Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

ATOM Modeling Pipeline (AMPL) for Drug Discovery

S. Ravichandran,
Data Scientist, BIDS, FNLCR

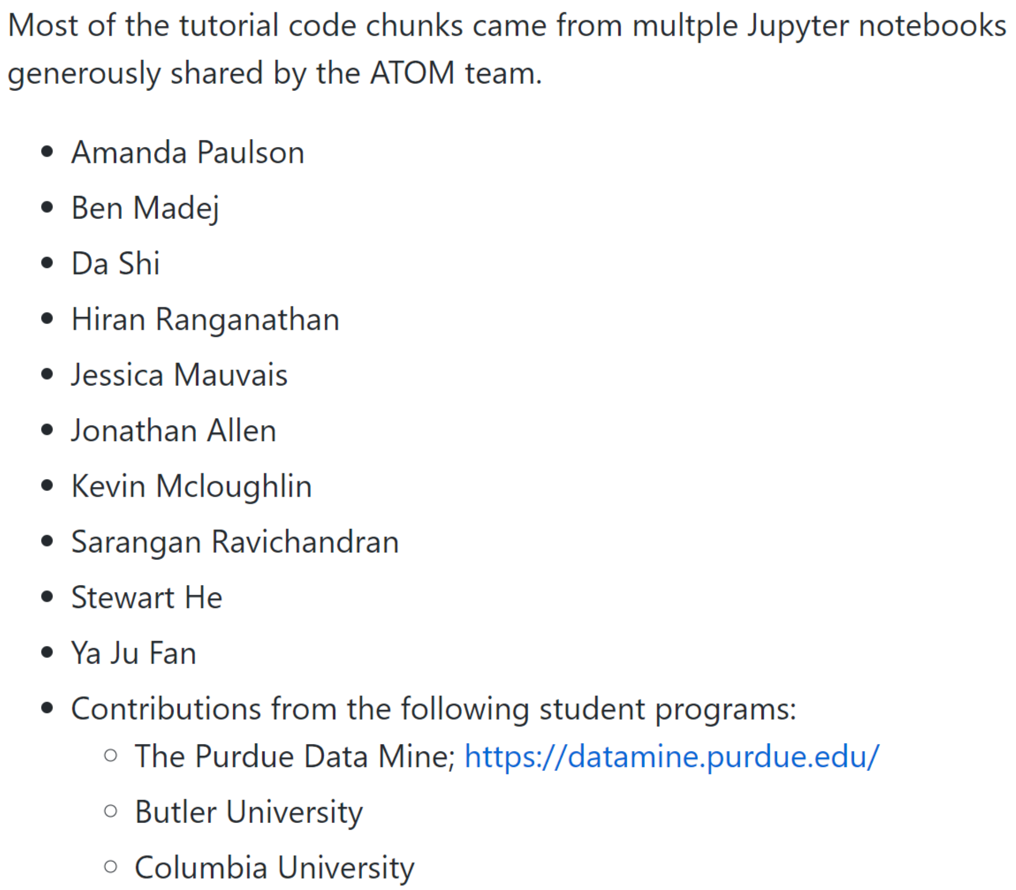
June 8, 2021

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Acknowledgements

- ATOM Team

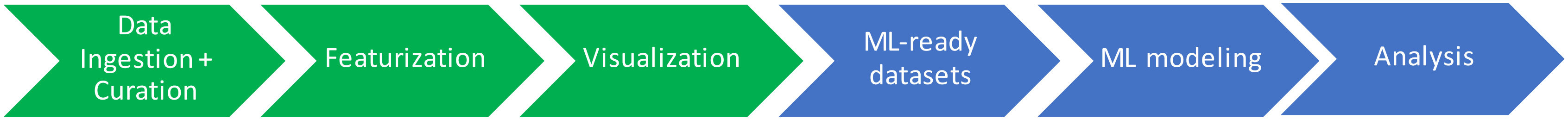


Most of the tutorial code chunks came from multple Jupyter notebooks generously shared by the ATOM team.

