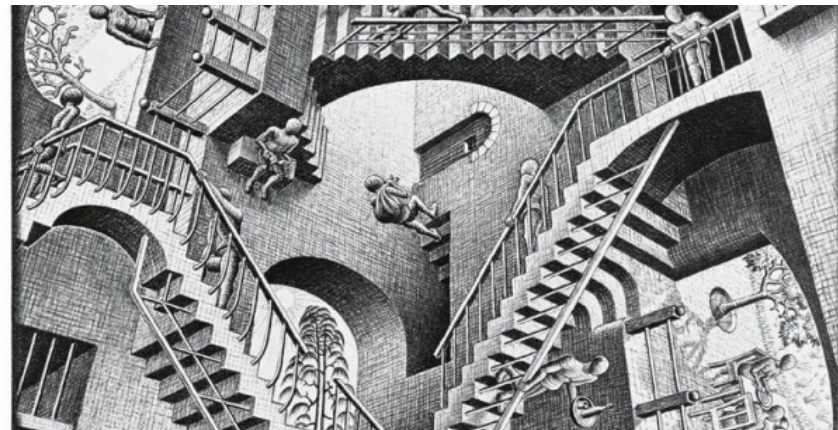


AI FOR DRUG DISCOVERY

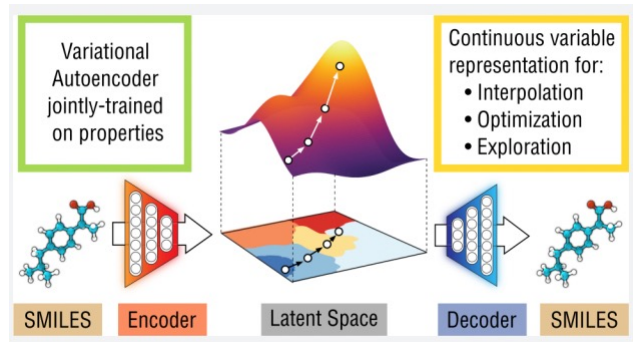


AUSTIN CLYDE
Assistant Computational Scientist
Data Science & Learning Division
Argonne National Laboratory

aclyde@anl.gov

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.



Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,^{†,‡,§,¶} Jennifer N. Wei,^{†,‡,§,¶} David Duvenaud,^{¶,¶} José Miguel Hernández-Lobato,^{S,¶} Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,^{‡,§} Jorge Aguilera-Iparraguirre,[‡] Timothy D. Hirzel,[‡] Ryan P. Adams,^{∇,||} and Alán Aspuru-Guzik^{*,†,‡,§,¶}

[†]Kyulux North America Inc., 10 Post Office Square, Suite 800, Boston, Massachusetts 02109, United States

[‡]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

[§]Department of Computer Science, University of Toronto, 6 King's College Road, Toronto, Ontario M5S 3H5, Canada

[§]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.

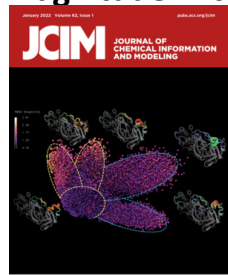
[∇]Google Brain, Mountain View, California, United States

^{||}Princeton University, Princeton, New Jersey, United States

^{*}Biologically-Inspired Solar Energy Program, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario M5S 1M1, Canada

Summary: AI and Drug Design Key Contributions

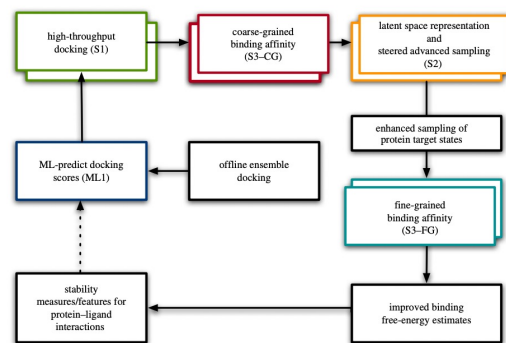
Accelerated virtual screening time by two orders of magnitude with no loss of detection



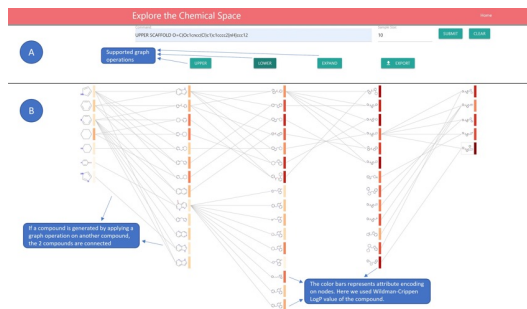
- Large-scale virtual screening workflows on national supercomputing infrastructure at scale
- Discovery of a protease inhibitor

Tiered-workflows for increased accuracy over standard VLS campaigns

10,000,000,000 compounds screened with AI models
Top 2.5%
250,000,000 poses docked
Top 2.5%
6,250,000 systems build and minimized
Top 2.5%
156,250 systems simulated
(that's about 12H on 1024 summit nodes)

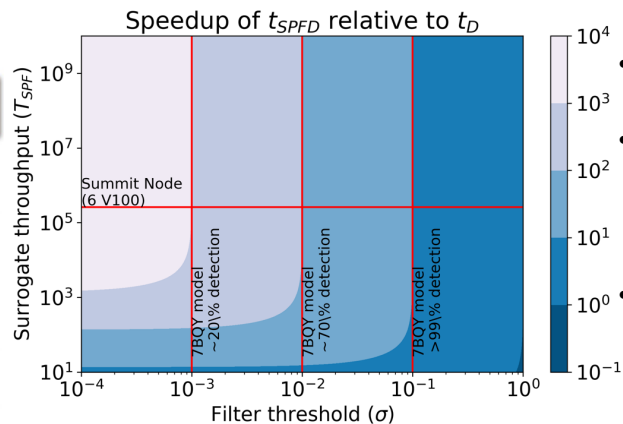


Sampling strategies with visualizations to drive HPC workflows



- 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



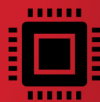
- VLS is bottlenecked by modeling, not compute
- Higher-throughput experimental techniques can drive deeper chemical probes
- Active-learning loops may be able to help with more complex simulations

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.

