INTEGRATING HIGH-PERFORMANCE SIMULATIONS AND LEARNING TOWARD IMPROVED CANCER THERAPY

AUSTIN CLYDE

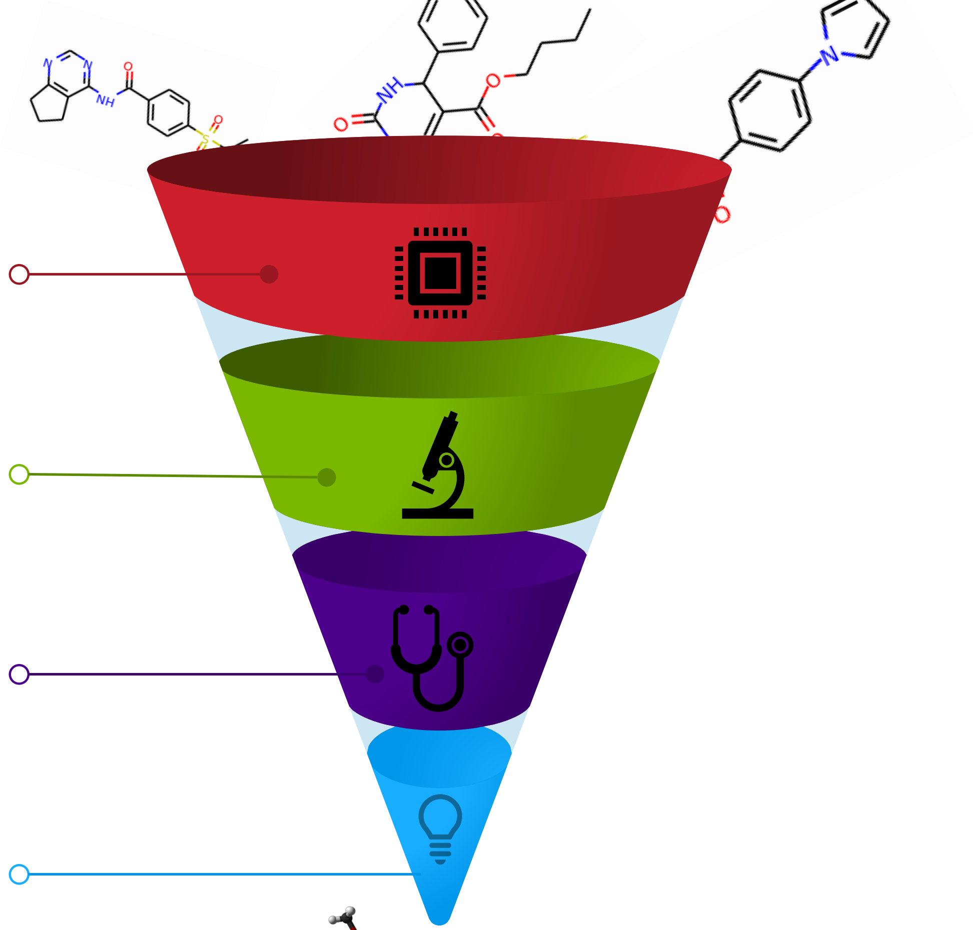
Ph.D. Student, University of Chicago
Computational science, Argonne National Laboratory

Austin Clyde[1*], Dave Wright[2*], Katya Ahmed[2], John D. Chodera[3], John Karanicolas[4],
Palani Kirubakaran[4], Hyungro Lee[5], Matteo Turilli[5], Shunzhou Wan[2],
Peter Coveney[2**], Shantenu Jha[5,6**], Rick Stevens[1**]

[1] *University of Chicago and Argonne National Laboratory*
[2] *University College London*
[3] *Memorial Sloan Kettering Cancer Center*
[4] *Molecular Therapeutics group, Fox Chase Cancer Center*
[5] *RADICAL, ECE, Rutgers University, Piscataway,NJ 08854, USA*
[6] *Brookhaven National Laboratory, Upton, NY, USA* **Joint First Authors* ***Senior Authors*

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Fifth Computational Approaches for Cancer Workshop (CAFCW19)
November 17th, 2019



DRUG DISCOVERY

~10^8 products

PRE CLINICAL

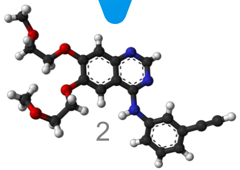
11,000 products

CLINICAL TRIALS

6,300 products

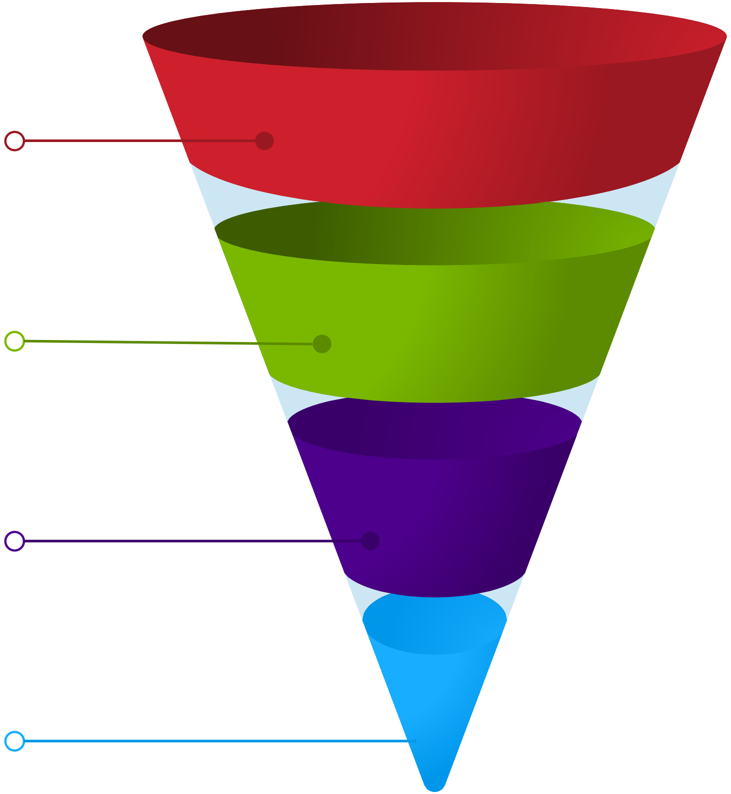
FDA APPROVAL

111 products



Target based compound screening

10^{60} estimated drug-like compounds



COMPOUND DISCOVERY

Mining massive building block or de-novo generated libraries

INTERESTING?

Does this compound inhibit or interact with the target?

TOXICOLOGY

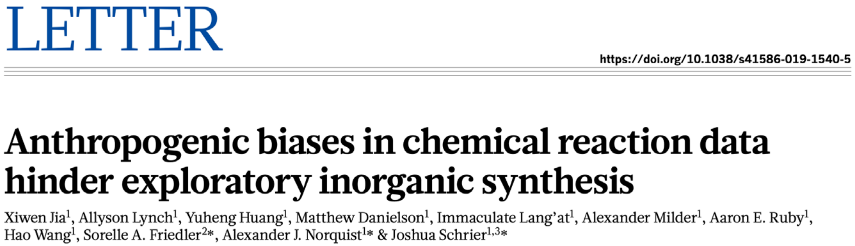
Is this compound reasonably safe?

SYNTHESIS

Can we buy it, is it from available building blocks, or do we need to hire a medicinal chemist?

GOAL:

Design an intelligent system to screen a space of drugs efficiently and intelligently.