- Amanda Paulson
- Ben Madej
- Da Shi
- Hiran Ranganathan
- Jessica Mauvais
- Jonathan Allen
- Kevin Mcloughlin
- Sarangan Ravichandran
- Stewart He
- Ya Ju Fan
- Contributions from the following student programs:
  - The Purdue Data Mine; https://datamine.purdue.edu/
  - Butler University
  - Columbia University

Agenda

- Introduction to AMPL (ATOM Modeling Pipeline)
- Why AMPL?
- Goal for today



Data Ingestion + Curation

Featurization

Visualization

ML-ready datasets

ML modeling

Analysis

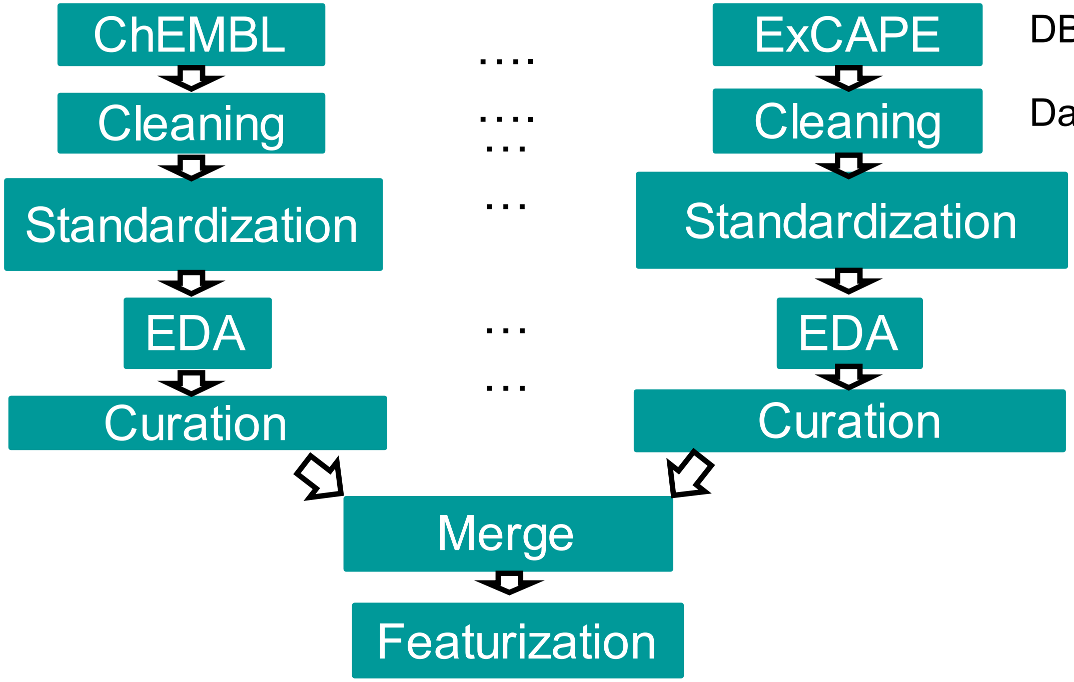
Data Sources

- ChEMBL: Manually curated repository of small molecules (EMBL/EBI)
  - ~1.9 M compounds; ~11K targets
  - https://www.ebi.ac.uk/chembl/
- ExCAPE-DB (EU program)
  - ~ 1M compounds/1.7K targets
  - https://solr.ideaconsult.net/search/excape/#
- Drug Target commons (Univ of Helsinki)
  - ~1.7M cpds; 13K targets
  - http://drugtargetcommons.fimm.fi/

Why combine data?

