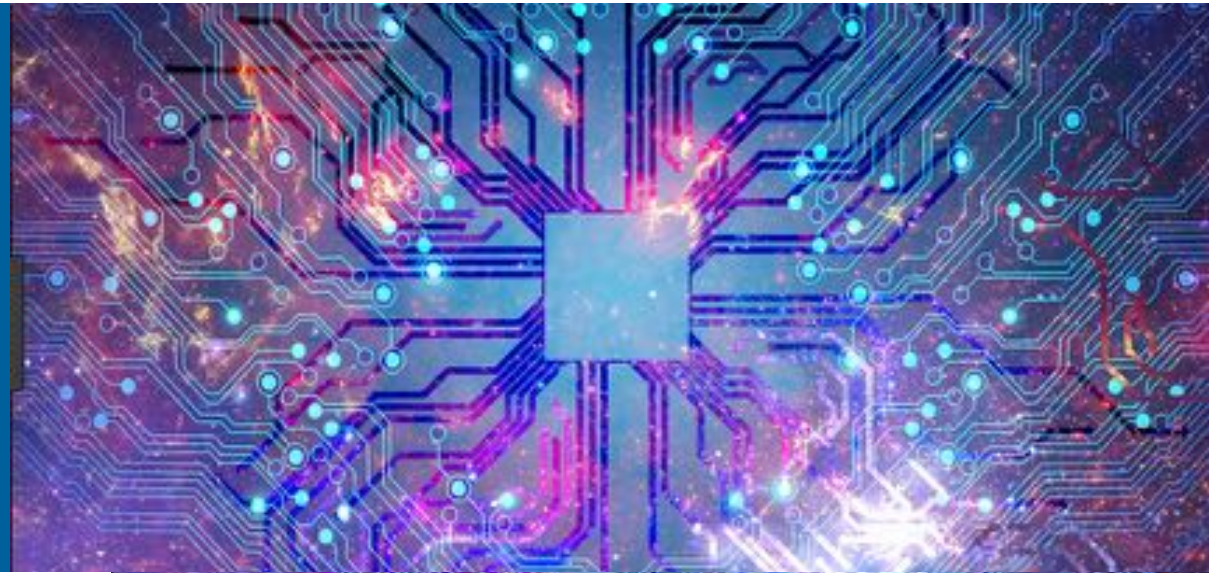
End-to-End Computational Drug Design for COVID-19: From Screening to Series and Back Again



AUSTIN CLYDE

Assistant computational scientist, Argonne National Laboratory

aclyde@anl.gov
www.cs.uchicago.edu/~aclyde

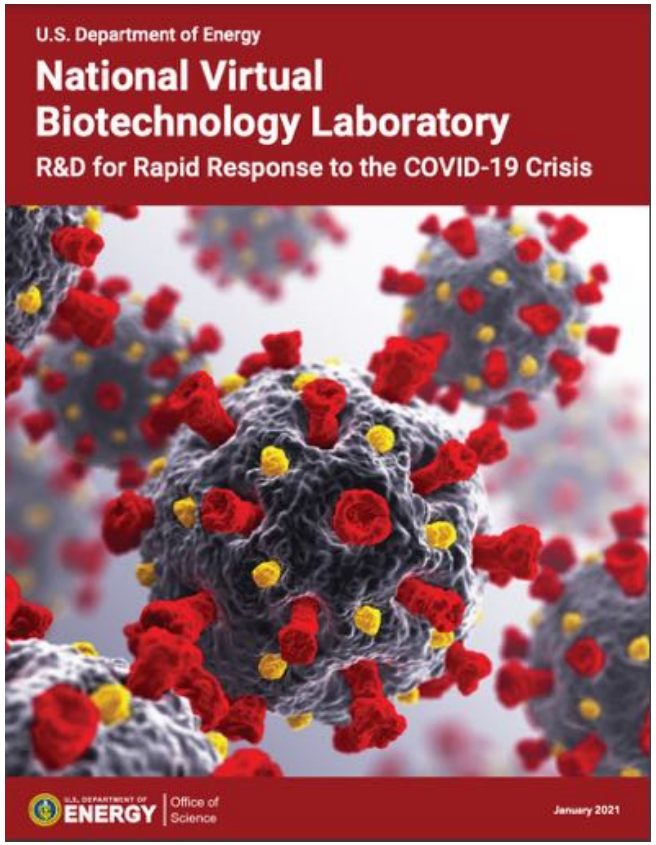
CompBioMed All Hands
June 22, 2022

1. Summary of COVID-19 Results
2. Take aways for building a computational bio preparedness program
3. What AI can do for late-stage drug discovery

National Virtual Biotechnology Laboratory

Research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

https://science.osti.gov/nvbl



NVBL Molecular Therapeutics Team

Ai Kagawa
Aidan Epstein
Alexander Partin
Alexander Batyuk
Andria Rodrigues
Andy DeGiovanni
Arvind Ramanathan
Austin Clyde
Babak Andi
Ben Brown
Bobbie-Jo Webb-Robertson
Brooke Harmon
Carlos Gamboa
Carlos Simmerling
Chris Mungall
Chris Ellis
Chris Stanley
Connor Cooper
Cornelius Gati
Dan Jacobson
Dan Faissol
Derek Jones
Ed Lau
Elijah Hoffman
Emily Dietrich
Fangqiang Zhu
Felice Lightstone
Garry Buchko
Gyorgy Babnigg
Henrique Pereira
Hubertus Van Dam
Hugh O'Neill
Hyunseung Yoo
Ian Foster
Irimpan Mathews
Jason Mcdermott
Jerry Parks
Jha Shantenu
Jim Brase
Joe Schoeniger
Jonathan Allen
Joshua Ladau
Jurgen Schmidt
Justin Reese
Katrina Waters
Kelly Williams
Kenneth Sale
Kerstin Kleese Van Dam
Kevin Mcloughlin
Kris Kulp
Li Tan
Magda Franco
Marisa Torres
Mark Steven Hunter
Marti Head
Matt Coleman
Michael Kent
Mitch Doktycz
Naoki Horikoshi
Neeraj Kumar
Nick Fischer
Oscar Negrete
Paul Adams
Quan Van Vuong
Richard Keith
Rick Stevens
Robert Netzor
Ryan Chard
Ryszard Michalczyk
Sam Chen
Sean McCorkle
Sean McSweeney
Sergio Wong
Simone Raugei
Sindhu Bhowmik
Soichi Wakatsuki
Srinivas Iyer
Stephan Irle
Stephanie Galanie
Stewart He
Tom Brettin
Tom Desautels
Tony Ferreira
Uma Ganapathy
Vilmos Kertesz
Yihui (Ray) Ren
Yue Yang

National Virtual Biotechnology Laboratory

Research was supported by the DOE Office of Science through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding provided by the Coronavirus CARES Act.

https://science.osti.gov/nvbl

Epidemiological Modeling

Manufacturing

Molecular Design for Therapeutics

COVID-19 Testing R&D

Viral Fate & Transport

High Performance Computing Simulation on Demand

Goal: Leverage the world-leading capabilities of the Department of Energy National Labs…

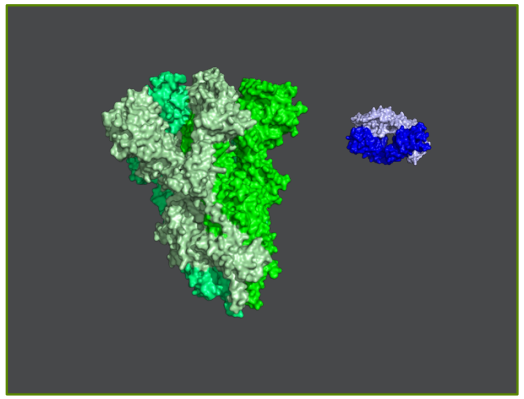
Chemical, biological, and analytical sciences

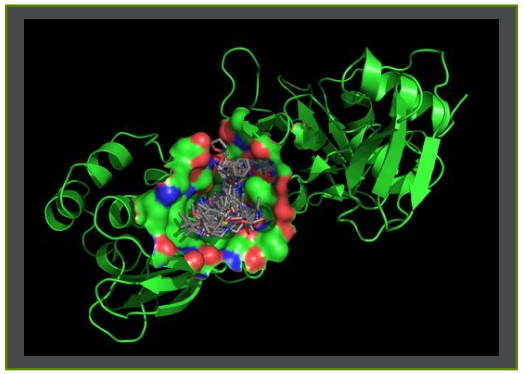
High performance computing

Light and neutron sources

COVID-19 Results

Computational and experimental design platforms





Starting points:

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

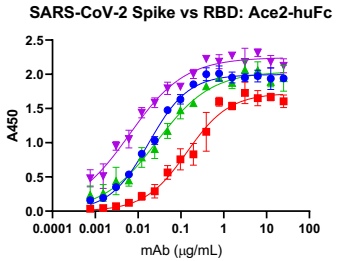
Structure and affinity

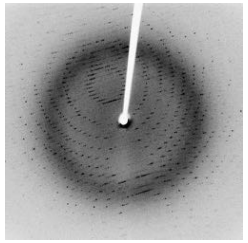
Physical, chemical properties

Safety and PK

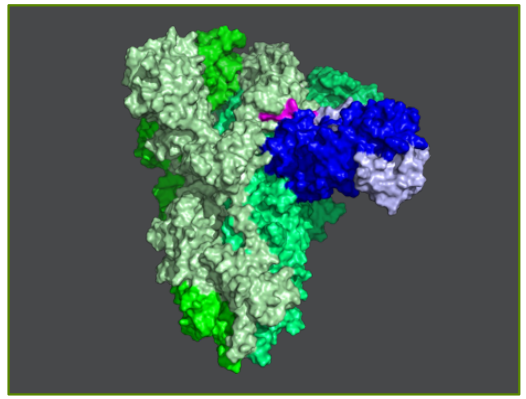
Physics-based and AI/ML models to assess design fitness

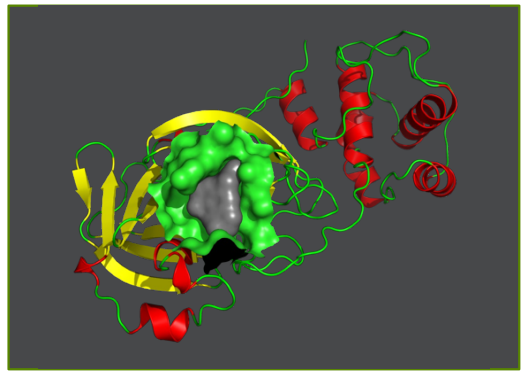
Structure- and sequence-driven generative models for iterative design





