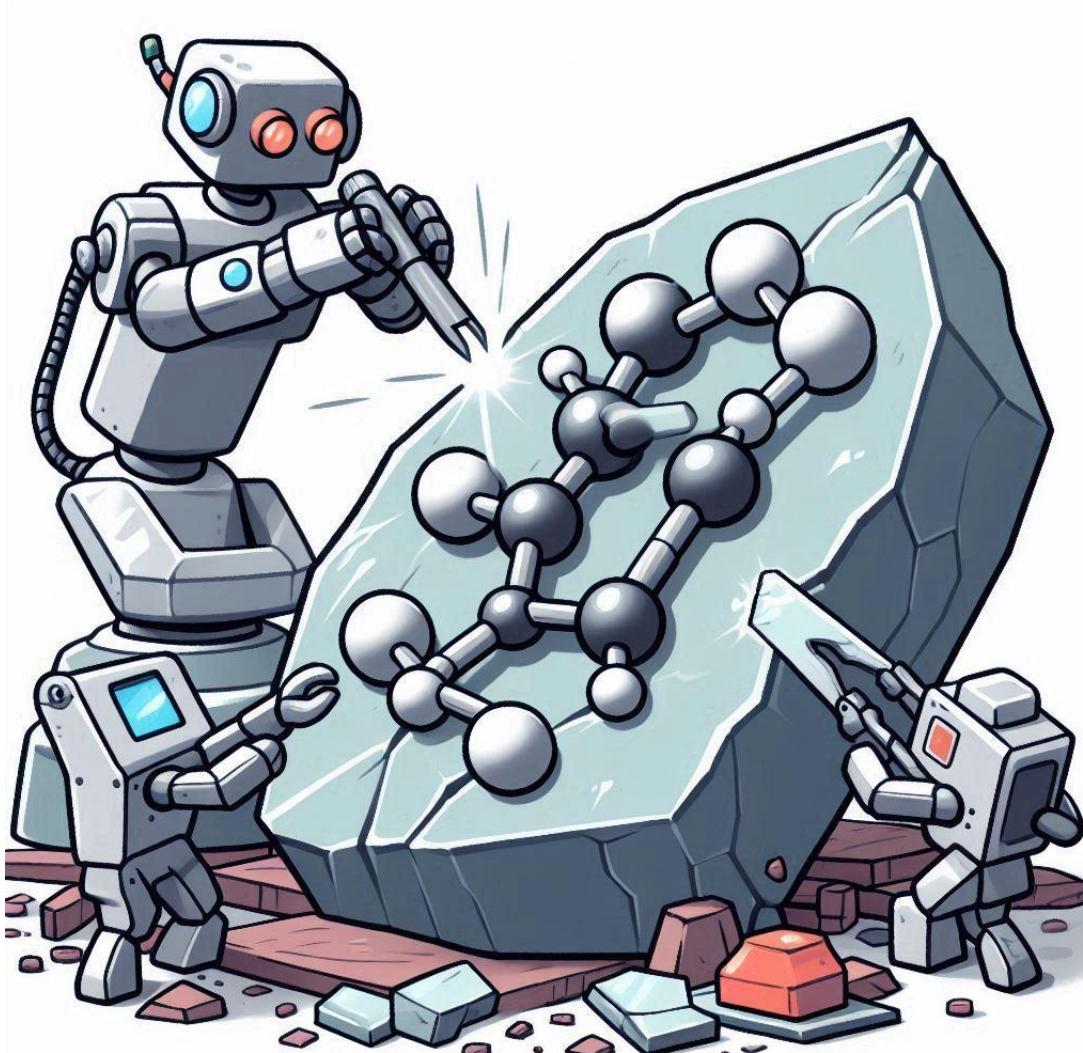


AI in Drug Discovery Workshop

Session 3 – Practical Applications

Molecule Generation



Large Language Models (LLMs) Learn Next Token Probabilities

Given the phrase “The cat is”, what is the next token?

In this case, a token is a word

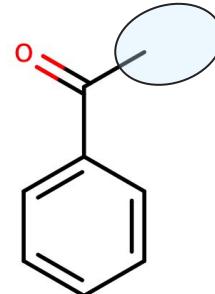
$$P(\text{blue}|\text{The, cat, is}) = .1$$

$$P(\text{black}|\text{The, cat, is}) = .85$$

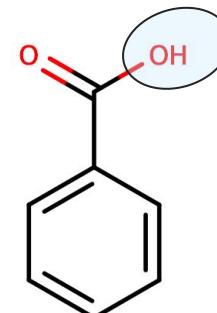
$$P(\text{green}|\text{The, cat, is}) = .05$$

Next Token Probabilities Can Also Be Applied to Chemistry

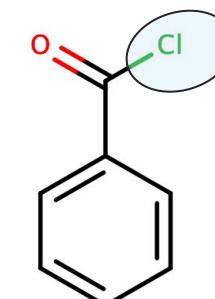
$P(C, \text{c}1\text{cccccc1C(=O)}) = 0.2$



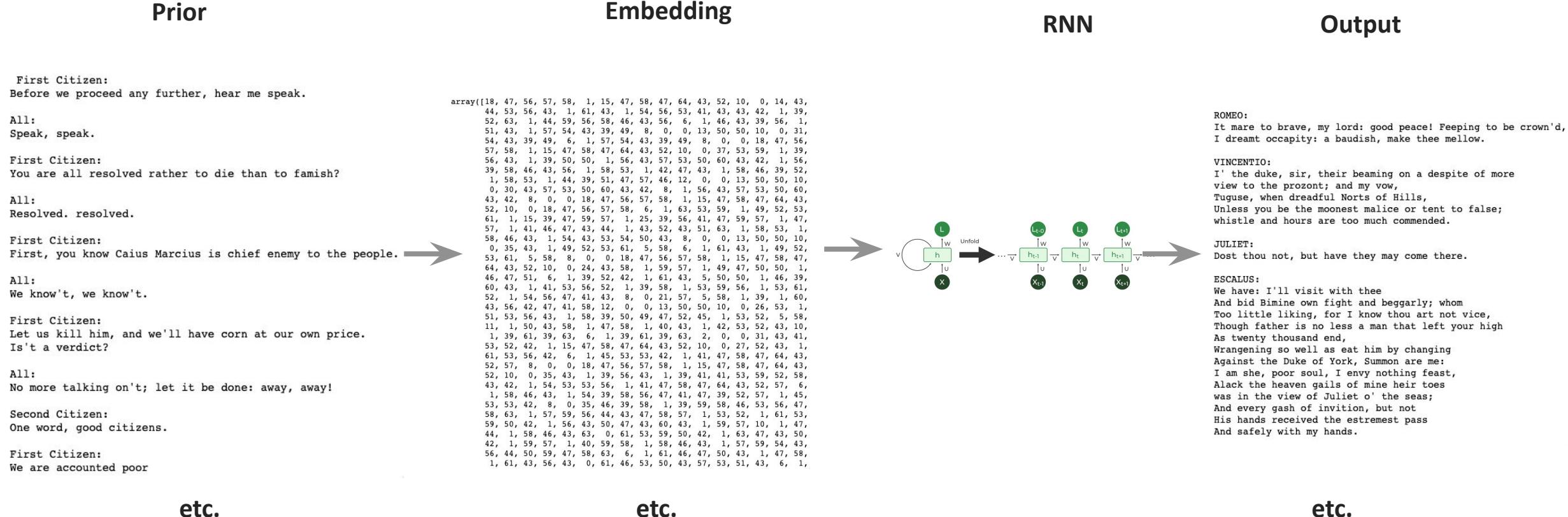
$P(O, \text{c}1\text{cccccc1C(=O)}) = 0.75$



$P(\text{Cl}, \text{c}1\text{cccccc1C(=O)}) = 0.05$



Training a Recurrent Neural Network (RNN) to Generate “Shakespeare”



Training a Recurrent Neural Network (RNN) to Generate “Shakespeare”

Prior

First Citizen:
Before we proceed any further, hear me speak.

All:
Speak, speak.

First Citizen:
You are all resolved rather to die than to famis

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcus is chief enemy to

All:
We know't, we know't.

First Citizen:
Let us kill him, and we'll have corn at our own
Is't a verdict?

All:
No more talking on't; let it be done: away, away

Second Citizen:
One word, good citizens.

First Citizen:
We are accounted poor

etc.

ROMEO:

It mare to brave, my lord: good peace! Feeping to be crown'd,
I dreamt occapity: a baudish, make thee mellow.

VINCENTIO:

I' the duke, sir, their beaming on a despite of more
view to the prozont; and my vow,
Tuguse, when dreadful Norts of Hills,
Unless you be the moonest malice or tent to false;
whistle and hours are too much commended.

JULIET:

Dost thou not, but have they may come there.

ESCALUS:

We have: I'll visit with thee
And bid Bimine own fight and beggarly; whom
Too little liking, for I know thou art not vice,
Though father is no less a man that left your high
As twenty thousand end,
Wrangening so well as eat him by changing
Against the Duke of York, Summon are me:
I am she, poor soul, I envy nothing feast,
Alack the heaven gails of mine heir toes
was in the view of Juliet o' the seas;
And every gash of invitision, but not
His hands received the estremest pass
And safely with my hands.

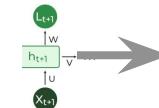
Output

ROMEO:
It mare to brave, my lord: good peace! Feeping to be crown'd,
I dreamt occapity: a baudish, make thee mellow.

VINCENTIO:
I' the duke, sir, their beaming on a despite of more
view to the prozont; and my vow,
Tuguse, when dreadful Norts of Hills,
Unless you be the moonest malice or tent to false;
whistle and hours are too much commended.

JULIET:
Dost thou not, but have they may come there.

ESCALUS:
We have: I'll visit with thee
And bid Bimine own fight and beggarly; whom
Too little liking, for I know thou art not vice,
Though father is no less a man that left your high
As twenty thousand end,
Wrangening so well as eat him by changing
Against the Duke of York, Summon are me:
I am she, poor soul, I envy nothing feast,
Alack the heaven gails of mine heir toes
was in the view of Juliet o' the seas;
And every gash of invitision, but not
His hands received the estremest pass
And safely with my hands.



etc.

Training a Recurrent Neural Network (RNN) to Generate SMILES

Prior

A large, solid gray arrow pointing to the right, indicating the direction of the next section.

Embedding

RNN

Output

C(=O)Oc1c(C)c2cc(C(F)(F)c1)cc(C)c1
 C(=O)C(C)NCC1c(c1c(C)c1)cc1
 C(C)C1ccc(NC(=O)N2c3cccc3C2C)c1CC
 O=(NC1)c2cccc(CCCC)c1cc(-c8ccco2)cc1Br
 OCC1OC(=O)c2c(F)c3c(SC4acc(C1)c1)4)ncn32)C(O)C10
 N#CC=C1cccc(OCC(=O)c1)c1nc2cccc2[nH]1
 CC(=O)OCC1(CC2C(C)C3=CC4C(CCC(OC5OC(CO)C(O)C50)(C(O)c40)C3(C)CCC2CC13CCC)=O)
 CCCN(C)C1CC2c(C)c1cc2C1
 CNS(=O)=O)Cc1nn(COC2cccccc2)n1
 C(C)(=O)c1Cc(C)NC2=C(=O)C(Cc3cc(CC)c3C2)c1cc1cccc(F)c1
 C(=O)c1N(=O)c2cccc(C(F)c2)c2Cc1O[N]1CCO
 =O;c1cc(c2[nH]j)c3cccc32)c1-c2cccc(=O)c2n1
 C1cc(c2C(C)C(F)c2C)c1
 Cc3cc(C(F)c3)cc2C1
 Br=c1cc(c2[nH]j)c3cccc32)c1
 C(=O)N(C1)C2CC3C1CC1C(O)(C3)C2)c1cc(N2C2CCS(=O)=O)c2nn1Cc1cccccc1
 O[!L]CCN[G]Cc3cc(C(F)c3)N[G]NNG3C2GG3c1cc2c1
 C(C)C1ccc(NC(C)C(=O)c1Cc(C)c1
 C(=O)=N)Cc1N(H)N(CCCN)CCCCN(=O)10
 Cc1cccc1-c1nc(C(C)N2cc2c(CCCC)c1)3cccc(N)c3)cc2)1o
 C2Cc1nc(N2cc2c(F#)F)c2)nc(N2CCOC2)1n
 CS(=O)=O)N1NC1CC1(CM(=O)c2c(Cl)c1)s2Cc1
 C(=O)C(C(=O)N(=O)C(C)N)=O)NC(COCOCOCOCOCOCOCNCNNNNNNNNNNNNNNNN
 CCO(=O)C(NC(=O)=O)c1N(c1ccccc1)C1OC(C(=O)OCC=Cc2cc(-c3cccc3)c2)C(C)(S)1
 CN(=O)C(NC(=O)=O)N2cc(c2Cc1)oc2)Nc1c2cccccc21
 CCCCN(CCCC)COC1cc(c1Cc1
 CCCS(=O)=O)N(N)c1nn2cc(O)ccc2s1
 Co1c(c1-c2cc3c1c2)NC(=O)c3=Cc2[nH]c(C)c(C)c2Cc1cc10
 C1Cc1(C)C(=O)c2Cc2-CN(C)C(=O)c12
 C(C)(=O)c1cc(Cc2cccc3c(C(=O)O)c3c2)c(=O)[nH]1
 G1c(C)C2CCCCCCC2)nc(-c2cc(Br)c2)n1-c1cc(C1)c1
 CCCC(C)O1cc(c1-c2cc(C(F)F)c1)nc(N(C)C)C(O)n2)cc1
 CCN(C)C1cc(c1(NC(=O)c2ccccc3c(CC)c2)cc1
 CCO1Cc(N2c(z=O)=H)cc(O)c2=CCO1
 C(C)(=O)CCCCC(COP(=O)(O)O)c1=O)O
 C(C)(=O)CCCCC(COP(=O)(O)O)c1=O)O
 Cc1cccc10cc(c1-c2cccc(C(F)c1)cc(C)c1)S(=O)(=O)N3CCOC3)c2)cl
 Cc1cccc10cc(c1-c2cccc(C(F)c1)cc(C)c1)S(=O)(=O)N3CCOC3)c2)cl
 CCO1Cc(N2c(z=O)C)Nc3c4c(C(=O)COP(=O)(O)c1=O)C
 C1c(-c1=O)cc(C)C2c1n1cc(c4(C)c1)cc12
 O=(C1CCOC1)N1CCM(c2nncc(C1)c2Bz)C1
 CC(=O)=O)cc(c1(C)C)sc1Cc(=O)(C(C)c1cc(C(C)C(C)c(=O)[nH]n1
 CN(C)N1CCc2ccccc(C(F)F)c1)CC1
 Co1c(ccc(C(=O)c2cccc2C(F)F)c1)cc1
 Cc1ccc(NC(=O)c2c[nH]j)ncn34CC4Cc1
 Cc1ccc(C(=O)c2c[nH]j)ncn32)C(O)c1
 Cc1ccc(C(=O)c2c[nH]j)ncn32)C(O)c1
 C1(C)C(=O)NCC(=O)N2cc(c1Sc3cc(C1)c3Cc(=O)N3CCOC3)c2)C1

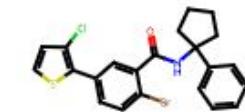
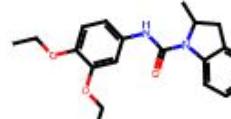
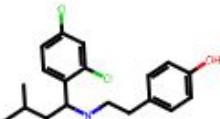
etc.

etc.

etc

Training a Recurrent Neural Network (RNN) to Generate SMILES

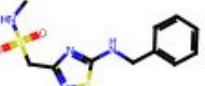
Prior



Output



etc.



etc

Augmented Likelihood Incorporates Scoring Into Molecule Generation

Prior

A thick, dark gray arrow pointing to the right, indicating a continuation or next step.

Embedding

RNN

Output

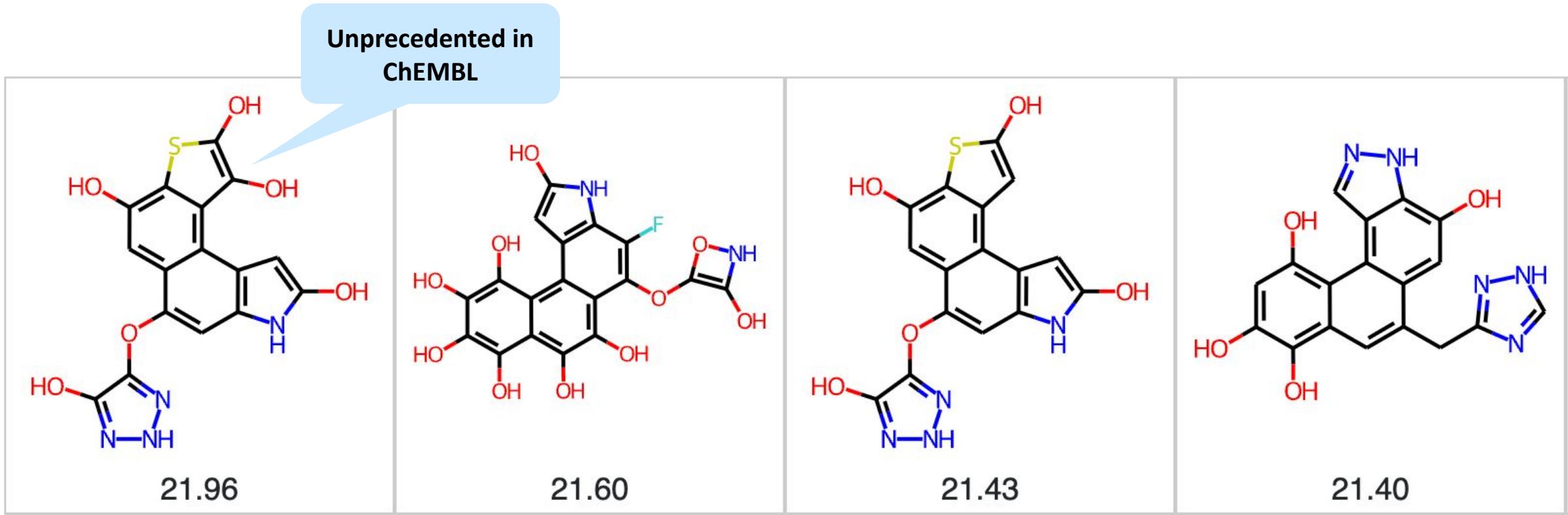
etc.

etc.

etc.

