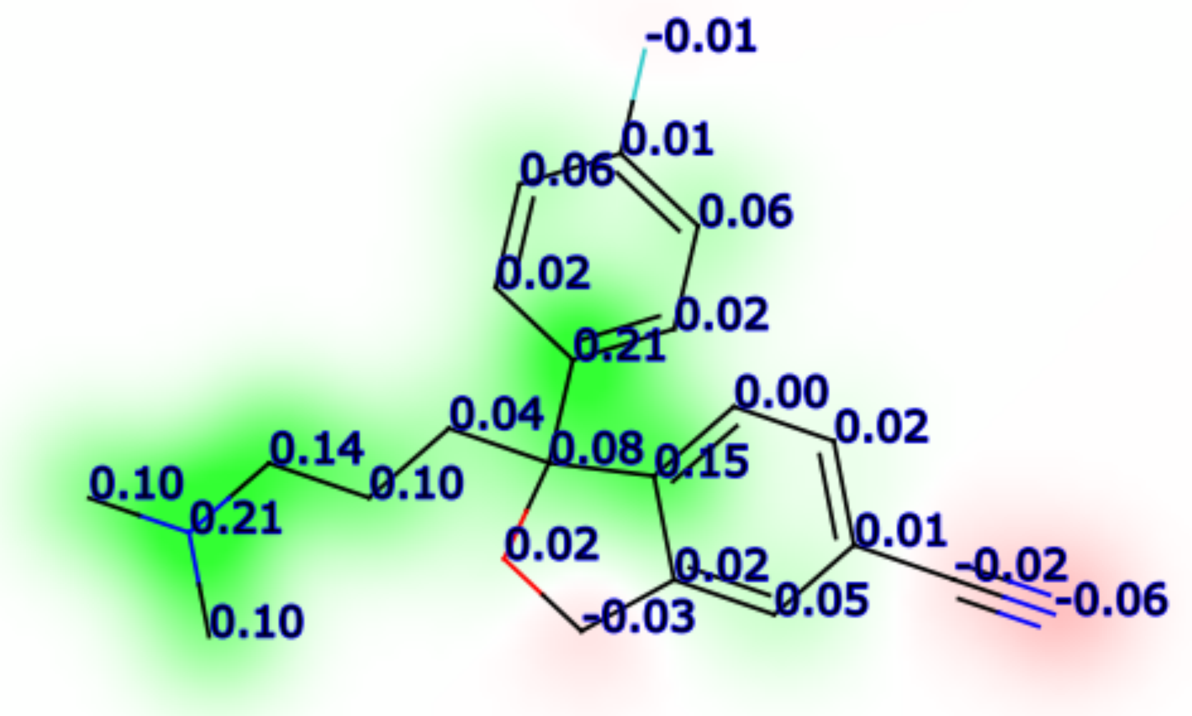
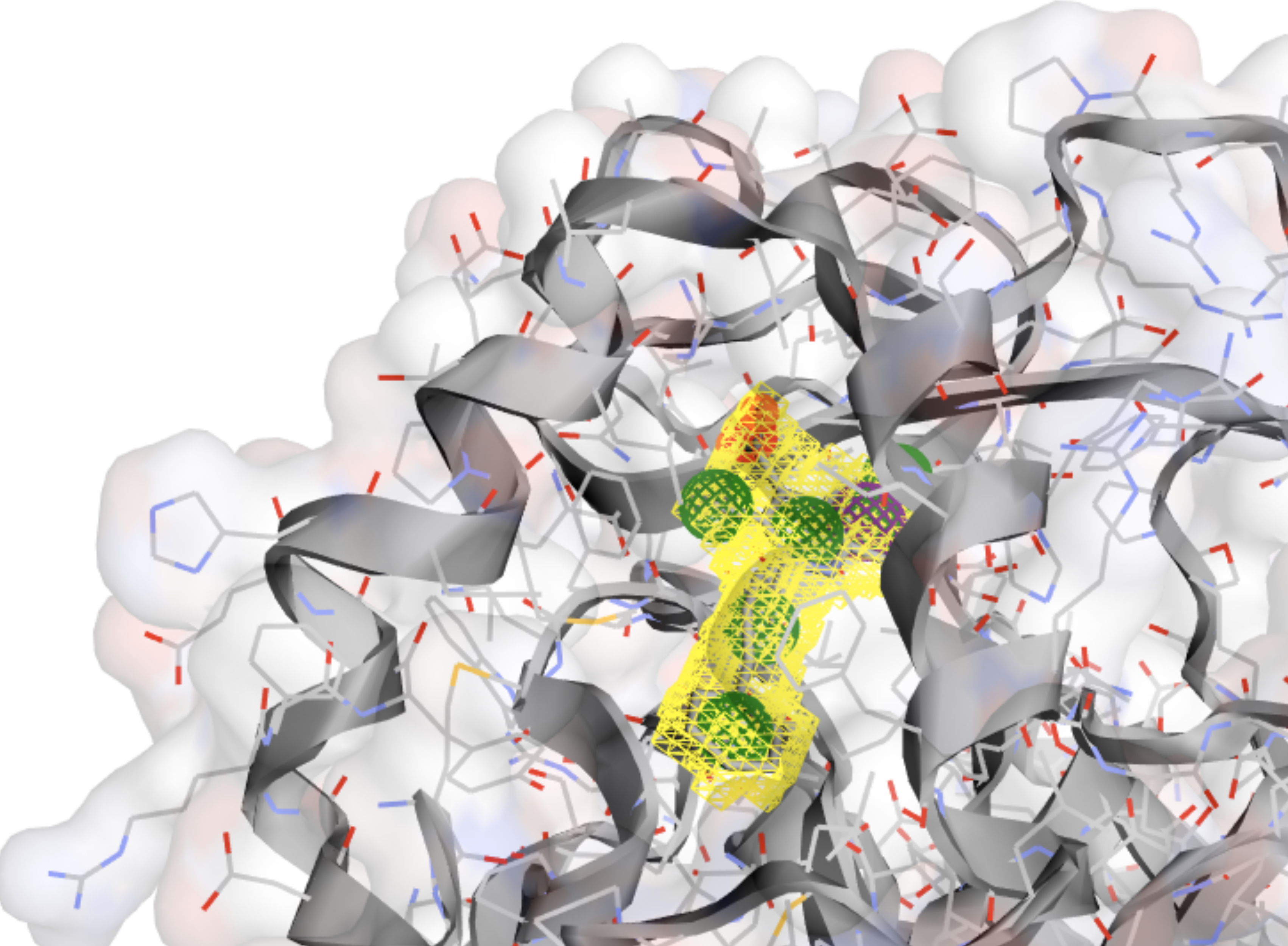
Computational Drug Discovery

David Ryan Koes

2/25/2026

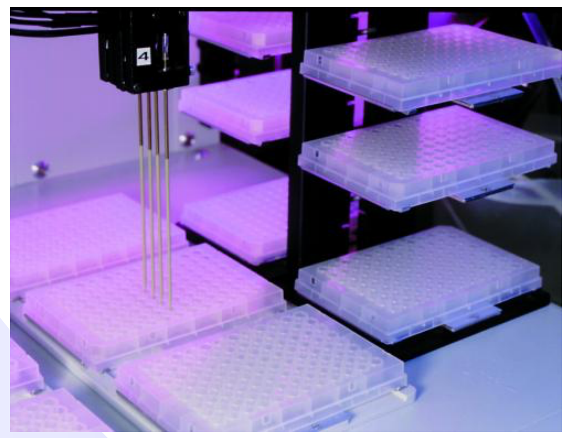


University of Pittsburgh

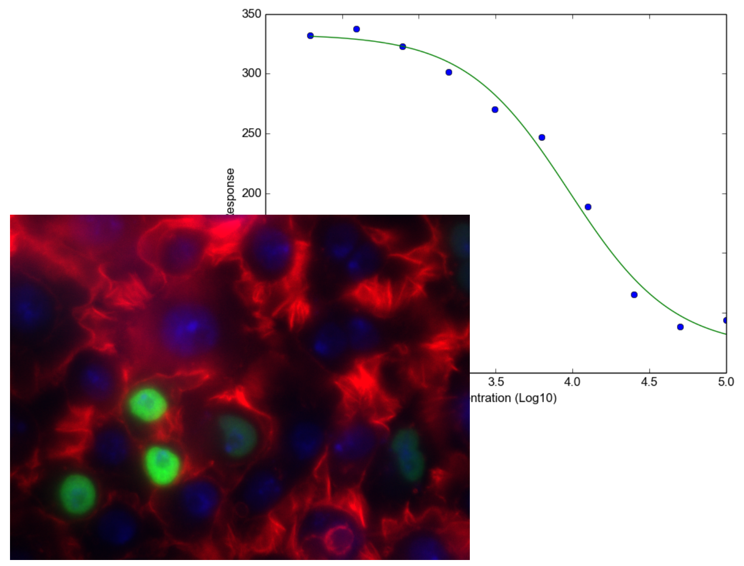


Drug Discovery

Target Identification



Screening



Lead Identification

Lead Optimization

Compounds

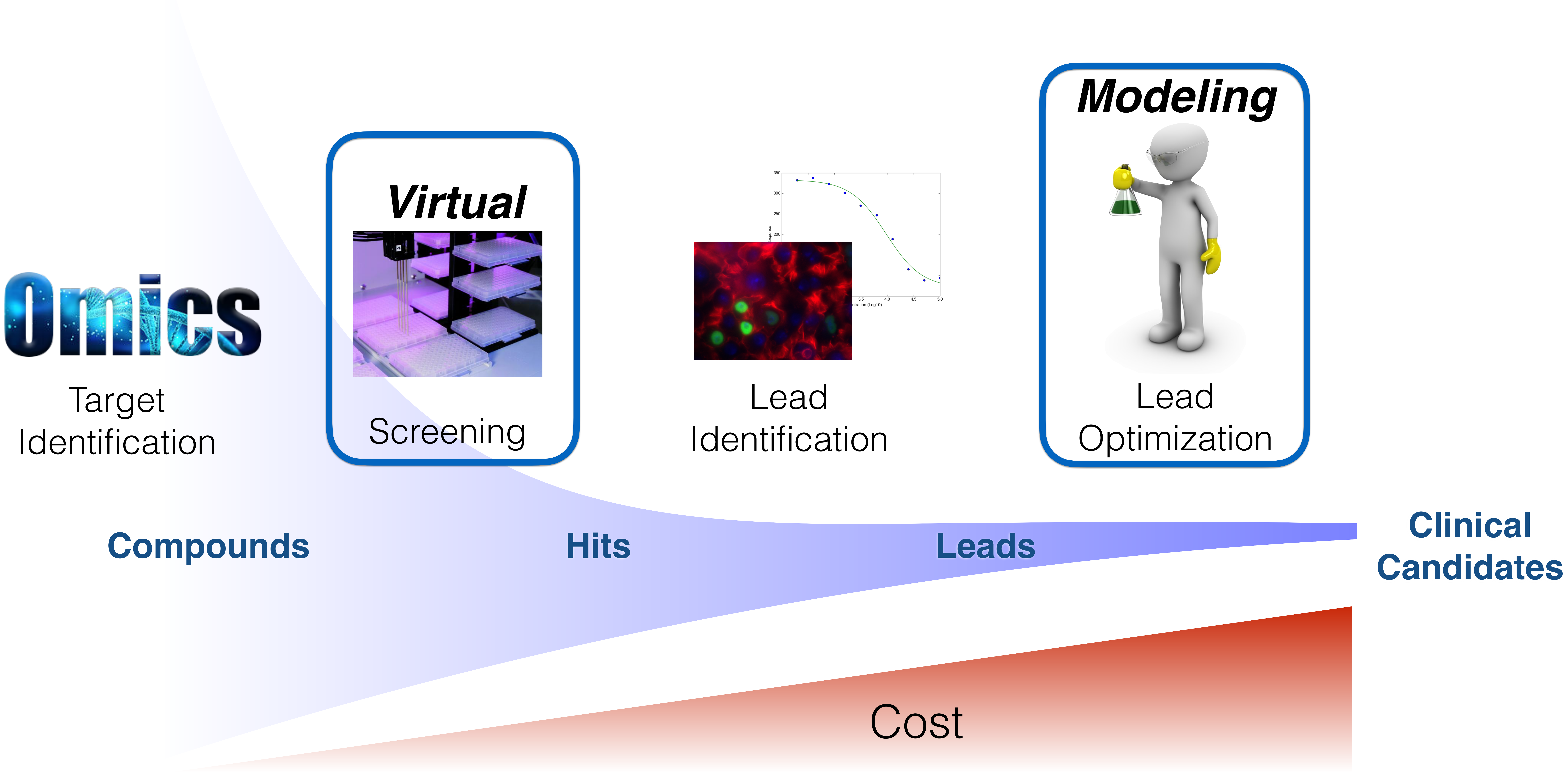
Hits

Leads

Clinical Candidates

Cost

Computational Drug Discovery



Kinds of Virtual Screening

ADMET

Ligand Based

- similarity to known binder
- QSAR
- pharmacophore

Receptor Based

- dock and score
- simulation

ADMET

Absorption

Distribution

Metabolism

Excretion

Toxicity

Will this be a usable drug?

Screening for ADMET:

Cytochrome P450 interaction

Lipinksi's Rule of Five

QSPR: Quantitative Structure Property Relationship

Kinds of Virtual Screening

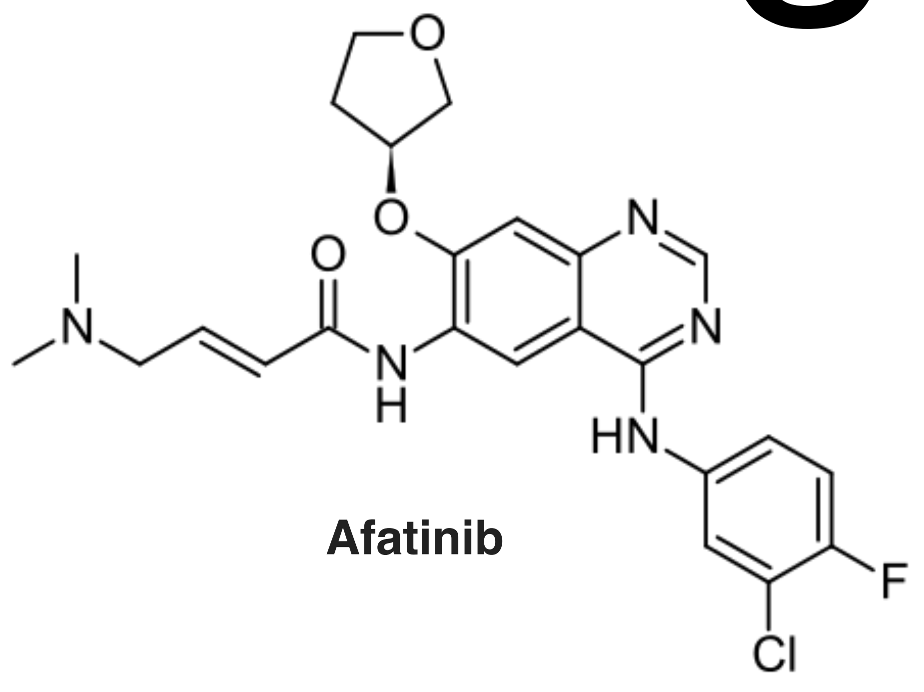
ADMET

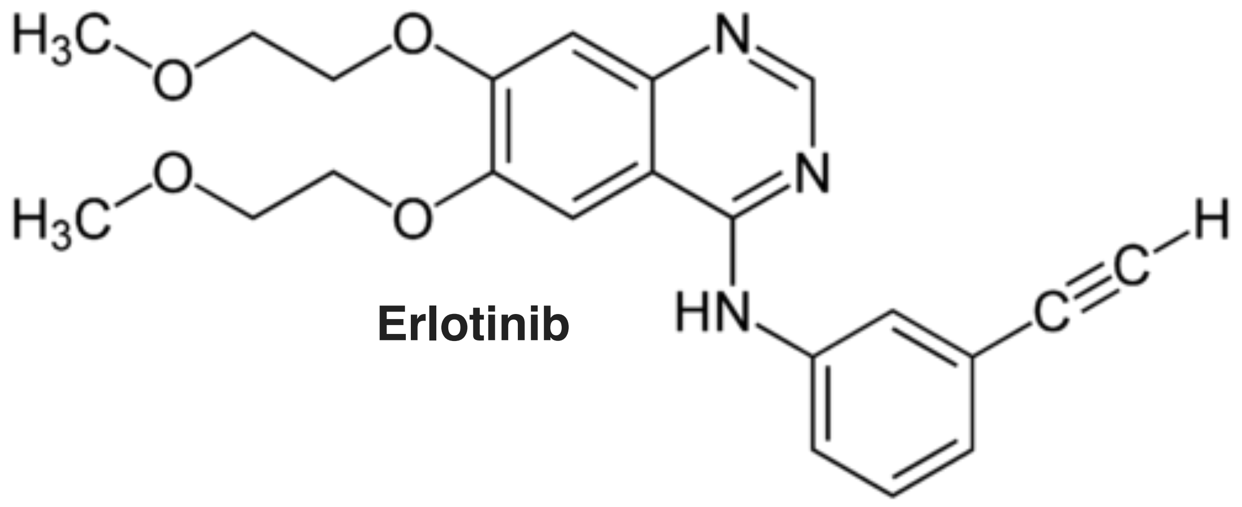
Ligand Based

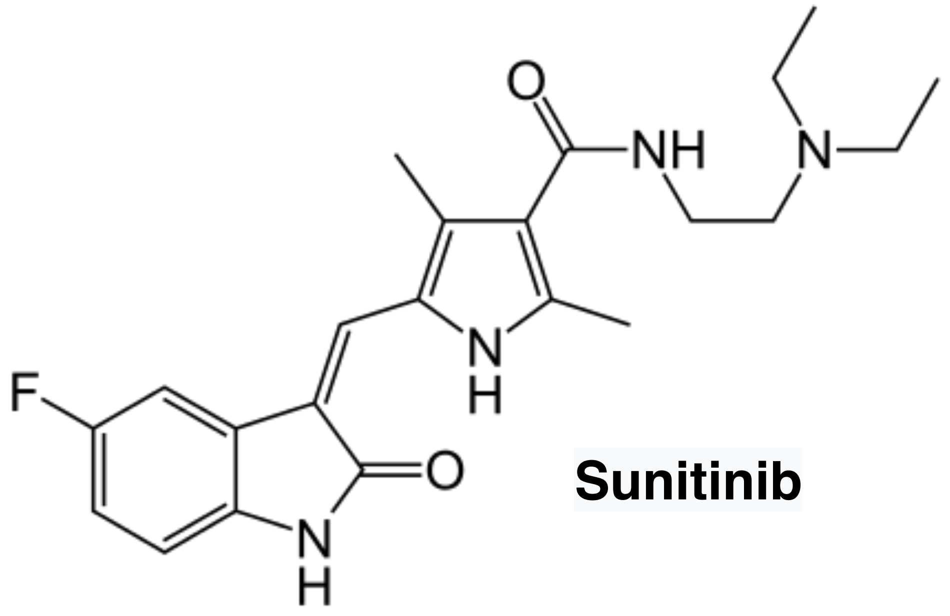
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based

- dock and score


Afatinib


Erlotinib


Sunitinib

Ligand Based: Similarity

Fingerprint Methods

- map molecules to a descriptor space:

  1D: molecule weight, #h-bonds, etc.

  2D: paths, bond distances between atom-pairs

- similarity is "distance" between descriptors

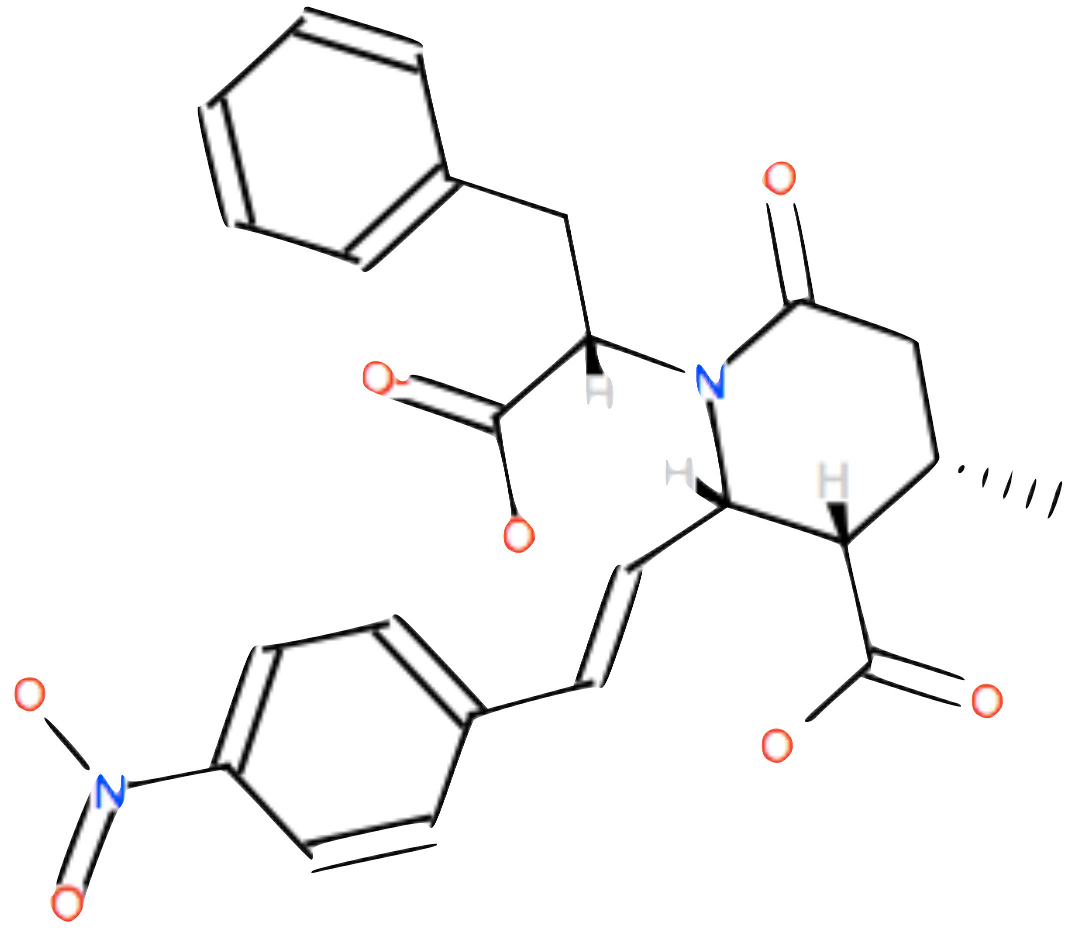
- for bit vectors, Tanimoto distance used

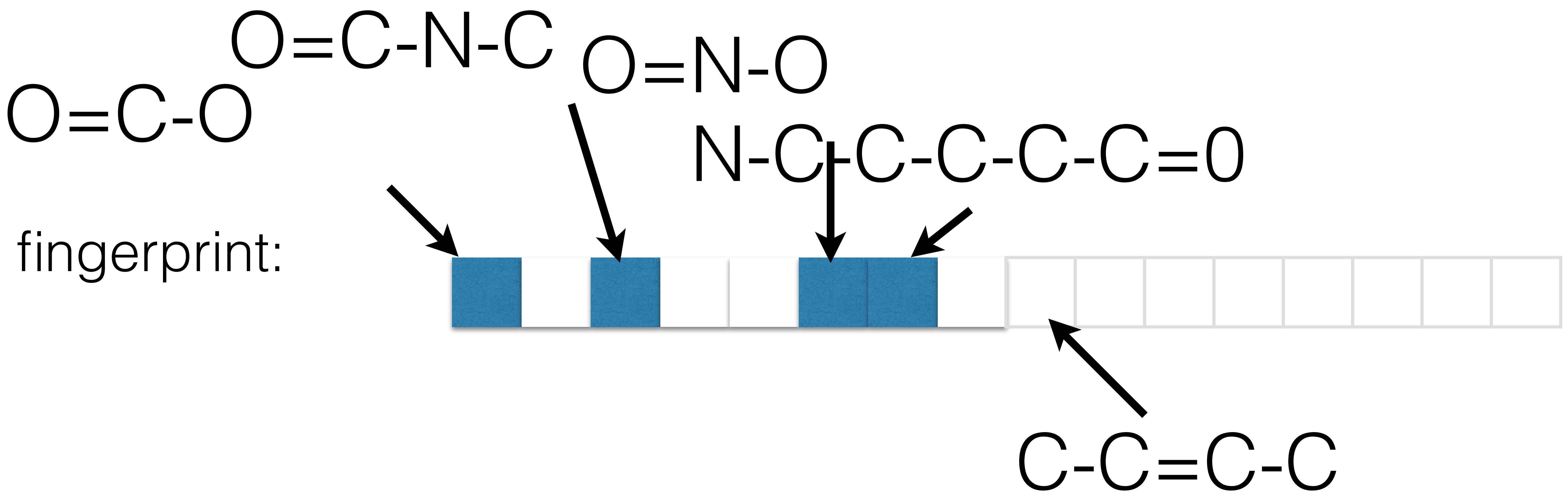
$$T(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Topological Fingerprints

Daylight/FP2 Fingerprints

- all paths up to 7 bonds long
- each path corresponds to bit position (**hashing**)
- fast similarity checking (Tanimoto)

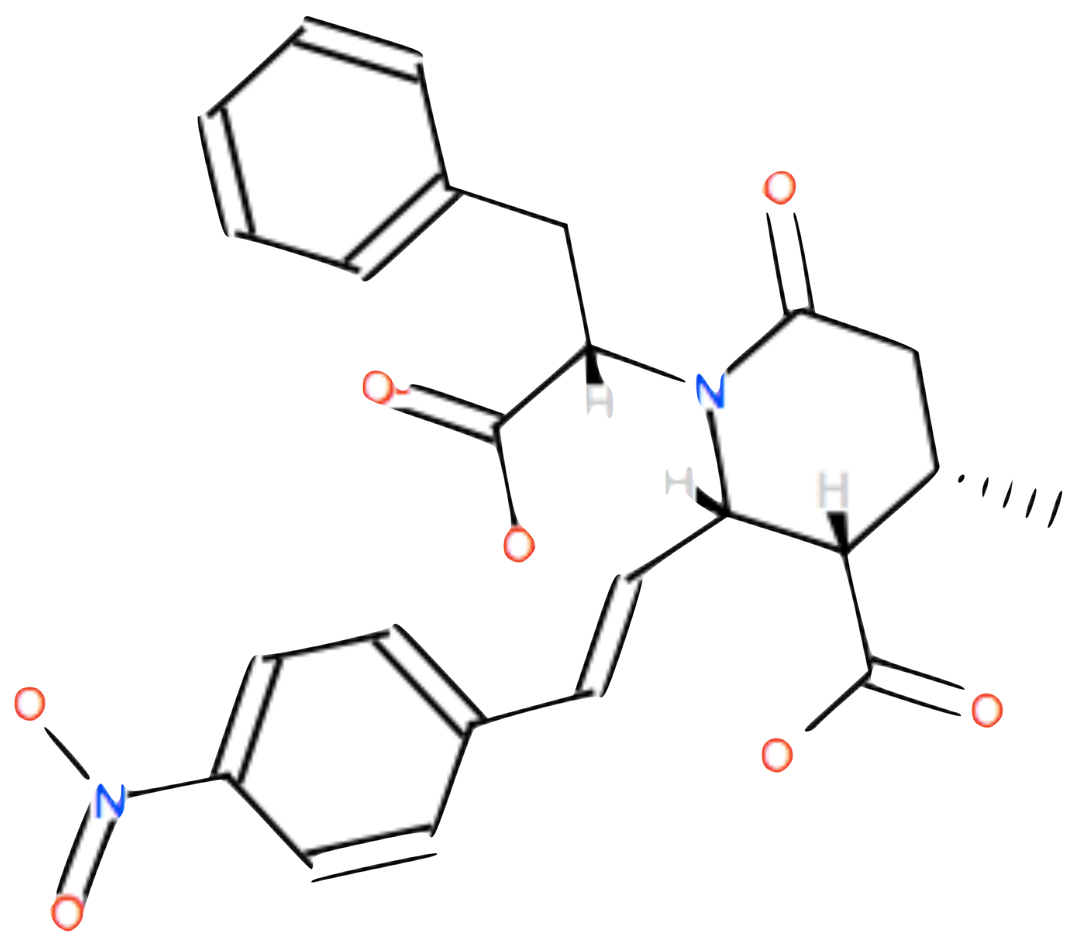


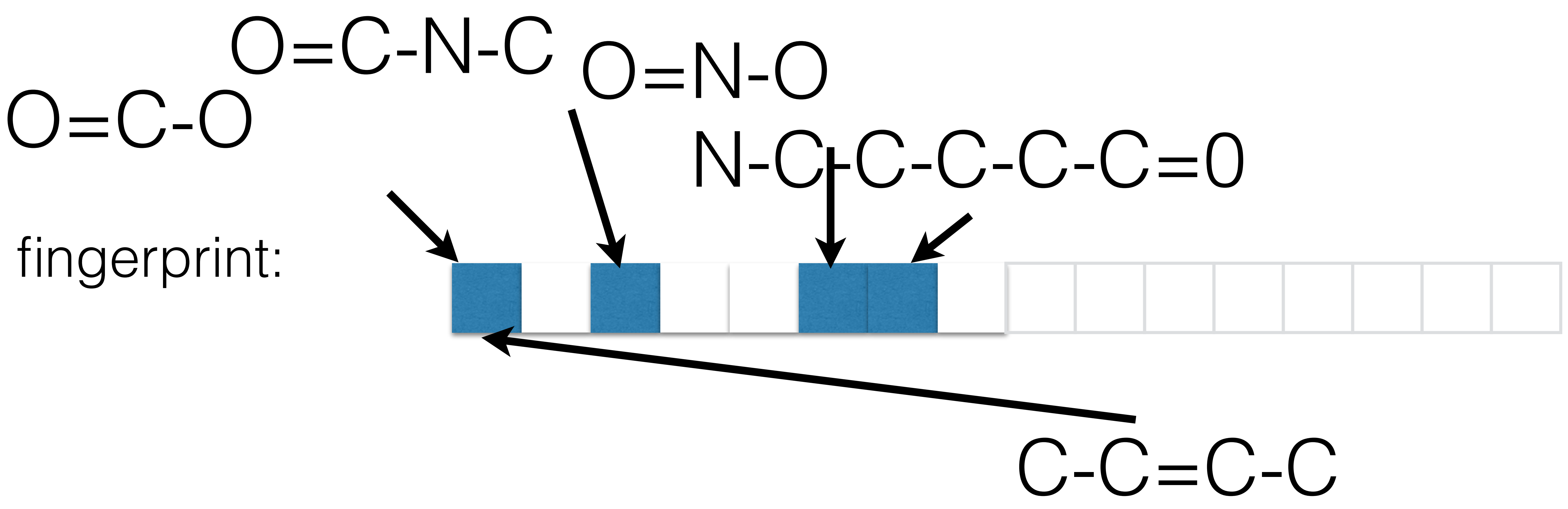


Topological Fingerprints

Daylight/FP2 Fingerprints

- all paths up to 7 bonds long
- each path corresponds to bit position (**hashing**)
- fast similarity checking (Tanimoto)

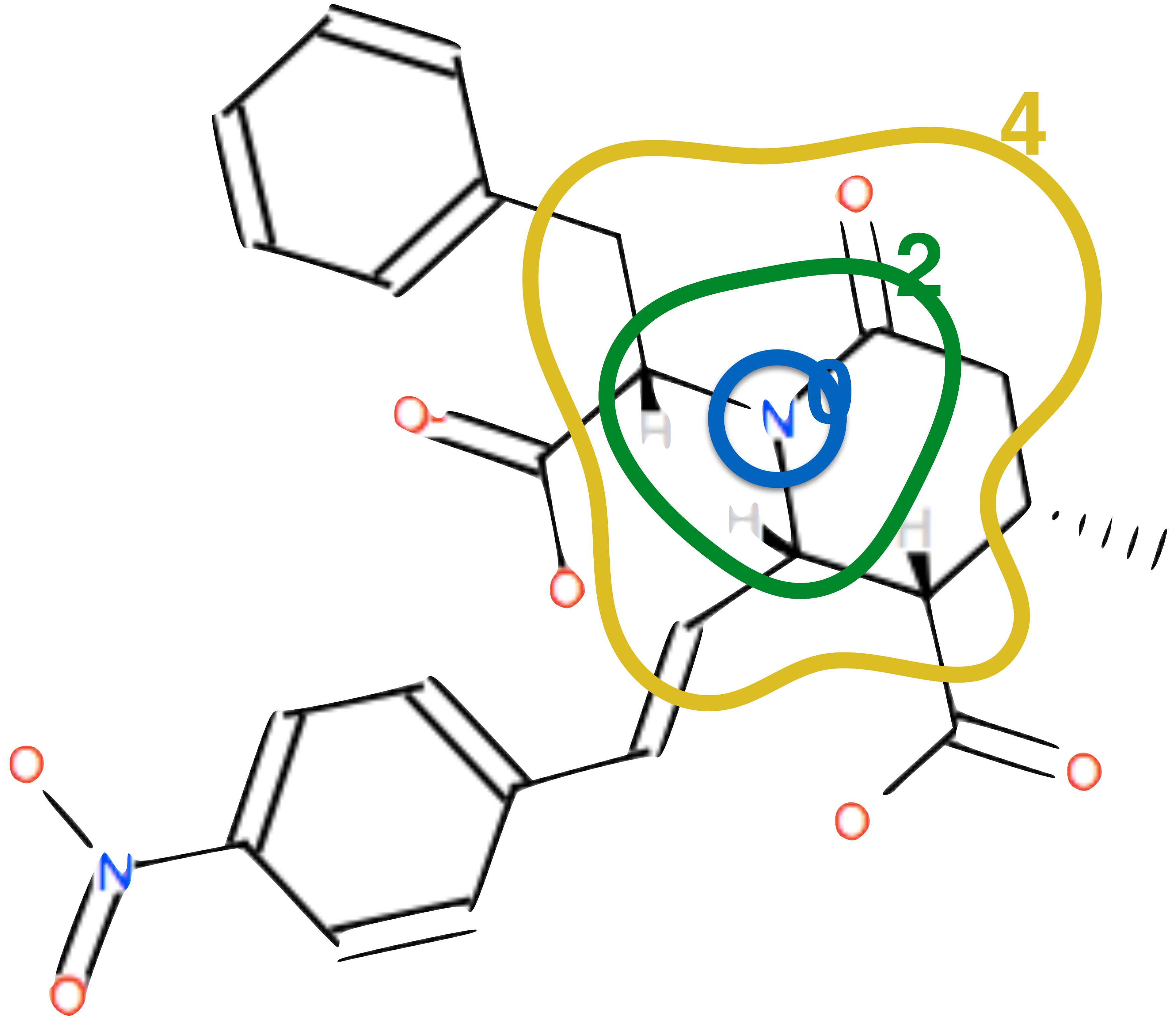




Topological Fingerprints

ECFP4

- all substructures with diameter 4 around every atom



Ligand Based: QSAR

Quantitative Structure/Activity Relationships

Properties

Compounds

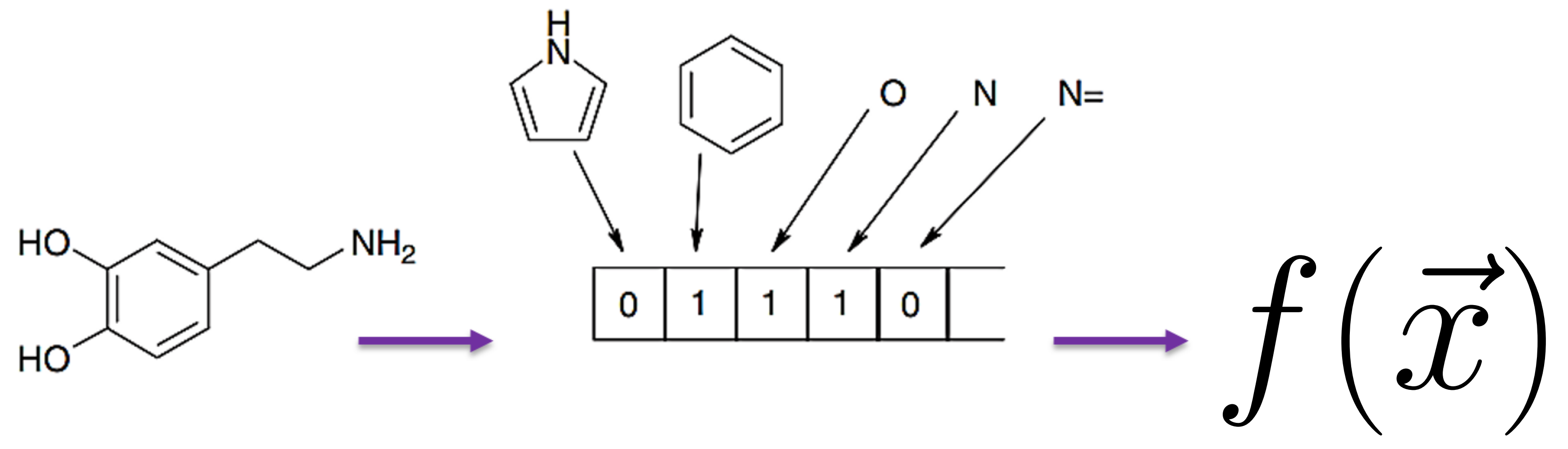
Cmpd	Cmpd	X				Residual
1	6a	H	1.07	0	0.79	0.28
2	6b	Cl	0.09	0.71	0.21	-0.12
3	6d	NO_2	0.66	-0.28	1.02	-0.36
4	6e	CN	1.42	-0.57	1.26	0.16
5	6f	C_6H_5	-0.62	1.96	-0.81	0.19
6	6g	$N(CH_3)_2$	0.64	0.18	0.65	-0.01
7	6h	I	-0.46	1.12	-0.12	-0.34

Biological Activity = Learned ~~linear~~ function of properties

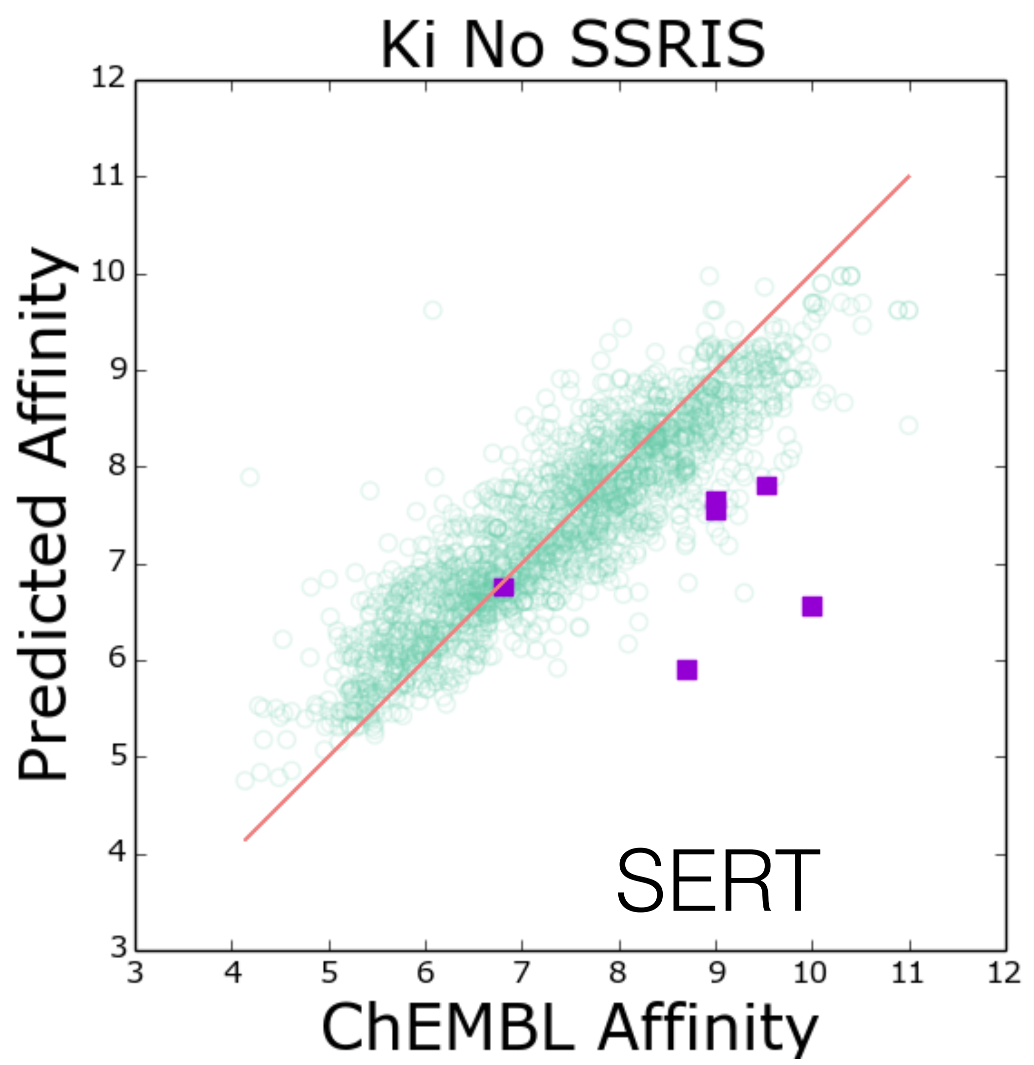
3D-QSAR: includes geometric/structural properties

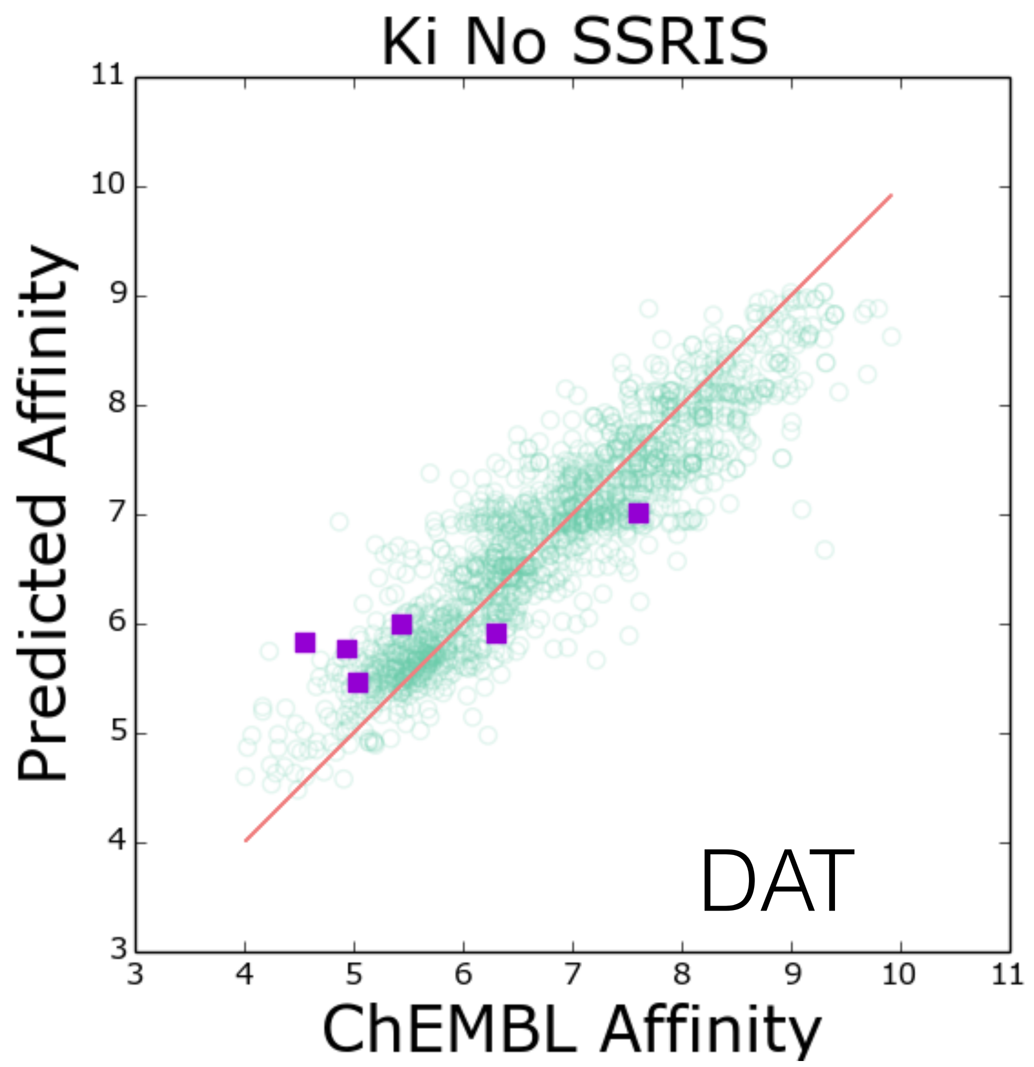
Ligand Based: QSAR

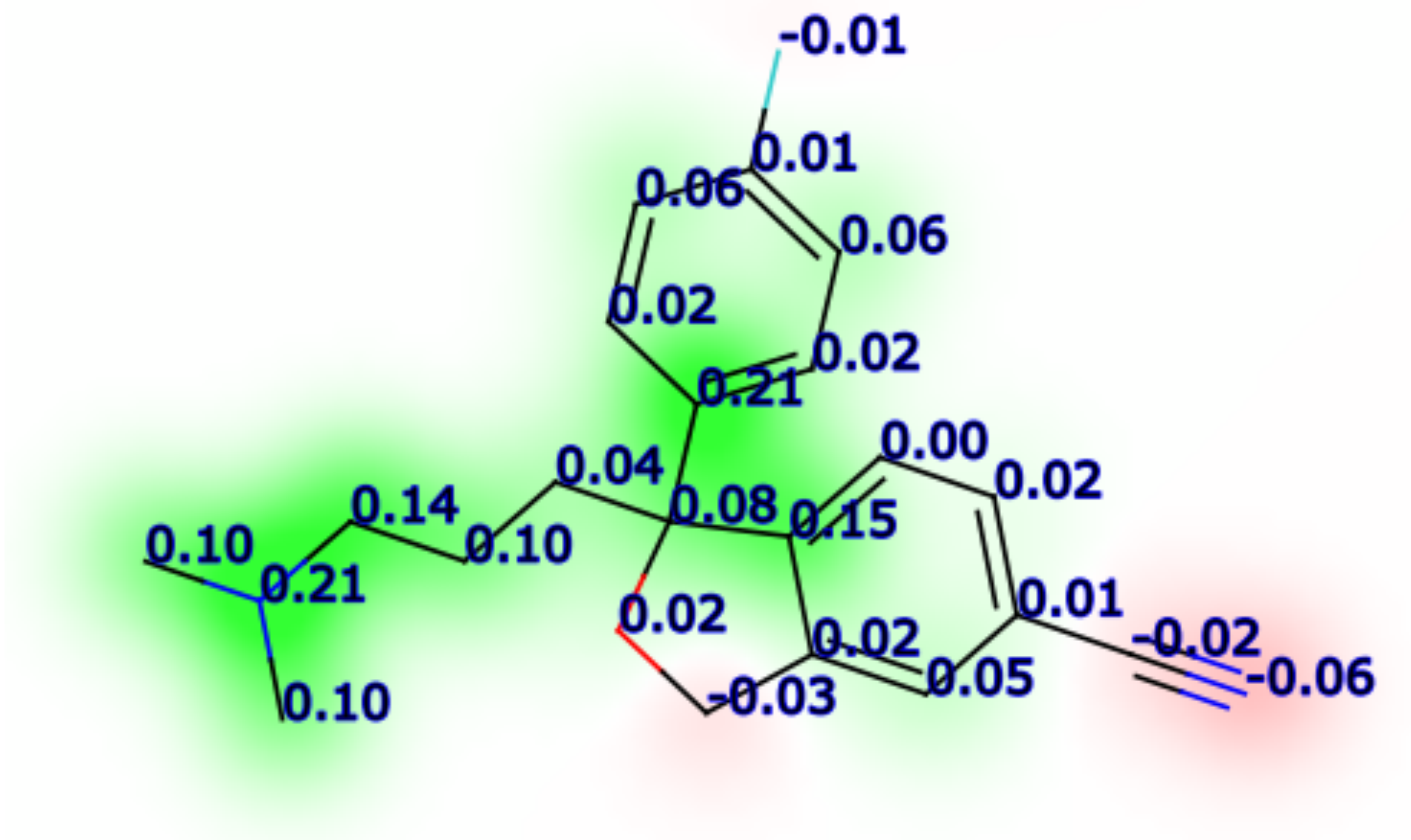
Quantitative Structure/Activity Relationships



$$f(\vec{x}) = w_1\vec{x}_1 + w_2\vec{x}_2 + w_3\vec{x}_3 + \ldots + b$$



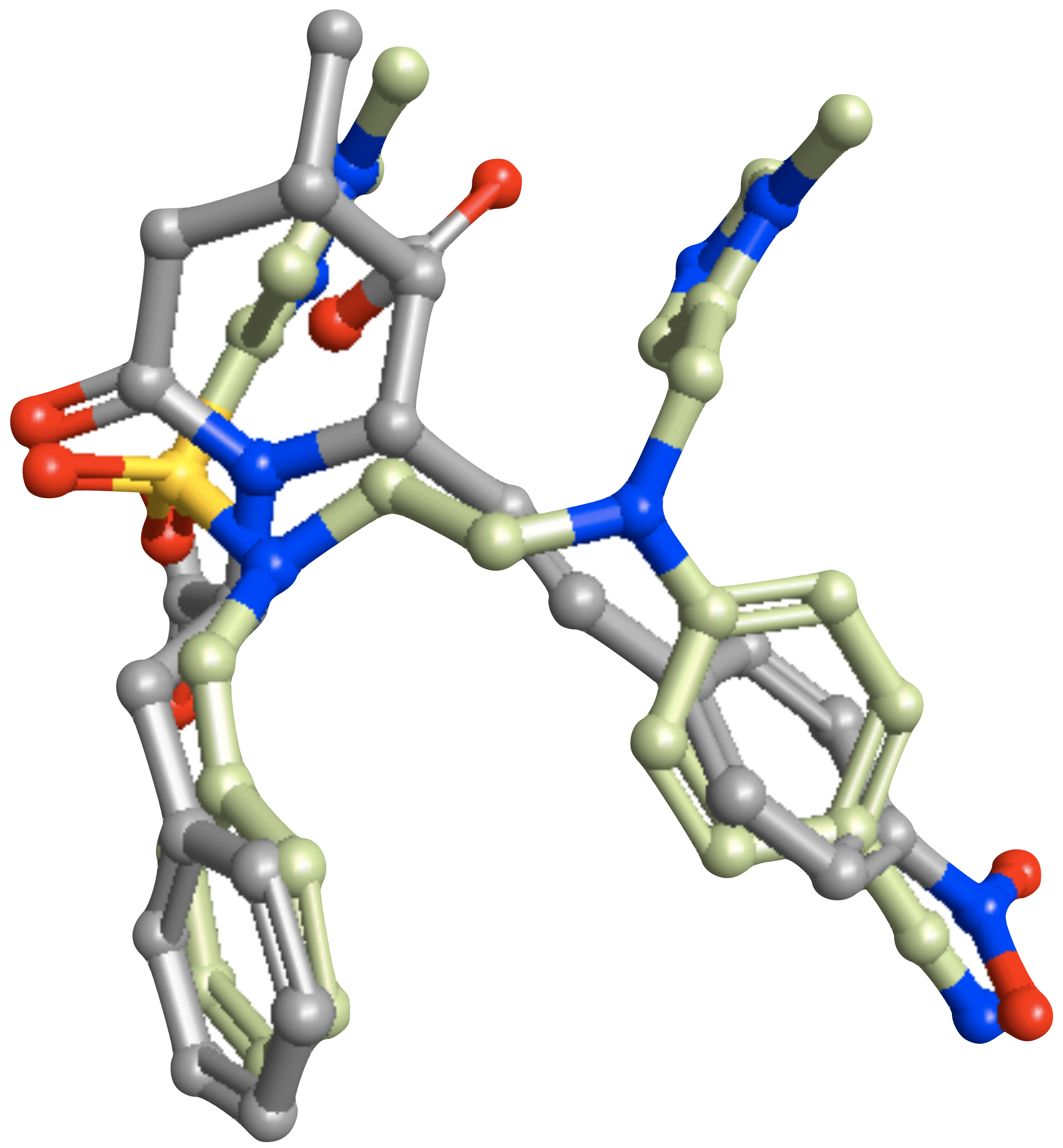


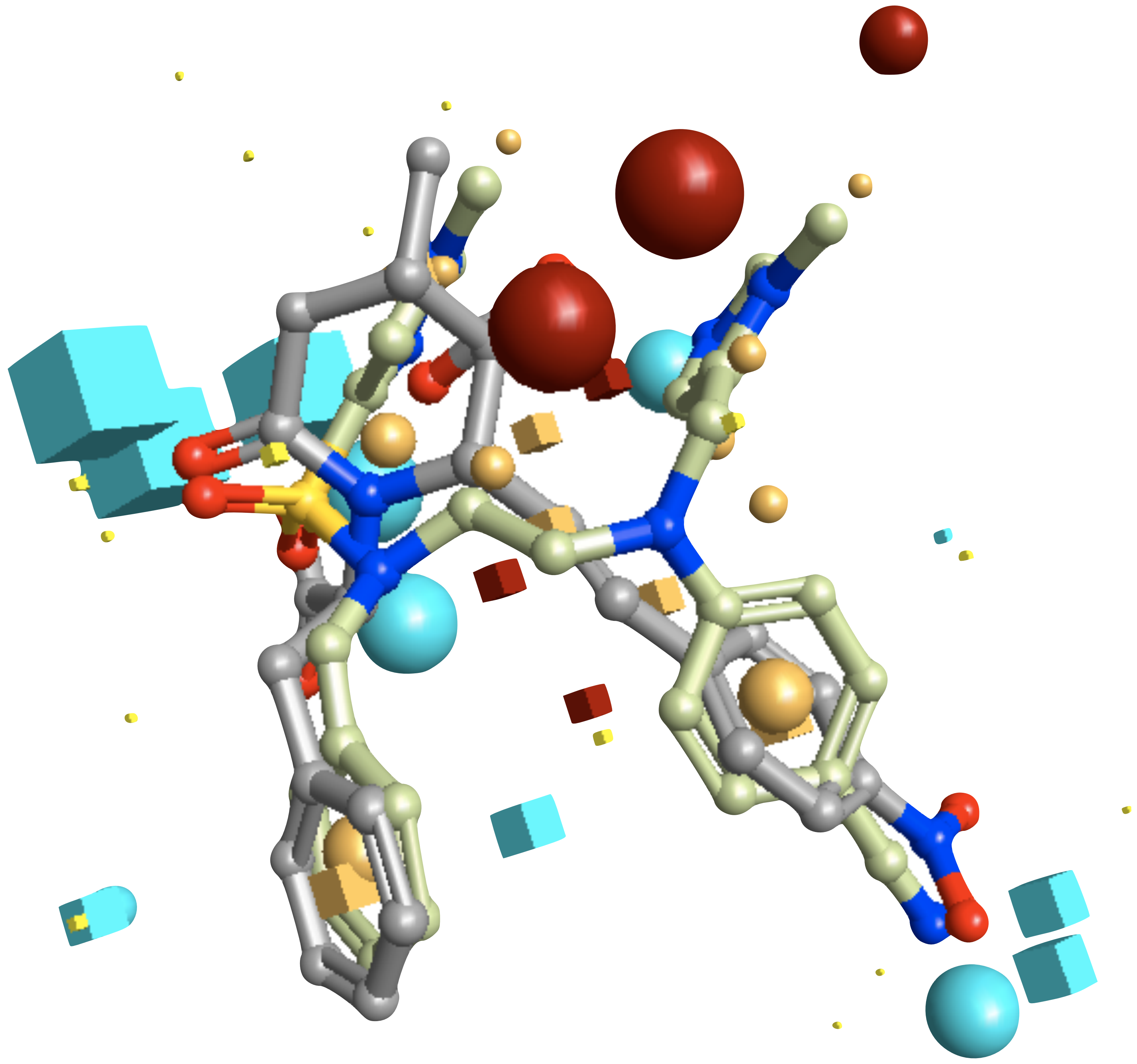


Ligand Based: Similarity

Superposition Methods

- compute “overlap” between molecules
- consider shape, electrostatics, **pharmacophores**





http://www.cresset-group.com/

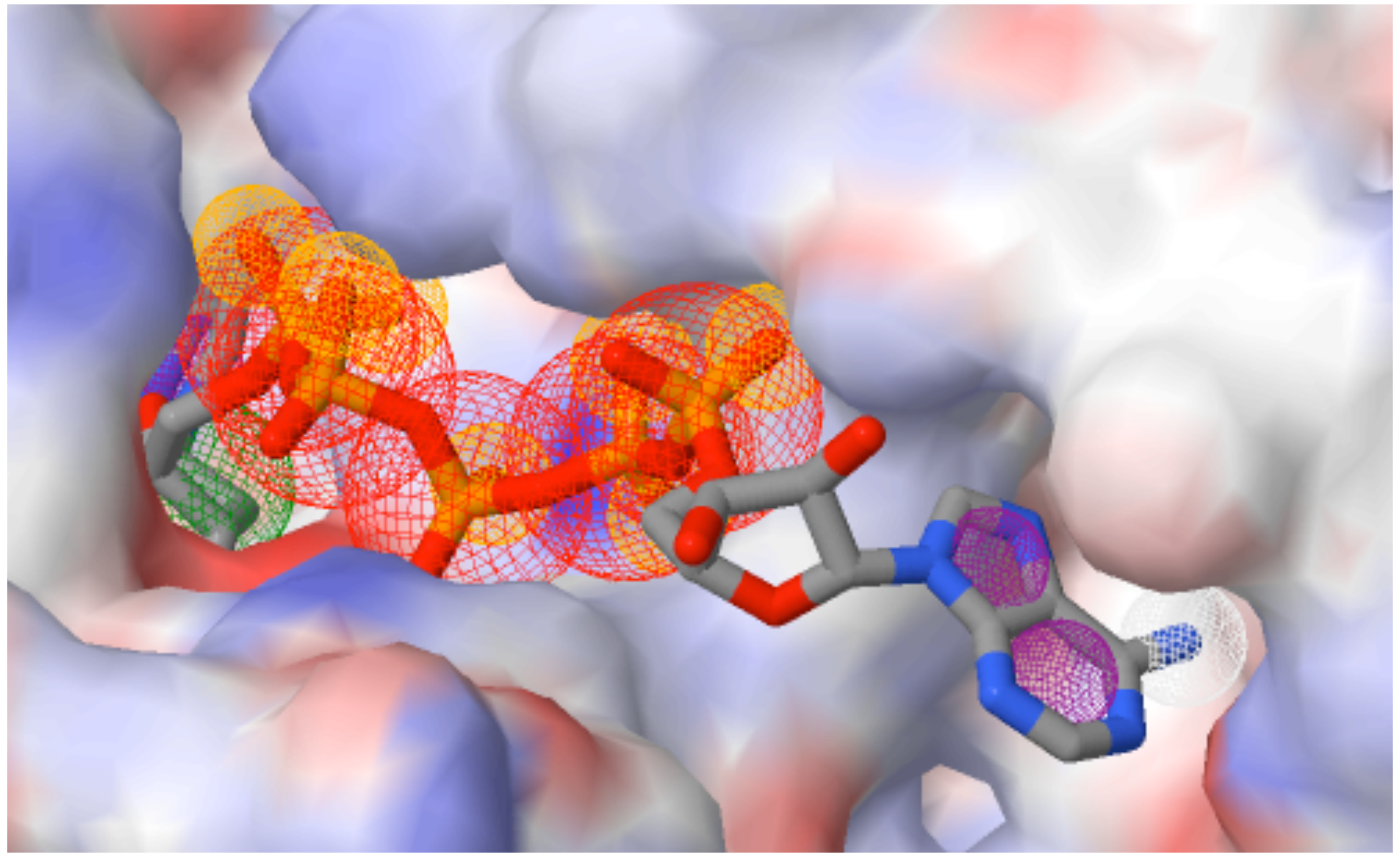
Ligand/Receptor Based: Pharmacophore

Pharmacophore:

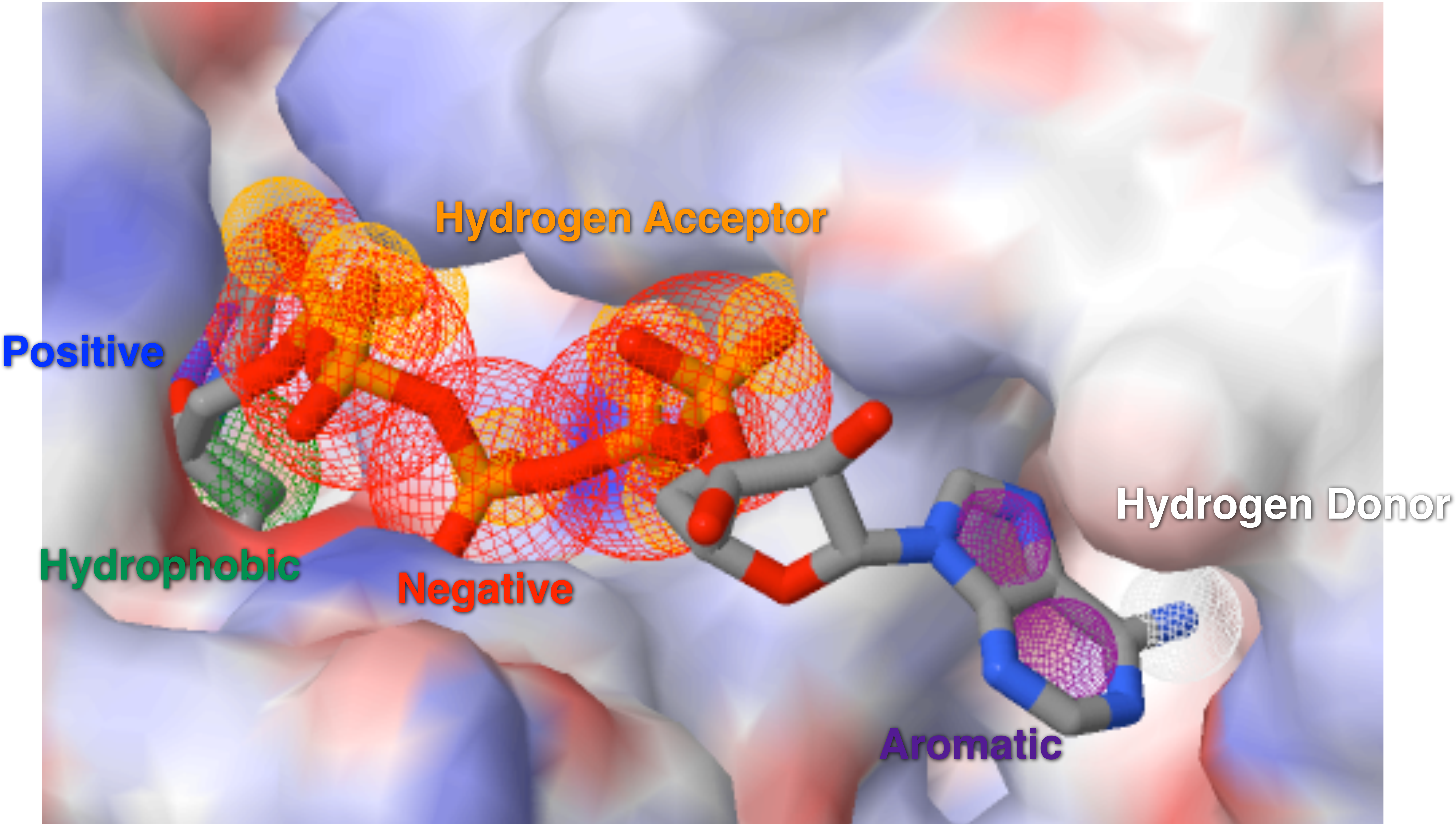
IUPAC: The ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response.