DRUG DISCOVERY

~10⁸ products



PRE CLINICAL

11,000 products



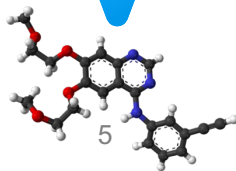
CLINICAL TRIALS

6,300 products



FDA APPROVAL

111 products

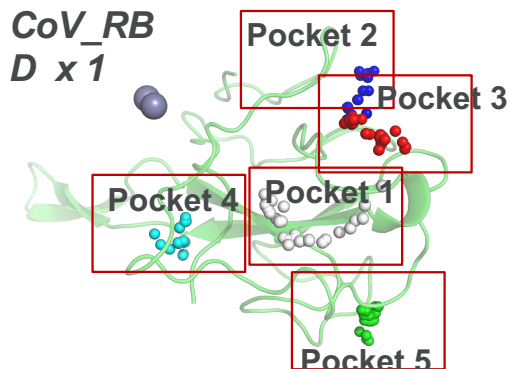
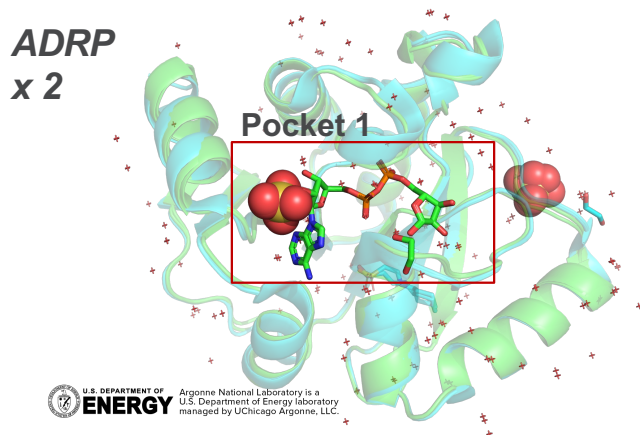
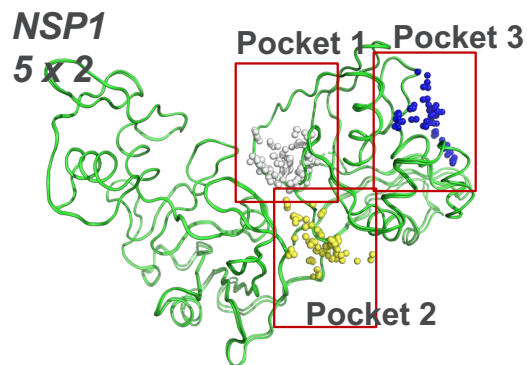
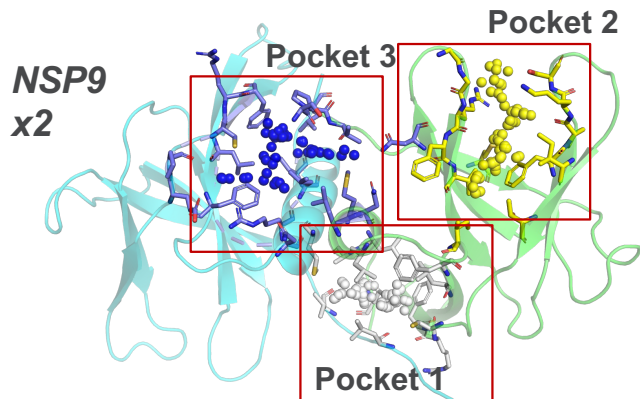


Drug Discovery Funnel

Automatic pocket detection

Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009). <https://doi.org/10.1186/1471-2105-10-168>

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

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

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

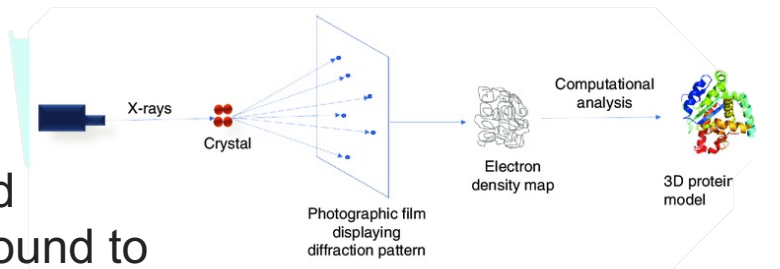
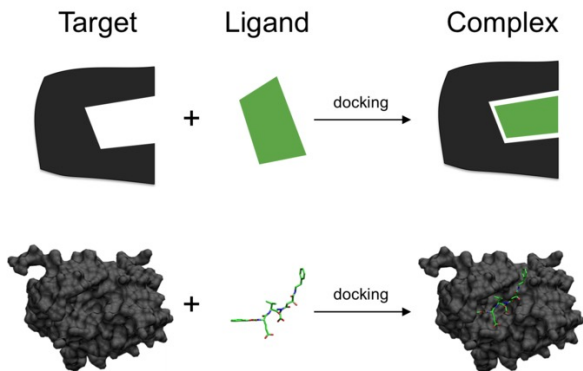