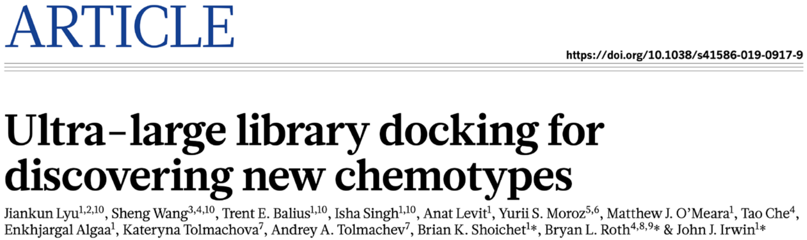


LETTER

https://doi.org/10.1038/s41586-019-1540-5

Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis

Xiwen Jia[1], Allyson Lynch[1], Yuheng Huang[1], Matthew Danielson[1], Immaculate Lang'at[1], Alexander Milder[1], Aaron E. Ruby[1], Hao Wang[1], Sorelle A. Friedler[2]*, Alexander J. Norquist[1]* & Joshua Schrier[1,3]*



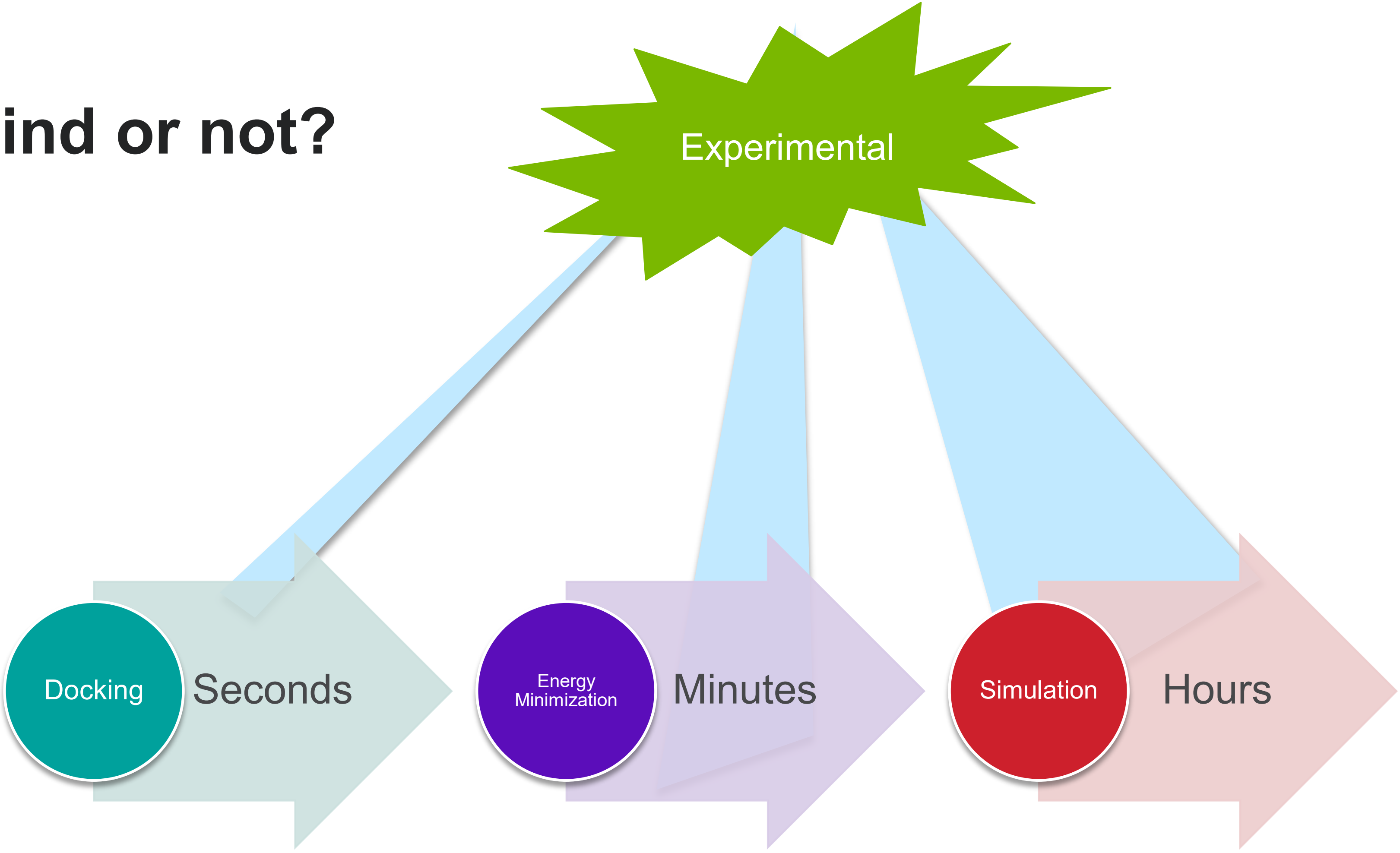
ARTICLE

https://doi.org/10.1038/s41586-019-0917-9

Ultra-large library docking for discovering new chemotypes

Jiankun Lyu[1,2,10], Sheng Wang[3,4,10], Trent E. Balius[1,10], Isha Singh[1,10], Anat Levit[1], Yurii S. Moroz[5,6], Matthew J. O'Meara[1], Tao Che[4], Enkhjargal Algaa[1], Kateryna Tolmachova[7], Andrey A. Tolmachev[7], Brian K. Shoichet[1]*, Bryan L. Roth[4,8,9]* & John J. Irwin[1]*

Bind or not?