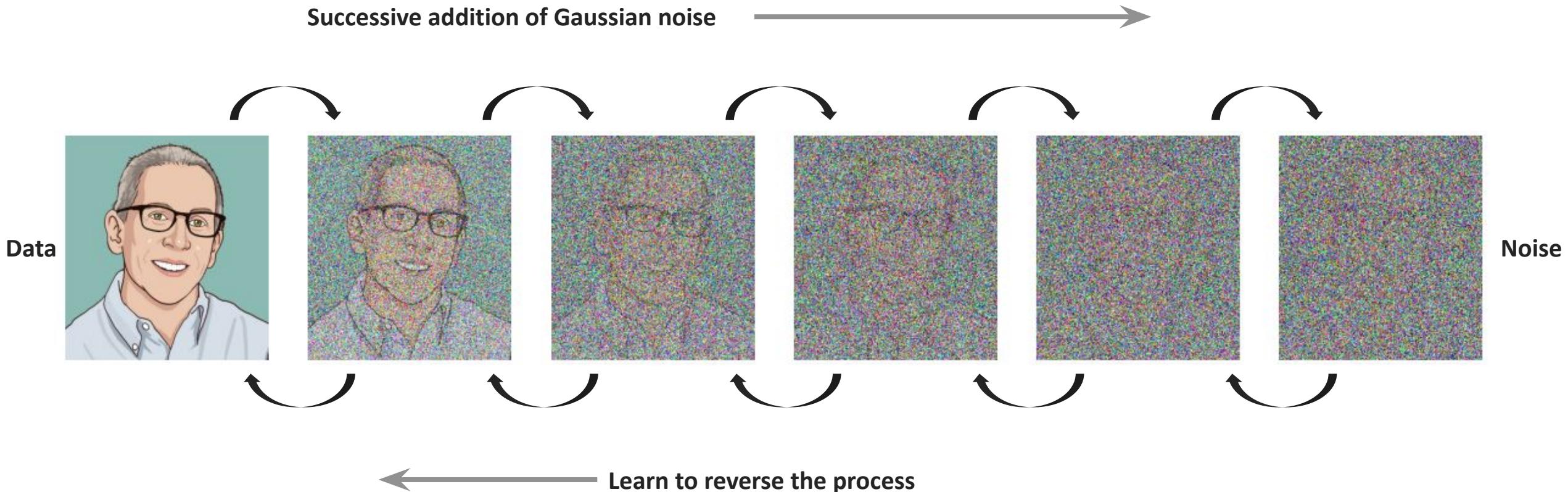
Augmented likelihood

Generative Molecules Can “Game” Scoring Functions



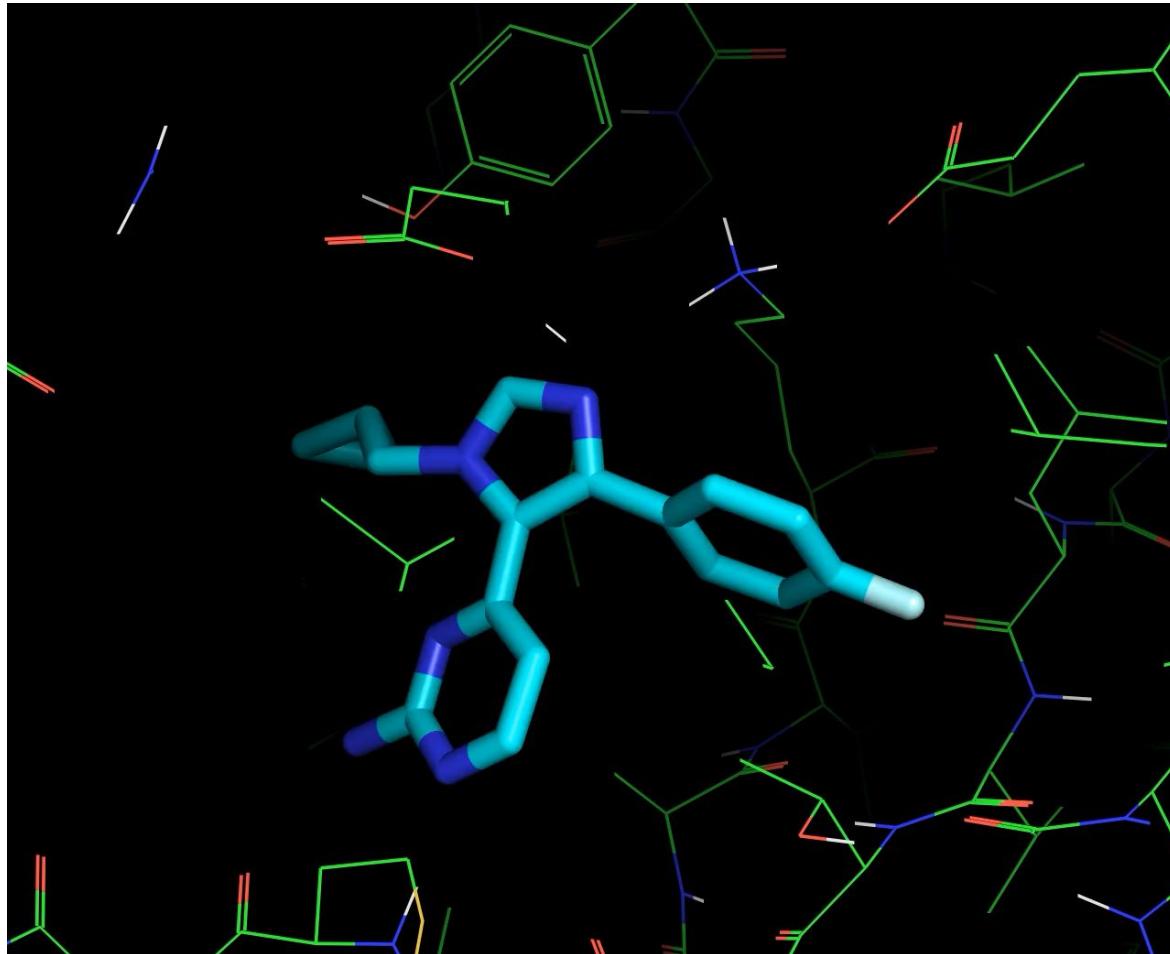
Dyachenko, Natalia V., et al. "Synthesis of fused heterocyclic systems via the Mallory photoreaction of arylthienylethenes." *Photochemical & Photobiological Sciences* 18.12 (2019): 2901-2911.

Diffusion Models Learn to Successively Denoise Data



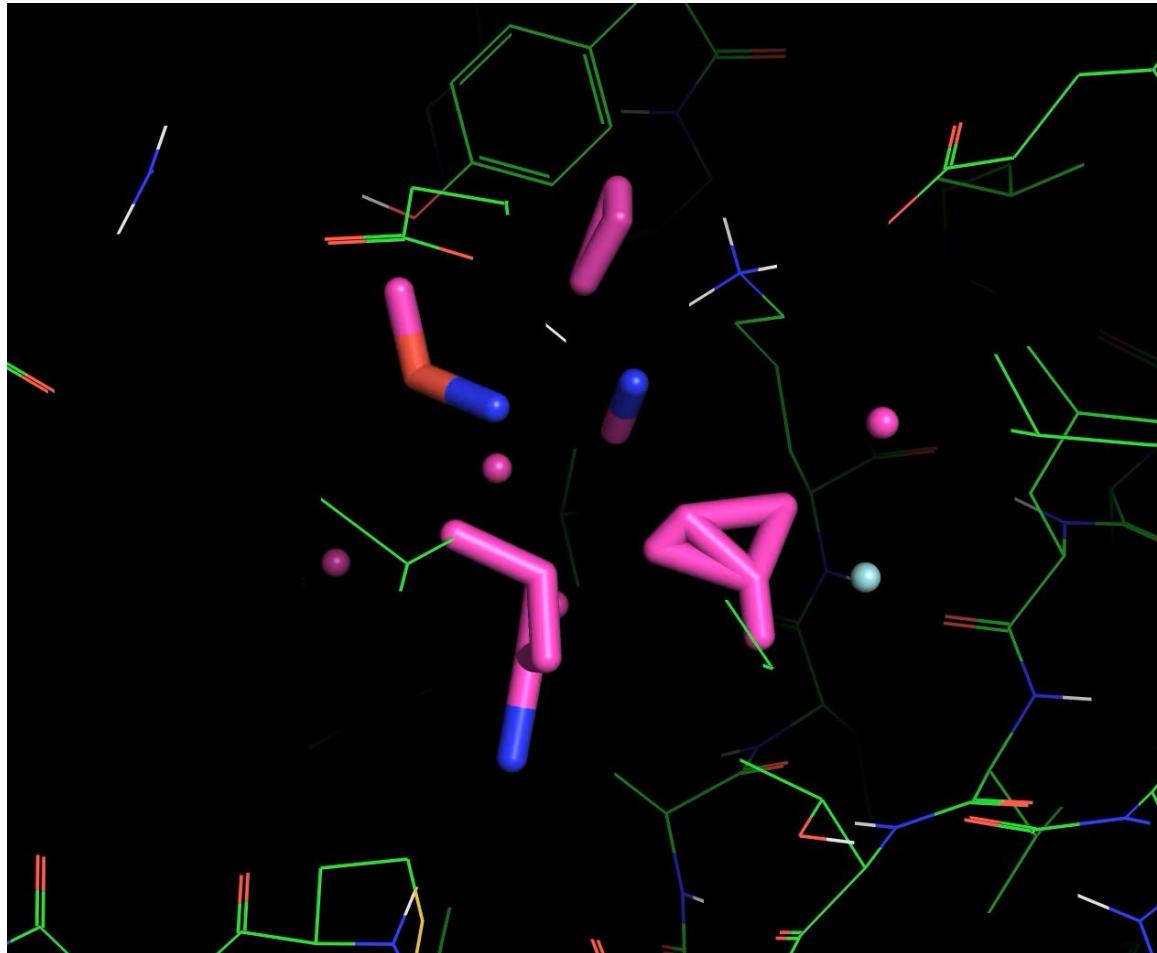
Once trained, the model can generate “data” from noise

Diffusion Can Also Be Used To Generate 3D Protein and Ligand Structures



Similar denoising strategies can be applied to atoms in a binding site

A Trained Diffusion Model Can Transform Atoms at Random Positions



Begin with atoms at random positions and use diffusion to generate a bound molecule

Molecule Generation is Not Without Its Problems



Article

Design of SARS-CoV-2 Main Protease Inhibitors Using Artificial Intelligence and Molecular Dynamic Simulations

Lars Elend ¹, Luise Jacobsen ², Tim Cofala ¹, Jonas Prellberg ¹, Thomas Teusch ³, Oliver Kramer ^{1,*} and Ilia A. Solov'yov ^{3,4,5,*}

¹ Computational Intelligence Lab, Department of Computer Science, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany; lars.elend@uni-oldenburg.de (L.E.); tim.cofala@uni-oldenburg.de (T.C.); jonas.prellberg@uni-oldenburg.de (J.P.)

² Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark; luja@sdu.dk

³ Department of Physics, Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany; thomas.teusch@uni-oldenburg.de

⁴ Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany

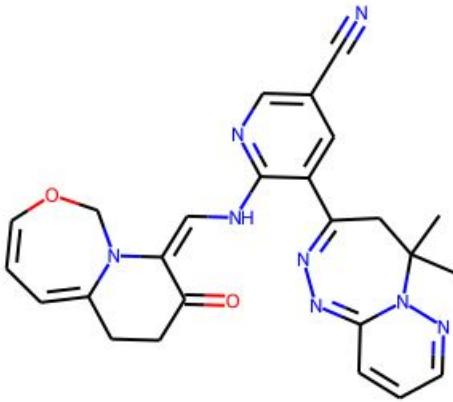
⁵ Center for Nanoscale Dynamics (CENAD), Carl von Ossietzky Universität Oldenburg, Institut für Physik, Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany

* Correspondence: oliver.kramer@uol.de (O.K.); ilia.solovyov@uni-oldenburg.de (I.A.S.); Tel.: +49-441-798-3817 (I.A.S.)

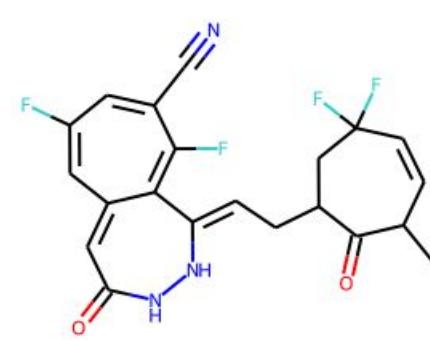
Highest Scoring Molecules Contain Questionable Functionality



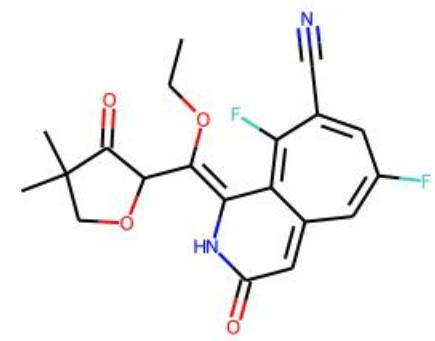
Lig 1



Lig 7



Lig 19

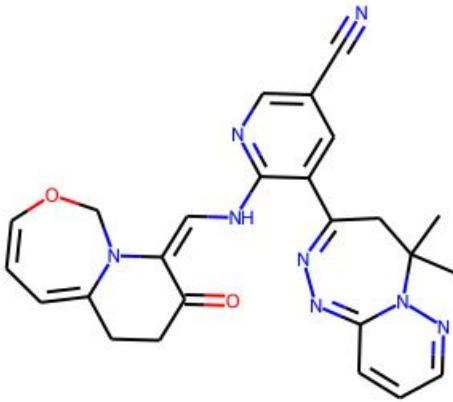


Lig 21

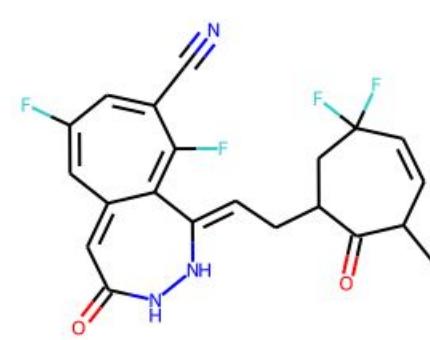
Highest Scoring Molecules Contain Questionable Functionality