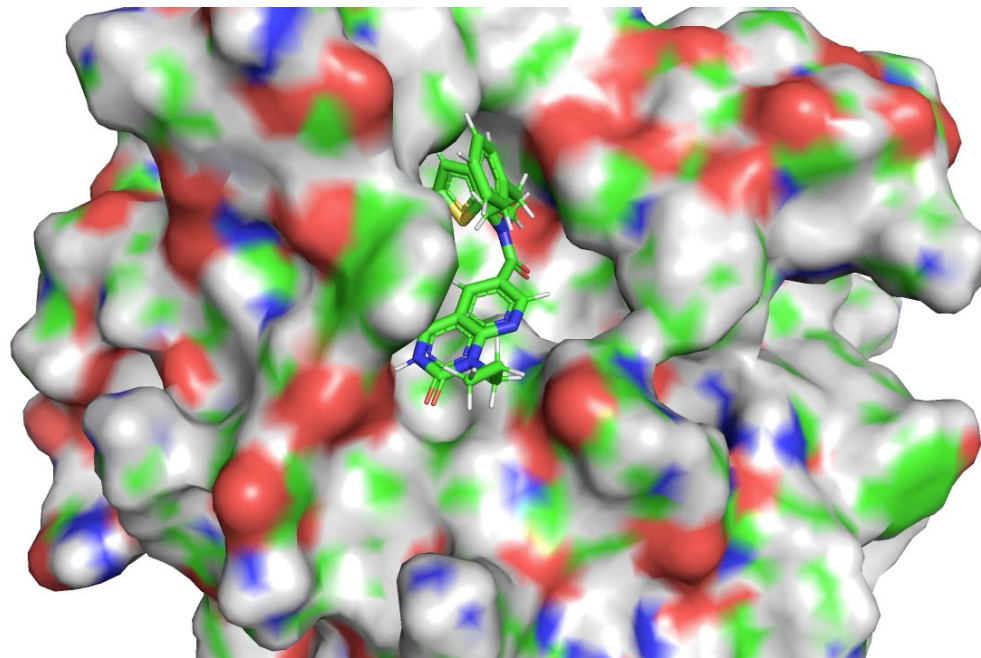
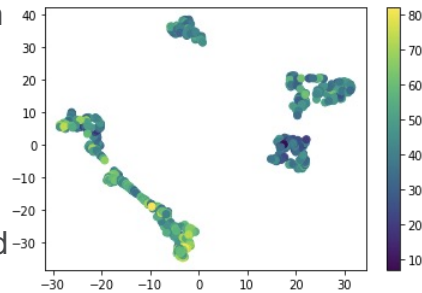
Diagram of X-ray crystallography



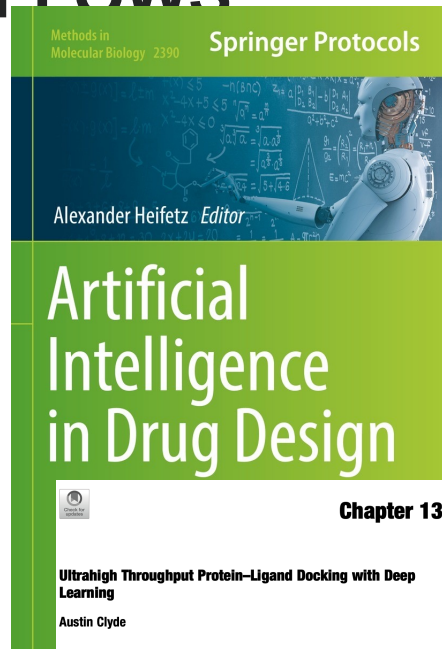
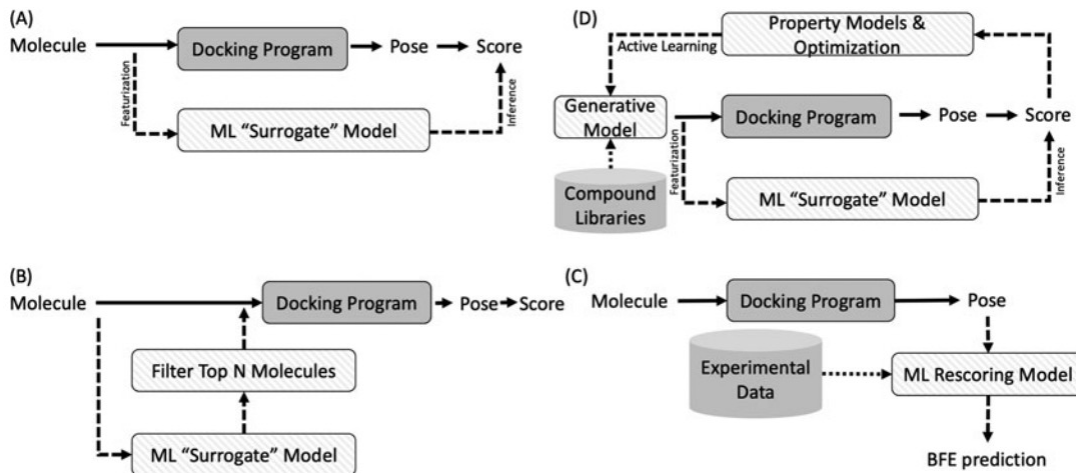
MD SIMULATION