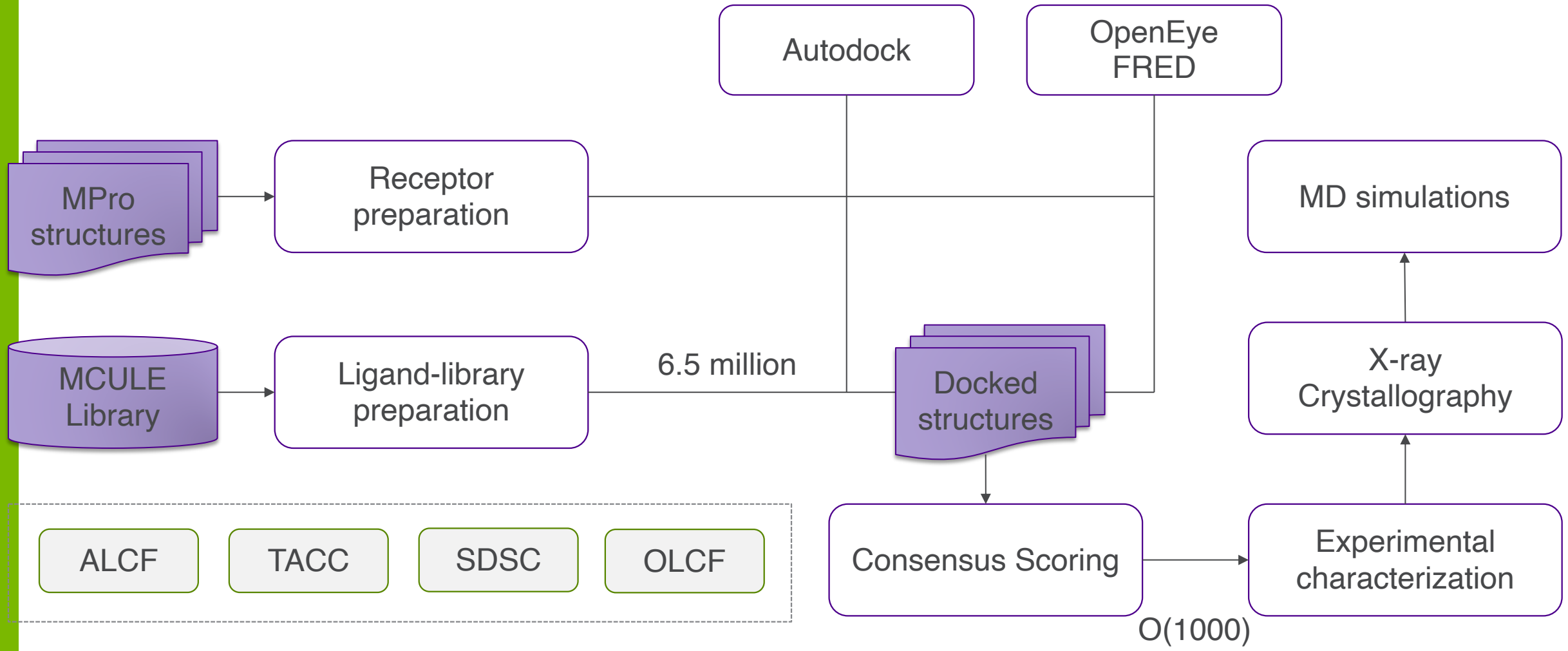
Platform capability build funded over time through DOE, LDRD, DARPA, DOD, and other funding sources



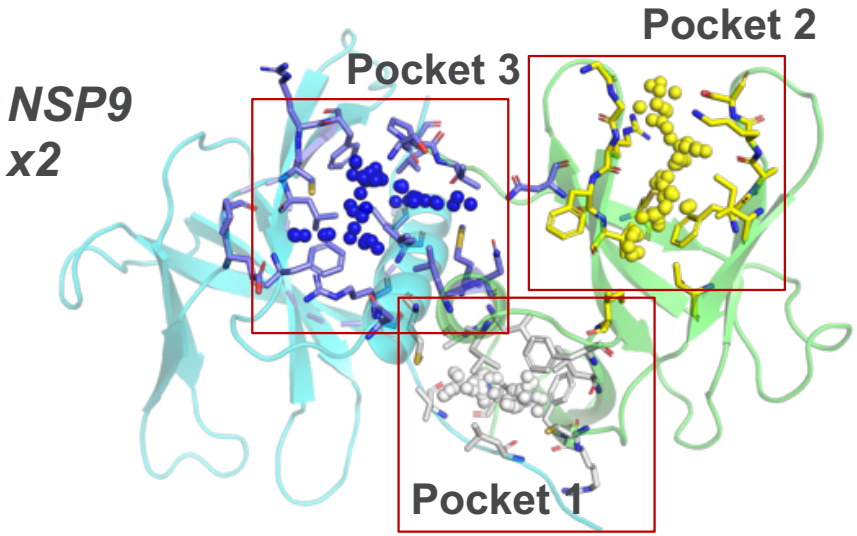


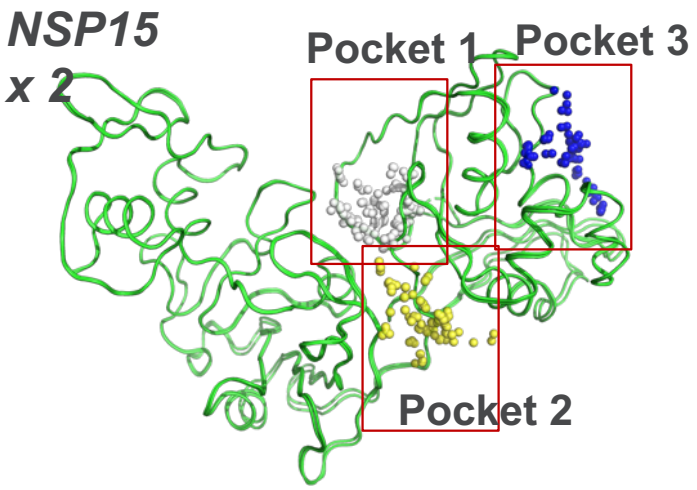
Outputs:

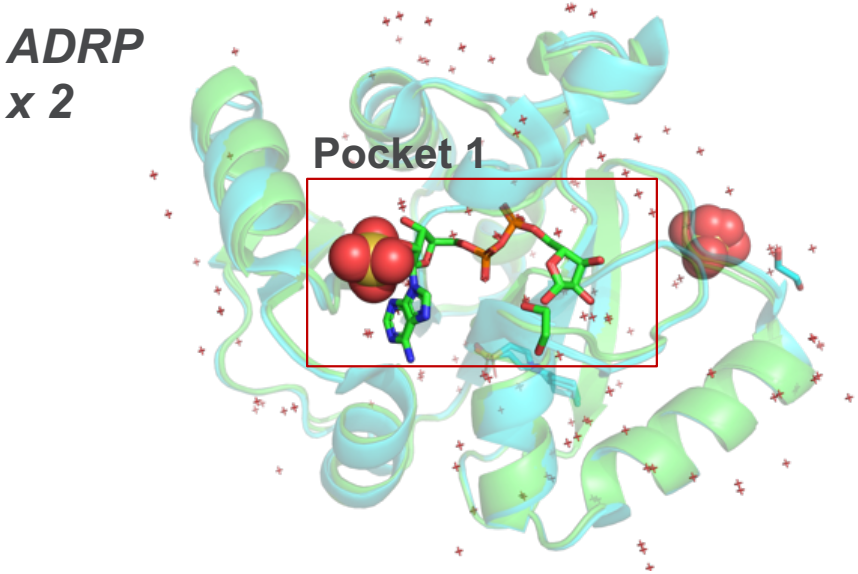
- Designs with probability of:
  - Desired activity
  - Desired biological effect
  - Good physical and safety parameters

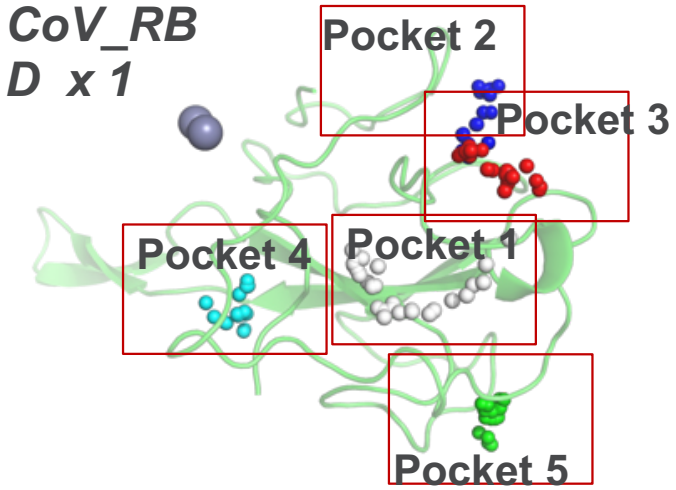

Autodock
OpenEye
FRED
MPro
structures
Receptor
preparation
MD simulations
MCULE
Library
Ligand-library
preparation
6.5 million
Docked
structures
X-ray
Crystallography
ALCF
TACC
SDSC
OLCF
Consensus Scoring
Experimental
characterization
O(1000)

Targets and binding sites







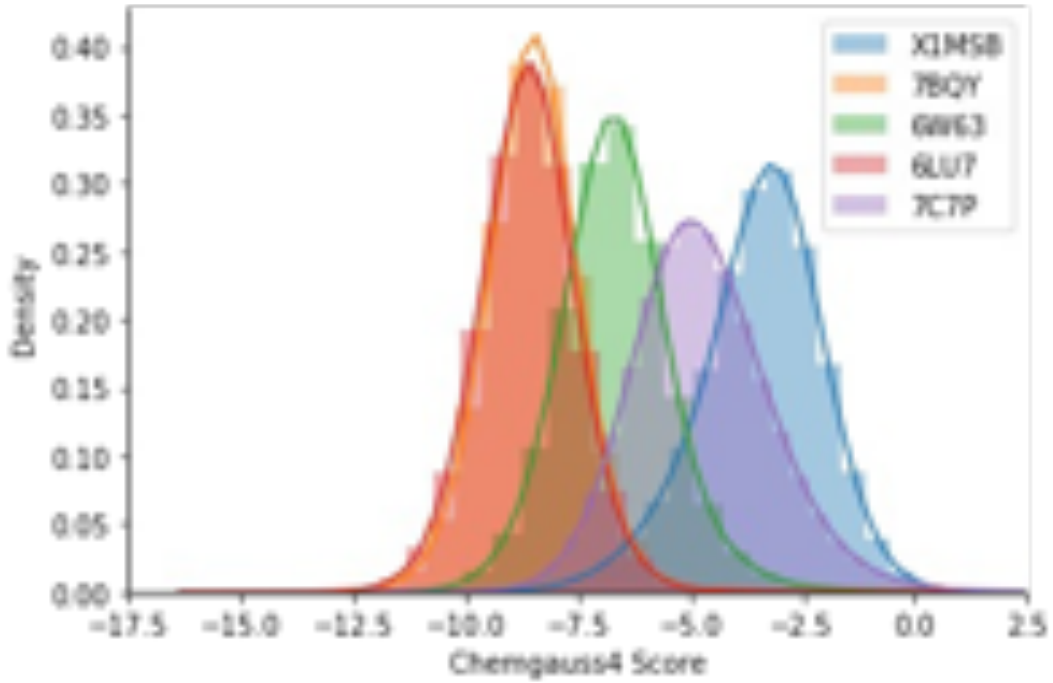


Automatic pocket detection

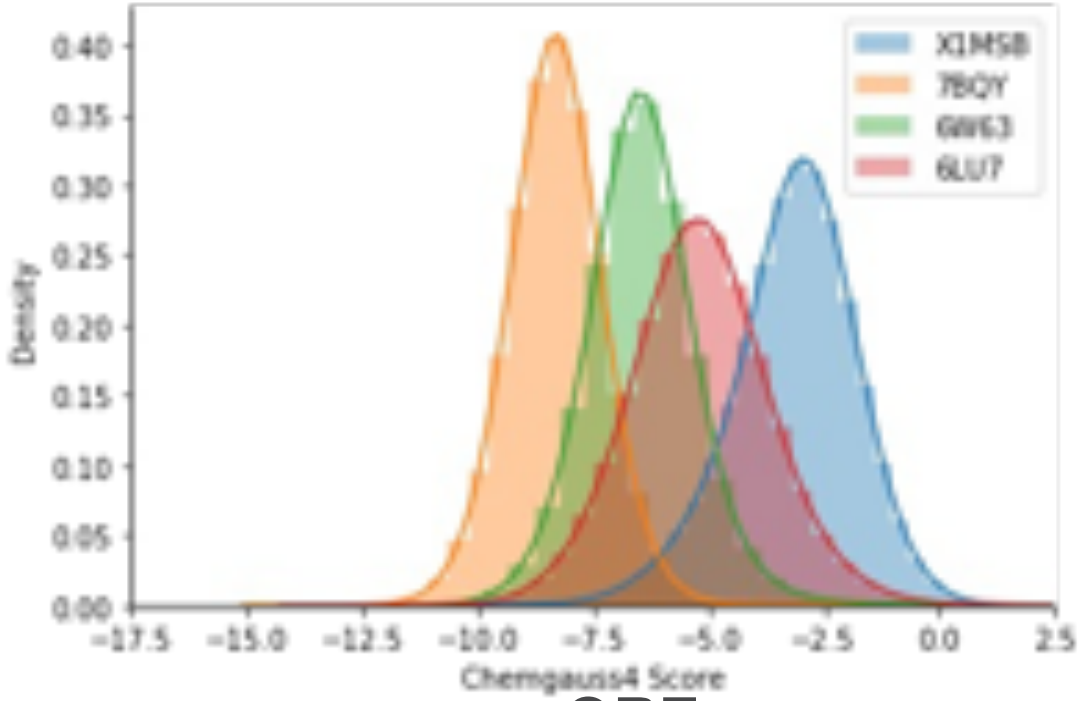
Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10,** 168 (2009). https://doi.org/10.1186/1471-2105-10-168