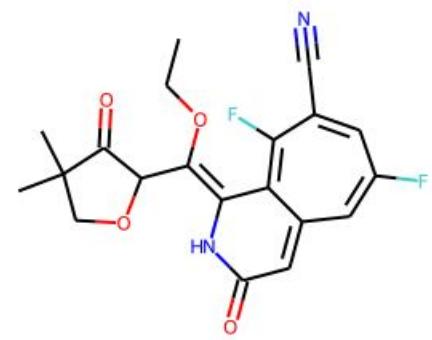
Lig 1



Lig 7

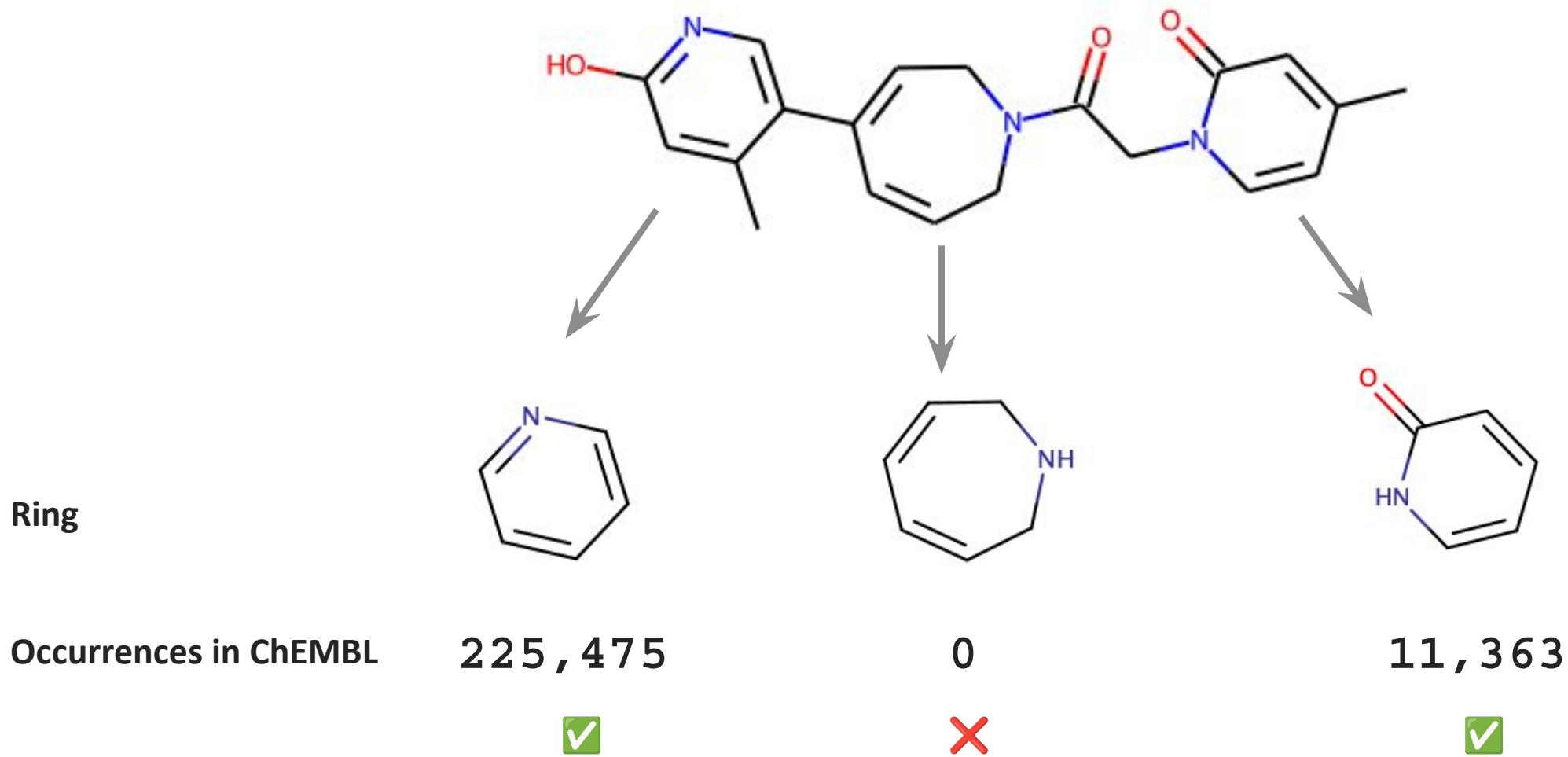


Lig 19

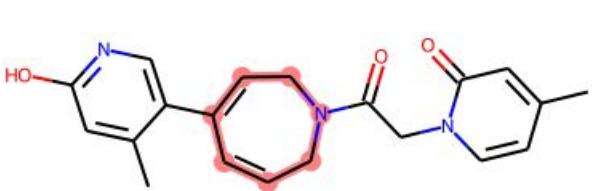


Lig 21

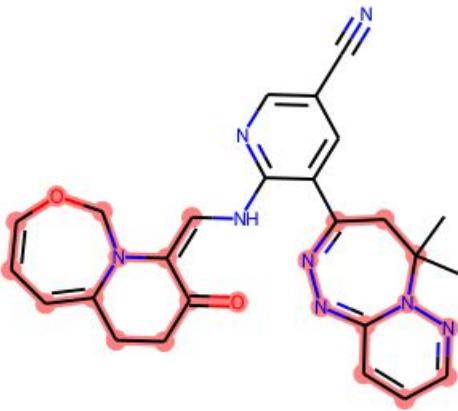
Evaluating Ring System Frequency



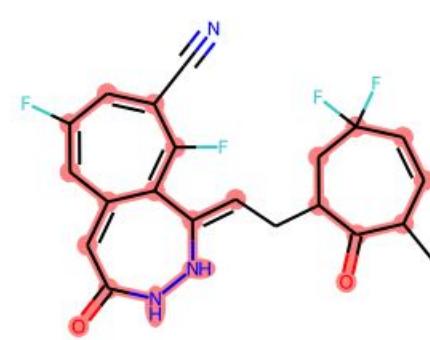
Top Molecules



Lig 1



Lig 7



Lig 19



Lig 21

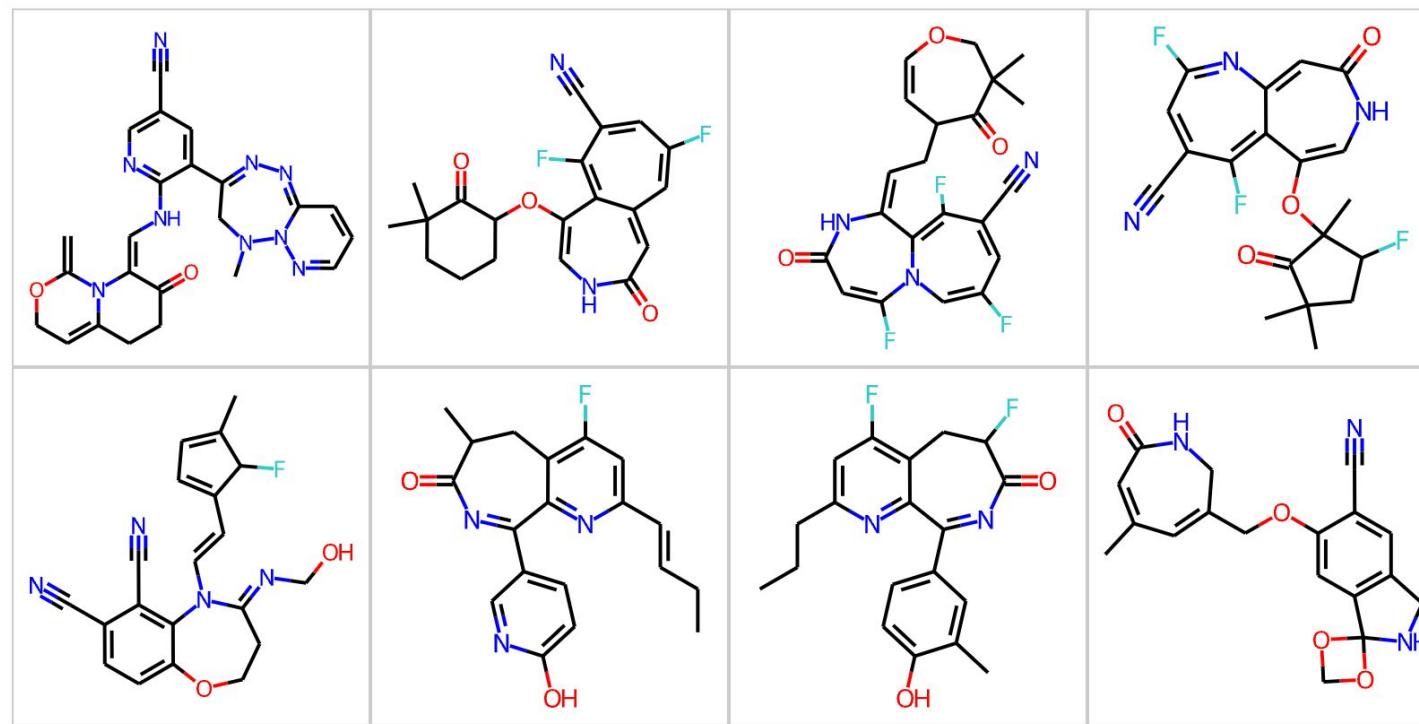
Evaluating Generative Model Output

144,350 molecules generated

23 molecules violated rules of valence

107,386 (74%) contained ring systems not found in ChEMBL

144,350 (79%) contained ring systems occurring < 10 times in ChEMBL



Article

Design of SARS-CoV-2 Main Protease Inhibitors Using Artificial Intelligence and Molecular Dynamic Simulations

Lars Elend ¹, Luise Jacobsen ², Tim Cofala ¹, Jonas Prellberg ¹, Thomas Teusch ³, Oliver Kramer ^{1,*} and Ilia A. Solov'yov ^{3,4,5,*}

¹ Computational Intelligence Lab, Department of Computer Science, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany; lars.elend@uni-oldenburg.de (L.E.); tim.cofala@uni-oldenburg.de (T.C.); jonas.prellberg@uni-oldenburg.de (J.P.)

² Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark; luja@sdu.dk

³ Department of Physics, Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany; thomas.teusch@uni-oldenburg.de

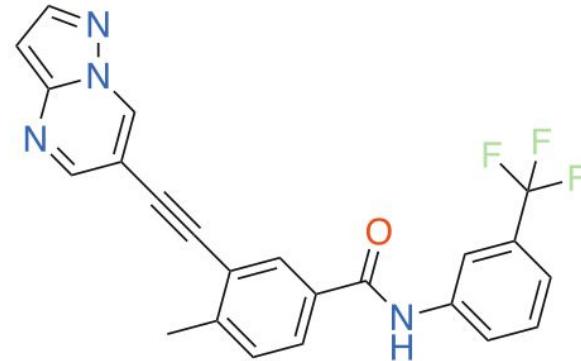
⁴ Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany

⁵ Center for Nanoscale Dynamics (CENAD), Carl von Ossietzky Universität Oldenburg, Institut für Physik, Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany

* Correspondence: oliver.kramer@uol.de (O.K.); ilia.solovyov@uni-oldenburg.de (I.A.S.); Tel.: +49-441-798-3817 (I.A.S.)

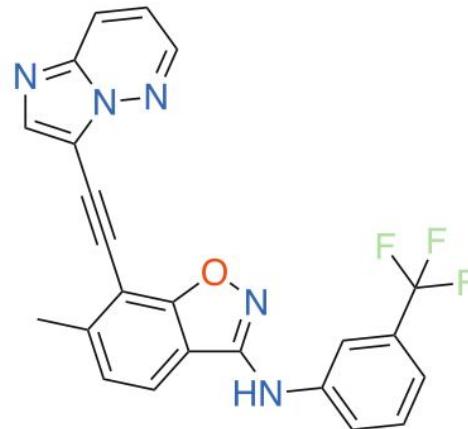
Molecules 2022, 27, 4020

Assessing the Novelty of Molecules Produced By Generative Models



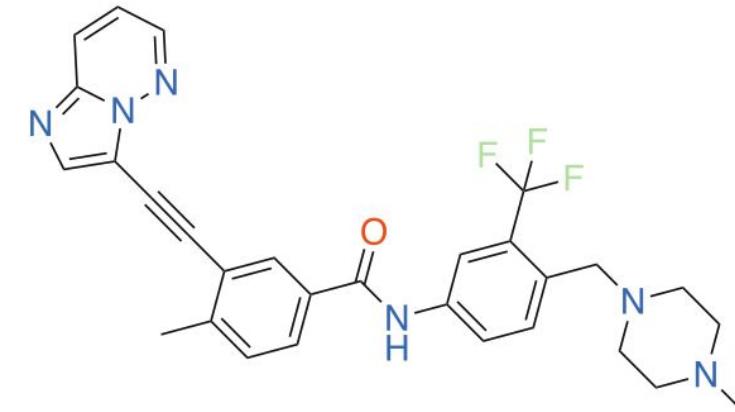
Gao et al.
Compound 7r
6 nM

Published in JMC 2013



Zhavoronkov et al.
Compound 1
10 nM

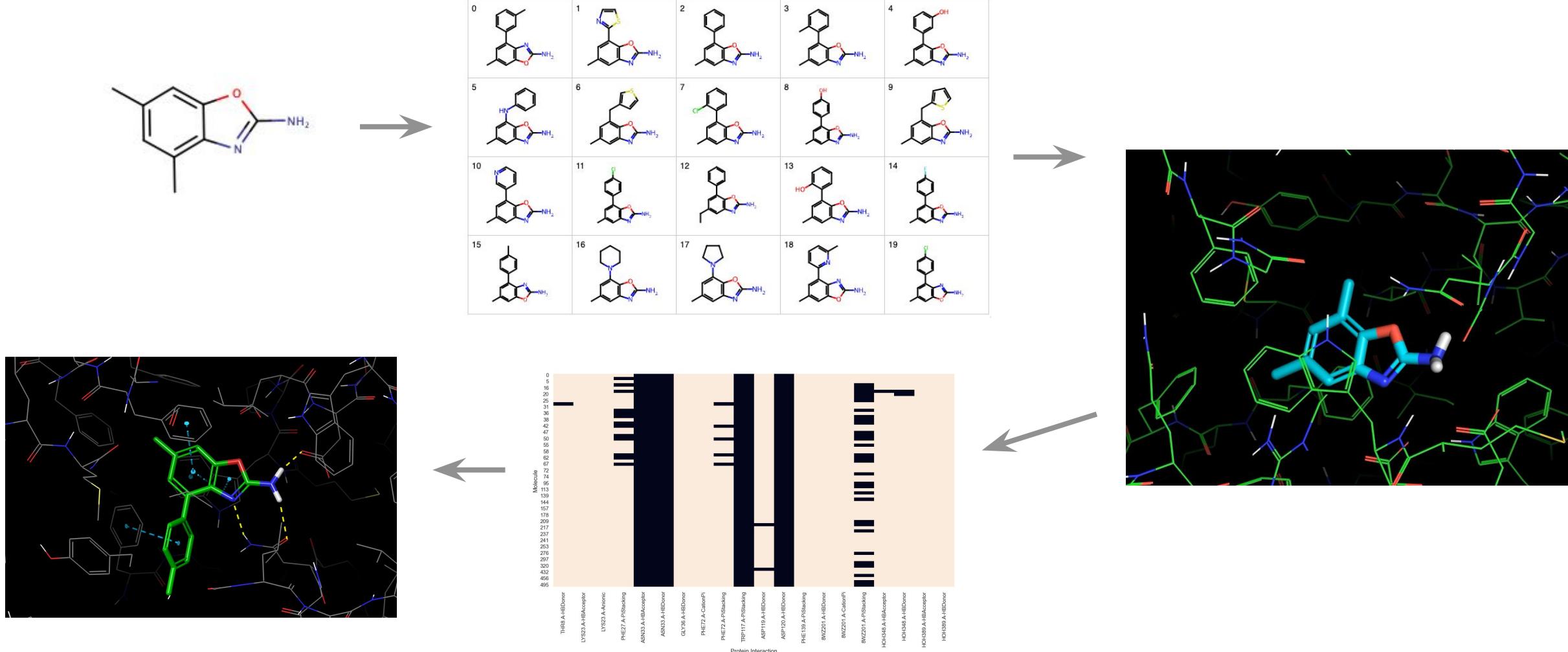
Generative Design
2020



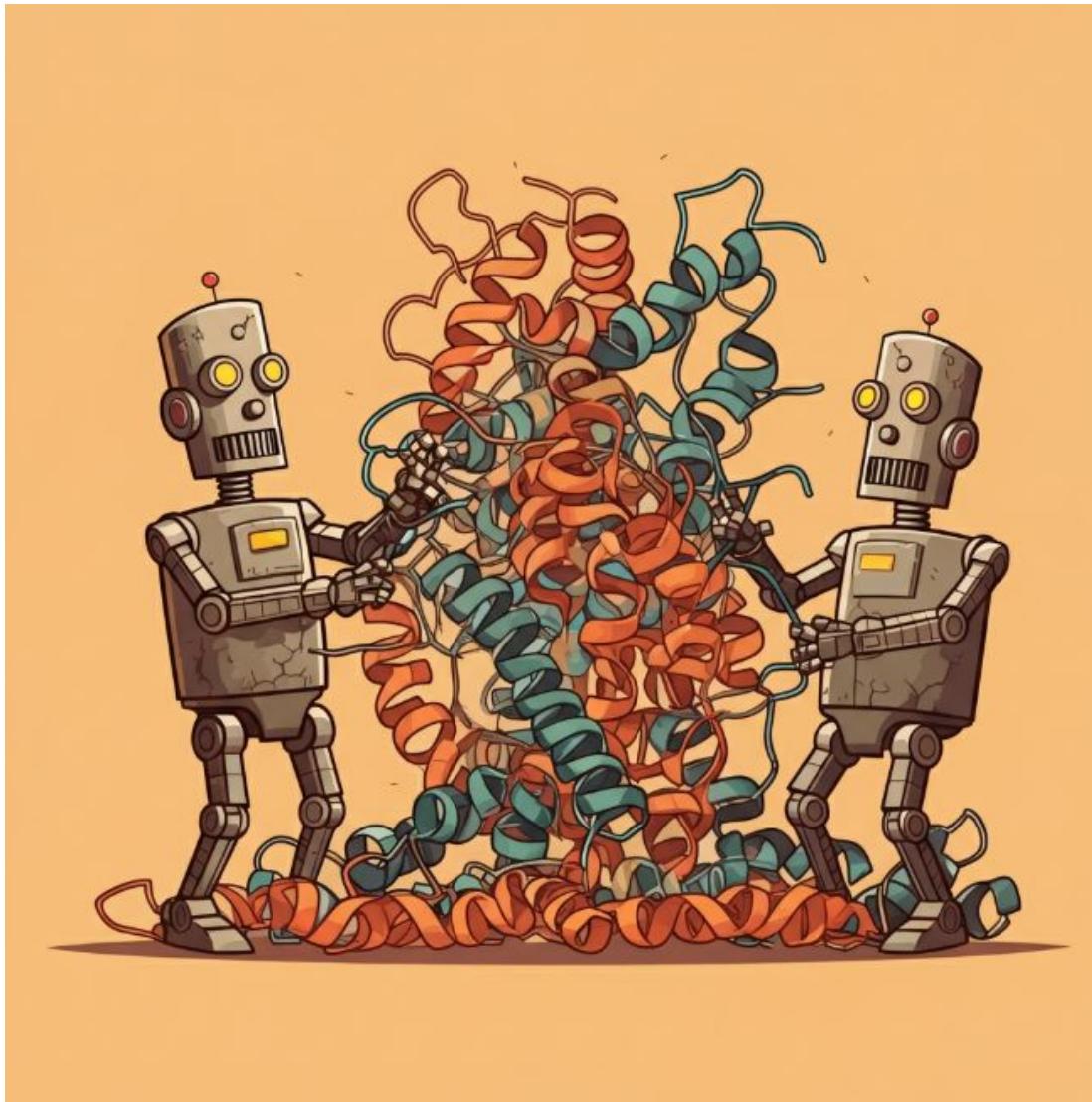
Ponatinib
9 nM

FDA Approved
2012

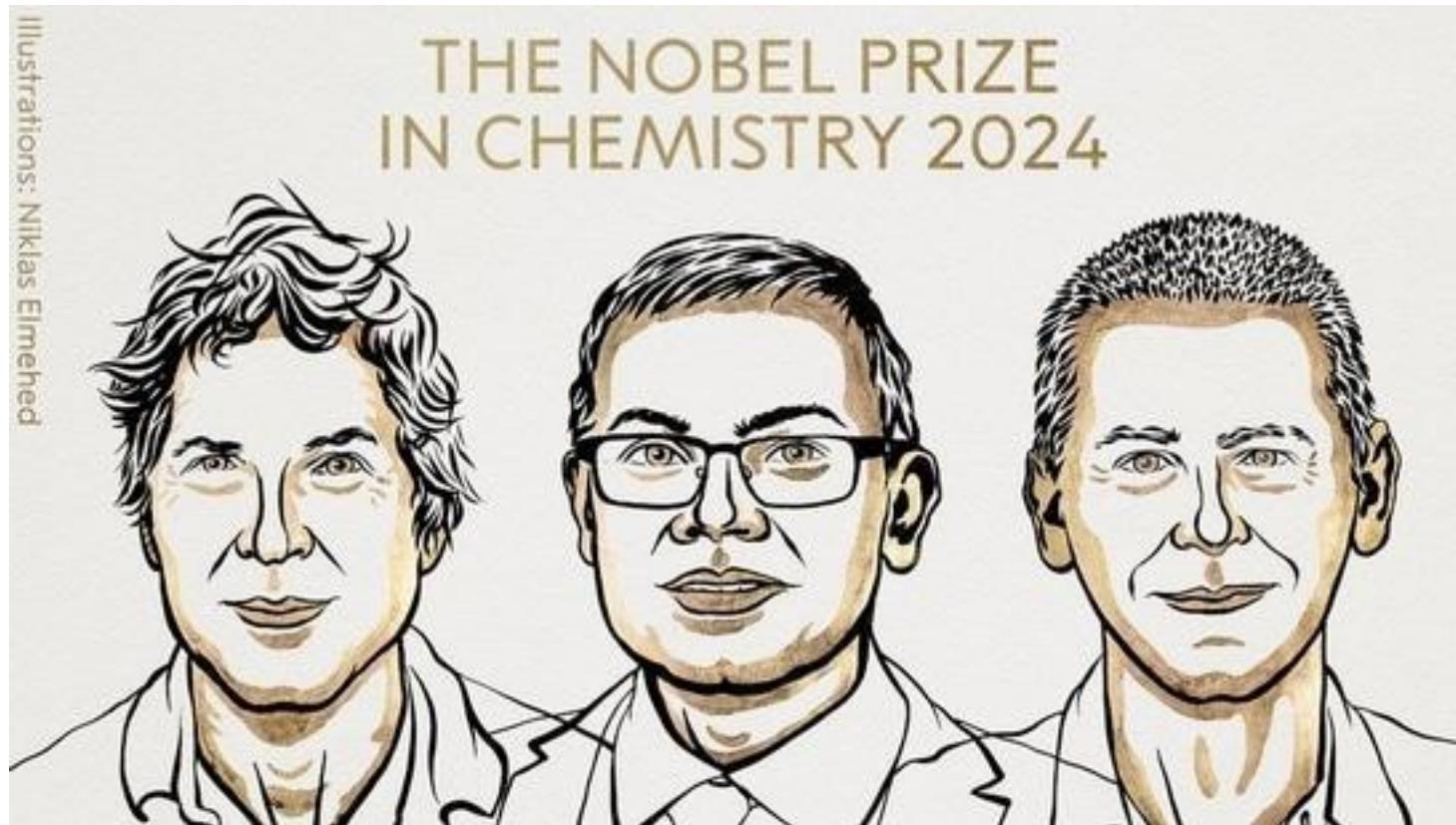
A Generative Design Workflow for Fragment Screening



Protein Structure Prediction and Cofolding

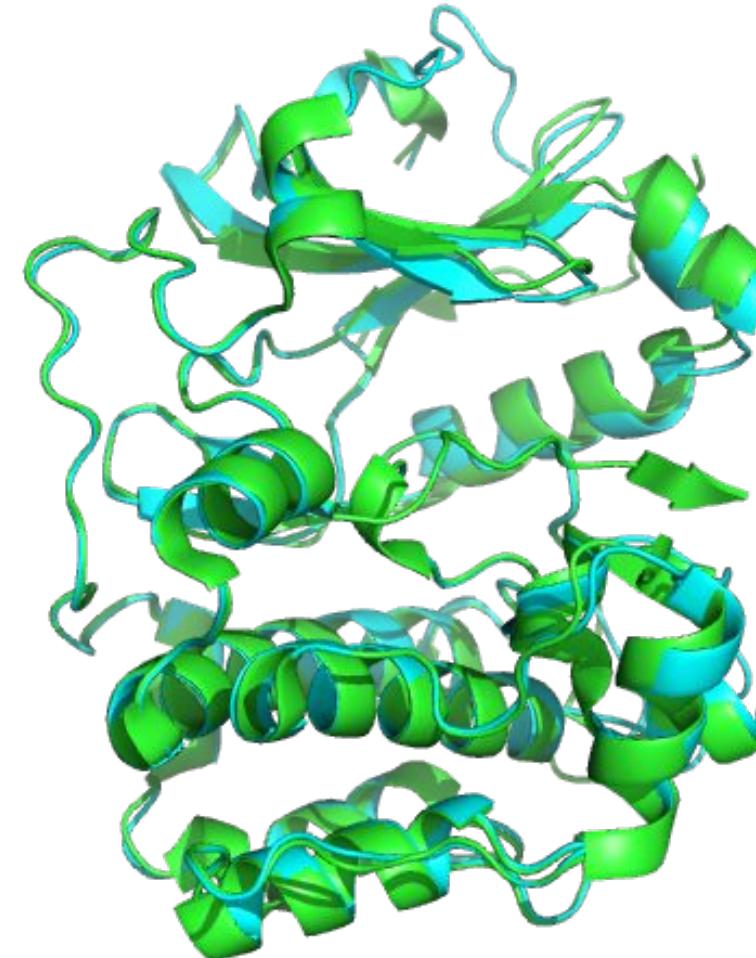


Protein Structure Prediction



Converting Amino Acid Sequence to 3D Structure

```
VTMNEFEYLKLLKGKTFGKVILVKEATGRYYAMK  
ILKKEVIVAKDEVAHTLTENRVLQNSRHPFLTALKYS  
FQTHDRLCFVMYEYANGGELFFHLSRERVFSEDRA  
RFYGAEVSAALDYLHSEKNVVYRDLKLENLMLDKD  
GHIKITDFGLCKEGIKDGATMKFCGTPEYLAPEVLE  
DNDYGRAVDWWGLGVVMYEMMCGRLPFYNNQ  
DHEKLFEIILMEEIRFPRTLGPEAKSLLSGLLKKDPK  
QRLGGGSEDAKEIMQHRFFAGIVWQHVYEKKLS  
PPFKPQVTSETDTRYFDEEFTAQRPHFPQFDYSA
```



AlphaFold and related techniques convert sequence to 3D structure

Can We Use AlphaFold2 Structures For Compound Design or Discovery?



