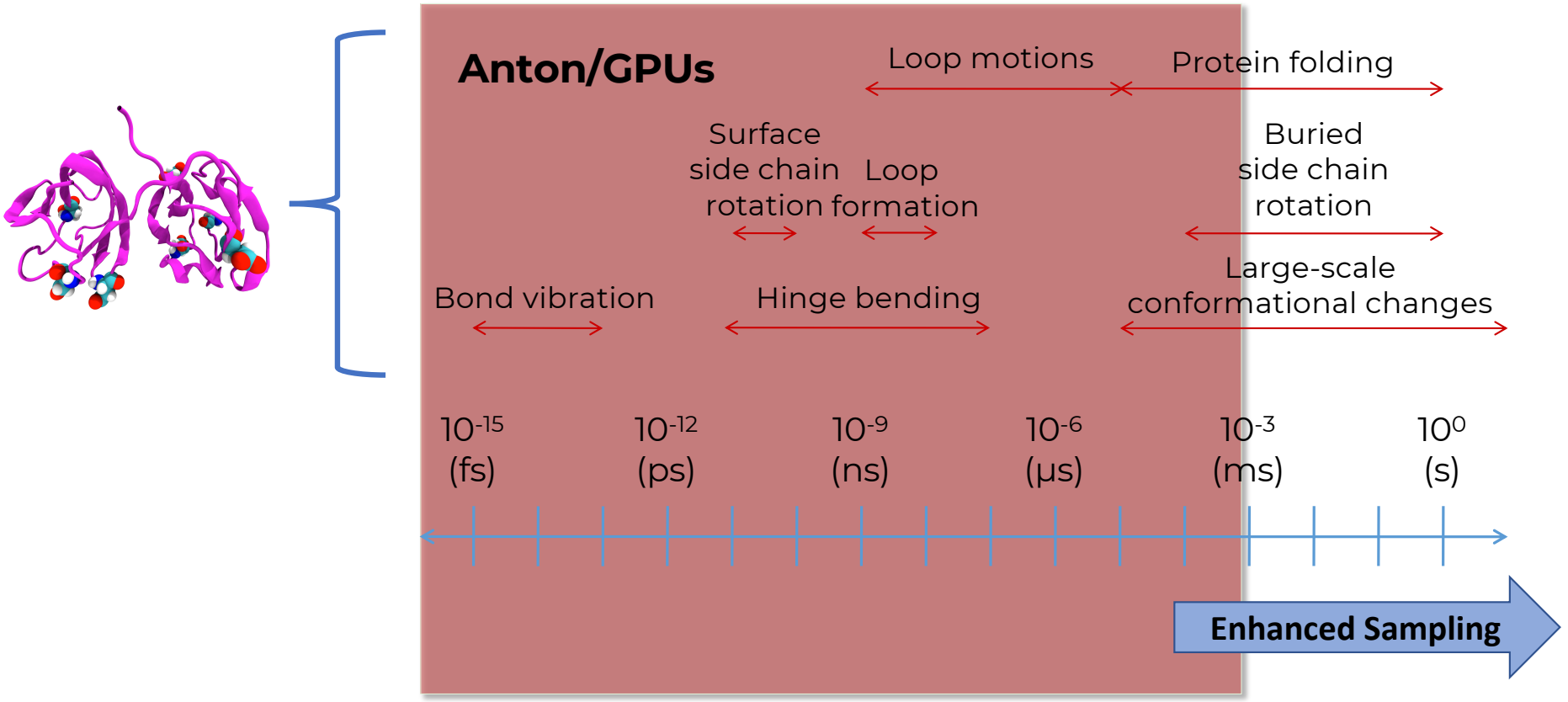
A ML-based Progress Coordinate for Enhanced Molecular Simulations