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


X1M5B
7BQY
6W63
6LU7
7C7P
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−17.5
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
Chemgauss4 Score

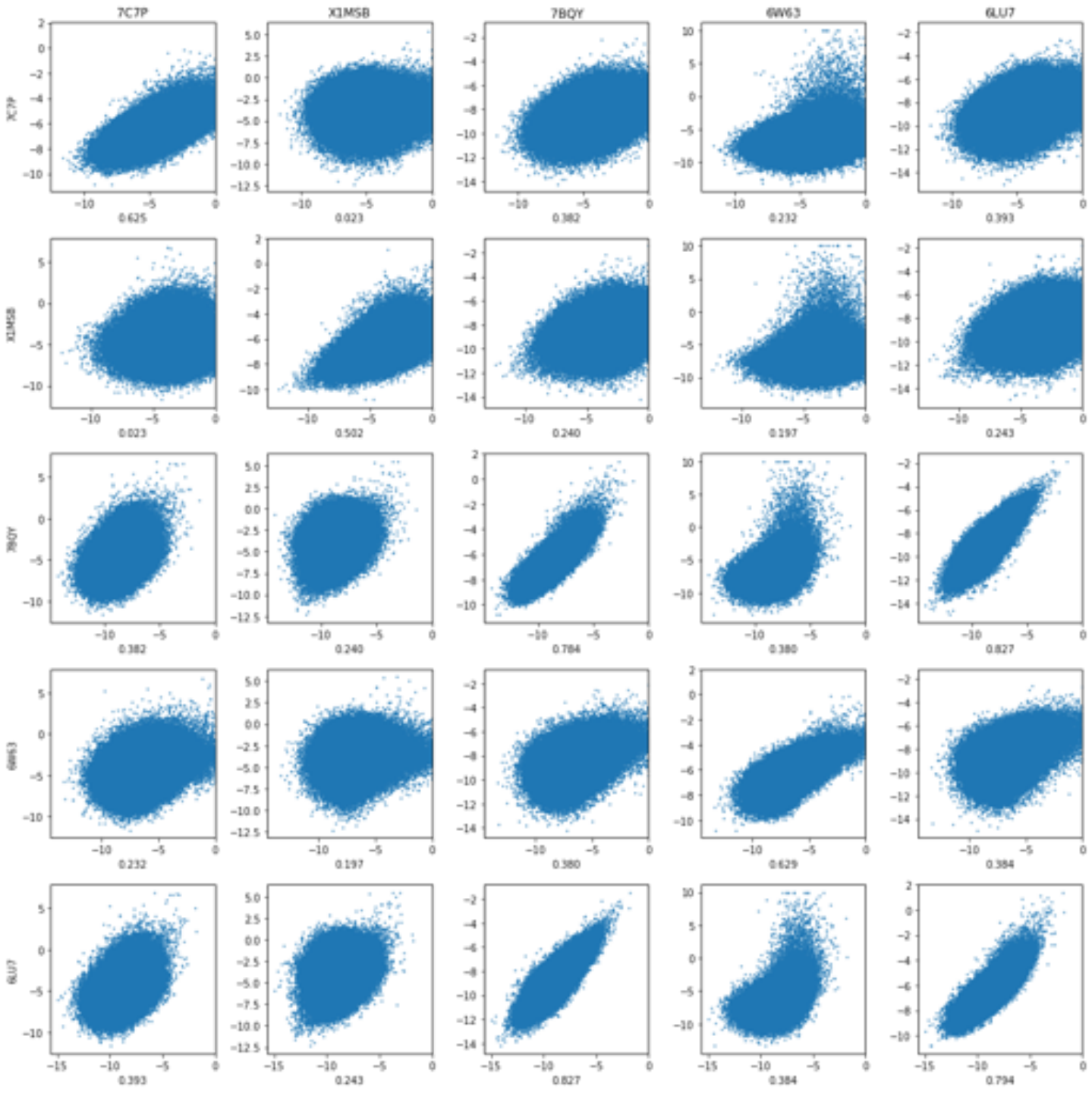
ORD


X1M5B
7BQY
6W63
6LU7
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−17.5
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
Chemgauss4 Score

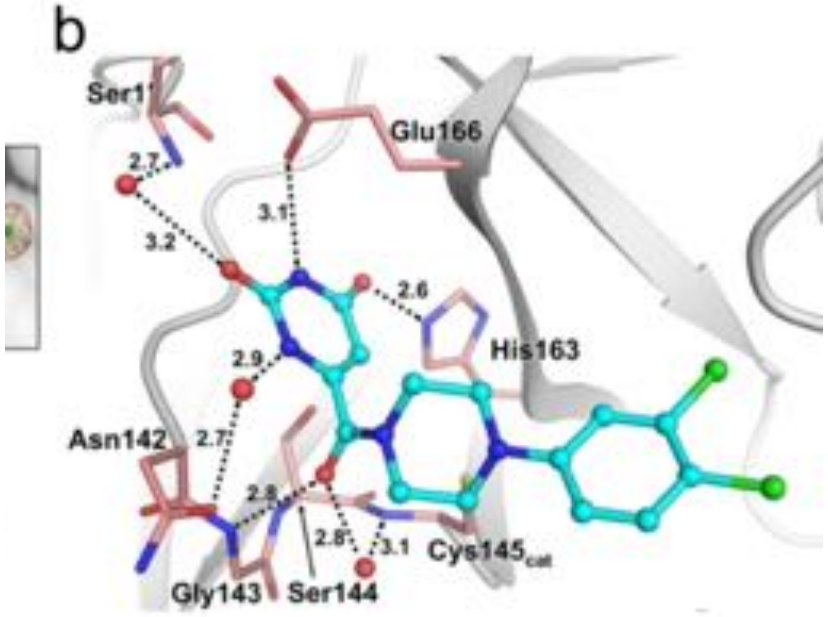
ORZ

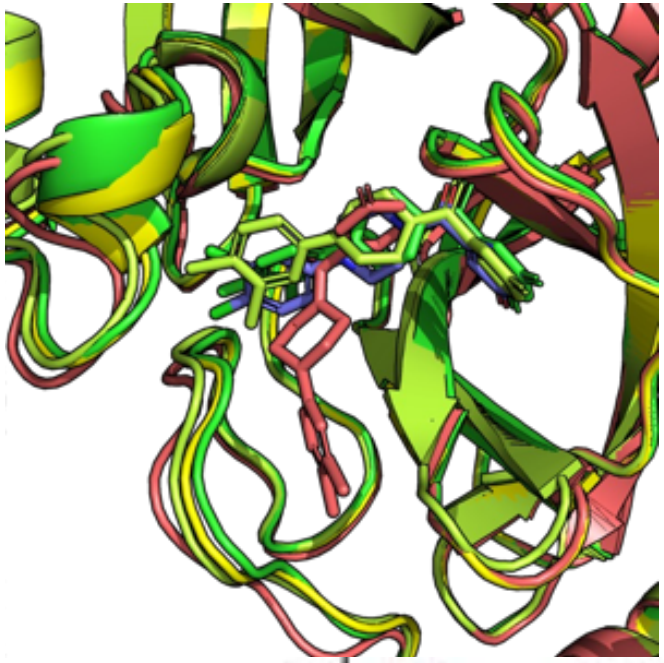
Diagonal is mean score

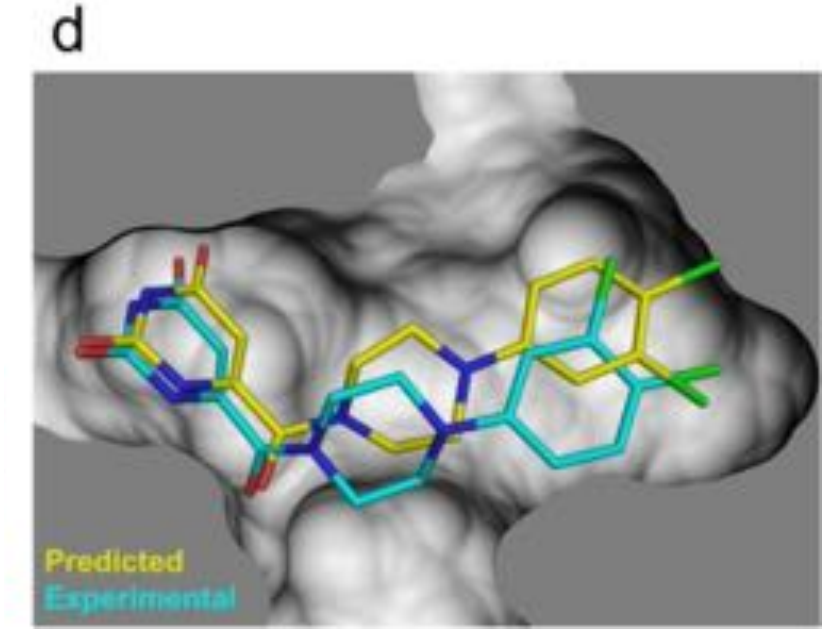
Lower number is spearman rank correlation

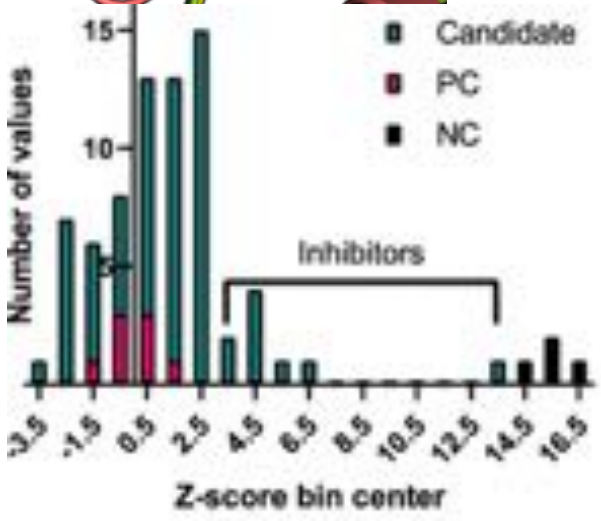


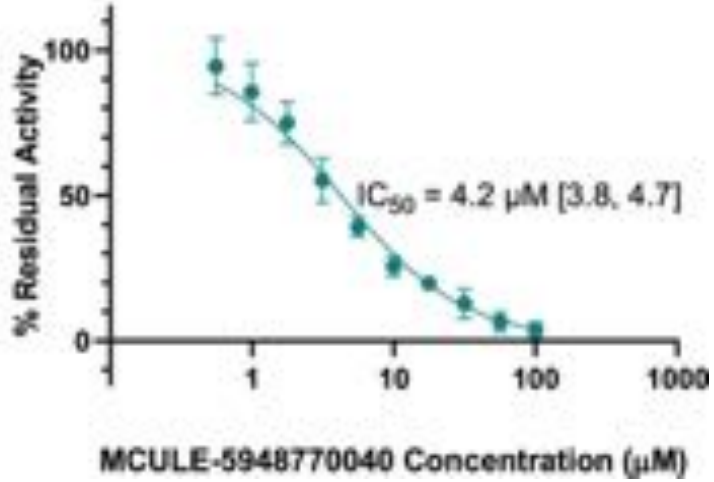
Four docking models…











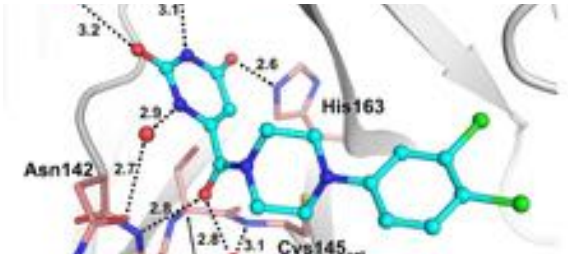
Clyde, Austin, Stephanie Galanie, Daniel W. Kneller, Heng Ma, Yadu Babuji, Ben Blaiszik, Alexander Brace et al. "High-throughput virtual screening and validation of a sars-cov-2 main protease noncovalent inhibitor." *Journal of chemical information and modeling* 62, no. 1 (2021): 116-128.