Common Features:

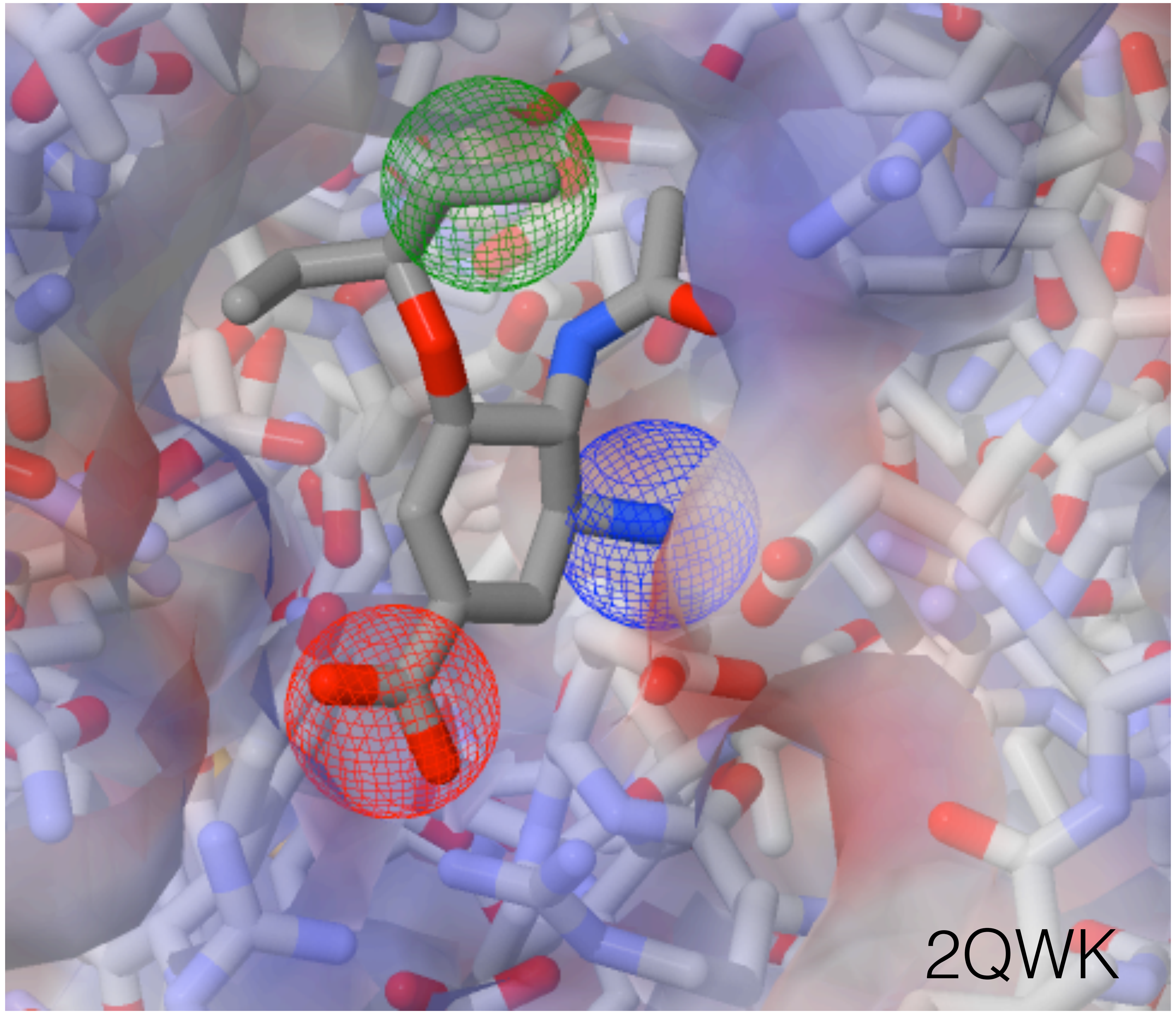
aromatic ring
hydrophobic area
positive ionizable
negative ionizable
hydrogen bond donor
hydrogen bond acceptor



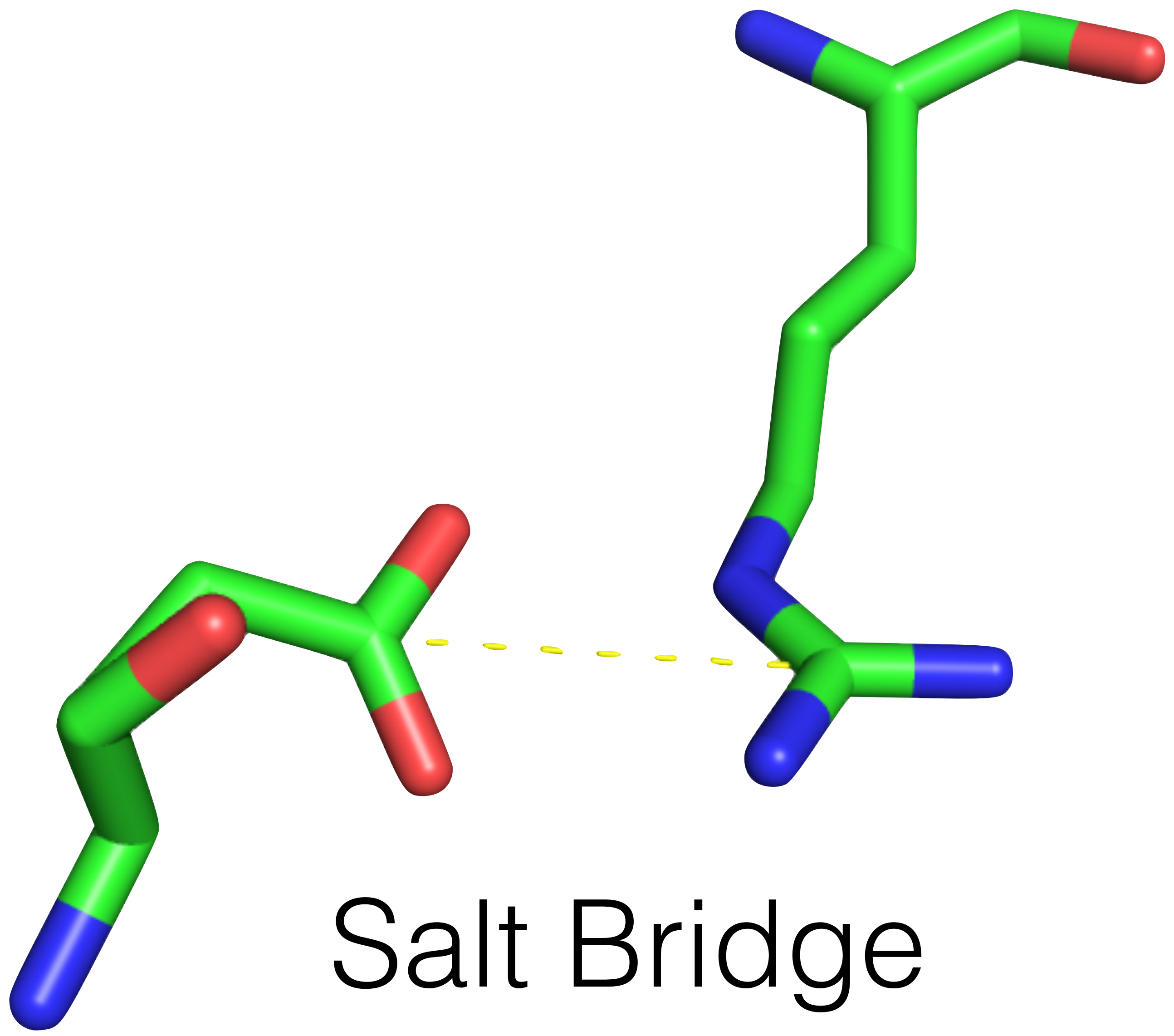
Pharmacophore Features



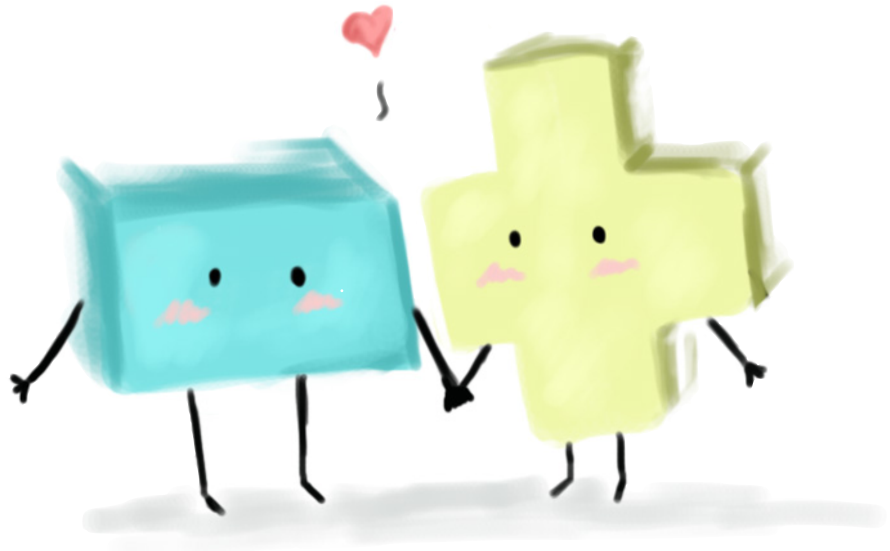
Charge-Charge



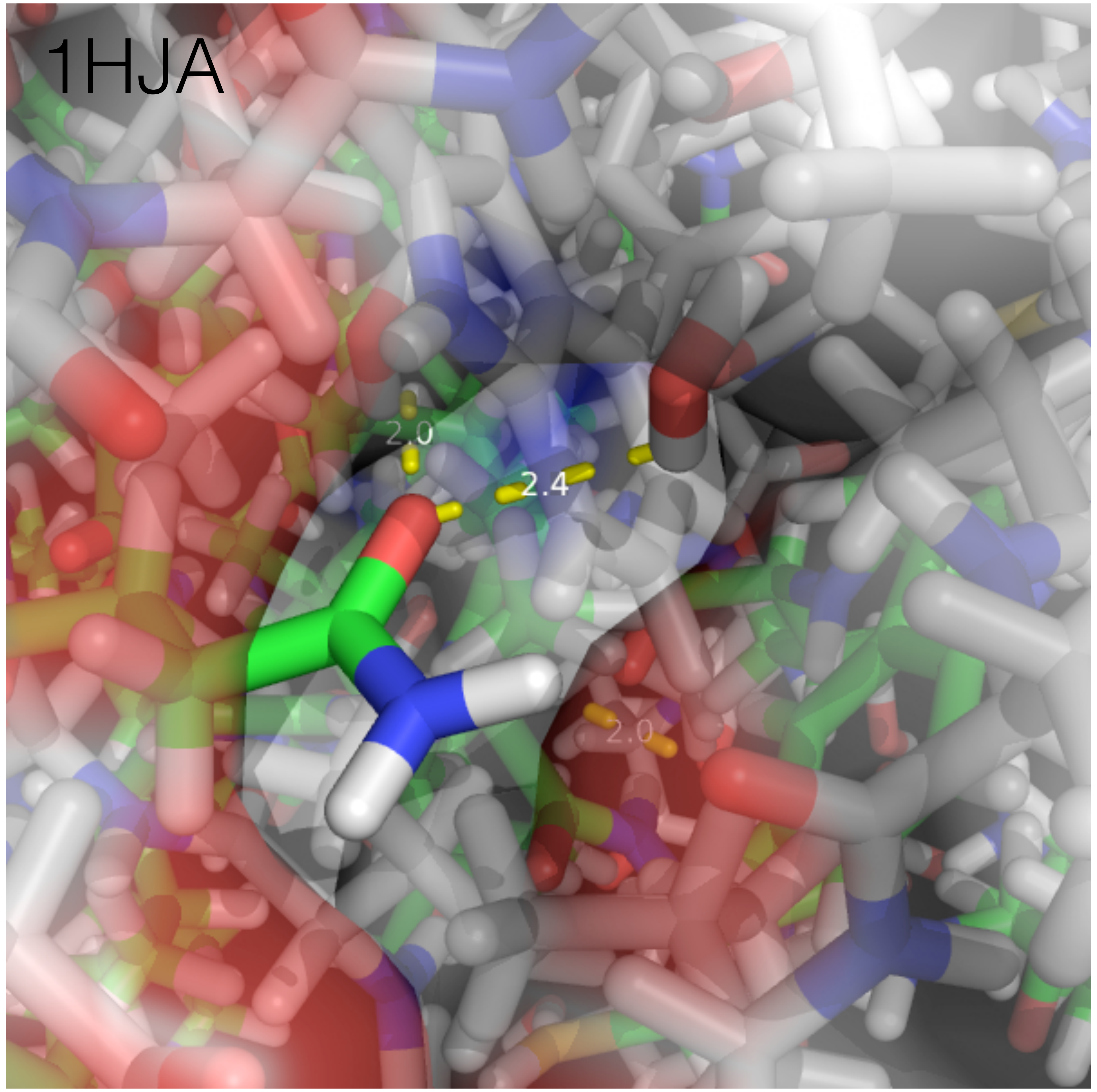
Inhibitor of the influenza virus neuraminidase (antiviral agent)



Salt Bridge



Hydrogen Bond



Turkey Ovomucoid Inhibitor

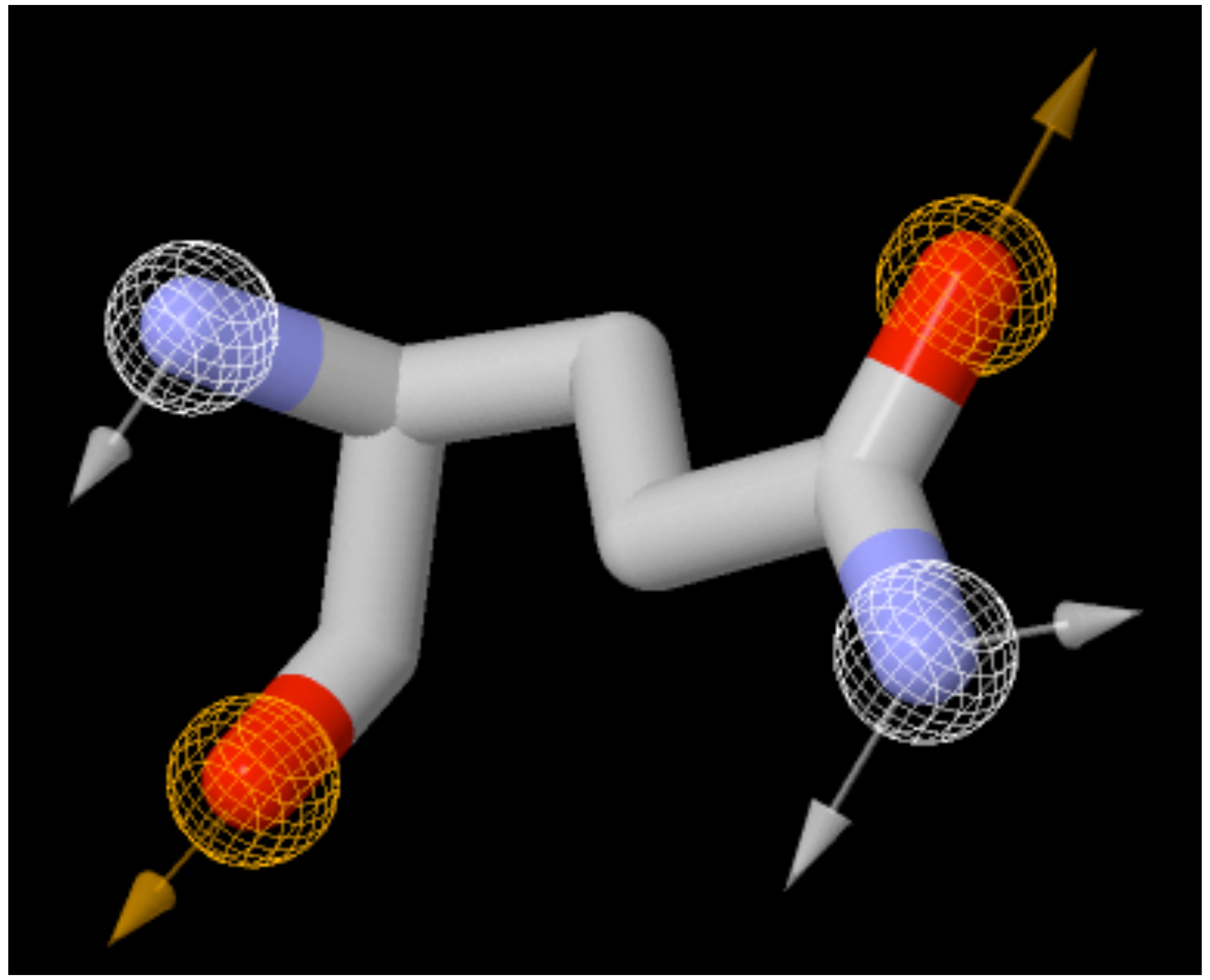
Distance:

D-A: 2.5Å – 3.5Å (4.0Å?)

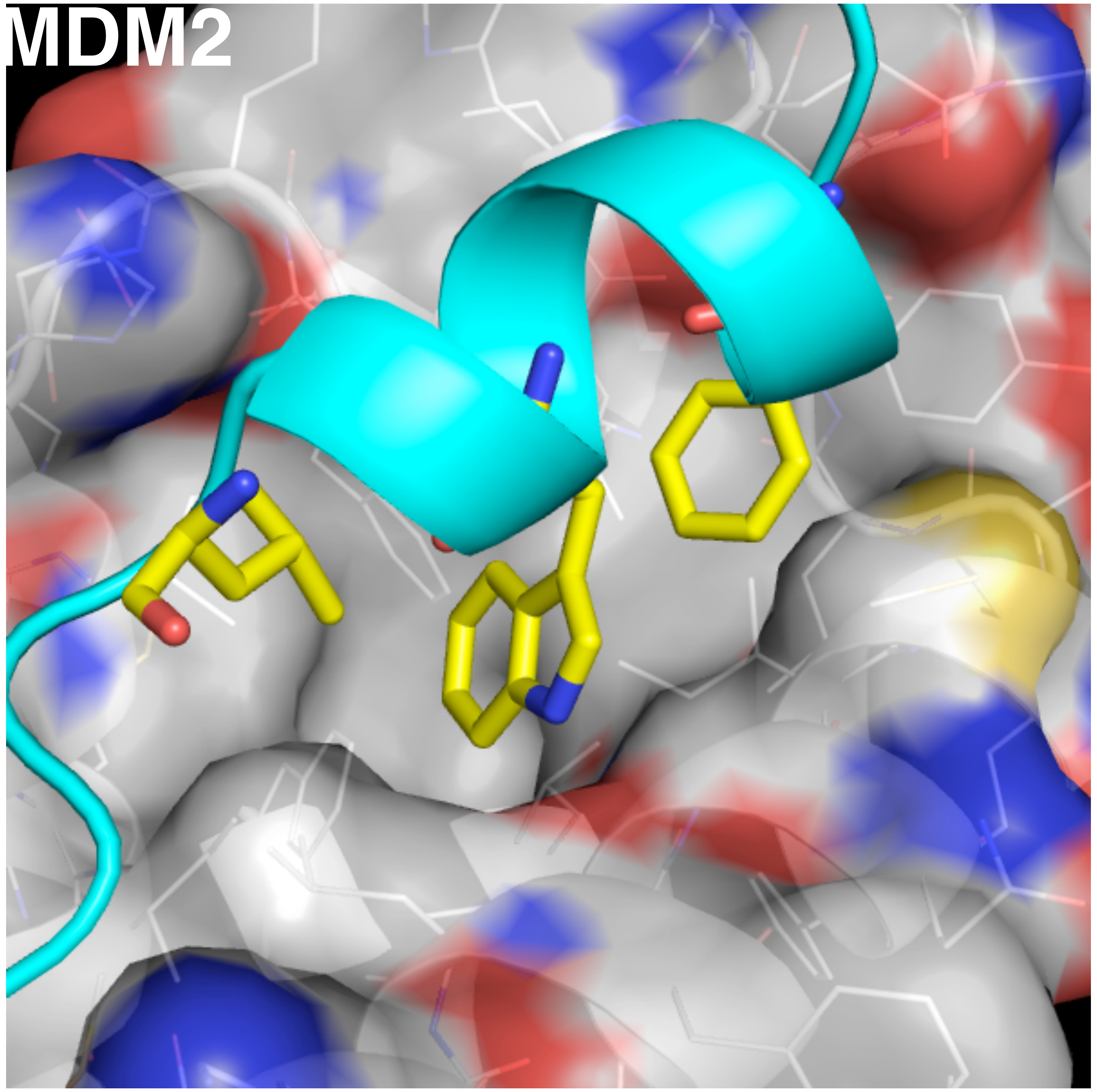
H-A: 1.5Å – 2.5Å

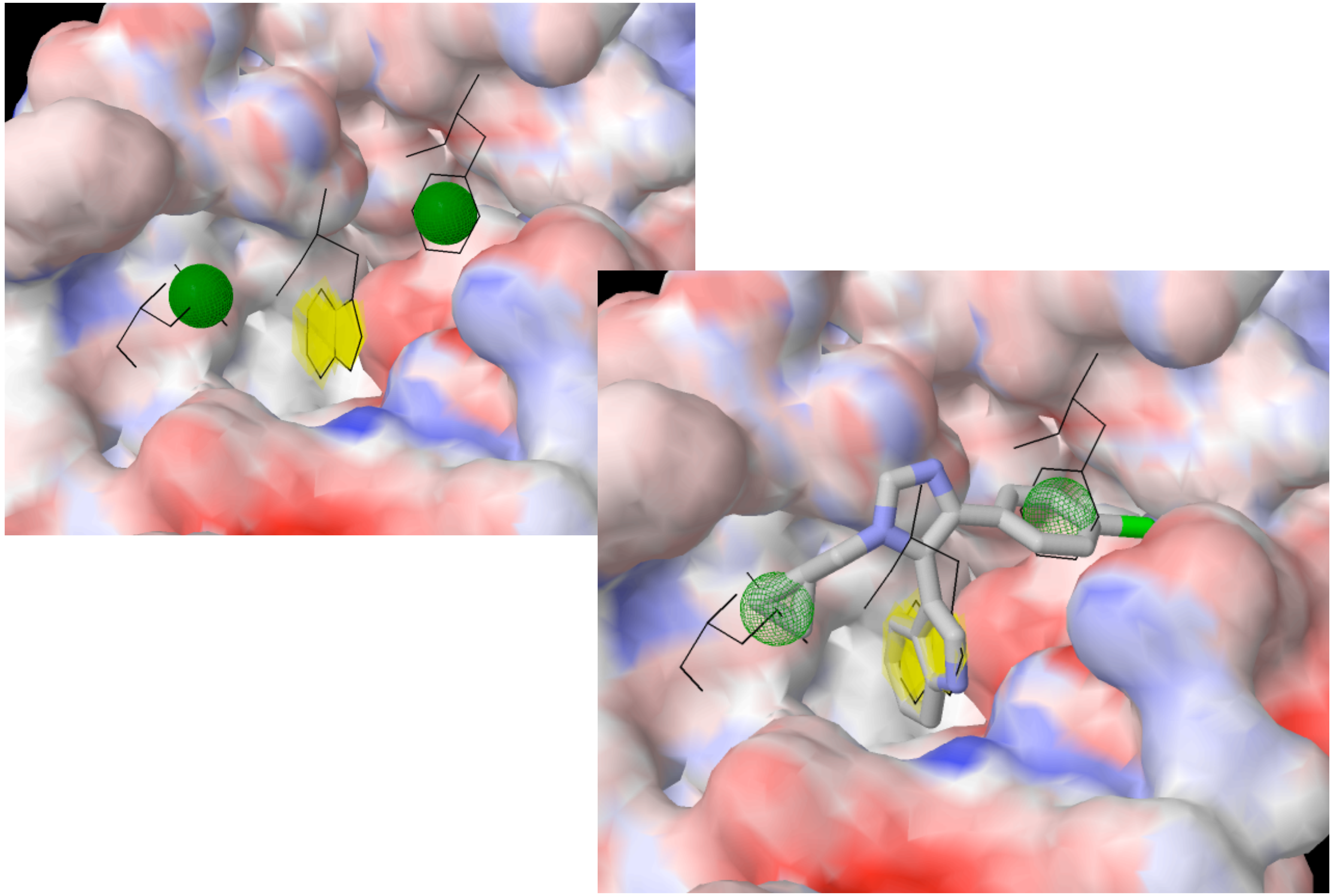
Angle:

Depends on context



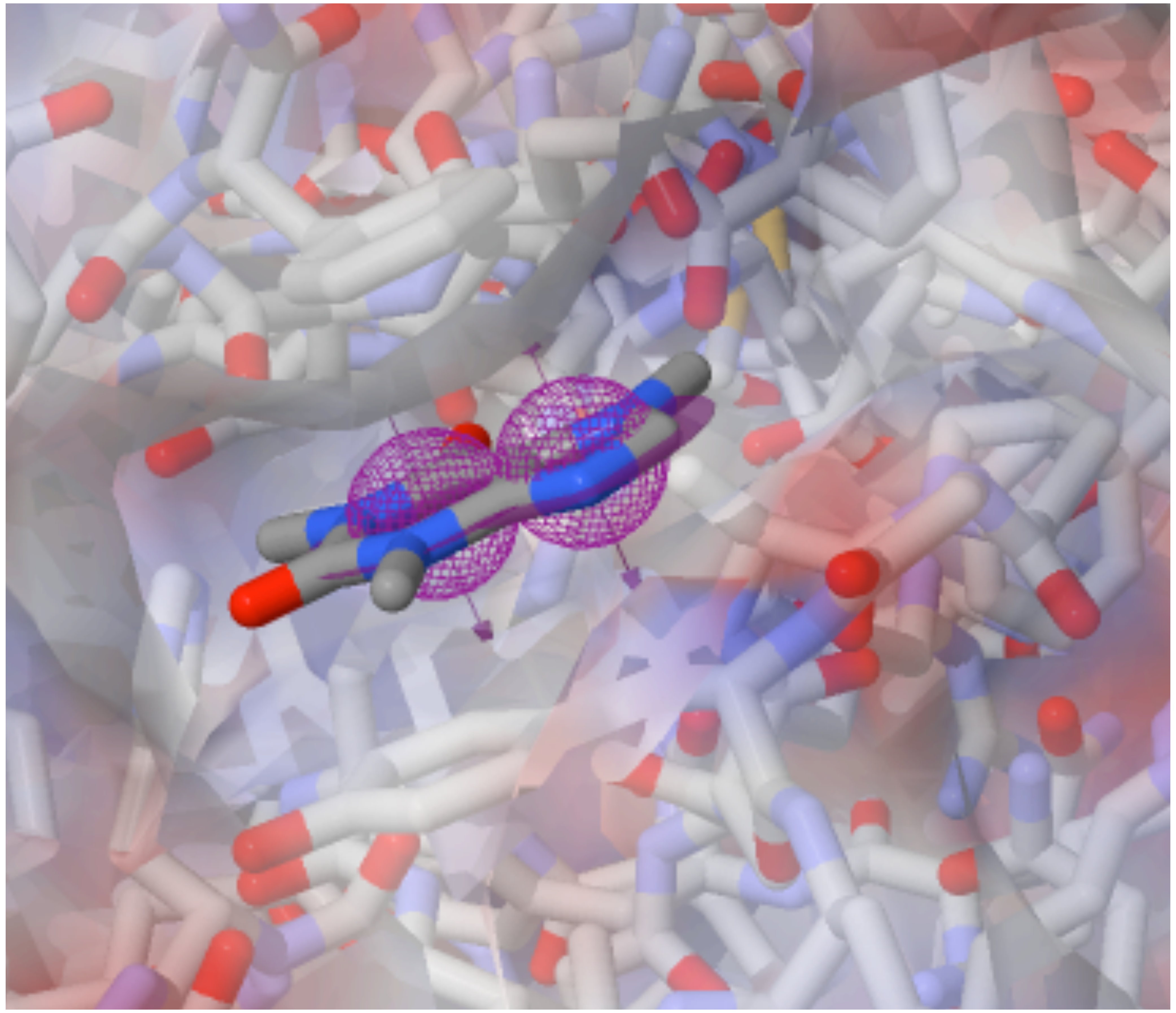
Hydrophobic





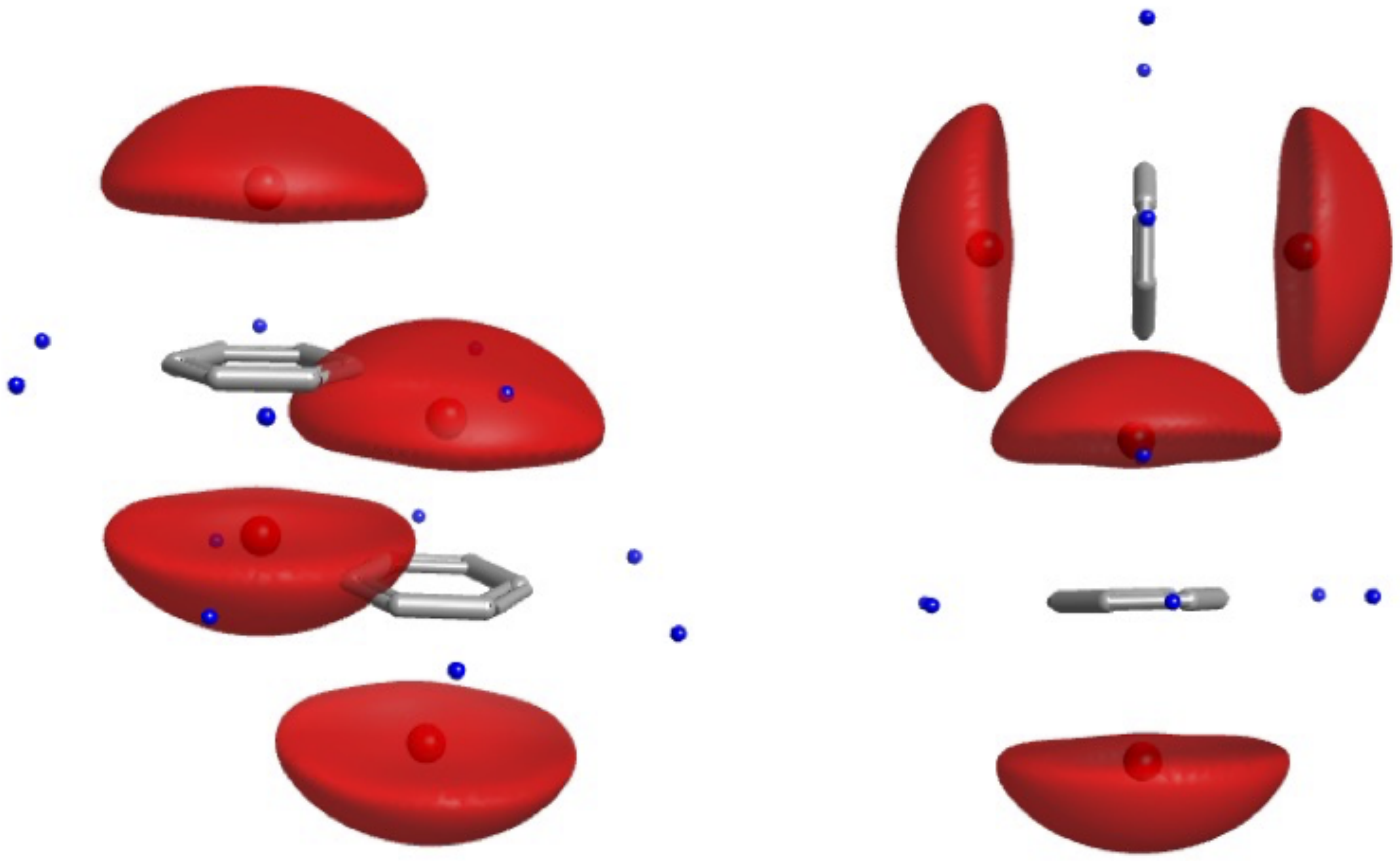
MDM2 (over expressed in >50% of cancers) down-regulates p53 (guardian of the genome)

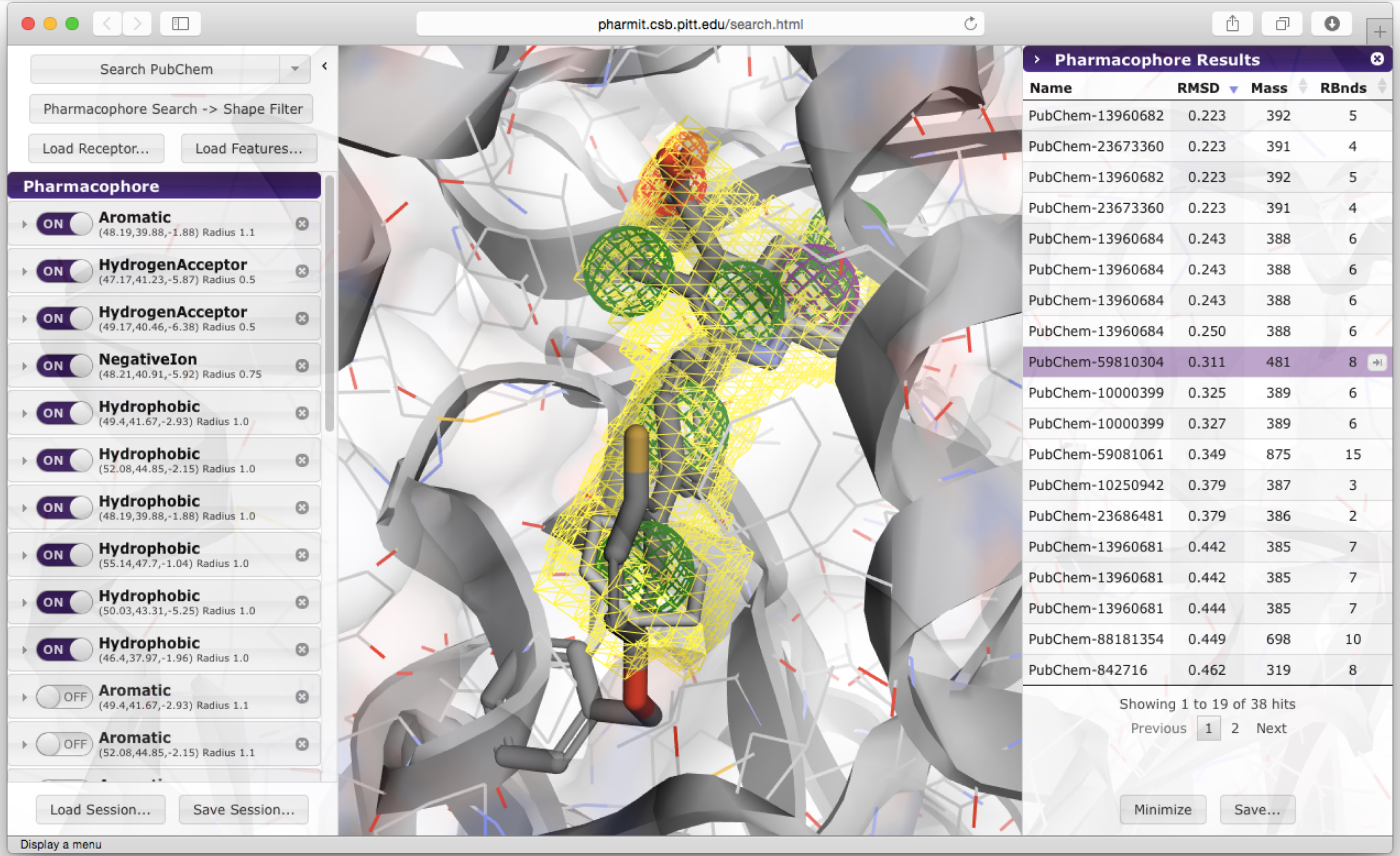
Aromatic



Human liver glycogen phosphorylase a complexed with caffeine

Rings offset
Interplanar distance: 3.3-3.8Å





http://pharmit.csb.pitt.edu

Kinds of Virtual Screening

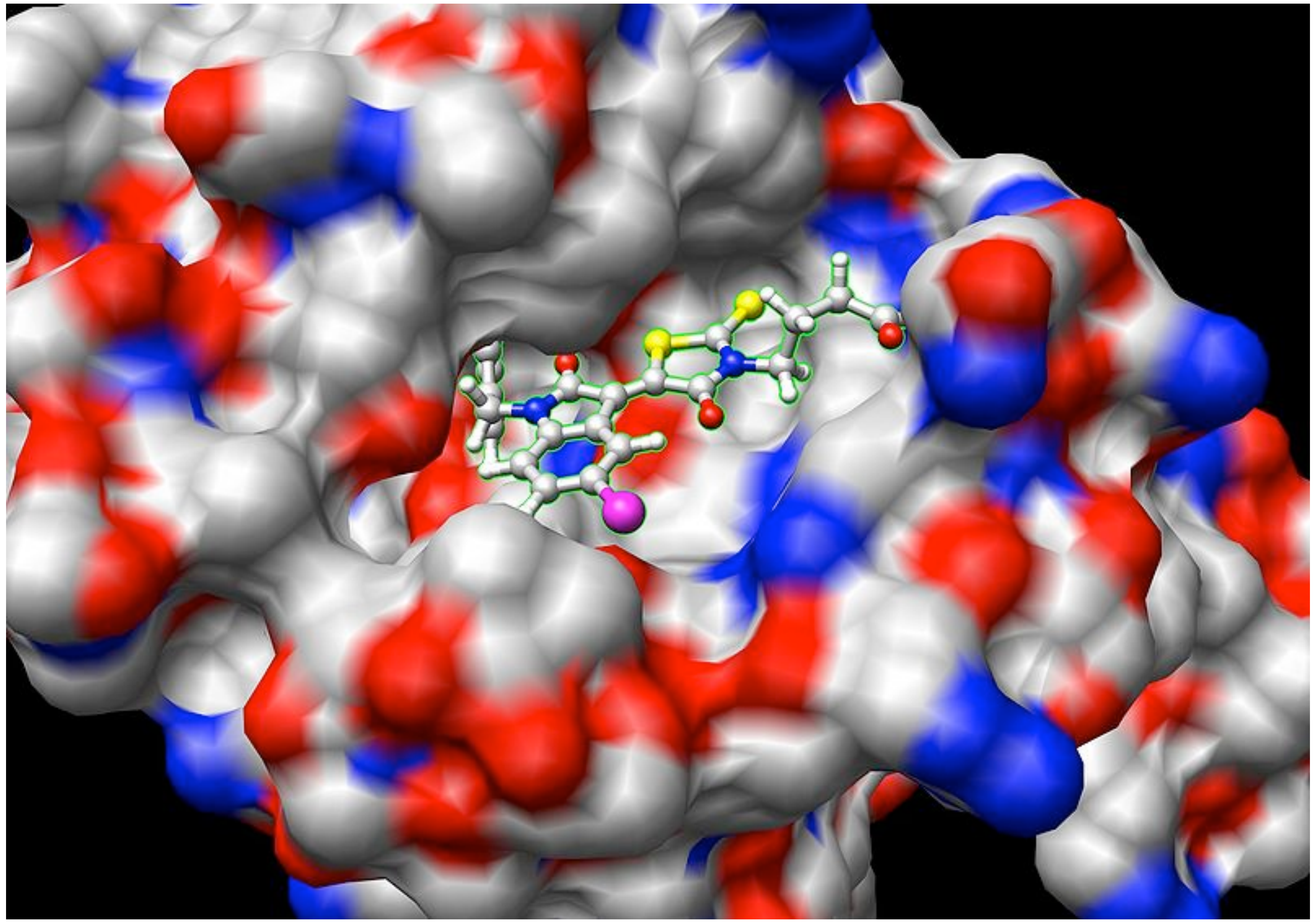
ADMET

Ligand Based

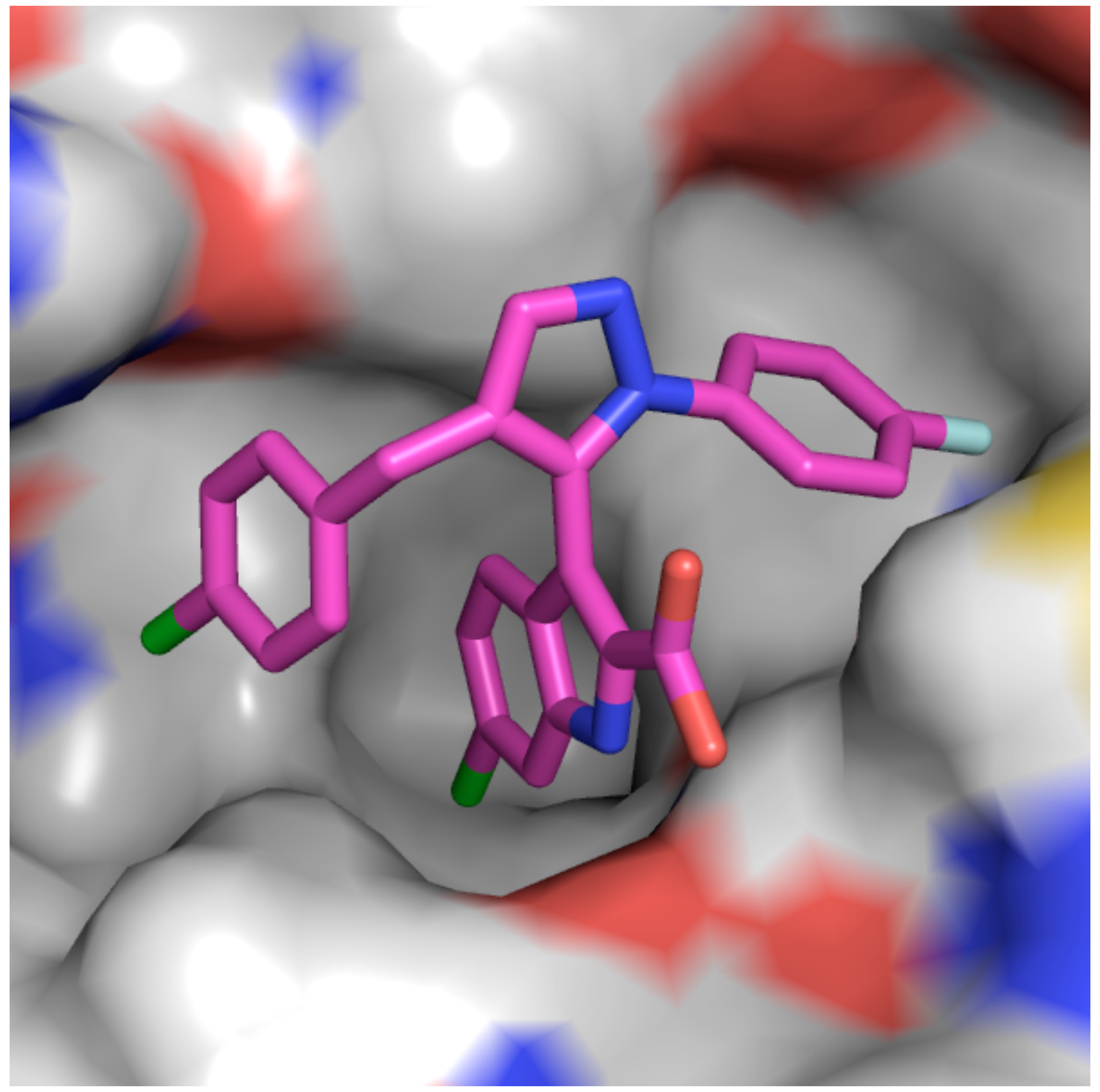
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based

