analysis, and then trained and evaluated single and multi-task models for NEK kinases.



From chemical structure and bioassay/property data to model to prediction

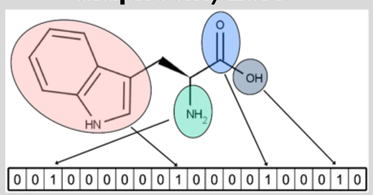
Fall 2022: Single-Task Models/Target Research

- Made heatmaps to compare and group NEK Kinase targets based on similarity
- Trained single-task (predicts on one task) random forests and XGBoost models with different featurizers

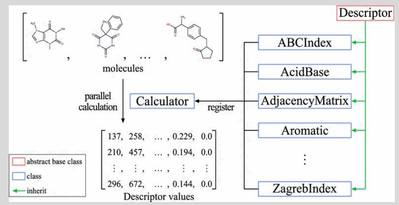
Spring 2022: Multi-Task Models

- Trained multi-task (predicts on multiple tasks) neural networks with different featurizers

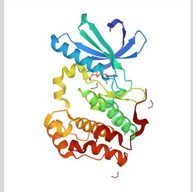
GraphConv/ECFP



Mordred/RDKit

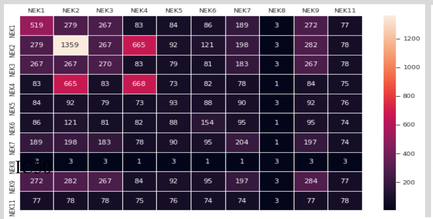


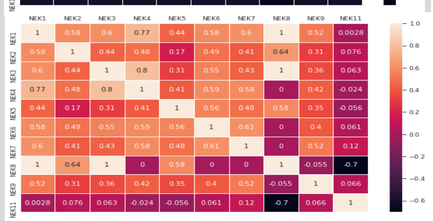
NEK kinase



RESULTS AND DISCUSSION

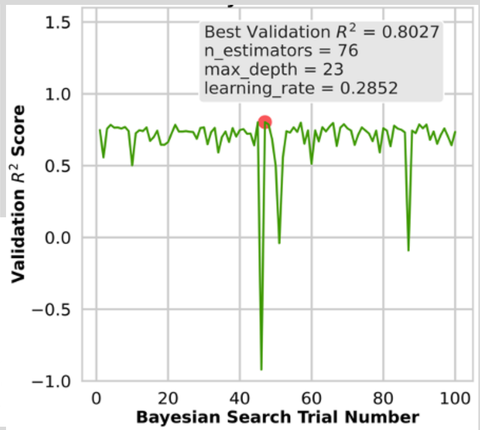
Number of overlapping structures available for each target





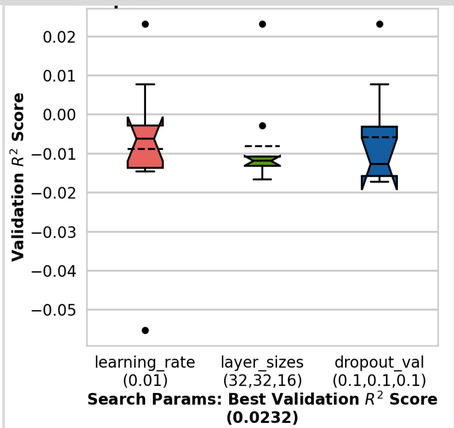
Best single-task model:

- XGBoost with the Mordred featurizer on NEK6 targets and IC50 endpoints
- Bayesian search → validation R^2 of 0.8027



Best multi-task model:

- GraphConv model with ECFP4 featurizer on all targets and endpoints
- Grid search → validation R^2 = 0.0232



CONCLUSIONS

- Mostly low R^2 values
  - XGBoost with the Mordred featurizer was exception
- Single-task models performed better
  - Limited data
- GPU essential for neural network training
  - 3x speedup with GPU

REFERENCES

1. Moriwaki, H., Tian, YS., Kawashita, N. *et al.* Mordred: a molecular descriptor calculator. *J Cheminform* **10**, 4 (2018). https://doi.org/10.1186/s13321-018-0258-y
2. Paul, S., Mytelka, D., Dunwiddie, C. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9**, 203–214 (2010). https://doi.org/10.1038/nrd3078
3. https://github.com/ATOMScience-org/AMPL/tree/master/atomsci

WHAT WE LEARNED

- Manipulating and curating data
- Splitting of SMILES (scaffold) to create training data
- Reading literature and using cheminformatics libraries such as DeepChem and RDKit
- Use cases of different models
- Optimizing hyperparameters
- GPU usage and creating environments

FUTURE DIRECTIONS

- Improve model training performance
- Predict molecules with best binding affinity to NEK Kinases using virtual screening

ACKNOWLEDGEMENTS

We would like to thank Dr. Jonathan Allen (LLNL), Dr. Stewart He (LLNL), Rose Wilfong (LLNL) the ATOM team, David Glass (Data Mine), and The Data Mine Staff