- ATOM team’s experience shows that the combined dataset (Union) models show robustness and performance than individual dataset

Overview



DB sources

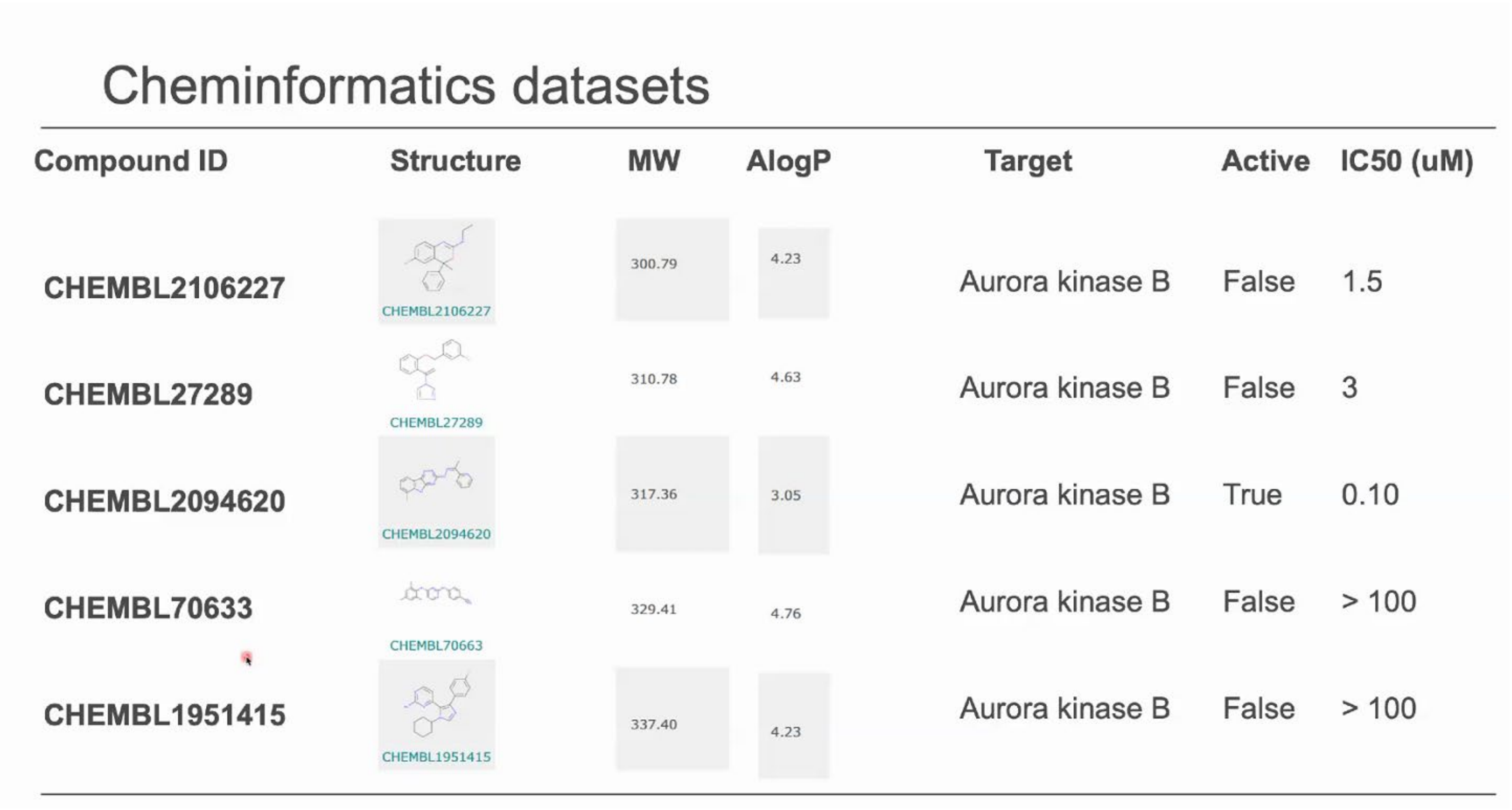
Data Tidying

base_smiles_from_smiles

diversity_plots*; heat map, tanimoto similarity*

Curate_data.aggregate_assay_data

A sample dataset



Cheminformatics datasets

Compound ID	Structure	MW	AlogP	Target	Active	IC50 (uM)
CHEMBL2106227	CHEMBL2106227	300.79	4.23	Aurora kinase B	False	1.5
CHEMBL27289	CHEMBL27289	310.78	4.63	Aurora kinase B	False	3
CHEMBL2094620	CHEMBL2094620	317.36	3.05	Aurora kinase B	True	0.10
CHEMBL70633	CHEMBL70663	329.41	4.76	Aurora kinase B	False	> 100
CHEMBL1951415	CHEMBL1951415	337.40	4.23	Aurora kinase B	False	> 100