09-615 Final Project | 08Dec2022

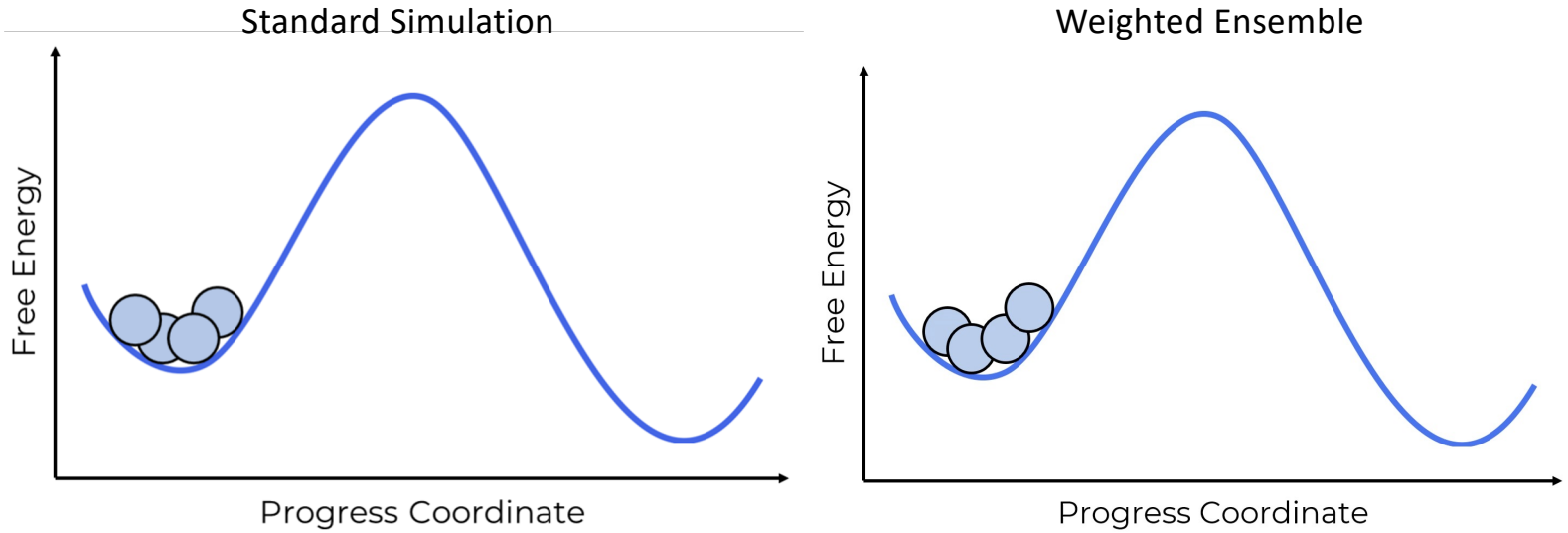
Darian Yang

Simulation Timescales are Limited



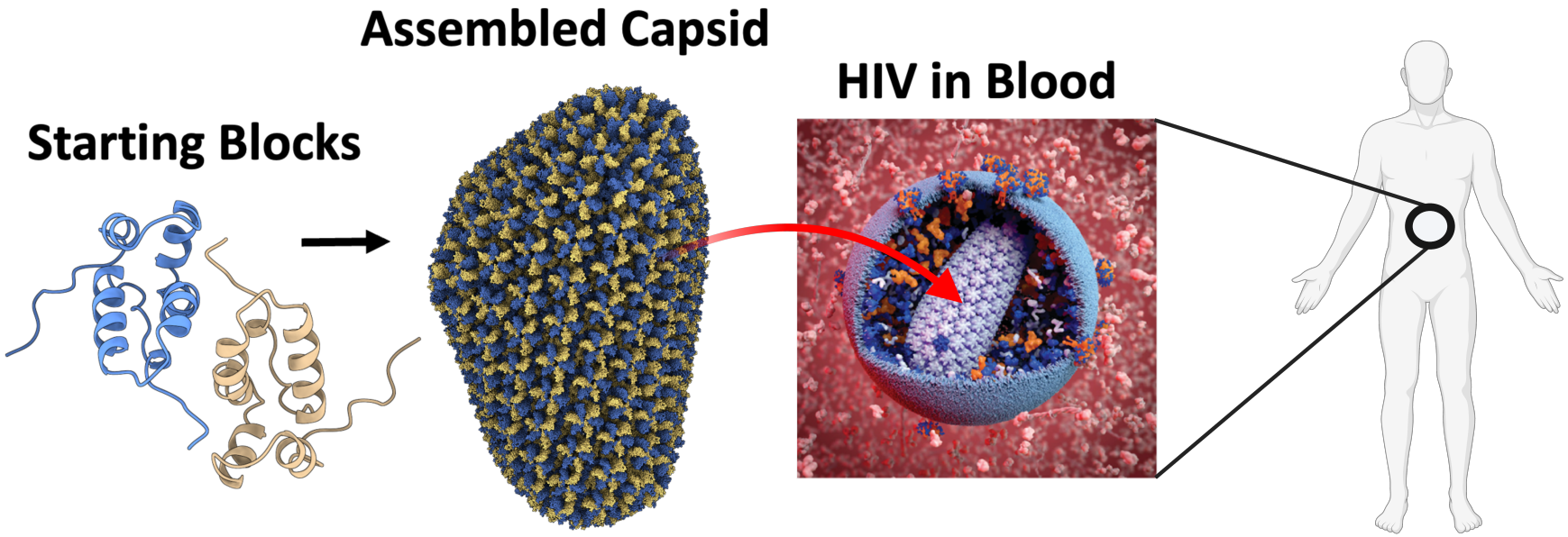
Zwier & Chong, *Curr. Opin. Pharmacol., 10: 745-752* (2010).