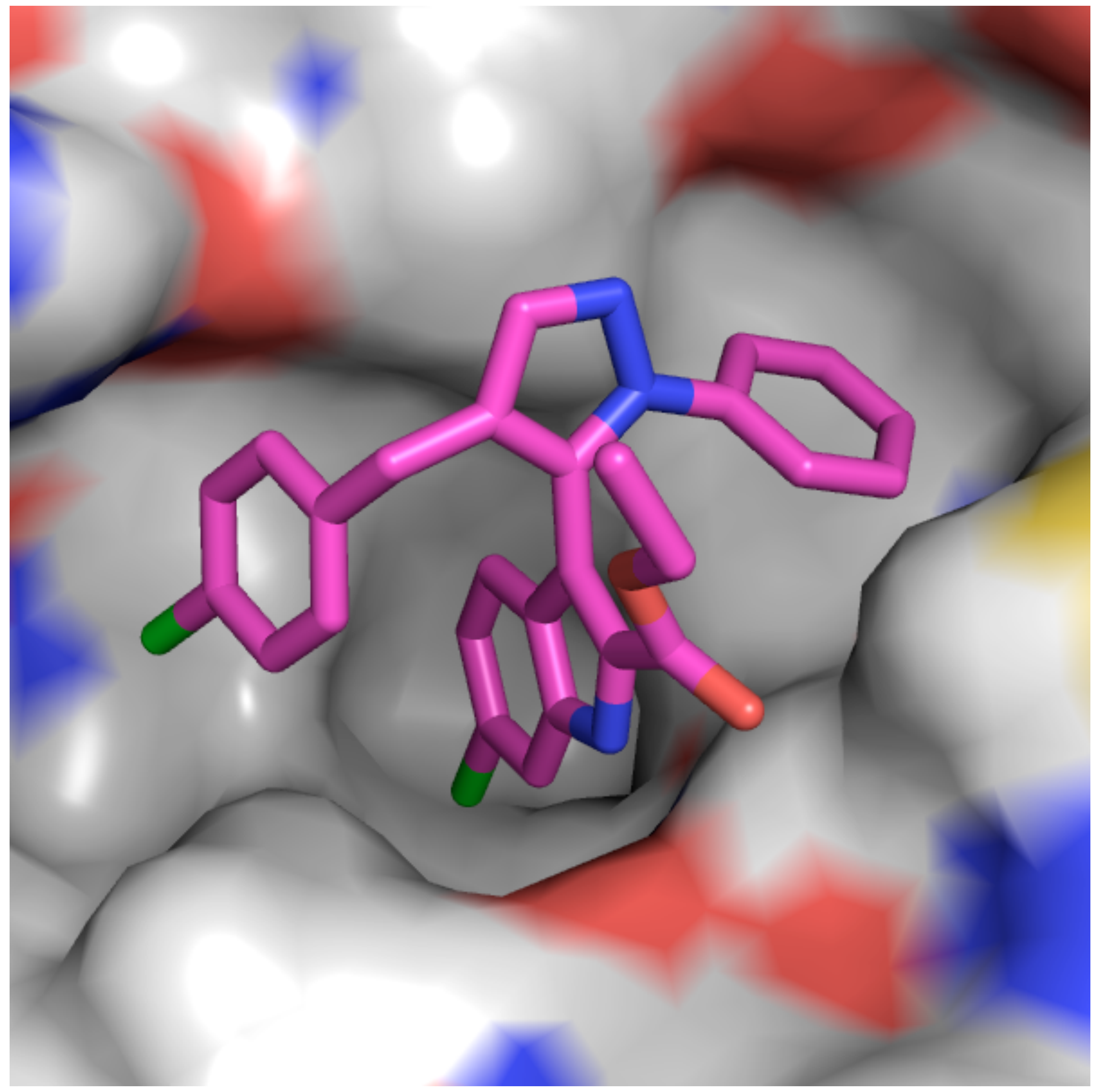
- **dock and score**



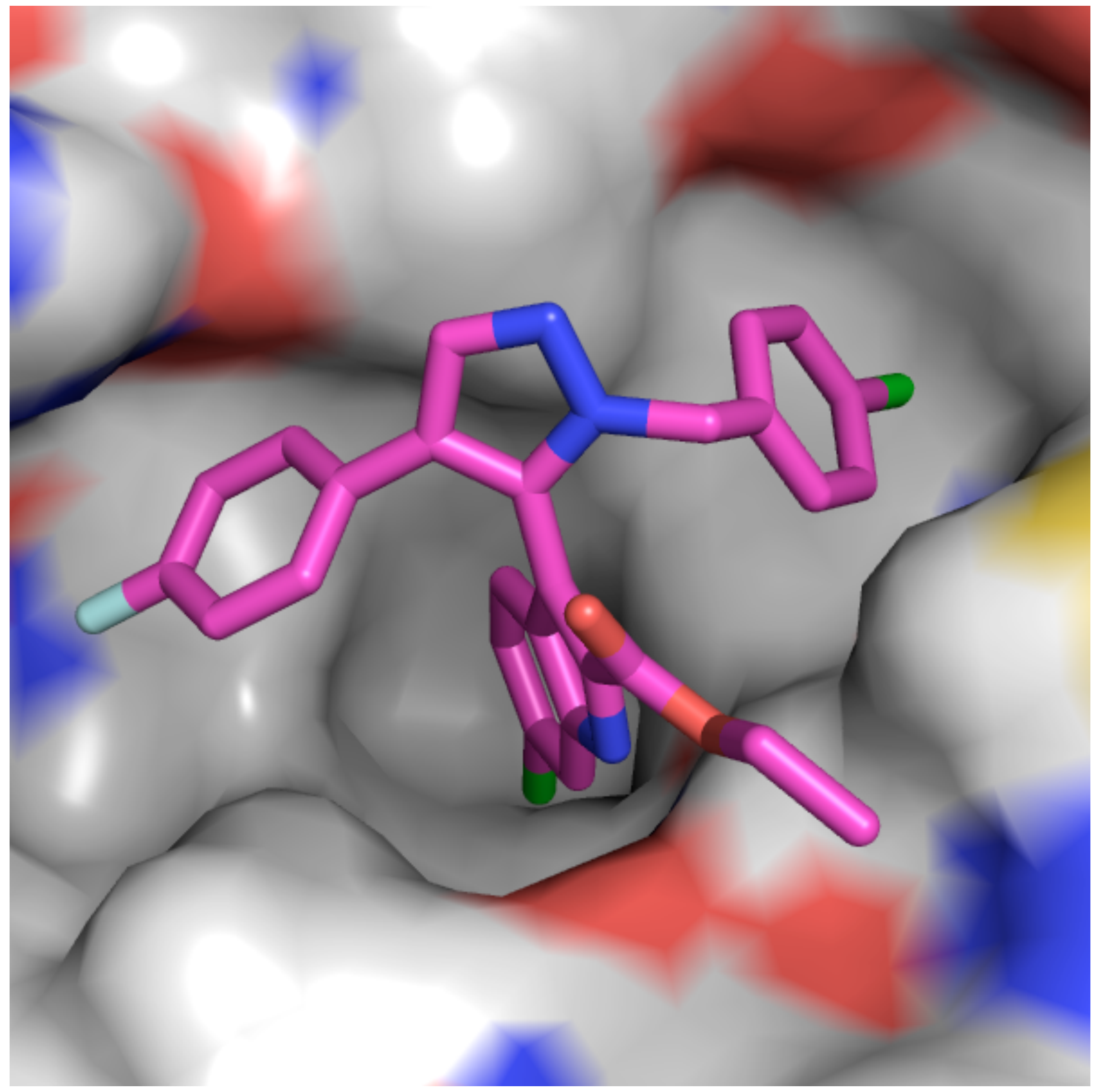
Pharmacophores Aren’t Enough



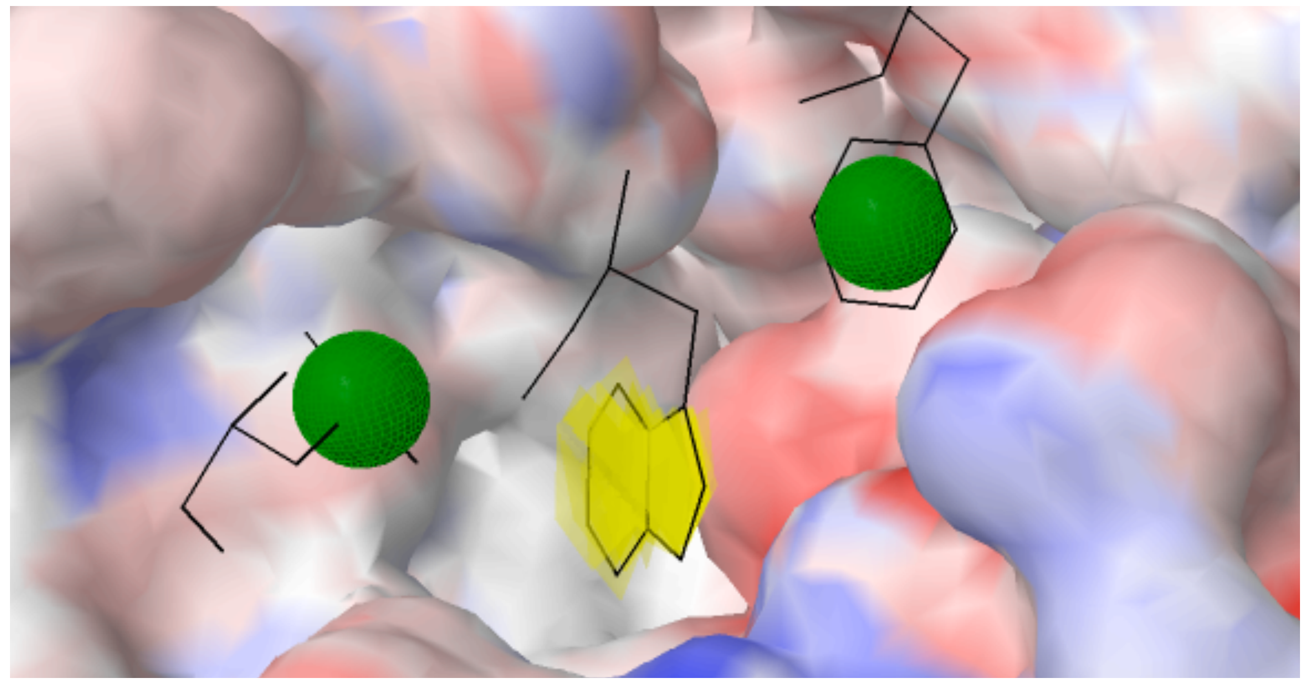
.2μM



50μM



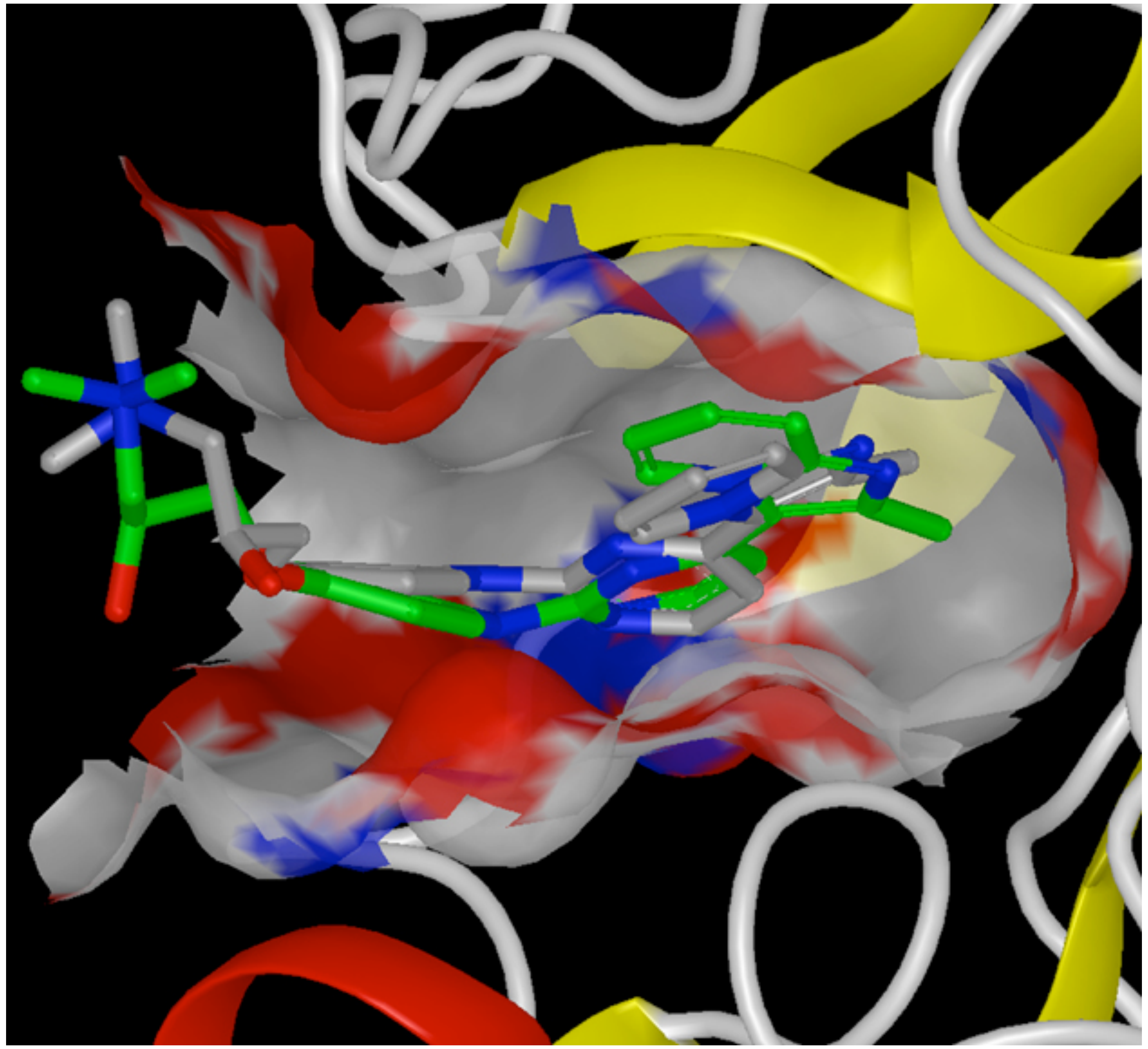
n.i.



Docking

Determine the **conformation** and **pose** of a ligand at a docking site

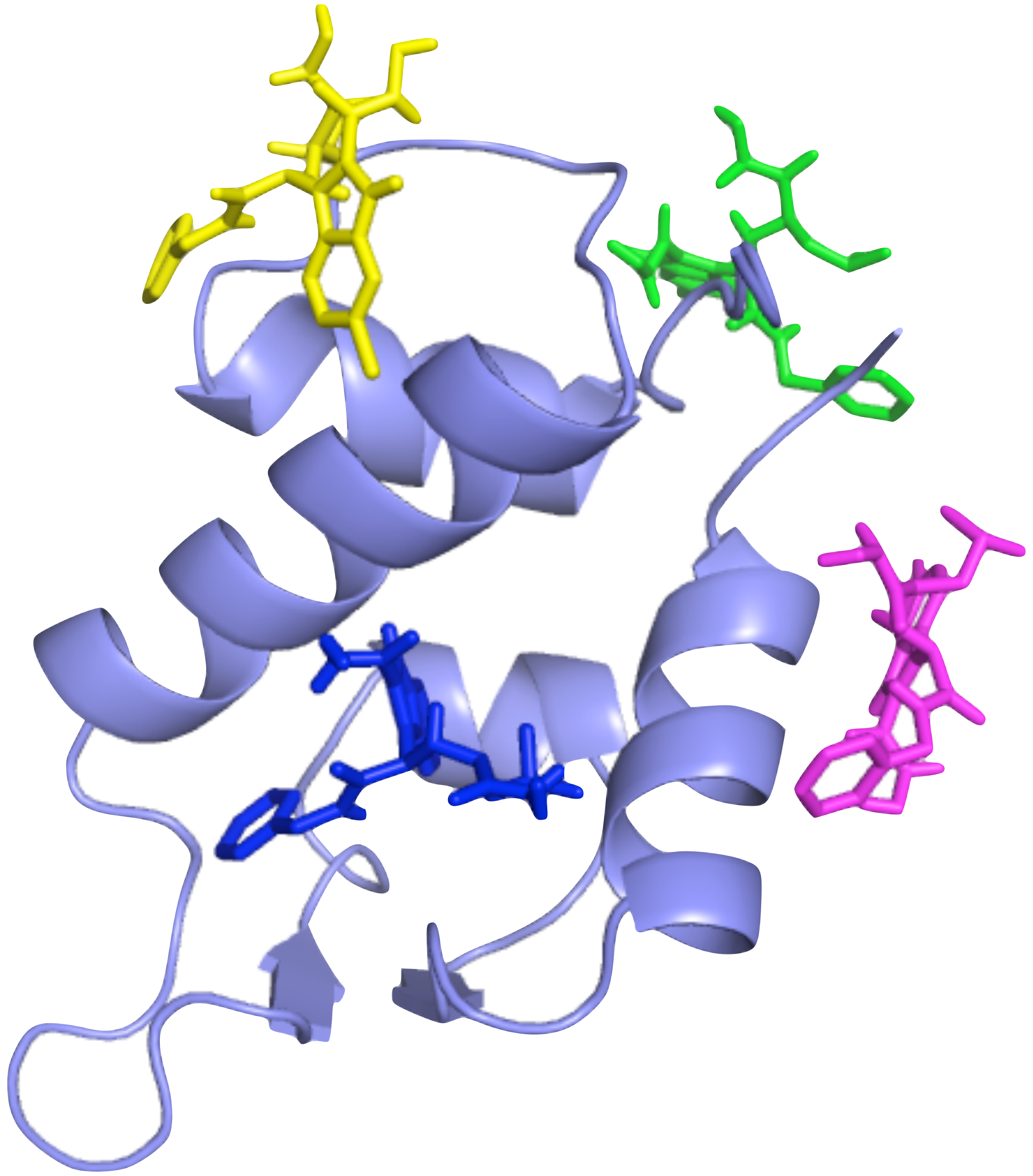
Challenge is to find conformation and pose with the best **score**



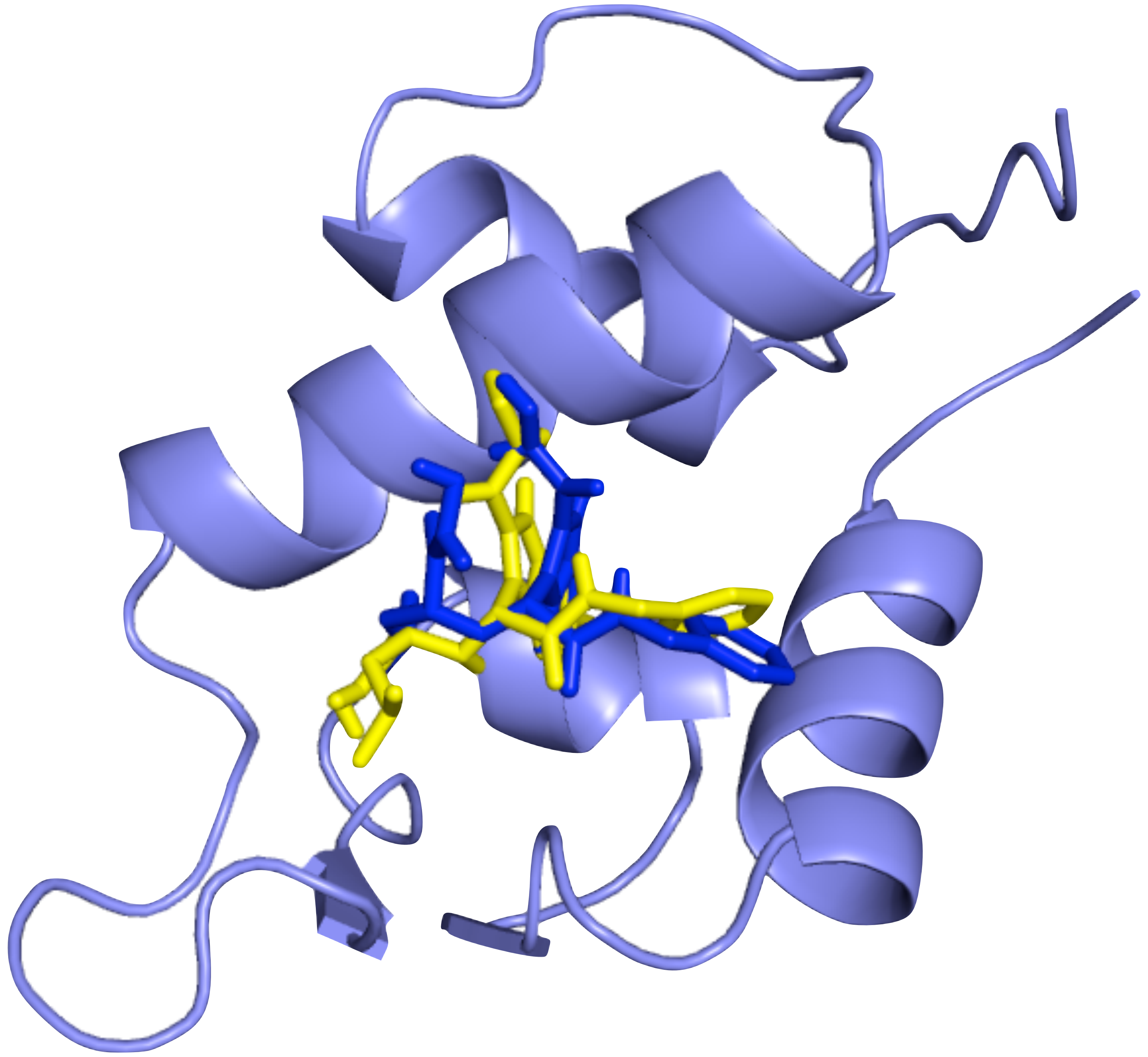
Two Phase Docking

1. Global Pose Estimation

2. Local Refinement



Stochastic

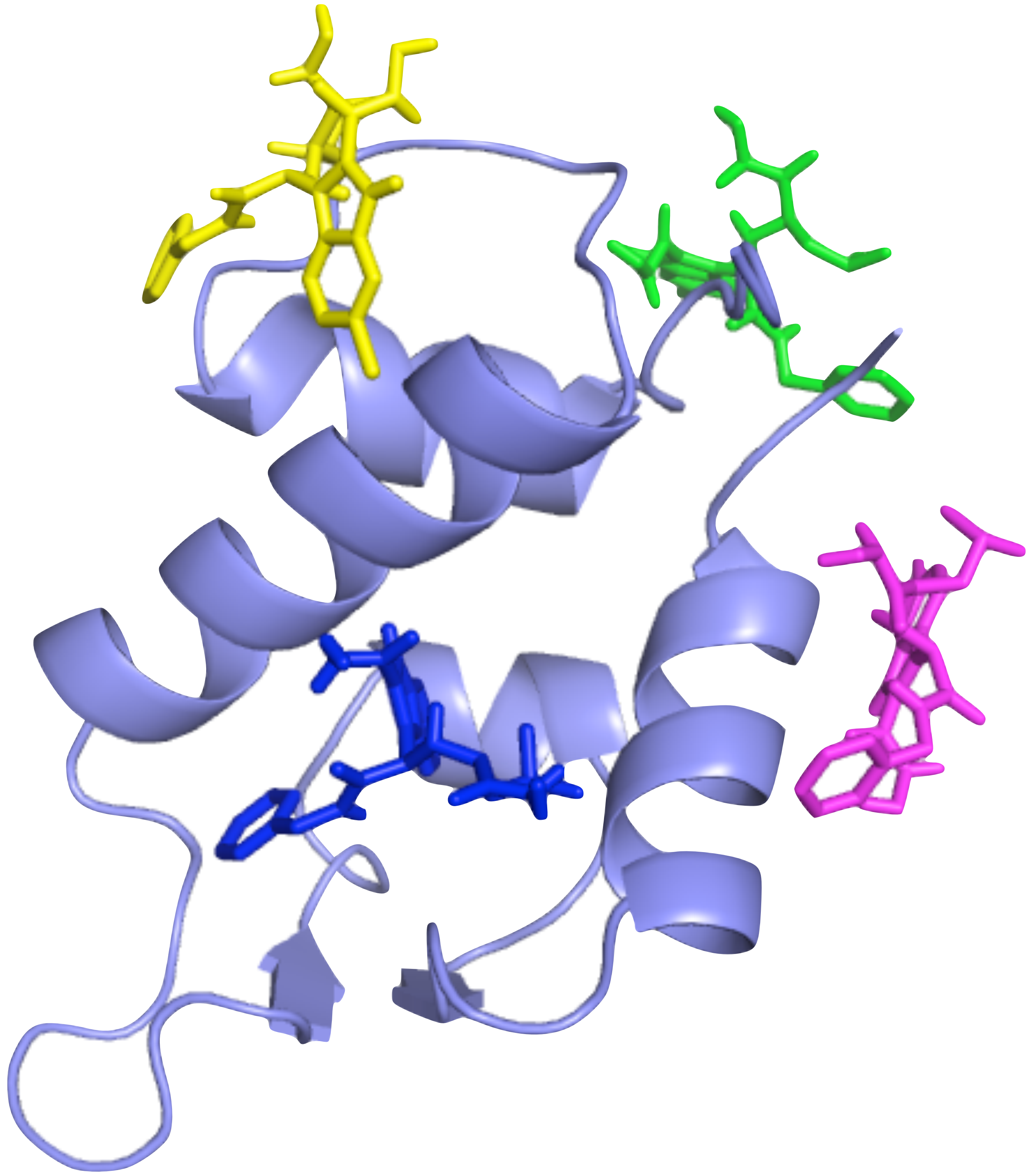


Minimization

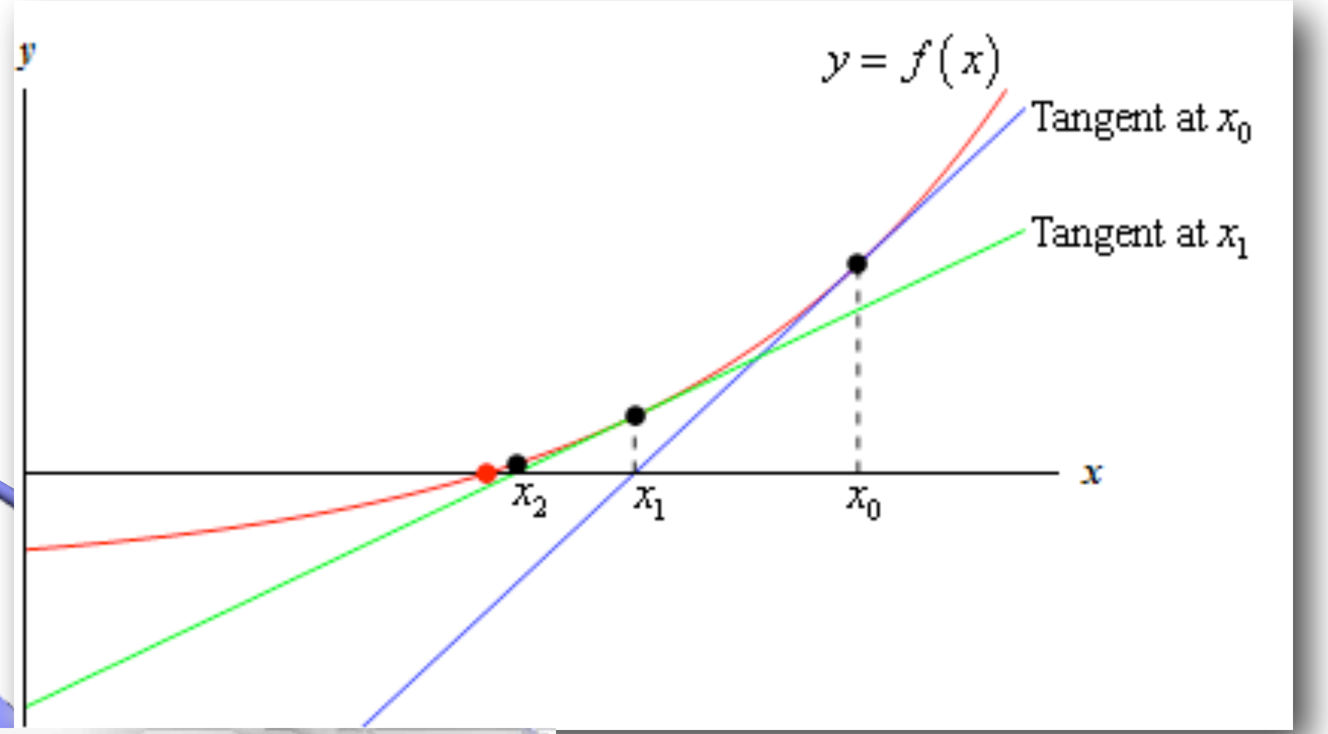
Two Phase Docking

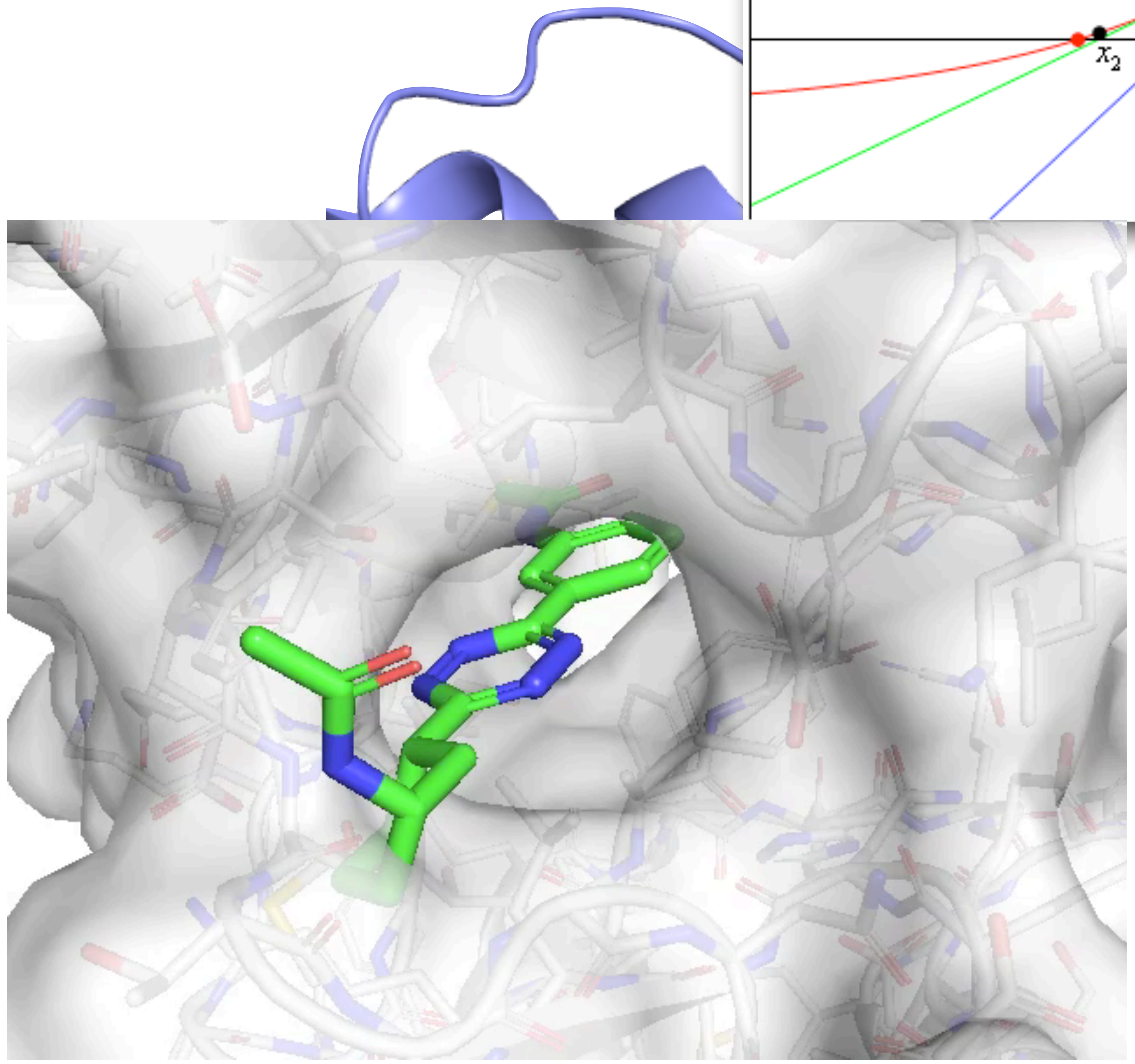
1. Global Pose Estimation

2. Local Refinement



Stochastic





Minimization

Scoring Goals

Affinity Prediction

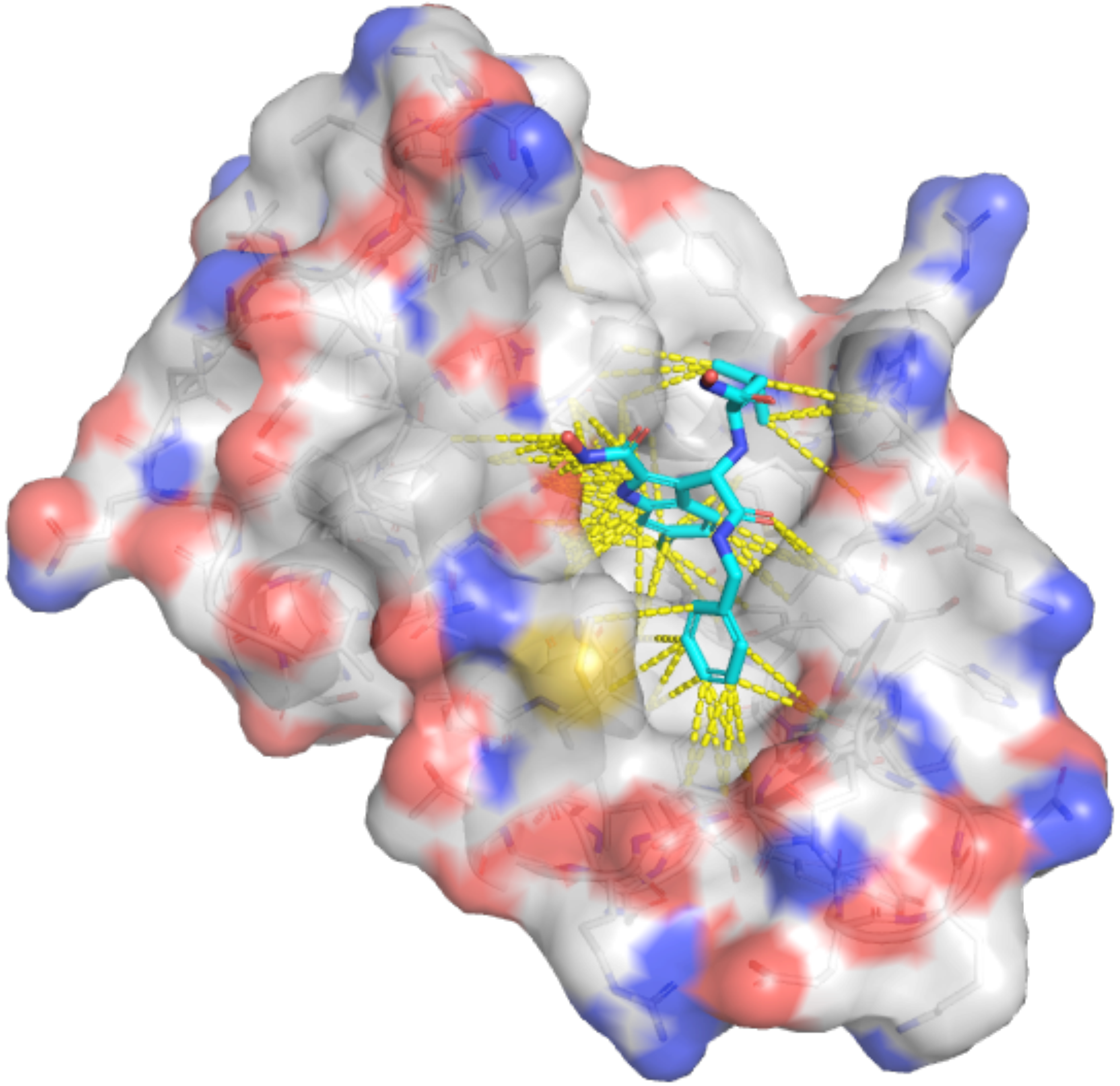
-how well does it bind?

Inactive/Active Discrimination

-does it bind?

Pose Prediction

-how does it bind?



Scoring Goals

Affinity Prediction

-how well does it bind?

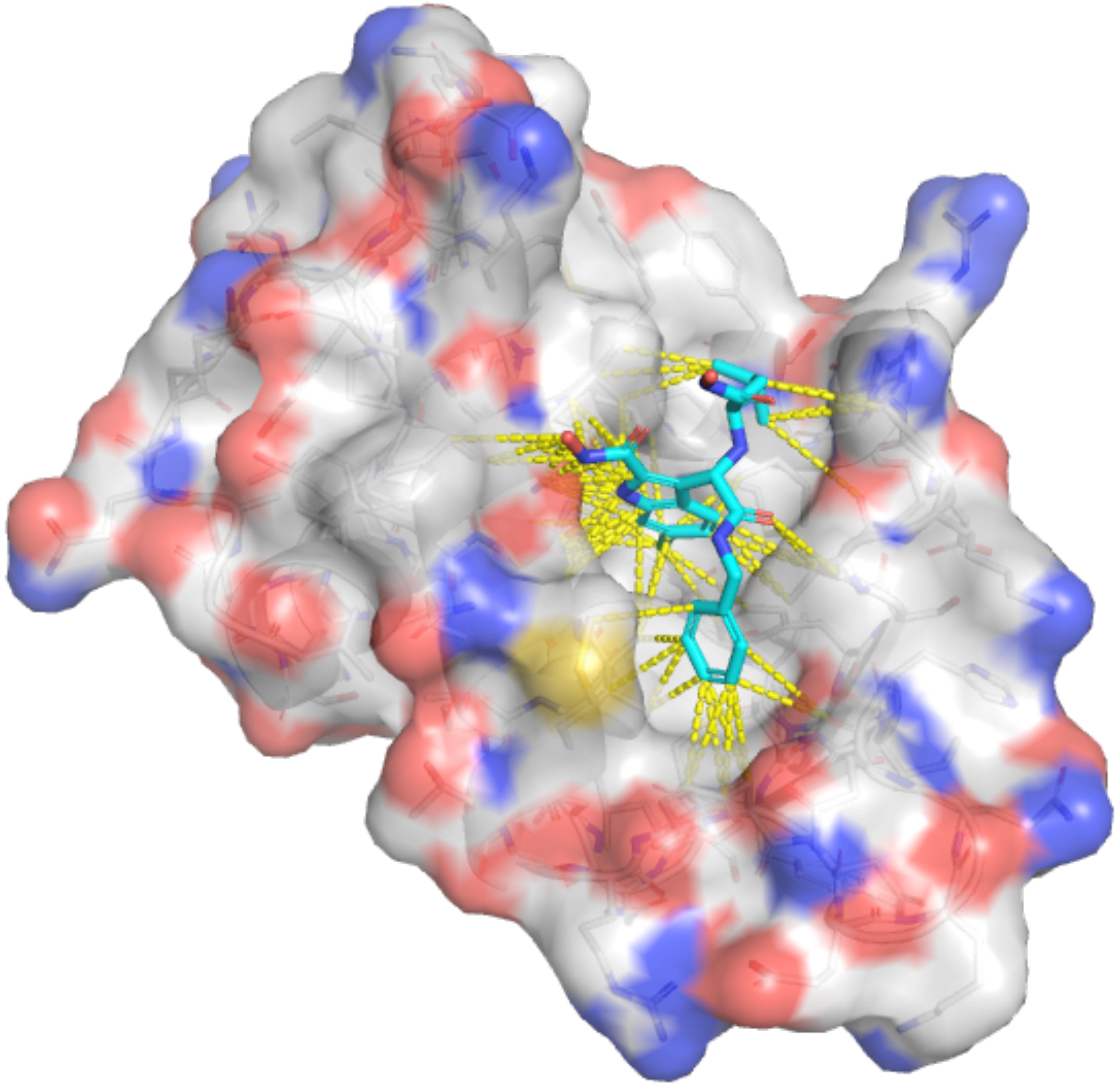
Inactive/Active Discrimination

-does it bind?

Pose Prediction

-how does it bind?

Speed



Scoring Goals

Affinity Prediction

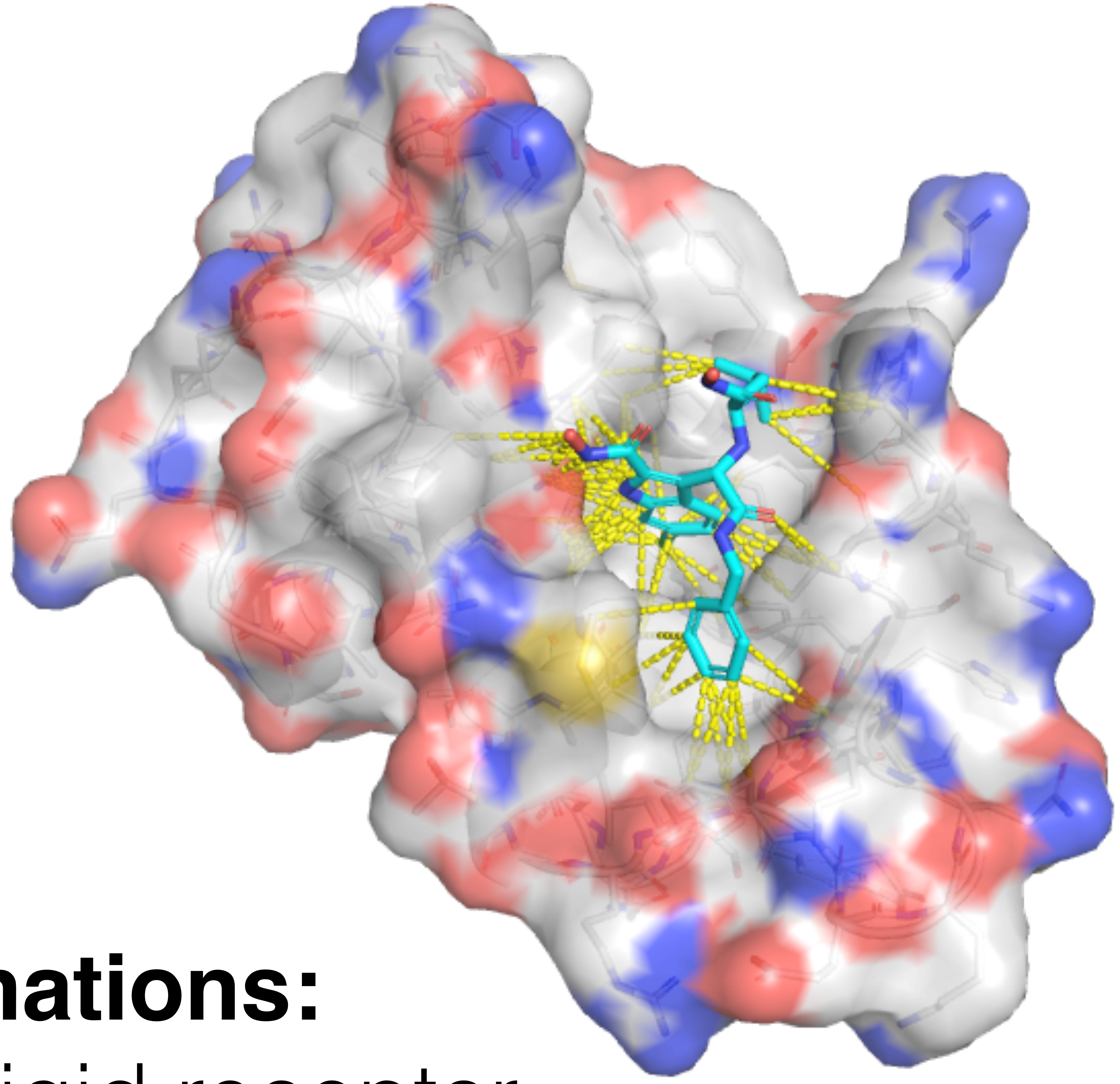
-how well does it bind?

Inactive/Active Discrimination

-does it bind?

Pose Prediction

-how does it bind?



Speed

Approximations:
Rigid or semi-rigid receptor
Implicit water model

Scoring Types

Force-field based

inter- and intra- molecular forces

van der Waals, electrostatic, torsional

Empirical

parameterized function is fit to binding energy data

Knowledge based

scoring function based on known structure, not physical principles

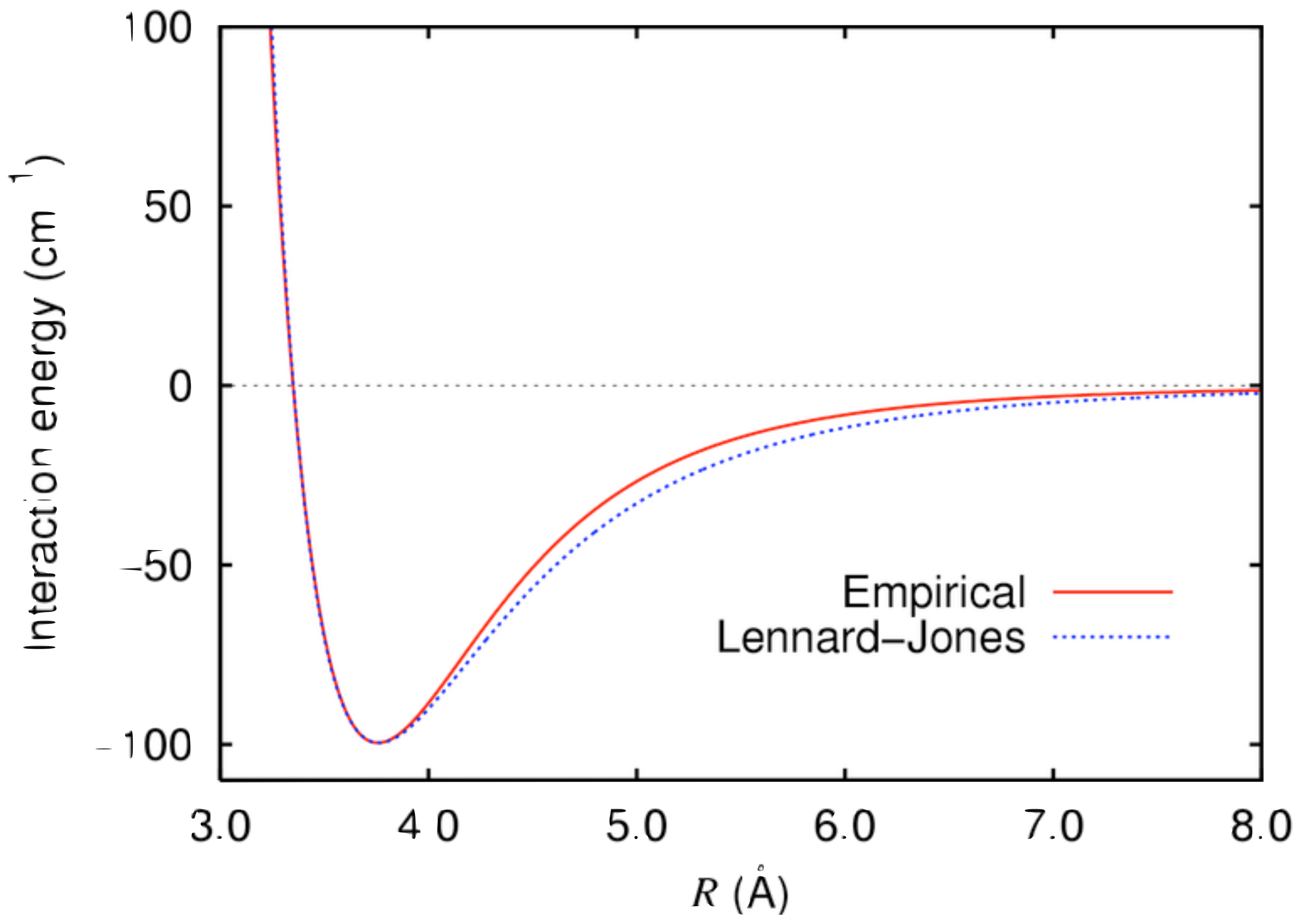
Consensus

combine multiple scoring functions

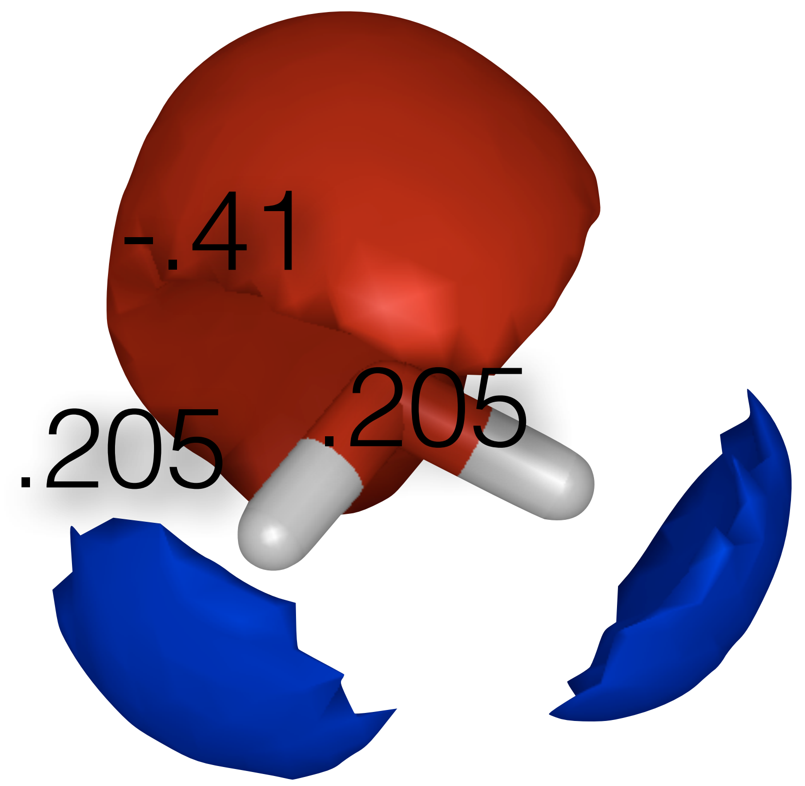
Force Field: Dock 4.0

Coulomb's Law
q: partial charges
D: dielectrict constant

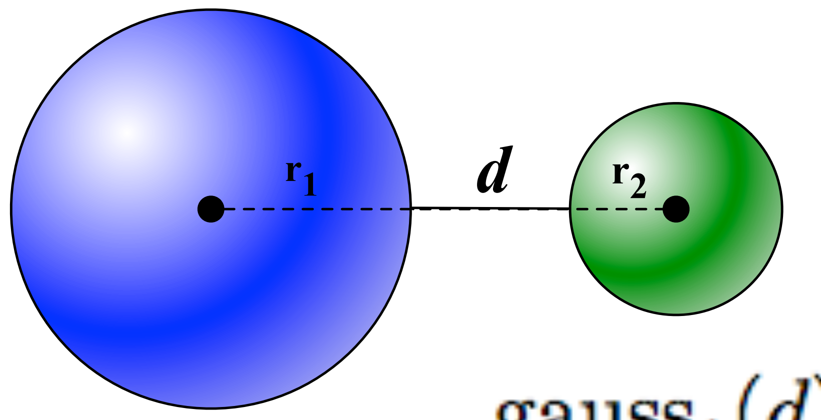
$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left(\frac{A_{ij}}{r_{ij}^{a}} - \frac{B_{ij}}{r_{ij}^{b}} + 332 \frac{q_i q_j}{D r_{ij}} \right)$$



van der Waals
a = 12, b = 6
Lennard-Jones potential



Empirical: AutoDock Vina



$$\mathrm{gauss}_1(d) = w_{\mathrm{guass}_1} e^{-(d/0.5)^2}$$

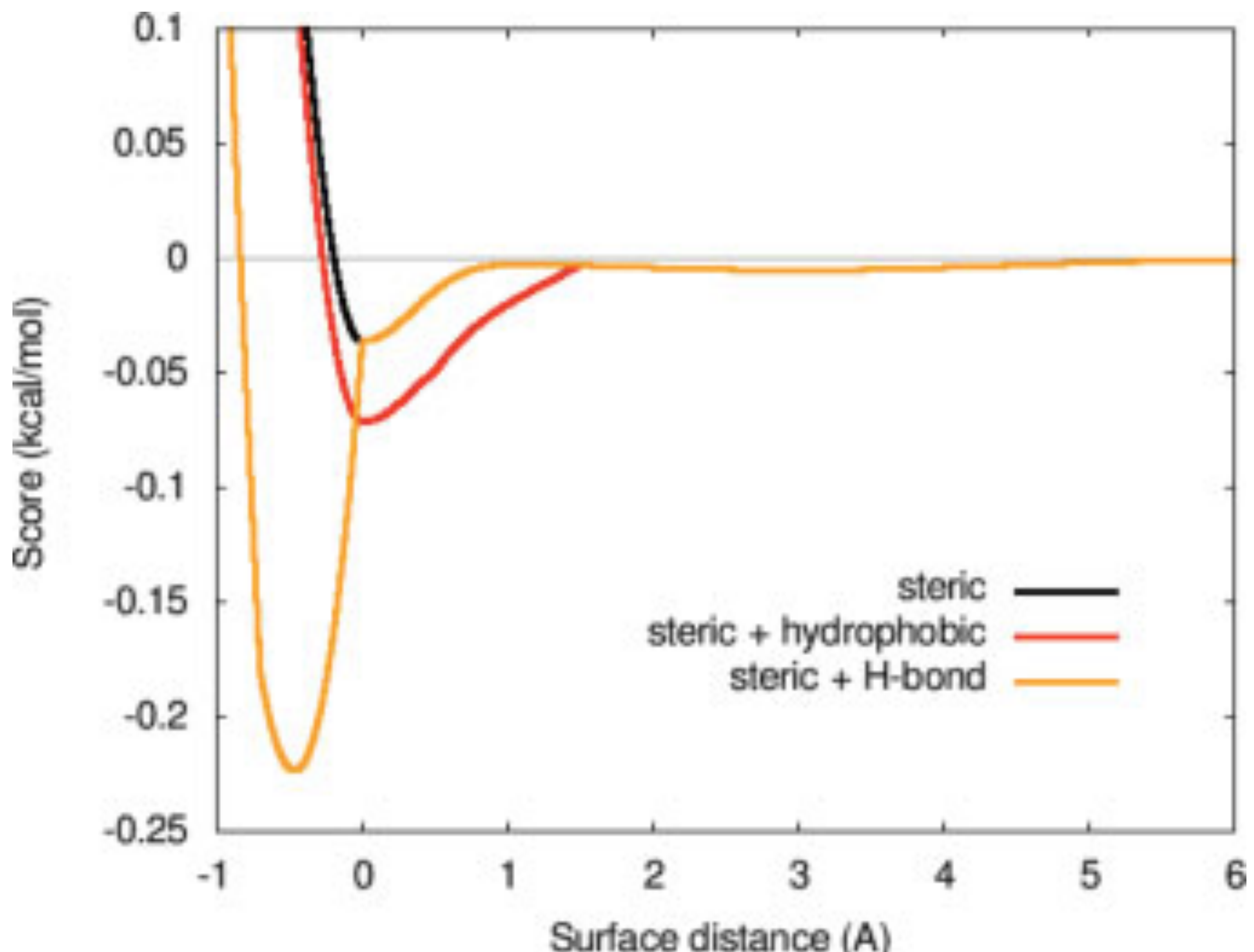
$$\mathrm{gauss}_2(d) = w_{\mathrm{guass}_2} e^{-((d-3)/2)^2}$$

$$\mathrm{repulsion}(d) = \begin{cases} w_{\mathrm{repulsion}} d^2 & d < 0 \\ 0 & d \geq 0 \end{cases}$$

$$\mathrm{hydrophobic}(d) = \begin{cases} w_{\mathrm{hydrophobic}} & d < 0.5 \\ 0 & d > 1.5 \\ w_{\mathrm{hydrophobic}}(1.5 - d) & \textit{otherwise} \end{cases}$$

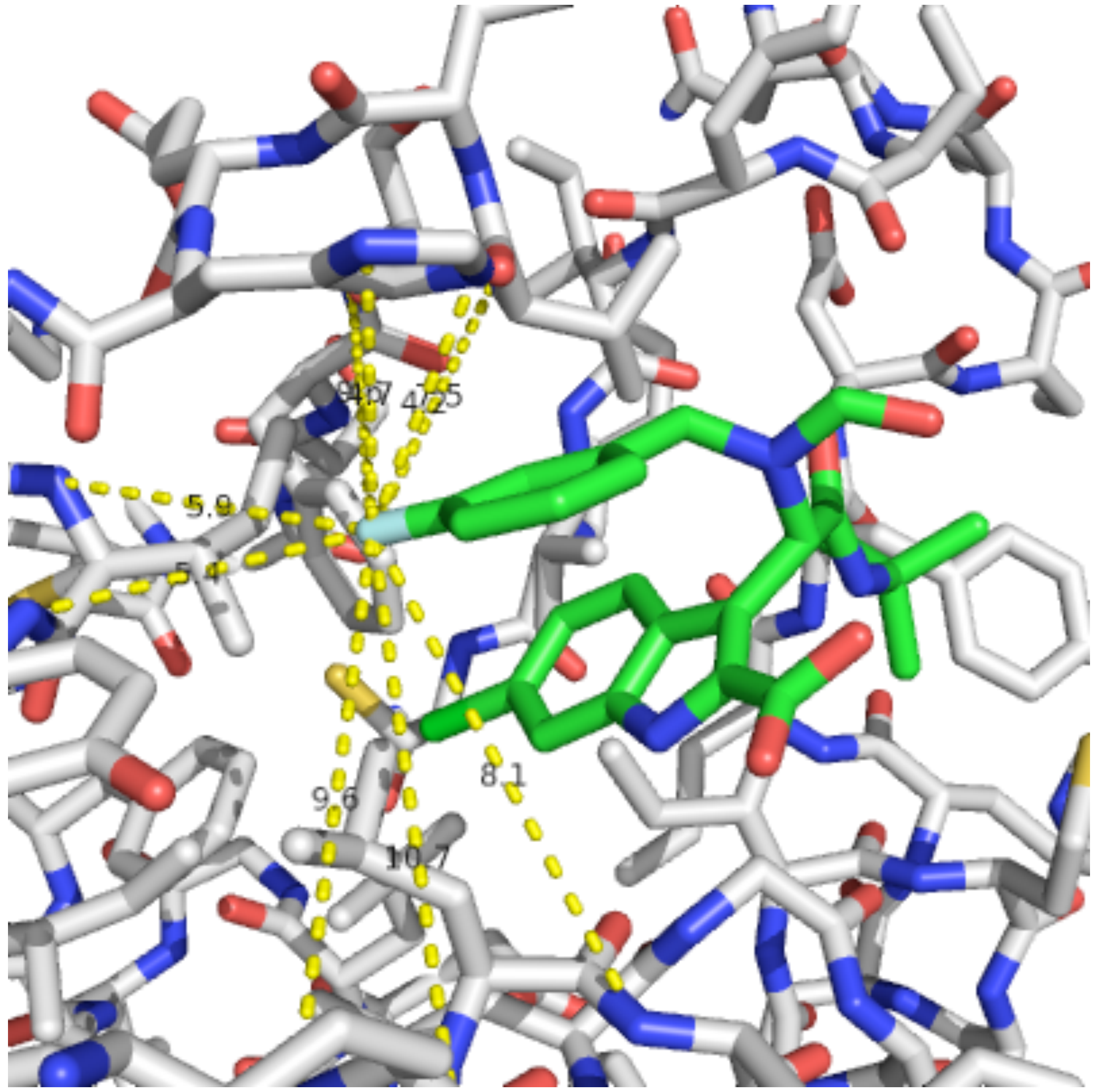
$$\mathrm{hbond}(d) = \begin{cases} w_{\mathrm{hbond}} & d < -0.7 \\ 0 & d > 0 \\ w_{\mathrm{hbond}}(-\frac{10}{7} d) & \textit{otherwise} \end{cases}$$

Weight	Term
−0.0356	gauss_1
−0.00516	gauss_2
0.840	Repulsion
−0.0351	Hydrophobic
−0.587	Hydrogen bonding
0.0585	N_{rot}



Knowledge Based: RF-Score

Pairwise Distance Counts (<12Å)



BIOINFORMATICS **ORIGINAL PAPER**

Vol. 26 no. 9 2010, pages 1169–1175
doi:10.1093/bioinformatics/btq112

Structural bioinformatics

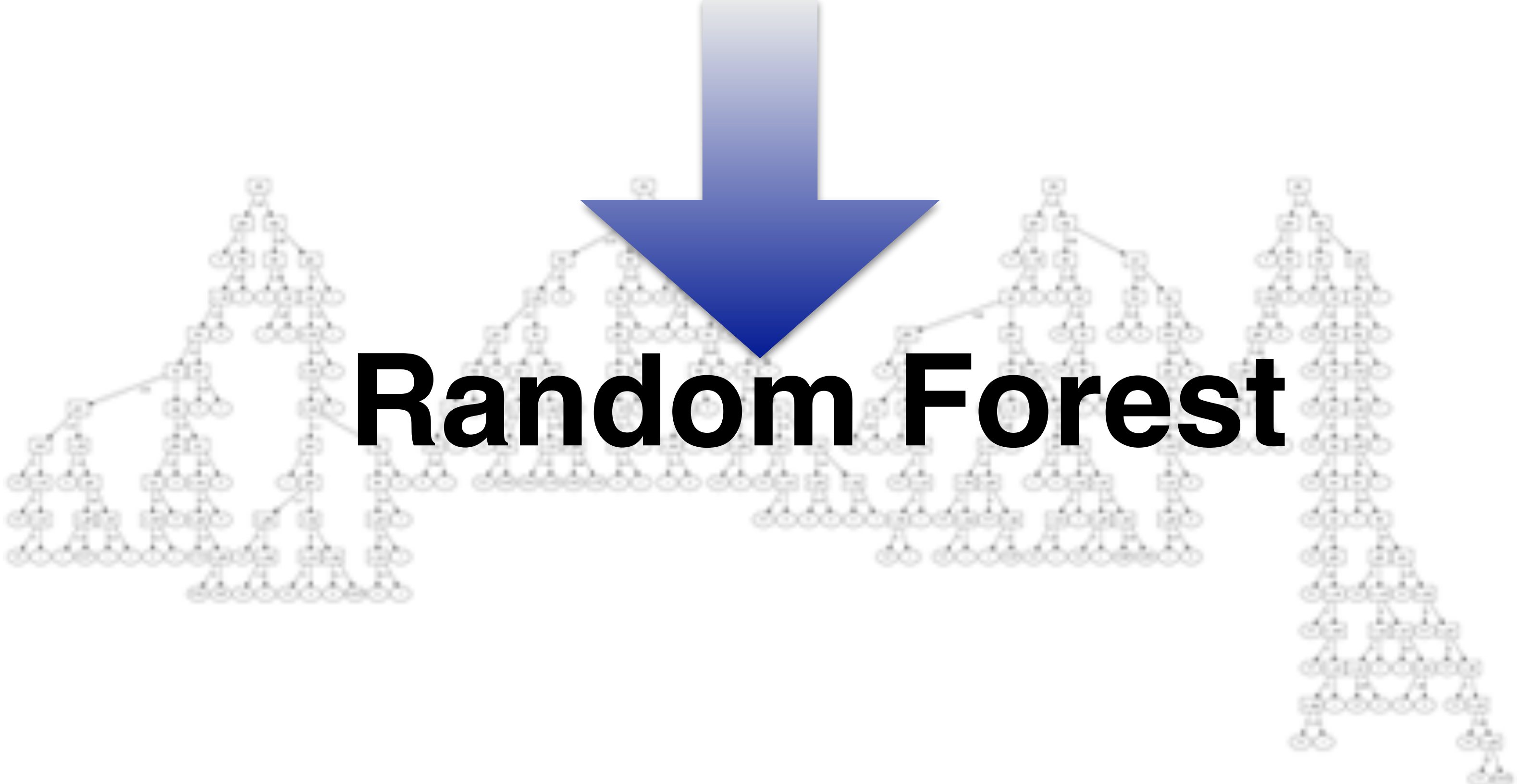
Advance Access publication March 17, 2010

A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking

Pedro J. Ballester[1,*,†] and John B. O. Mitchell[2,*]
[1]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW and [2]Centre for Biomolecular Sciences, University of St Andrews, North Haugh, St Andrews KY16 9ST, UK
Associate Editor: Burkhard Rost

Protein

Ligand	C	N	O	S
C				
N				
O				
S				
P				
F		9		
Cl				
Br				
I				



Can we do better?