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

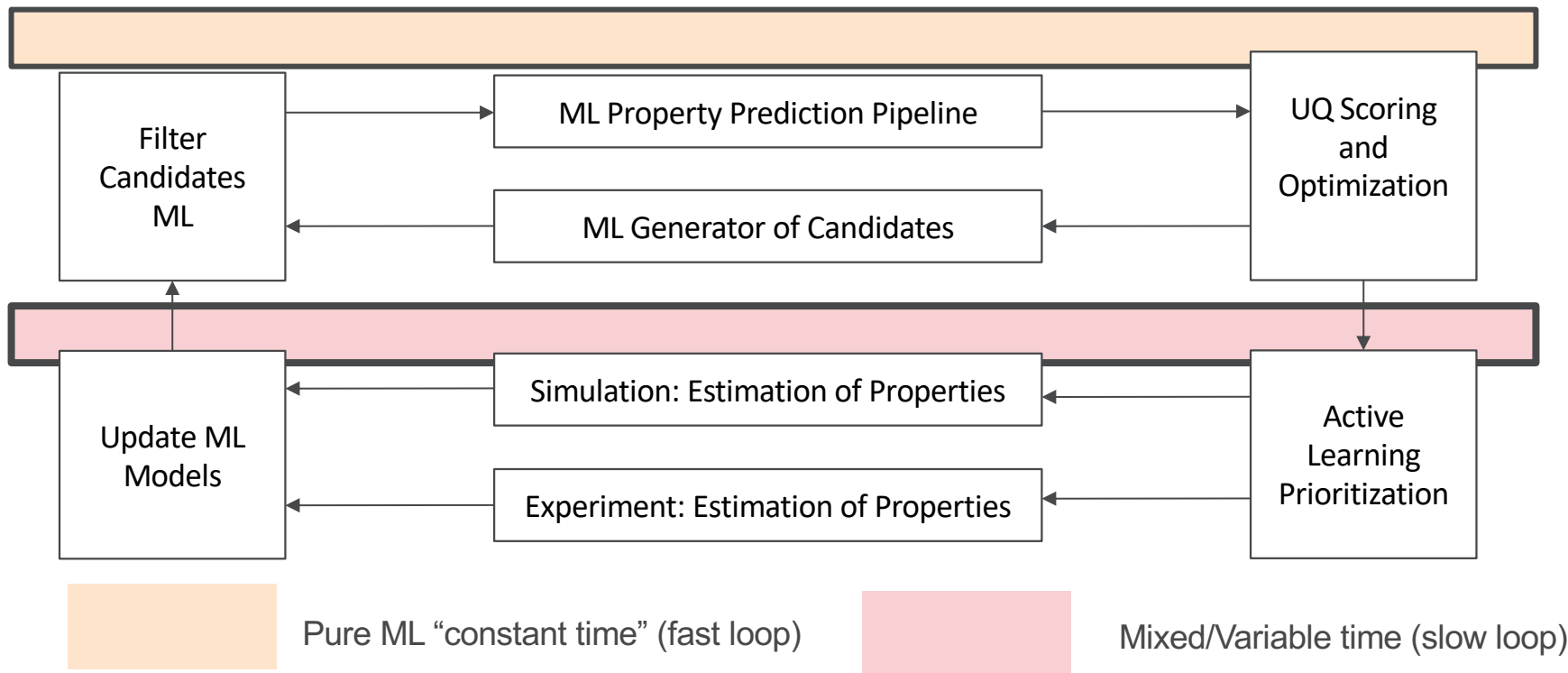


Abstract

Ultrahigh-throughput virtual screening (uHTVS) is an emerging field linking together classical docking techniques with high-throughput AI methods. We outline mechanistic docking models' goals and successes. We present different AI accelerated workflows for uHTVS, mainly through surrogate docking models. We showcase a novel feature representation technique, molecular depictions (images), as a surrogate model for docking. Along with a discussion on analyzing screens using regression enrichment surfaces at the tens of billion scale, we outline a future for uHTVS screening pipelines with deep learning.

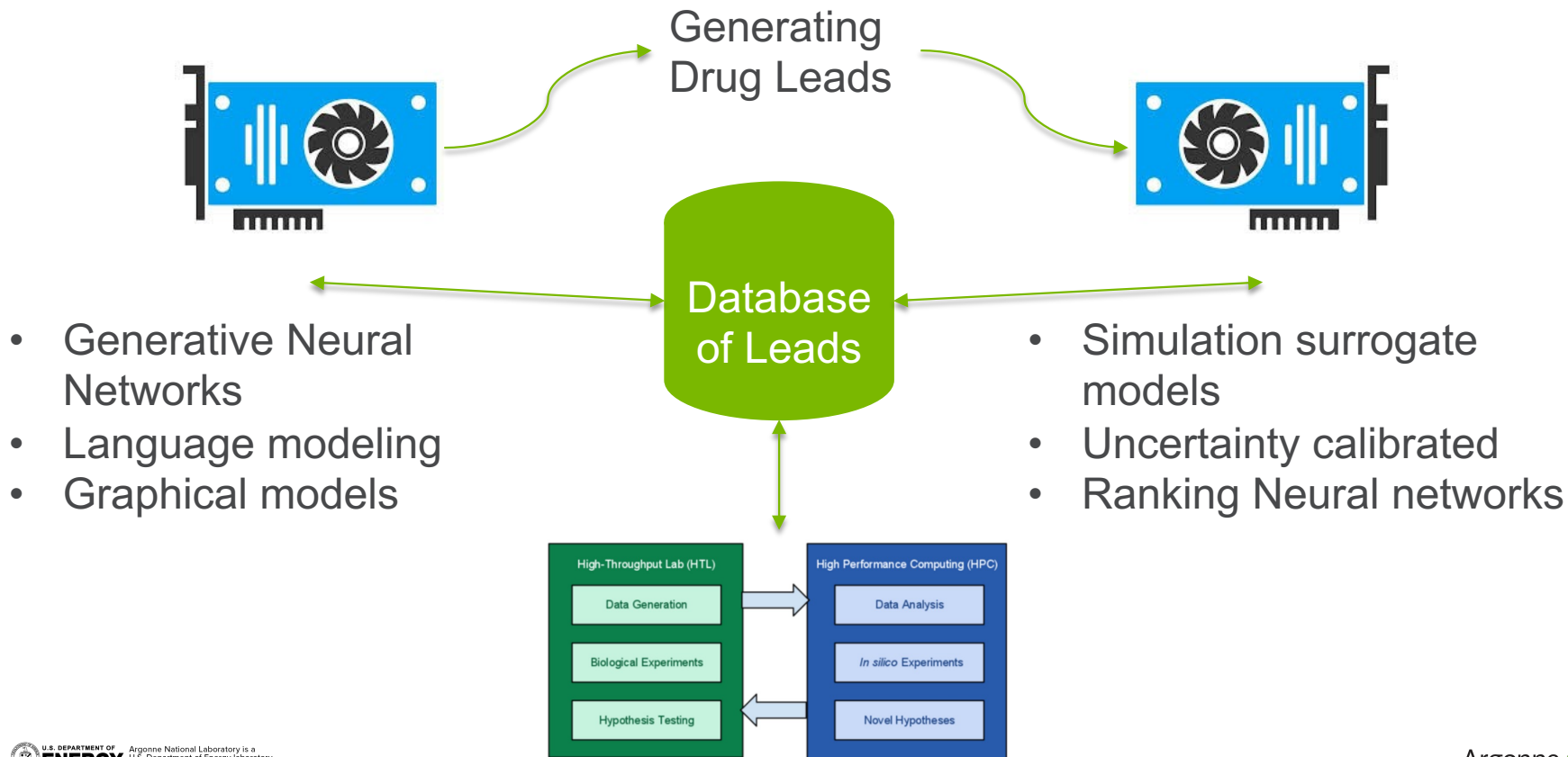
Key words Drug discovery, Protein-ligand docking, Deep learning, Graph convolution, Virtual screening, Chemical screening