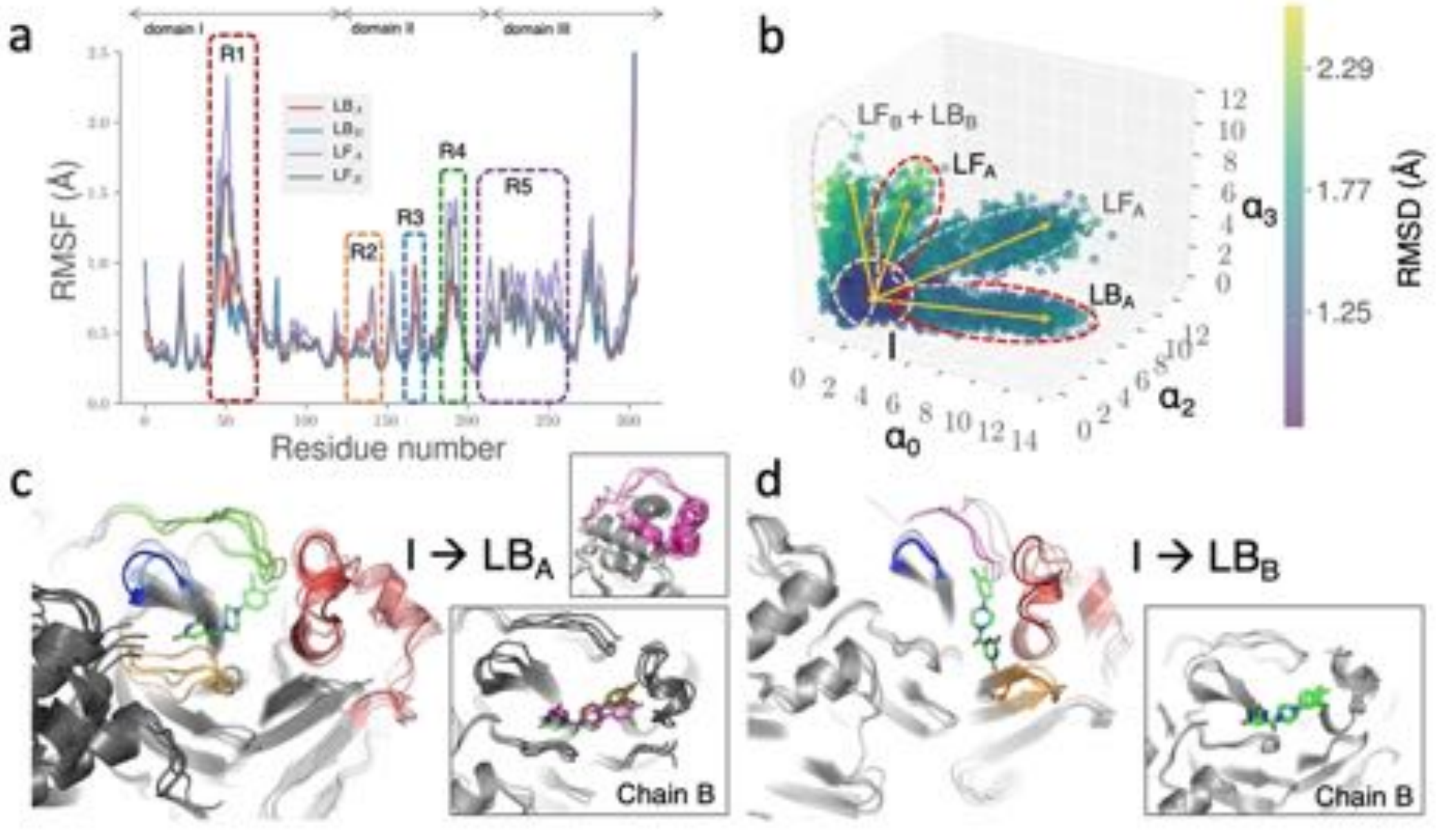
Cool facts

- Summit
- Theta
- Frontera
- …

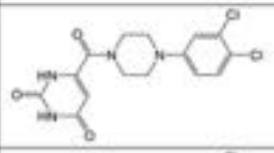
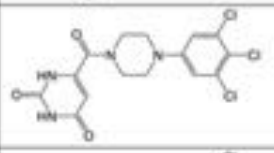
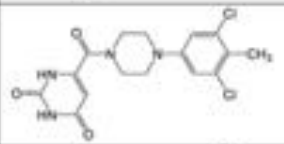
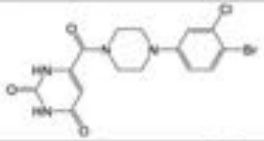
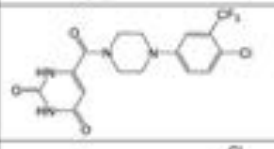
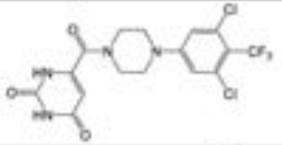
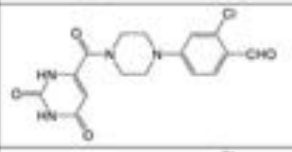
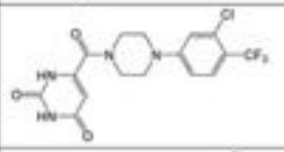
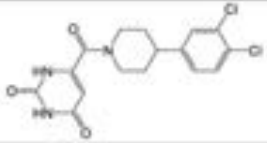
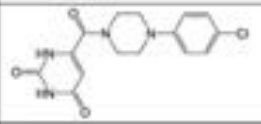
ML models screened over a billon compounds

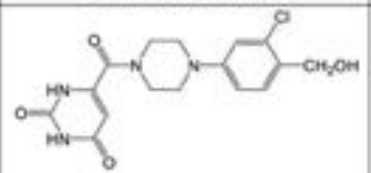
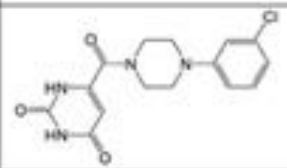
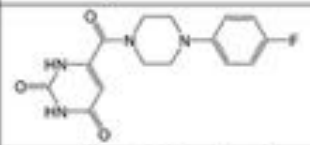
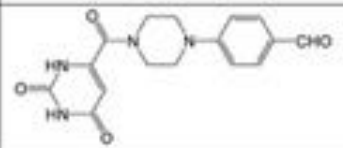
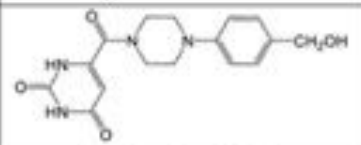
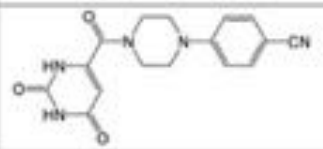
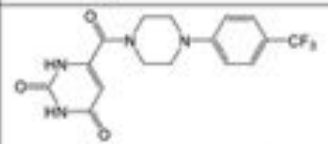
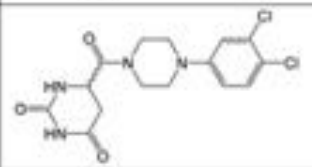
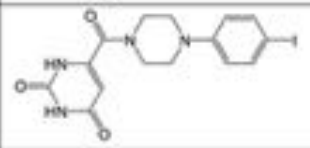
- 336,000,000 docking scores computed
  - ~67 billion conformer-protein optimization
  - Retained all computed structural data for community sharing
- >50 hits on a full virus assay from computational pipeline
- 1 crystalized, assayed, interesting compound targeting the main protease. (forthcoming publication)

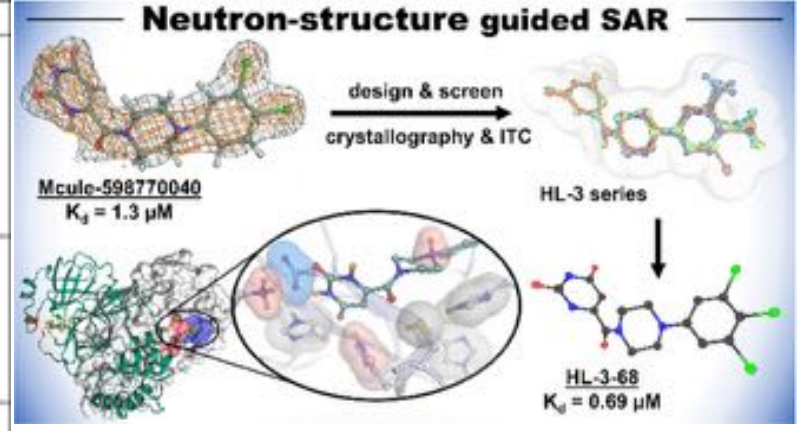



a
domain I
domain II
domain III
R1
R2
R3
R4
R5
RMSF (Å)
Residue number
b
LF_B + LB_B
LF_A
LF_A
LB_A
I
α0
α2
α3
RMSD (Å)
2.29
1.77
1.25
c
I → LB_A
Chain B
d
I → LB_B
Chain B

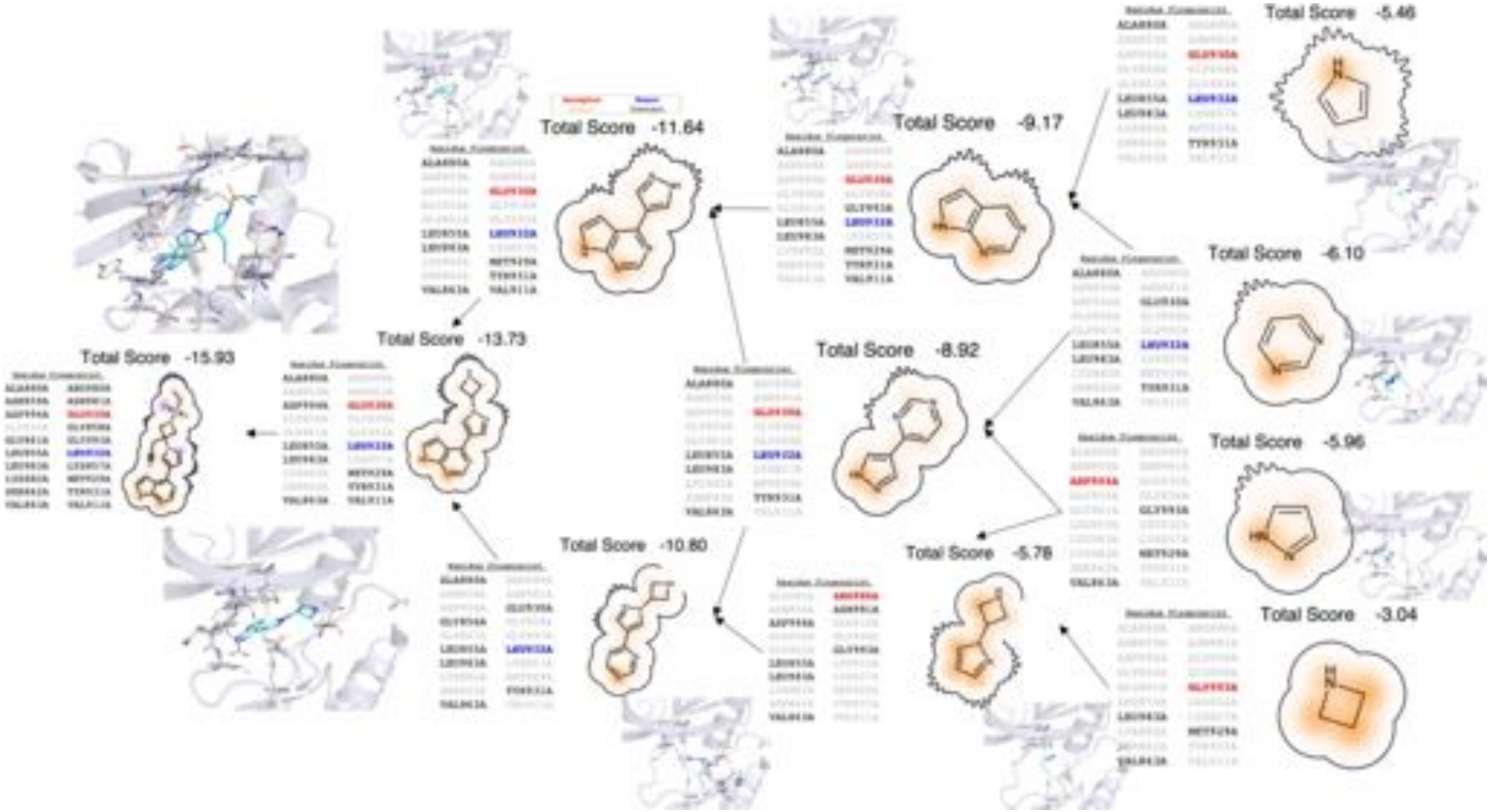
SAR

Compound	Chemical Structure	IC50, μM	PDB ID
Compound 1 (Mcule-5948770040)		0.68 [0.48, 0.97][b,40]	7LTJ
HL-3-68		0.29 [0.22, 0.40]	7RLS
Mcule-CSR-494190-S1		0.29 [0.19, 0.43]	7RM2
HL-3-78		0.61 [0.37, 0.96]	7RMB
HL-3-52		1.4 [0.80, 2.3]	7RME
HL-3-87		1.4 [0.9, 2.2]	N/D[c]
HL-3-70		6.2 [4.8, 8.0]	7RMT
HL-3-63		6.4 [4.3, 9.5]	7RMZ
HL-3-69		8.8 [6.3, 13]	7RN4
HL-3-45		> 20	7RNH