A Splitting Approach to Enhance Sampling



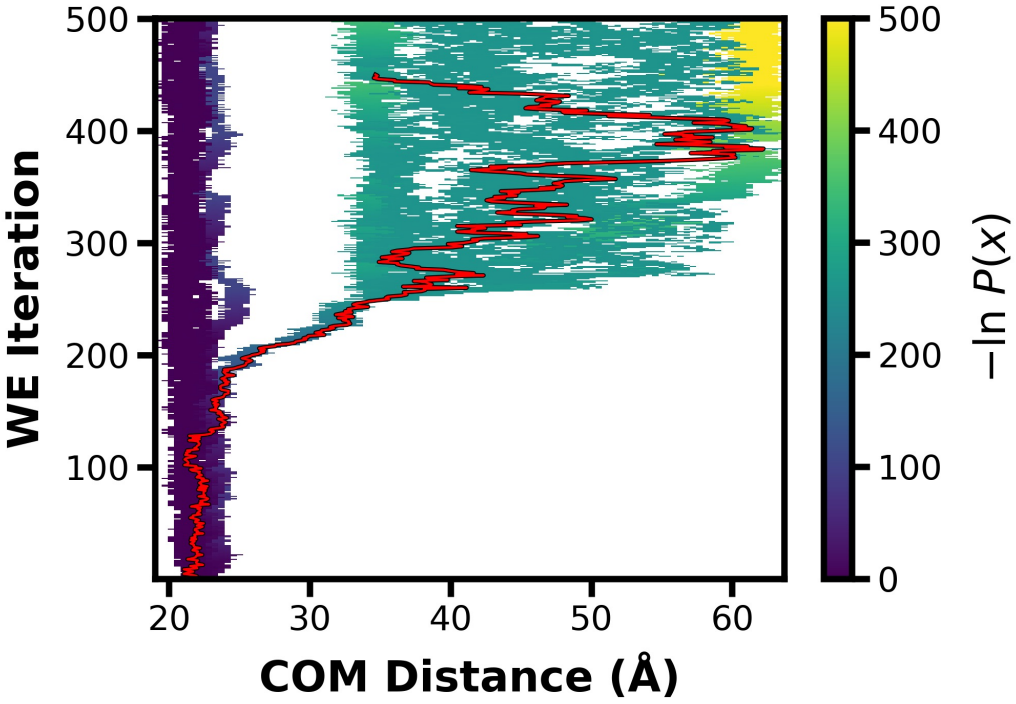
Huber and Kim, Biophys. Journal (1996)
Zuckerman and Chong, Ann. Rev. Biophys. (2017)