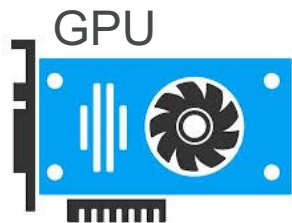
LAYERED WORKFLOW



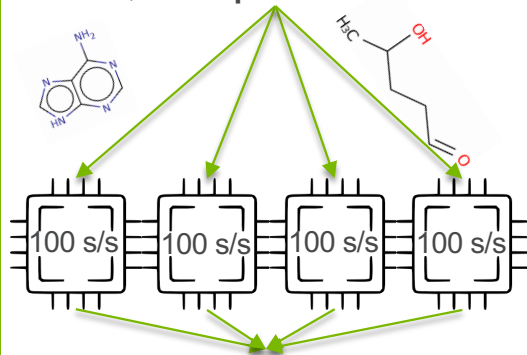
DRUG DISCOVERY

HIGH THROUGHPUT SCREENING





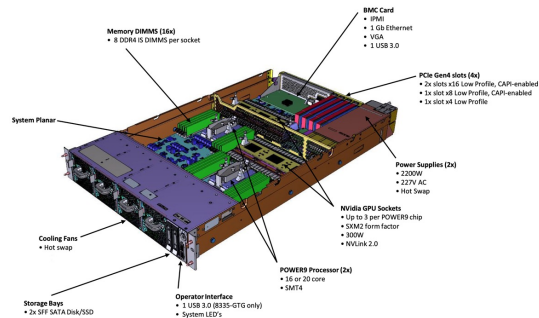
10,000 per second



Super fast, modern generative algorithms

Single threaded algorithms for CPU post-processing

Even slower simulations

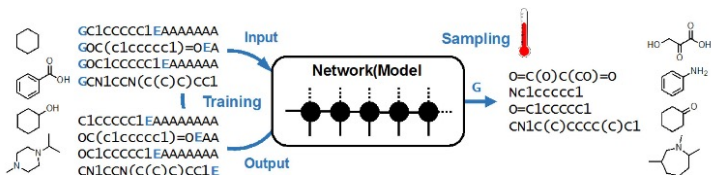


IBM AC922, 6 GPU node. Balanced
Heavily towards GPU, not CPU

5000 Seconds per smiles

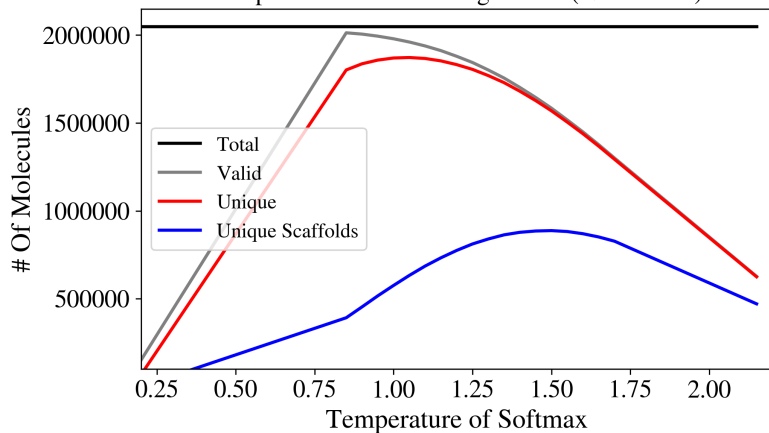
1 SMILE per second

RNN SMILES Modeling

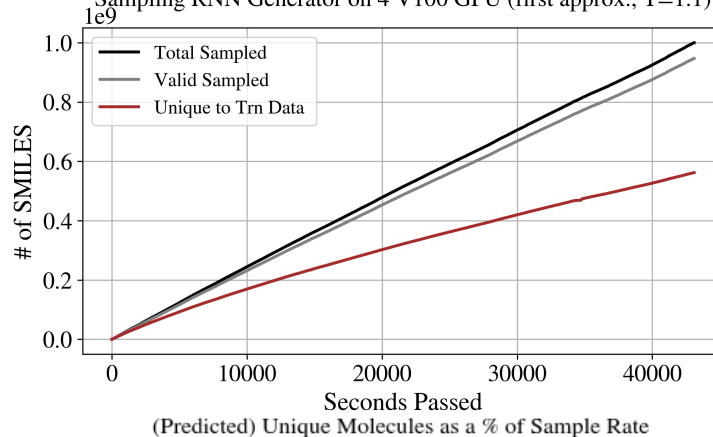


Gupta, Anvita, et al. "Generative recurrent networks for de novo drug design." *Molecular informatics* 37.1-2 (2018): 1700111.