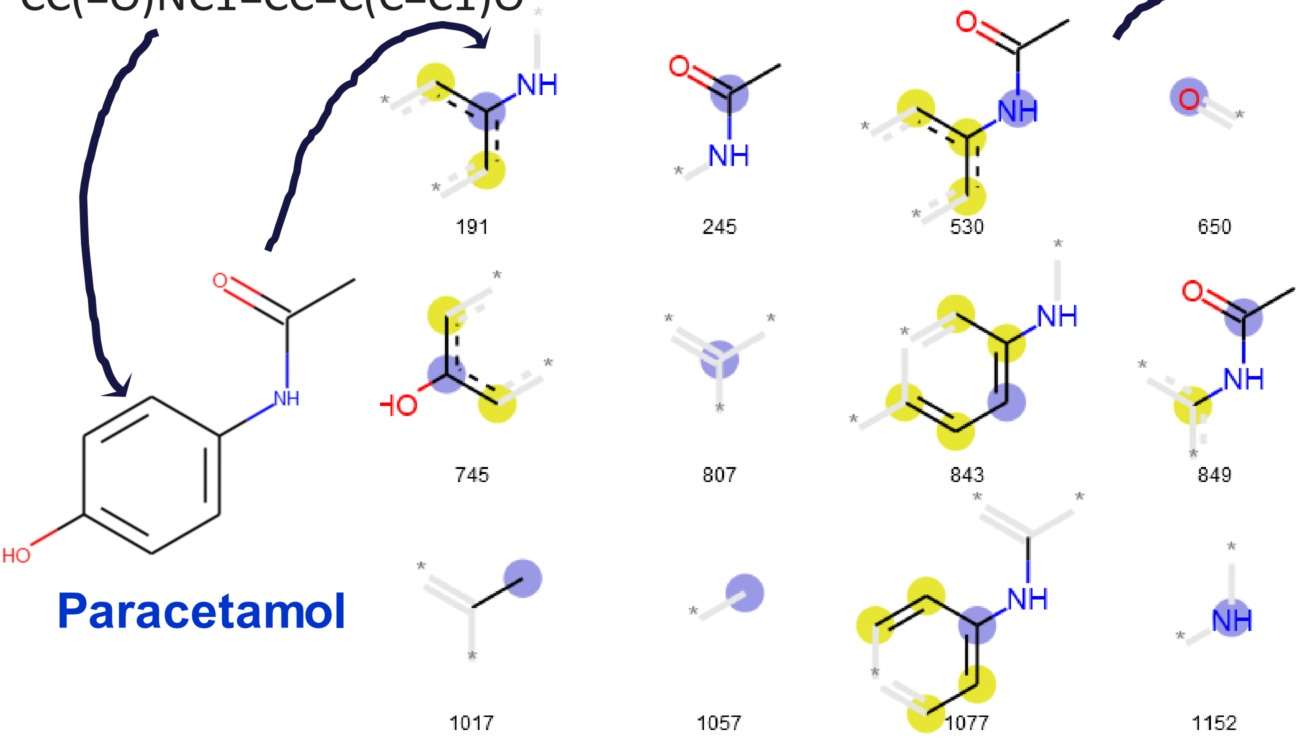
Featurizing a molecule: Fingerprints

- Fingerprints
  - Molecules → fixed-length binary vectors (0s and 1s). indicating presence/absence of certain molecular features
  - One can compare fingerprints of two molecules and identify similarity

		Properties or Fingerprint						Outcome
ID	**SMILES**	**Bit0**	**Bit1**	**Bit2**	**Bit3**	**Bit4**	**Bit5**	**Class**
1	SMILES1	1	1	0	1	0	1	cns
2	SMILES2	0	0	0	1	1	0	cns
3	SMILES3	1	0	0	1	0	0	Cardiovascular
3	SMILES4	1	0	0	1	1	0	Antineoplastic
4	SMILES5	1	1	0	1	1	1	Dermatologic
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…