RETURN TO ISSUE | < PREV MACHINE LEARNING AND... NEXT >

Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures

Anna M. Díaz-Rovira, Helena Martín, Thijs Beuming, Lucía Díaz, Victor Guallar*, and Soumya S. Ray*

Cite this: *J. Chem. Inf. Model.*, 2023, 63, 6, 1668–1674

Publication Date: March 9, 2023

https://doi.org/10.1021/acs.jcim.2c01270

Copyright © 2023 American Chemical Society

Request reuse permissions

Article Views
2994
Altmetric
38
Citations
1
LEARN ABOUT THESE METRICS

PROTEIN SCIENCE

TOOLS FOR PROTEIN SCIENCE | Open Access | ⓘ ⓘ ⓘ ⓘ

Evaluation of AlphaFold2 structures as docking targets

Matthew Holcomb, Ya-Ting Chang, David S. Goodsell, Stefano Forli ✉

First published: 07 December 2022 | https://doi.org/10.1002/pro.4530 | Citations: 4

Review Editor: Nir Ben-Tal

Funding information National Institute of General Medical Sciences, Grant/Award Number: GM069832

NO!

NO!

How accurately can one predict drug binding modes using AlphaFold models?

Masha Karelina, Joseph J. Noh, Ron O. Dror

doi: https://doi.org/10.1101/2023.05.18.541346

This is a preprint. It has not been certified by a journal but peer reviews are

0 4 0 0 0 0 105

NO!

AlphaFold2 structures template ligand discovery

Jiankun Lyu, Nicholas Kapolka, Ryan Gumper, Assaf Alon, Liang Wang, Manish K. Jain, Ximena Barros-Álvarez, Kensuke Sakamoto, Yoojoong Kim, Jeffrey DiBerto, Kuglae Kim, Tia A. Tummino, Sijie Huang, John J. Irwin, Olga O. Tarkhanova, Yurii Moroz, Georgios Skiniotis, Andrew C. Kruse, Brian K. Shoichet, Bryan L. Roth

doi: https://doi.org/10.1101/2023.12.20.572662

This article is a preprint and has not been certified by peer review [what does this mean?]

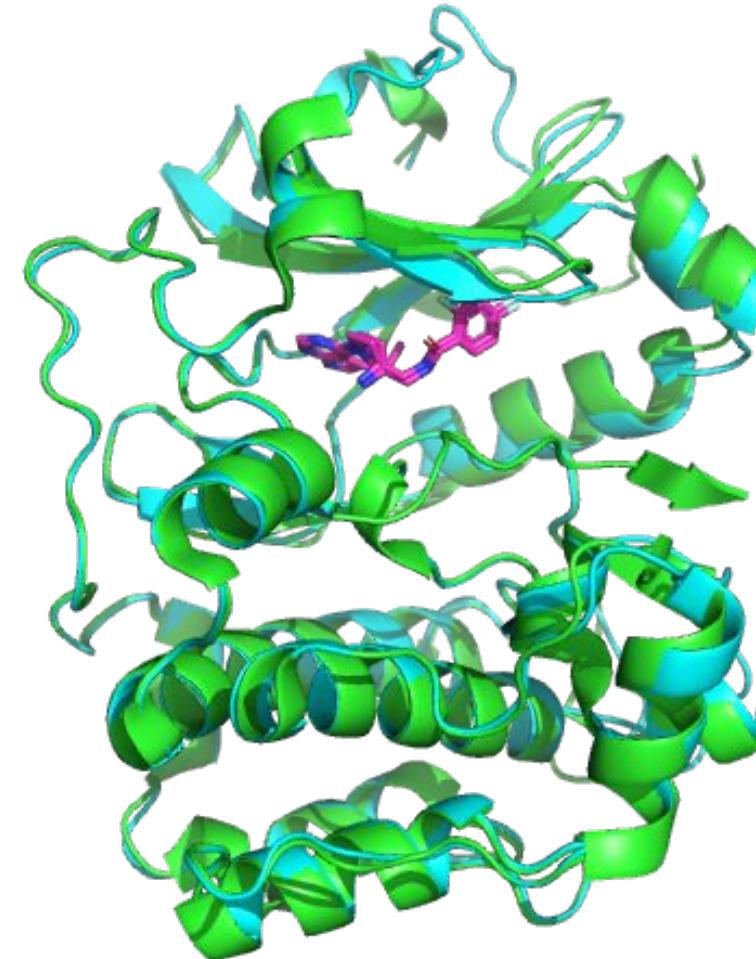
0 0 0 1 0 0 139

Yes!

Performance of AF2 structures for docking is roughly equivalent to that of apo structures

Co-Folding Enables the Prediction of Ligand Poses With Protein Structures

```
VTMNEFEYLKLLKGKTFGKVILVKEATGRYYAMK  
ILKKEVIVAKDEVAHTLTENRVLQNSRHPFLTALKYS  
FQTHDRLCFVMYEYANGGELFFHLSRERVFSEDRA  
RFYGAEVSAALDYLHSEKNVVYRDLKLENLMLDKD  
GHIKITDFGLCKEGIKDGATMKFCGTPEYLAPEVLE  
DNDYGRAVDWWGLGVVVMYEMMCGRLPFYNQ  
DHEKLFLILMEEIRFPRTLGPEAKSLLSGLLKKDPK  
QRLGGGSEDAKEIMQHRFFAGIVWQHVYEKKLS  
PPFKPQVTSETDTRYFDEEFTAQRPHFPQFDYSA  
  
CCc1c[nH]c2c1c(ncn2)N3CCC(C3)(CNC(=O)c4ccc(cc4F)F)N
```



Co-folding was introduced with DragonFold, AlphaFold 3, and RoseTTaFold All Atom

Attack of the (AlphaFold3) Clones

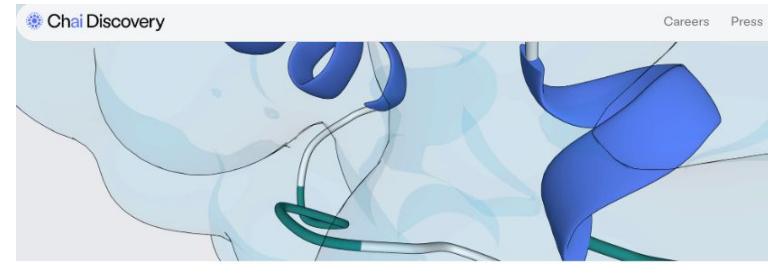


OpenFold



<https://openfold.io/>

Chai-1



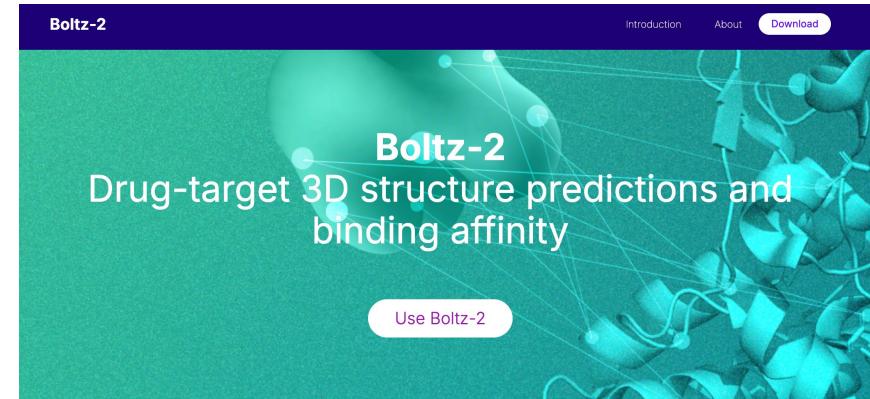
Introducing Chai-1: Decoding the molecular interactions of life

We're excited to release Chai-1, a new multi-modal foundation model for molecular structure prediction that performs at the state-of-the-art across a variety of tasks relevant to drug discovery. Chai-1 enables unified prediction of proteins, small molecules, DNA, RNA, covalent modifications, and more.

The model is available for free via a [web interface](#), including for commercial applications such as drug discovery. We are also releasing the model weights and inference code as a [software library](#) under an Apache 2.0 License.

<https://www.chaidiscovery.com/blog/introducing-chai-1>

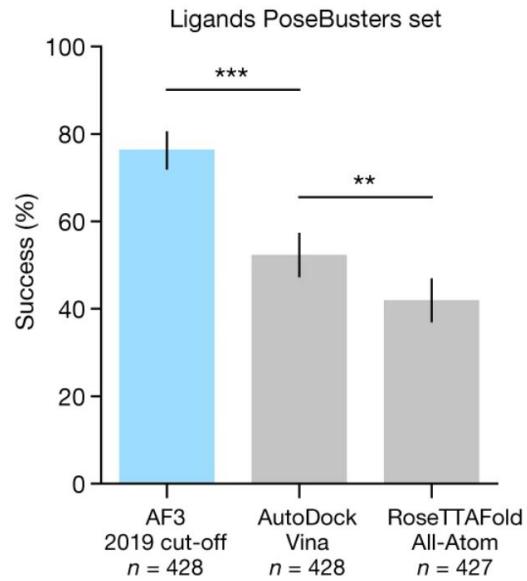
Boltz-2



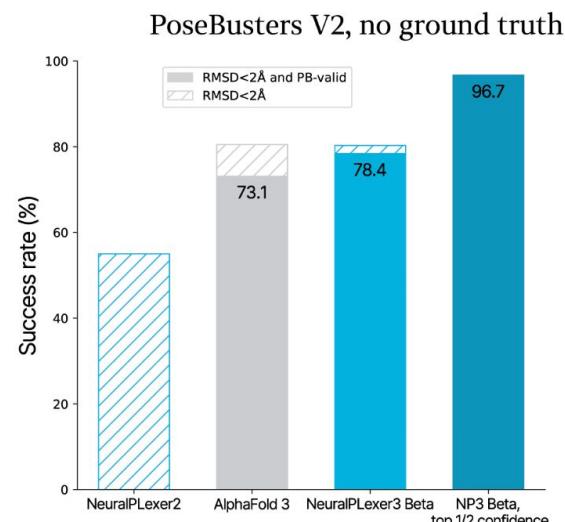
<https://www.rxxr.ai/boltz-2>

AlphaFold3 license restrictions motivated the development of programs with similar functionality

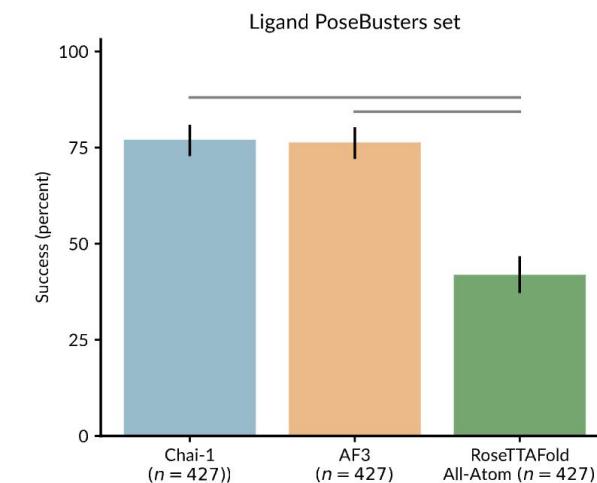
PoseBusters Has Become a Standard Benchmark for Co-folding Methods



AlphaFold 3



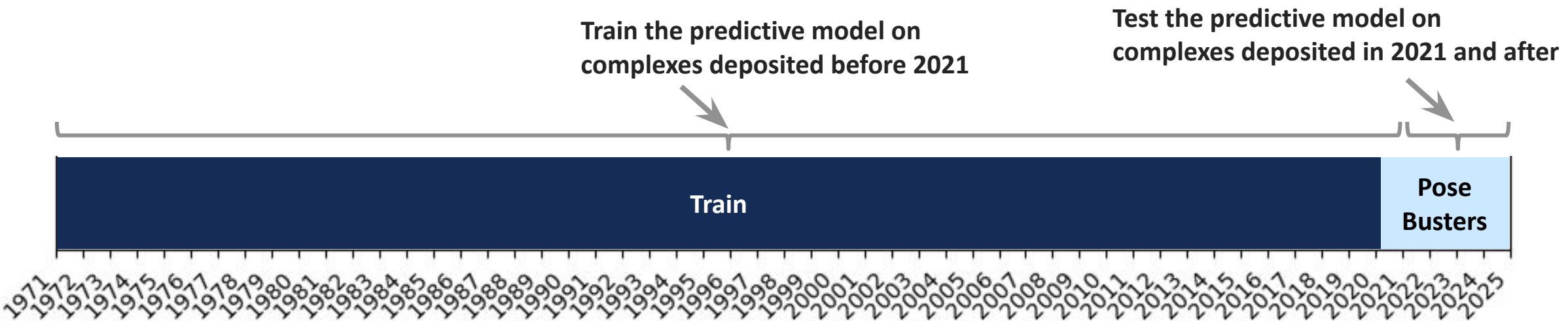
NeuralPlexer



Chai-1

Success rates for co-folding benchmarks may not reflect real-world performance

Time Splits Allow Training Data to Leak Into Predictions



Many test set complexes may have close analogs in the training set

The Most Popular Slide at the 2025 CADD GRC

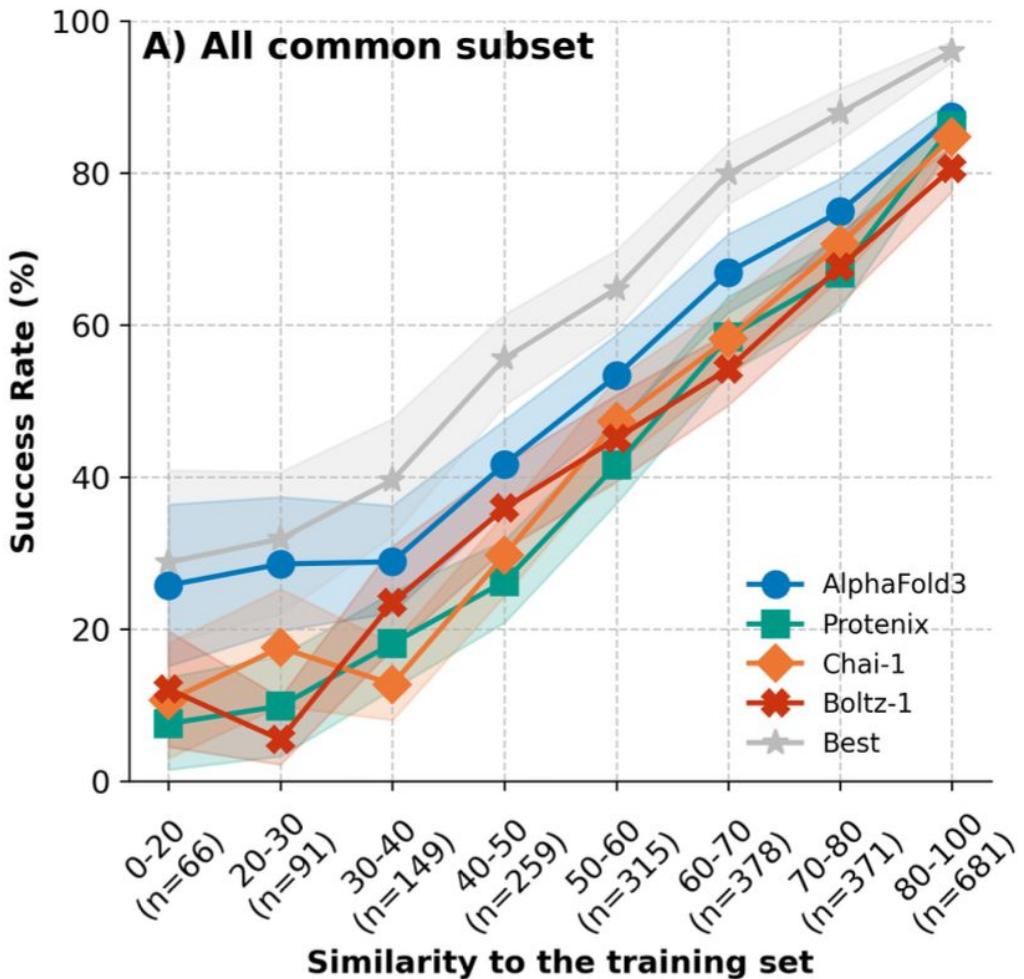
HAVE PROTEIN-LIGAND CO-FOLDING METHODS MOVED BEYOND MEMORISATION?