Compound	Chemical Structure	IC50, μM	PDB ID
HL-3-71		> 20	7RNK
HL-3-46		> 20	N/D
HL-3-43		> 20	N/D
HL-3-44		> 20	N/D
HL-3-49		> 20	N/D
HL-3-62		> 20	N/D
HL-3-50		> 20	N/D
HL-3-51 HL-3-53		> 20	N/D
HL-3-65		> 20	N/D



Kneller, Daniel W., Hui Li, Stephanie Galanie, Gwyndalyn Phillips, Audrey Labbé, Kevin L. Weiss, Qiu Zhang et al. "Structural, electronic, and electrostatic determinants for inhibitor binding to subsites S1 and S2 in SARS-CoV-2 main protease." *Journal of medicinal chemistry* 64, no. 23 (2021): 17366-17383.



Scaffold-Induced Molecular Subgraphs (SIMSG): Effective Graph Sampling Methods for High-Throughput Computational Drug Discovery. Clyde, A., Shah, Ashka, Zvyagin, M., Ramanathan, A., Stevens, R., Virtual 13 November 2020.

Basic Meet Operators:

Predecessor

Scaffold

Basic Join Operators (generative)

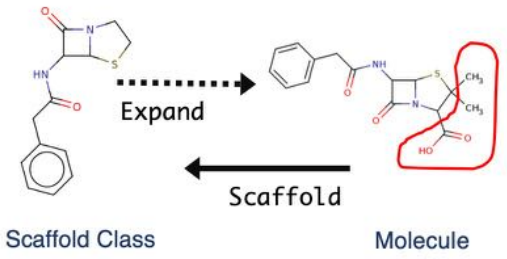
$\texttt{Successor}_\Phi$

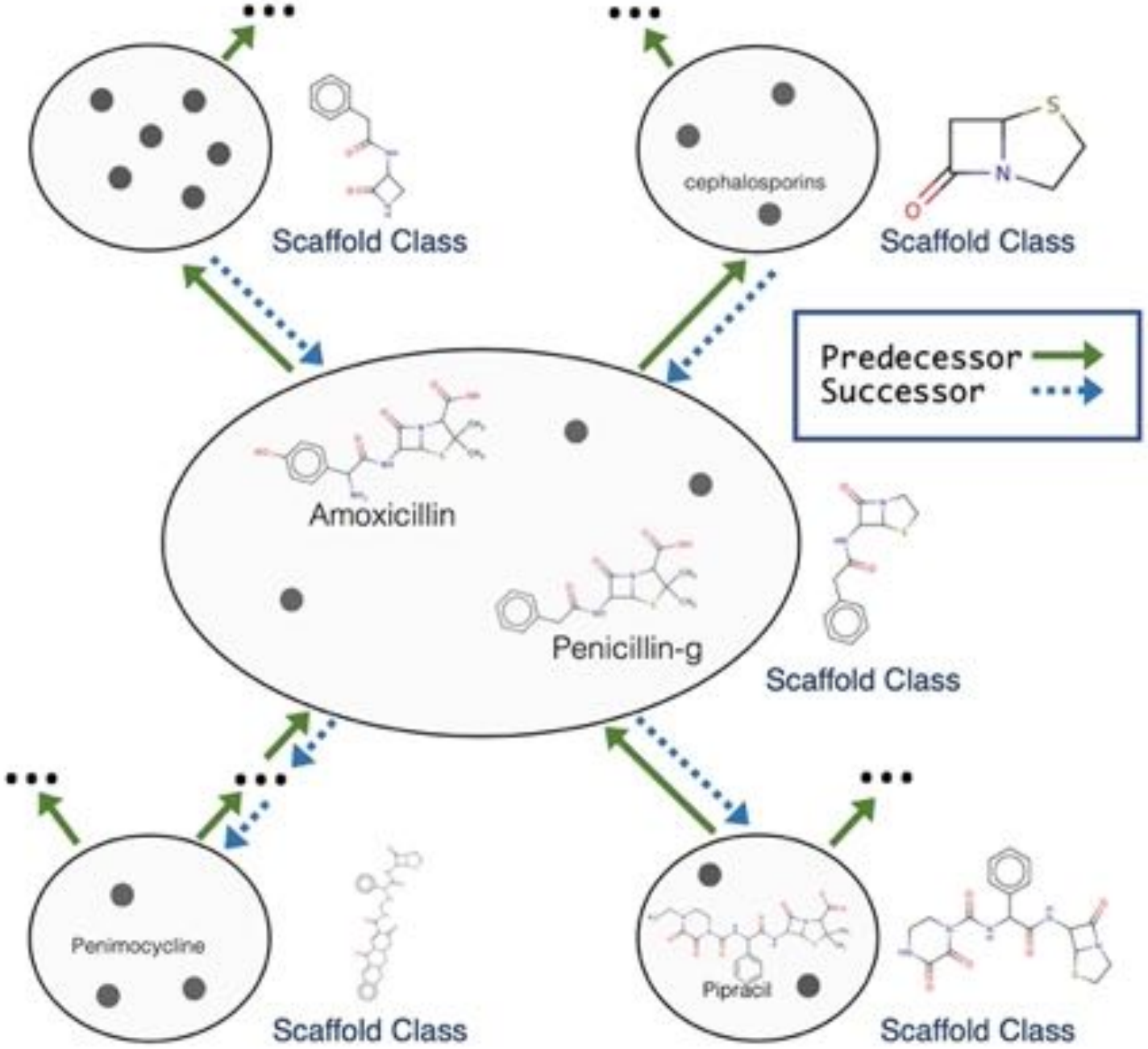
$\texttt{Expand}_\Phi$

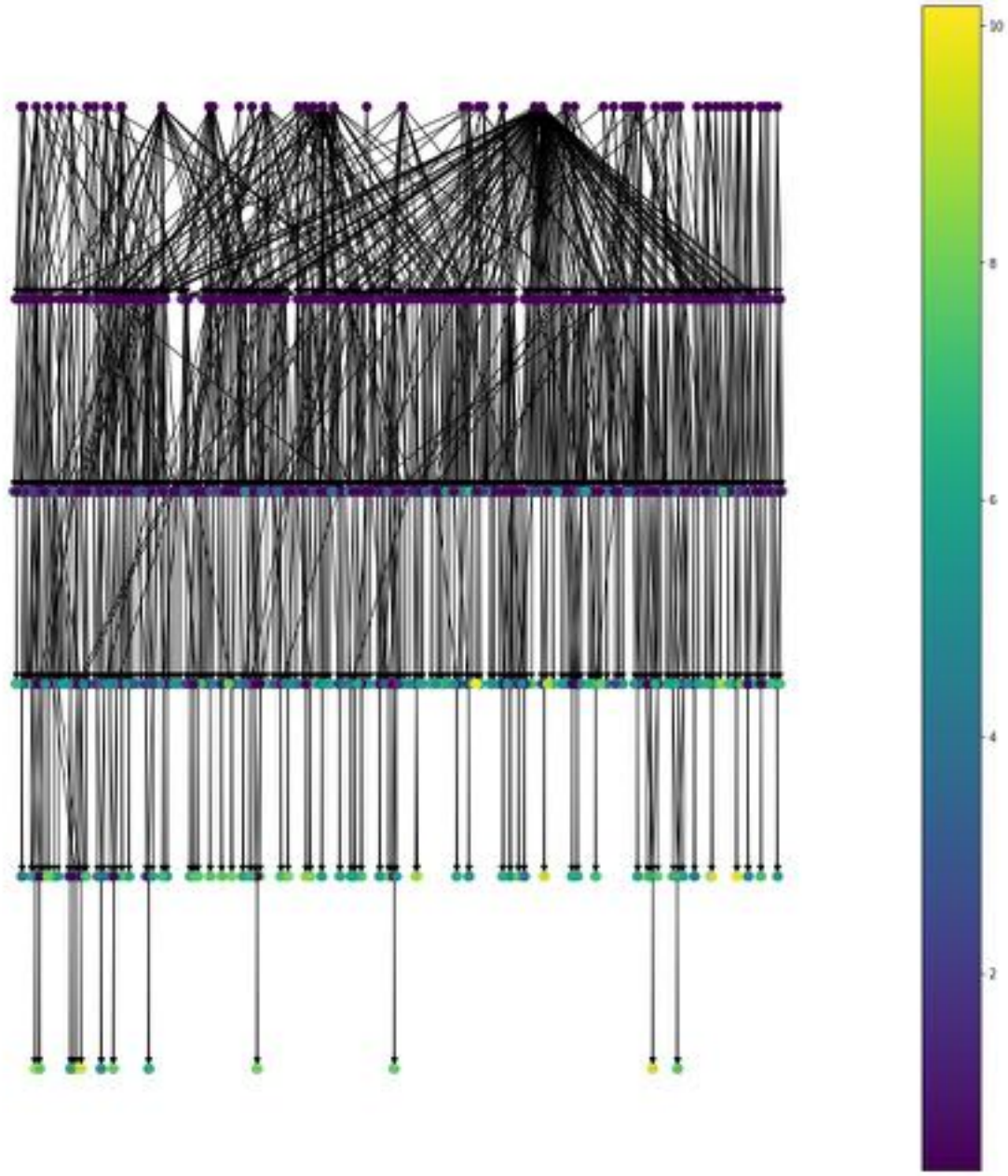
$$\texttt{UpperCone}_\Phi(S) = \{A : S \prec A\}$$

