AI FOR DRUG DISCOVERY

AUSTIN CLYDE
Assistant Computational Scientist
Data Science & Learning Division
Argonne National Laboratory
aclyde@anl.gov

ALCF
September 27th Virtual
Deep learning and chemistry

- In 2006, virtual screening could computationally screen roughly $10^5$ compounds.
- In 2018, deep-learning was applied directly to native molecular representation.
- In 2020, virtual screening with deep learning screened over $10^{10}$ compounds in a few days.
Summary: AI and Drug Design Key Contributions

**Accelerated virtual screening time by two orders of magnitude with no loss of detection**
- Large-scale virtual screening workflows on national supercomputing infrastructure at scale
- Discovery of a protease inhibitor

**Tiered-workflows for increased accuracy over standard VLS campaigns**

**Sampling strategies with visualizations to drive HPC workflows**
- Viz platform for chemical space
- Uses LLMs to navigate and generate large graph structure efficiently
- Based on an atlas of chemical space through scaffolds/shape

**Workflow and economic analysis**
- VLS is bottlenecked by modeling, not compute
- Higher-throughput experimental techniques can drive deeper chemical probes
- Active-learning loops may be able to help with more complex simulations

---

<table>
<thead>
<tr>
<th>10,000,000,000 compounds screened with AI models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top: 2.5%</td>
</tr>
<tr>
<td>250,000,000 poses docked</td>
</tr>
<tr>
<td>Top: 2.5%</td>
</tr>
<tr>
<td>6,250,000 systems build and minimized</td>
</tr>
<tr>
<td>Top: 2.5%</td>
</tr>
<tr>
<td>156,250 systems simulated</td>
</tr>
<tr>
<td>(that's about 12H on 1024 summit nodes)</td>
</tr>
</tbody>
</table>
Drug discovery and basic science

• Cost per new drug range from less than $1 billion to more than $2 billion per drug

• The federal government is the primary funder of basic research in biomedical sciences. That research ultimately increases the supply of new drugs because drug companies rely on the findings from that research—for example, the identification of disease targets toward which new drug therapies can be aimed

• Between 2010 and 2016, every drug approved by the FDA was in some way based on biomedical research funded by NIH.

Source: Research and Development in the Pharmaceutical Industry, Congressional Budget Office August 2021
Drug Discovery Funnel

DRUG DISCOVERY
~$10^8$ products

PRE CLINICAL
11,000 products

CLINICAL TRIALS
6,300 products

FDA APPROVAL
111 products
TARGETS AND BINDING SITES

Pocket 1
Score: 0.915
Druggability Score: 0.920
Number of Alpha Spheres: 80
Total SASA: 16.657
Polar SASA: 2.165
Apolar SASA: 14.492
Volume: 599.003
Mean local hydrophobic density: 18.690
Mean alpha sphere radius: 3.963
Mean alp. sph. solvent access: 0.523
Apolar alpha sphere proportion: 0.363
Hydrophobicity score: 33.000
Volume score: 3.143
Polarity score: 4
Charge score: 0
Proportion of polar atoms: 39.583
Alpha sphere density: 5.345
Cent. of mass - Alpha Sphere max dist: 14.313
Flexibility: 0.118

Pocket 2
Score: 0.689
Druggability Score: 0.834
Number of Alpha Spheres: 67
Total SASA: 8.089
Polar SASA: 3.259
Apolar SASA: 4.831
Volume: 367.098
Mean local hydrophobic density: 20.545
Mean alpha sphere radius: 3.909
Mean alp. sph. solvent access: 0.483
Apolar alpha sphere proportion: 0.328
Hydrophobicity score: 27.125
Volume score: 2.875
Polarity score: 3
Charge score: 1
Proportion of polar atoms: 40.541
Alpha sphere density: 3.665
Cent. of mass - Alpha Sphere max dist: 10.679
Flexibility: 0.124

Automatic pocket detection
STRUCTURAL DOCKING

Exhaustive shape fitting

- docking is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex.
- It is a simple problem to understand, figure out how to fit the ligand onto the protein.
- The search space in theory consists of all possible orientations and conformations of the protein paired with the ligand.