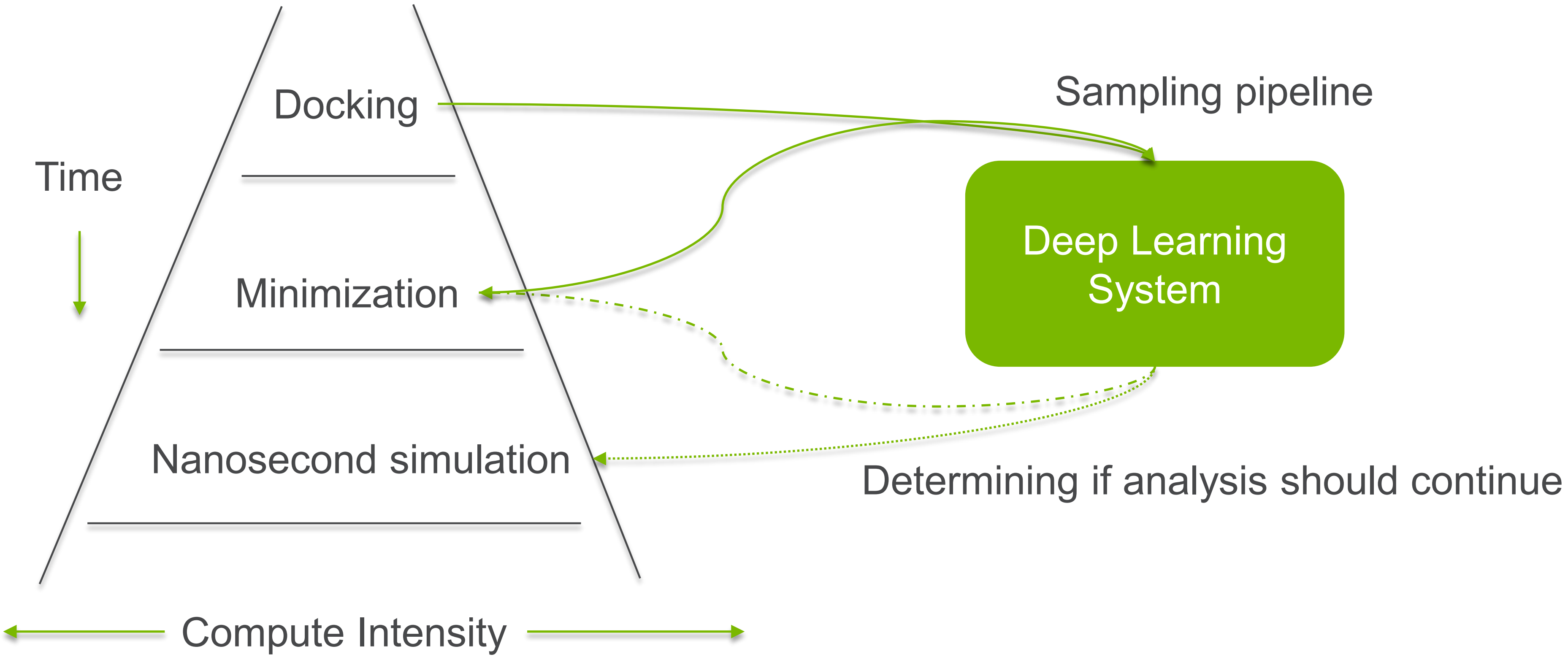


CPU

CUDA

CUDA

A pipeline unit



Docking

Minimization

Nanosecond simulation

Time

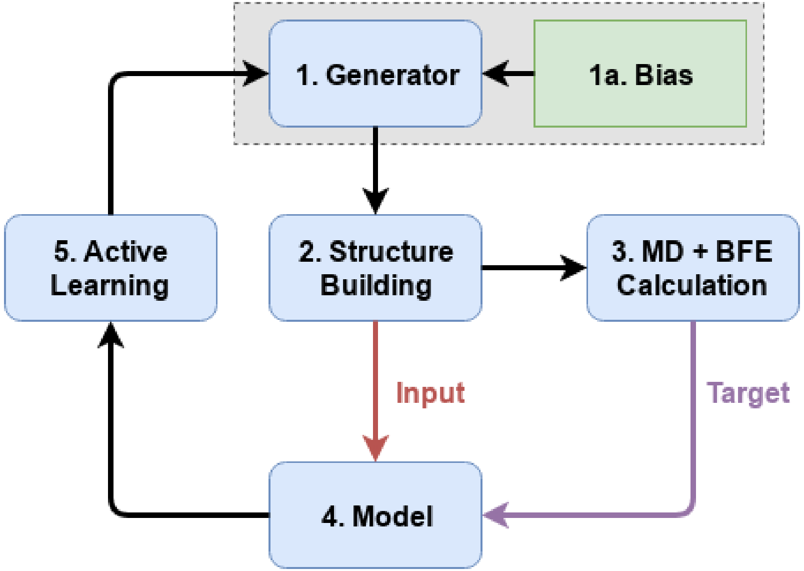
Compute Intensity

Sampling pipeline

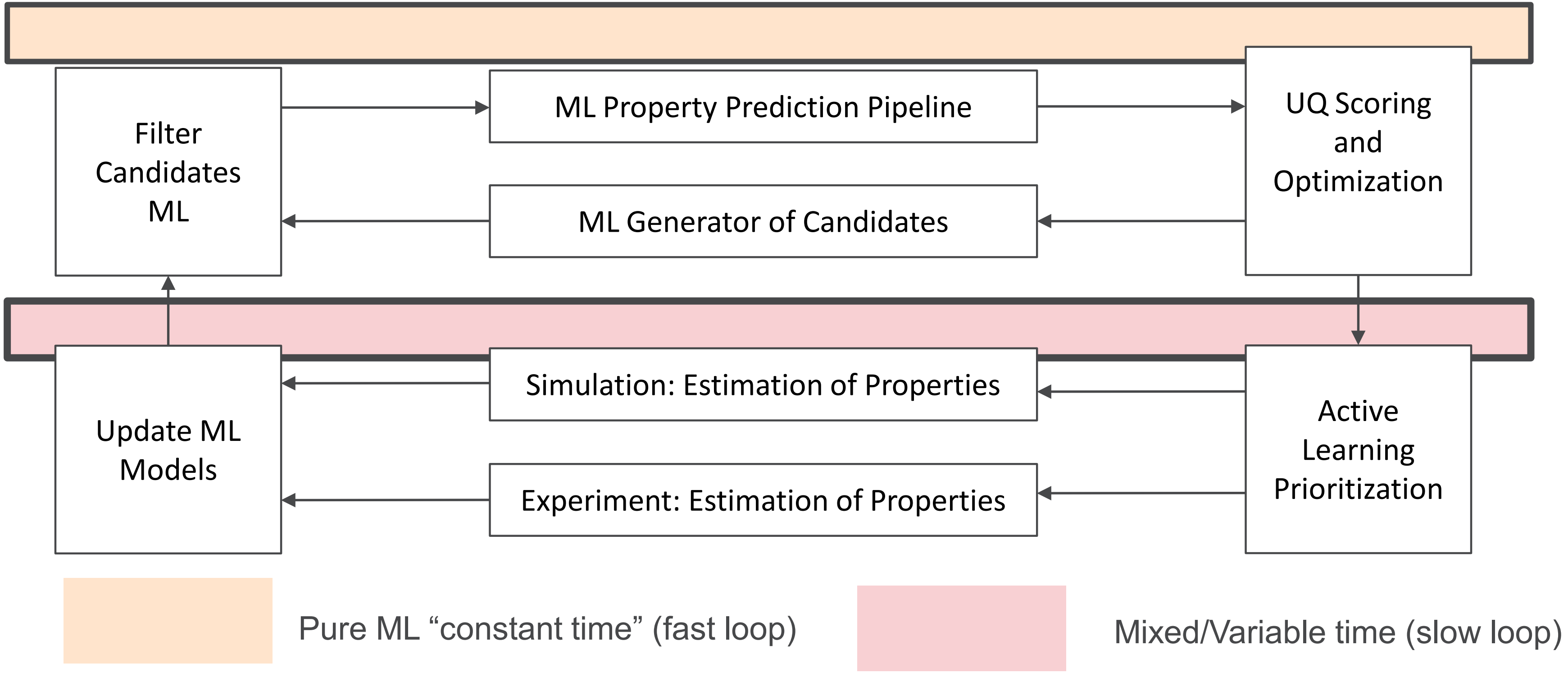
Deep Learning System

Determining if analysis should continue

Pipelining discovery and screening



LAYERED WORKFLOW



Filter Candidates ML

ML Property Prediction Pipeline

UQ Scoring and Optimization

ML Generator of Candidates

Update ML Models

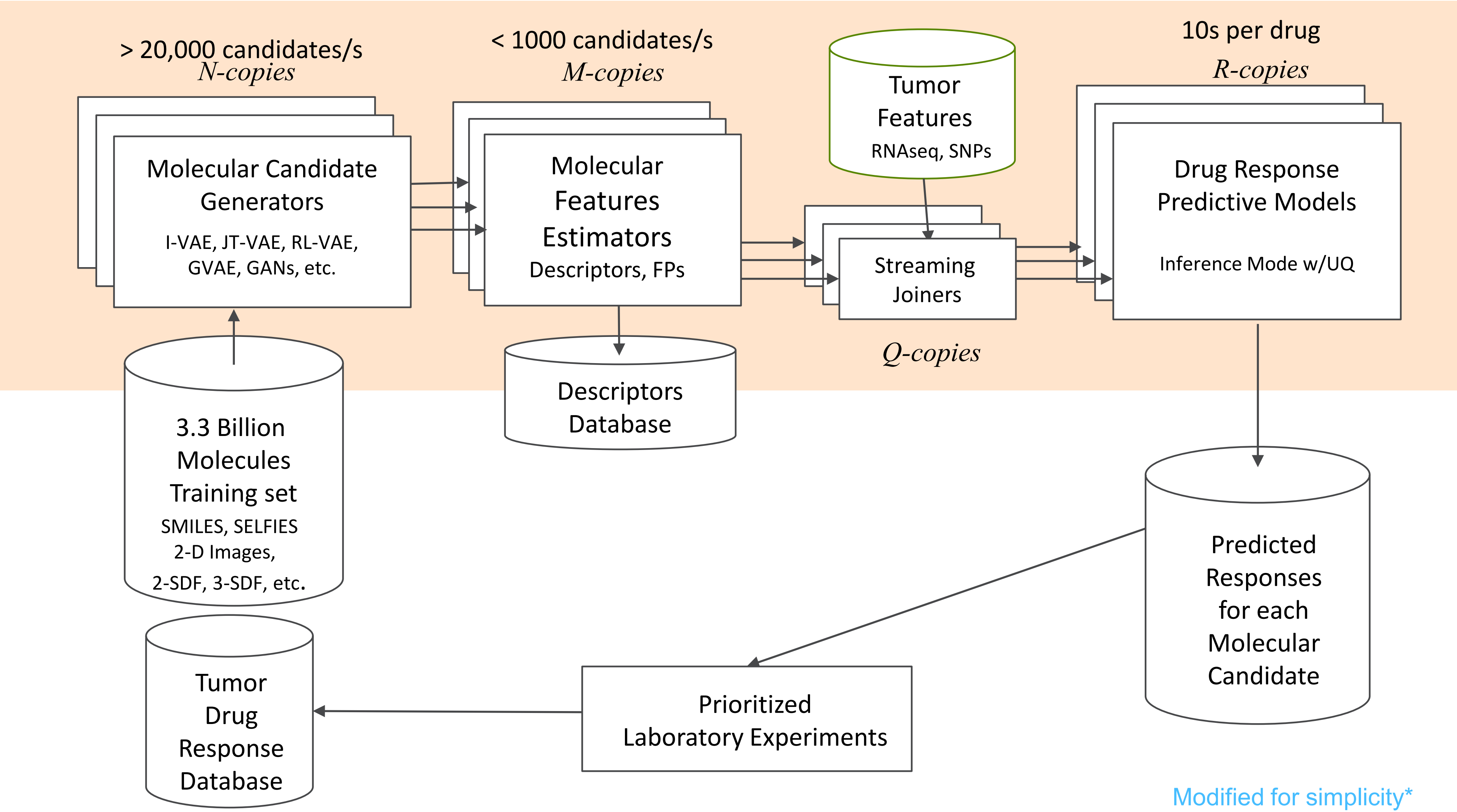
Simulation: Estimation of Properties

Experiment: Estimation of Properties

Active Learning Prioritization

Pure ML “constant time” (fast loop)

Mixed/Variable time (slow loop)



> 20,000 candidates/s
N-copies

Molecular Candidate Generators

I-VAE, JT-VAE, RL-VAE, GVAE, GANs, etc.