Peter Škrinjar
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
peter.skrinjar@unibas.ch

Janani Durairaj
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
janani.durairaj@unibas.ch

Jérôme Eberhardt
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
jerome.eberhardt@unibas.ch

Torsten Schwede
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
torsten.schwede@unibas.ch



Co-folding success rate correlates with training set similarity

Benchmarking Co-folding at Orthosteric and Allosteric Sites

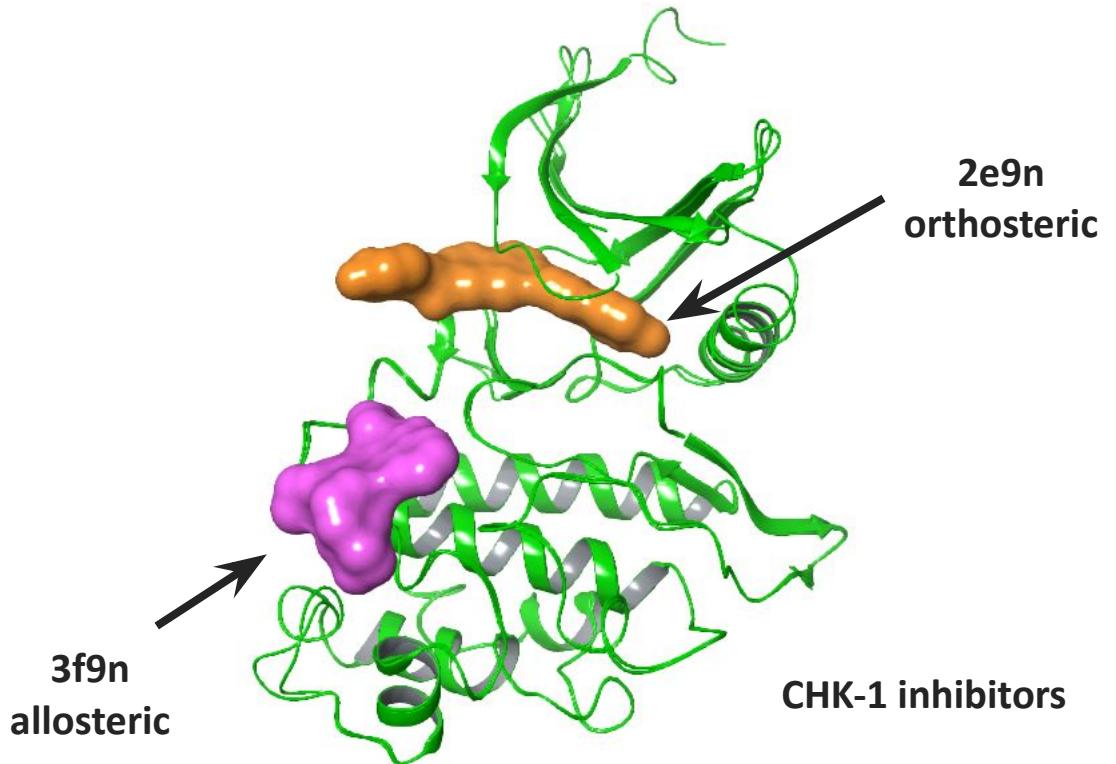
Article

Co-folding, the Future of Docking – Prediction of Allosteric and Orthosteric Ligands

Eva Nittinger, Özge Yoluk, Alessandro Tibo, Gustav Olanders, Christian Tyrchan

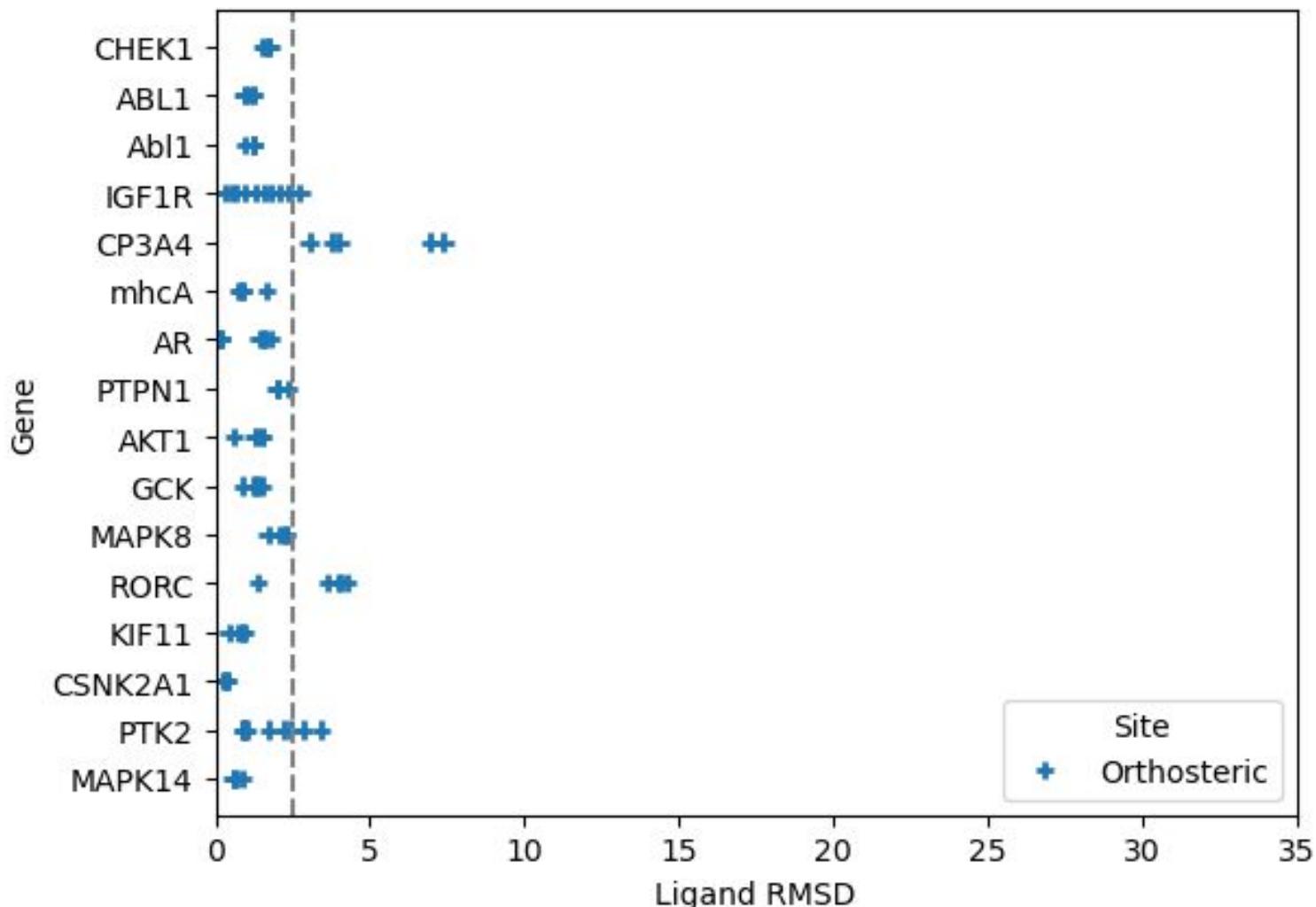
This is a preprint; it has not been peer reviewed by a journal.

<https://doi.org/10.21203/rs.3.rs-6526650/v1>
This work is licensed under a CC BY 4.0 License

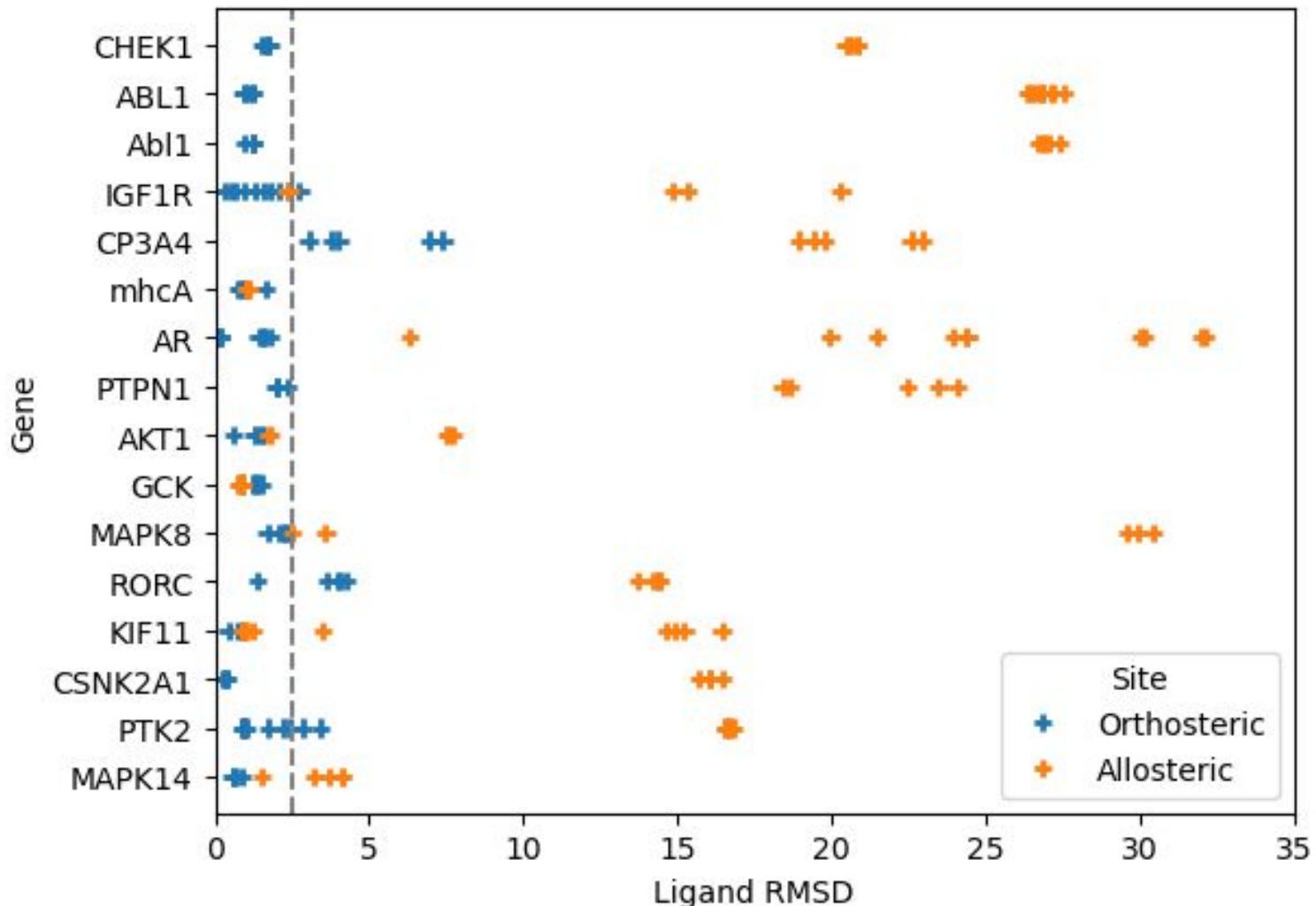


16 pairs protein/ligand complex with both allosteric and orthosteric binders

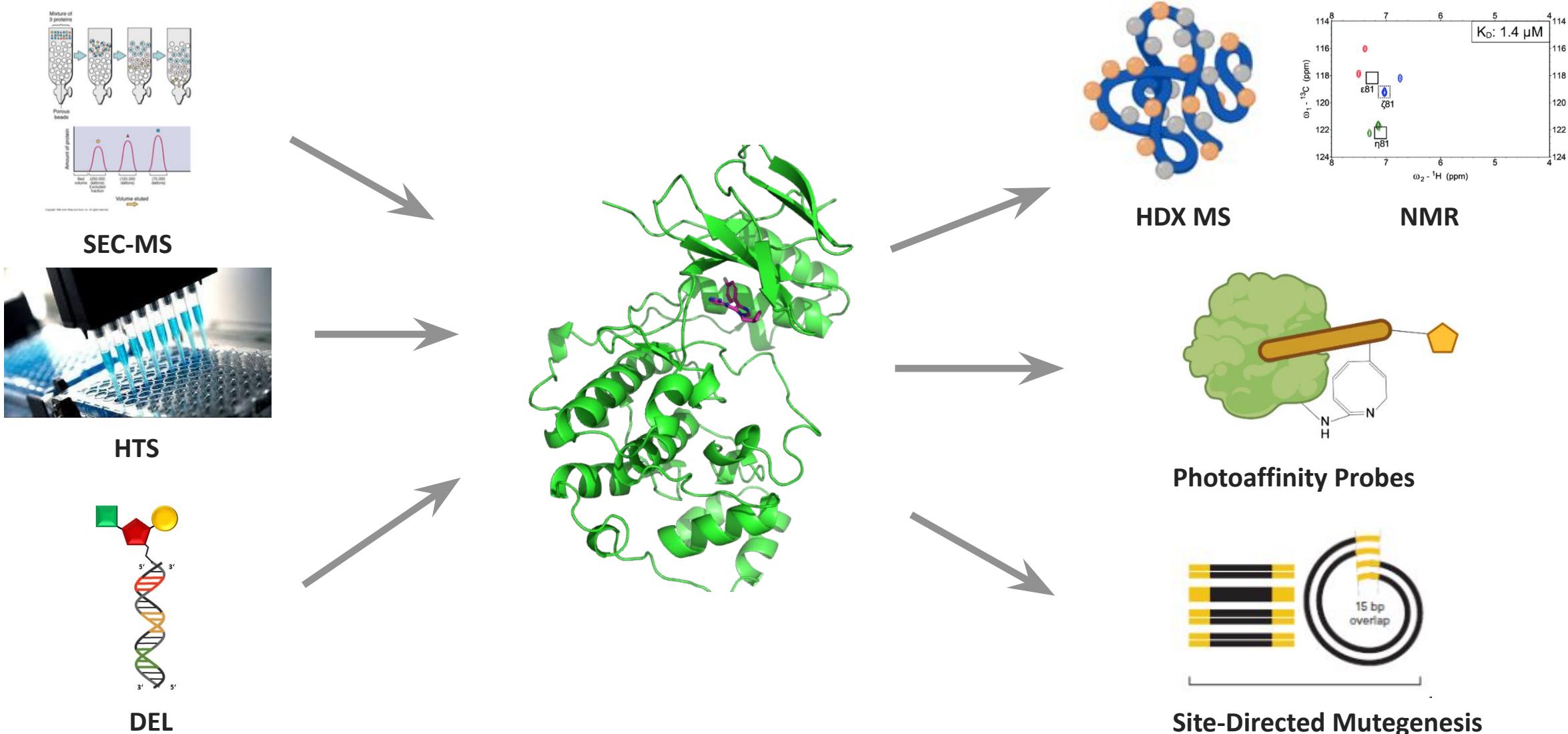
Co-Folding Performs Well With Orthosteric Binding Sites



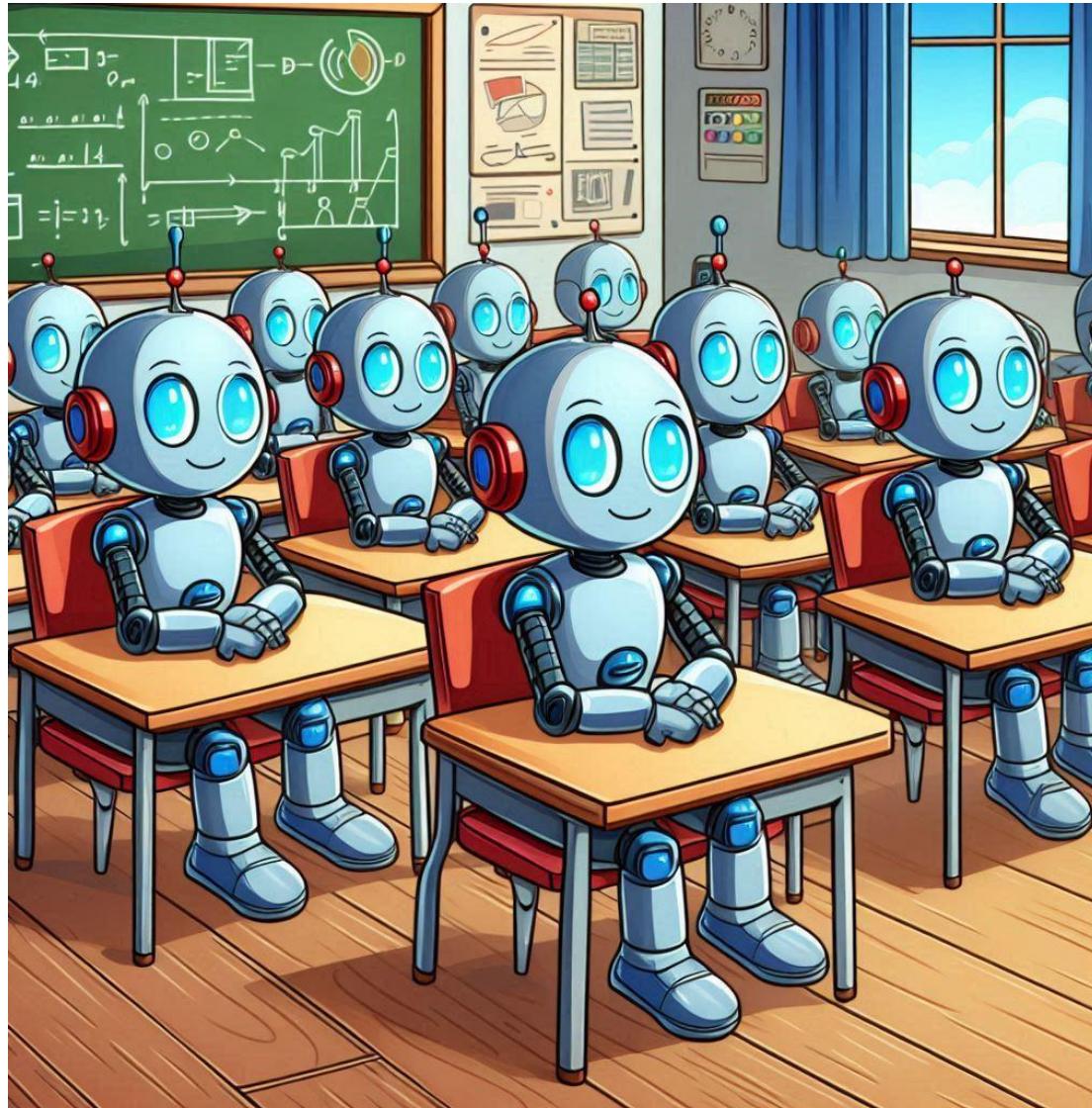
Co-Folding Performs Poorly With Allosteric Binding Sites



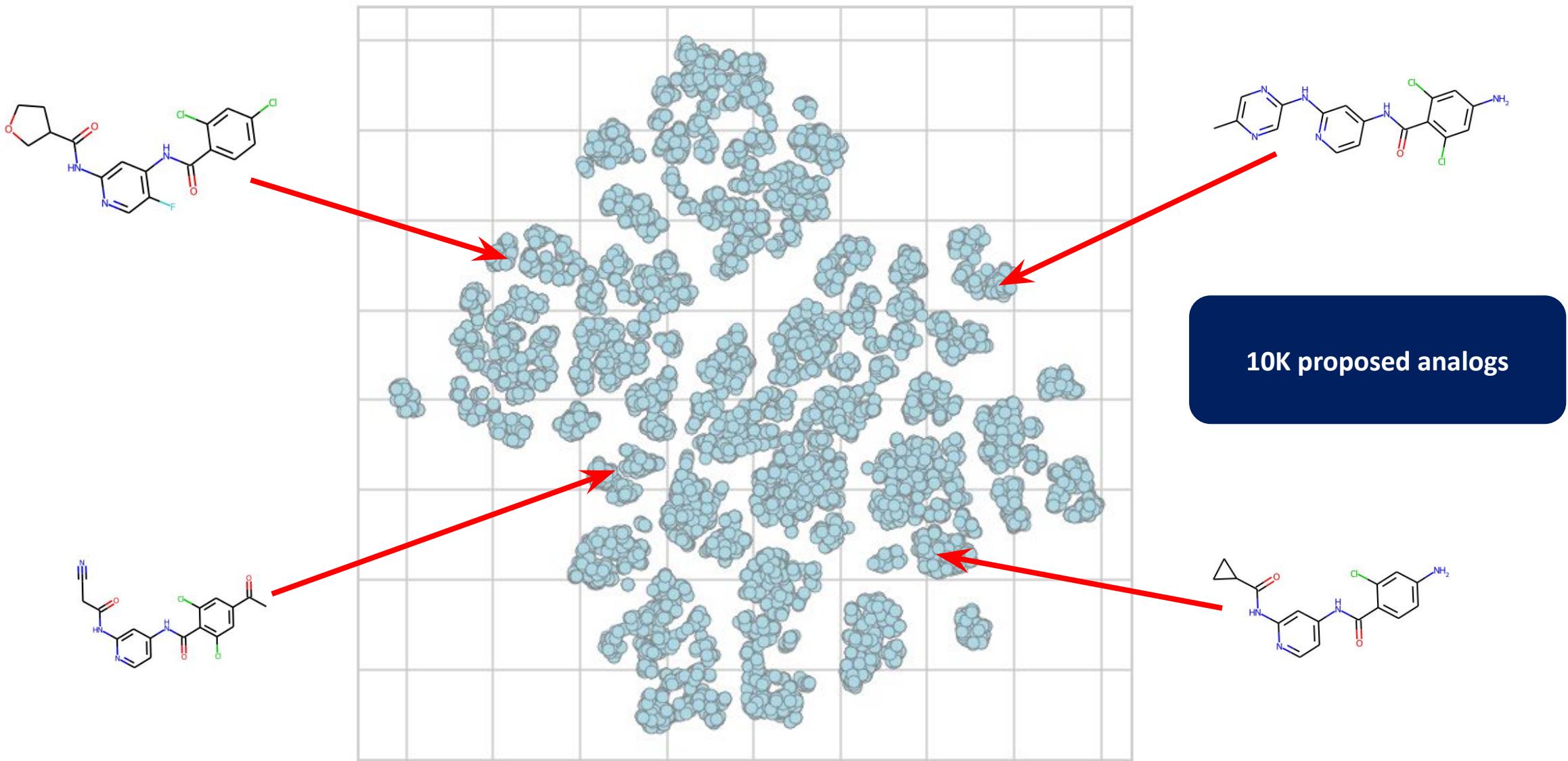
Co-Folding Generates Hypotheses to Bridge Experiments