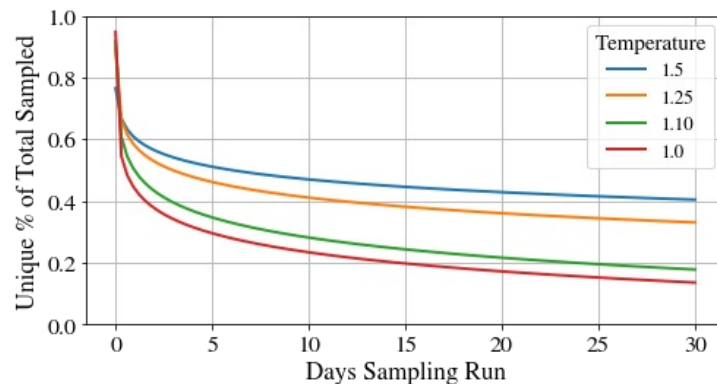
Samples from RNN on single GPU (<6 minutes)



Sampling RNN Generator on 4 V100 GPU (first approx., T=1.1)



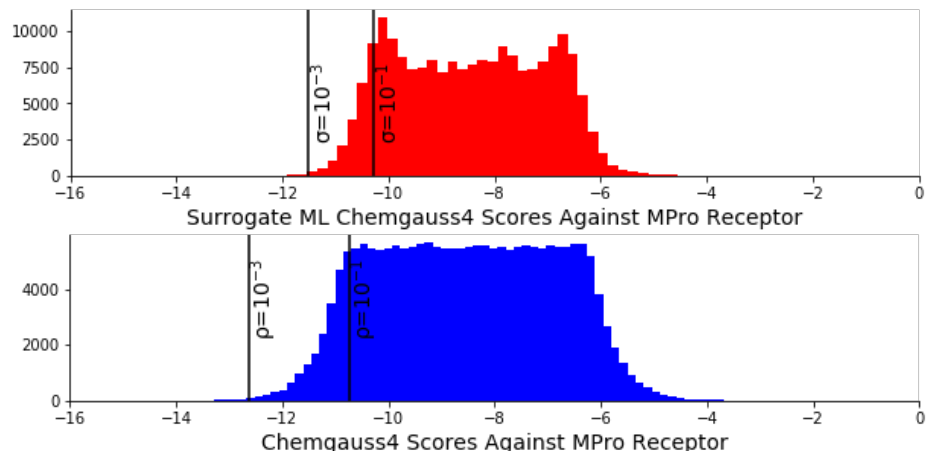
(Predicted) Unique Molecules as a % of Sample Rate



WORKFLOW ANALYSIS

Surrogate Prefilter then Dock (SPFD)

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

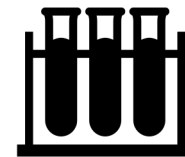
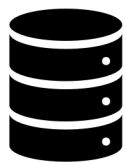