Accurate pose prediction, binding discrimination, **and** affinity prediction without sacrificing performance?

Can we do better?

Accurate pose prediction, binding discrimination, **and** affinity prediction without sacrificing performance?

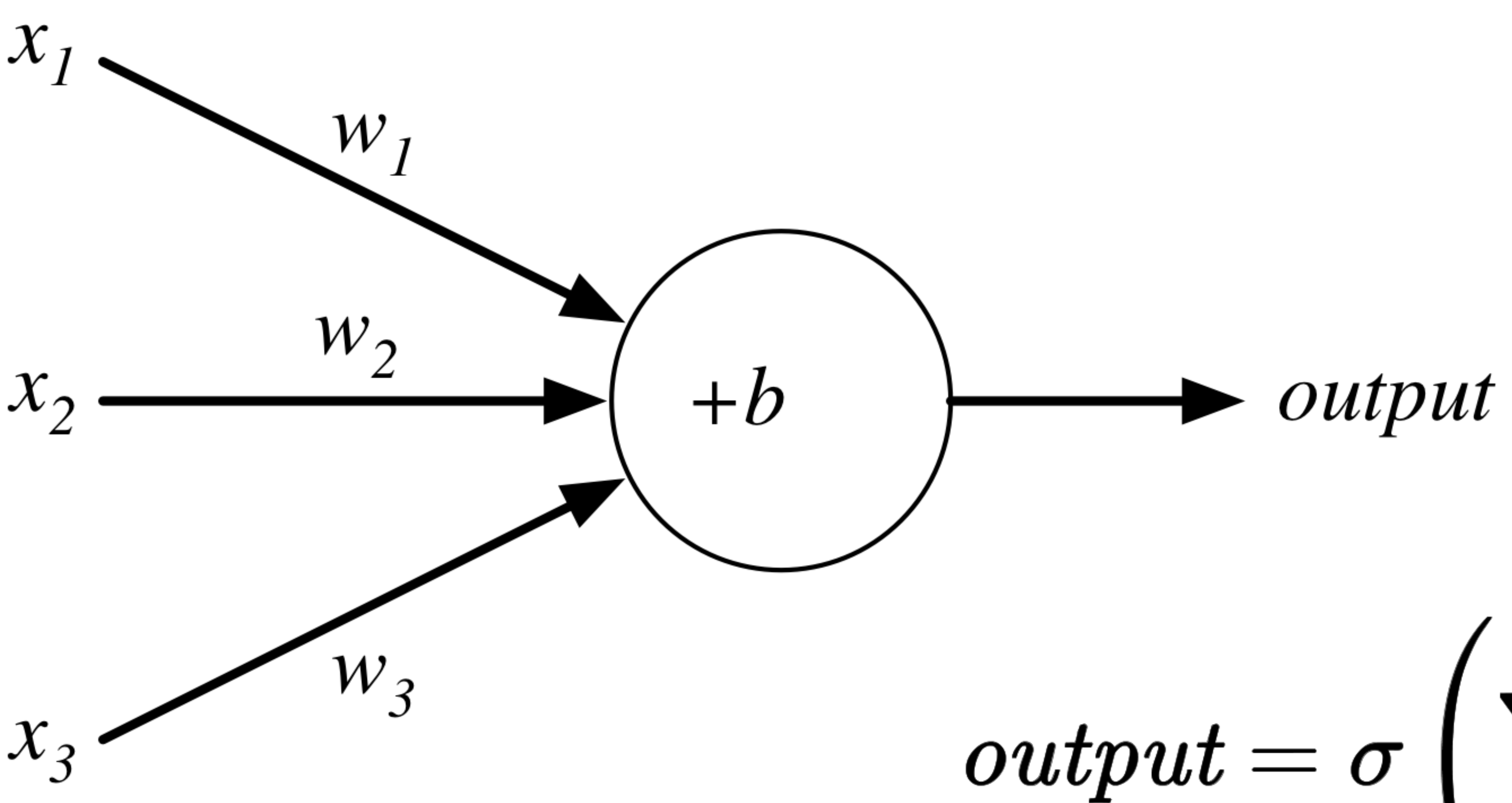
Key Idea: Leverage "big data"

- 231,655,275 bioactivities in PubChem
- 125,526 structures in the PDB
- 16,179 annotated complexes in PDBbind

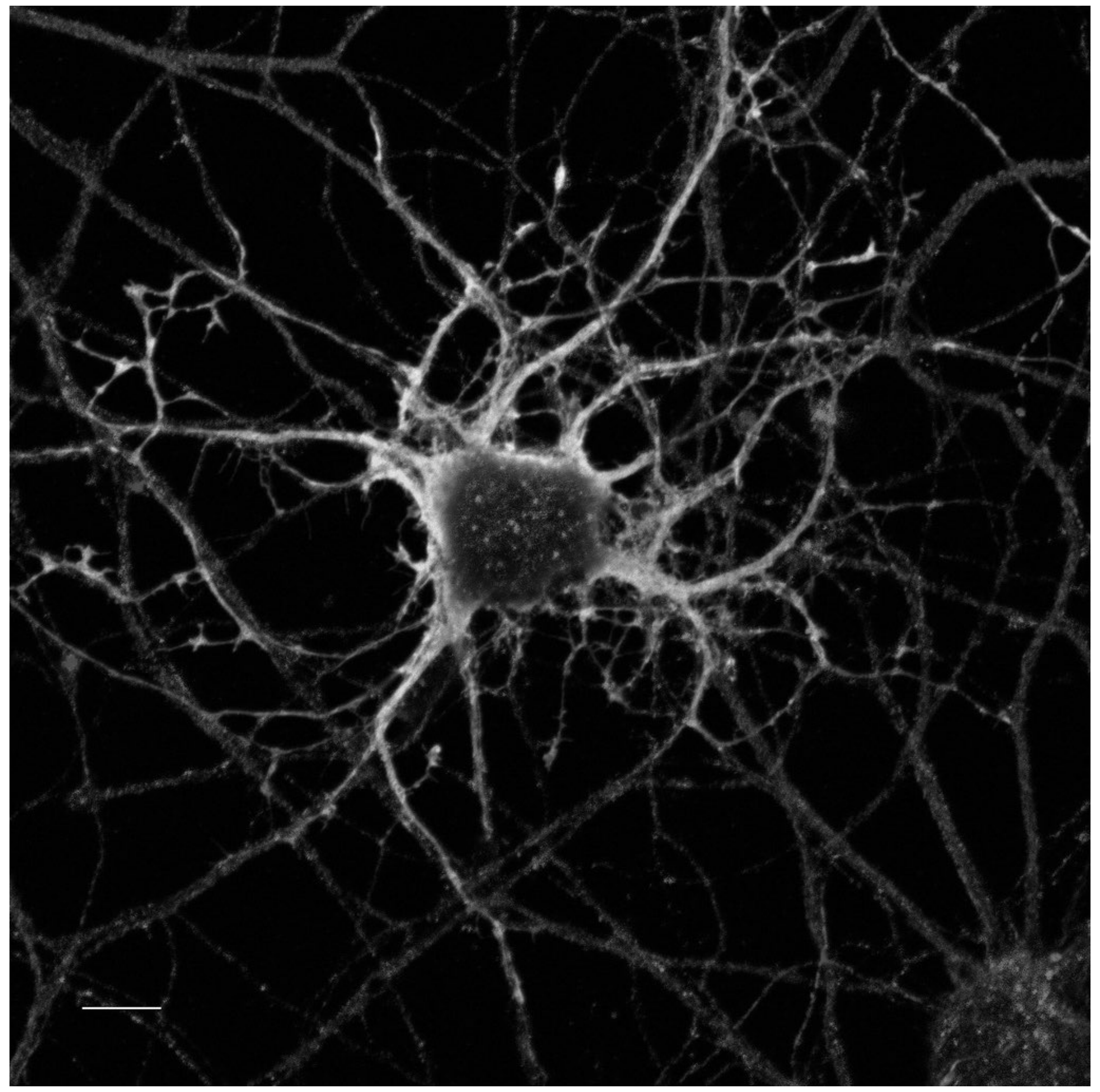
Machine Learning

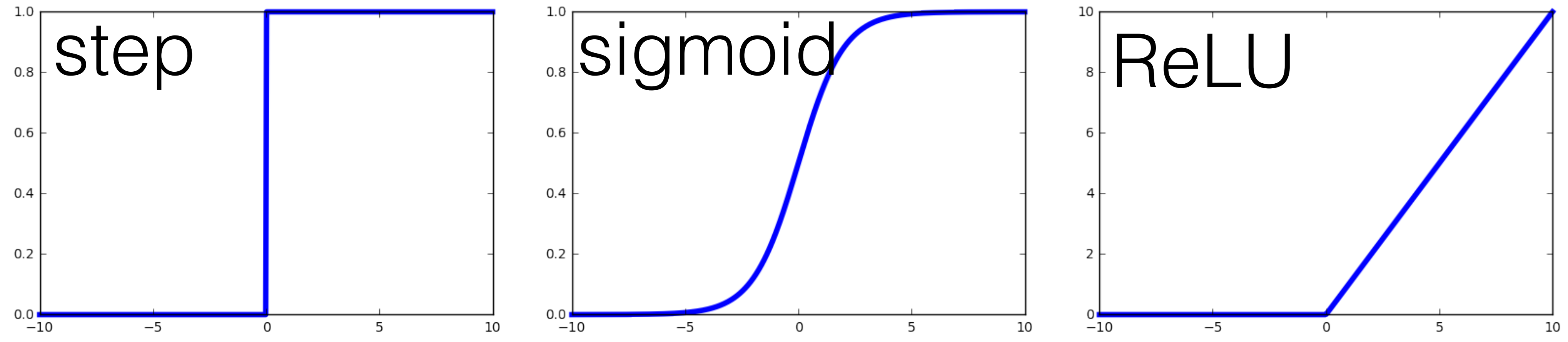
Features $X \rightarrow$ **Model** $\rightarrow y$ Prediction

Neural Networks

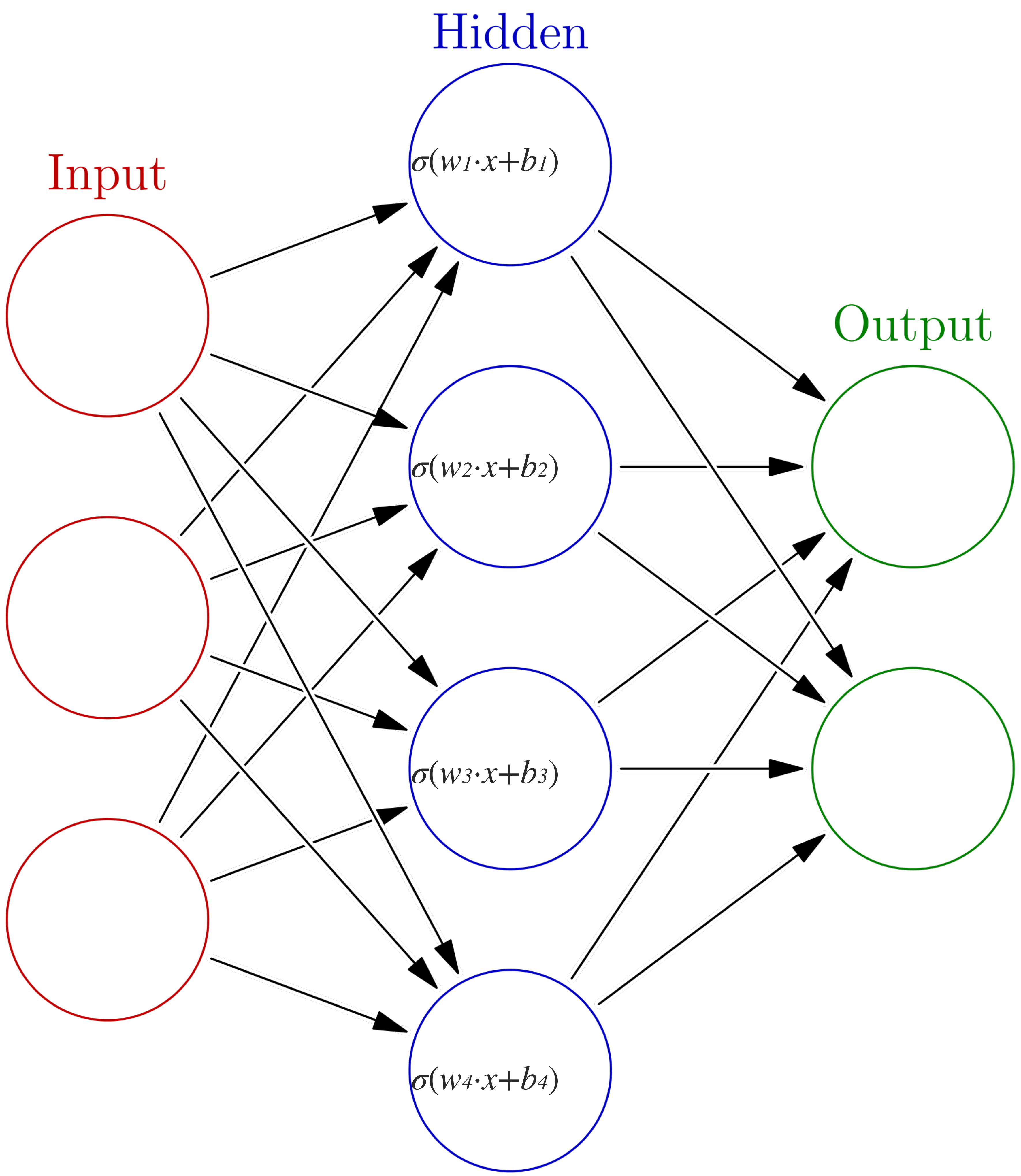


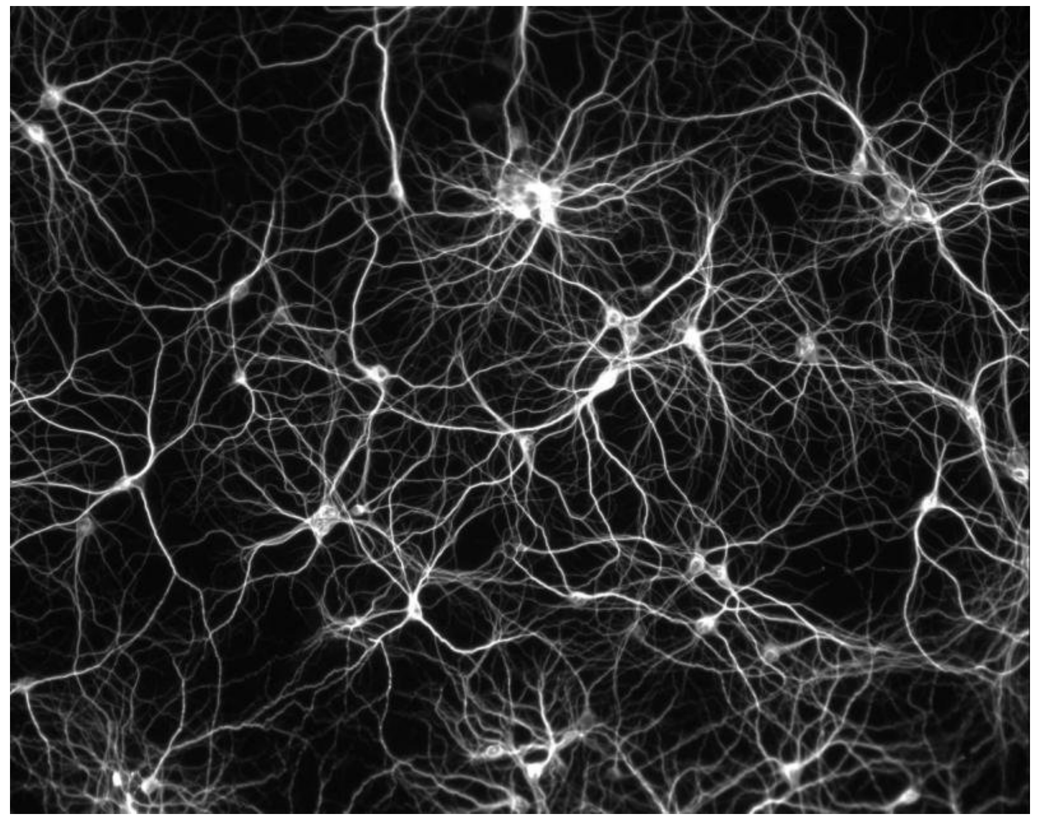
$$output = \sigma\left(\sum_i w_i x_i + b\right)$$





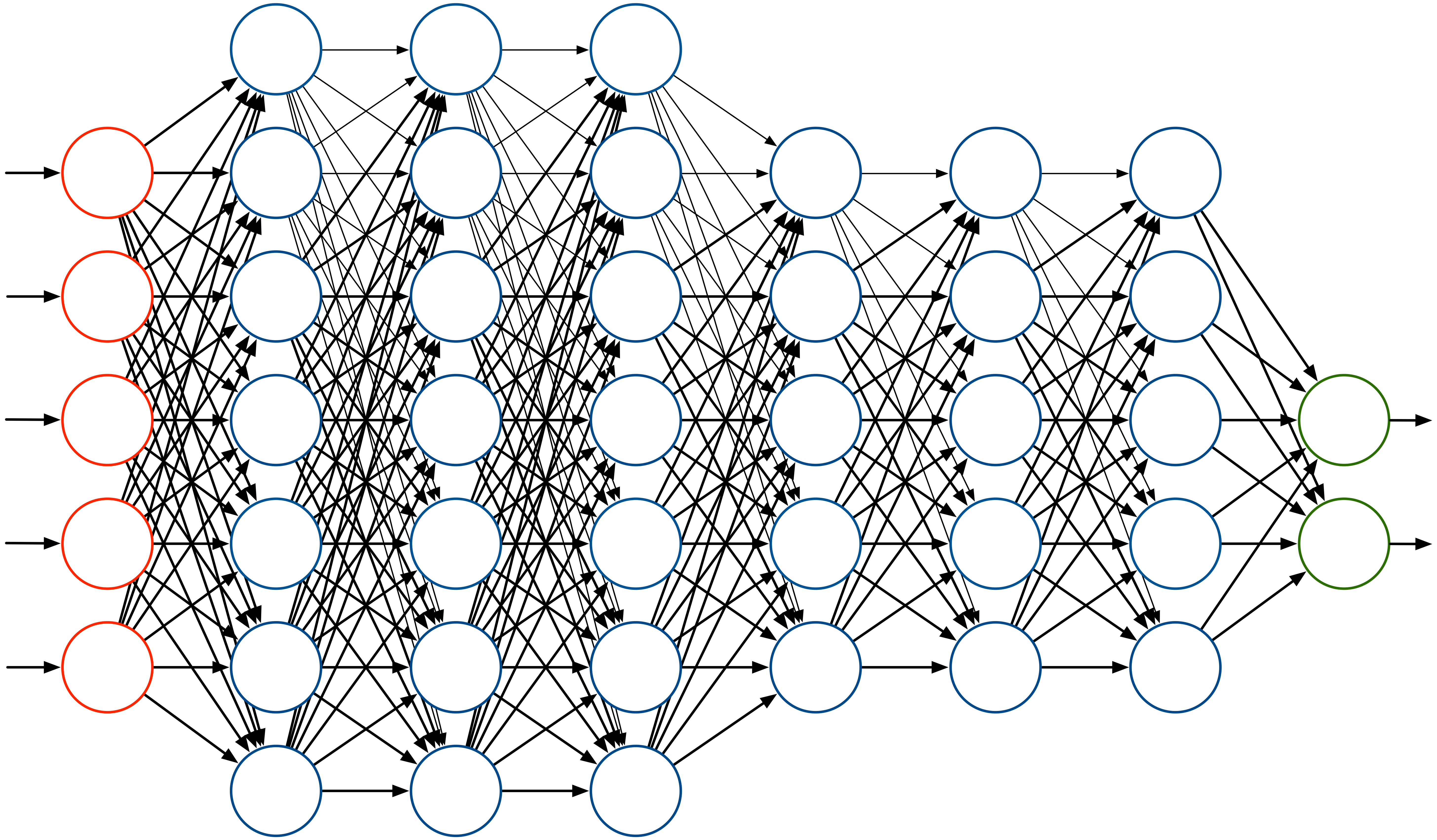
Neural Networks



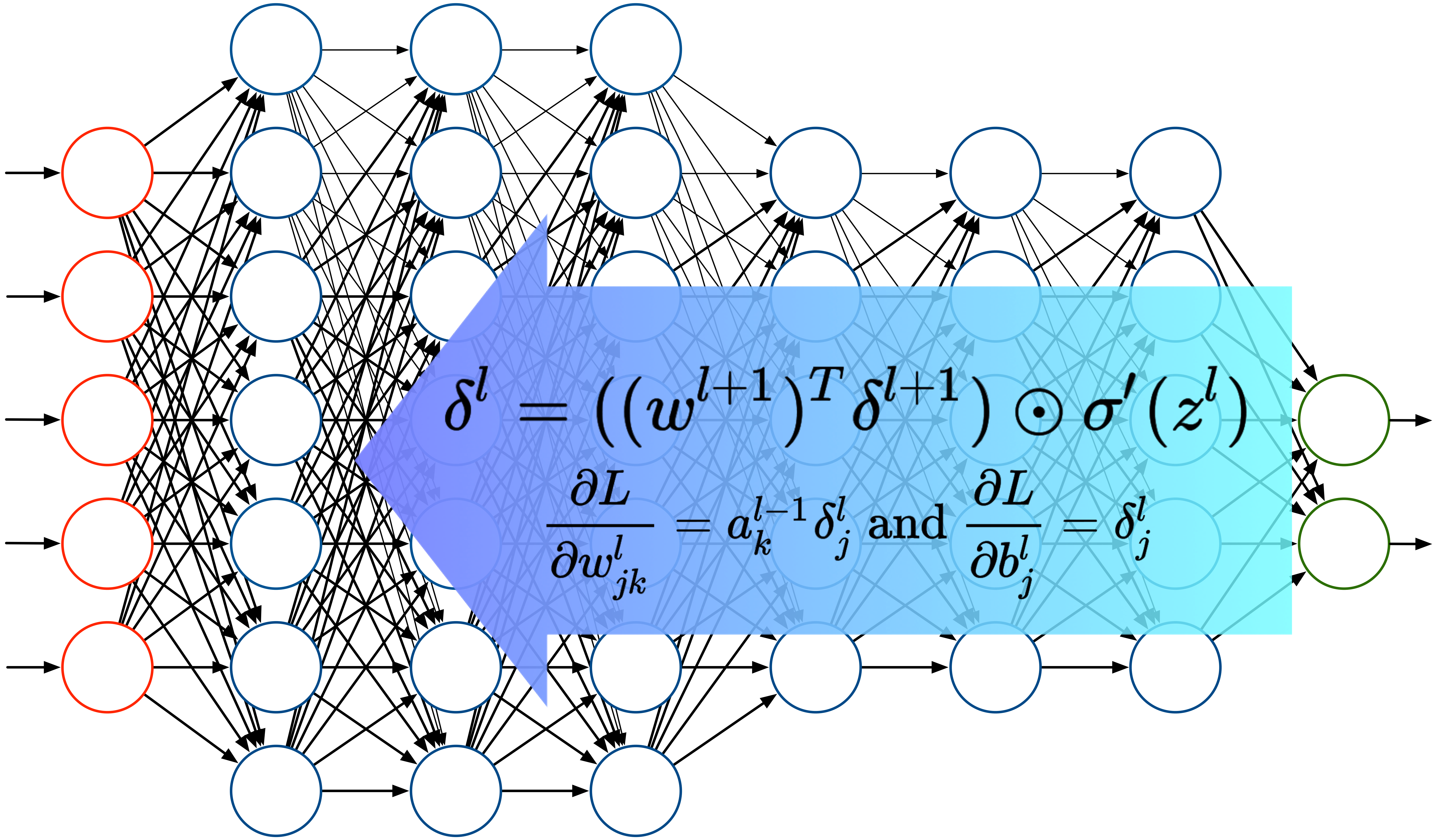


The **universal approximation theorem** states that, under reasonable assumptions, a feedforward **neural network** with a finite number of nodes **can approximate any continuous** function to within a given error over a bounded input domain.

Deep Learning



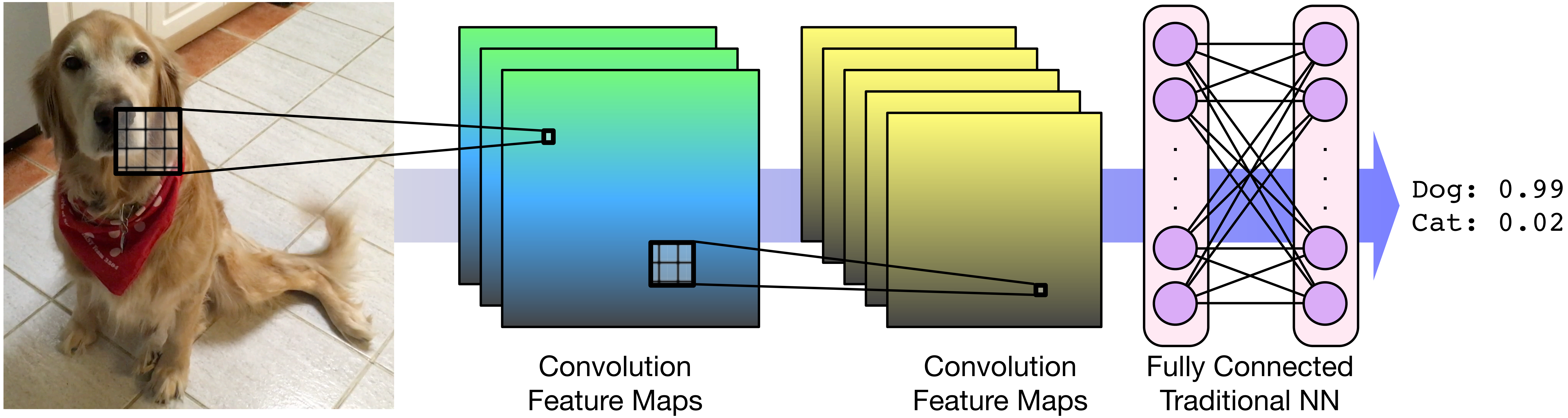
Deep Learning



$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

$$\frac{\partial L}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j \text{ and } \frac{\partial L}{\partial b^l_j} = \delta^l_j$$

Convolutional Neural Networks



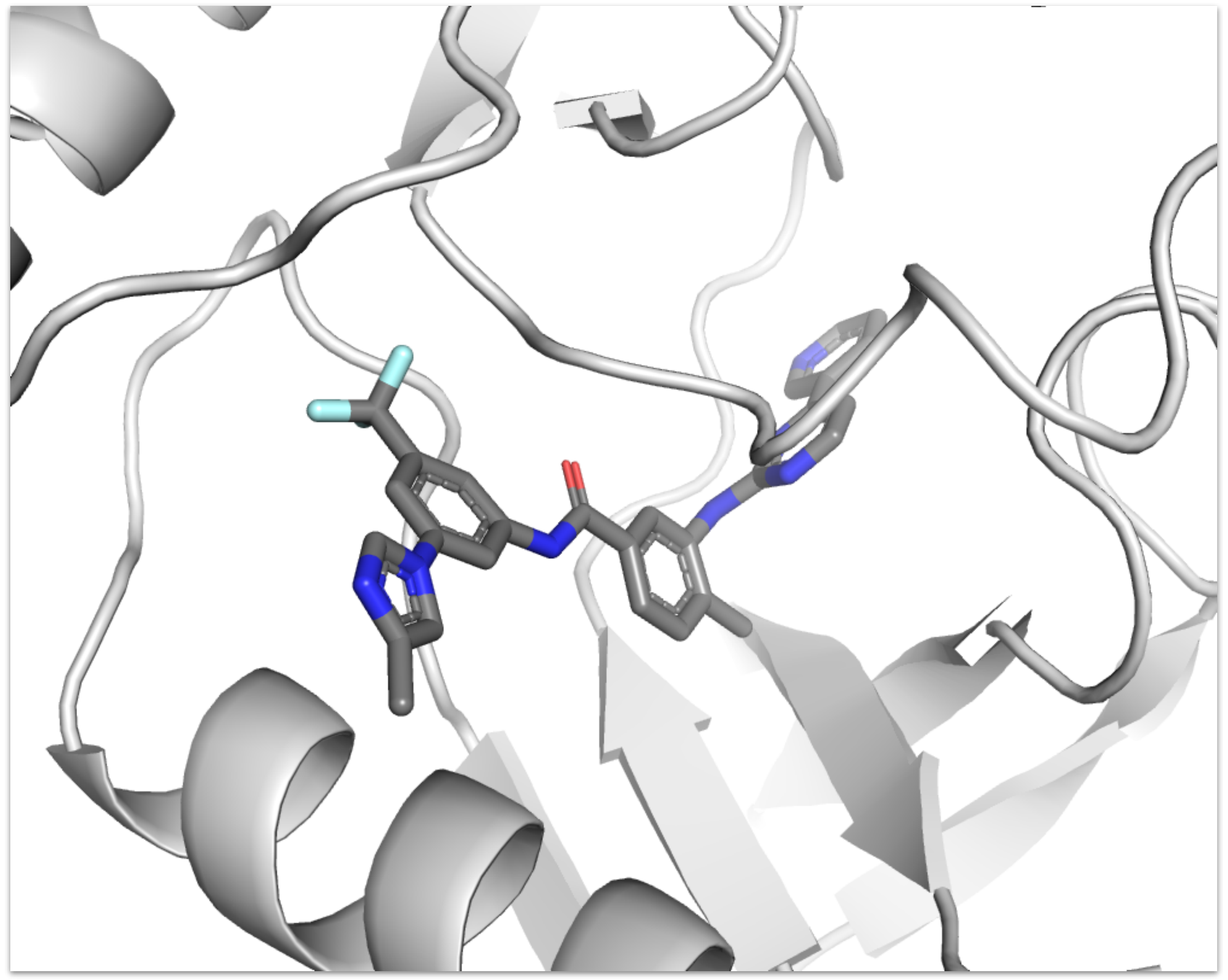
Convolutional Filters

-1	-1	-1
0	0	0
1	1	1

-1	0	1
-1	0	1
-1	0	1

-1	-1	-1
-1	8	-1
-1	-1	-1

CNNs for Protein-Ligand Scoring



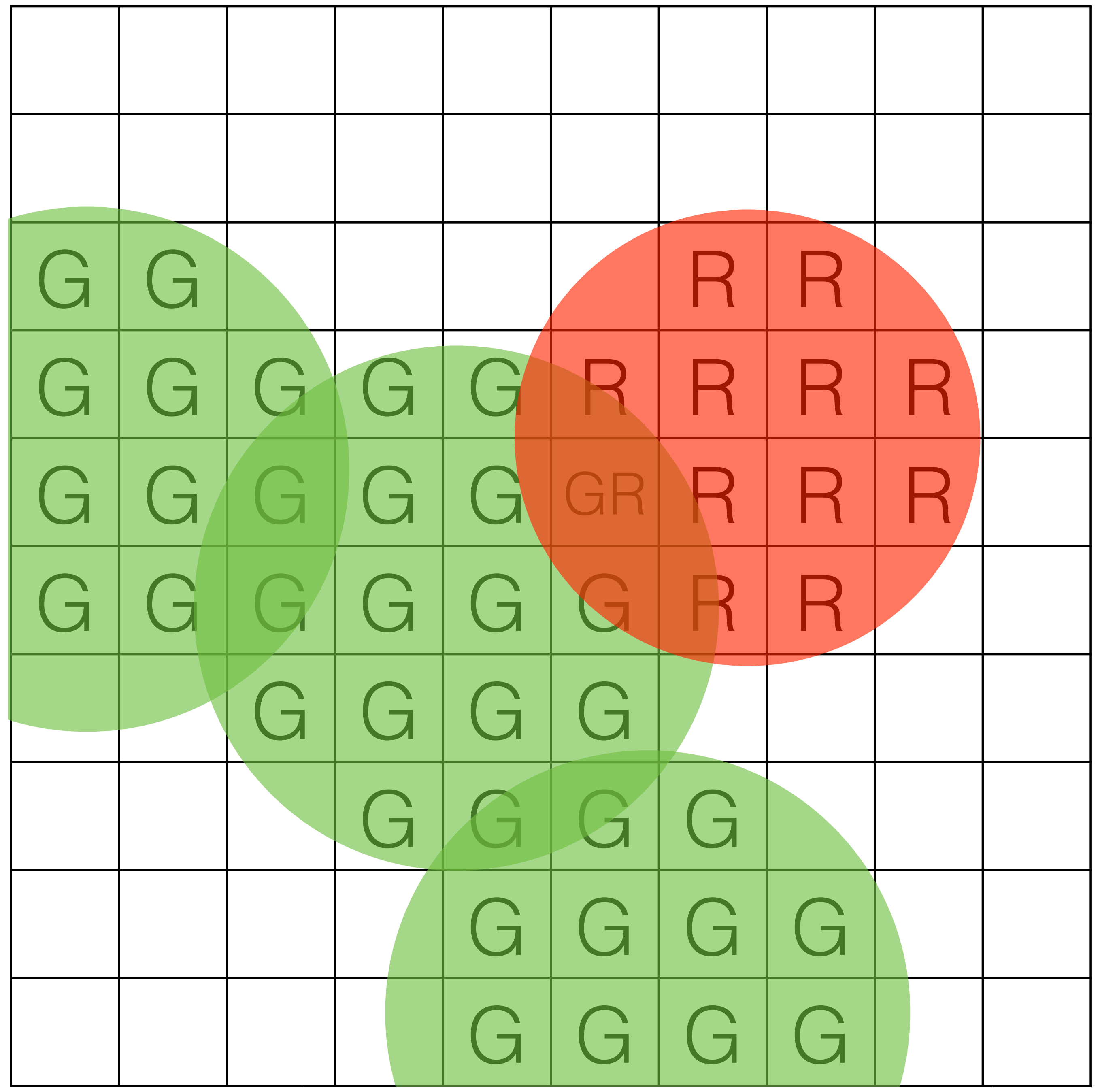
CNN

Pose Prediction

Binding Discrimination

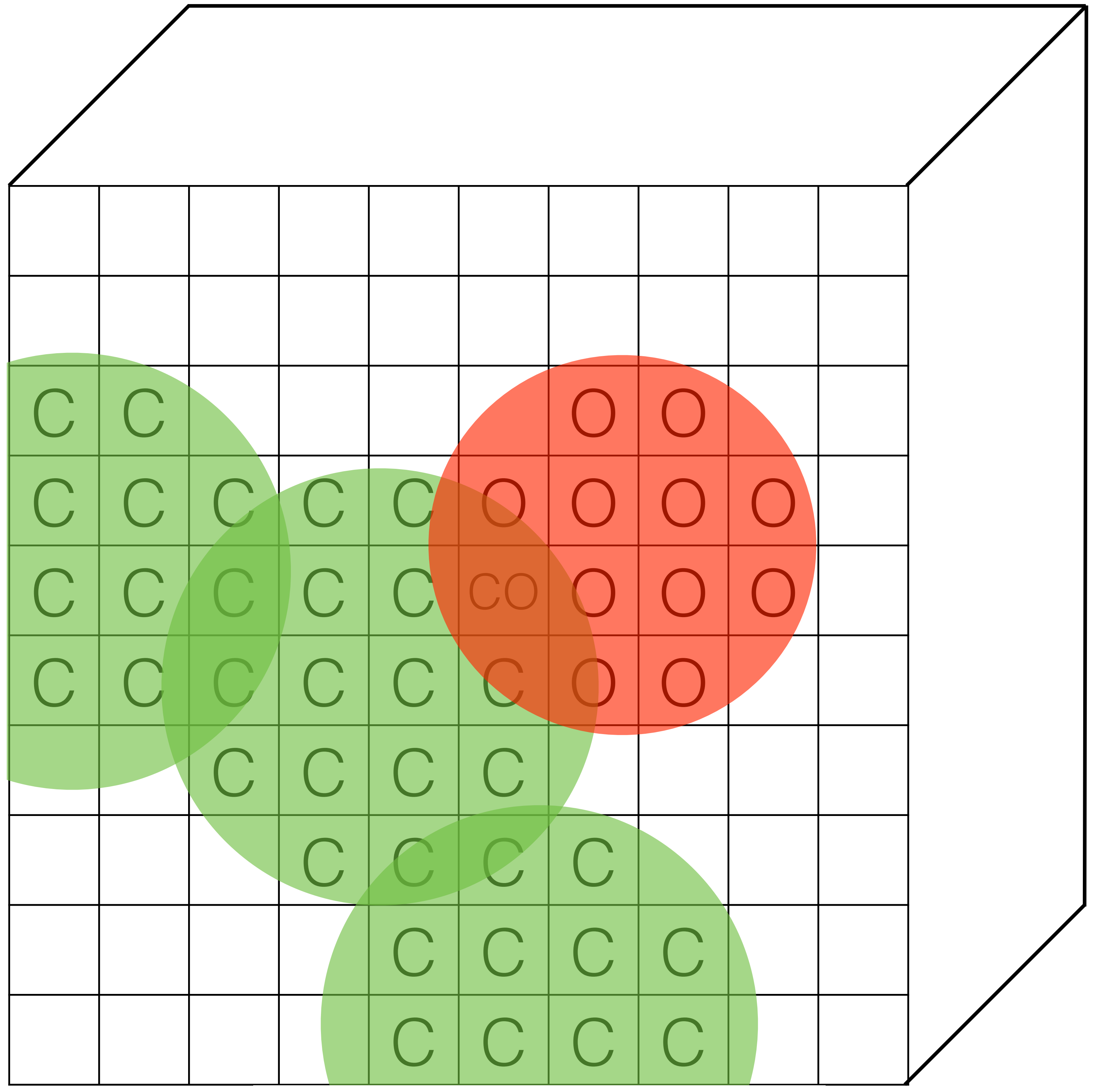
Affinity Prediction

Protein-Ligand Representation



(R,G,B) pixel

Protein-Ligand Representation

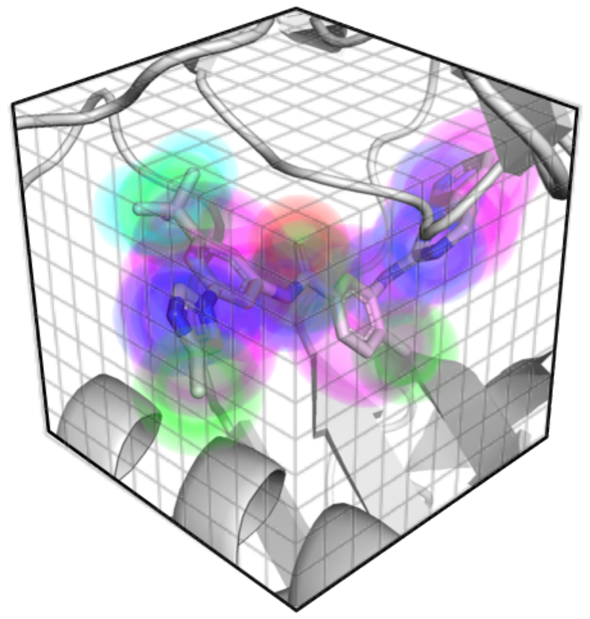


(R,G,B) pixel →

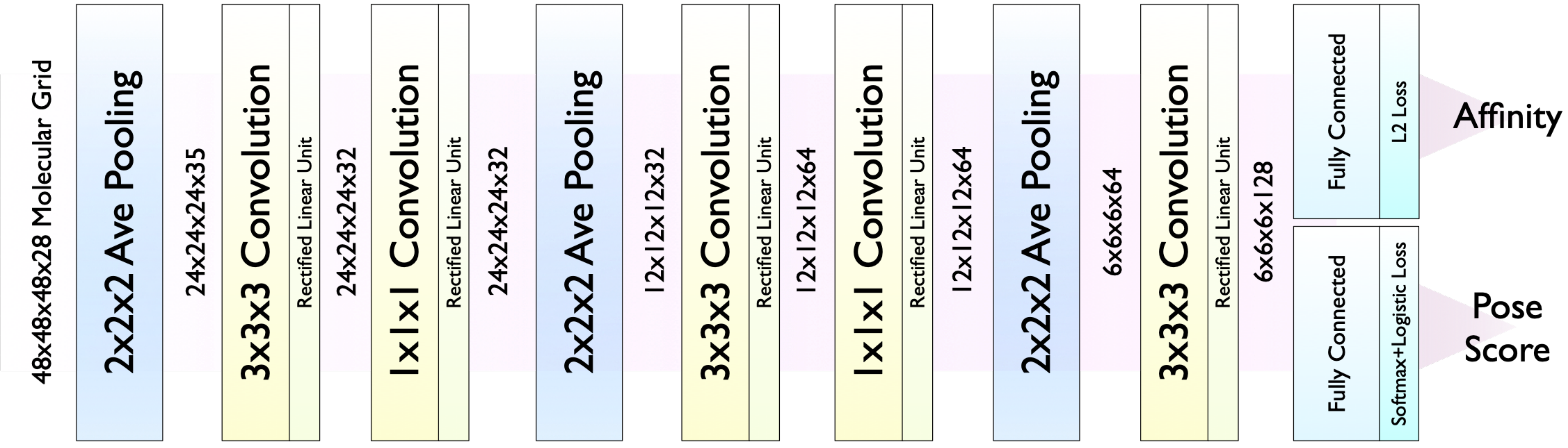
(Carbon, Nitrogen, Oxygen,…) **voxel**

The only parameters for this representation are the choice of **grid resolution**, **atom density**, and **atom types**.