< 1000 candidates/s
M-copies

Molecular Features Estimators

Descriptors, FPs

Tumor Features

RNAseq, SNPs

Streaming Joiners

Q-copies

10s per drug
R-copies

Drug Response Predictive Models

Inference Mode w/UQ

3.3 Billion Molecules Training set

SMILES, SELFIES
2-D Images,
2-SDF, 3-SDF, etc.

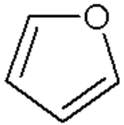
Descriptors Database

Predicted Responses for each Molecular Candidate

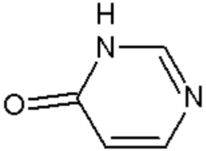
Prioritized Laboratory Experiments

Tumor Drug Response Database

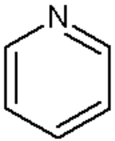
Modified for simplicity*



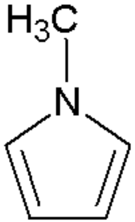
c1ccco1



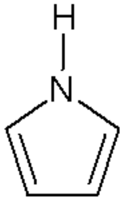
O=C1C=CN=CN1



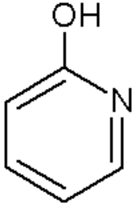
c1ccccn1



Cn1cccc1



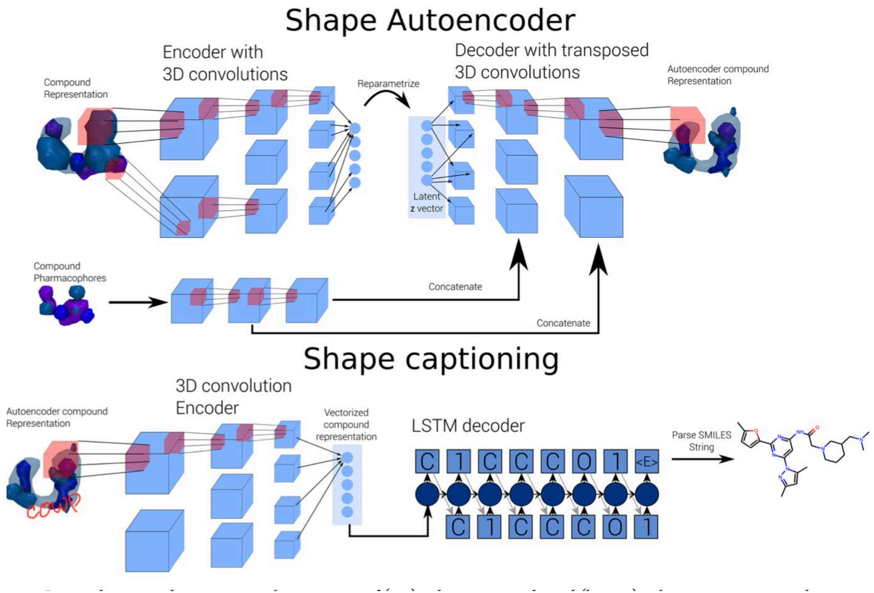
c1cccn1



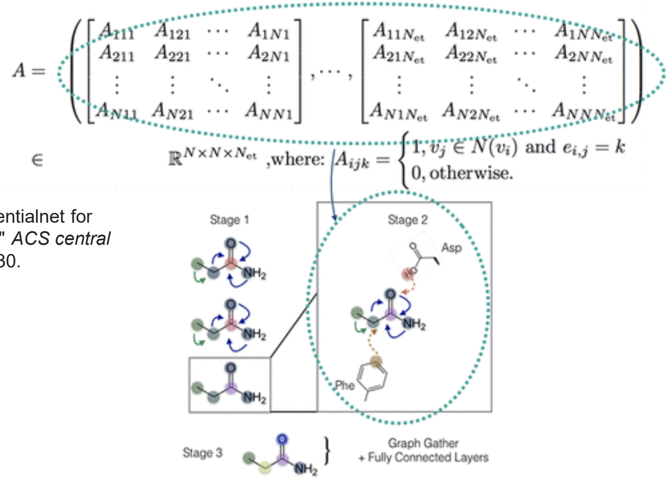
Oc1ccccn1

PROPERTY PREDICTIONS

Images, 3D surfaces



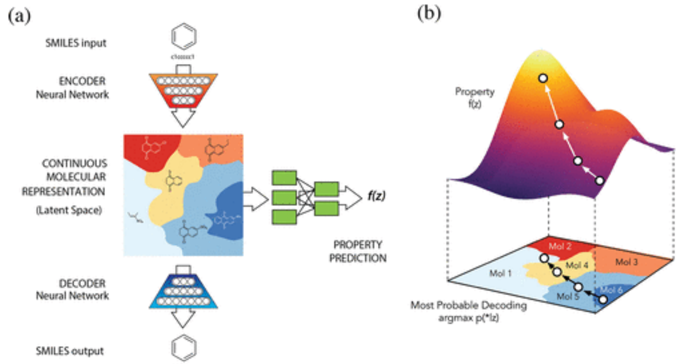
Skalic, Miha, et al. "Shape-Based Generative Modeling for de Novo Drug Design." *Journal of chemical information and modeling* 59.3 (2019): 1205-1214.



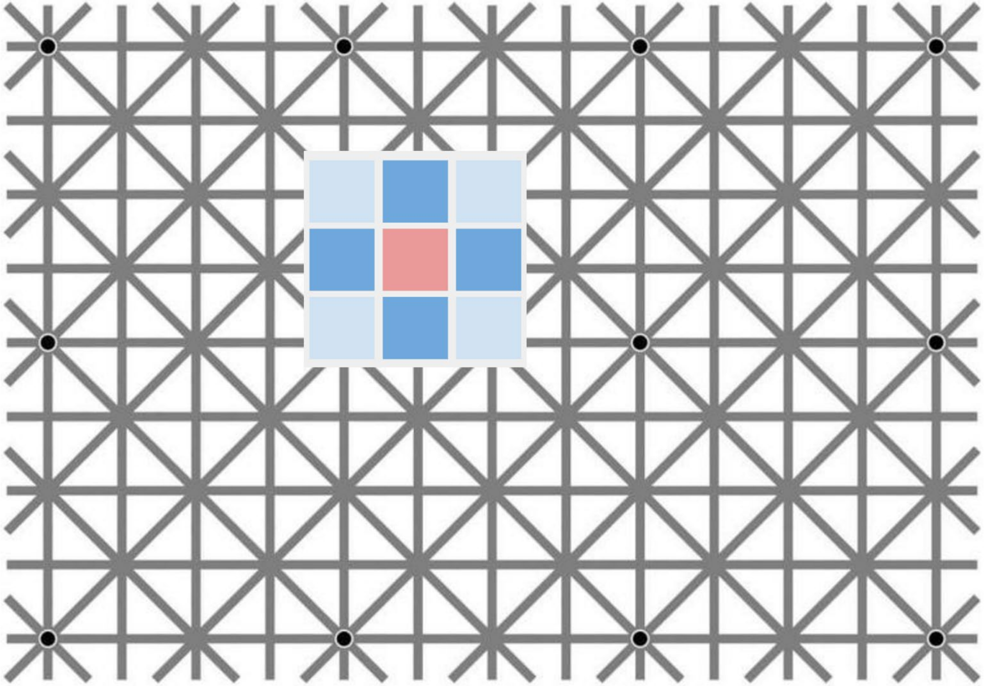
$$A = \left(\begin{bmatrix} A_{111} & A_{121} & \cdots & A_{1N1} \\ A_{211} & A_{221} & \cdots & A_{2N1} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N11} & A_{N21} & \cdots & A_{NN1} \end{bmatrix}, \cdots, \begin{bmatrix} A_{11N_{et}} & A_{12N_{et}} & \cdots & A_{1NN_{et}} \\ A_{21N_{et}} & A_{22N_{et}} & \cdots & A_{2NN_{et}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N1N_{et}} & A_{N2N_{et}} & \cdots & A_{NNN_{et}} \end{bmatrix} \right)$$