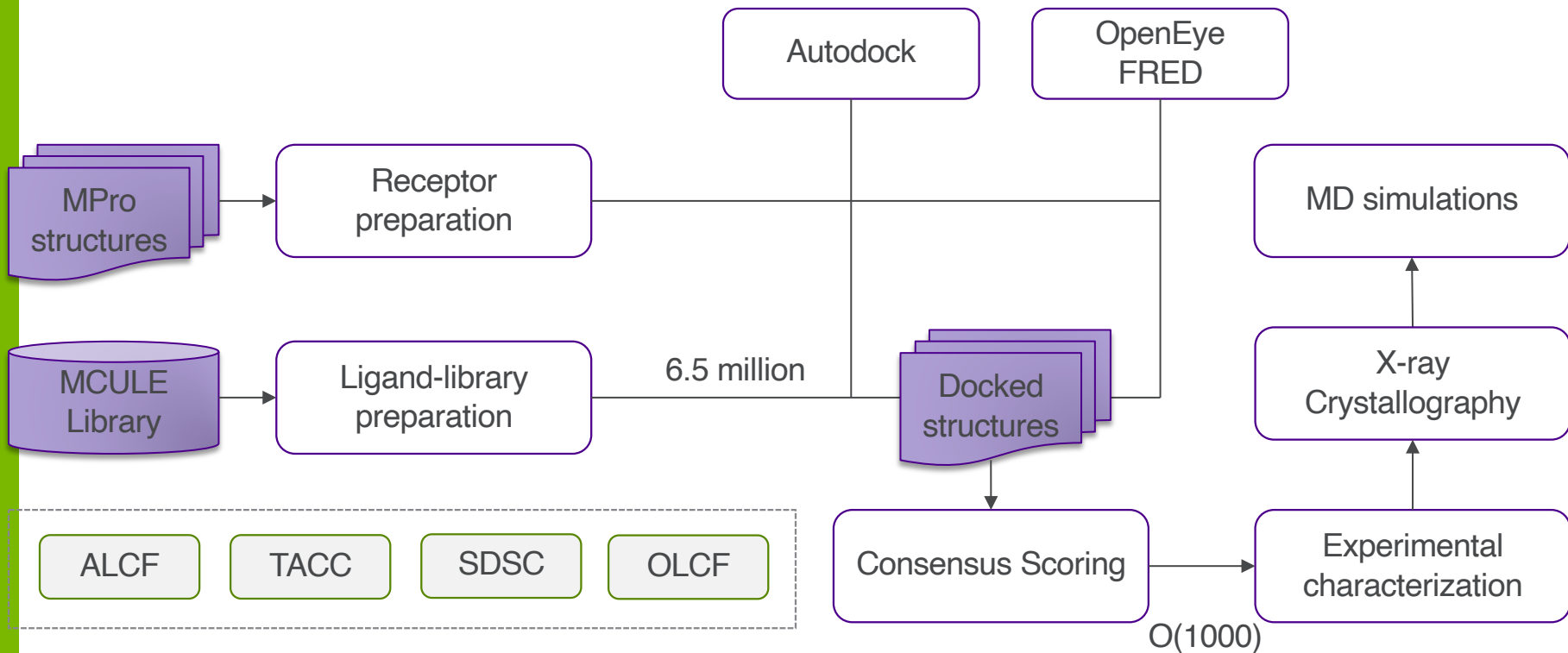
L Molecules

σL Hits

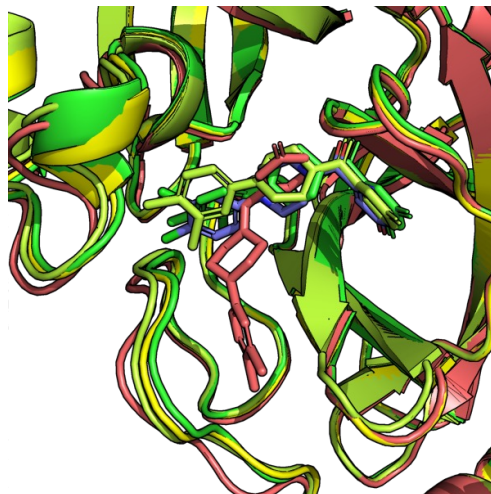
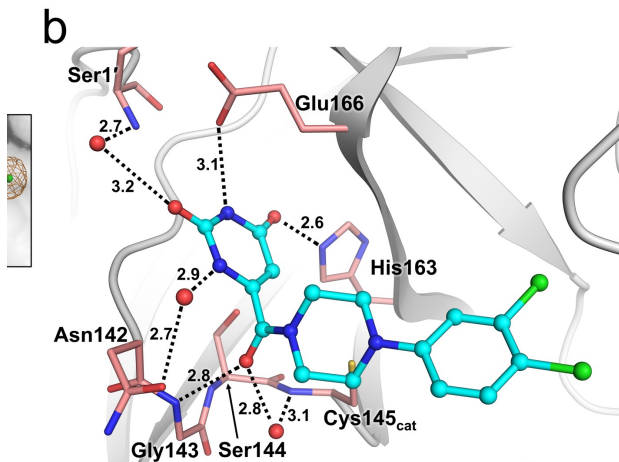
ρL Hits

Active leads



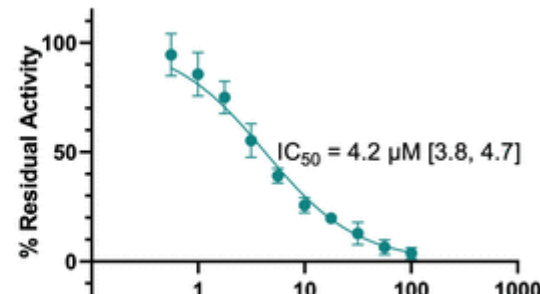
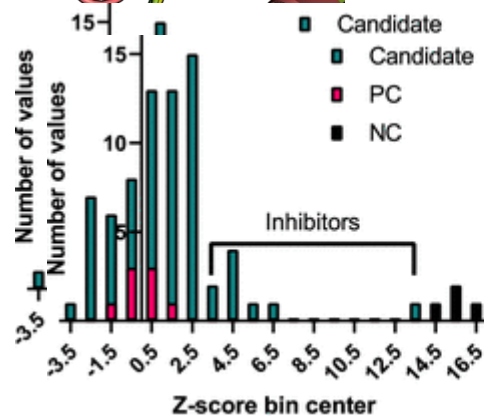
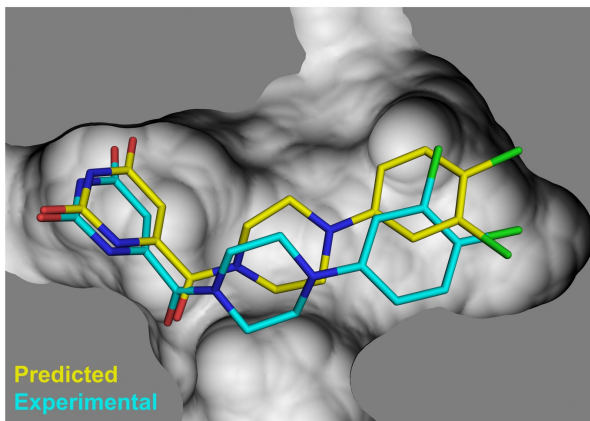


Four docking models...



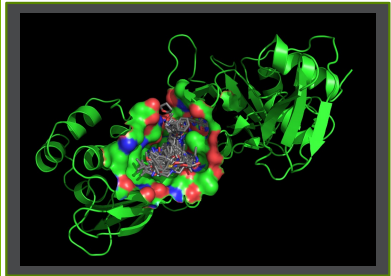
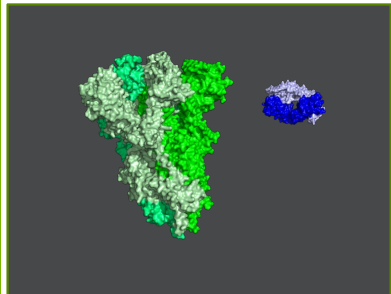
Clyde, Austin, Stephanie Galanie, Daniel W. Kneller, Heng Ma, Ya Blaiszik, Alexander Brace et al. "High-throughput virtual screening sars-cov-2 main protease noncovalent inhibitor." *Journal of chemical modeling* 62, no. 1 (2021): 116-128.

d



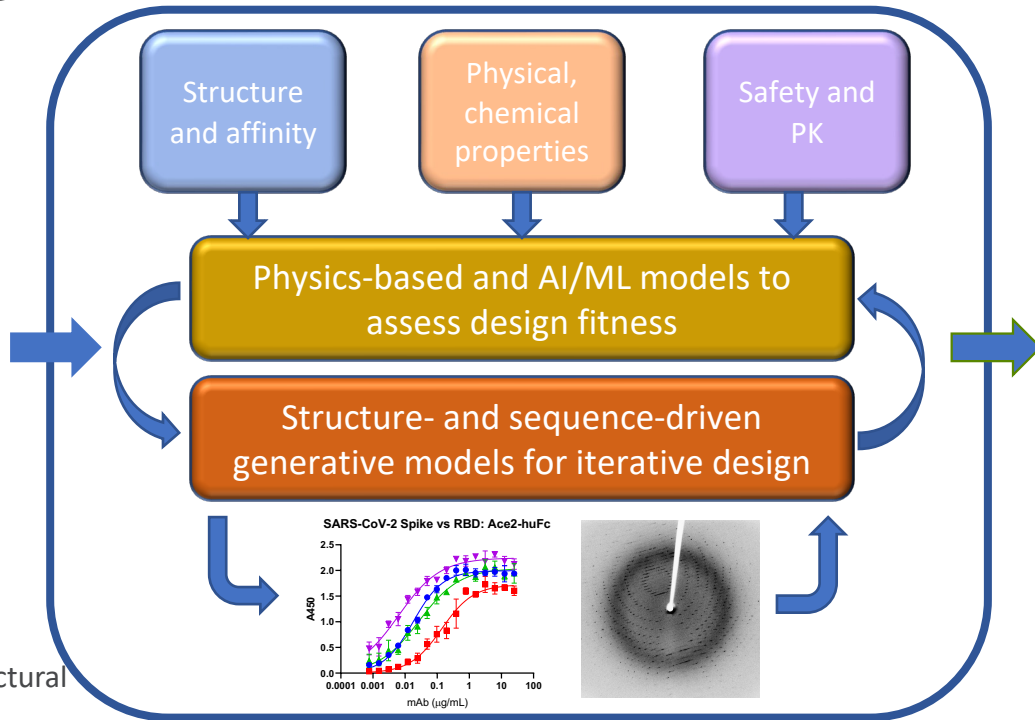
Blaiszik, Alexander Brace et al. "High-throughput virtual screening and validation of a sars-cov-2 main protease noncovalent inhibitor." *Journal of chemical information and modeling* 62, no. 1 (2021): 116-128.