or $\texttt{LowerCone}(S) = \{B : B \prec S\}$

$$\hat{H} = \bigcap_{i \in I^*} \texttt{UpperCone}_\Phi(S_i^h)$$



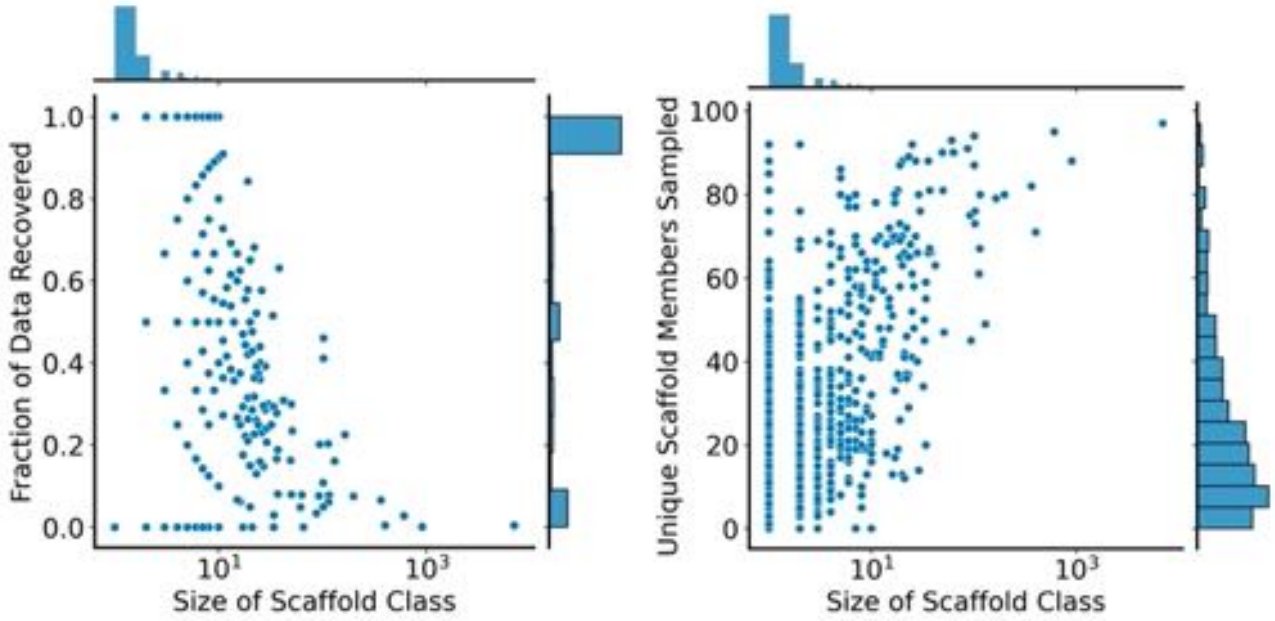


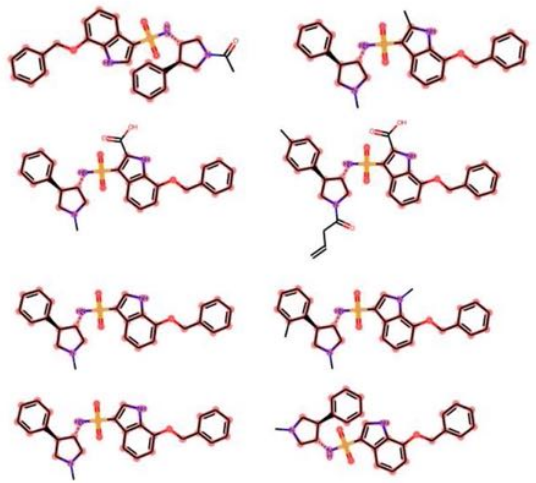

10
8
6
4
2

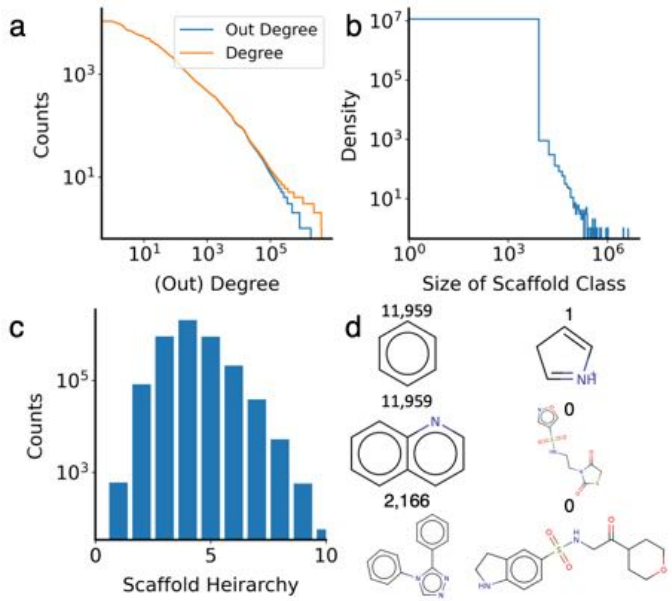
Training a Transformer for Operators

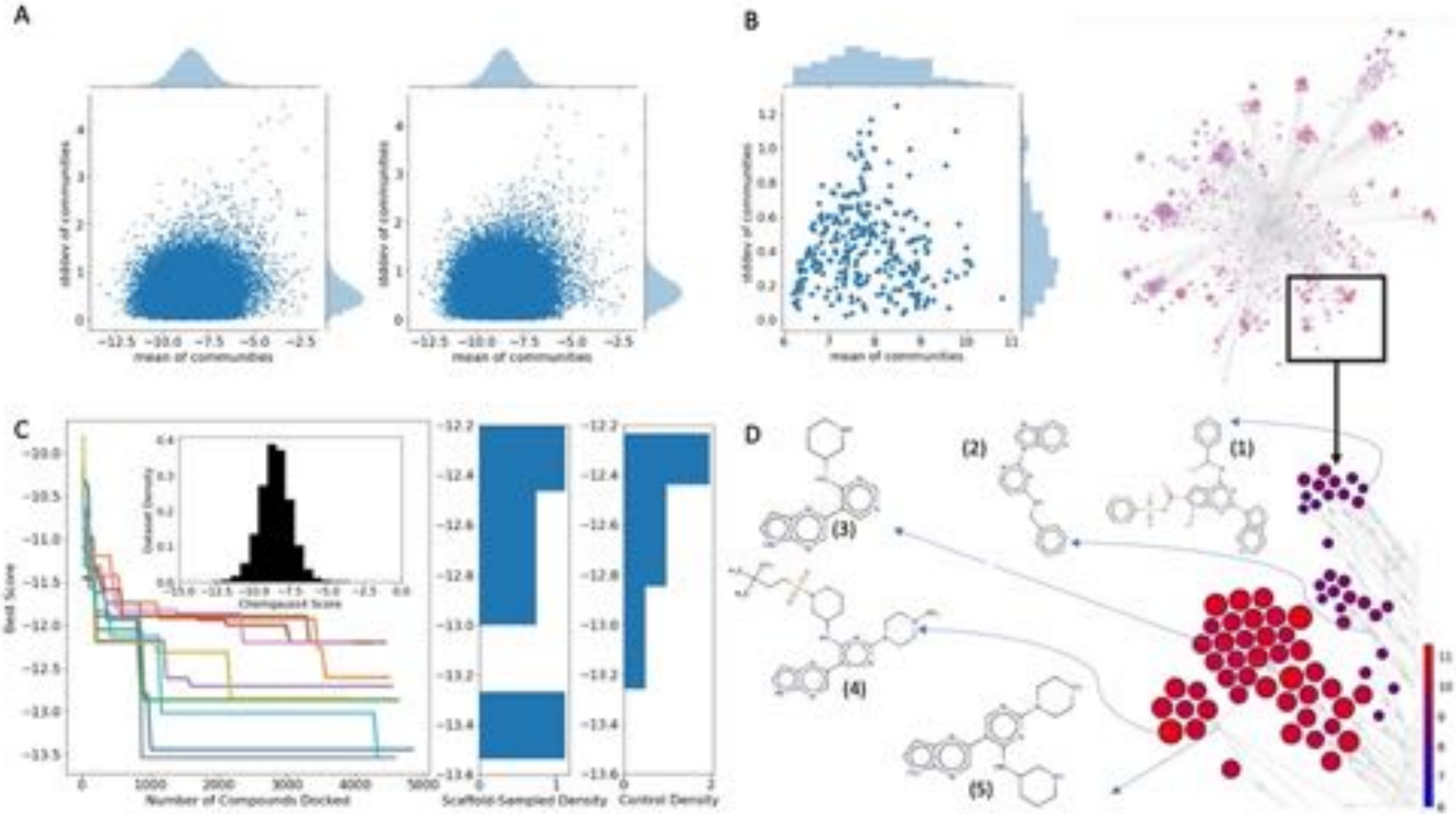
Scaffold	Class Size (Data)	Unique Sampled	Overlap (Recall)
c1ccc(COc2ccccc2)cc1	373,939	168,261	4,146 (1.1%)
O=S(=O)(c1ccccc1)N1CCCCCC1	88,608	145,904	20,097 (22.7%)
O=S(=O)(NCCc1ccccc1)c1ccccc1	911,360	176,539	23,715 (2.6%)
c1ccncc1	818,230	183,838	23,999 (3.0%)
O=S(=O)(NS(=O)(=O)c1cccnc1)c1ccccc1	203,891	173,599	20,331 (10.0%)

Model	SMILES Validity	Type Accuracy	Correctness Accuracy
$\texttt{Successor}_\Phi$	98.9%	98.9%	97.9%
$\texttt{Predecessor}$	99.8%	99.8%	94.0%
$\texttt{Expand}_\Phi$	98.6%	-	96.9%