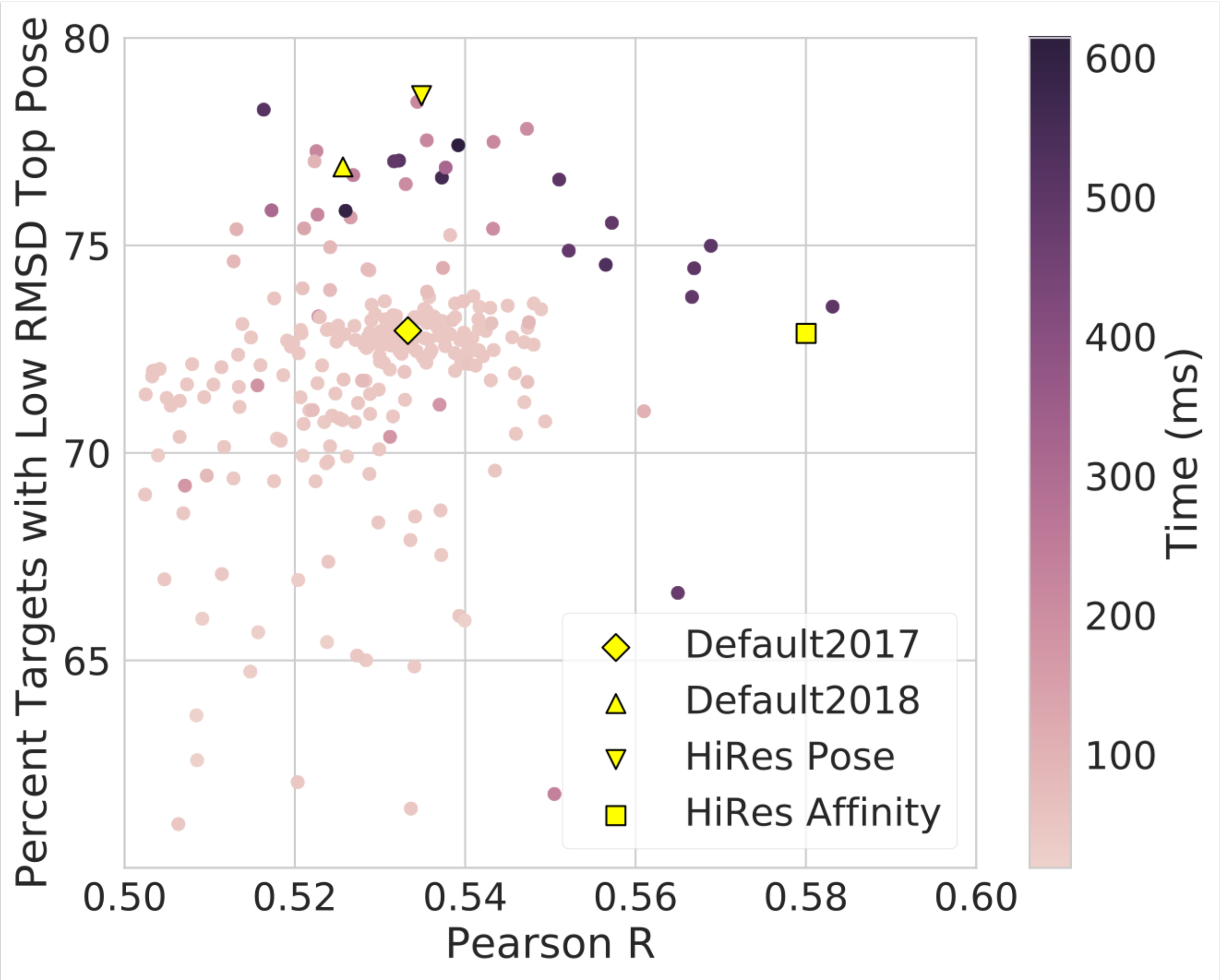
Protein Ligand Scoring



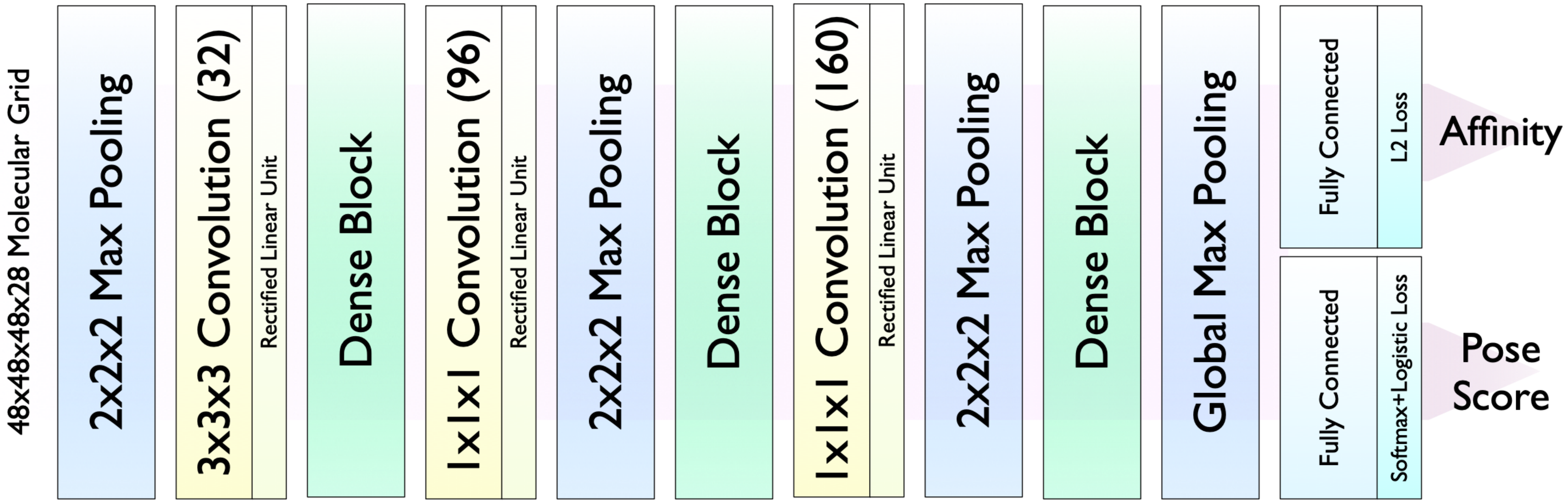
Def2018

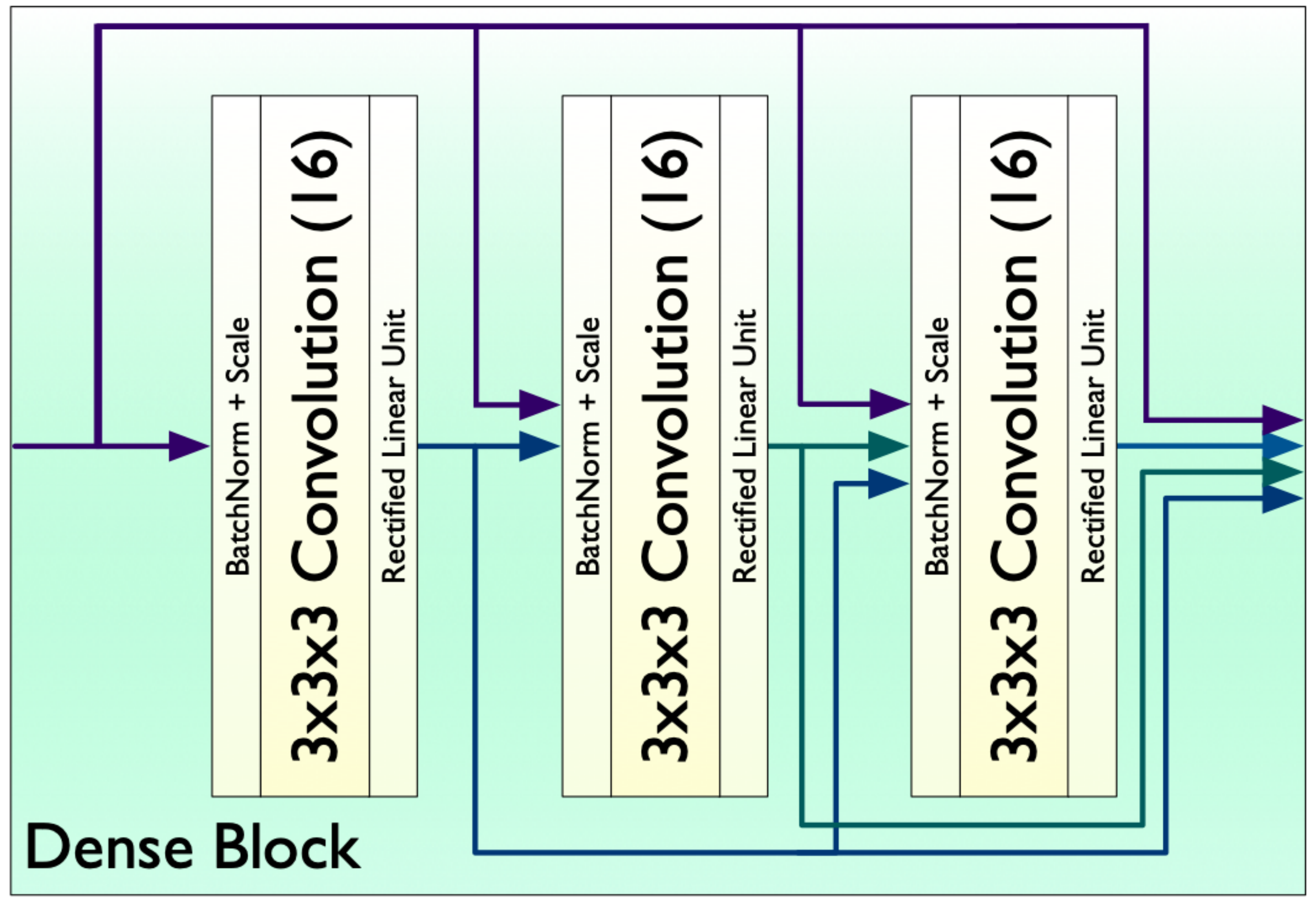


388,736 Parameters



Dense





684,640 Parameters

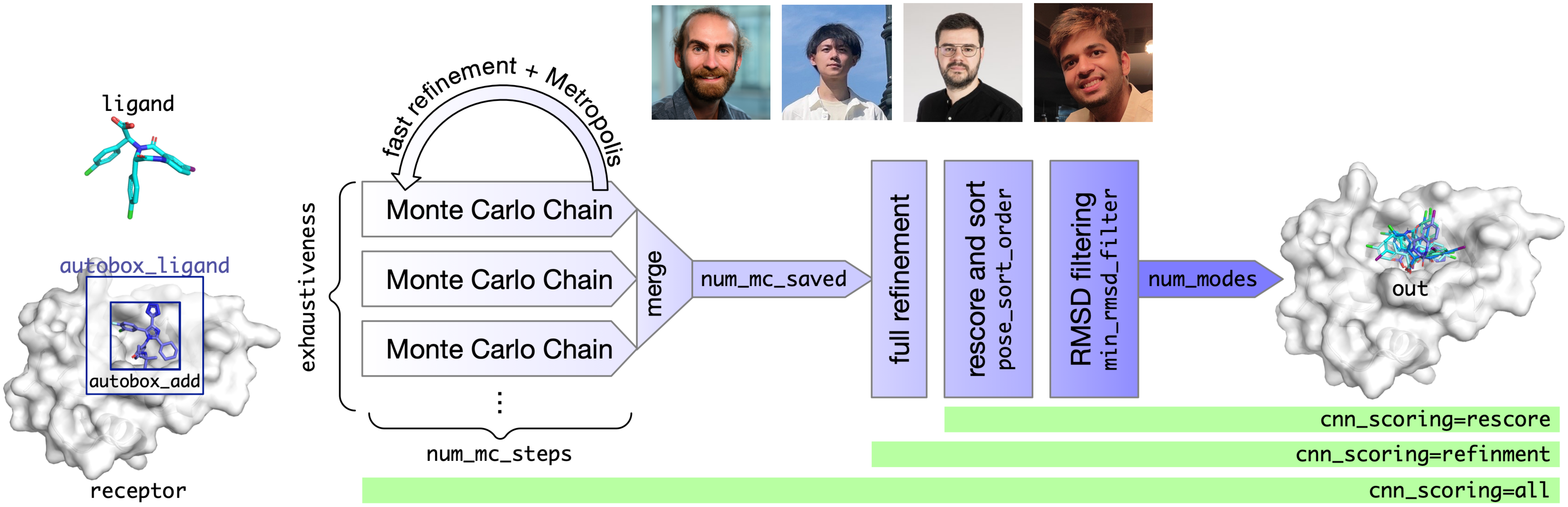
GNINA 1.3

https://github.com/gnina/gnina

GNINA 1.3: the next increment in molecular docking with deep learning

Andrew T. McNutt, Yanjing Li, Rocco Meli, Rishal Aggarwal & David Ryan Koes

Journal of Cheminformatics **17**, Article number: 28 (2025) | Cite this article

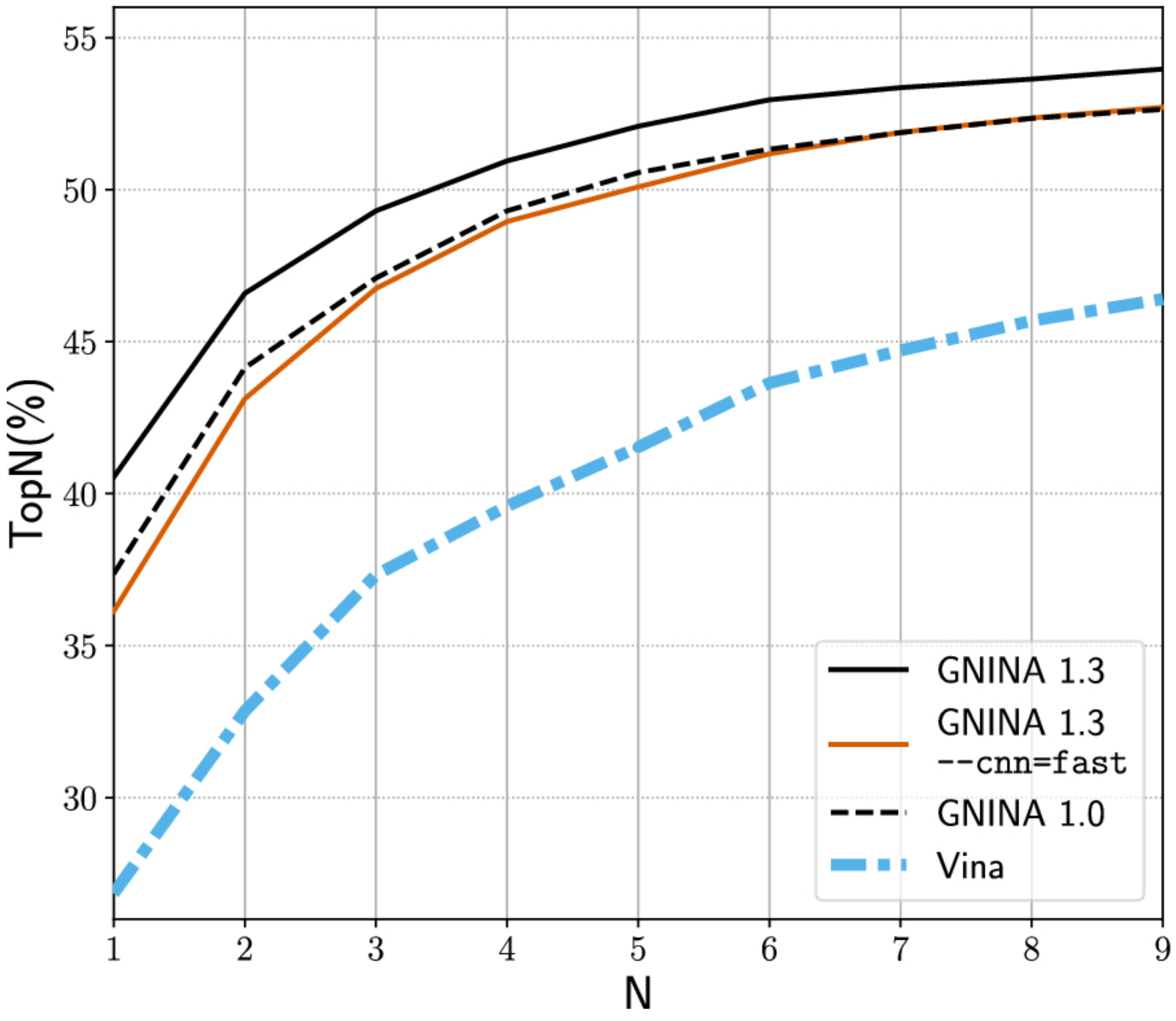


Caffe → Torch

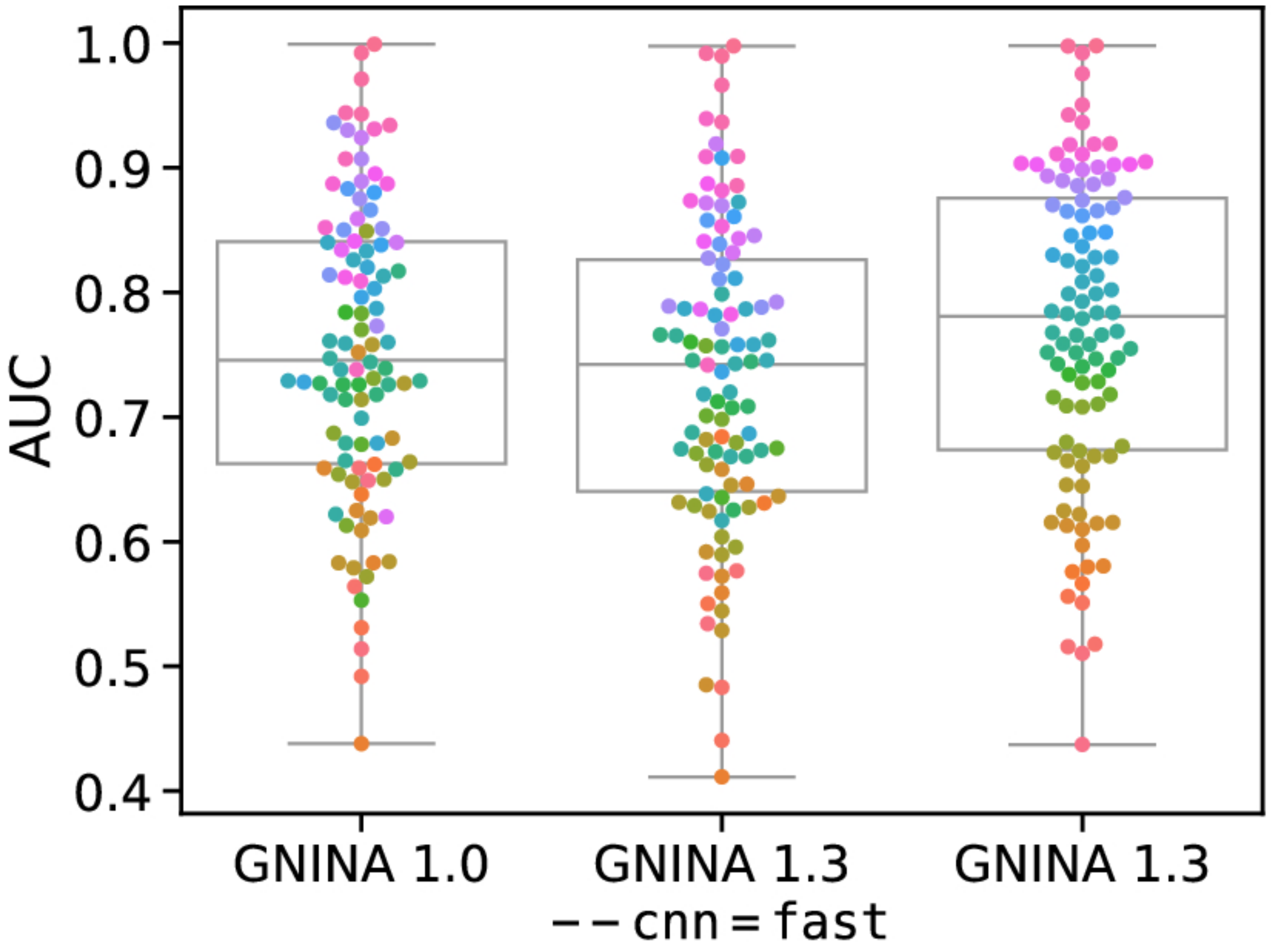
easy covalent docking

retrained models

GNINA 1.3 Performance

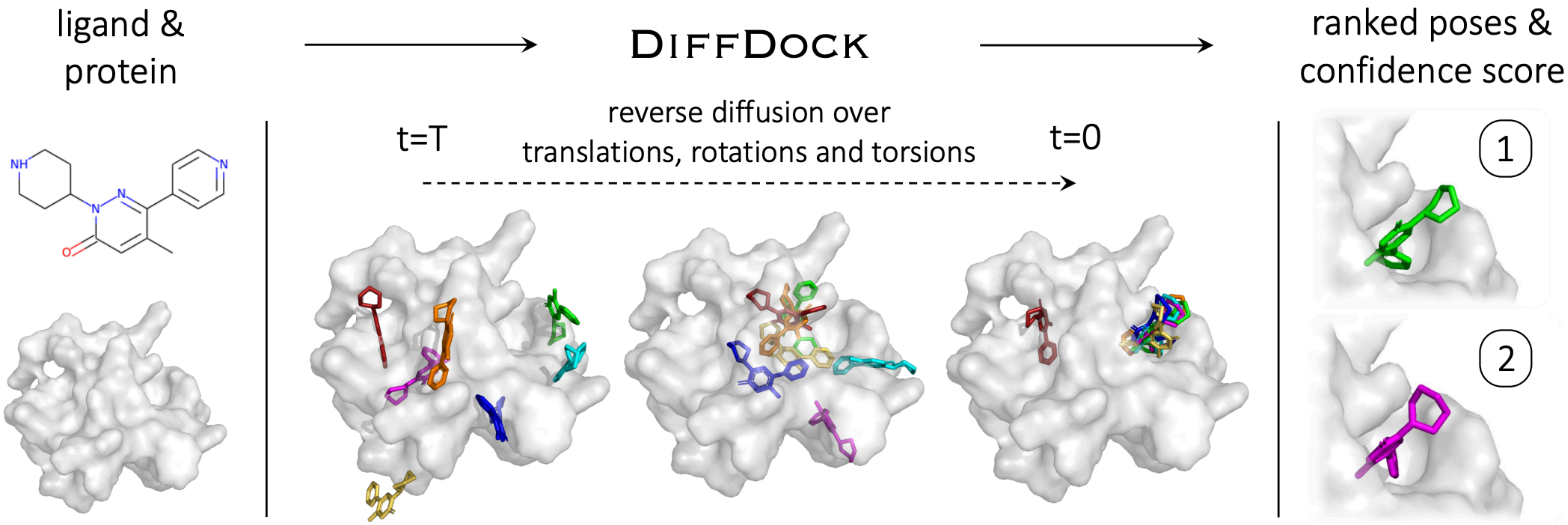


Crossdocking

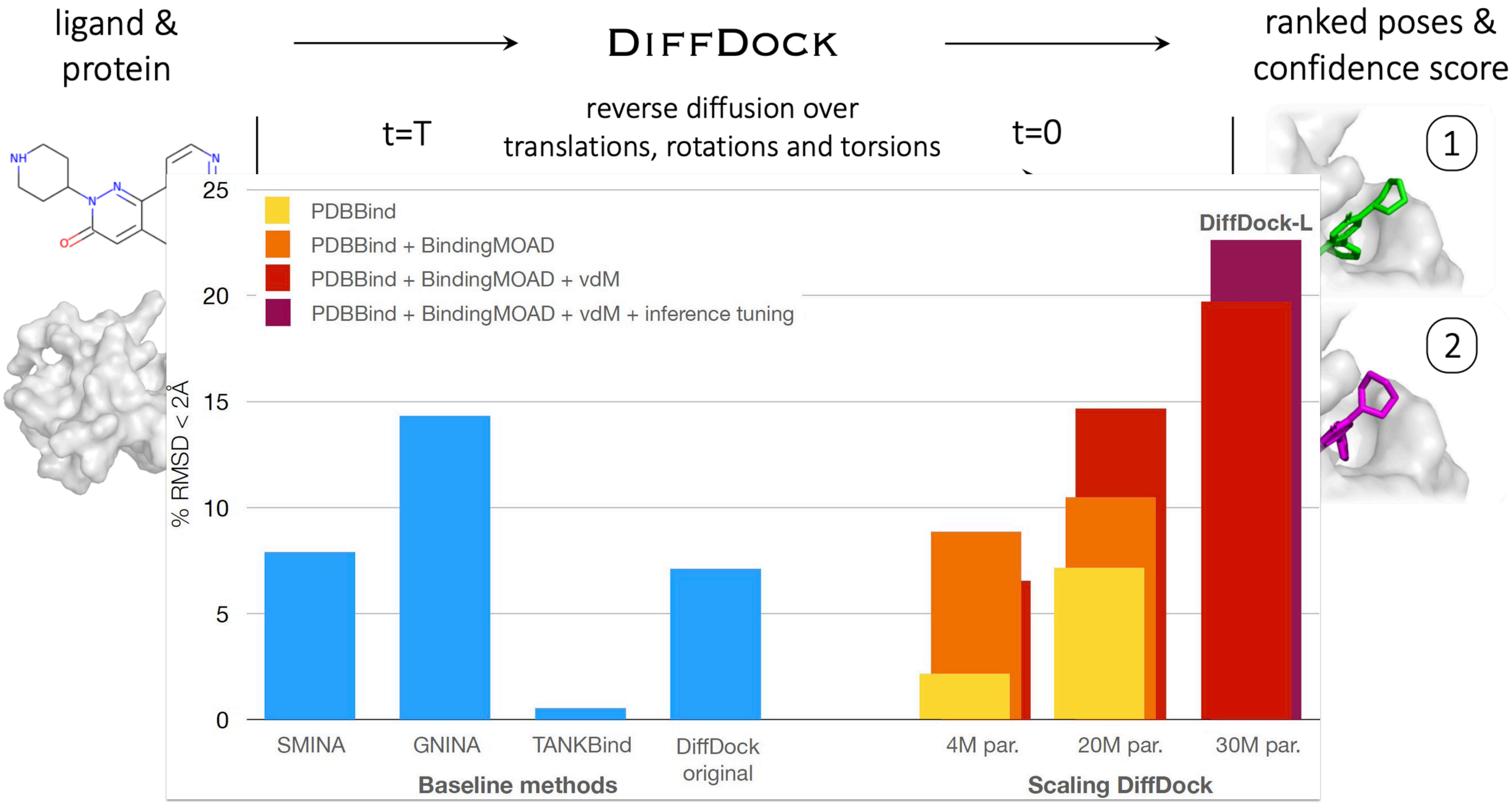


Virtual Screening (DUD-E)

Beyond Scoring

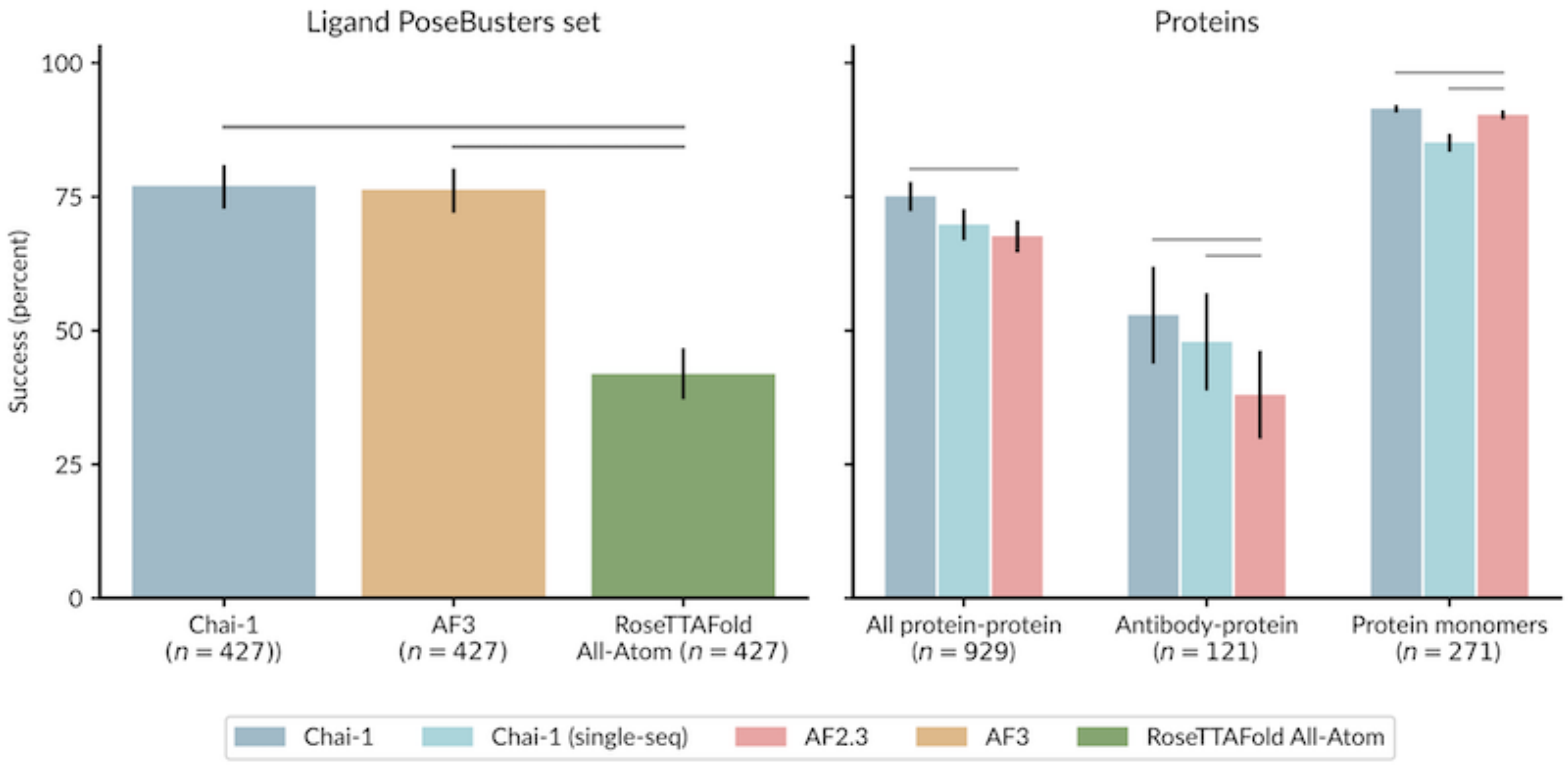


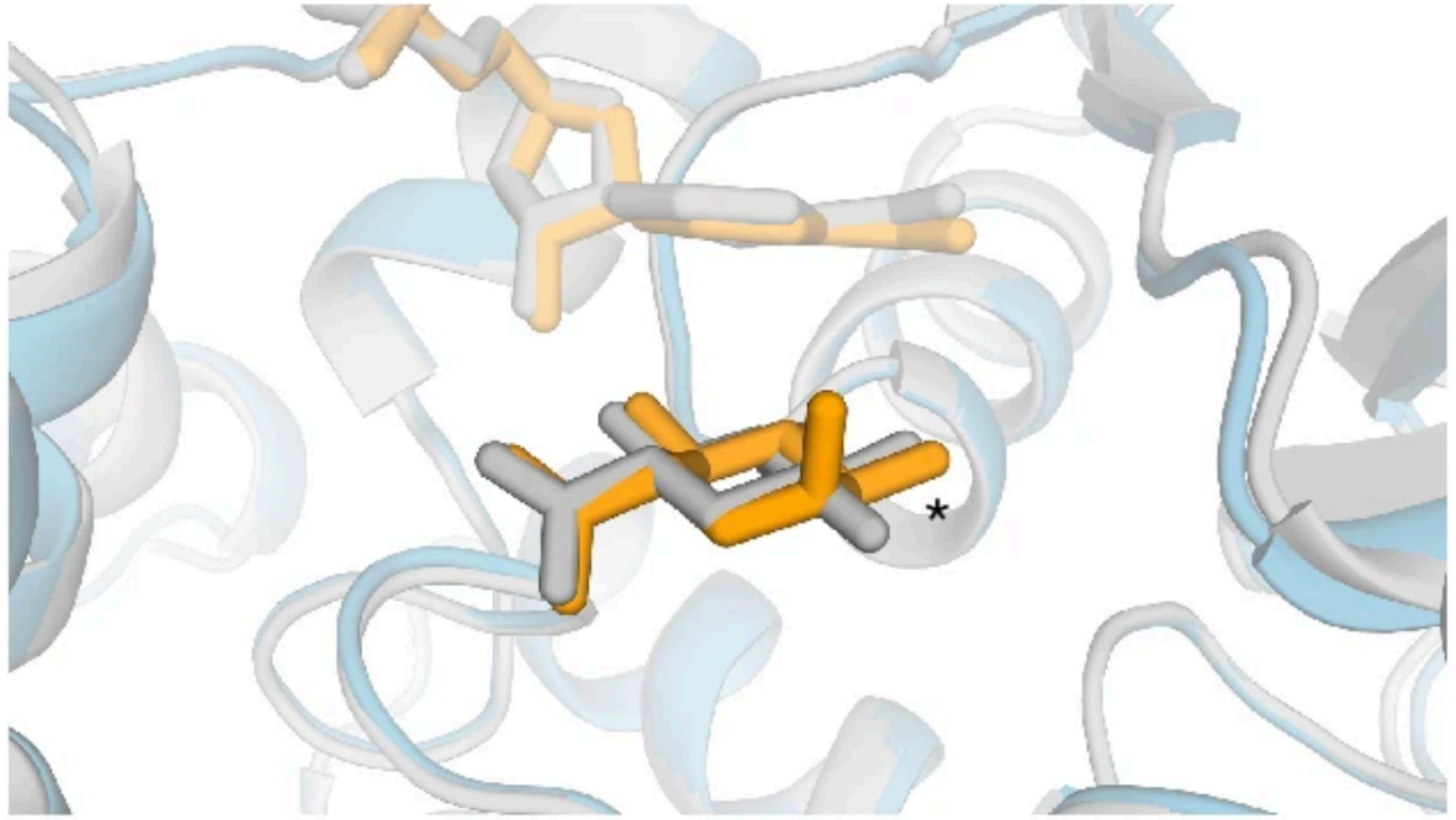
Beyond Scoring



Cofolding

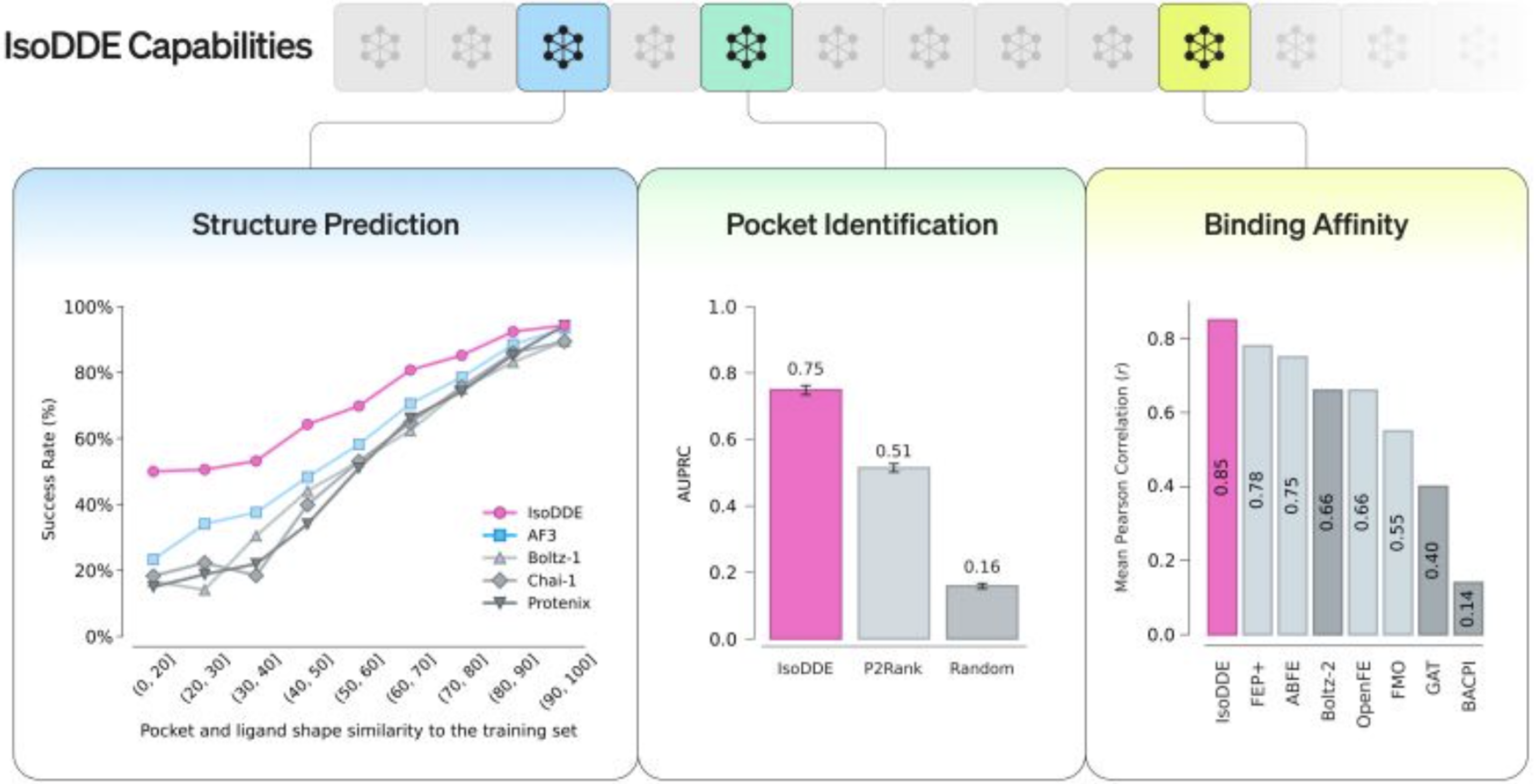
Cofolding models predict the structure of complexes directly from protein sequence and ligand identity.





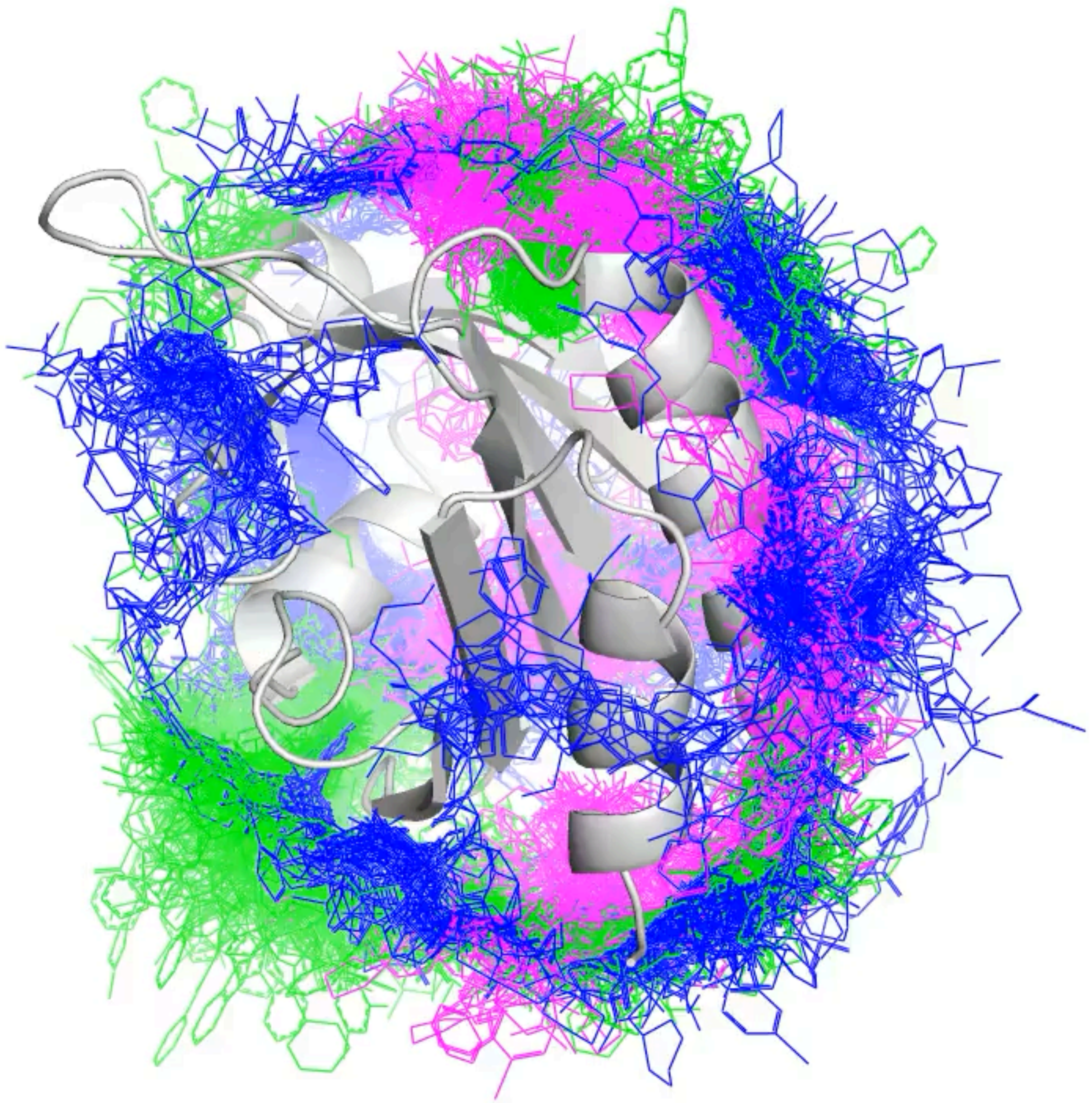
Isomorphic Labs
73,687 followers
1w •

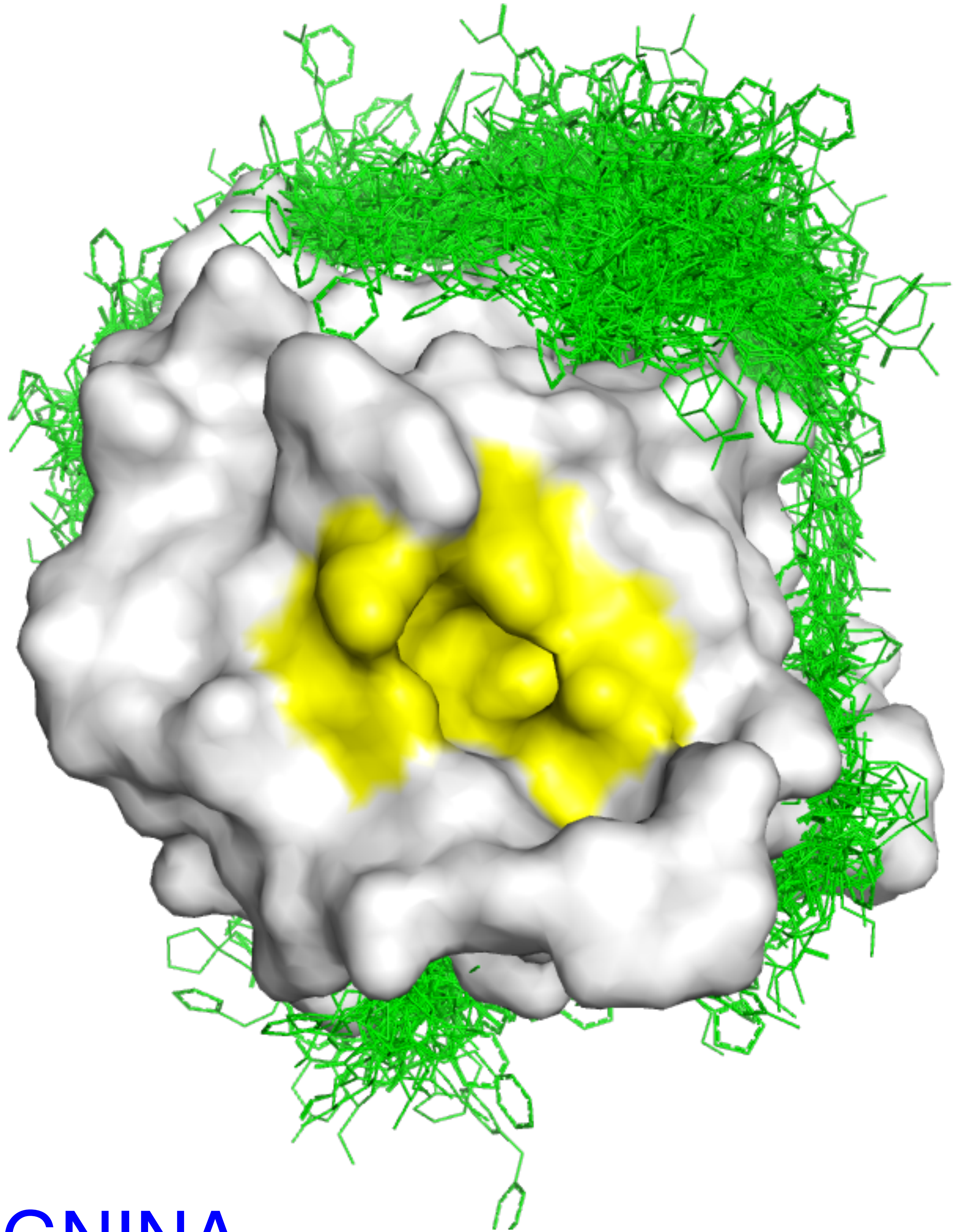
Isomorphic Labs has reached an exciting new milestone in computational drug discovery. In a technical report published today, we show how our drug design engine represents a massive leap forward in accurately predicting ...more



You and 1,890 others   16 comments · 203 reposts

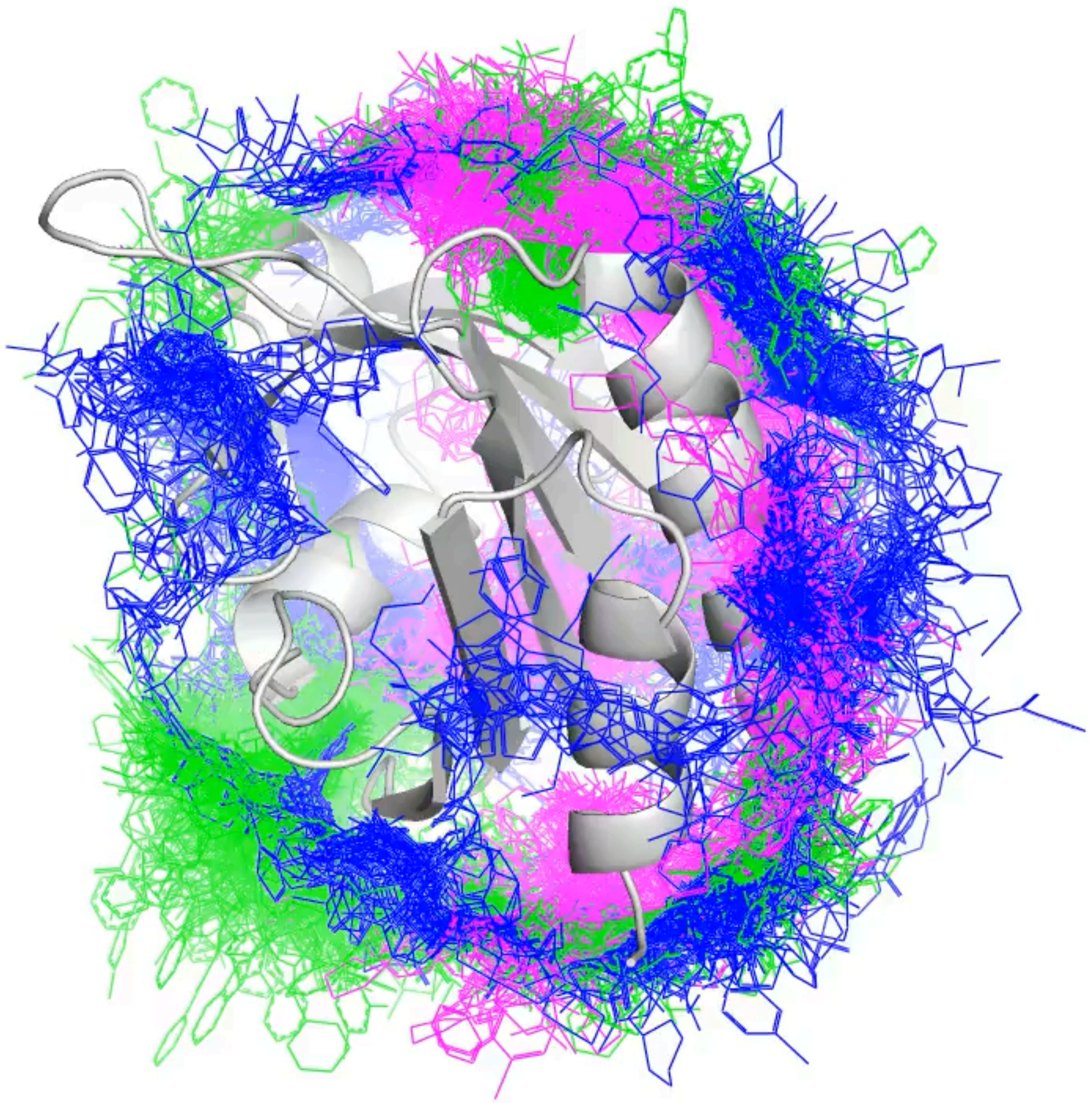
ML Models Have a Data Bias

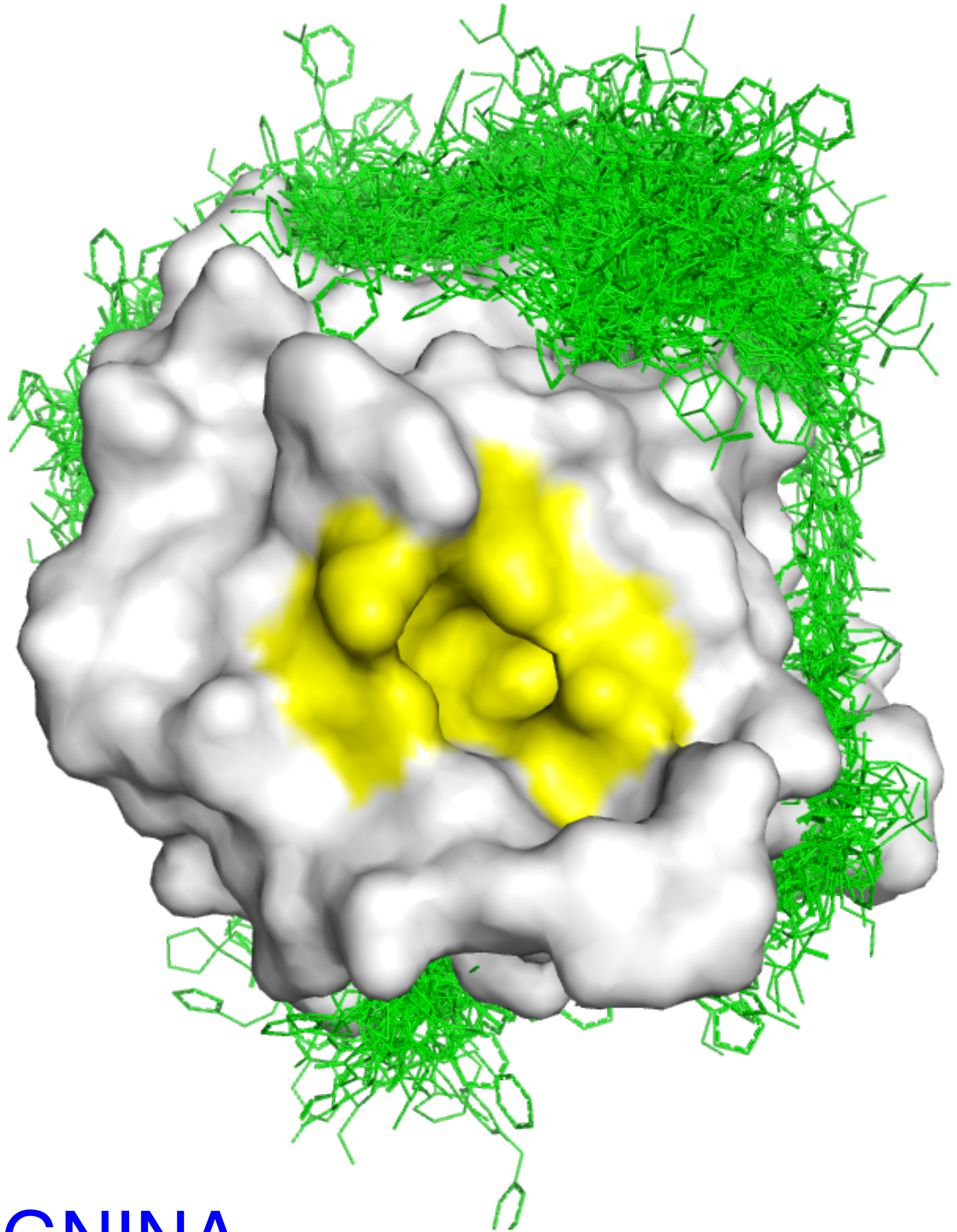




Boltz DynamicBind GNINA

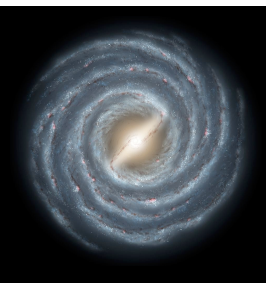
ML Models Have a Data Bias

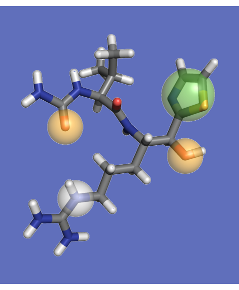
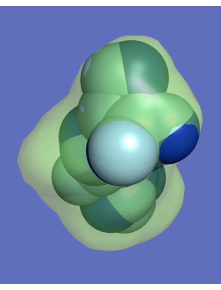


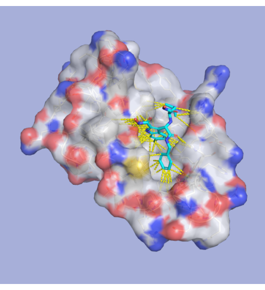
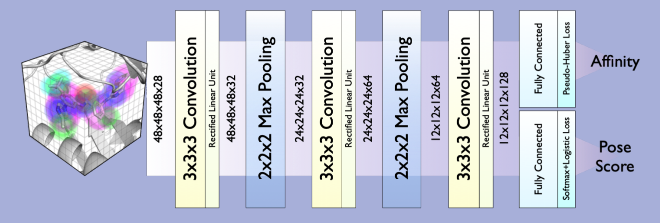


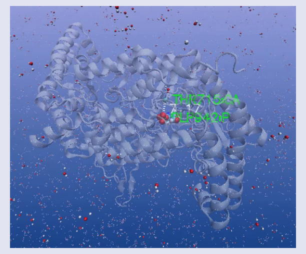
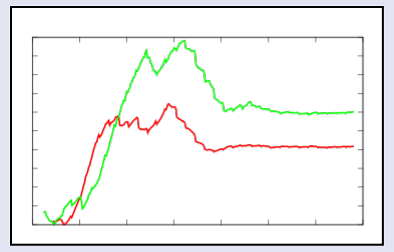
Boltz DynamicBind GNINA

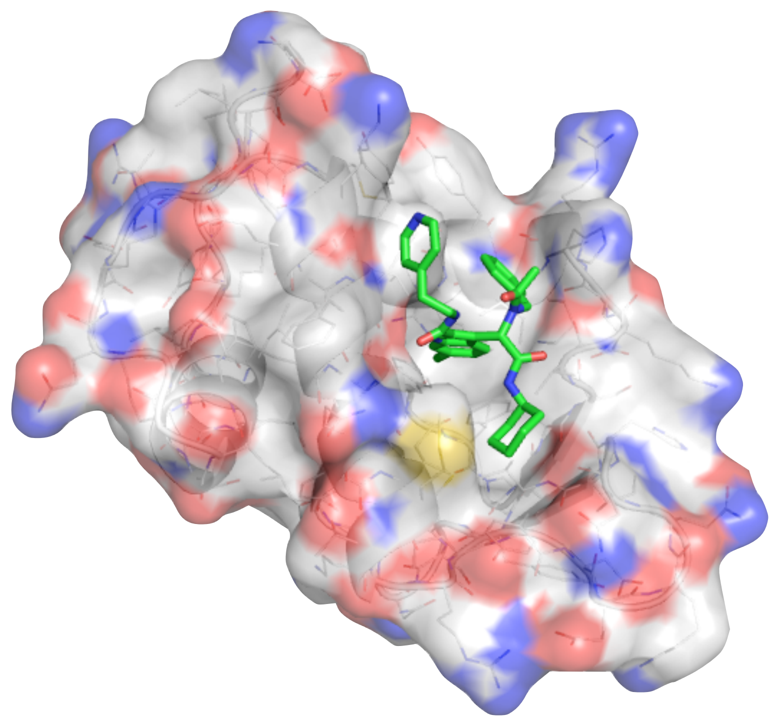
Drug Discovery Funnel

Purchasable  **Accessible**

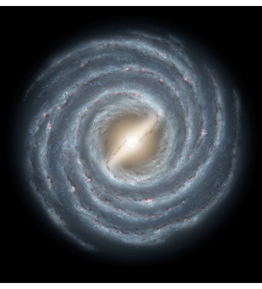
 **Matching** 

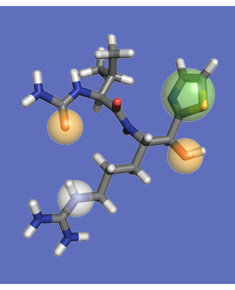
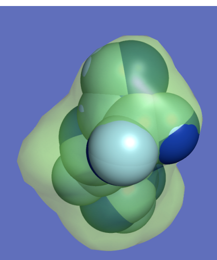
 **Scoring** 

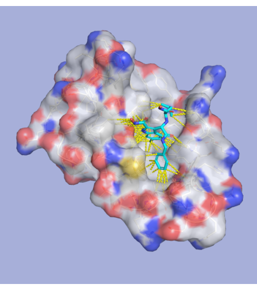
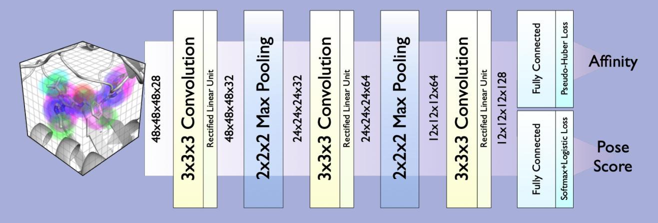
 **Dynamics** 

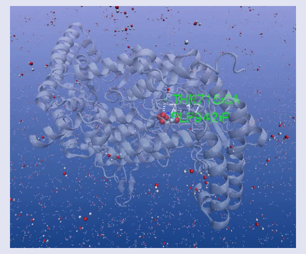
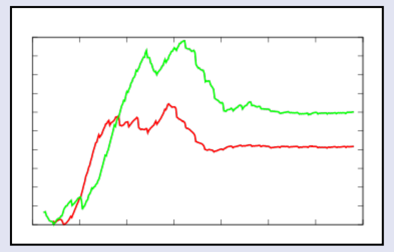


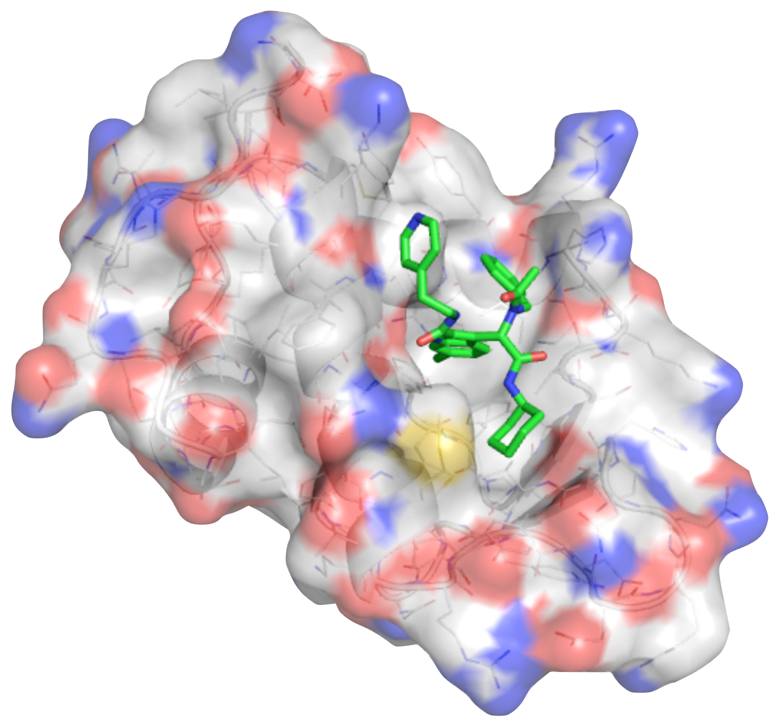
Drug Discovery Funnel

Purchasable  **Accessible**

 **Matching** 

 **Scoring** 

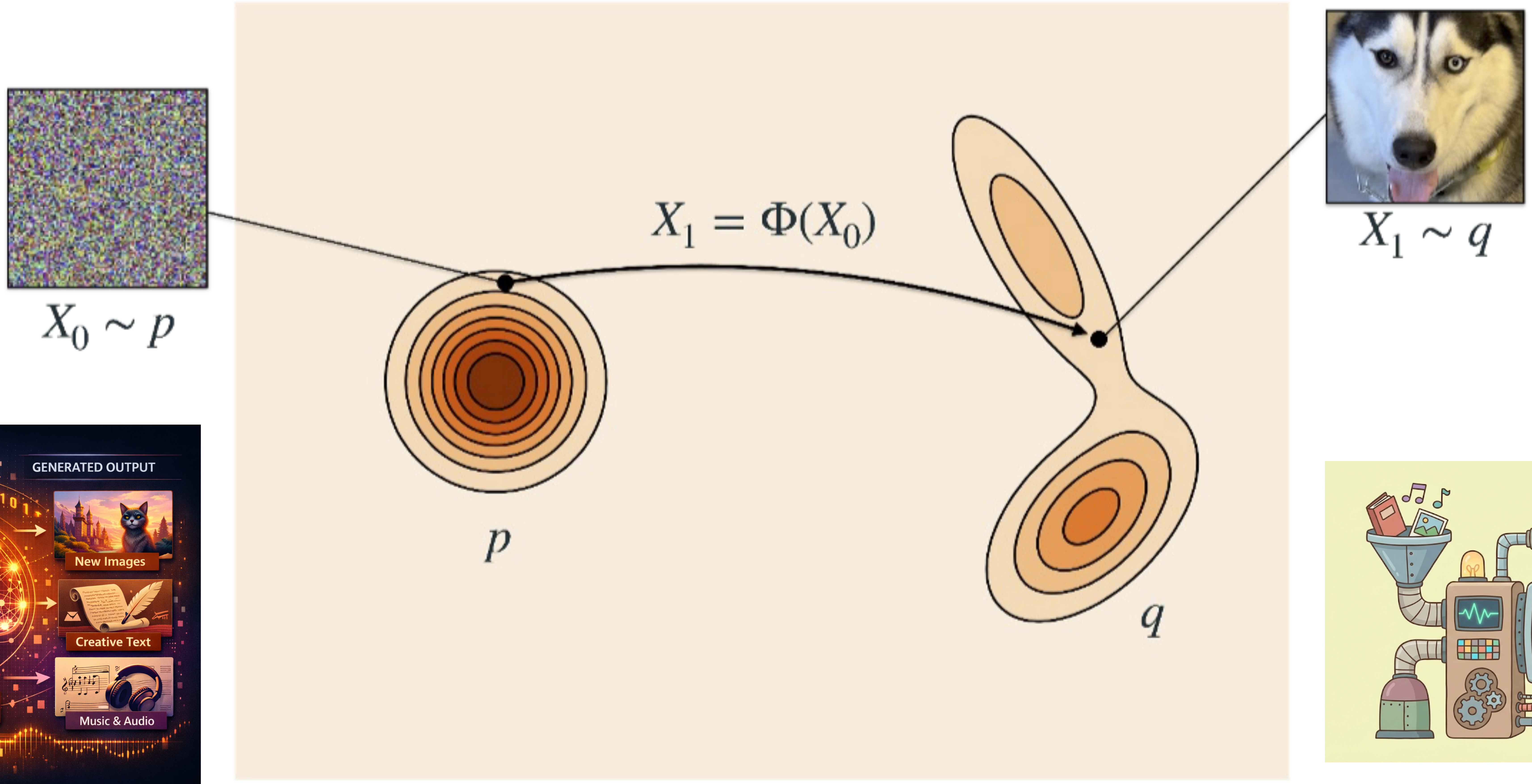
 **Dynamics** 

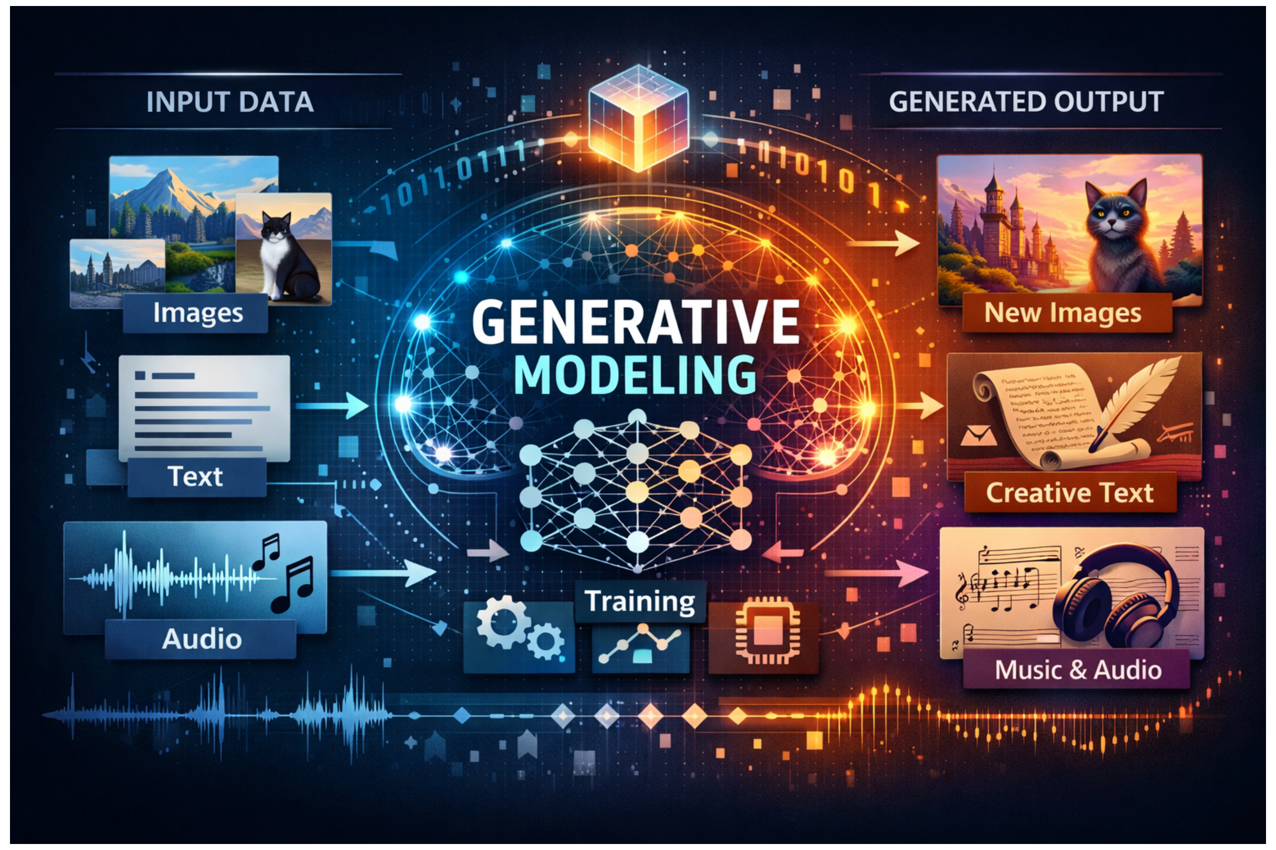


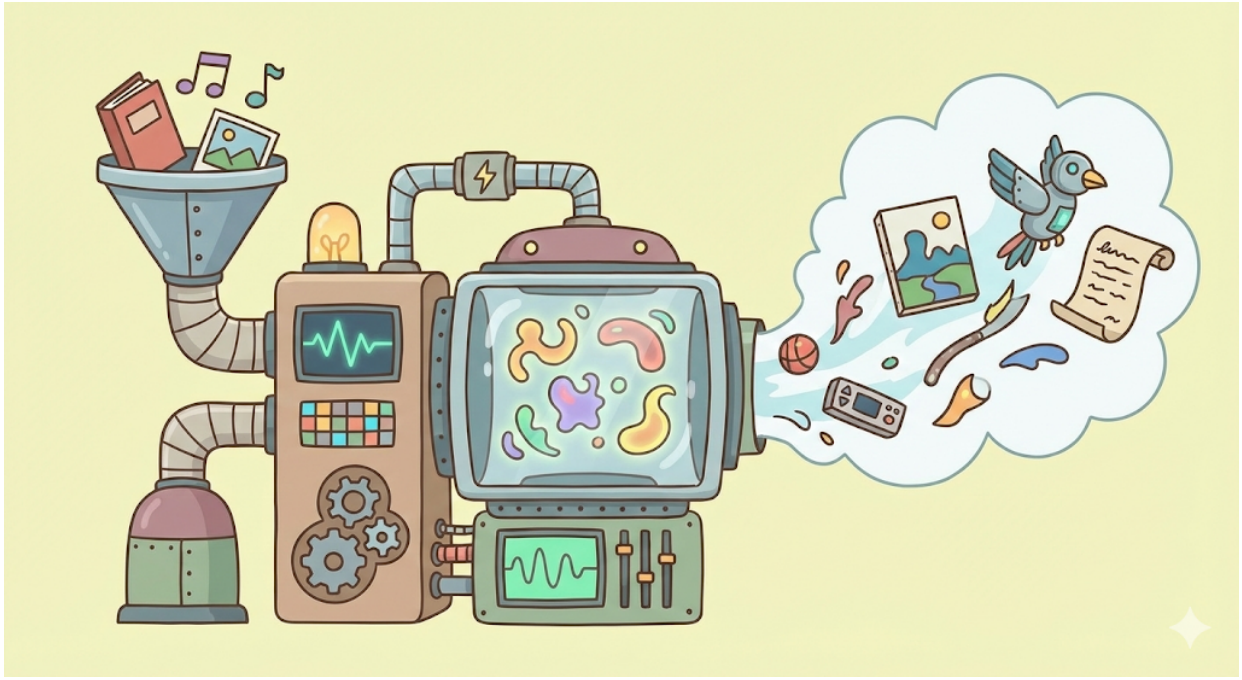
Generative Modeling

Generative Modeling

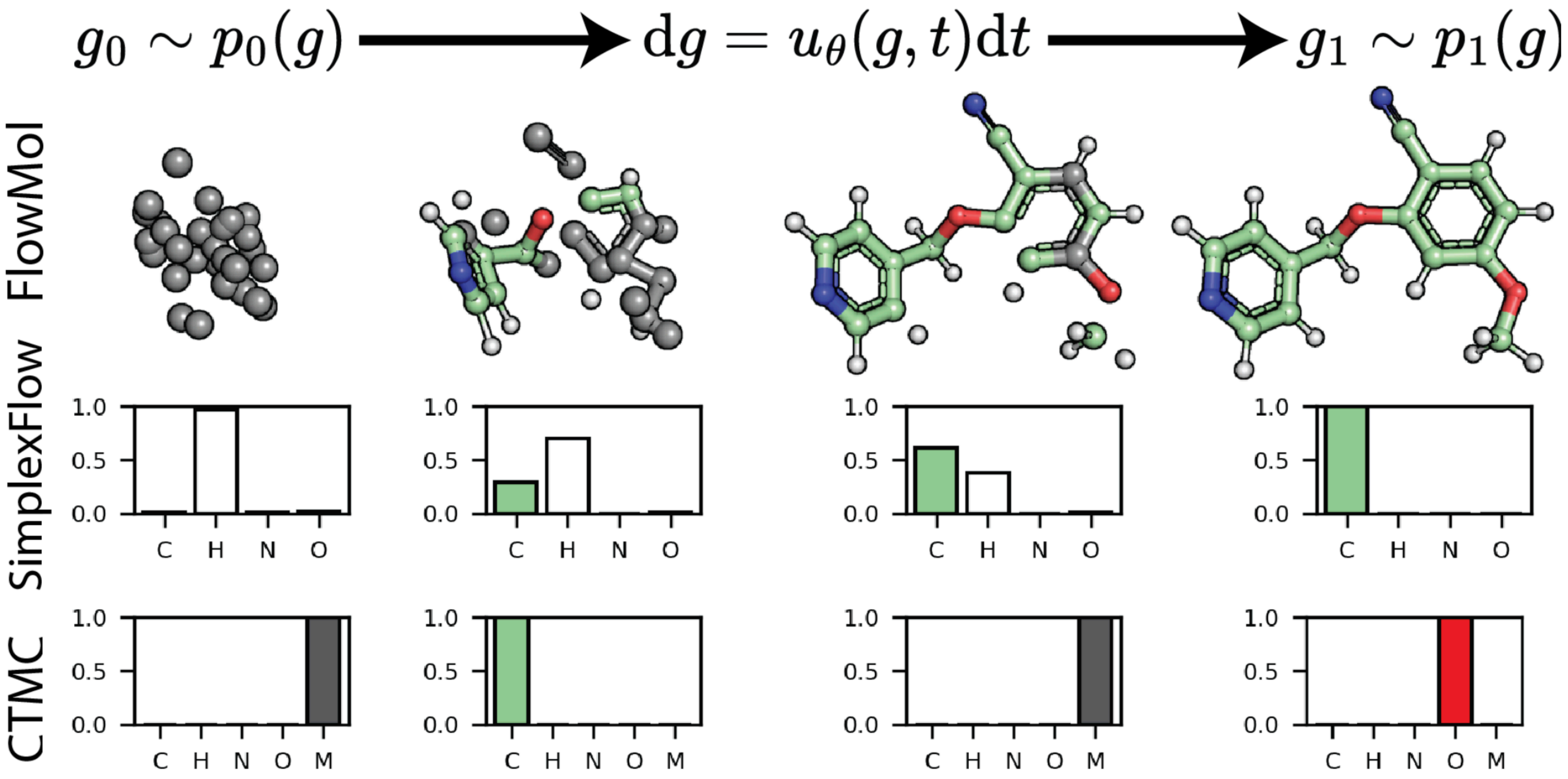
We learn a model Φ that maps between a distribution we know how to sample and one we don't (but have samples from).