COMPUTATIONAL AND EXPERIMENTAL DESIGN PLATFORMS

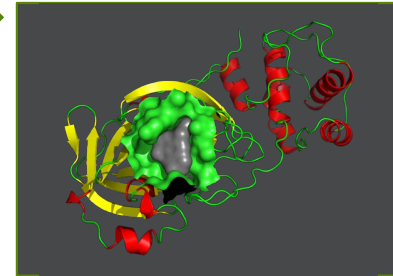
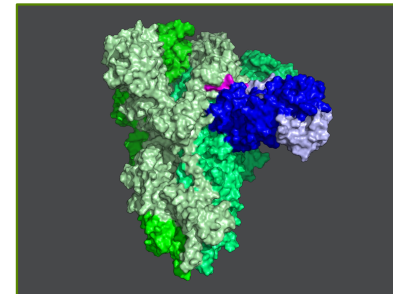


Starting points:

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



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



Outputs:

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

Command:

UPPER SCAFFOLD O=C(Oc1cnc(Cl)c1)c1cccc2[nH]ccc12

Sample Size:

10

SUBMIT

CLEAR


A

Supported graph operations

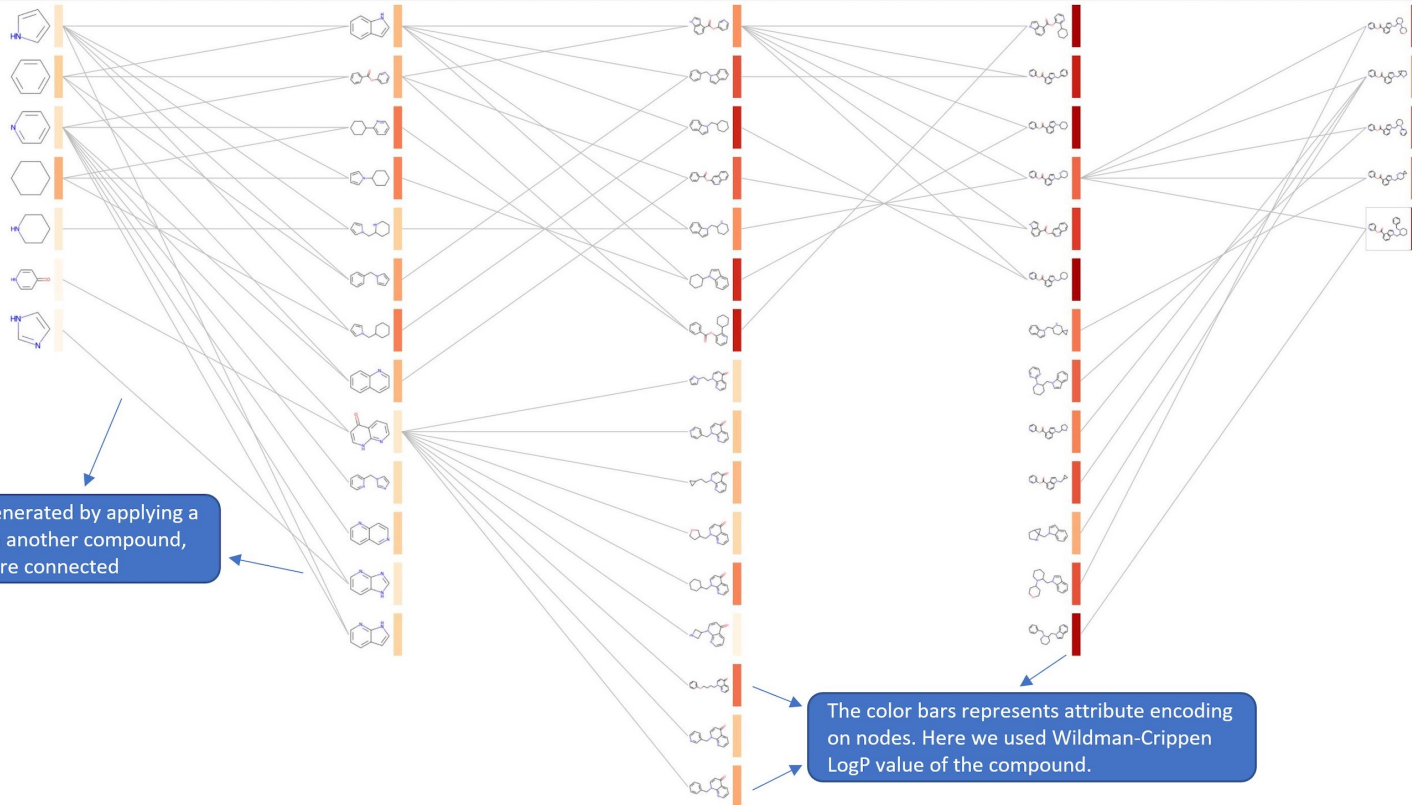
UPPER

LOWER

EXPAND

 EXPORT

B



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

The color bars represents attribute encoding on nodes. Here we used Wildman-Crippen LogP value of the compound.

THANKS!

aclyde@anl.gov



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