Active Learning



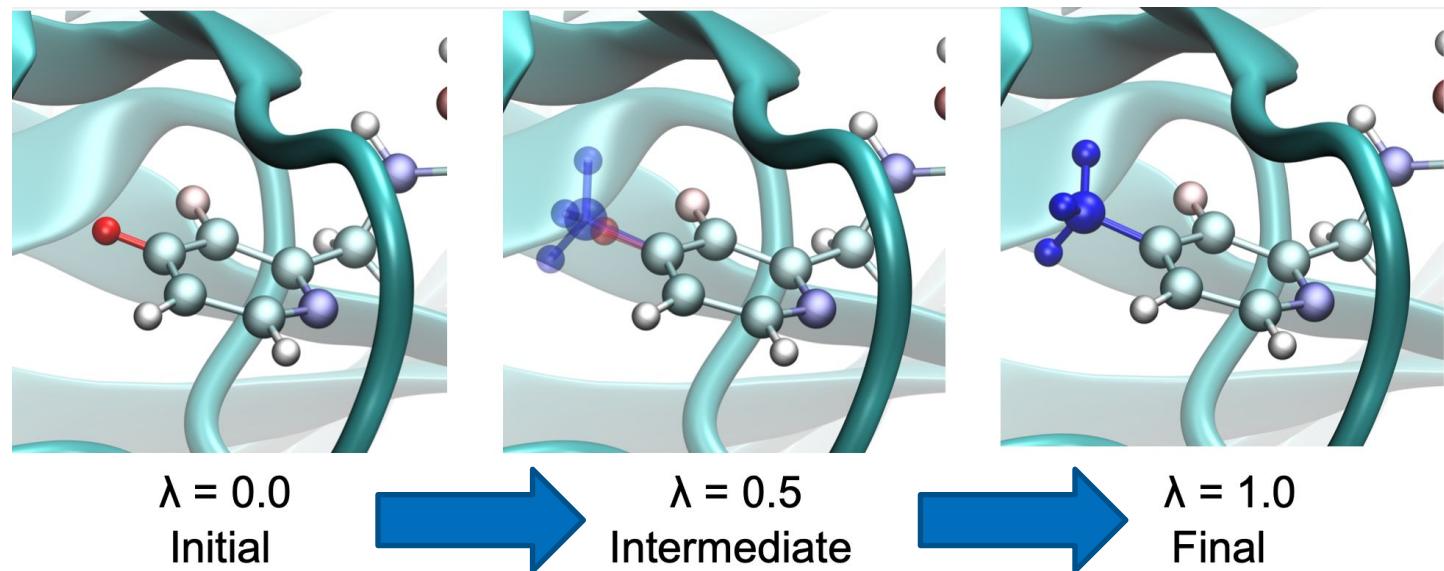
Prioritizing Molecules For Synthesis



Free Energy Perturbation (FEP): Transformation

FEP evaluates binding energy by morphing a known molecule into a new molecule

1. Alchemically morph one molecule to another
2. Run simulations using Molecular Dynamics
3. Compute free energy of transforming the molecule



Calculations typically take 4-8 hrs/molecule
10K calculations would require ~9 GPU years

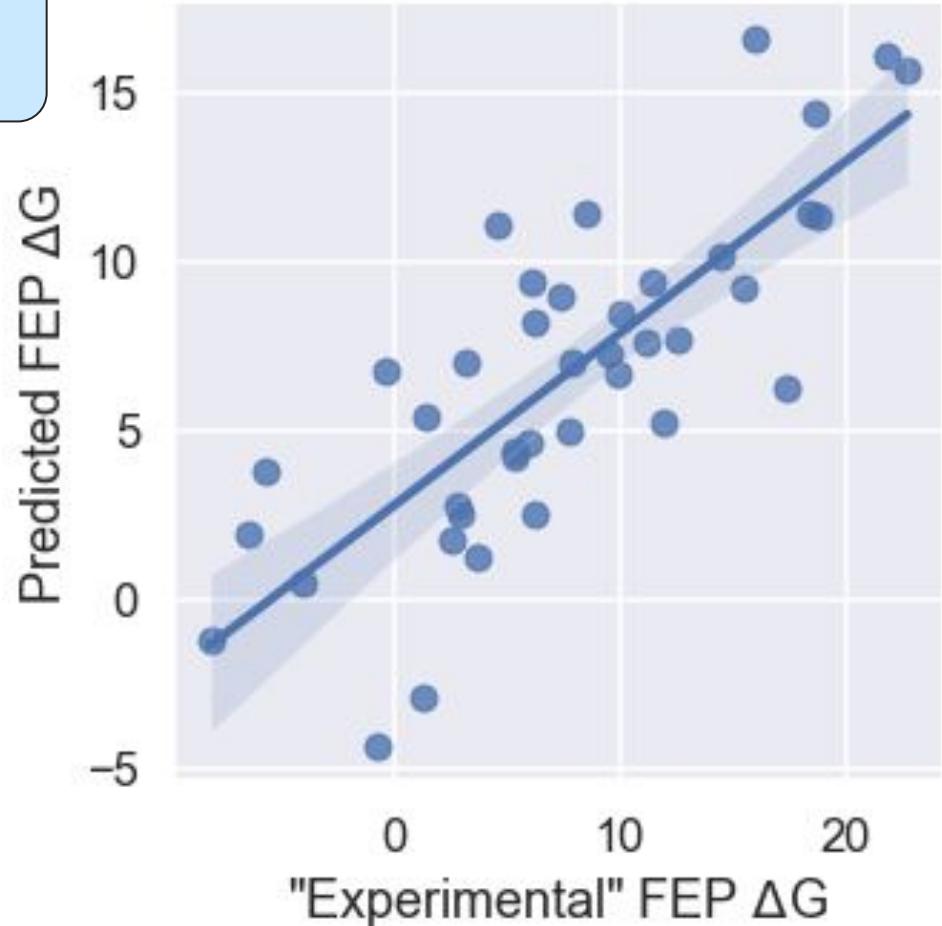
Build a Machine Learning Model to Predict FEP ΔG

Time to process 10K molecules

Method	Time (sec)
FEP	2.9×10^8
Machine Learning Model	7.5

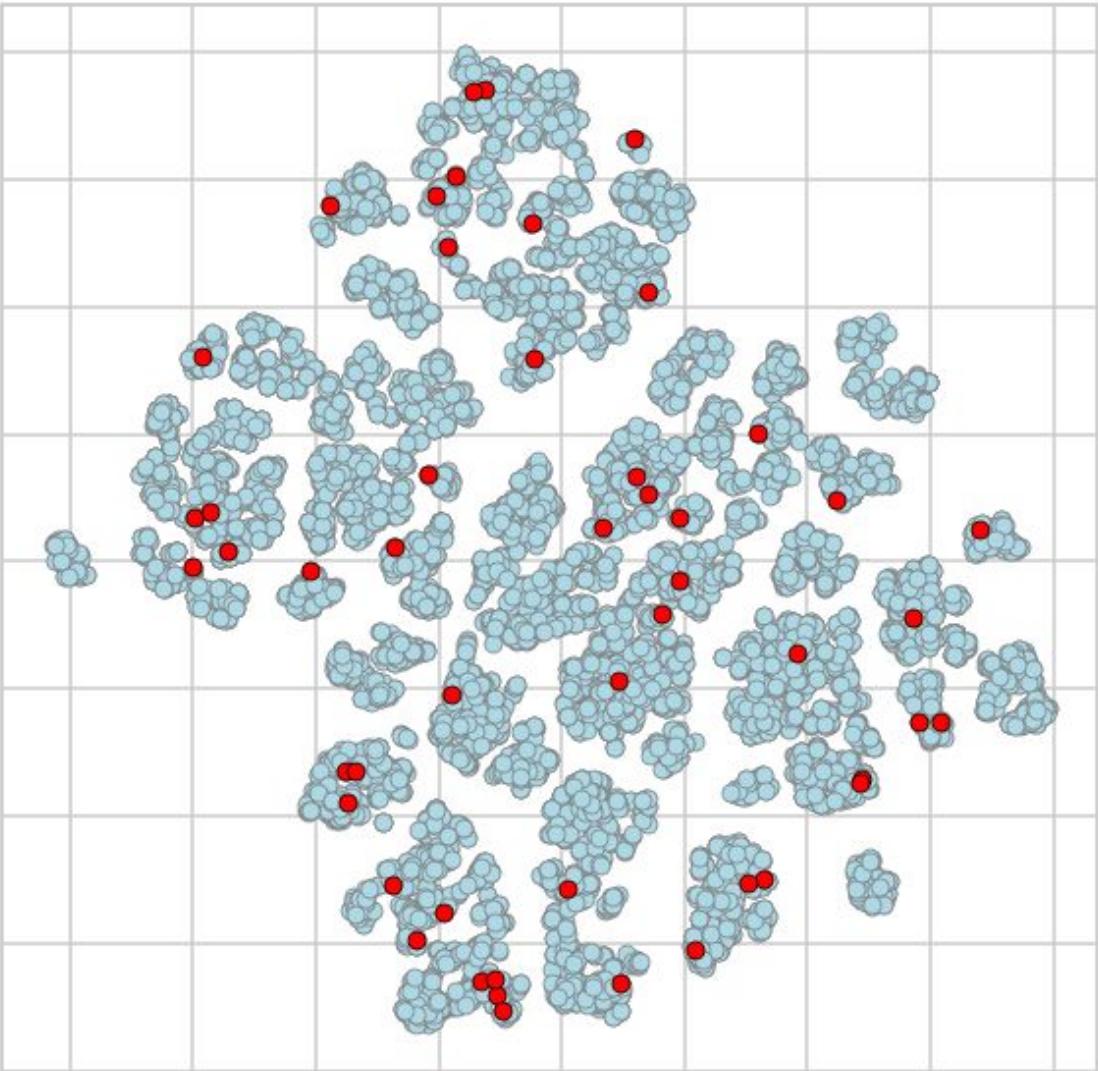
9 GPU years

Machine learning model is 39 million times faster than FEP*



*Training on 100 molecules + inference on 10K molecules

Sample 50 of 10K molecules and Run FEP



FEP



	SMILES	dG bind
112	COc1c(F)cccc1C(=O)Nc1cc(NC(=O)C2COC2)ncc1F	-12.83
275	CN(C)c1nccc(Nc2cc(NC(=O)c3c(F)ccc(F)c3Cl)ccn2)n1	-10.16
563	CCNC(=O)Nc1cc(NC(=O)c2cccc(O)c2Cl)ccn1	-7.68
589	O=C(CC1CC1)Nc1cc(NC(=O)c2cc(O)ccc2Cl)c(F)cn1	-7.51
1170	COC(=O)c1cccc(Nc2cc(NC(=O)c3cccc3Cl)c(F)cn2)n1	-4.58
1342	CNC(=O)Nc1cc(NC(=O)c2c(F)cccc2OC)ccn1	-3.84
1657	CC(C)NC(=O)Nc1cc(NC(=O)c2c(F)ccc(F)c2Cl)ccn1	-2.71
1664	Cc1cccc(Cl)c1C(=O)Nc1cc(Nc2ccn2)ncc1F	-2.69
1974	O=C(Nc1cc(NC(=O)C2CCCC2)nc1F)c1c(F)cccc1Br	-1.80
2071	O=C(Nc1cc(NC(=O)c2c(Cl)cccc2Br)ccn1)NC1CC1	-1.51
2372	Cc1cc(Cl)c(C(=O)Nc2ccnc(Nc3cc(C)nc(CO)n3)c2)c(...	-0.74
2542	Cc1cc(Nc2cc(NC(=O)c3cccc3C)ccn2)nc(NC2CC2)n1	-0.32
2792	CSCC(=O)Nc1cc(NC(=O)c2c(Cl)cc(N)cc2Cl)c(F)cn1	0.31
2815	COC(=O)Nc1cc(NC(=O)c2cccc(F)c2Cl)c(F)cn1	0.37
2880	COc1ccc(C(=O)Nc2cc(NC(=O)C3CC3(F)F)nc2F)c(Cl)c1	0.54
3265	COc1ccc(Cl)c(C(=O)Nc2ccnc(NC(=O)C3CC3CO)c2)c1	1.45
3355	O=C(Nc1ccnc(NC(=O)C2CC2(F)F)c1)c1ccc(O)cc1F	1.65
3568	CC1CC1C(=O)Nc1cc(NC(=O)c2c(F)cc(F)cc2F)ccn1	2.19
3829	O=C(Nc1cc(Nc2ccn2)nc1F)c1cccc(Cl)c1	2.76
3857	COc1ccc(F)c(C(=O)Nc2cc(NC(=O)C3CC3C)nc2F)c1	2.82
4644	COc1cccc(C(=O)Nc2cc(NC(=O)C3CC(=O)C3)nc2F)c1F	4.45
4729	Cc1nc(Cl)cc(Nc2cc(NC(=O)c3ccc(N)cc3)ccn2)n1	4.65
4901	CN(C)C(=O)c1cccc(Nc2cc(NC(=O)c3cccc(N)c3Cl)ccn1)	5.00
4987	COc1ccc(Nc2cc(NC(=O)c3c(Cl)cccc3Br)ccn2)nc1	5.13

etc.

Generate Molecular Descriptors From Chemical Structures

SMILES		Molecular Descriptors																					
	SMILES	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93	94	95	96	97	98	99	
112	COc1c(F)cccc1C(=O)Nc1cc(NC(=O)C2COC2)ncc1F	2.266	-2.240	2.246	-2.381	6.062	-0.133	19.144	10.089	1.784	855.897	...	23.320	0.0	4.984	5.918	0.000	30.957	47.660	0.0	5.75	15.370	
275	CN(C)c1ncnc(Nc2cc(NC(=O)c3c(F)ccc(F)c3Cl)ccn2)n1	2.192	-2.121	2.244	-2.152	6.342	0.102	35.496	10.155	1.797	1030.708	...	40.780	0.0	14.952	0.000	0.000	29.629	64.946	0.0	0.00	15.533	
563	CCNC(=O)Nc1cc(NC(=O)c2cccc(O)c2Cl)ccn1	2.133	-2.093	2.269	-2.281	6.354	0.102	35.496	10.165	2.122	736.412	...	35.044	0.0	10.301	0.000	6.924	17.178	47.115	0.0	5.75	15.950	
589	O=C(CC1CC1)Nc1cc(NC(=O)c2cc(O)ccc2Cl)c(F)cn1	2.243	-2.100	2.322	-2.159	6.341	-0.116	35.496	10.163	1.742	839.883	...	34.921	0.0	4.984	5.918	19.262	10.634	46.866	0.0	5.75	10.634	
1170	COC(=O)c1cccc(Nc2cc(NC(=O)c3cccc3Cl)c(F)cn2)n1	2.139	-2.103	2.221	-2.138	6.340	0.059	35.496	10.177	1.811	1045.192	...	40.801	0.0	9.968	0.000	0.000	17.743	76.825	0.0	0.00	10.634	
1342	CNC(=O)Nc1cc(NC(=O)c2c(F)cccc2OC)ccn1	2.175	-2.107	2.225	-2.246	6.062	0.102	19.142	10.142	2.210	736.412	...	23.444	0.0	10.301	0.000	0.000	24.791	47.909	0.0	5.75	20.687	
1657	CC(C)NC(=O)Nc1cc(NC(=O)c2c(F)ccc(F)c2Cl)ccn1	2.193	-2.112	2.239	-2.339	6.342	0.102	35.496	10.155	2.159	814.717	...	35.044	0.0	10.301	0.000	19.889	10.634	52.683	0.0	0.00	15.950	
1664	Cc1cccc(Cl)c1C(=O)Nc1cc(Nc2ccn2)nc1F	2.155	-2.112	2.235	-2.151	6.342	0.102	35.496	10.087	1.847	899.917	...	34.831	0.0	14.952	0.000	6.924	10.634	71.020	0.0	0.00	10.634	
1974	O=C(Nc1cc(NC(=O)C2CCCO2)ncc1F)c1c(F)cccc1Br	2.241	-2.121	2.233	-2.270	9.103	-0.124	79.919	10.164	1.771	836.892	...	39.250	0.0	4.984	0.000	18.946	17.240	52.133	0.0	0.00	10.634	
2071	O=C(Nc1cc(NC(=O)c2c(Cl)cccc2Br)ccn1)NC1CC1	2.229	-2.101	2.257	-2.303	9.103	0.102	79.919	10.177	1.693	775.147	...	50.974	0.0	10.301	0.000	18.883	10.634	51.587	0.0	0.00	15.950	
2372	Cc1cc(Cl)c(C(=O)Nc2ccncc(Nc3cc(C)nc(CO)n3)c2)c(...	2.158	-2.111	2.274	-2.134	6.401	0.102	35.497	10.124	1.804	1019.578	...	46.432	0.0	14.952	0.000	20.454	10.634	69.219	0.0	0.00	10.634	
2542	Cc1cc(Nc2cc(NC(=O)c3cccc3C)ccn2)nc(NC2CC2)n1	2.218	-2.108	2.243	-2.132	6.049	0.102	16.149	10.091	1.486	1015.283	...	29.179	0.0	14.952	0.000	32.731	15.950	65.482	0.0	0.00	15.950	
2792	CSCC(=O)Nc1cc(NC(=O)c2c(Cl)cc(N)cc2Cl)c(F)cn1	2.178	-2.111	2.288	-2.171	7.988	-0.113	35.497	10.154	2.248	812.169	...	63.971	0.0	4.984	0.000	0.000	28.376	45.822	0.0	0.00	16.367	
2815	COC(=O)Nc1cc(NC(=O)c2cccc(F)c2Cl)c(F)cn1	2.159	-2.095	2.225	-2.159	6.341	0.102	35.496	10.168	2.248	770.928	...	35.107	0.0	4.984	0.000	0.000	17.743	52.683	0.0	0.00	10.634	
2880	COc1ccc(C(=O)Nc2cc(NC(=O)C3CC3(F)F)ncc2F)c(Cl)c1	2.609	-2.104	2.486	-2.189	6.342	-0.119	35.496	10.165	1.745	923.537	...	34.921	0.0	4.984	5.918	12.343	17.743	46.866	0.0	5.75	15.370	
3265	COc1ccc(Cl)c(C(=O)Nc2ccncc(NC(=O)C3CC3CO)c2)c1	2.419	-2.098	2.382	-2.170	6.341	-0.117	35.496	10.163	1.685	842.936	...	34.921	0.0	4.984	11.836	6.421	24.350	47.115	0.0	5.75	15.370	
3355	O=C(Nc1ccncc(NC(=O)C2CC2(F))c1)c1ccc(O)cc1F	2.609	-2.085	2.485	-2.185	6.045	-0.119	19.287	10.157	1.690	857.514	...	23.320	0.0	4.984	5.918	12.343	10.634	47.909	0.0	5.75	10.634	
3568	CC1CC1C(=O)Nc1cc(NC(=O)c2c(F)cc(F)cc2F)ccn1	2.387	-2.090	2.379	-2.176	6.046	-0.117	19.149	10.146	1.737	833.777	...	23.320	0.0	4.984	11.836	13.345	10.634	53.477	0.0	0.00	10.634	
3829	O=C(Nc1cc(Nc2ccn2)ncc1F)c1cccc(Cl)c1	2.113	-2.093	2.204	-2.132	6.306	0.102	35.496	10.179	1.800	875.352	...	34.831	0.0	14.952	0.000	0.000	10.634	71.523	0.0	0.00	10.634	
3857	COc1ccc(F)c(C(=O)Nc2cc(NC(=O)C3CC3C)ncc2F)c1	2.388	-2.102	2.380	-2.175	6.048	-0.117	19.144	10.152	1.773	872.755	...	23.320	0.0	4.984	11.836	13.345	17.743	47.660	0.0	5.75	15.370	