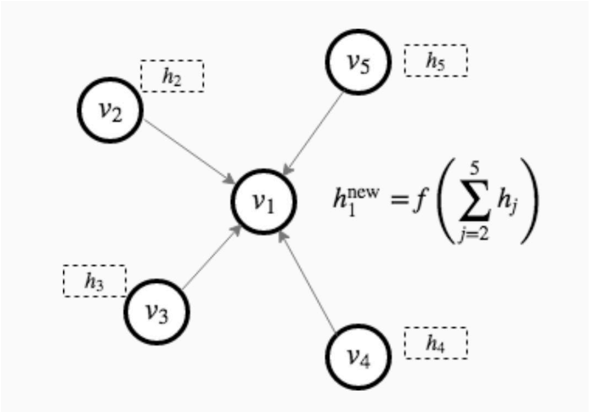
$$\in \mathbb{R}^{N \times N \times N_{et}}, \text{where: } A_{ijk} = \begin{cases} 1, v_j \in N(v_i) \text{ and } e_{i,j} = k \\ 0, \text{otherwise.} \end{cases}$$

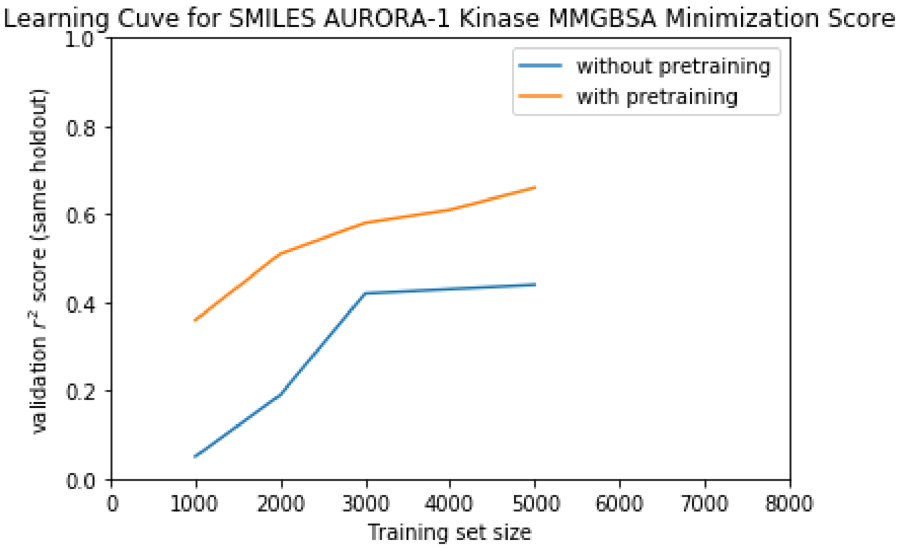
Feinberg, Evan N., et al. "Potentialnet for molecular property prediction." *ACS central science* 4.11 (2018): 1520-1530.

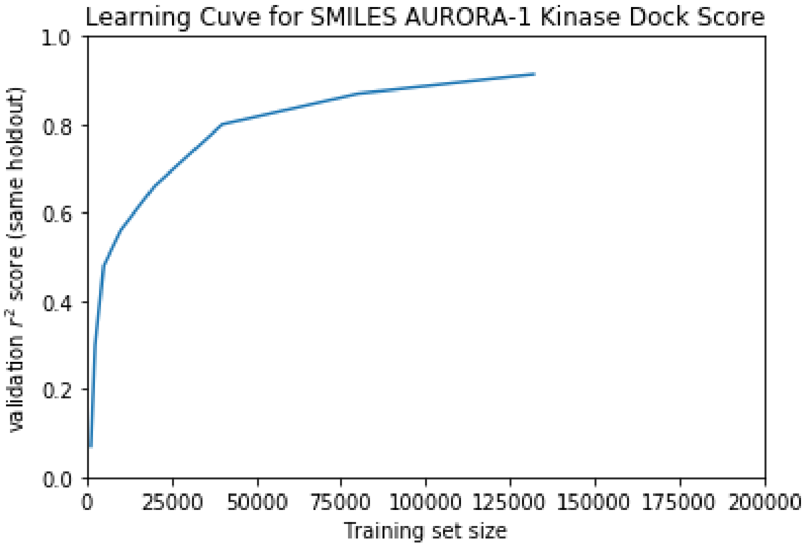


Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.




v_2
h_2
v_5
h_5
v_1
$h_1^{\text{new}} = f\left(\sum_{j=2}^{5} h_j\right)$
h_3
v_3
v_4
h_4


Learning Cuve for SMILES AURORA-1 Kinase MMGBSA Minimization Score
without pretraining
with pretraining
validation r^2 score (same holdout)
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
5000
6000
7000
8000
Training set size

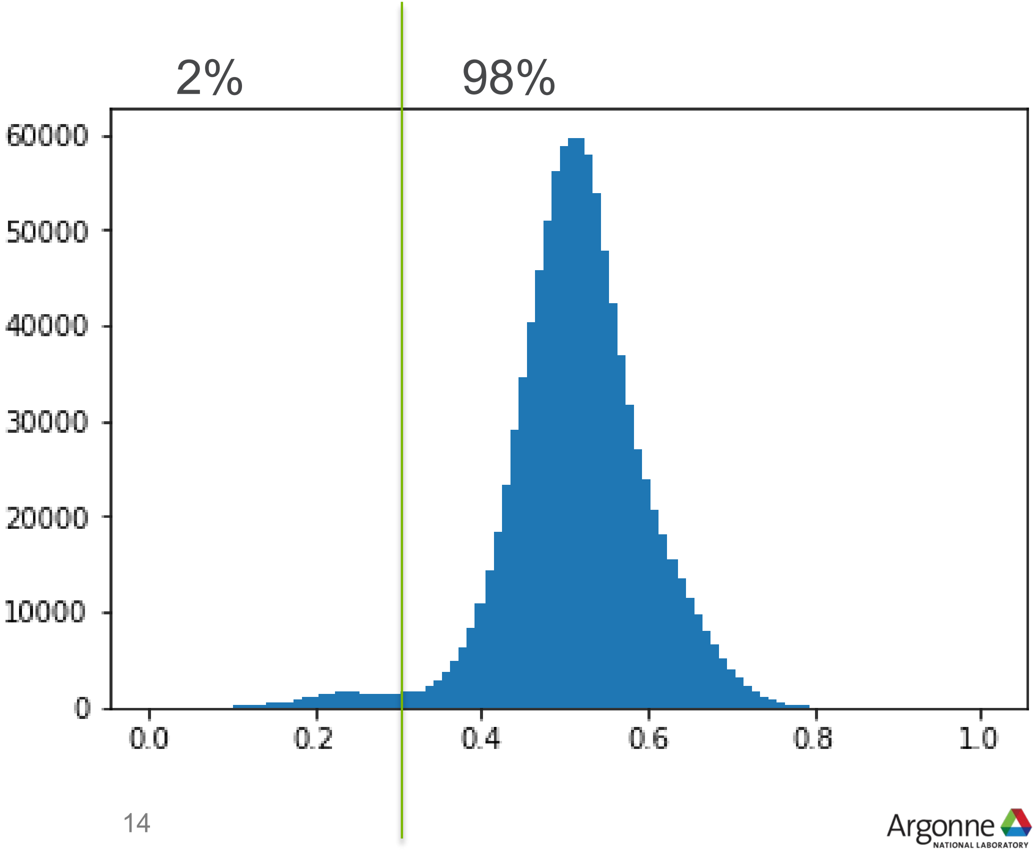

Learning Cuve for SMILES AURORA-1 Kinase Dock Score
validation r^2 score (same holdout)
1.0
0.8
0.6
0.4
0.2
0.0
0
25000
50000
75000
100000
125000
150000
175000
200000
Training set size

EXAMPLE: ML FOR DOCKING SCORING

Interested in the left tail

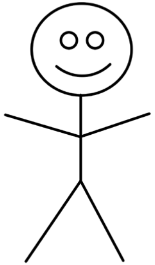
What is r2 score if we just guess everything in that right tail is clipped at the normal distribution? 0.75

Your balanced accuracy? 50%



- Each experiment cost $1,000
- Your boss wants to find leads at the very early stages.