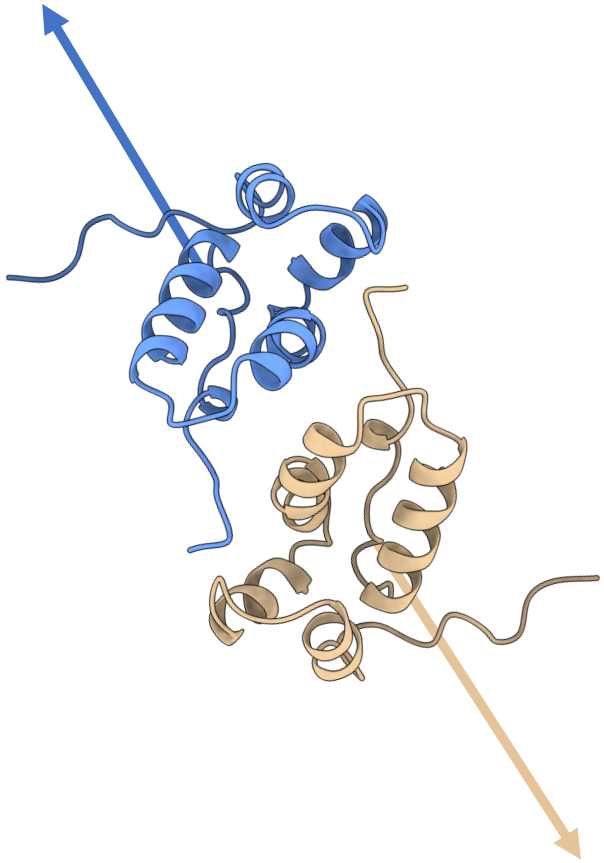
Test System:
Human Immunodeficiency Virus 1 (HIV-1) Capsid Protein



Johnson *et al.* "3D molecular models of whole HIV-1 virions generated with cellPACK", ***Faraday Discuss.***, 2014, **169**, 23-44.

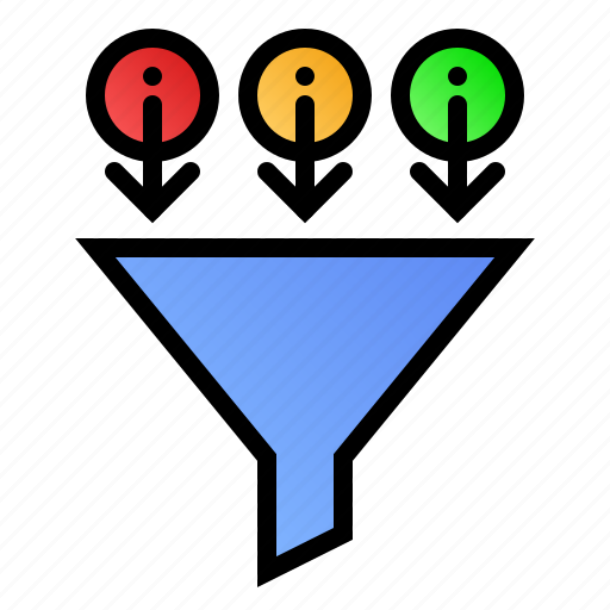
Dataset Curation





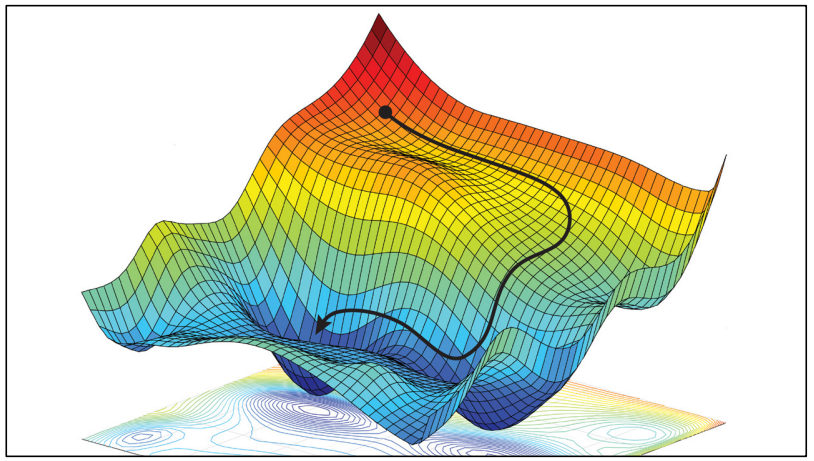
Exploratory Data Analysis and Processing

- Original Dataset:
  - 17912 simulation trajectories with 100 frames each
- Processed Dataset:
  - 50 Columns = Calculated 50 features
    - Various distances, angles, etc.
    - |last – first frame| at every iteration (then scaled)
  - 17912 Rows = each short trajectory (10 true)