SMILES → Fingerprint

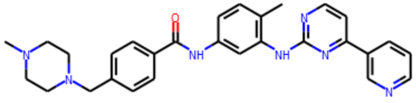
CC(=O)NC1=CC=C(C=C1)O



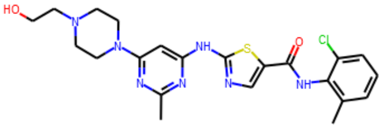
```
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 10 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 01 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 01 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 01 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 10 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 01 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 01 00 00 01 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 10 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 10 00 00 00 00 00 0 00 00 00 01 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 01 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 01 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 01
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 10 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 01 00 00 00 0 00 00 00 00 00 00 00 0 00 00 01 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00 00 00 00 00 1 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 10 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 01 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00 00
0 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 0 00 00 00 00 00 00 00 00 0 00 00 00 00 00
0 0 00 00 00
```

Featurizing a molecule: Molecular descriptors

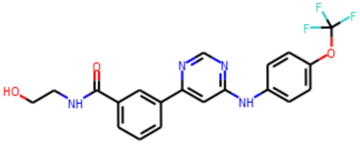
- Physicochemical properties
  - Molecular weight, # of Hydrogen bond donors, log partition coefficient etc.
- Mordred
  - ~1800 descriptors
  - Open source software
  - Implemented in AMPL



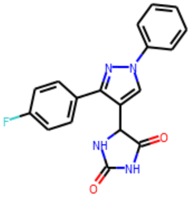
Imatinib



Dasatinib



GNF



DPH

	ABC	ABCGG	nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A	LogEE_A	...	SRW10
0	29.198227	19.516970	0	2	49.161634	2.372244	4.744487	49.161634	1.328693	4.541483	...	10.415502
1	25.731643	19.151718	0	1	42.312870	2.394767	4.762938	42.312870	1.282208	4.422390	...	10.323283
2	23.132682	16.941805	0	0	38.063201	2.370962	4.741923	38.063201	1.268773	4.312334	...	10.143881
3	19.924959	16.140292	0	0	32.867760	2.498596	4.828813	32.867760	1.314710	4.170130	...	10.150621

https://mordred-descriptor.github.io/documentation/master/descriptors.html

Computing Environment

BIOWULF

HIGH PERFORMANCE COMPUTING AT THE NIH

COLAB	NIH HPC Biowulf
Serverless	Can ask for resources
Resources are !unlimited and !guaranteed	Resources are guaranteed
Browser-based	Mostly command-line
Good for short jobs; explaining AMPL capabilities	Long jobs (HPO)
Audience: Interns, Workshop attendee (Educational)	Research

Curation

- Data Curation
  - Organization and integration of data from multiple sources
- Potent Targets
  - Dose-response measurements (Kd, Ki, IC50 and activity) in biochemical assays
    - <= 100 nM
  - Dose response measurements (activity %, % inhibition etc)
    - Different cutoffs for biochemical and cell-based assays
  - Multiple assays (different studies or data resources)
    - Median bioactivity
- Mutation data

Modeling Steps

- Data cleaning, tidying
- Data curation
- Feature engineering
  - Outcome variable IC50 → pIC50
  - Numerical → categorical
- EDA
  - FP → Tonimoto → tSNE
- Aggregate assay
- Diversity plots
- Featurization

Bio-assay with a specific target protein

- Identifying drugs or compounds primary targets and off-targets is a critical task in drug discovery
  - Kinases
    - Target promiscuity → polypharmacological effects
- Understanding this concept can help us with the drug repurposing efforts
- Many groups collect and curate Target-drug data
  - Diversity of experiments
    - Different bioassay, bioactivity endpoints etc. makes the problem challenging

Useful links

- https://github.com/ATOMconsortium/AMPL
- https://github.com/ravichas/AMPL-workshop-1
- https://github.com/ATOMconsortium/AMPL/tree/Tutorials/atomsci/ddm/examples/tutorials
- https://hpc.nih.gov/apps/ampl.html

- Workshop materials
- https://github.com/ravichas/AMPL-workshop-1