Here is $100,000, find five interesting beta lactamase inhibitors



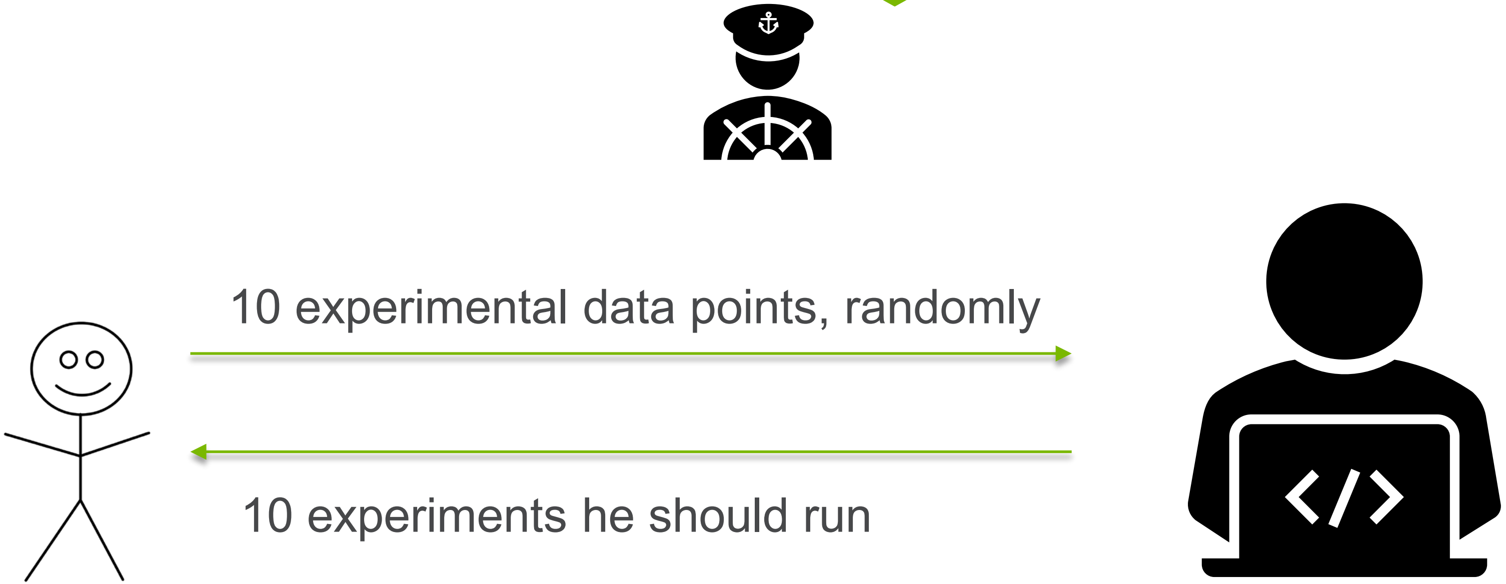
He finds 2 interesting compounds

Bob, the experimentalist

Bob’s experiments cost a lot of money
Bob has been working for the company for 10 years

- Each experiment cost $1,000
- Your boss wants to find leads at the very early stages.

Here is $100,000, find five interesting beta lactamase inhibitors



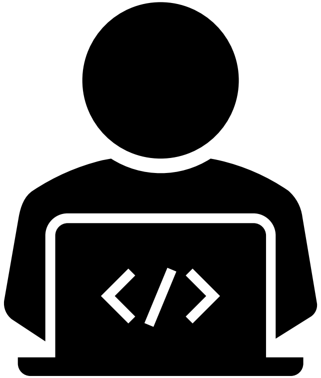
10 experimental data points, randomly

10 experiments he should run

Alice, the ML hacker

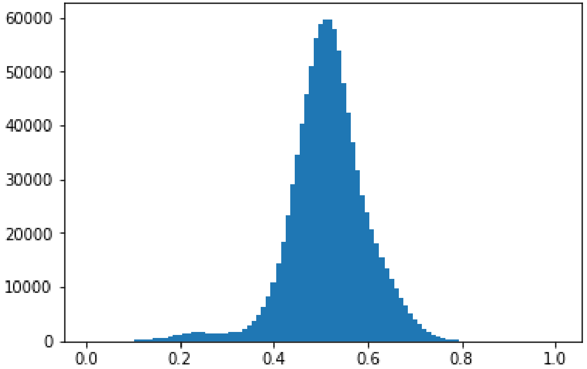
Bob runs 20 experiments, cost $20,000 –but he found 5 leads!

The r^2 value was 0.2

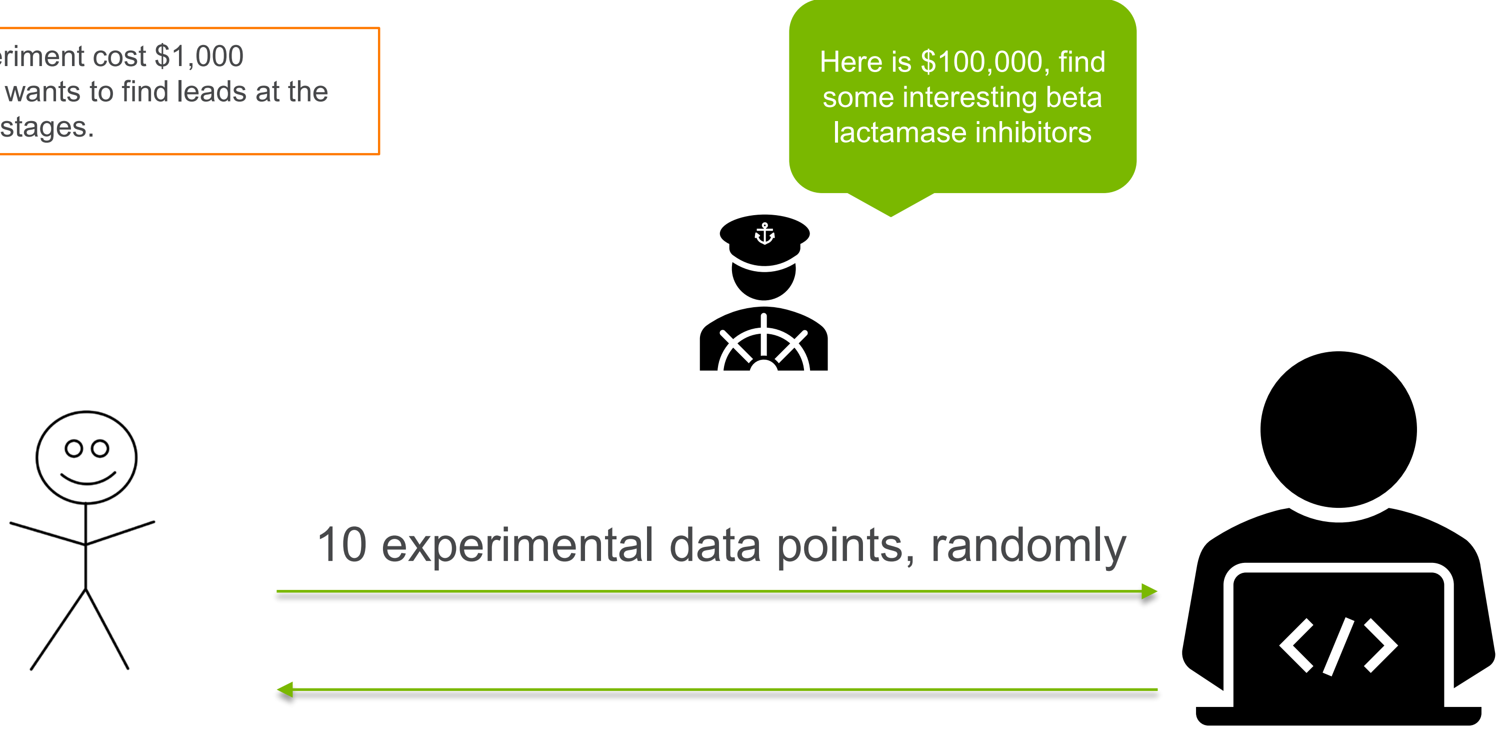


Alice, the ML hacker

Metrics measure distance in spaces, not real life goals, objectives
Dreams desires, etc! Especially, not on skewed distributions



- Each experiment cost $1,000
- Your boss wants to find leads at the very early stages.