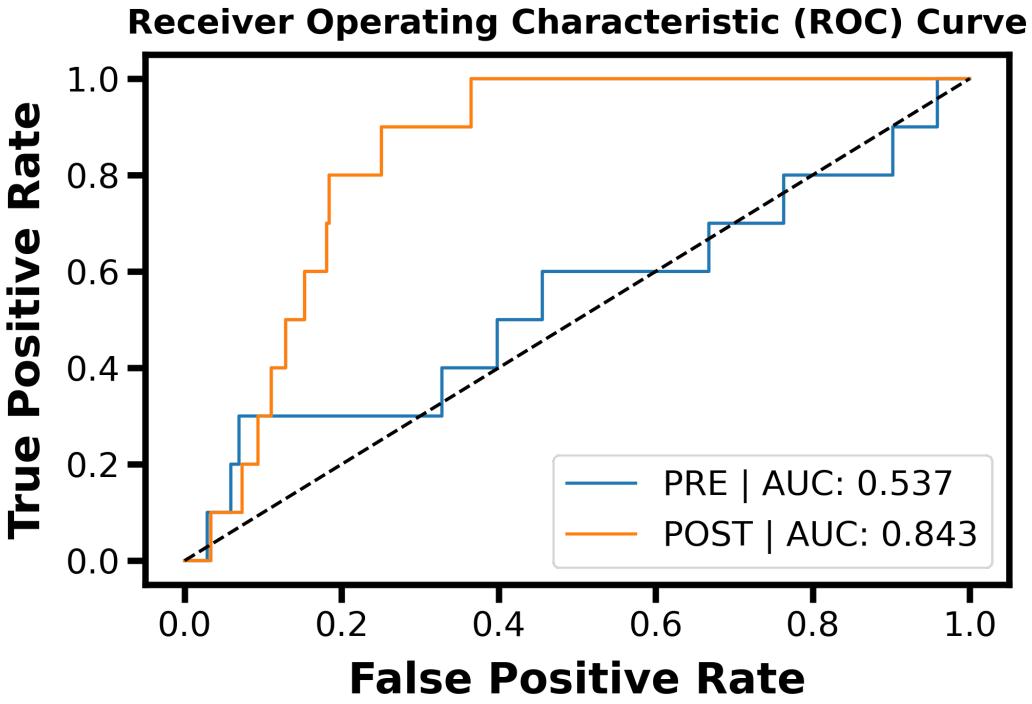
A ML-based Progress Coordinate

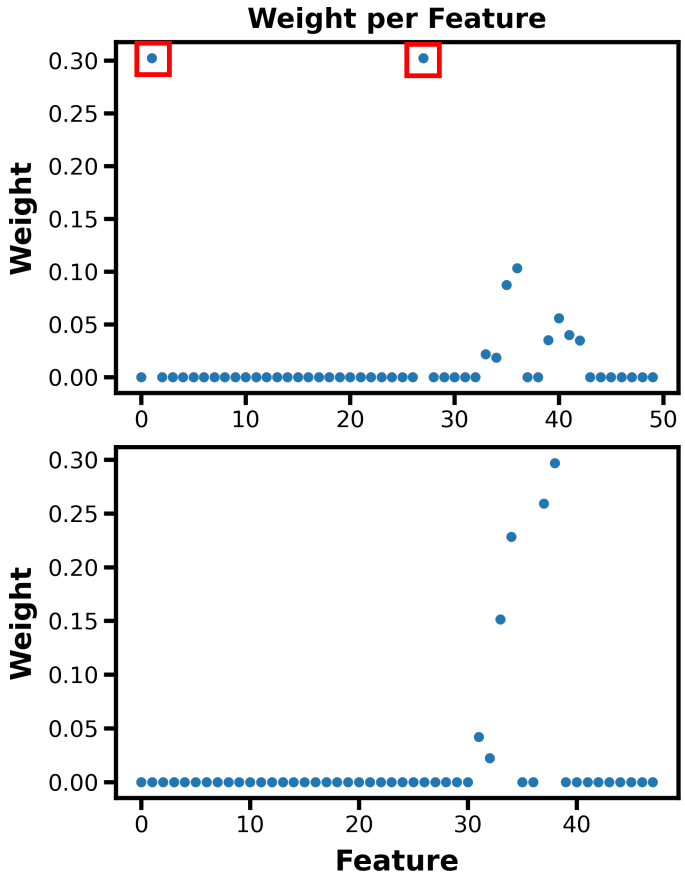
- Optimization of feature weights
  - Loss function : ROCAUC score
- Output is weights for each feature
  - New progress coordinate:
    - Linear combination of weighted features