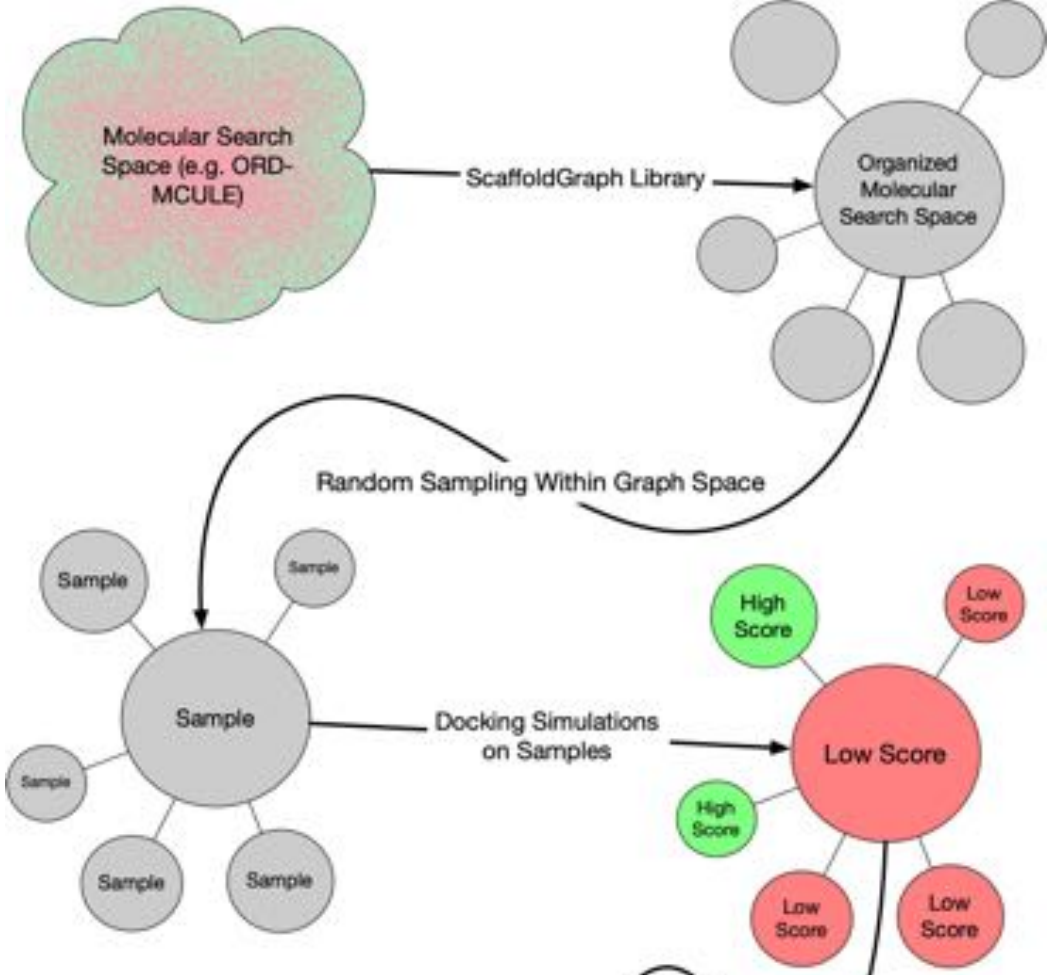


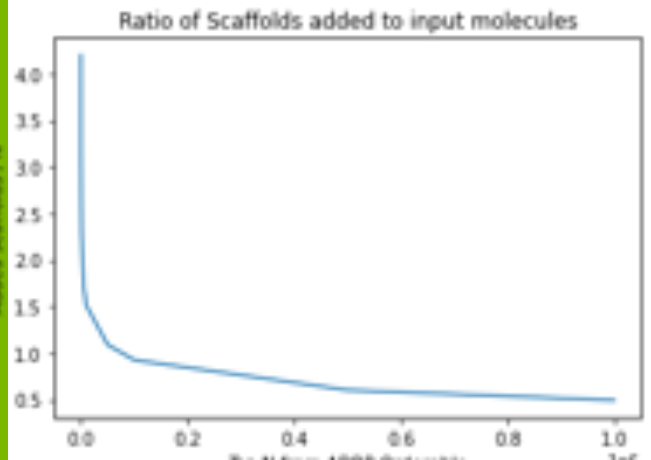


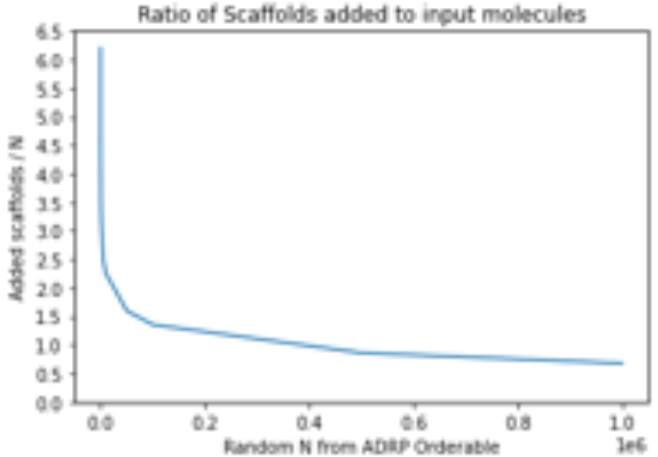

A
B
C
D
mean of communities
Best Score
Number of Compounds Docked
Scaffold-Sampled Density
Control Density
(1)
(2)
(3)
(4)
(5)

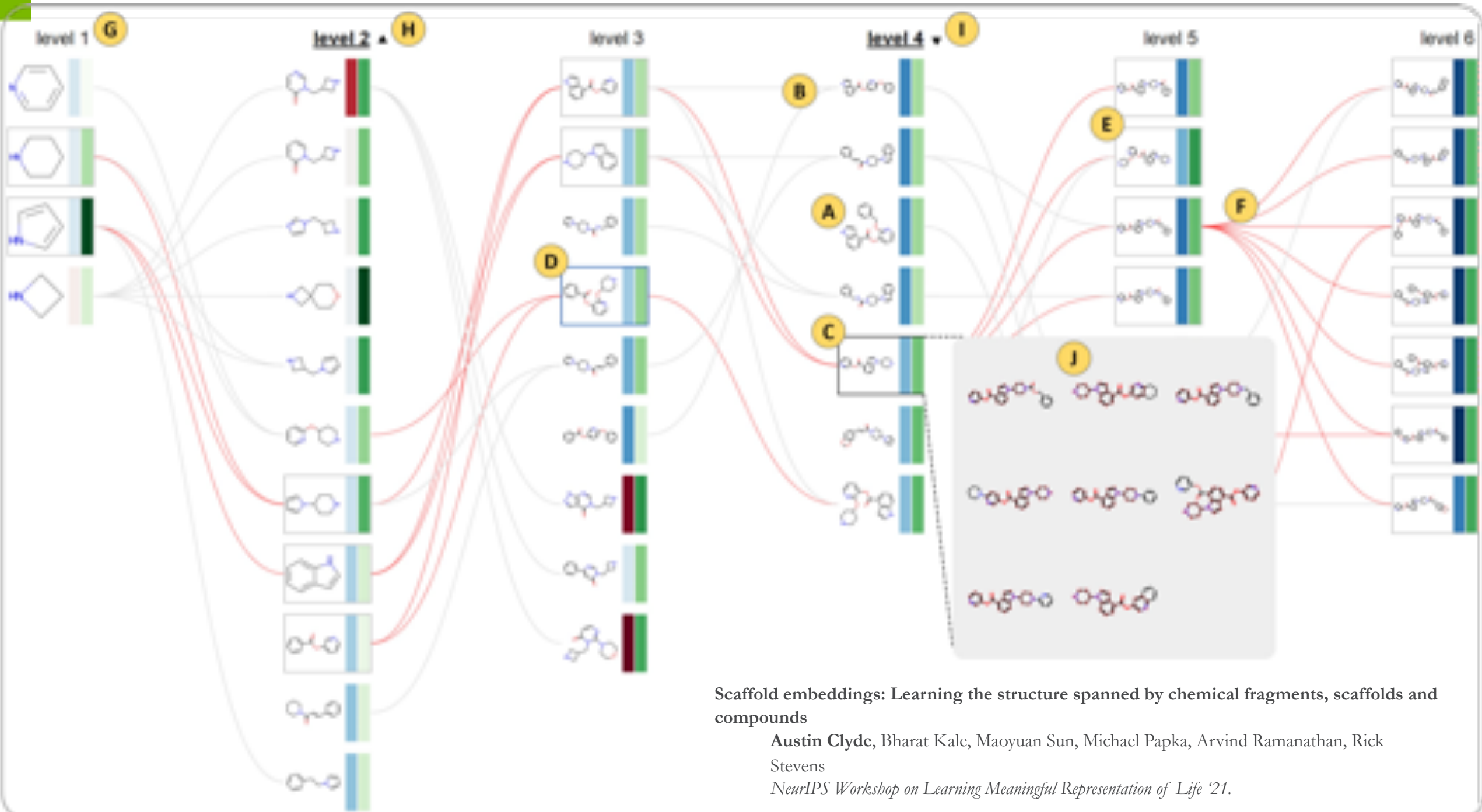
Scaffold Induced Graph Sampling

How can we leverage innate structure on the problem space to transform this problem?









Scaffold embeddings: Learning the structure spanned by chemical fragments, scaffolds and compounds

Austin Clyde, Bharat Kale, Maoyuan Sun, Michael Papka, Arvind Ramanathan, Rick Stevens

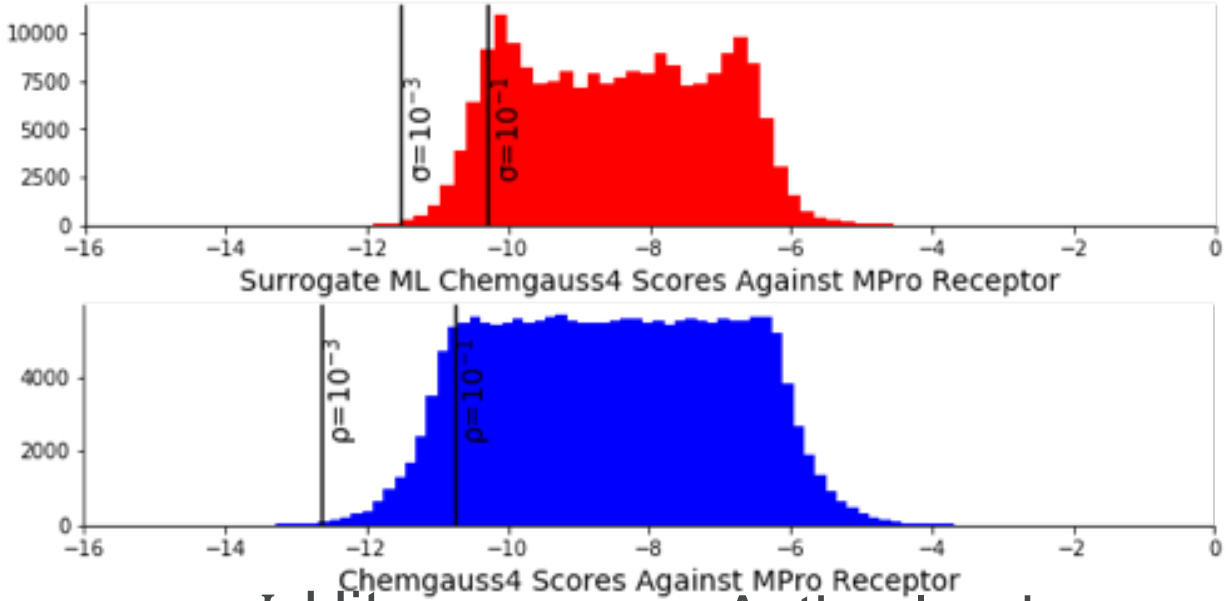
NeurIPS Workshop on Learning Meaningful Representation of Life '21.

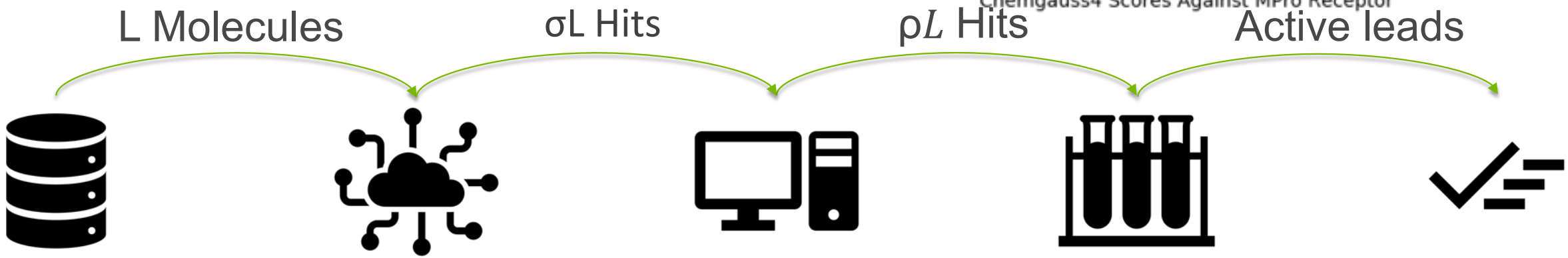
Infrastrcutre

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?