Here is $100,000, find some interesting beta lactamase inhibitors

10 experimental data points, randomly

10 experiments he should run

Best x% of of your experimental values

Best x% of of your predicted values

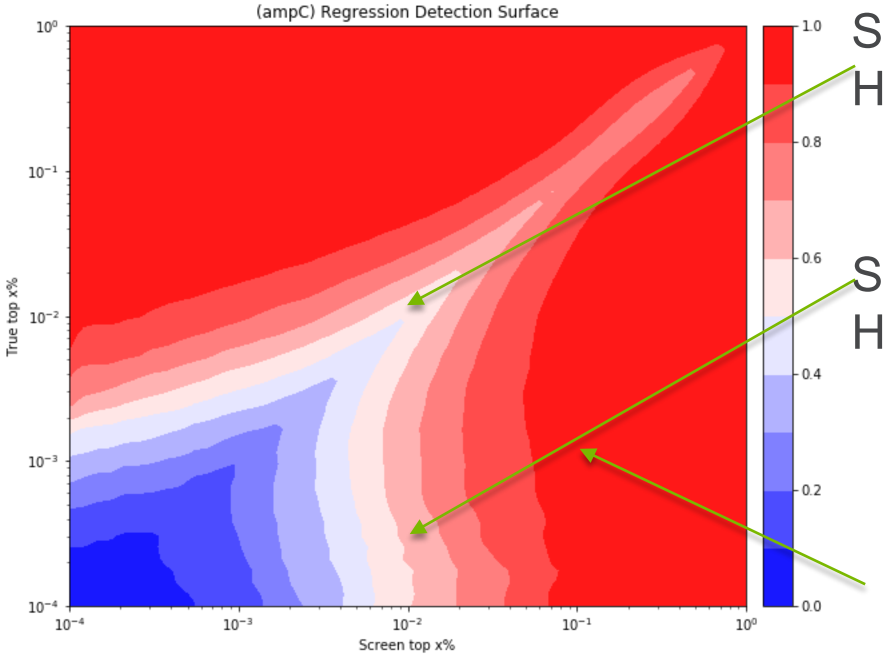
$$EF^{(\mathrm{COUNT})}_{x\%} = \frac{|\ \mathrm{TopR}(y,x) \cap \mathrm{TopR}(\hat{y},x)\ |}{xN}$$

How many values?

What if we replace the need to simulate every molecule?

Replicating Lyu et al. Giga-Docking with 200x less CPU compute

Trained message-passing network with 500K ampC



Screen 1% of molecules, you'll Have 50% of the true top 1%

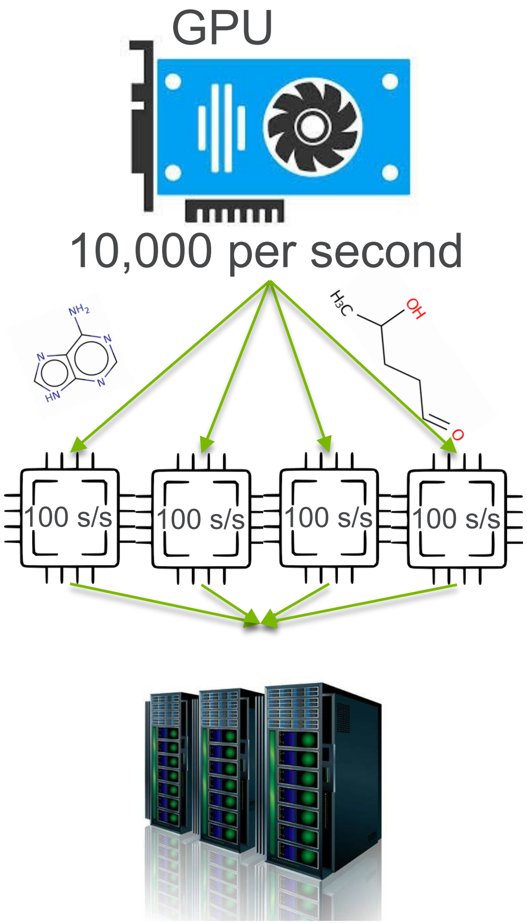
Screen 1% of molecules, you'll Have 70% of the true top 0.05%

Screen 10% of molecules, get all of the top 0.1%

U.S. DEPARTMENT OF ENERGY Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

*preliminary work, first approximation of a good model

Argonne NATIONAL LABORATORY


GPU

10,000 per second

100 s/s

100 s/s

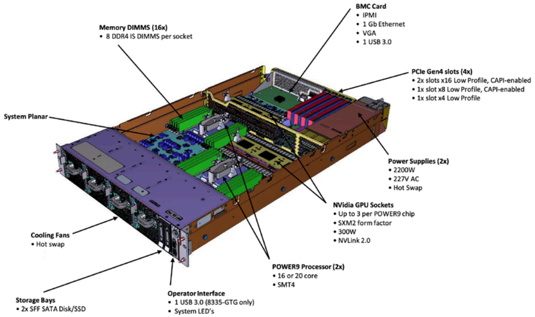
100 s/s

100 s/s

Super fast, modern generative algorithms

Single threaded algorithms for CPU post-processing

Even slower simulations

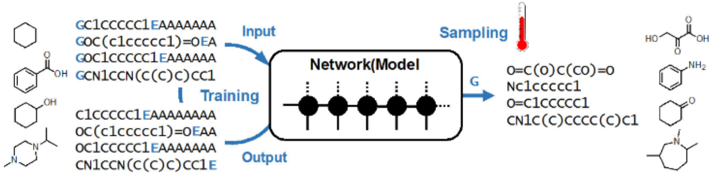


IBM AC922, 6 GPU node. Balanced
Heavily towards GPU, not CPU

5000 Seconds per smiles

1 SMILE per second

RNN SMILES Modeling

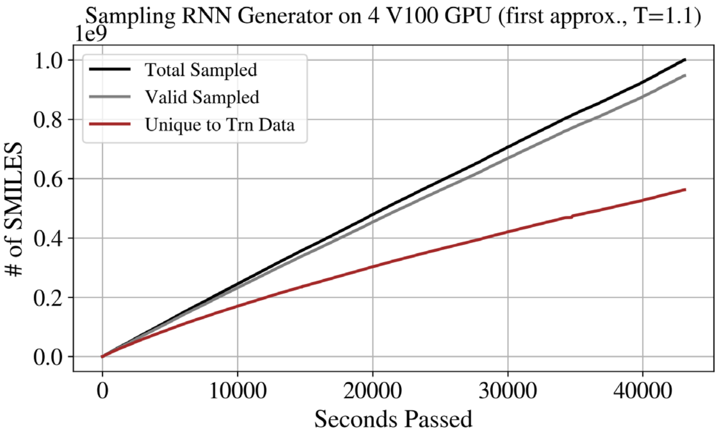


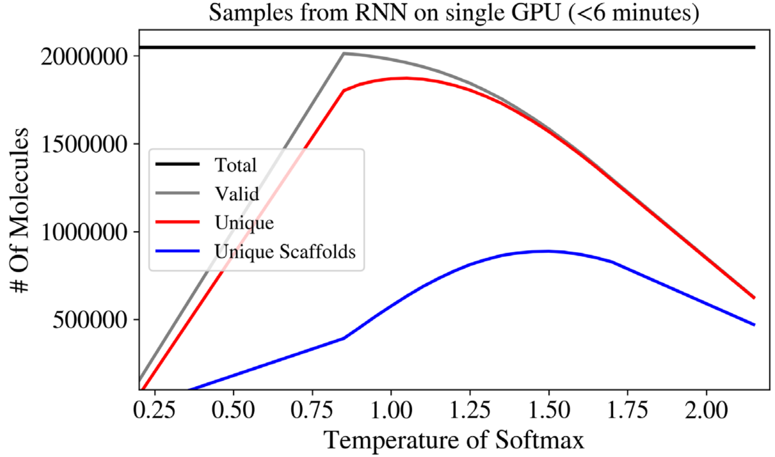
GC1CCCCC1EAAAAAAA
GOC(c1ccccc1)=OEA
GOC1CCCCC1EAAAAAA
GCN1CCN(C(C)C)CC1

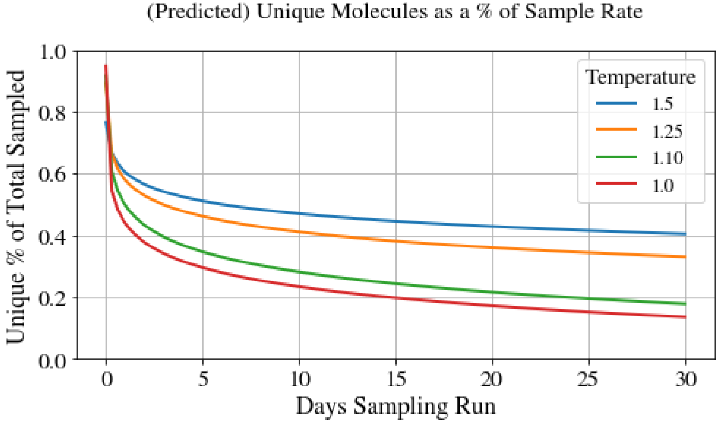
C1CCCCC1EAAAAAAAA
OC(c1ccccc1)=OEAA
OC1CCCCC1EAAAAAAA
CN1CCN(C(C)C)CC1E

O=C(O)C(CO)=O
Nc1ccccc1
O=C1CCCCC1
CN1C(C)CCCC(C)C1

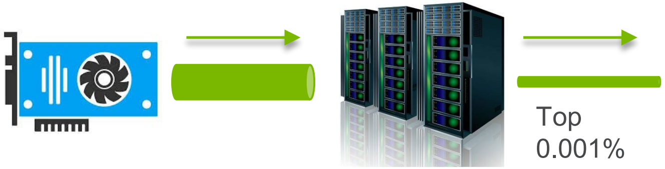
Gupta, Anvita, et al. "Generative recurrent networks for de novo drug design." Molecular informatics 37.1-2 (2018): 1700111.