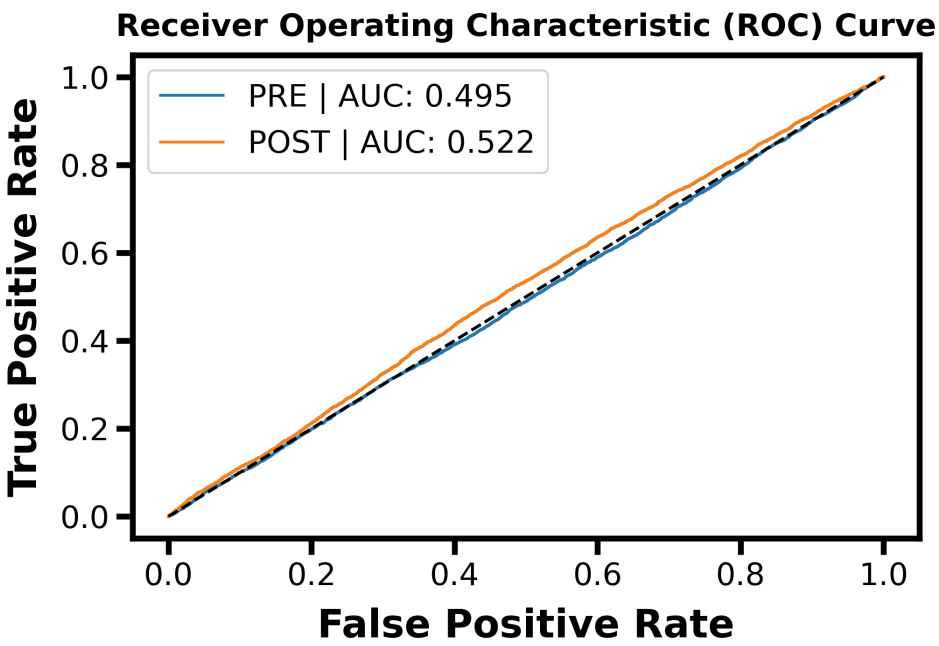
https://www.science.org/content/article/ai-researchers-allege-machine-learning-alchemy

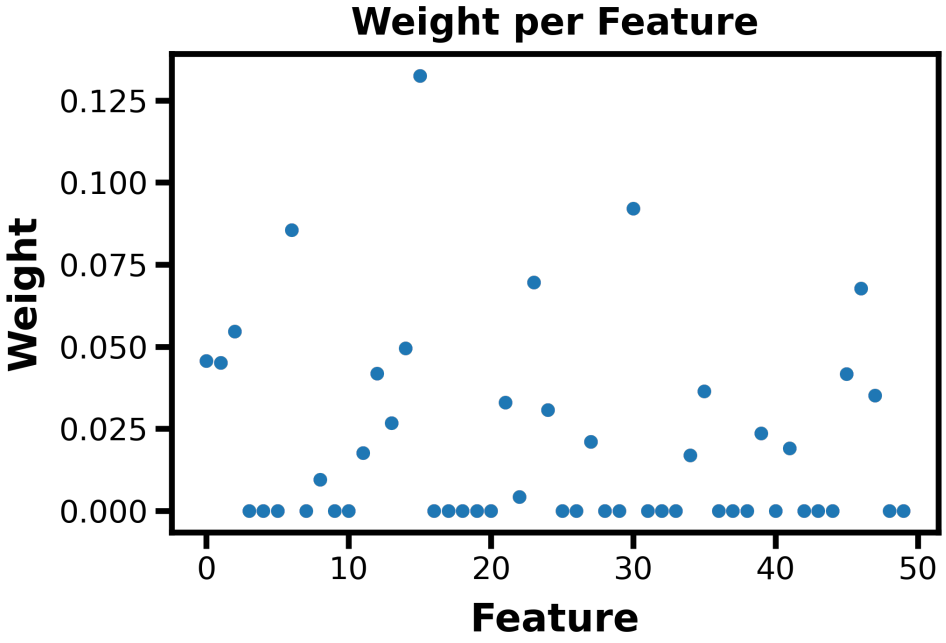
Optimization results





Optimization results (random)





Next steps:

- Improve/change loss function or scoring metric
- Larger and more diverse input dataset
  - Or change how ML input array is created
  - Adding history of trajectory for success labeling
- Testing the optimized progress coordinate in a follow-up simulation