Build a Machine Learning Models to Predict FEP ΔG From Molecular Descriptors

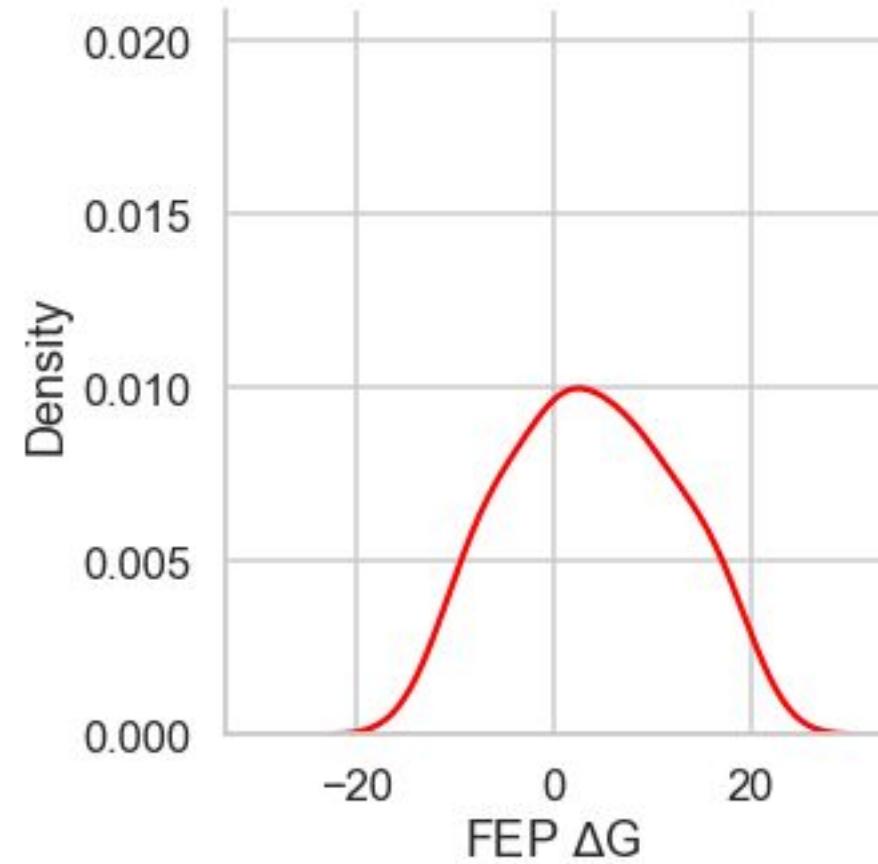
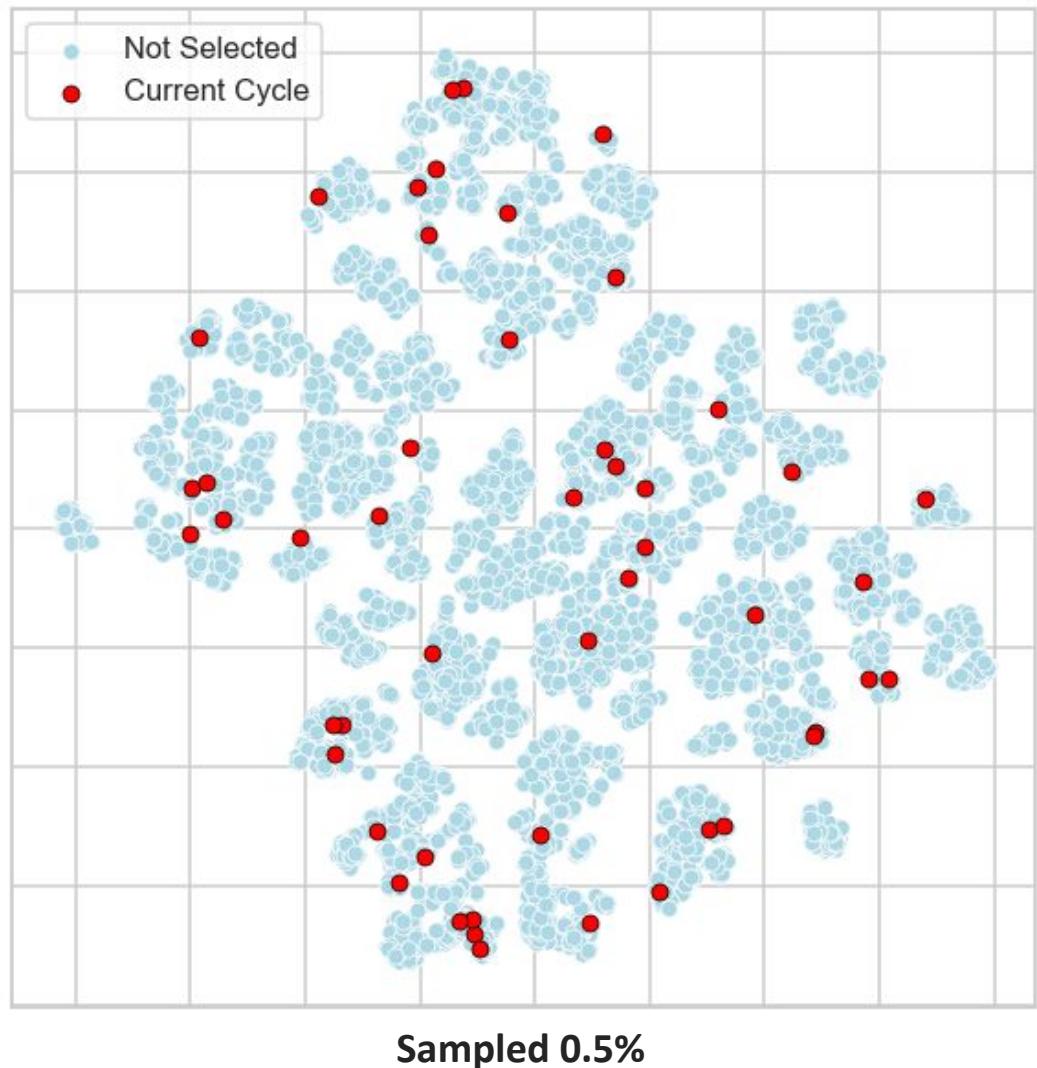
$\Delta G = f(\text{molecular descriptors})$

Predict ΔG for the remaining 9,950 molecules

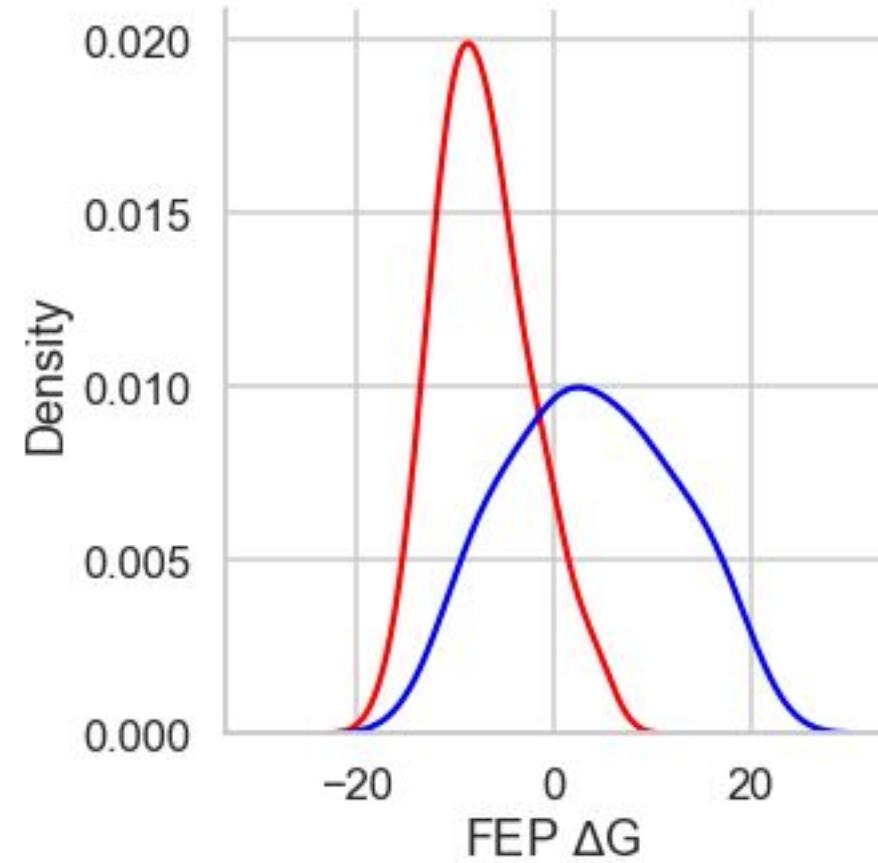
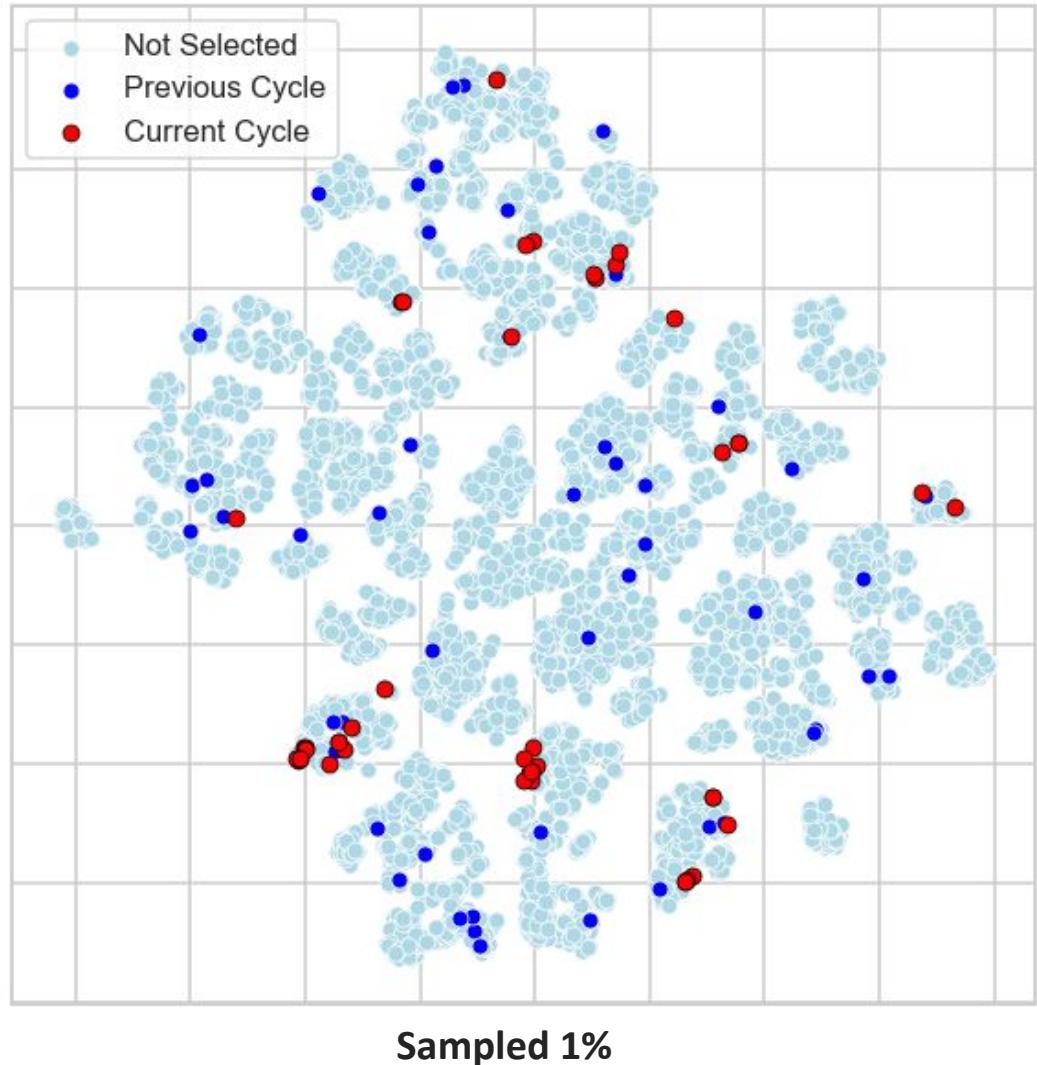
Run FEP on the 50 molecules with the best predicted ΔG

Repeat

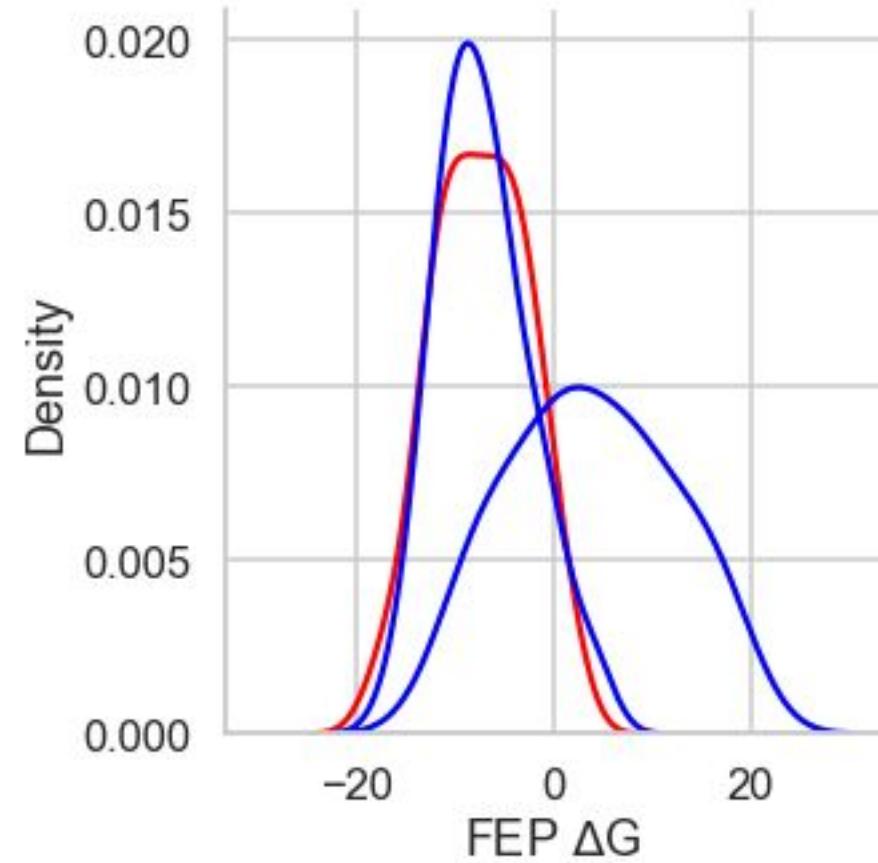
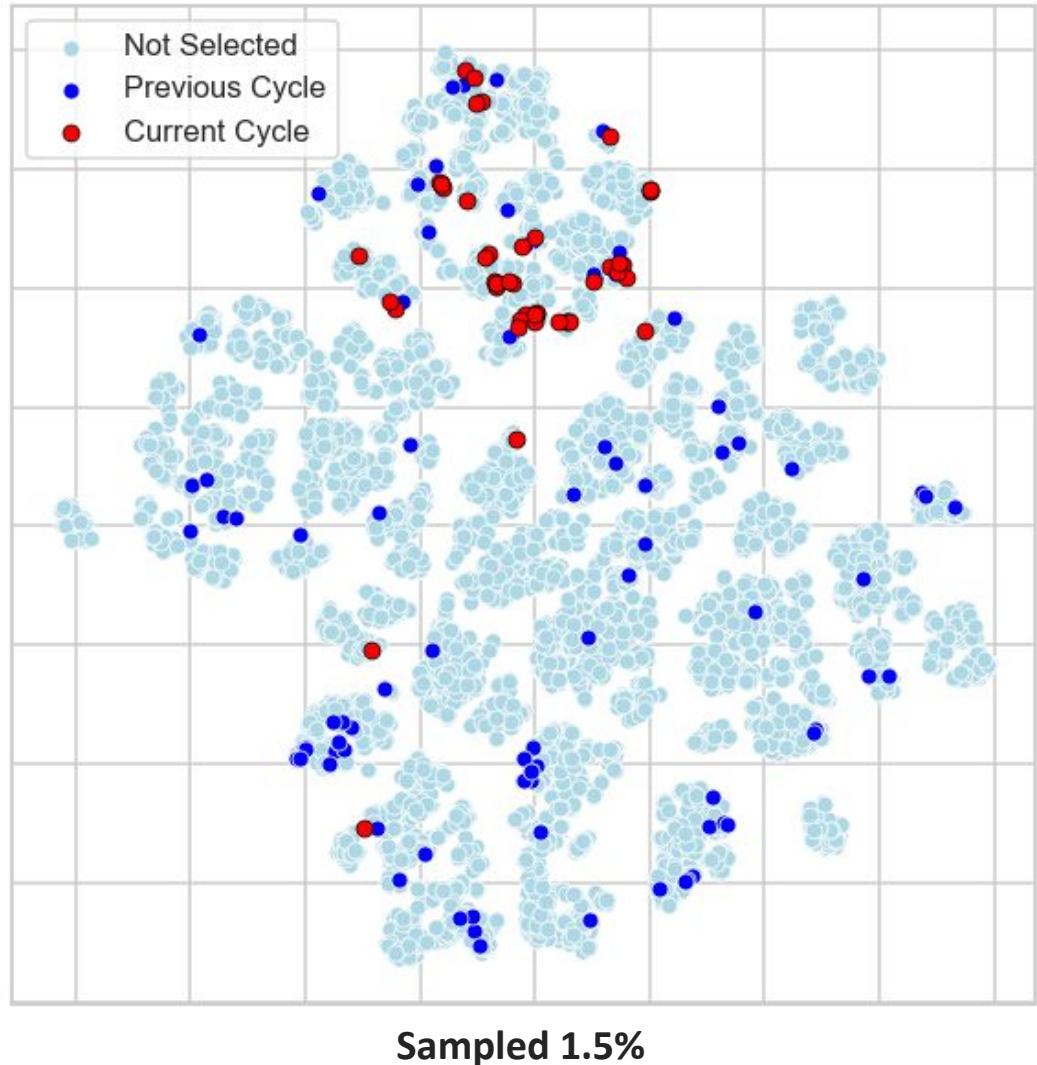
Initial Sample - 1 of the Top 100 Molecules Found