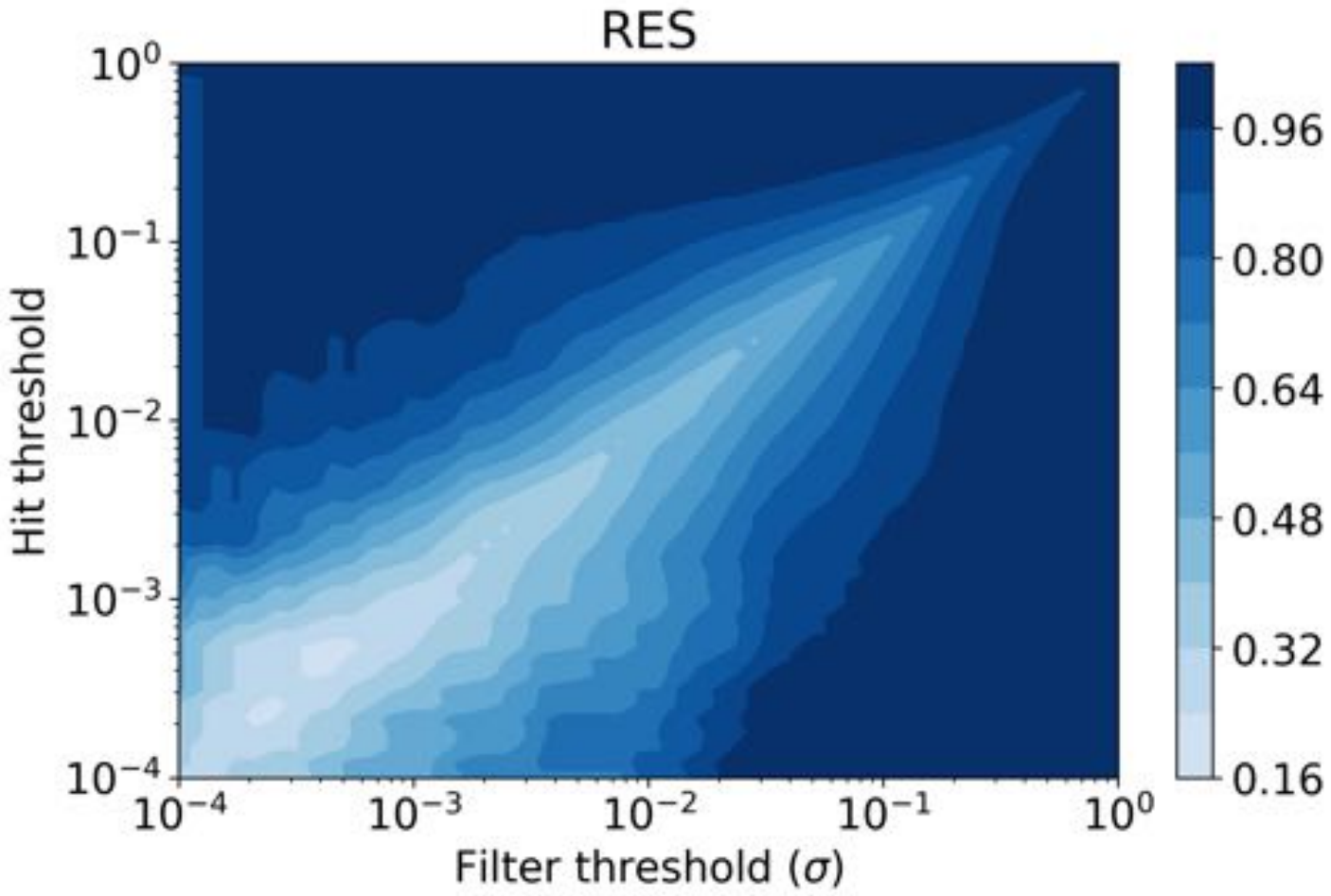


Regression Enrichment Surfaces

Relate model accuracy as a function of σ and ρ

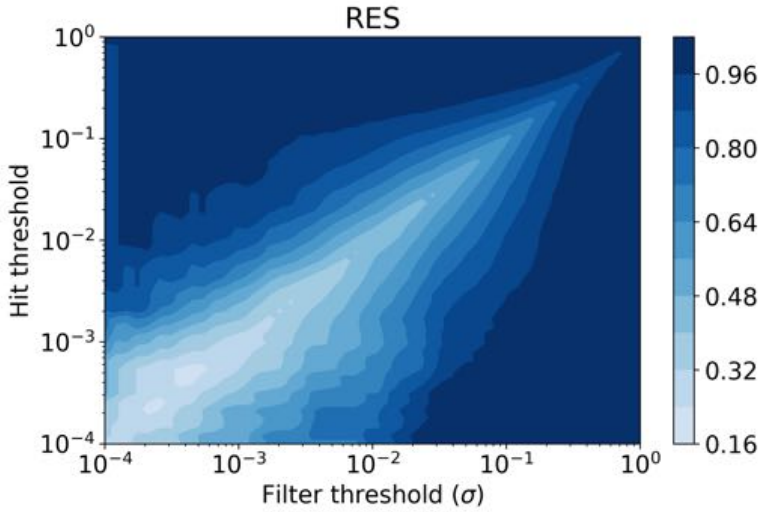
- Rank all compounds from surrogate R^{SPF}
- Rank all compounds form docking, R^{D}
- Let R_x for $0 \leq x \leq 1$ be x highest ranking compounds

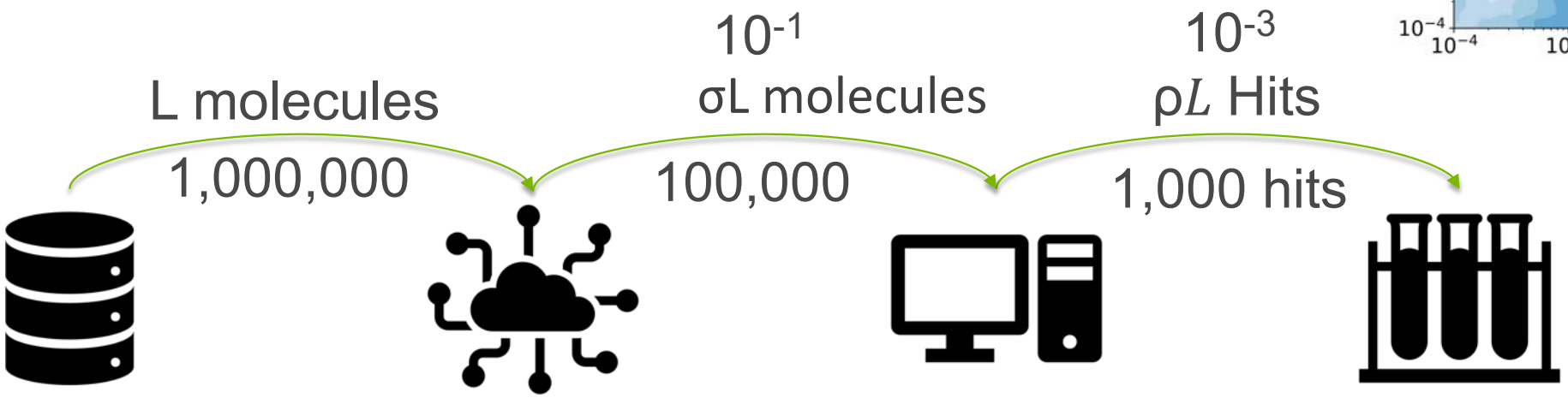
$$\text{Enrichment}(R^{\text{SPF}}, R^{\text{D}}, \sigma, \rho) = \frac{|R_\sigma^{\text{SPF}} \cap R_\rho^{\text{D}}|}{\min(\sigma, \rho)|L|}$$



Workflow analysis

Surrogate Prefilter then Dock (SPFD)





A compound library

ML prefilter

Docking protocol

Assays

Based on the RES plot, (10^{-1},10^{-3}), our surrogate model will not miss anyone

Let T_D = 1.37

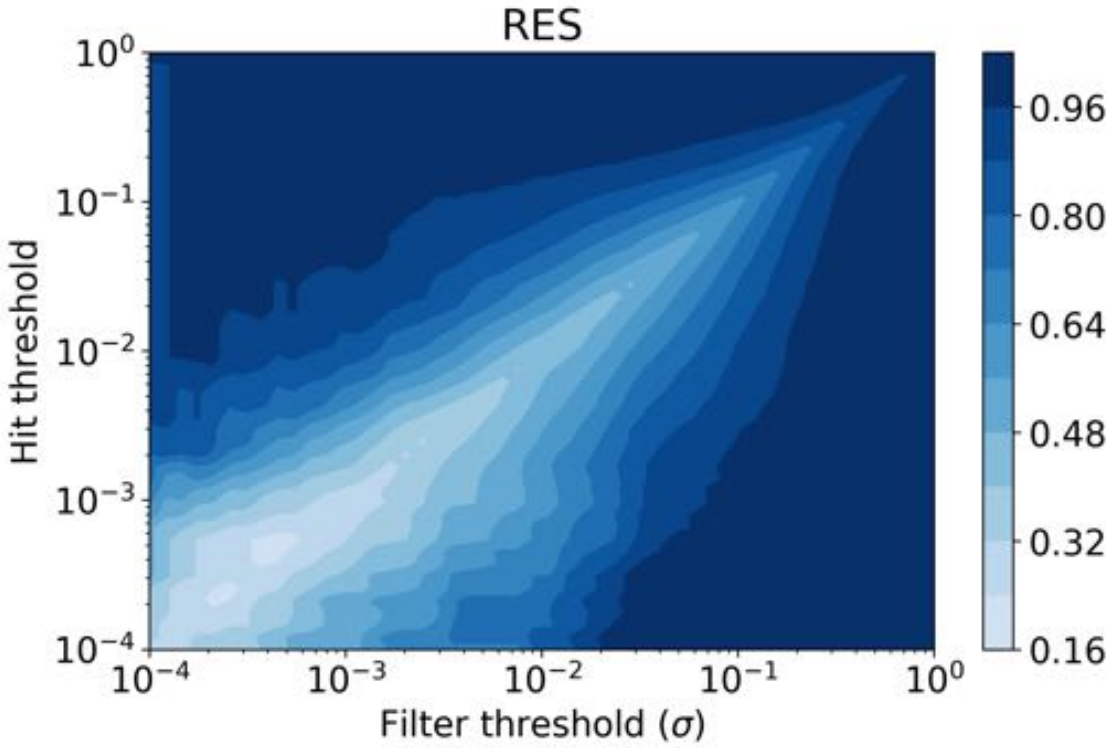
Let T_{SPF} = 250,000

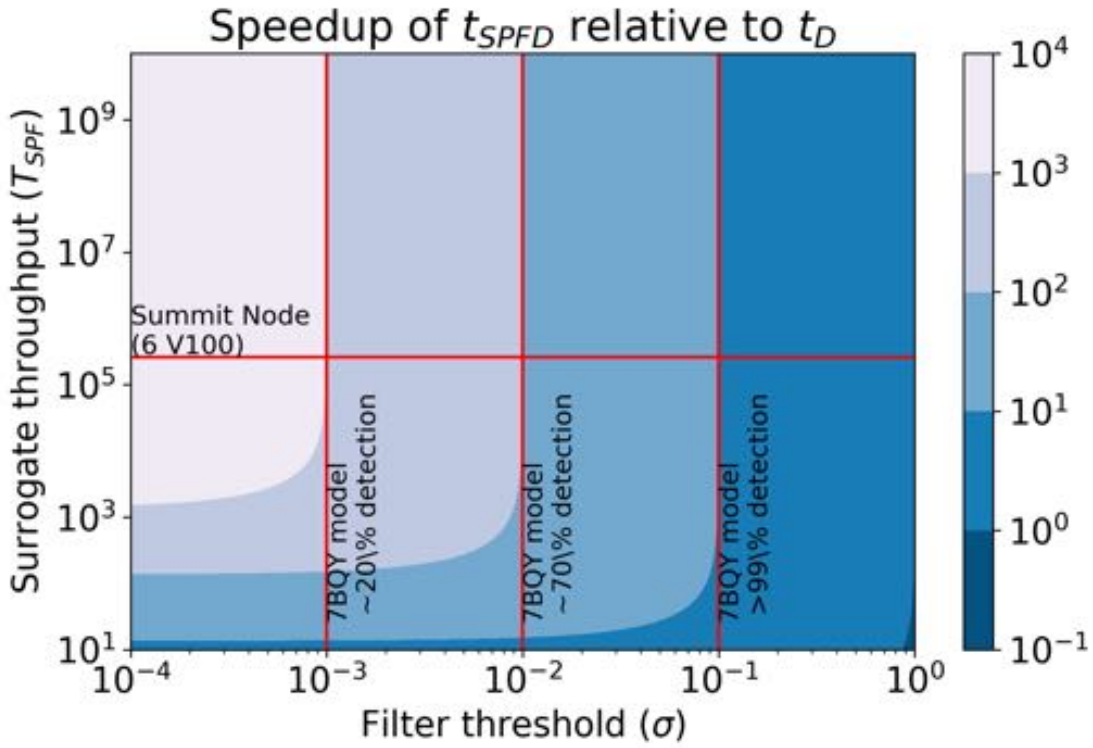
$$Time\ SPFD = \frac{L}{T_{SPFD}} + \frac{\sigma L}{T_D}\frac{1{,}000{,}000}{250{,}000} + \frac{10^{-1}\cdot 1{,}000{,}000}{1.37} = 73\text{k node seconds}$$

$$Time\ D = \frac{\sigma L}{T_D} = \frac{1{,}000{,}000}{1.37} = 730\text{k node seconds}$$

SPFD

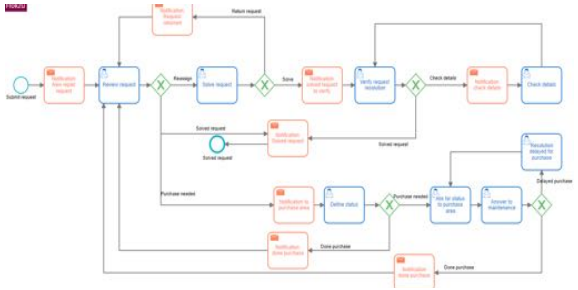
RES + SPFD couples performance, time, and accuracy





Takeaways

Out of box ML models can increase the throughput of your workflow by at least a factor of 10



The key to understanding ML is to first understand the workflow

Surrogate models are a small step towards integrating AI into science, but are relatively safe given we can analyze them inside of workflows we understand