Unconditional Generation with FlowMol

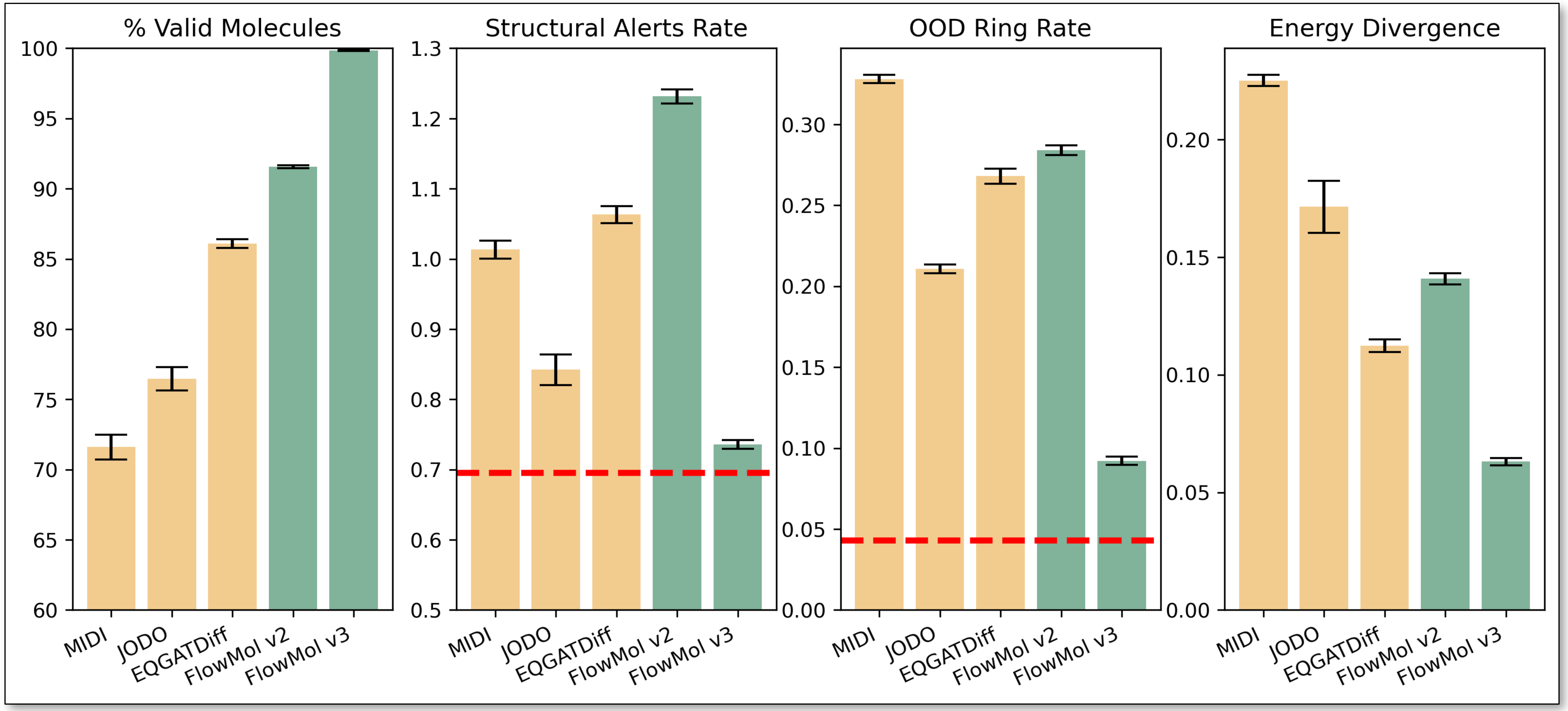


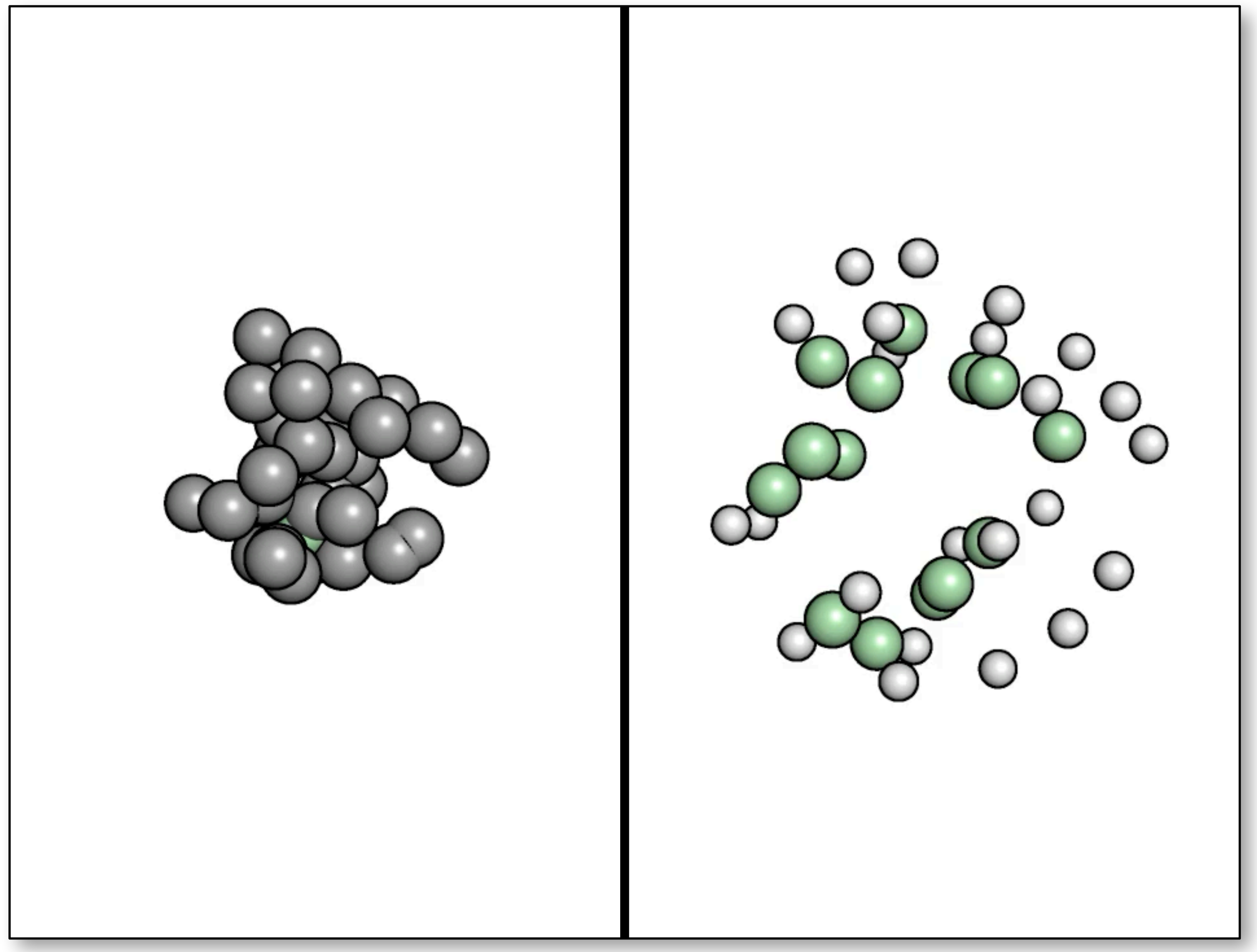
Exploring Discrete Flow Matching for 3D De Novo Molecule Generation

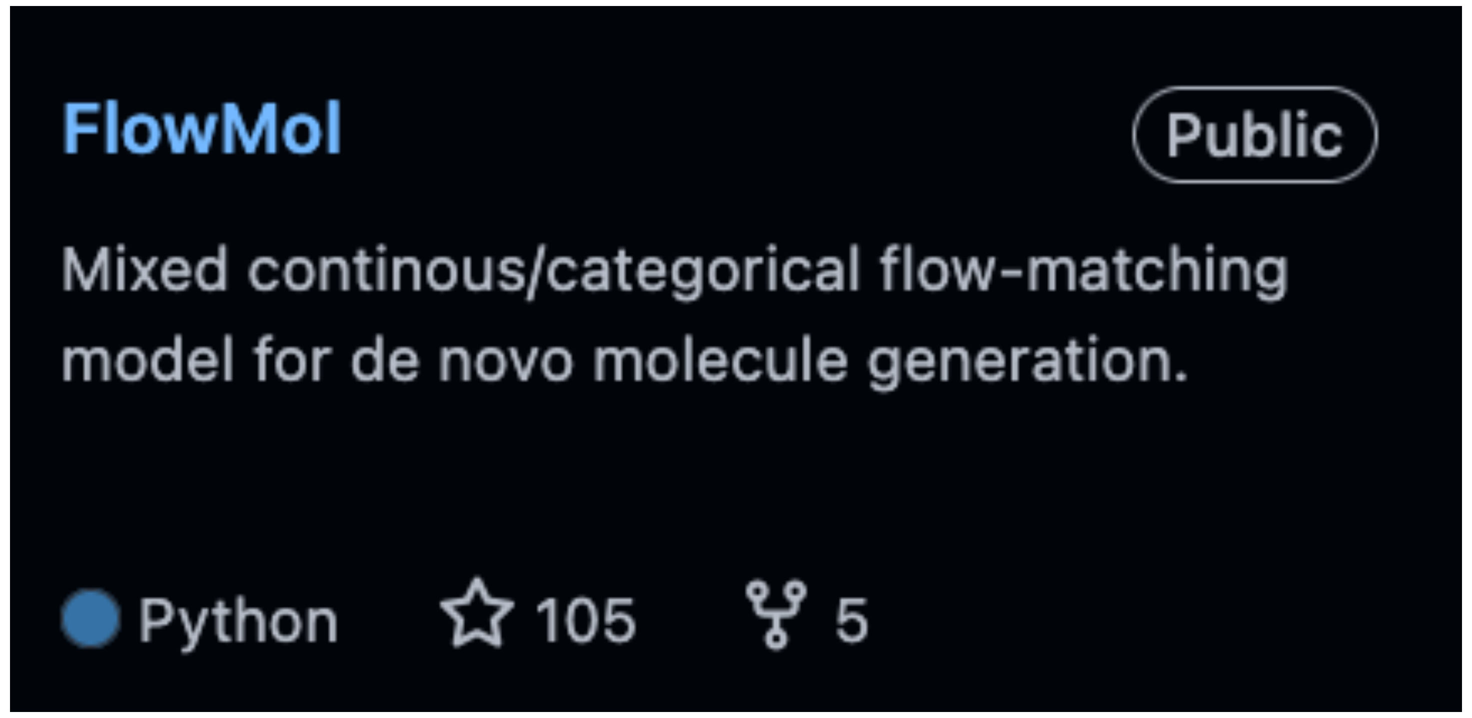
Ian Dunn
Dept. of Computational & Systems Biology
University of Pittsburgh
Pittsburgh, PA 15260
ian.dunn@pitt.edu

David Ryan Koes
Dept. of Computational & Systems Biology
University of Pittsburgh
Pittsburgh, PA 15260
dkoes@pitt.edu

FlowMol v3

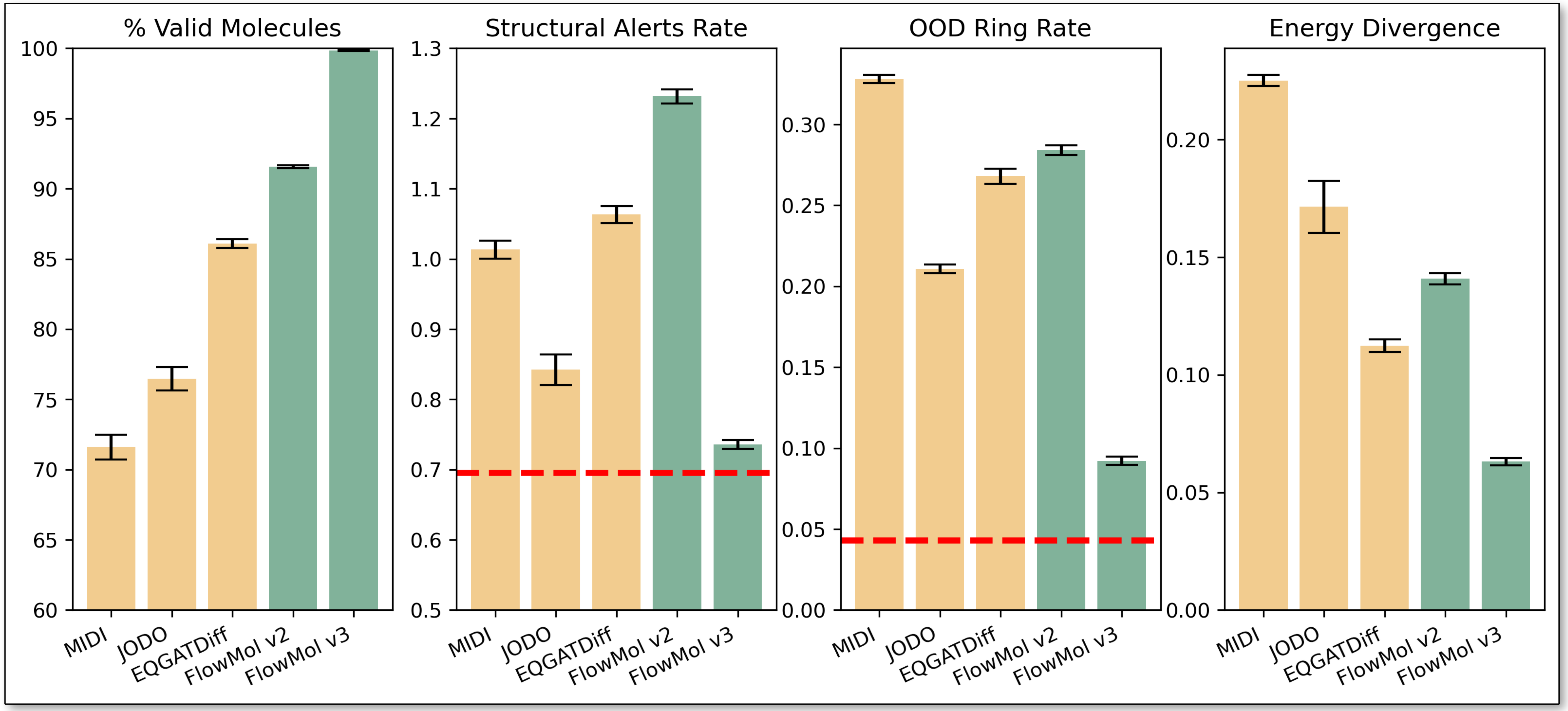


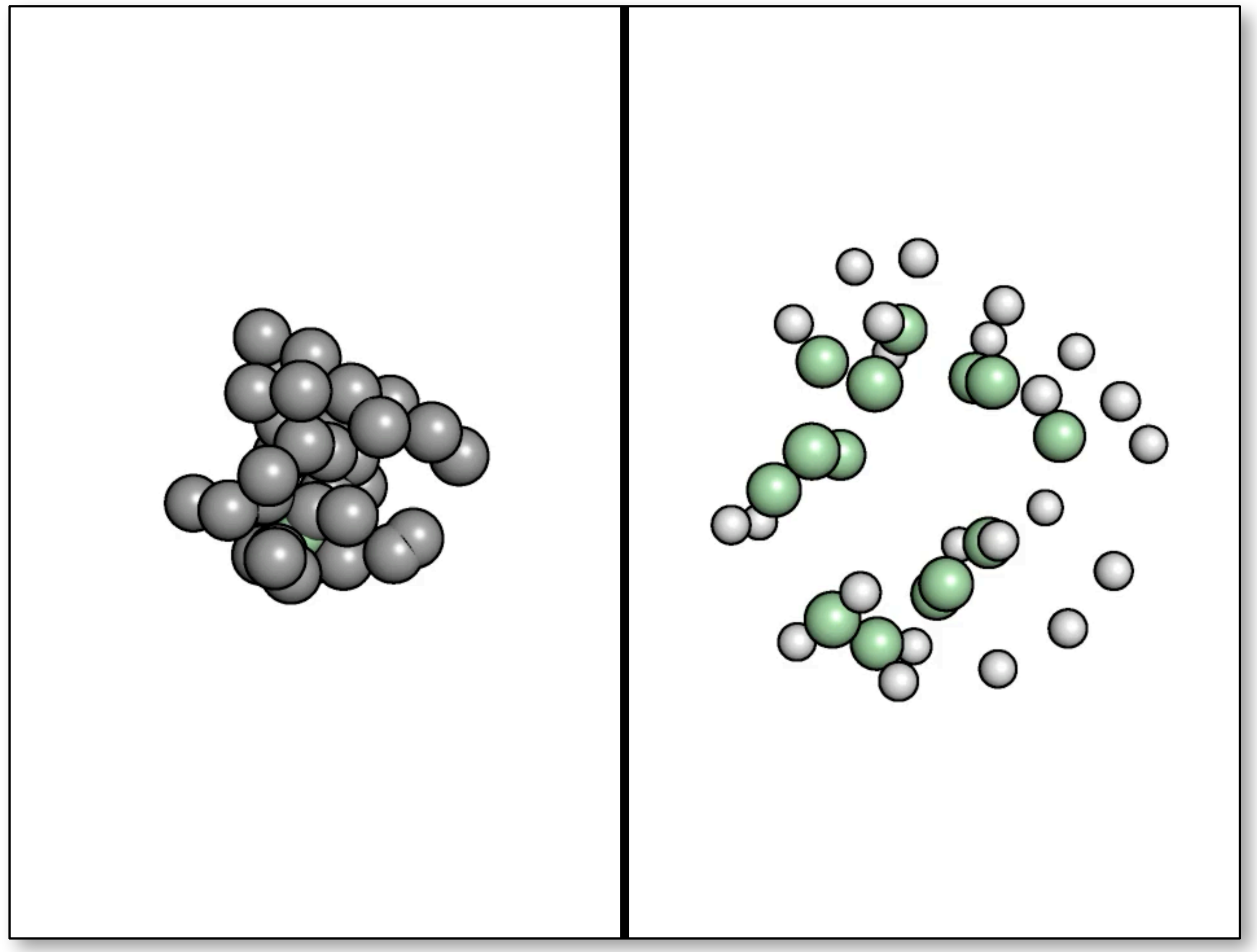


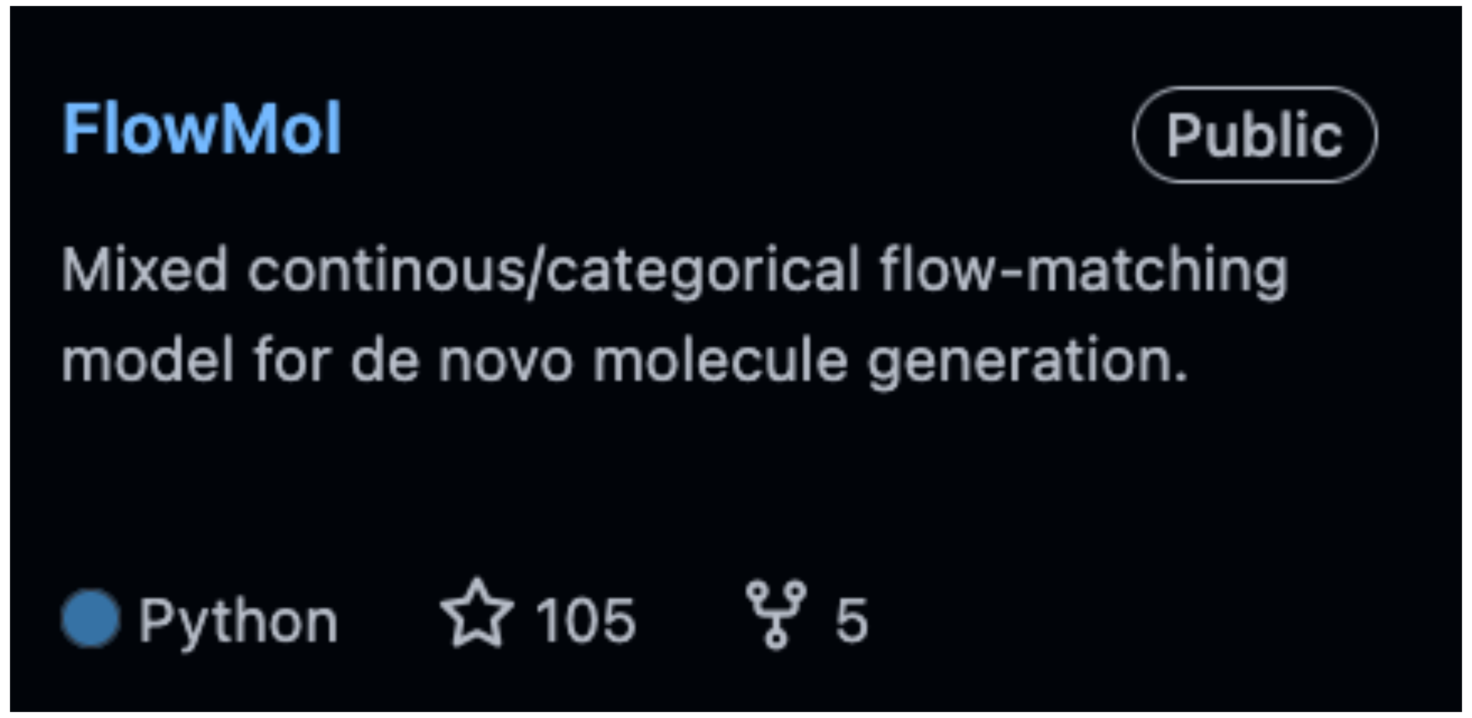


- State of the art validity
- Improves chemical plausibility and synthetic accessibility

FlowMol v3







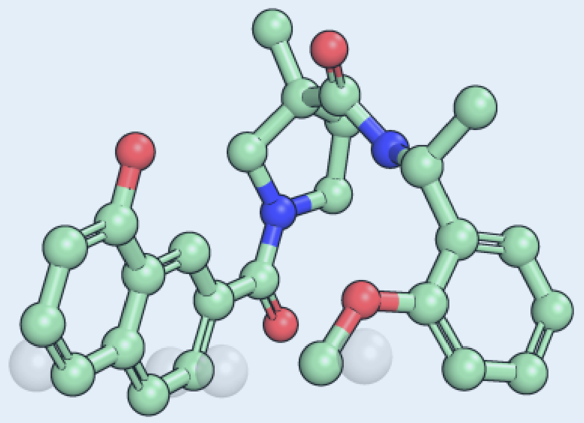
- State of the art validity
- Improves chemical plausibility and synthetic accessibility

OMTRA

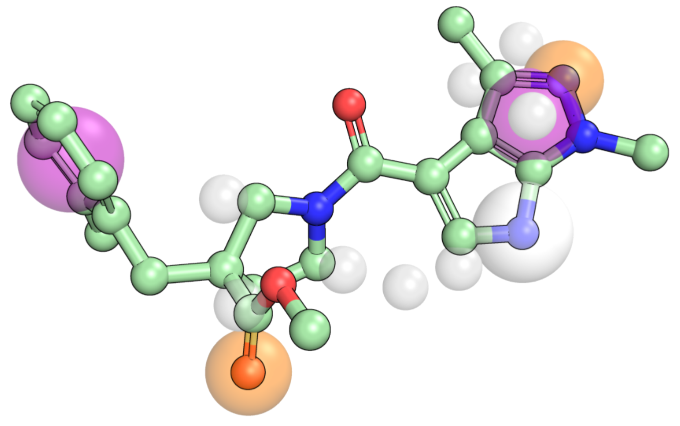
FlowMol3

De novo Generation

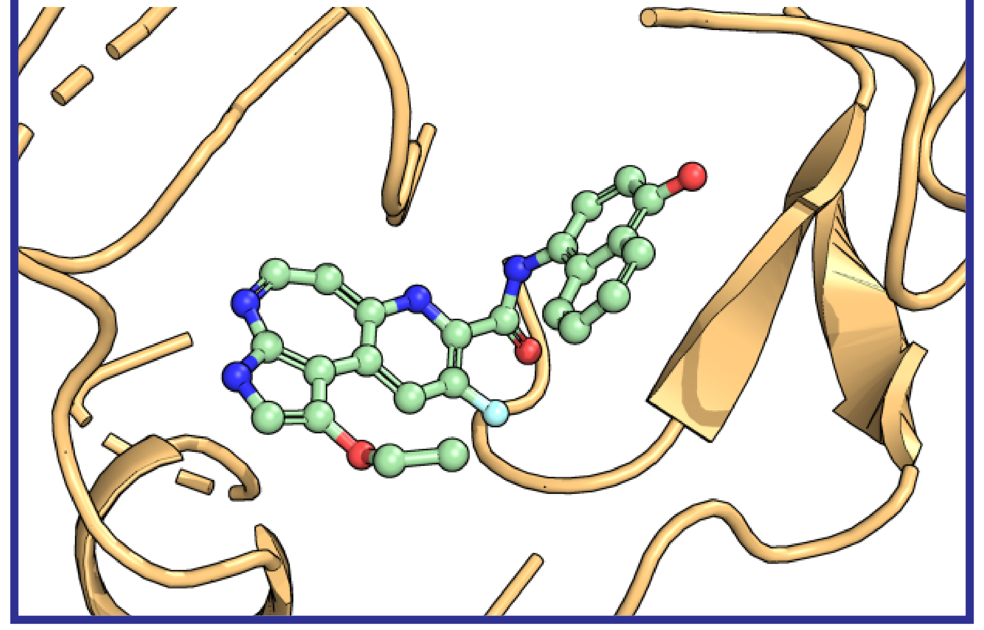
Unconditional



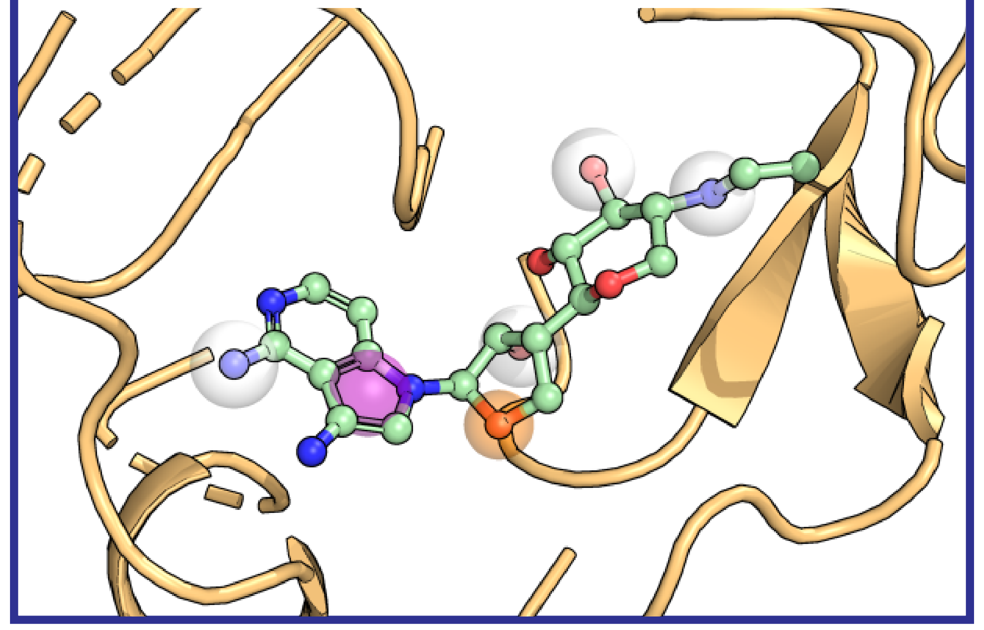
Pharmacophore Conditioned



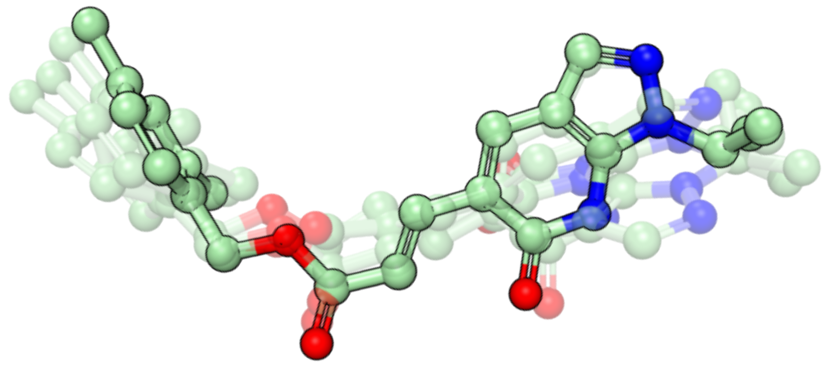
Pocket Conditioned



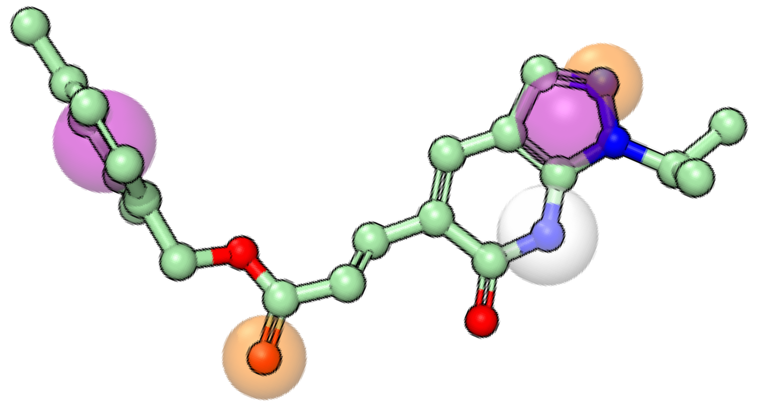
Pocket & Pharm. Conditioned



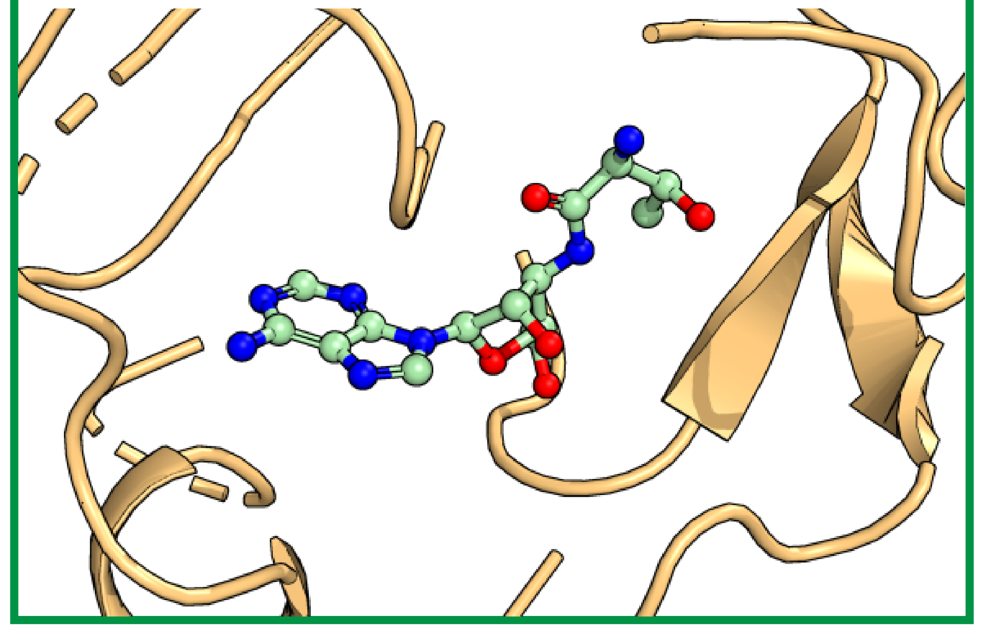
Unconditional



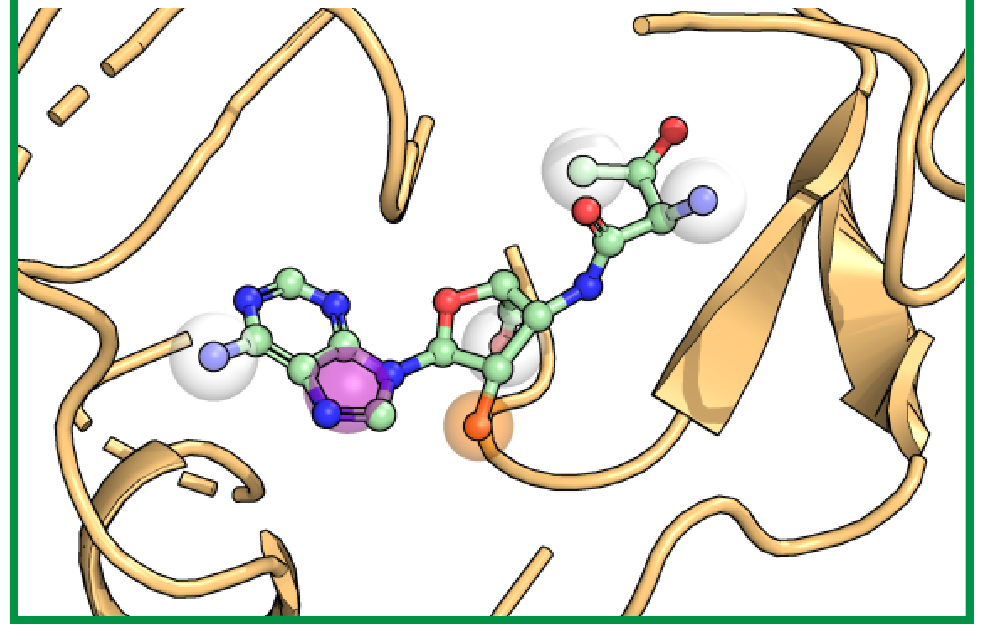
Pharmacophore Conditioned



Pocket Conditioned



Pocket & Pharm. Conditioned

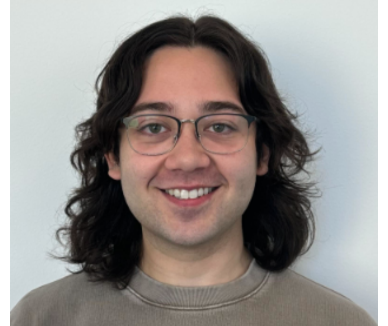


Conformer Generation

Docking

Ian Dunn

Tyler Katz



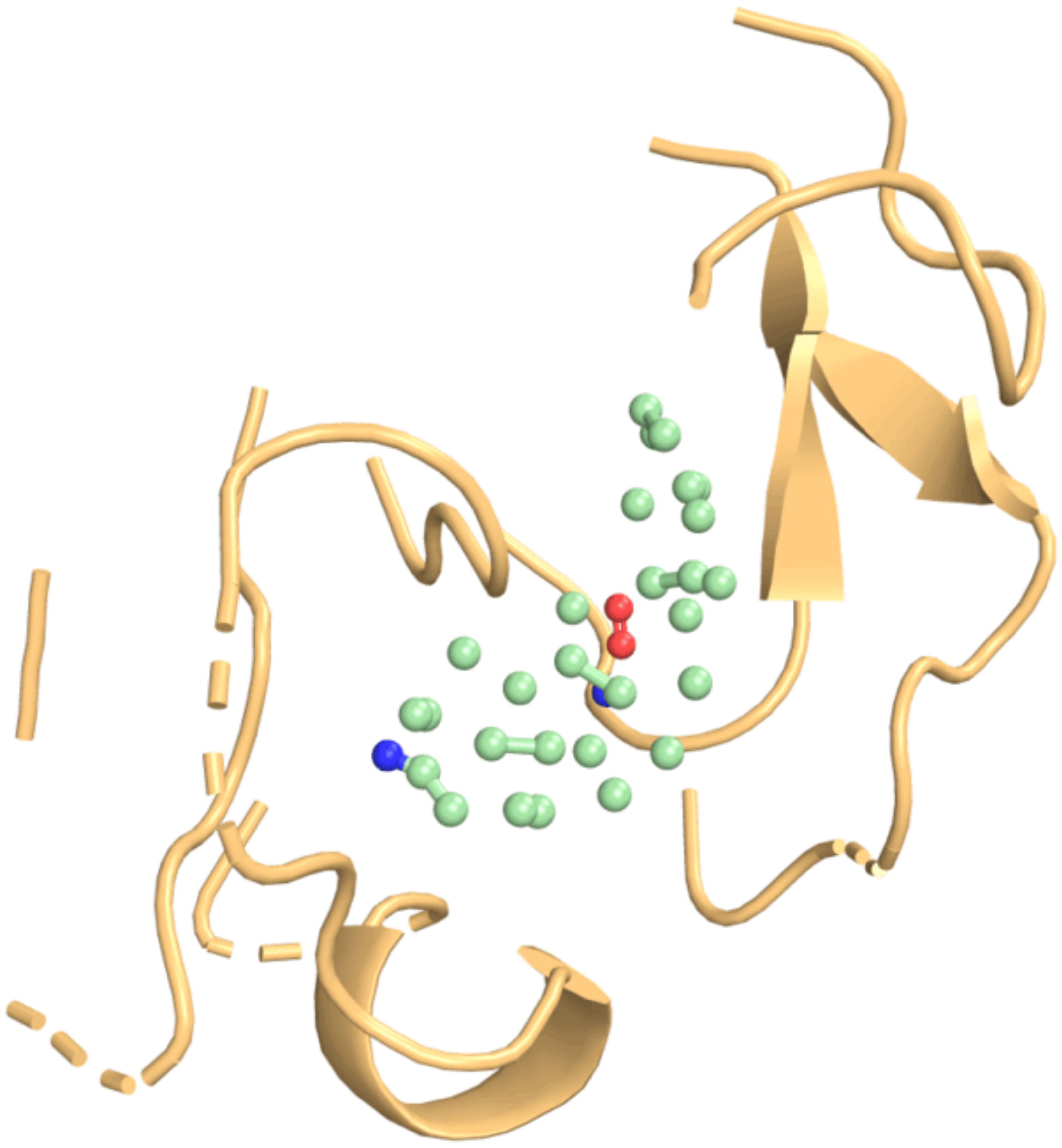
Liv Toft

Riya Shah

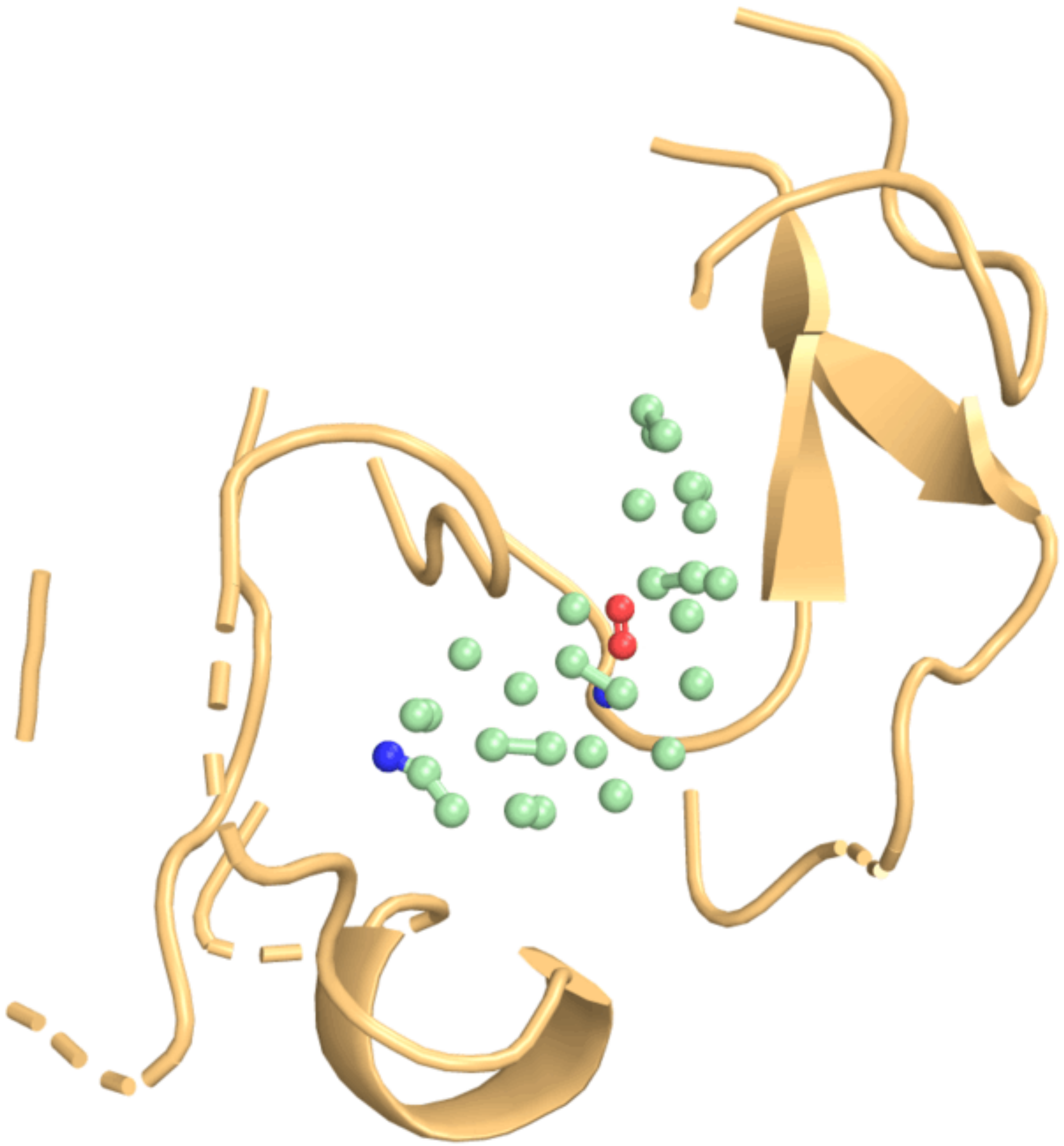
Juhi Gupta

Ramith Hettiarachchi

OMTRA: De Novo Design

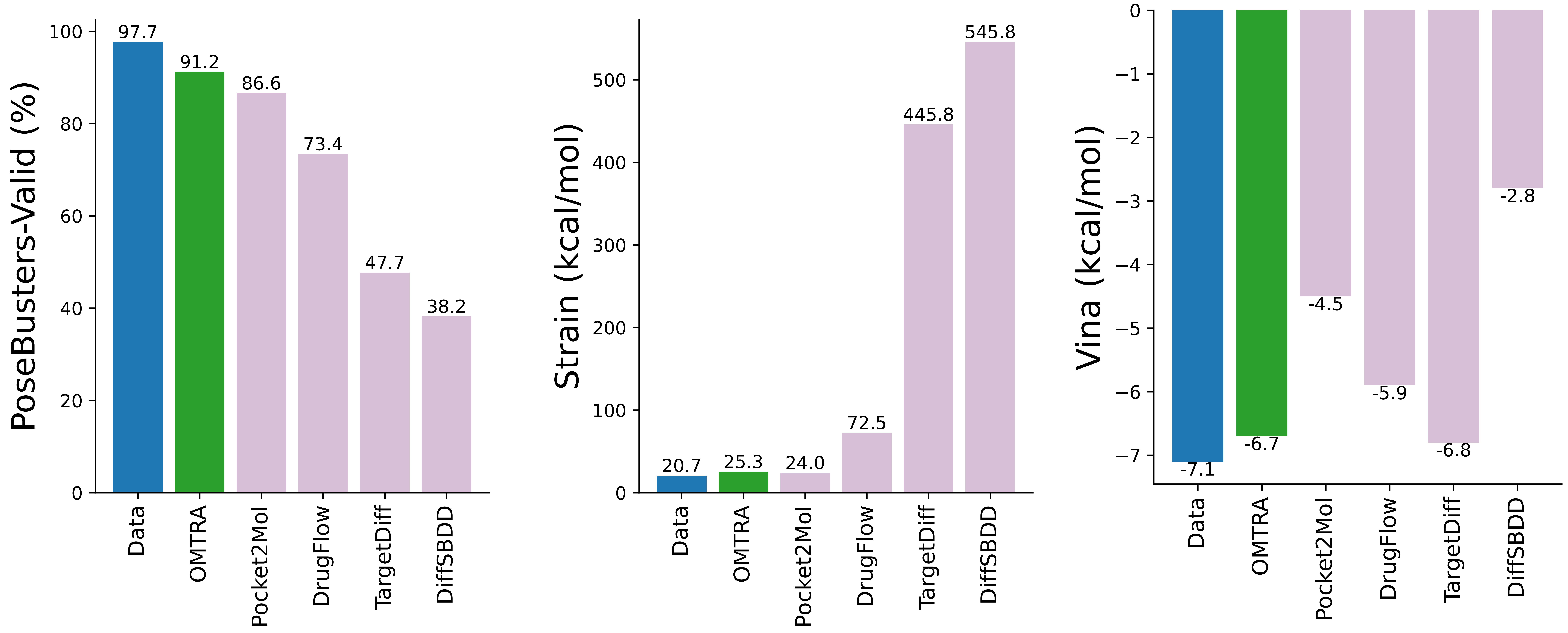


OMTRA: De Novo Design

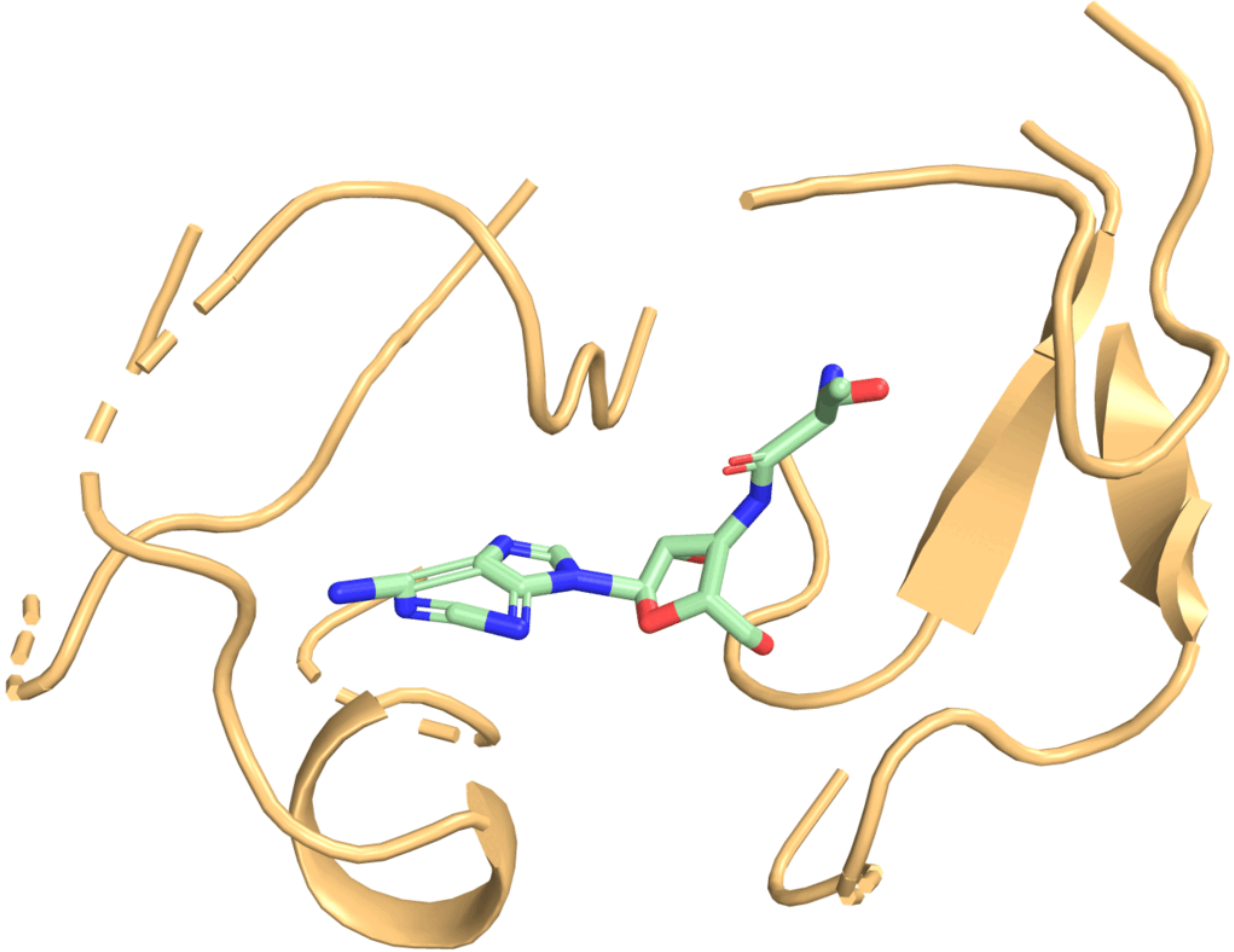


OMTRA: De Novo Design

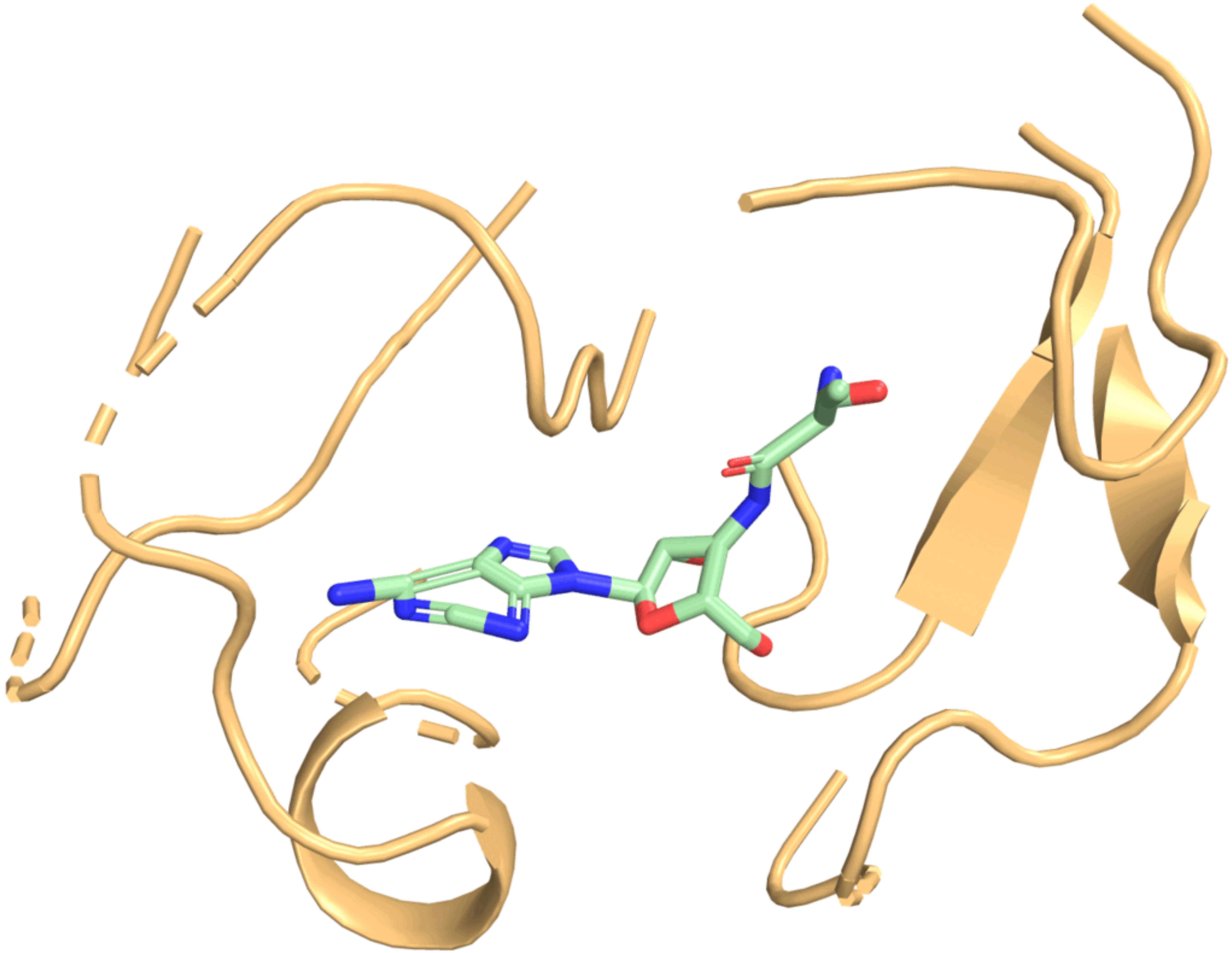
Evaluated on Luo et al CrossDocked test set.