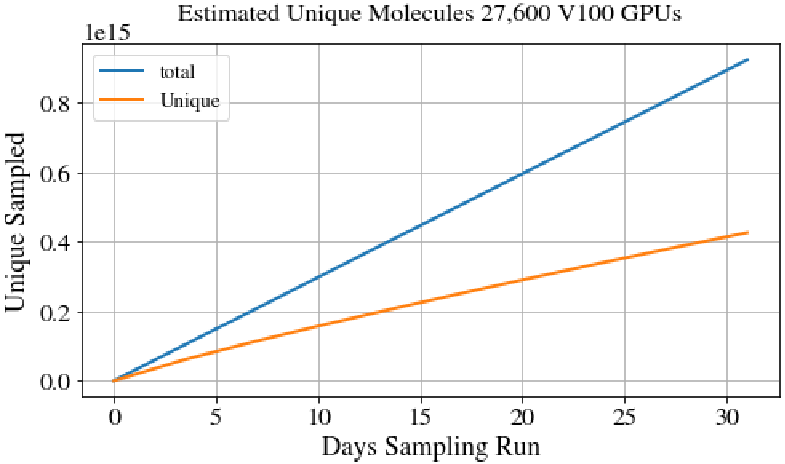


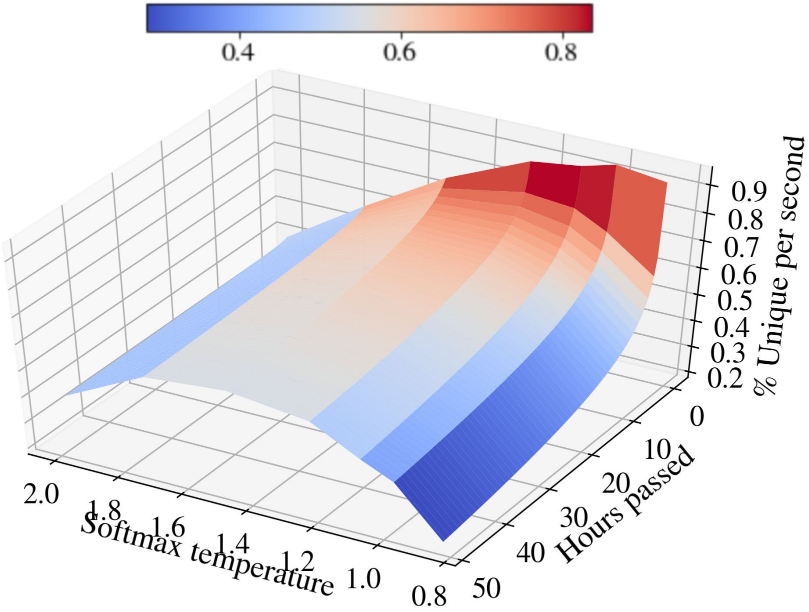
In order to keep GPUs and CPUs hot, unique stream of molecules needs to stay constant



Top 0.001%

Database, Experiments





$$p(s)_i = \frac{e^{-\beta s_i}}{\sum_{j=0}^{K} e^{-\beta s_j}}$$

DRUG DISCOVERY

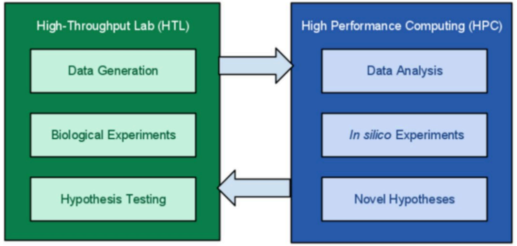
HIGH THROUGHPUT SCREENING

Generating Drug Leads

Database of Leads

- Generative Neural Networks
- Language modeling
- Graphical models

- Simulation surrogate models
- Uncertainty calibrated
- Ranking Neural networks



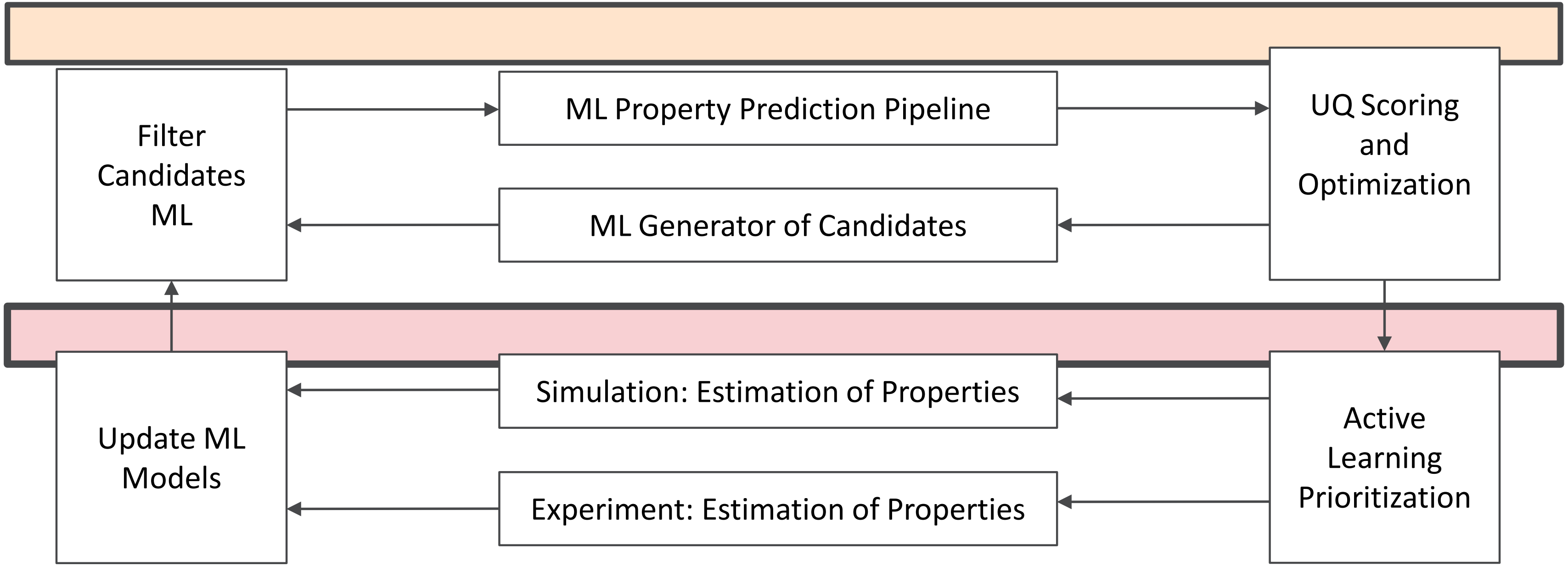
High-Throughput Lab (HTL)

- Data Generation
- Biological Experiments
- Hypothesis Testing

High Performance Computing (HPC)

- Data Analysis
- *In silico* Experiments
- Novel Hypotheses

LAYERED WORKFLOW



Filter Candidates ML

ML Property Prediction Pipeline

UQ Scoring and Optimization

ML Generator of Candidates

Update ML Models

Simulation: Estimation of Properties

Experiment: Estimation of Properties

Active Learning Prioritization

🟧 Pure ML “constant time” (fast loop)

🟥 Mixed/Variable time (slow loop)

THANKS!