Inputs: molecular dataset (2D SMILES strings), target protein structure, search parameters, scoring function $f$

For ligand pose from strategy

Result = $\max(\text{Result}, f(\text{Protein pose, Ligand pose}))$
MD SIMULATION

Molecular dynamics simulation can be used to observe and model these different states of a protein.

(Right) Protein states can be clustered and modeled. Each point represents a unique 3D structure that the protein took on during its interactions with itself and the ligand.
AI AND VIRTUAL SCREENING HPC WORKFLOWS

(A) Molecule → Docking Program → Pose → Score

ML “Surrogate” Model

(D) Property Models & Optimization

Active Learning

Generative Model

Docking Program → Pose → Score

Compound Libraries

ML “Surrogate” Model

(B) Molecule → Docking Program → Pose → Score

ML “Surrogate” Model

(C) Molecule → Docking Program → Pose

Experimental Data

ML Rescoring Model

BFE prediction

Chapter 13
Ultrahigh Throughput Protein–Ligand Docking with Deep Learning
Austin Clyde

Abstract

Ultrahigh-throughput virtual screening (UH-TV) is an emerging field bringing together classical docking techniques with high-throughput AI methods. We describe mechanisms, docking models, goals, and outcome. We present different AI accelerated workflows for UH-TV, mostly through surrogate docking models. We showcase a novel framework implemented in our molecular docking platform, as an example of the end-to-end process from target identification to ligand prioritization. We demonstrate an integrated UH-TV pipeline with machine learning modules integrated in the target of billions scale, we outline a future for AI-driven screening platforms with deep learning.

Key words: Drug discovery, Protein-ligand docking, Deep learning, Graph convolution, Virtual screening, Chemical screening.
Simulation: Estimation of Properties
Update ML Models
Filter Candidates ML
ML Property Prediction Pipeline
ML Generator of Candidates
UQ Scoring and Optimization
Simulation: Estimation of Properties
Experiment: Estimation of Properties
Active Learning Prioritization

pure ML “constant time” (fast loop)
mixed/variable time (slow loop)
DRUG DISCOVERY

• Generative Neural Networks
• Language modeling
• Graphical models

HIGH THROUGHPUT SCREENING

• Simulation surrogate models
• Uncertainty calibrated
• Ranking Neural networks

Generating Drug Leads

Database of Leads

High-Throughput Lab (HTL)
- Data Generation
- Biological Experiments
- Hypothesis Testing

High Performance Computing (HPC)
- Data Analysis
- In silico Experiments
- Novel Hypotheses
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
WORKFLOW ANALYSIS

Surrogate Prefilter then Dock (SPFD)

- With TD we understand that $\rho_L$ hits generally gets an active lead rate around X%.
- How can we be sure the top $\sigma_L$ compounds that come from the model capture all those $\rho_L$ compounds we want?

L Molecules $\rightarrow$ $\sigma_L$ Hits $\rightarrow$ $\rho_L$ Hits $\rightarrow$ Active leads
Four docking models...

Starting points:
- Crystal structures and structural models
- Multiple antibody templates
- Databases of purchasable small molecules

Outputs:
- Designs with probability of:
  - Desired activity
  - Desired biological effect
  - Good physical and safety parameters

Platform capability build funded over time through DOE, LDRD, DARPA, DOD, and other funding sources
Explore the Chemical Space

Supported graph operations:
- Upper
- Lower
- Expand

If a compound is generated by applying a graph operation on another compound, the 2 compounds are connected.

The color bars represent attribute encoding on nodes. Here we used Wildman-Crippen LogP value of the compound.
THANKS!
aclyde@anl.gov