OMTRA: Docking

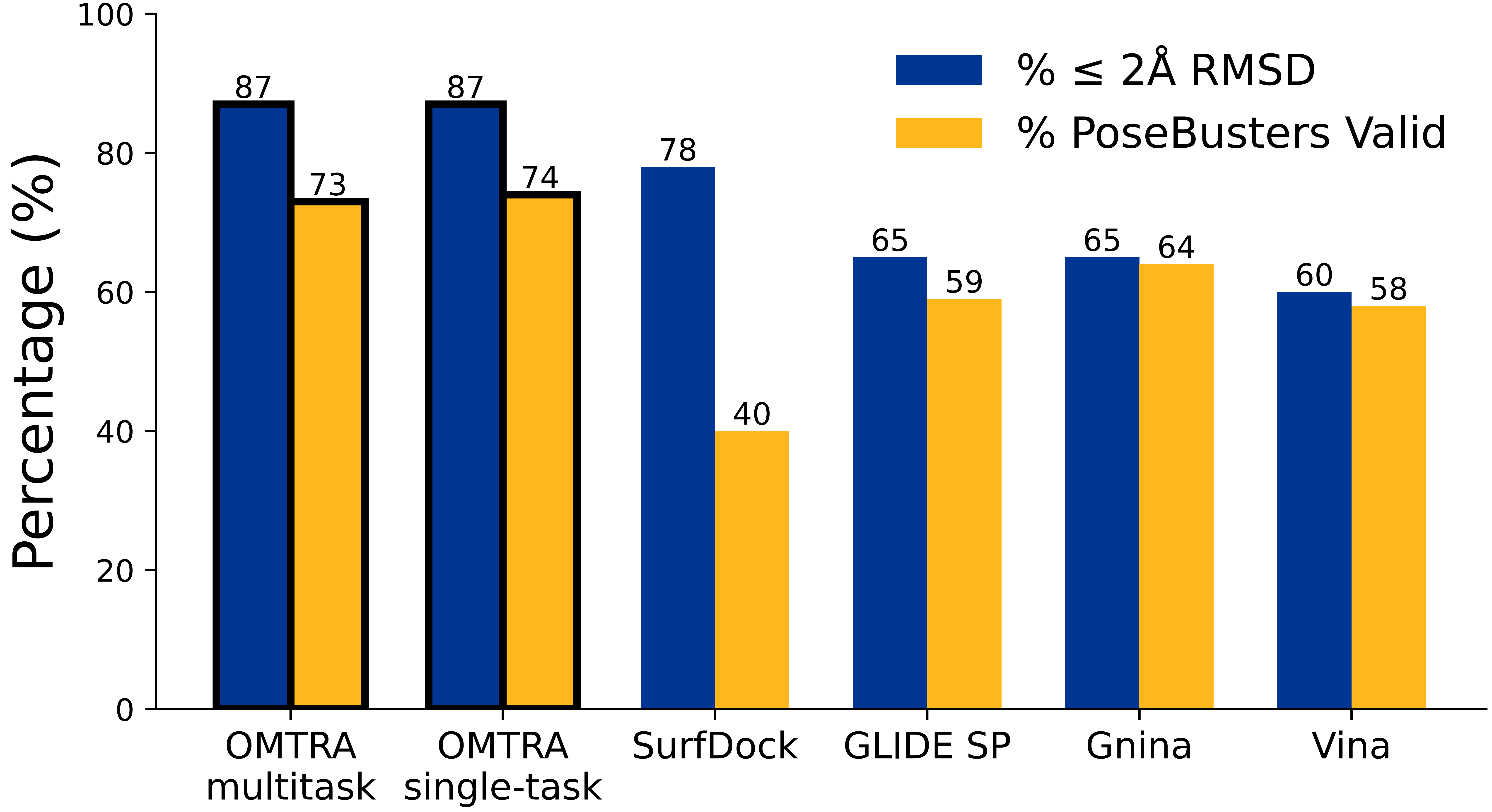


OMTRA: Docking

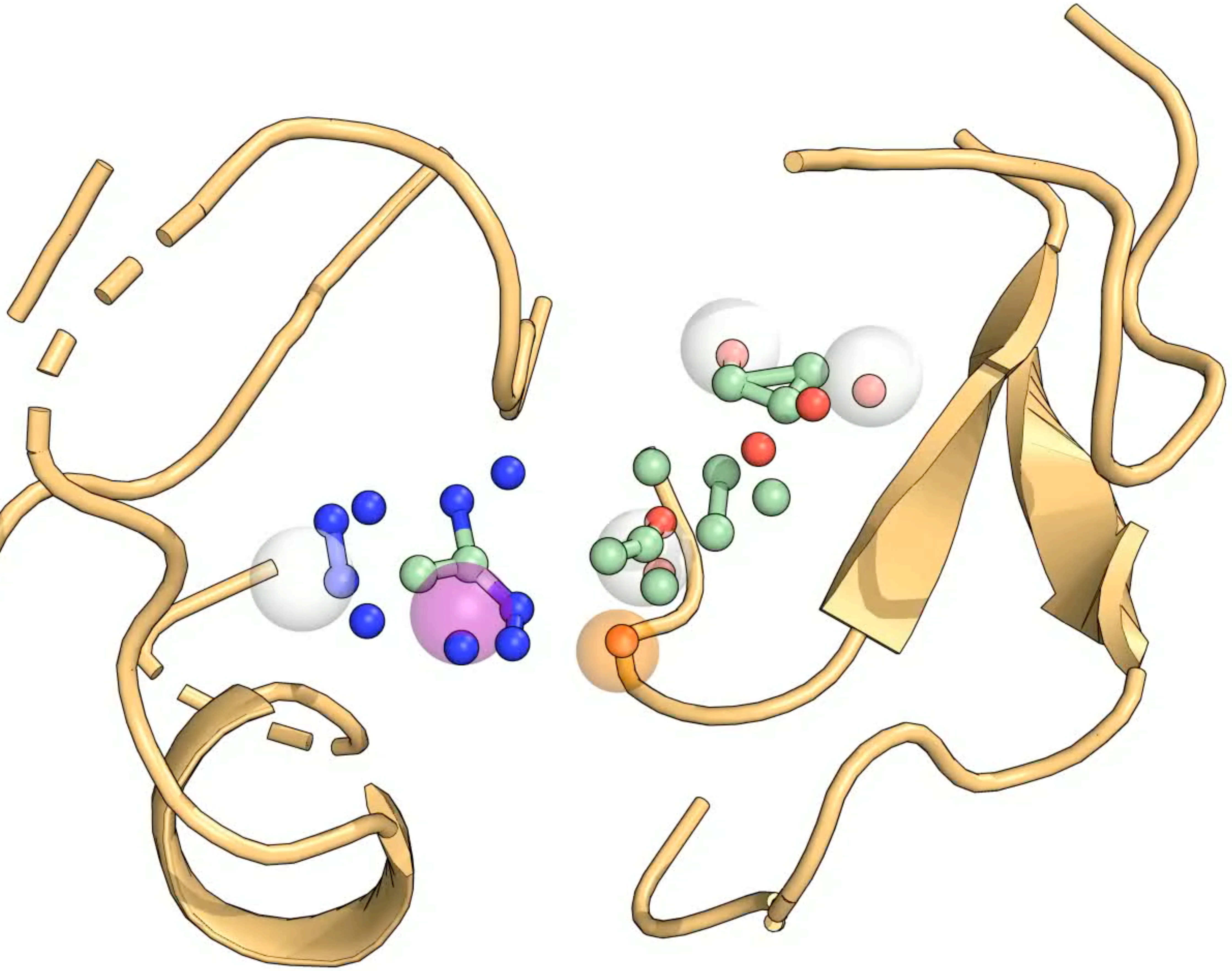


OMTRA: Docking

Evaluated on PoseBusters test set.

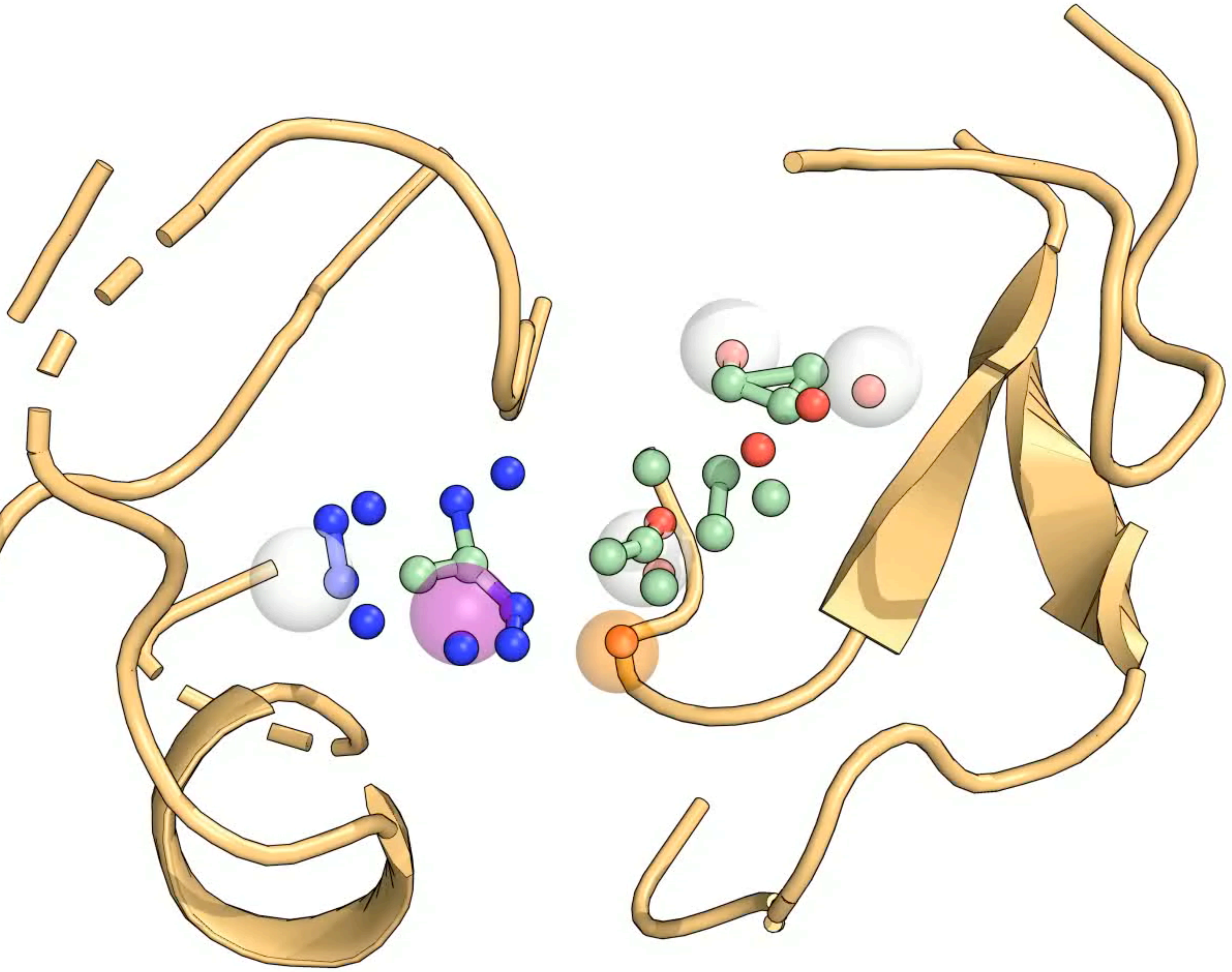


OMTRA: Pharmacophore Conditioning



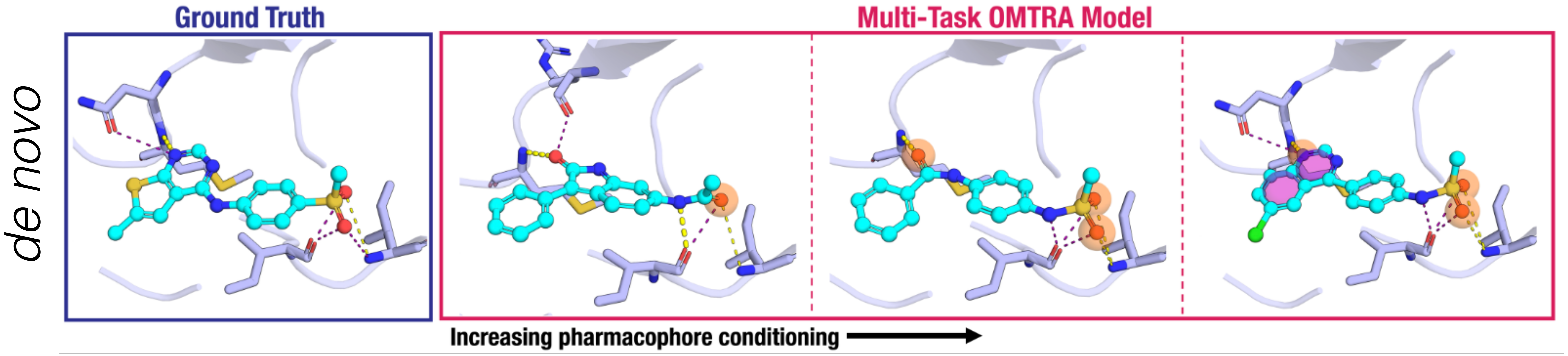
		Prot Conditioning	Prot + Pharm Conditioning
de novo design	%PB-Valid	67.5	66.2
	interaction recovery	51.0	67.4
	% Pharm Matches	-	96.9
docking	% RMSD $\leq$ 2Å	93.0	99.0
	%PB-Valid	73.0	81.0
	% Pharm Matches	-	99.5

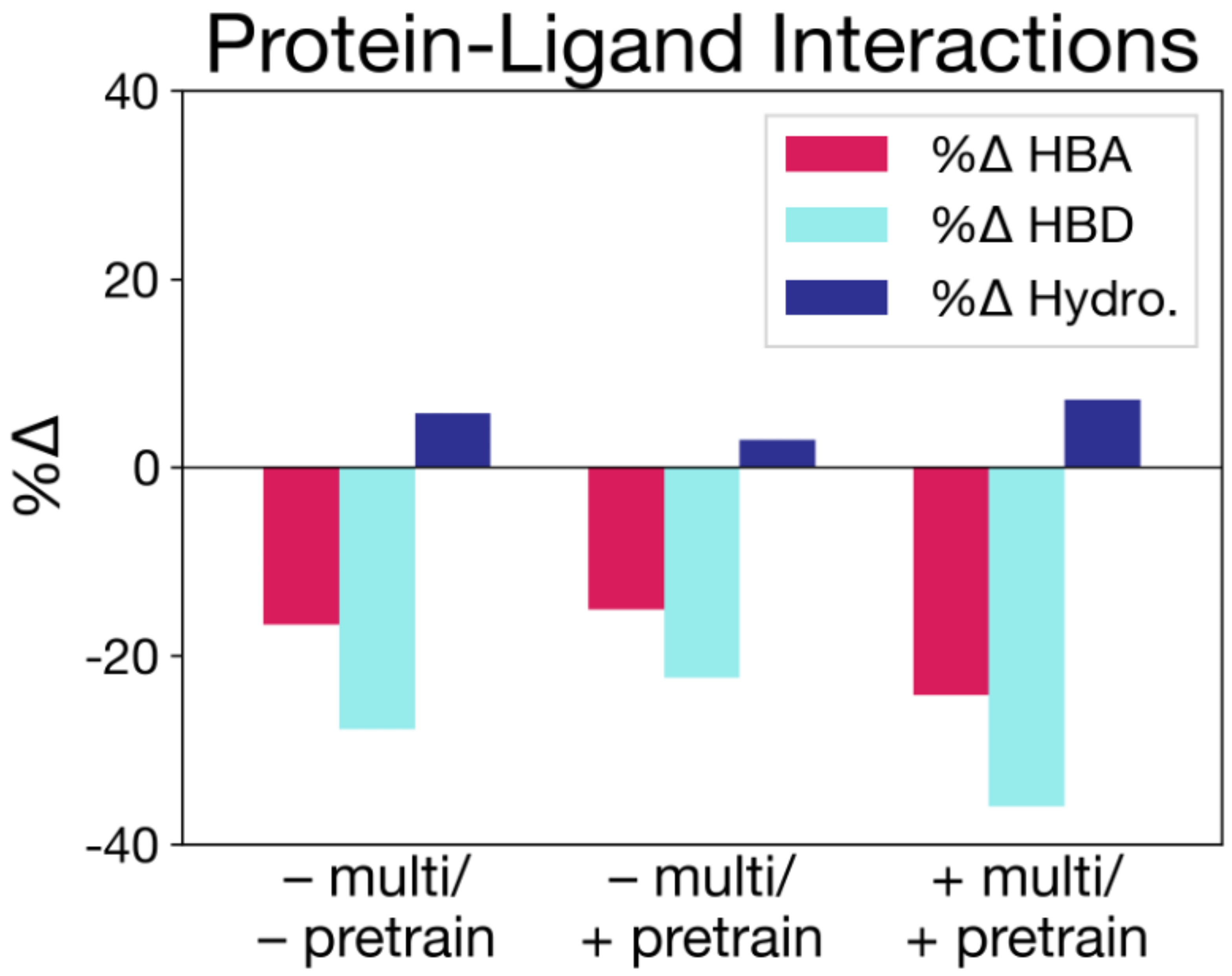
OMTRA: Pharmacophore Conditioning

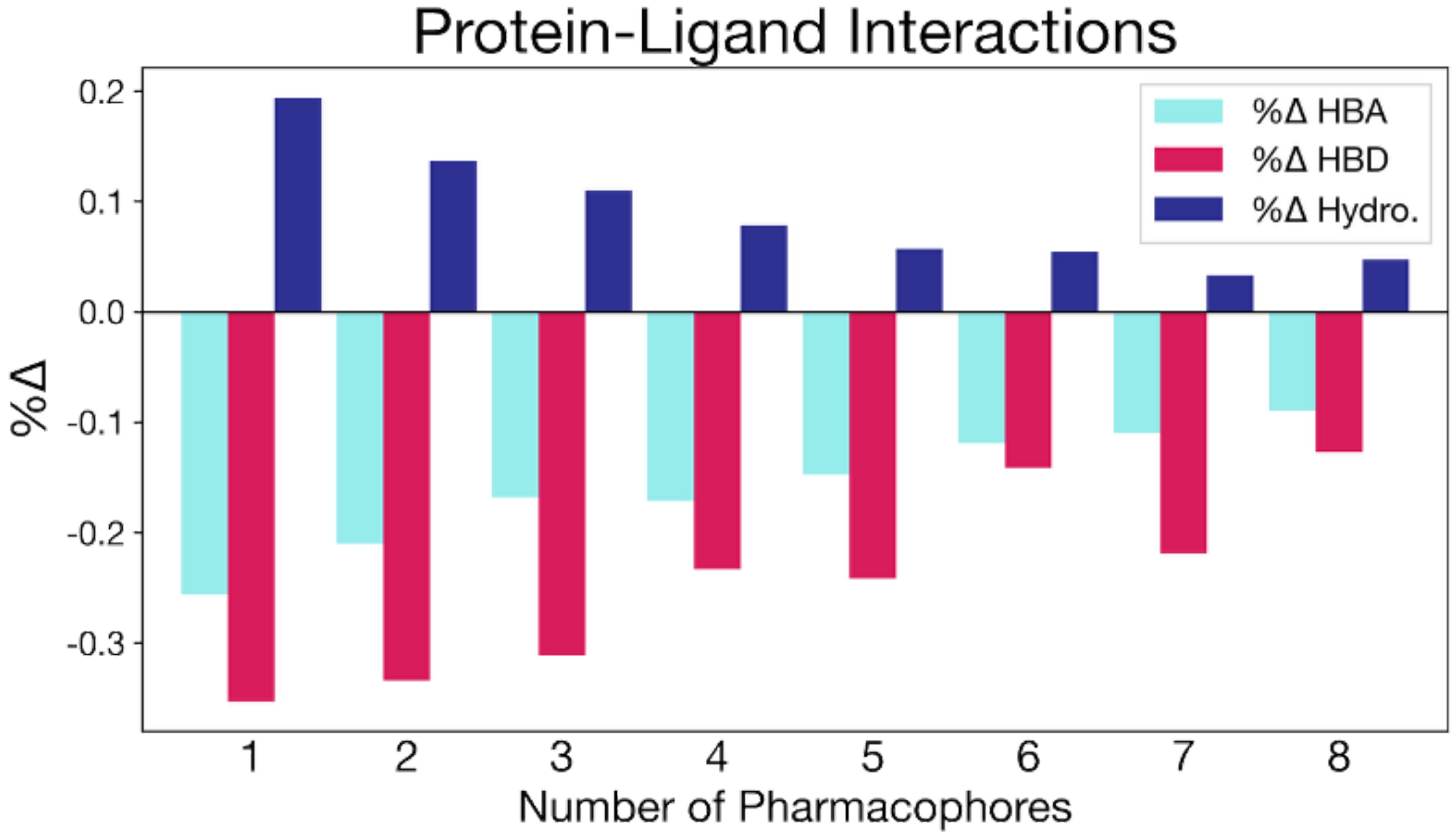


		Prot Conditioning	Prot + Pharm Conditioning
de novo design	%PB-Valid	67.5	66.2
	interaction recovery	51.0	67.4
	% Pharm Matches	-	96.9
docking	% RMSD $\leq$ 2Å	93.0	99.0
	%PB-Valid	73.0	81.0
	% Pharm Matches	-	99.5