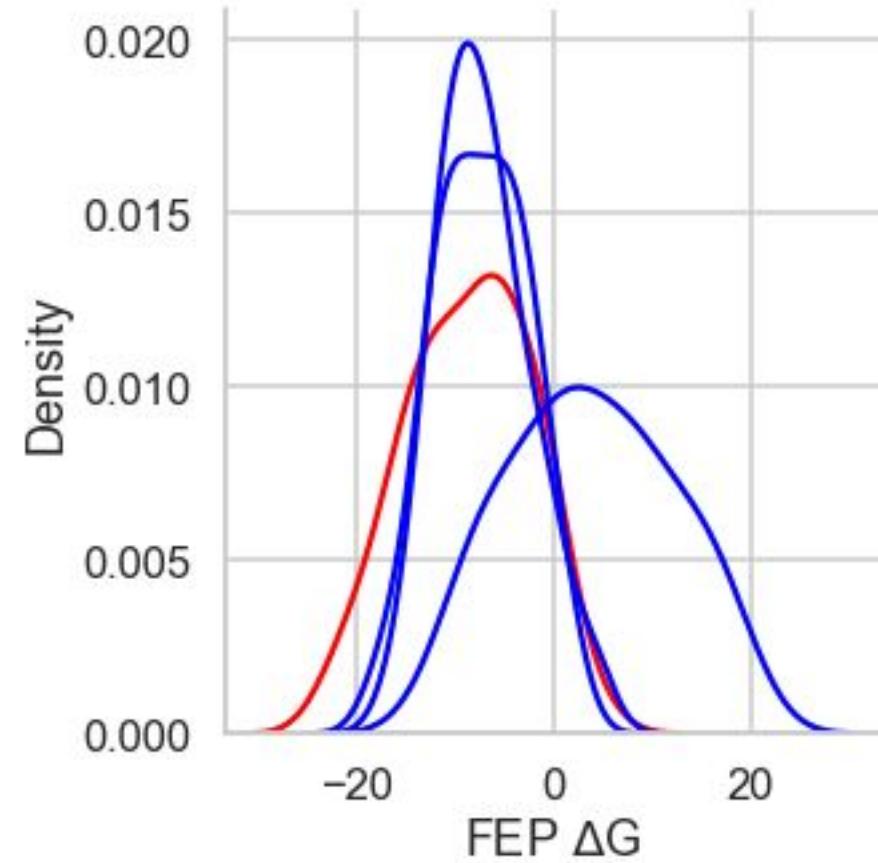
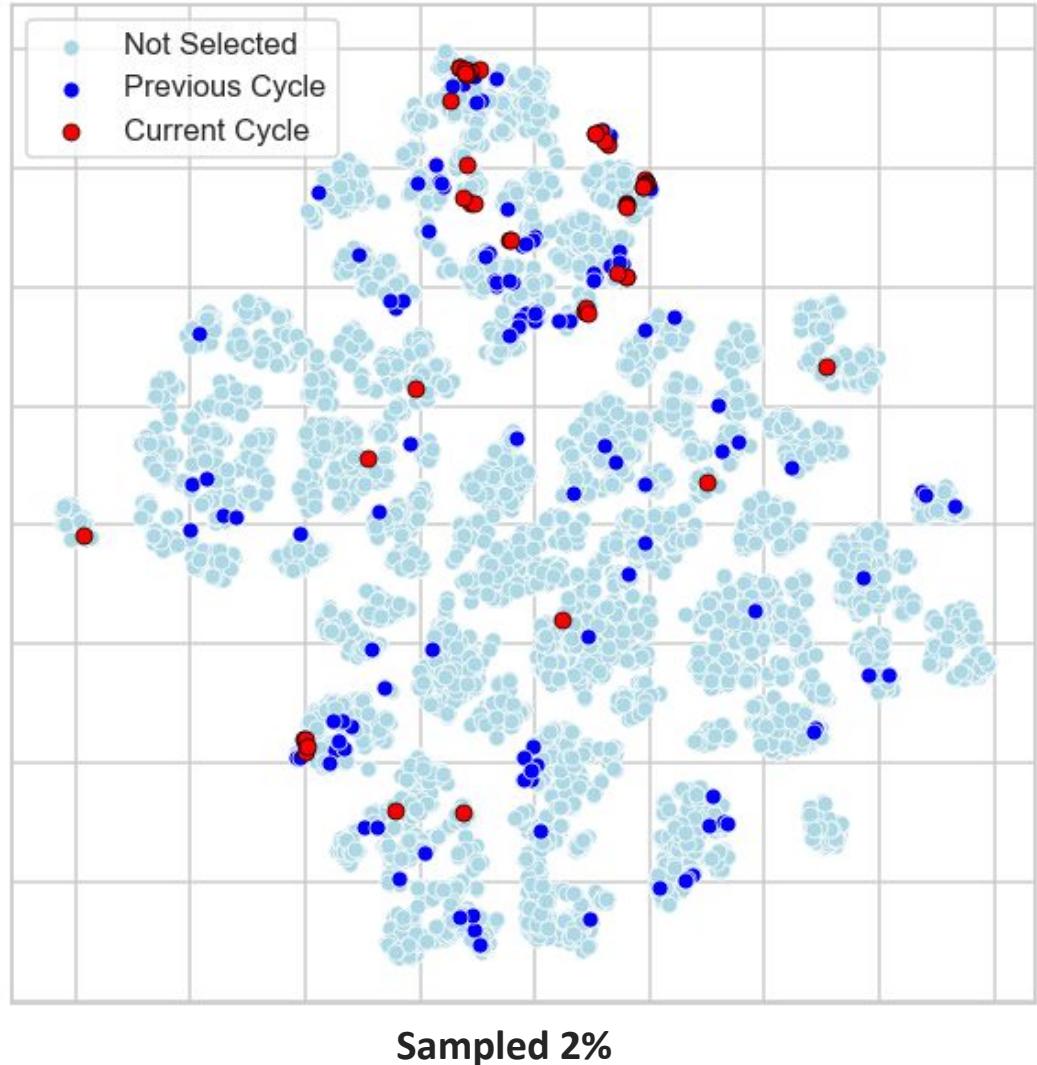
Cycle 1 - 16 of the Top 100 Molecules Found



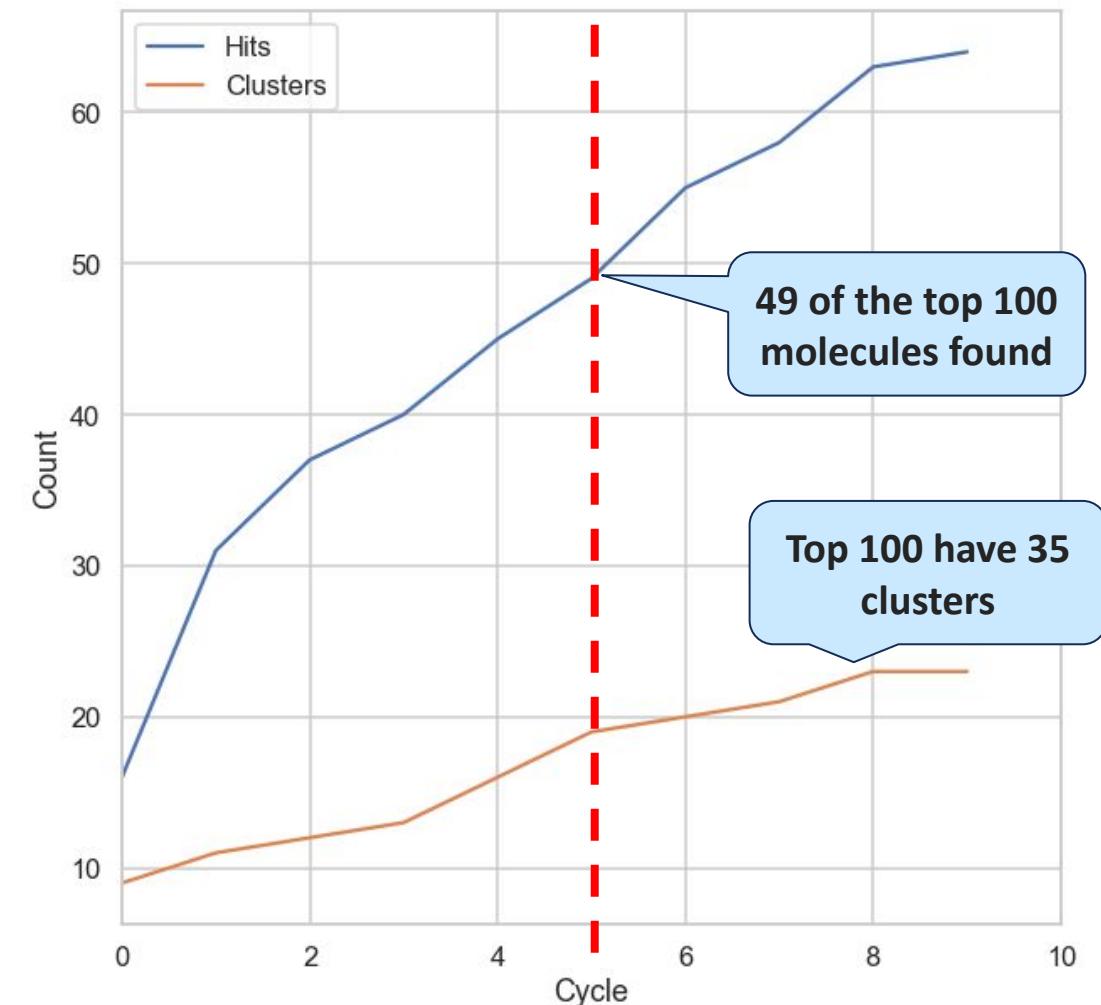
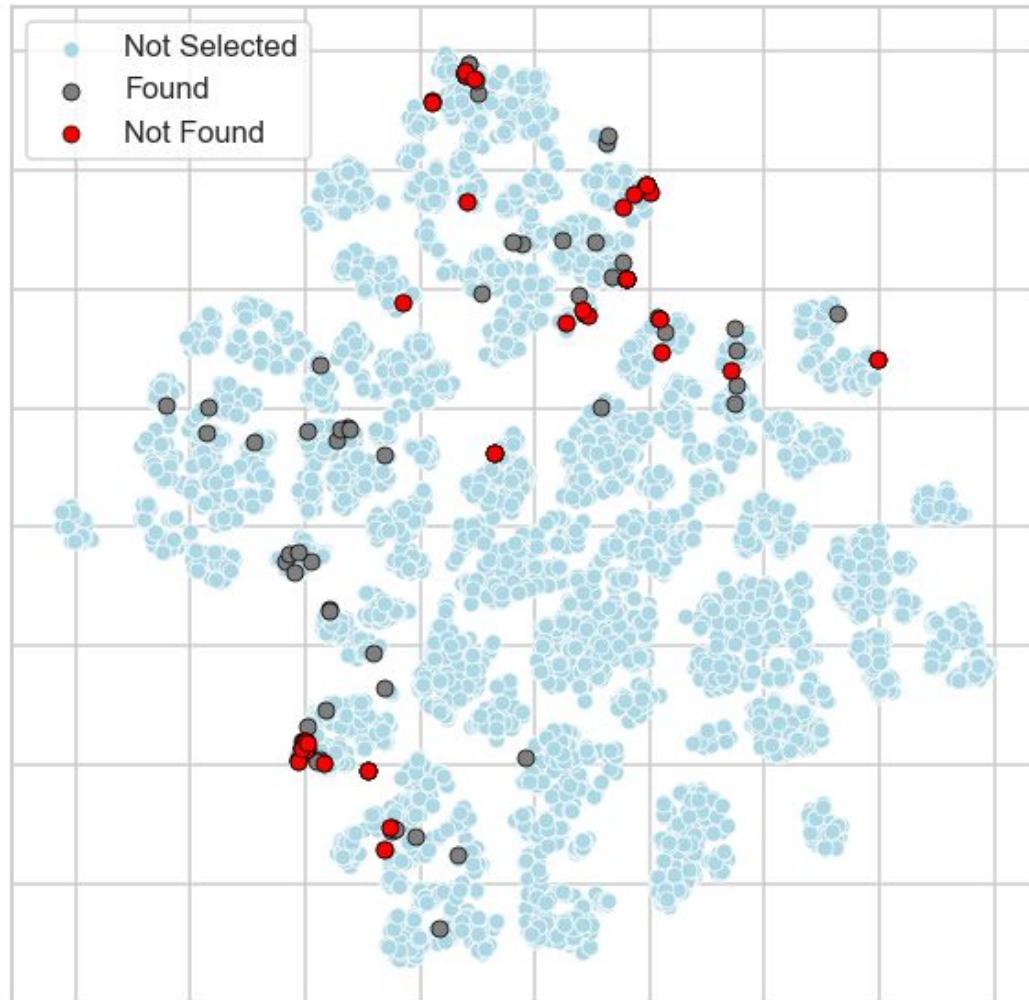
Cycle 2 - 31 of the Top 100 Molecules Found



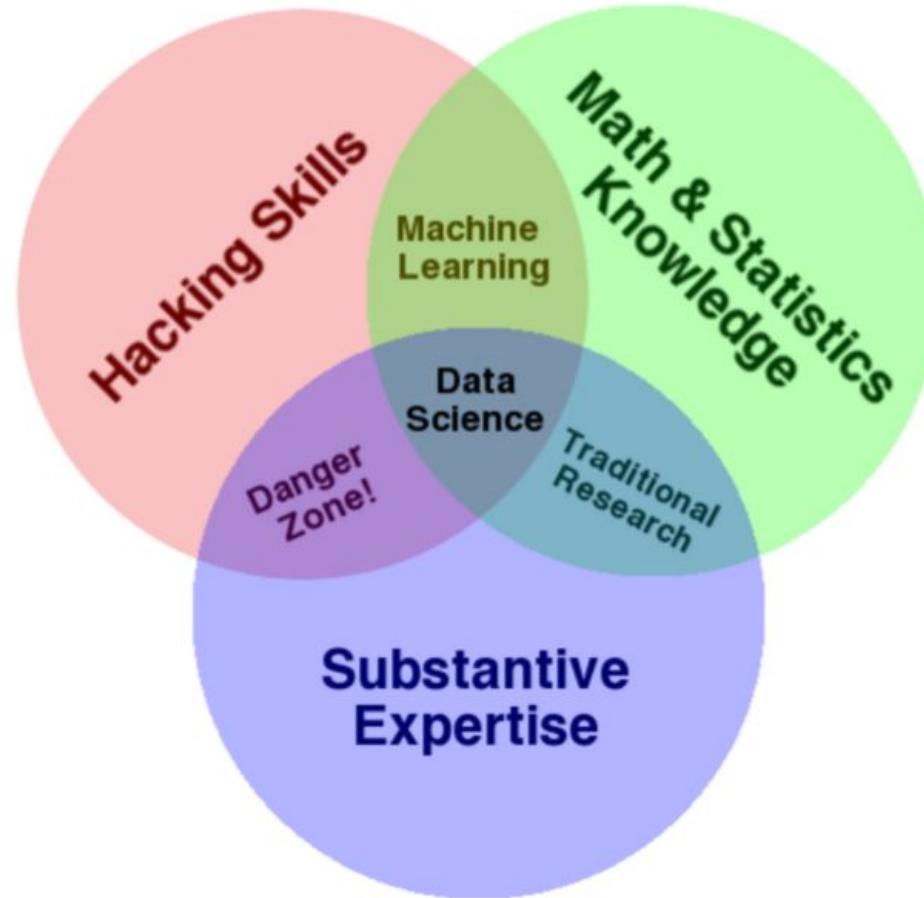
Cycle 3 - 37 of the Top 100 Molecules Found



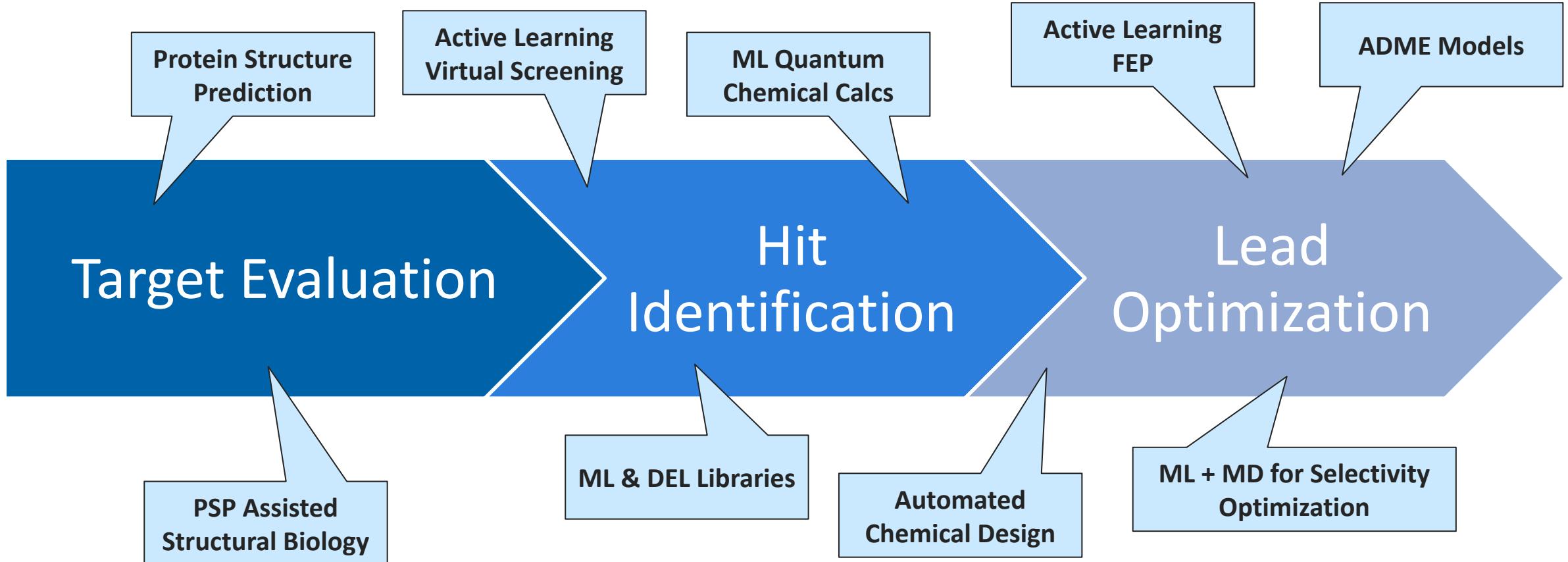
Comparing Top 100 Compounds Found and Not Found in 300 Samples (3%)



What Do We Need To Succeed?



ML Has Impact Across the Drug Discovery Process



ML is One Component in a Collaboration Between Experiment and Computation