OMTRA: Pharmacophore Conditioning





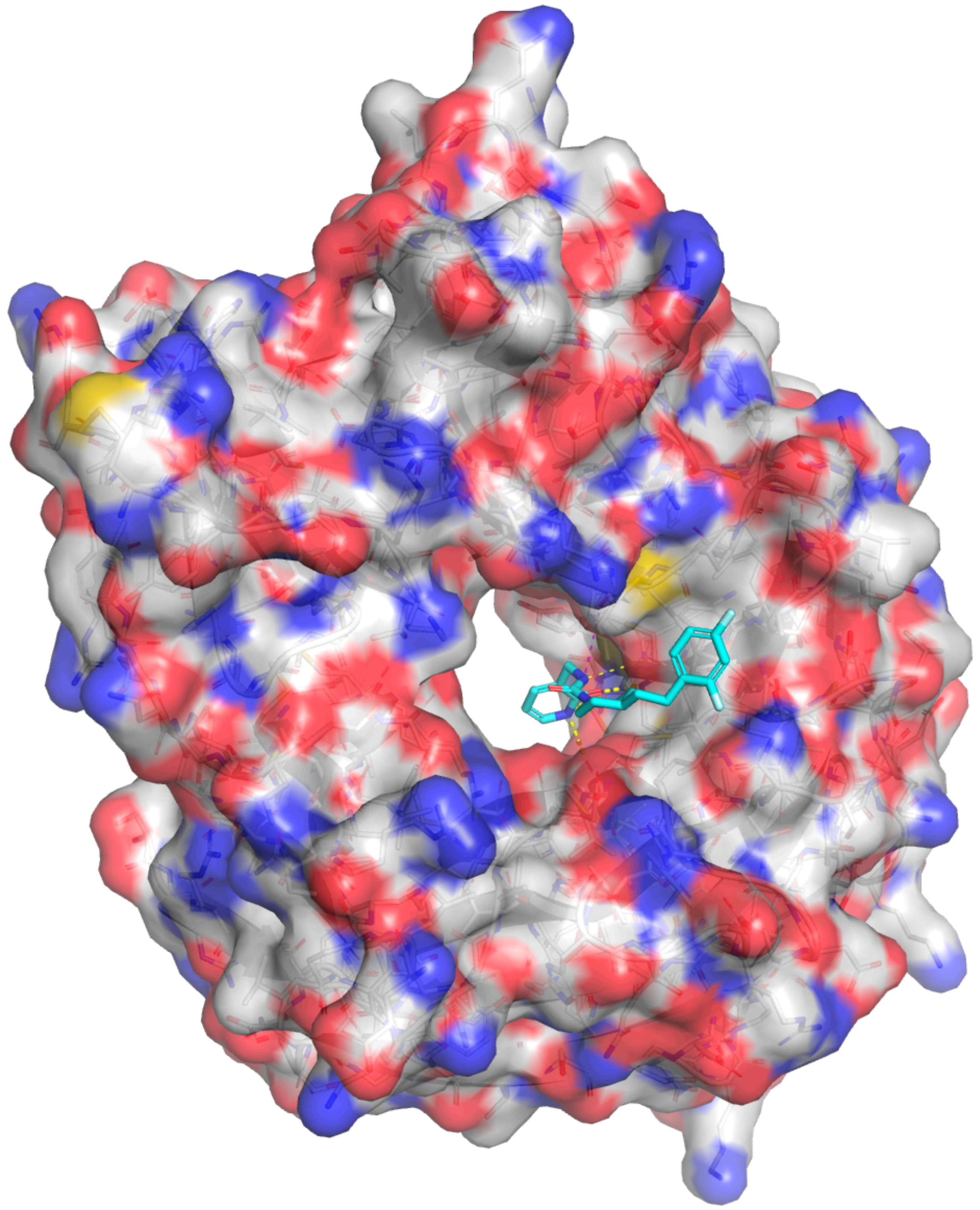


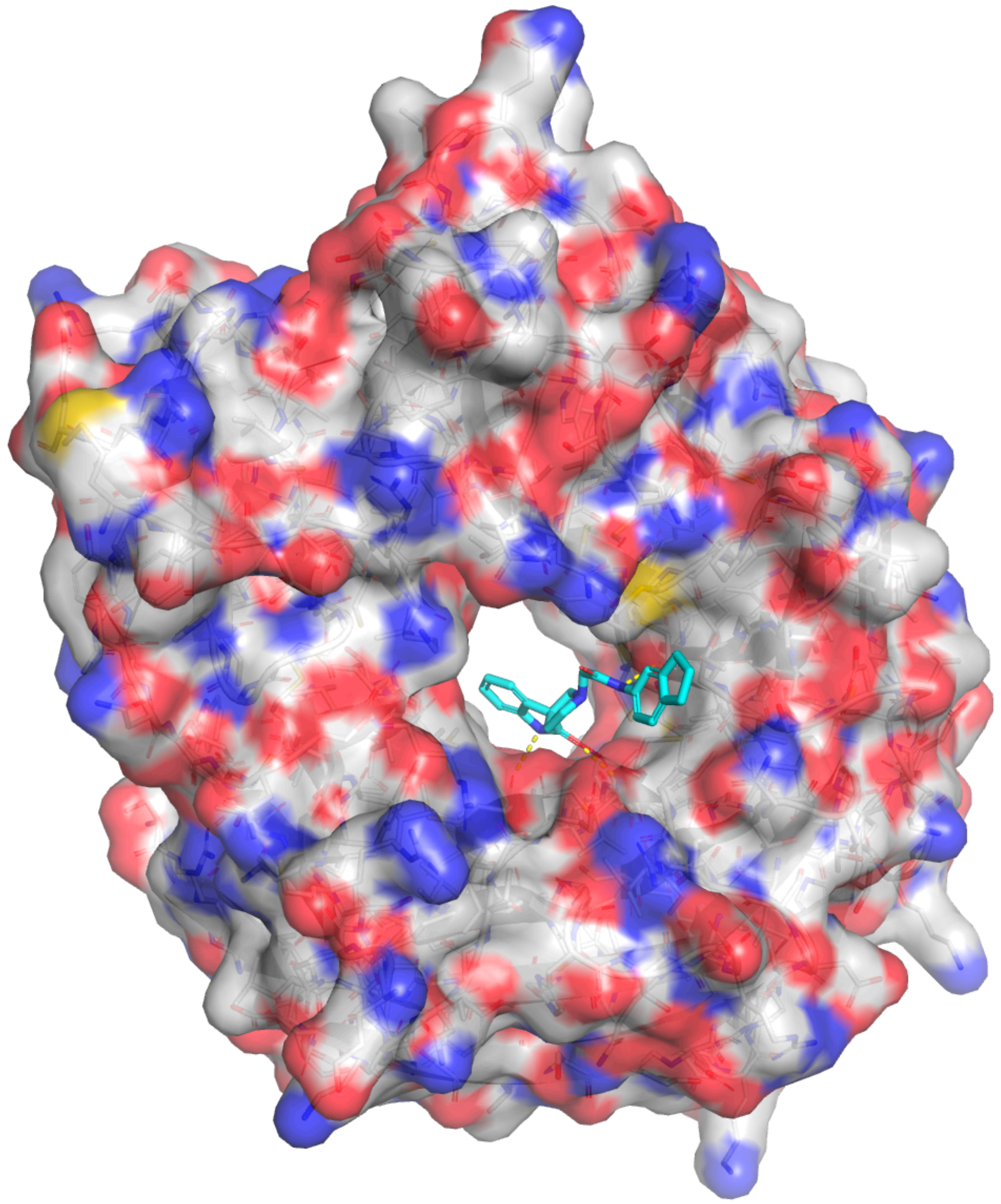
Case Studies

CACHE CHALLENGE

CRITICAL ASSESSMENT OF COMPUTATIONAL HIT-FINDING EXPERIMENTS

CACHE Challenge #1

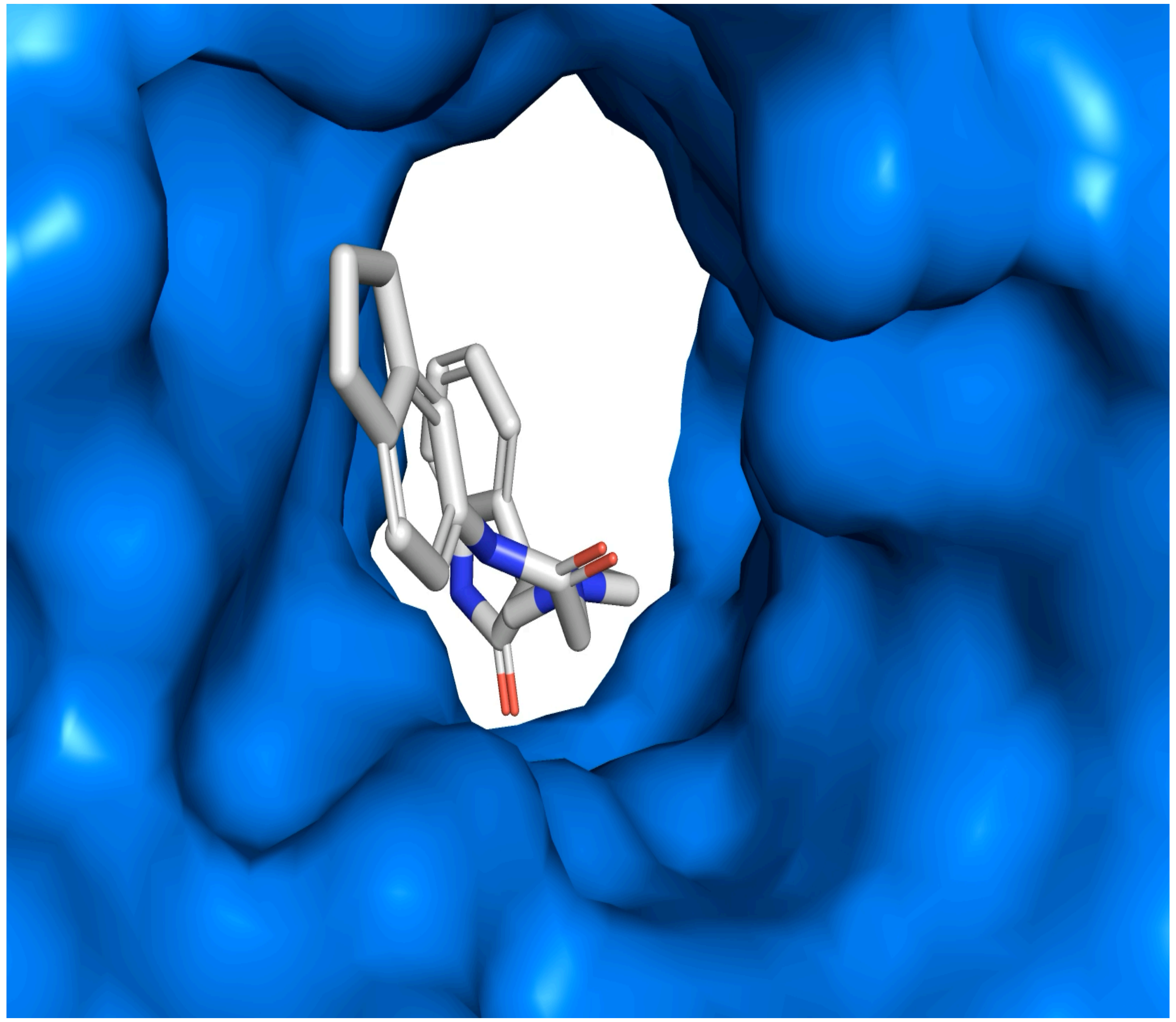




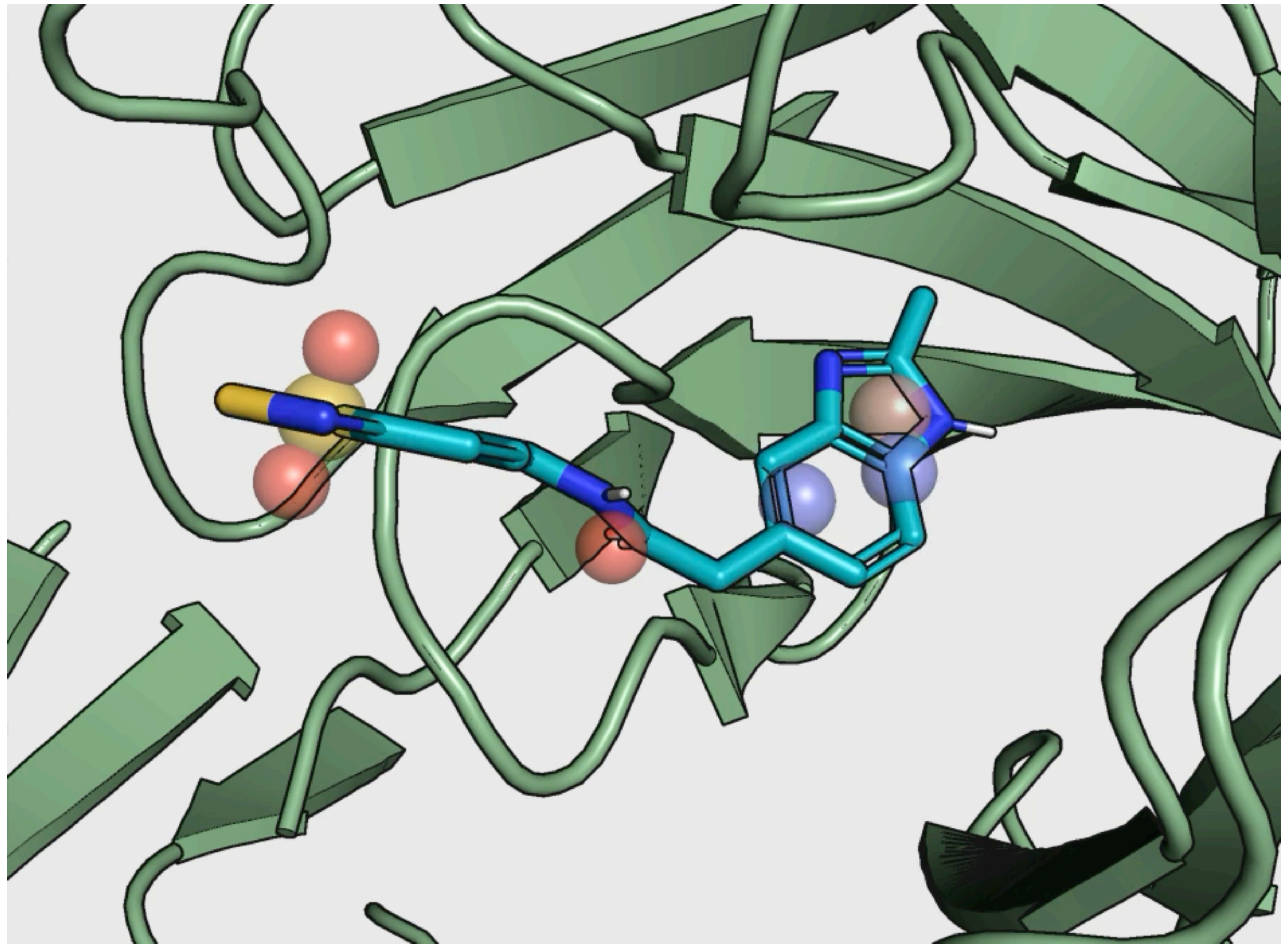
Ian Dunn

A Tale of Two Methods

Large-Scale Docking with GNINA

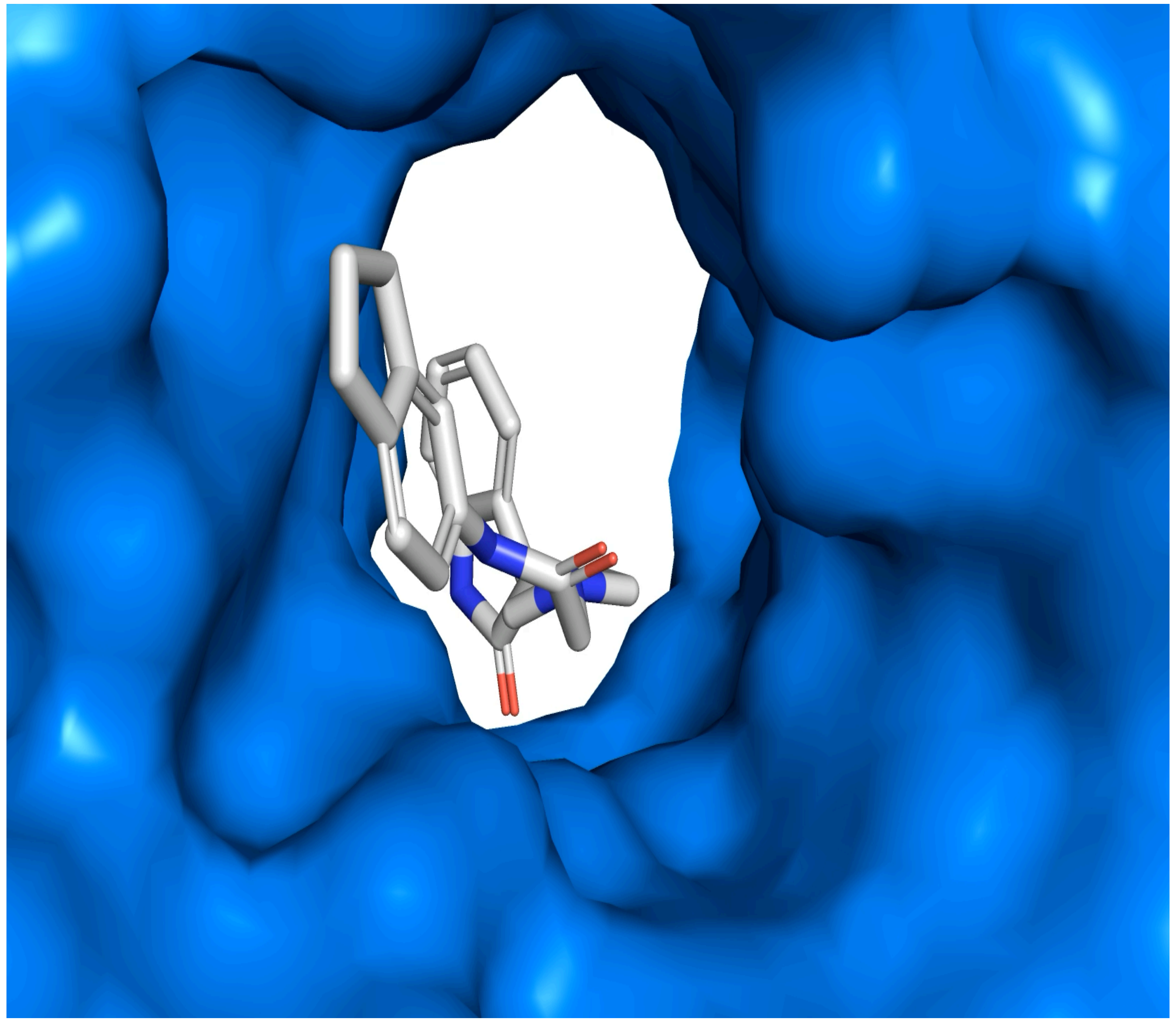


Pharmacophore Screening with Pharmit

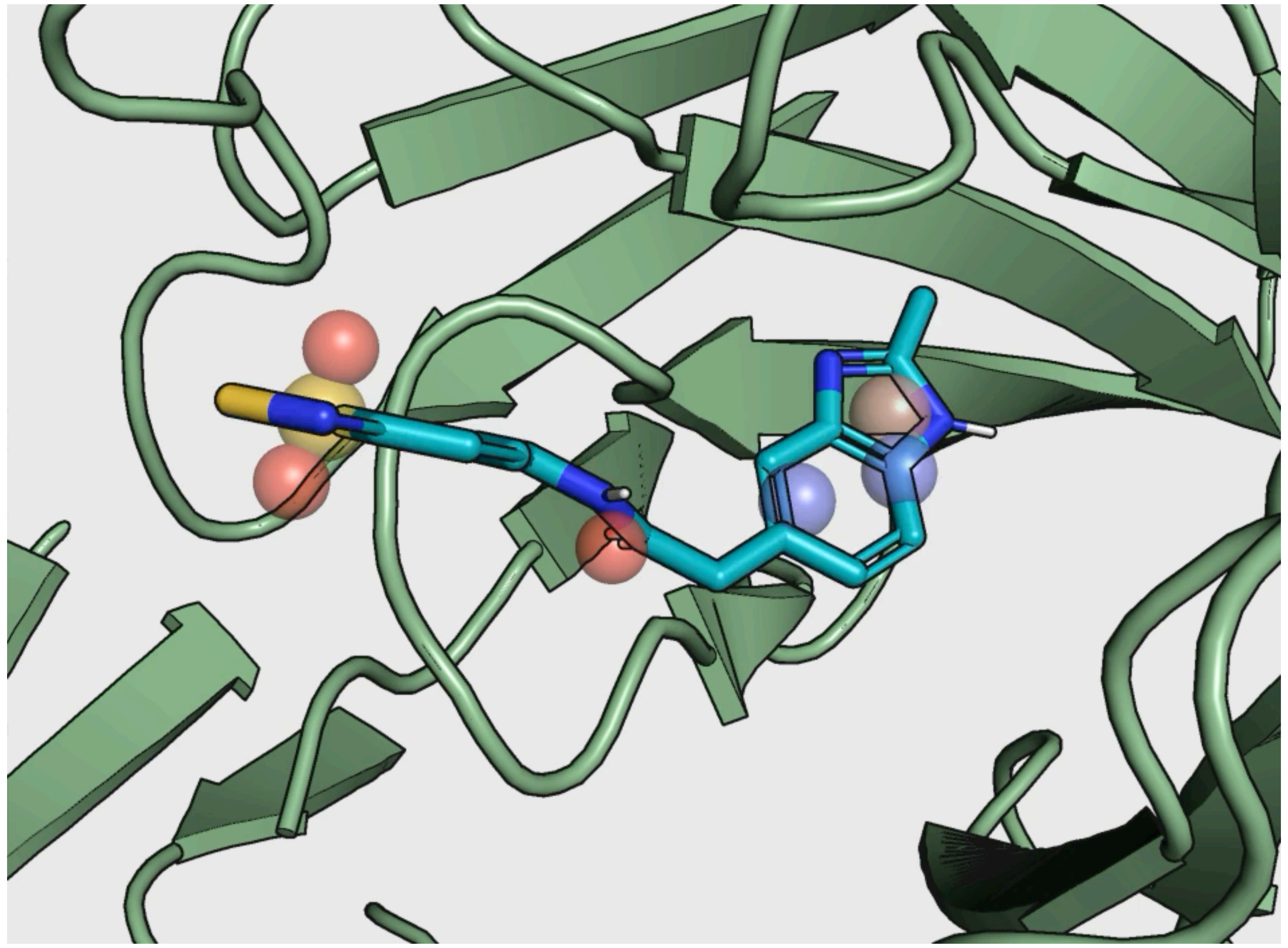


A Tale of Two Methods

Large-Scale Docking with GNINA

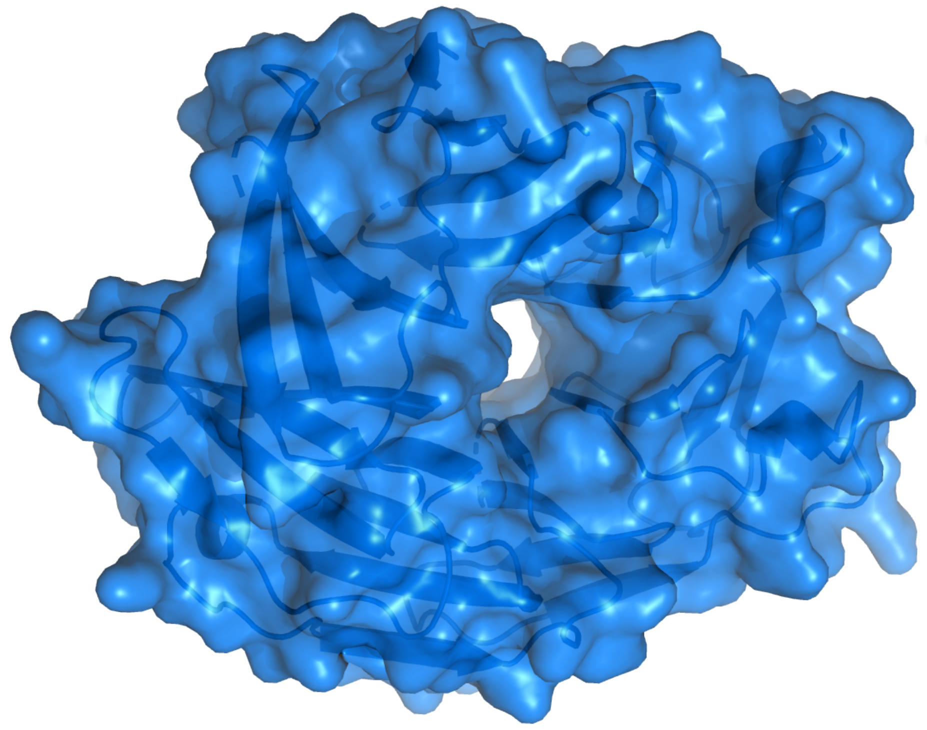


Pharmacophore Screening with Pharmit



High-throughput Docking Pipeline

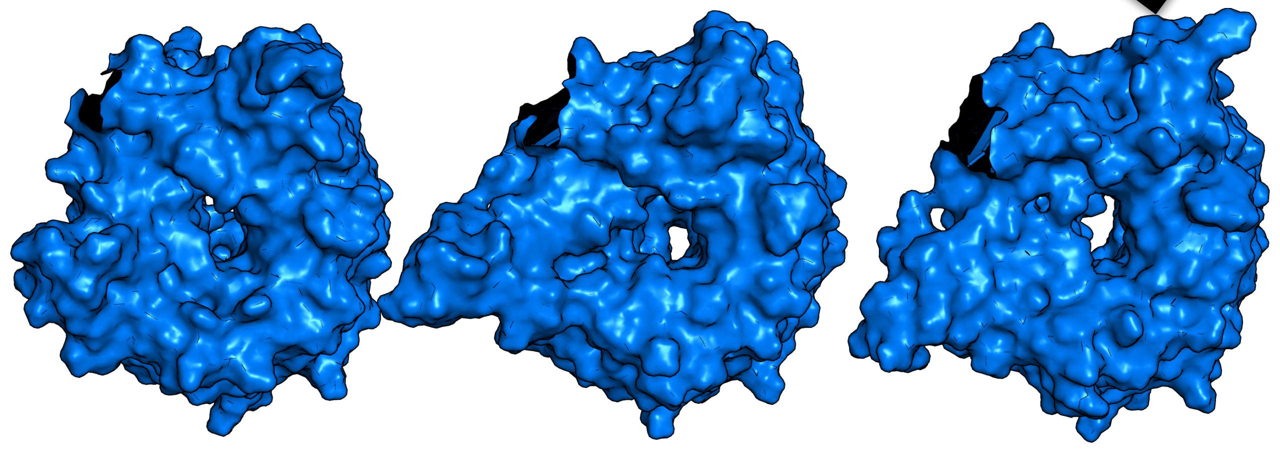
molport
~7 million molecules



Crystal Structure

Docking with GNINA

Top 1k Molecules
by GNINA score

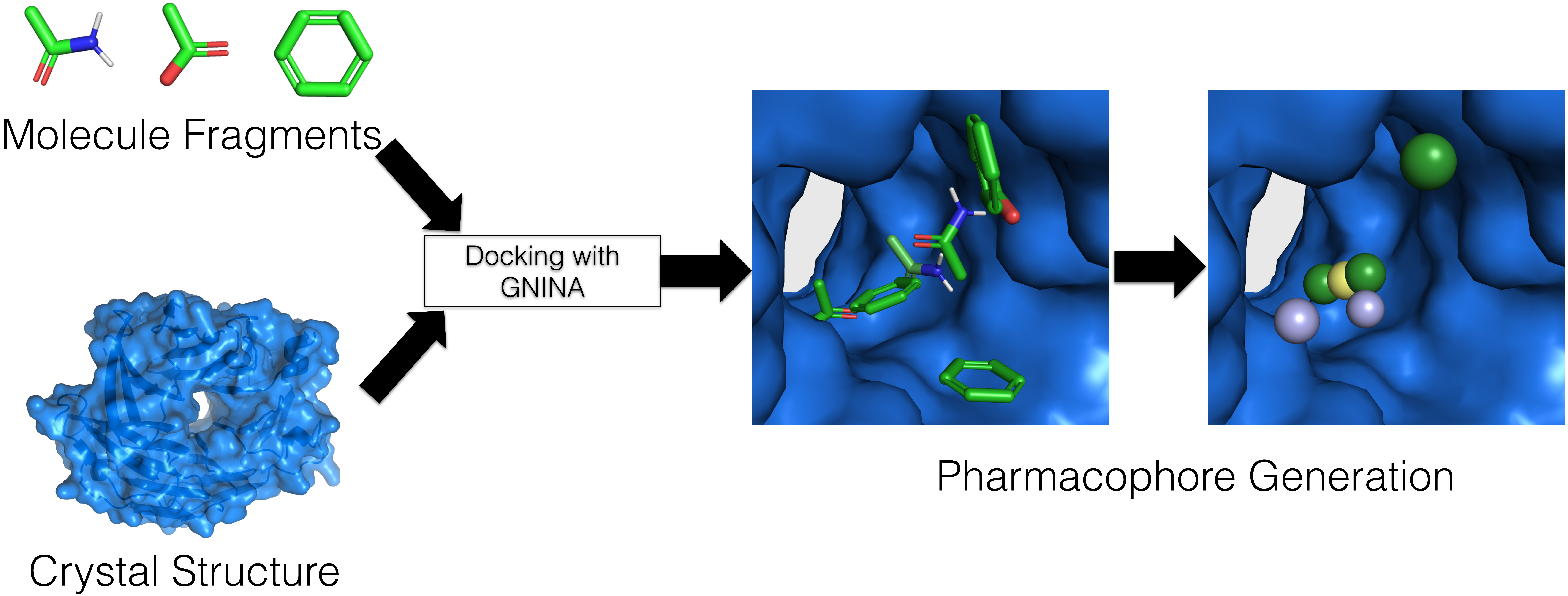


MD Ensemble Structures

Docking with GNINA

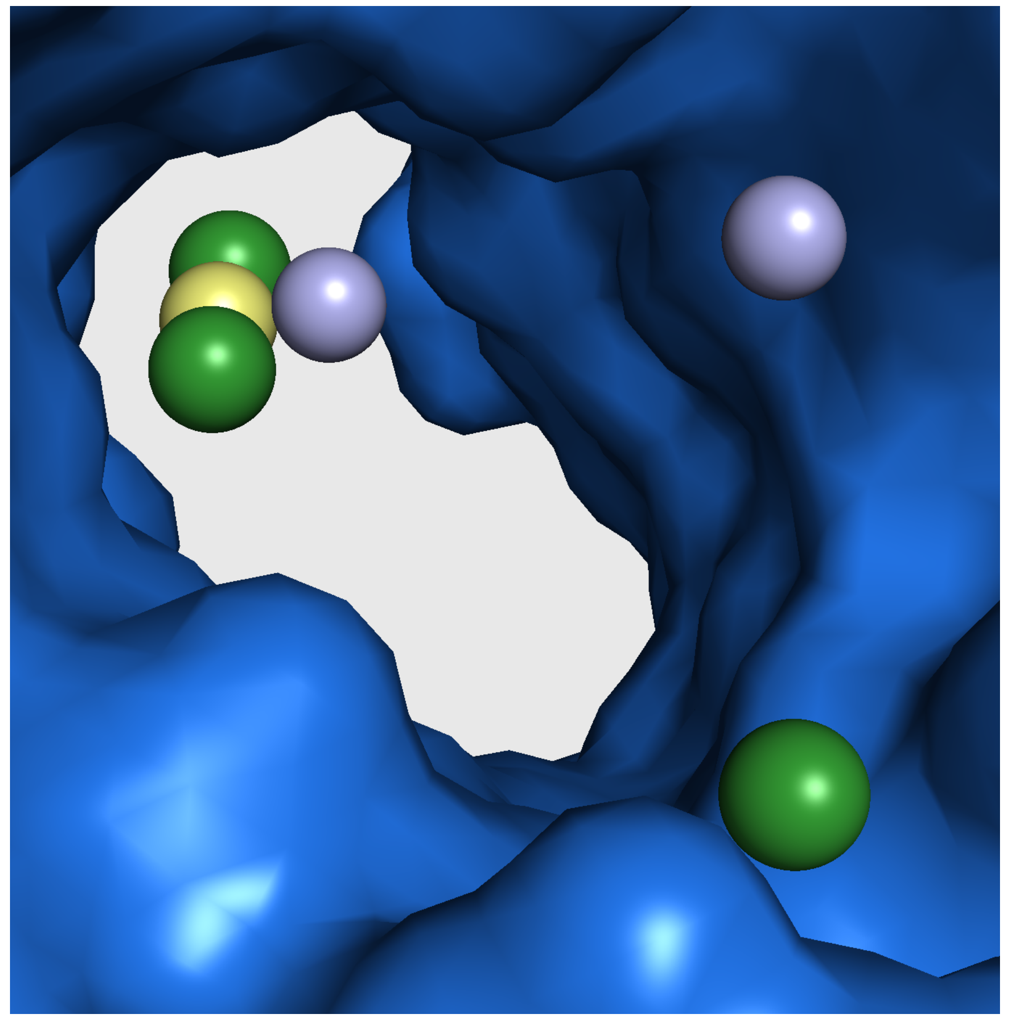
1000 molecules,
+
GNINA Scores

Pharmacophore Generation via Fragment Docking

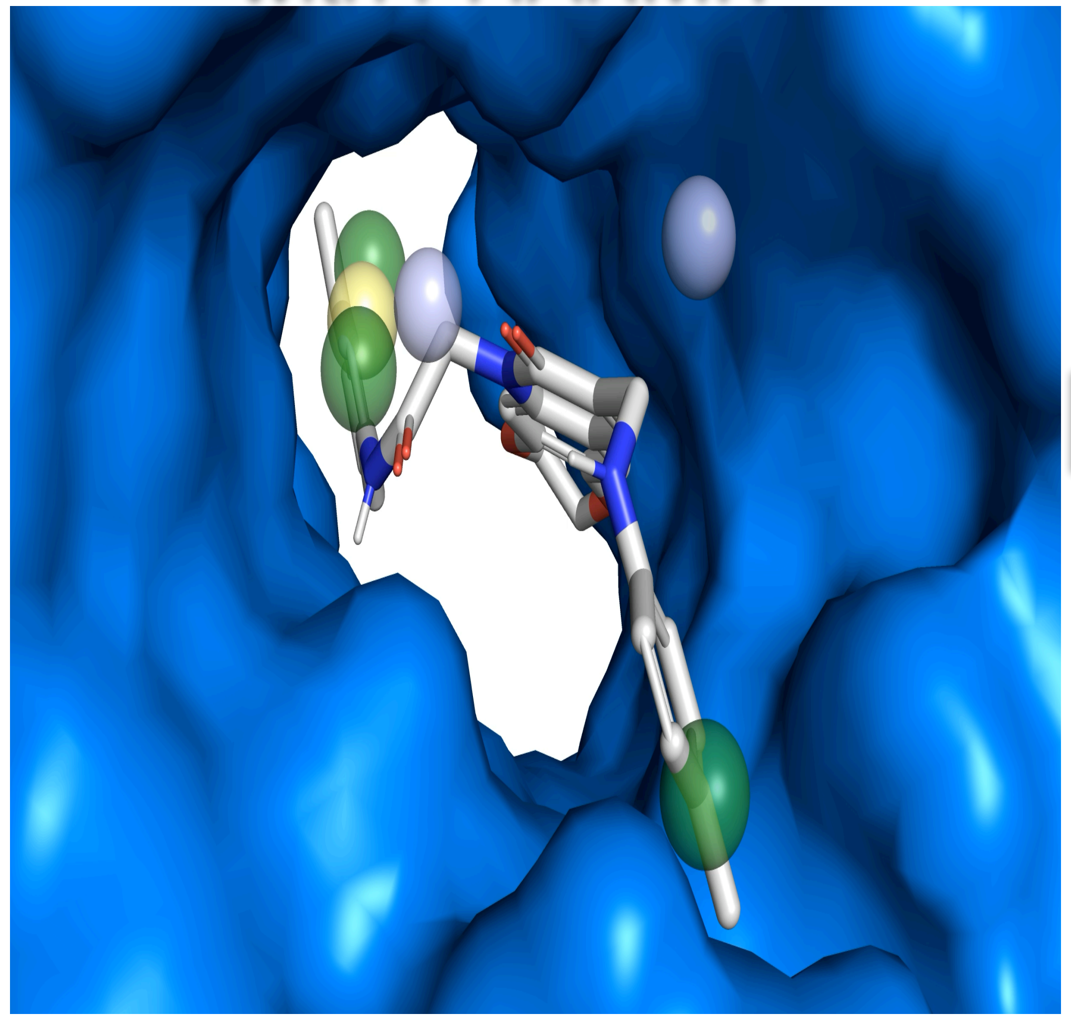


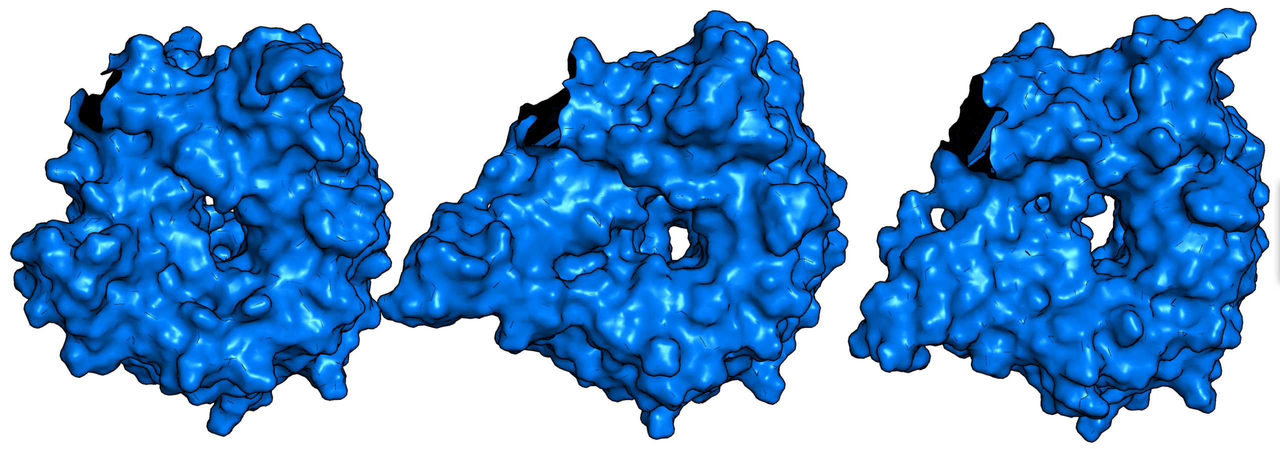
Pharmacophore Pipeline

Pharmacophore Generation



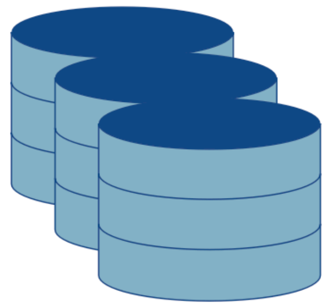
Pharmacophore Screening with PHARMIT





MD Ensemble Docking with GNINA

3572 molecules + GNINA scores

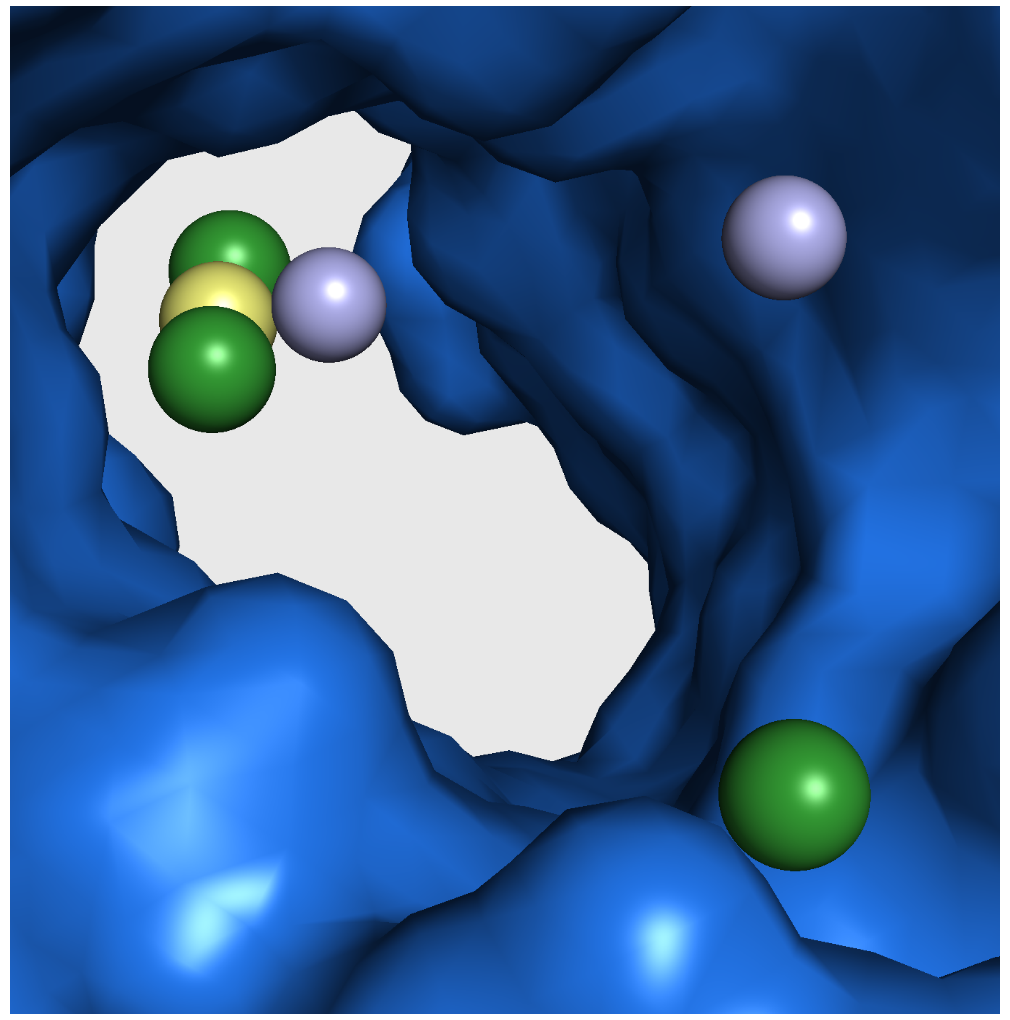


- ZINC20: 20 mil molecules
- MCULE: 45 mil molecules
- MCULE-ULTIMATE: 126 mil molecules

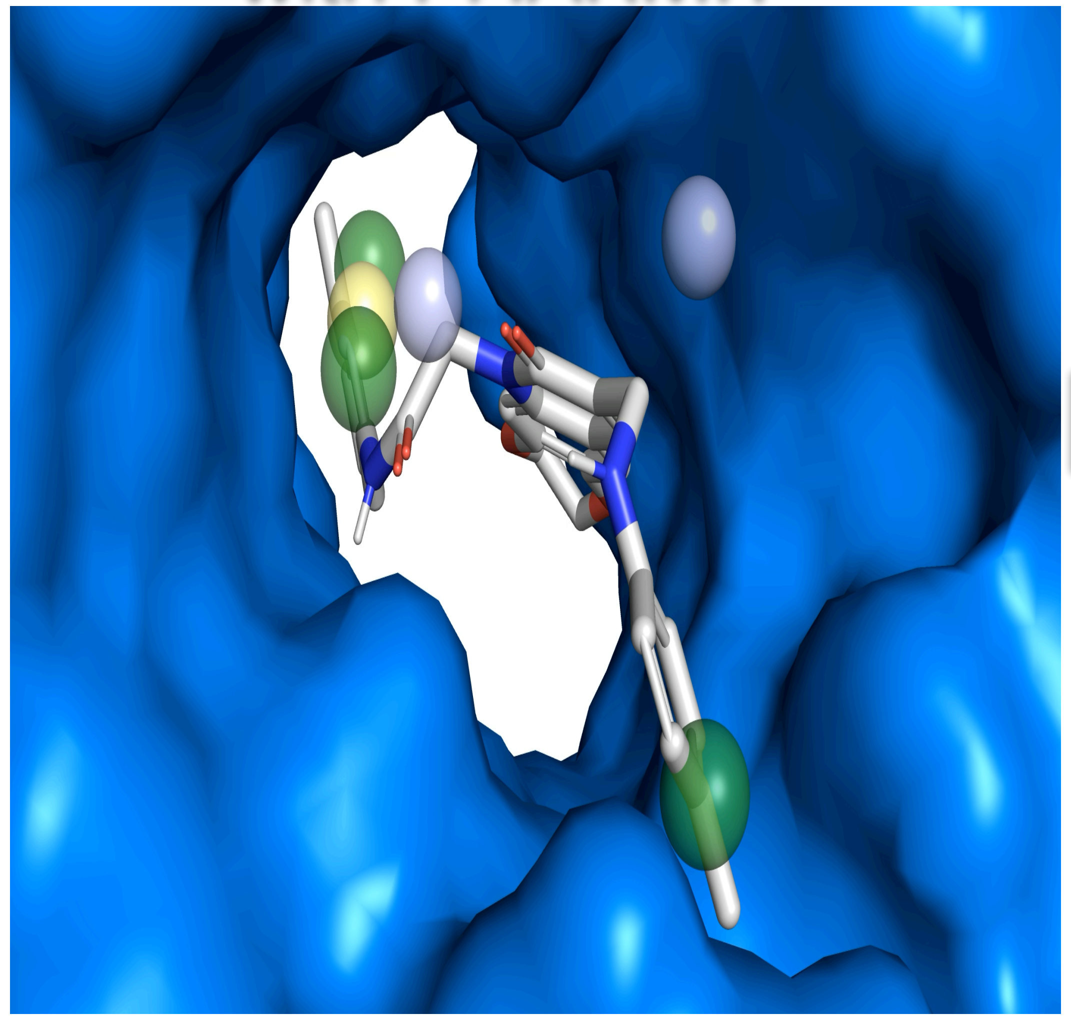
Molecule Libraries

Pharmacophore Pipeline

Pharmacophore Generation

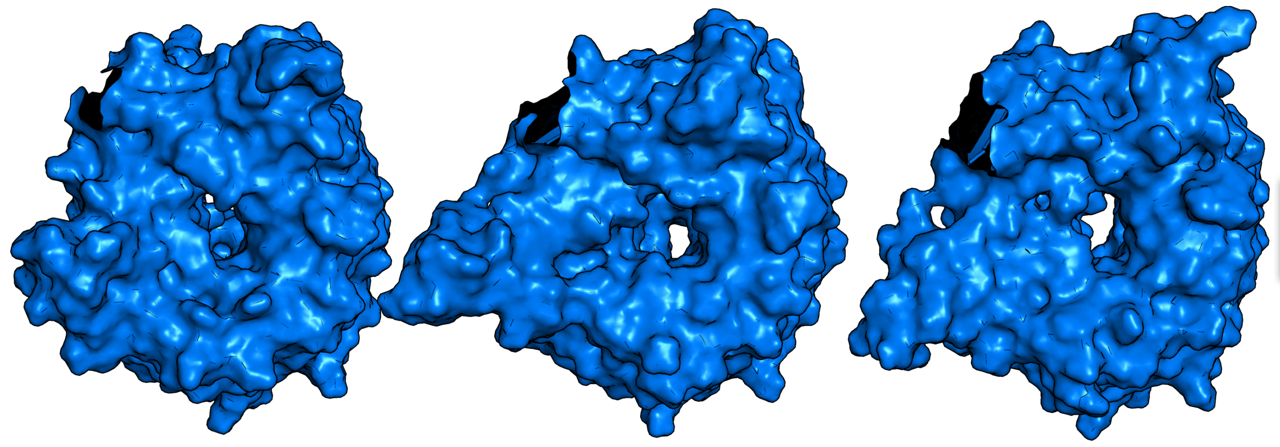


Pharmacophore Screening with PHARMIT



- ZINC20: 20 mil molecules
- MCULE: 45 mil molecules
- MCULE-ULTIMATE: 126 mil molecules

Molecule Libraries

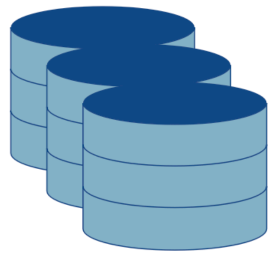


MD Ensemble Docking with GNINA

3572 molecules + GNINA scores

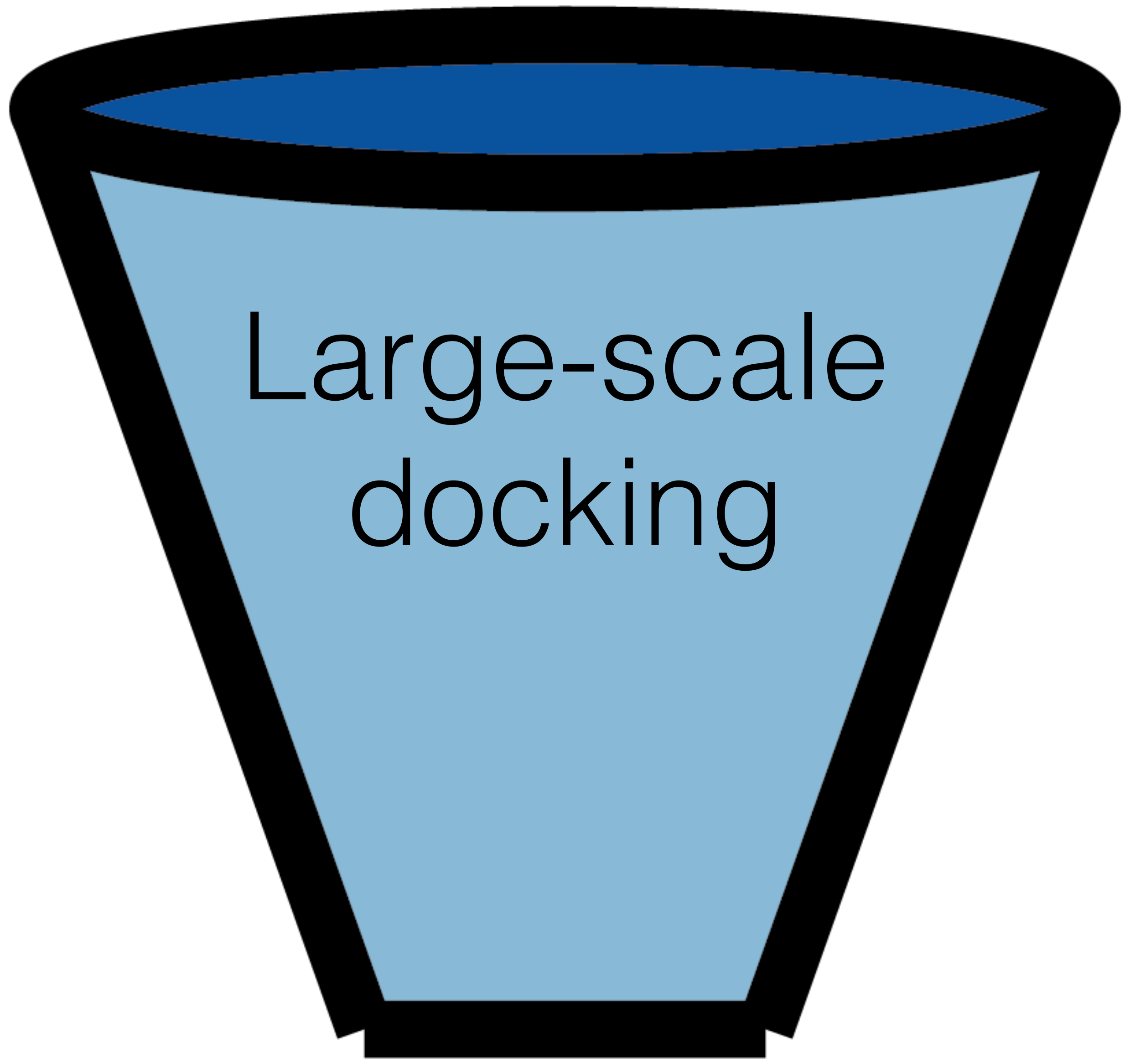
Round 1 Submission

molport

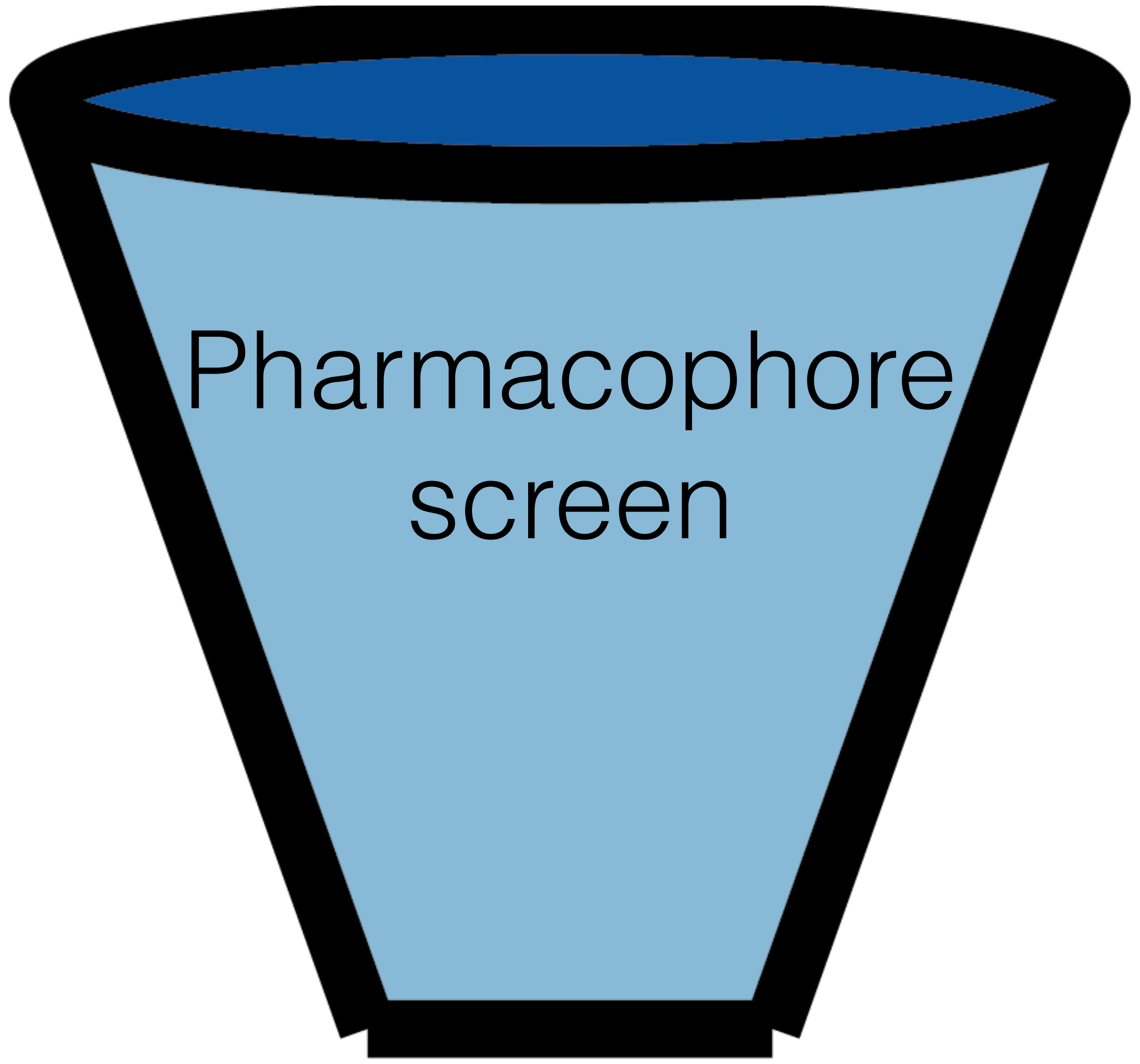


- ZINC20: 20 mil molecules
- MCULE: 45 mil molecules
- MCULE-ULTIMATE: 126 mil molecules

Molecule Libraries



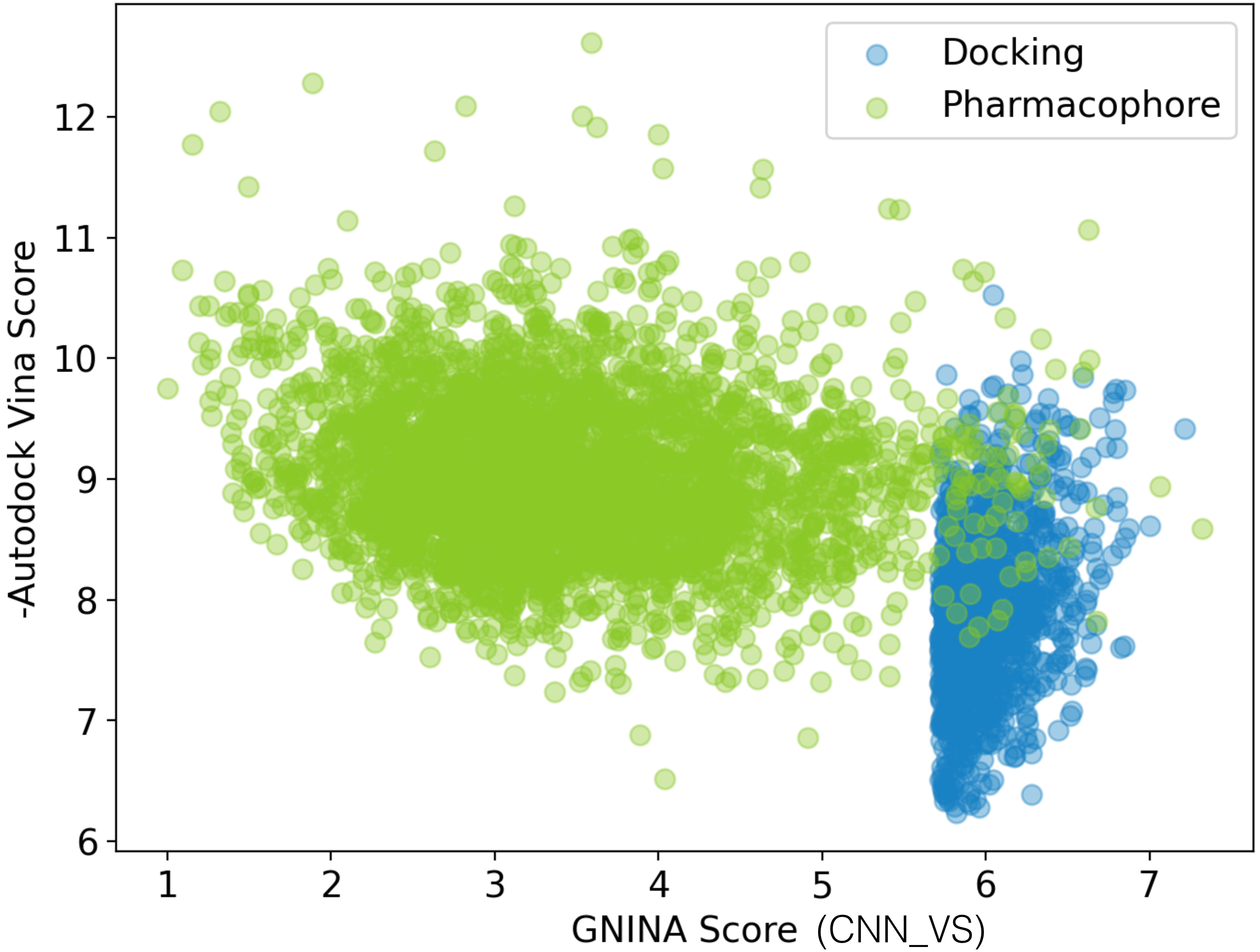
2 screening methods
2 scoring methods



1k ligands
gnina scores
vina scores

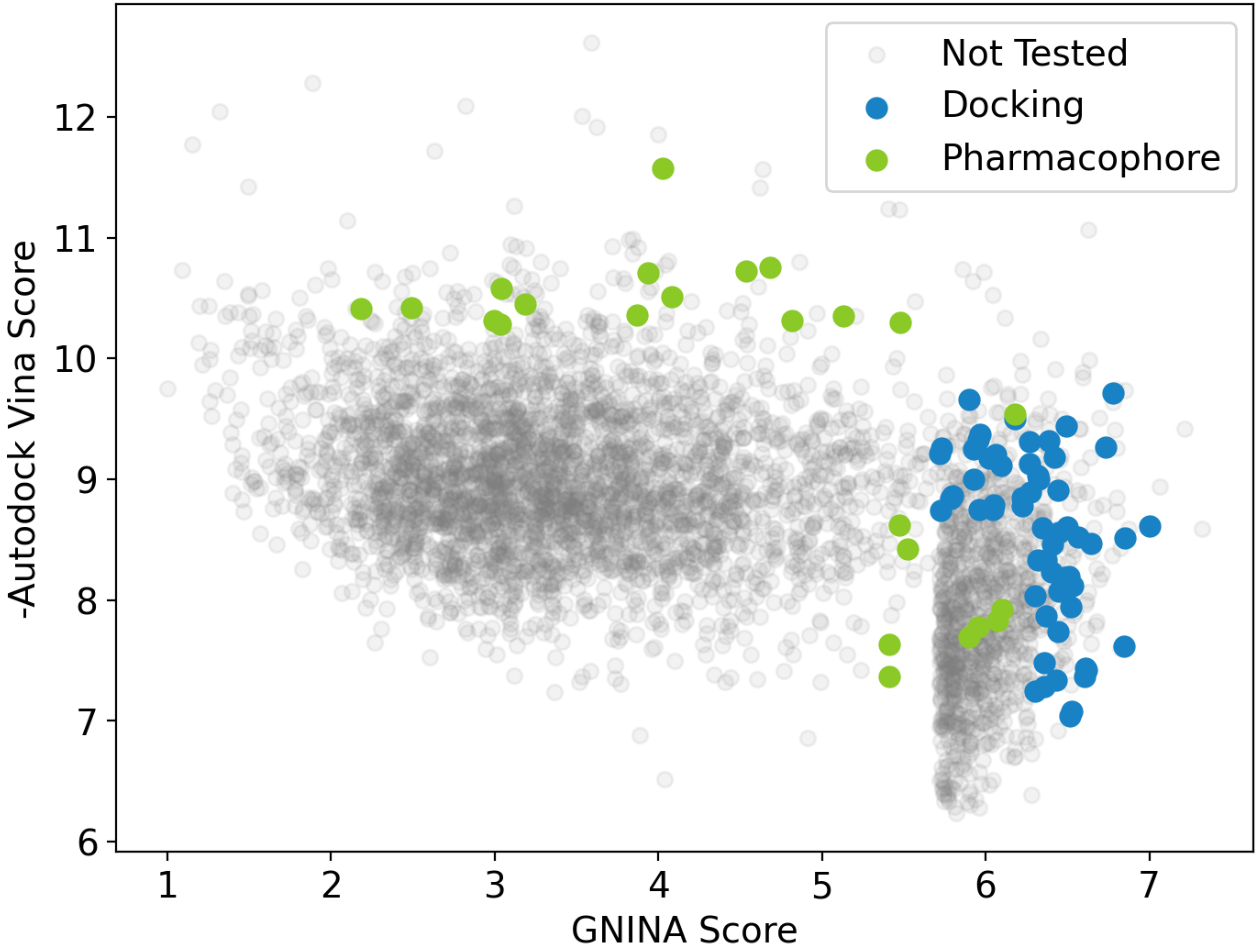
3.5k ligands
gnina scores
vina scores

Round 1 Results



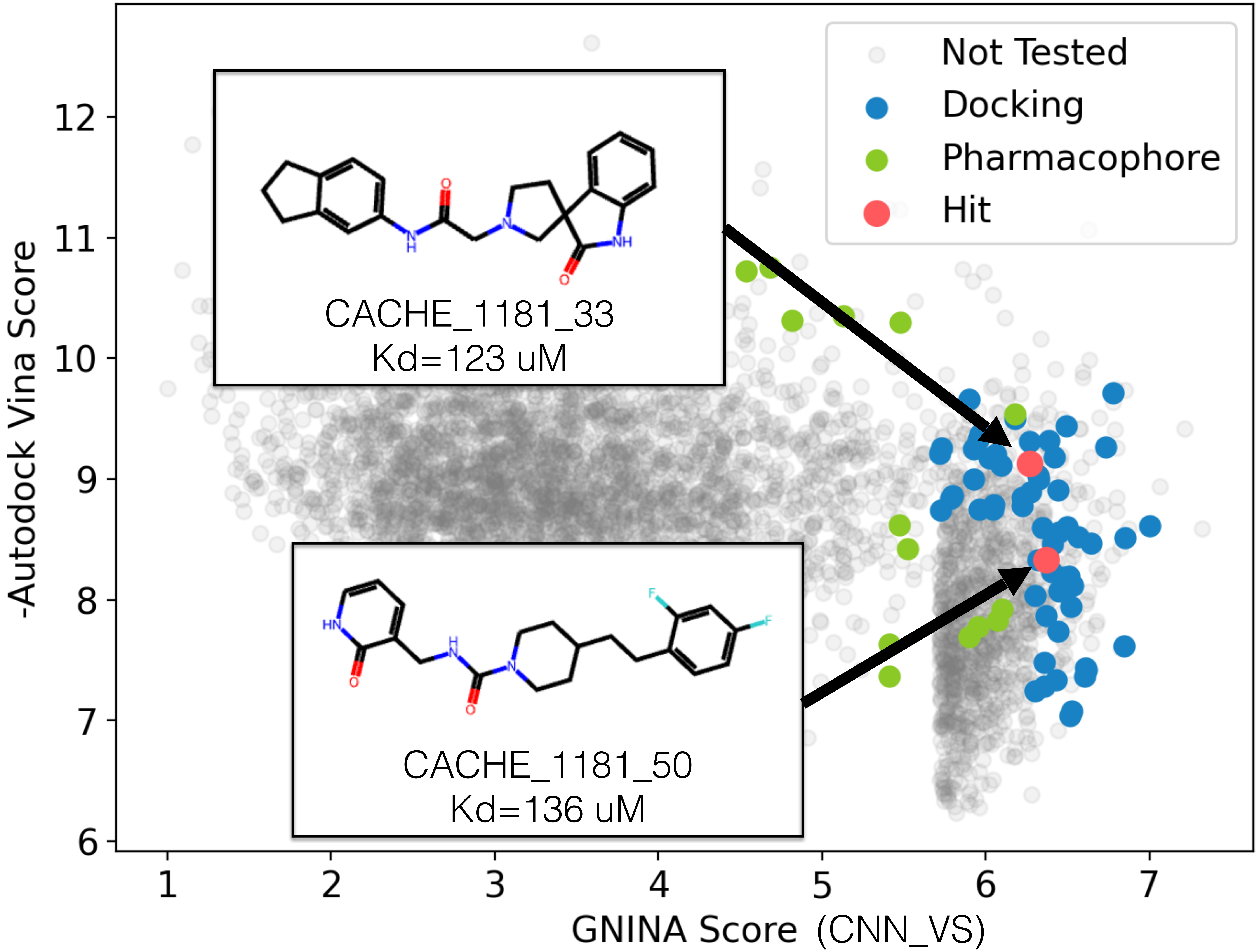
- Selection limited/ skewed by database availability
- 84 ligands tested
  - 59 from docking
  - 24 from pharm screen

Round 1 Results



- Selection limited/ skewed by database availability
- 84 ligands tested
  - 59 from docking
  - 24 from pharm screen

Round 1 Results

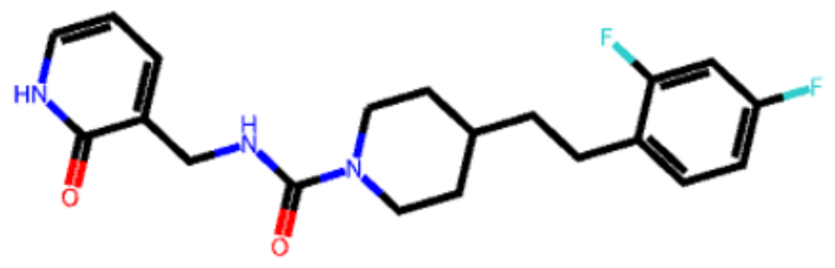


- 2/84 were hits
  - Both from docking

Round 2: Hit Optimization

Hit Optimization Pipeline

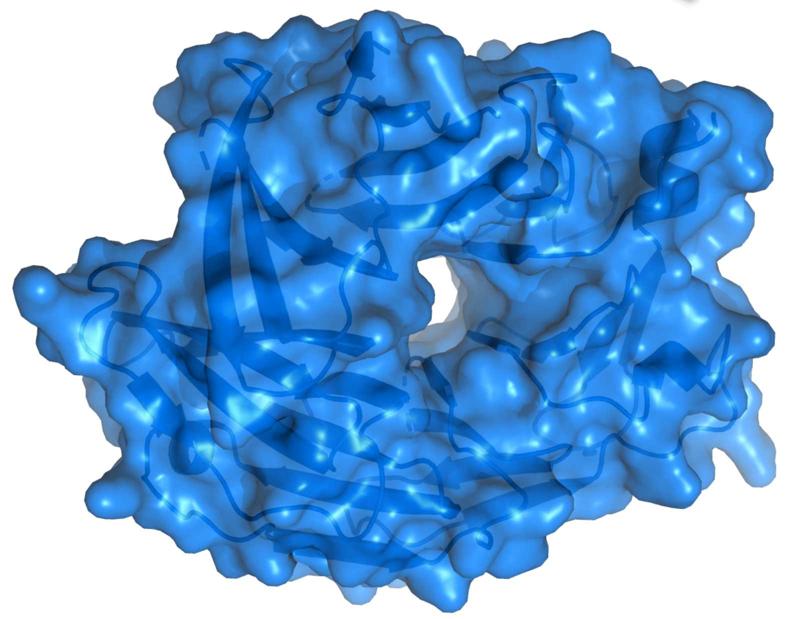
Parent Compound



Similarity screen against
Enamine REAL
Return 5000 most similar
ligands by tanimoto score

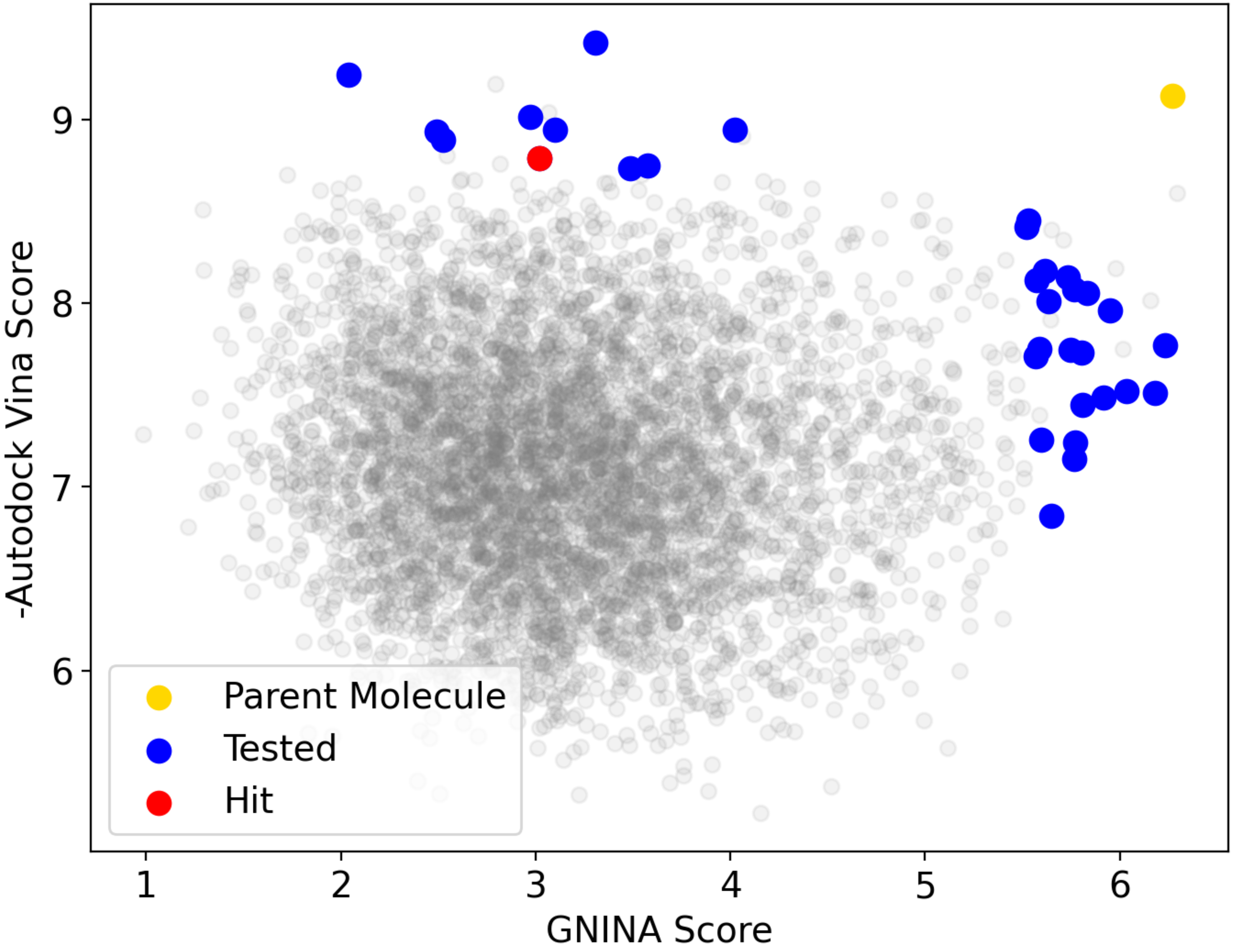
Docking with
GNINA

5000 molecules
+
GNINA scores

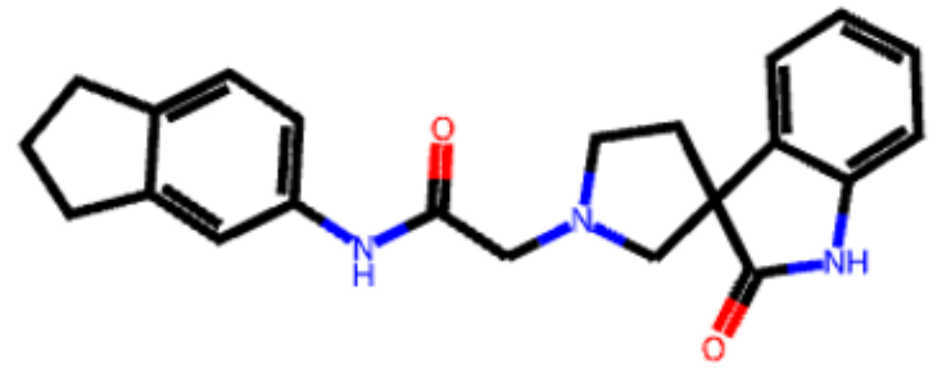


Crystal Structure

Hit Optimization Results

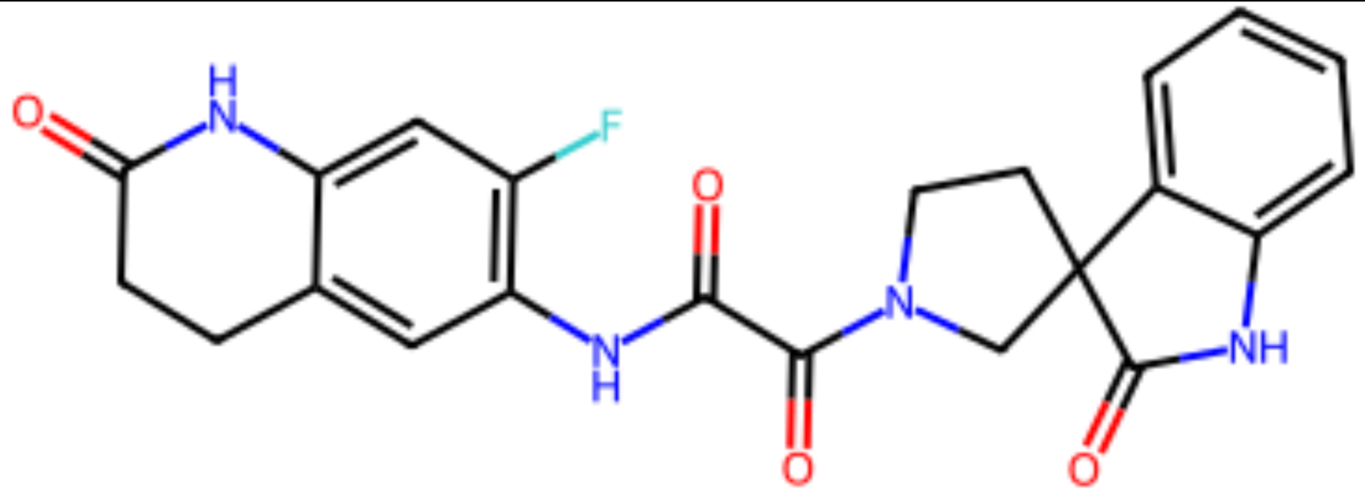


Parent Compound



CACHE_1181_33
Kd=123 uM

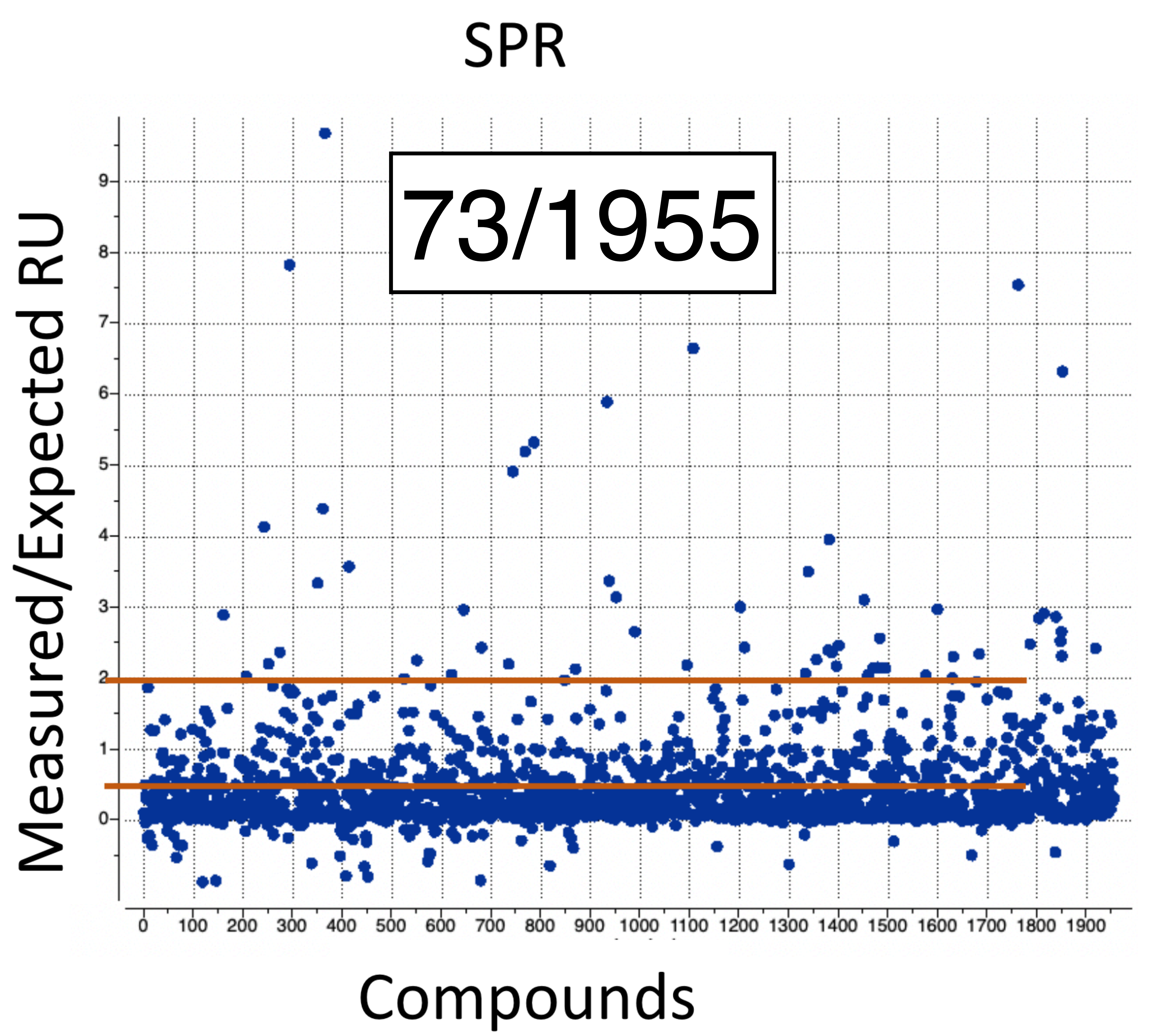
Hit Compound



CACHE-HO_1181_24
Kd=56 uM

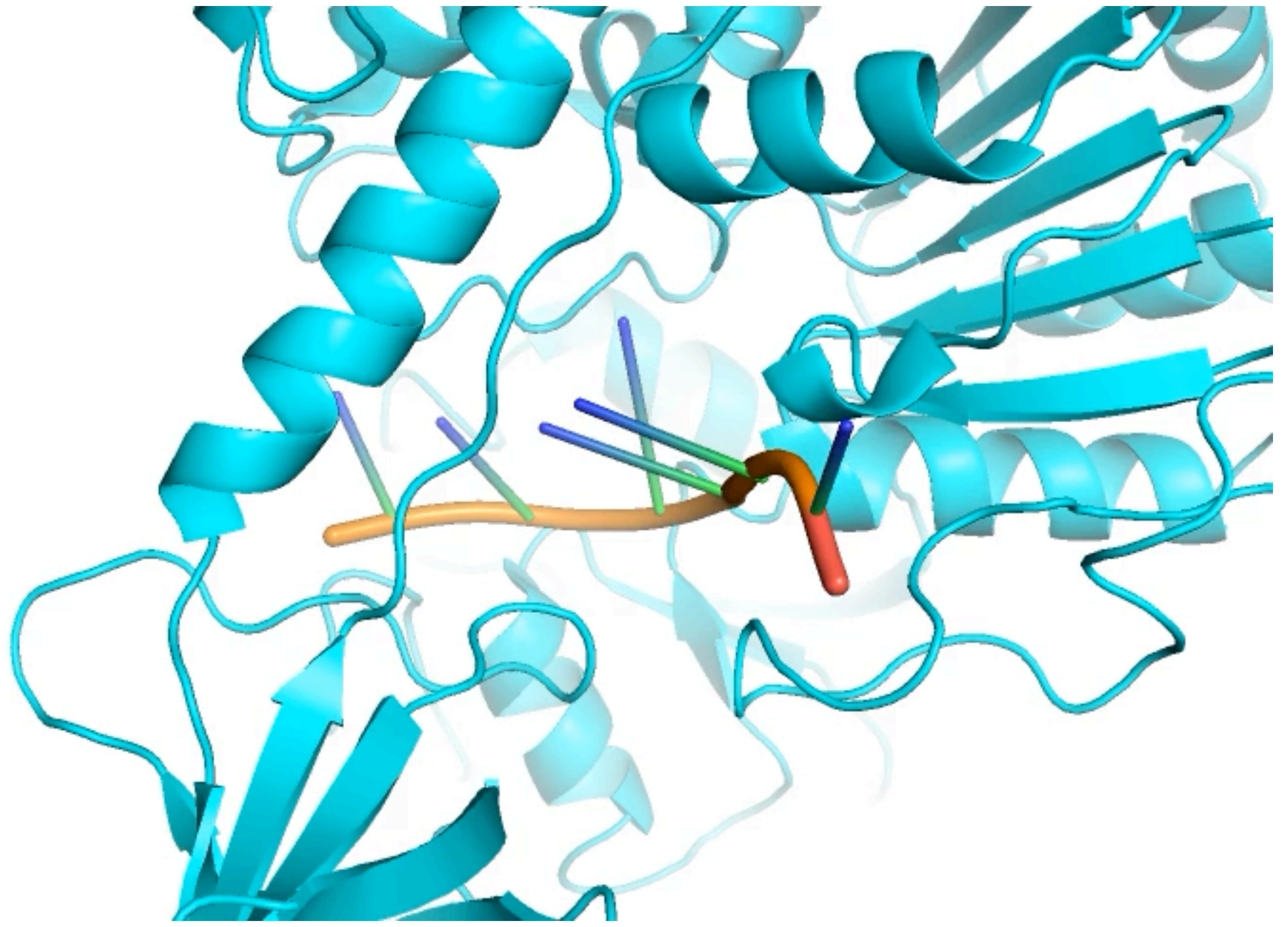
Final Results

Participant	Participant ID	Aggregated score
David Koes, University of Pittsburgh	1181	18
Olexandr Isayev & Maria Kurnikova, Carnegie Mellon University & Artem Cherkasov, University of British Columbia	1209	18
Christina Schindler, Merck KGaA	1193	17
Dmitri Kireev, University of Missouri	1183	16
Christoph Gorgulla, St. Jude Children's Research Hospital and Harvard University	1195	16
Didier Rognan, Université Strasbourg	1202	16
Pavel Polishchuk, Palacky University	1210	16
Kam Zhang, Centre for Biosystems Dynamic Research, RIKEN	1188	15
Shuangjia Zheng, Shanghai Jiao Tong University (previously Galixir)	1187	14
Carlos Zepeda, Treventis/UHN	1200	14
Fabian Liessmann, Leipzig University	1201	14
	1179	13



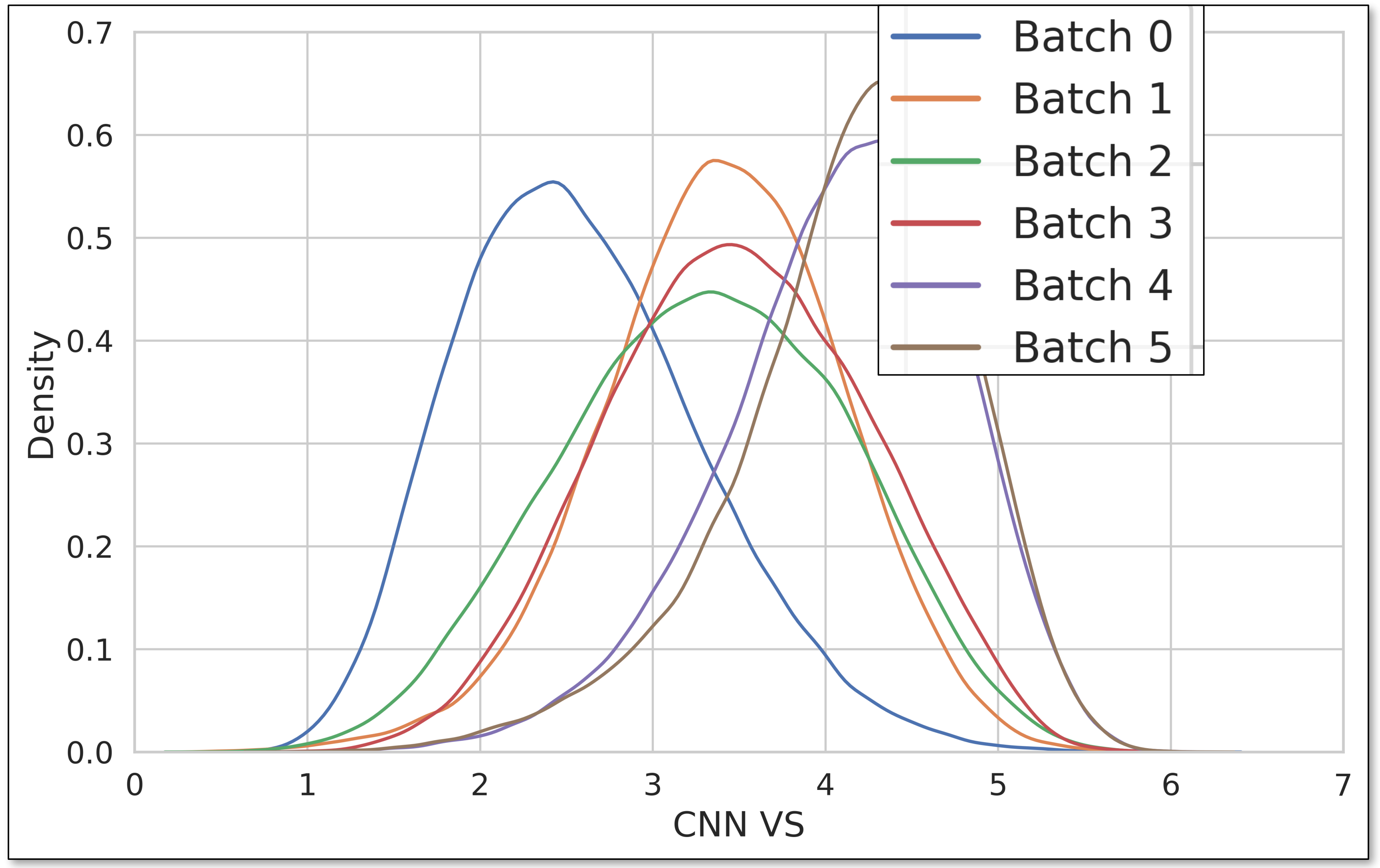
CACHE Challenge #2

- RNA binding site of SARS-COV2 NSP13
- “Deep Docking” of Enamine (4B)



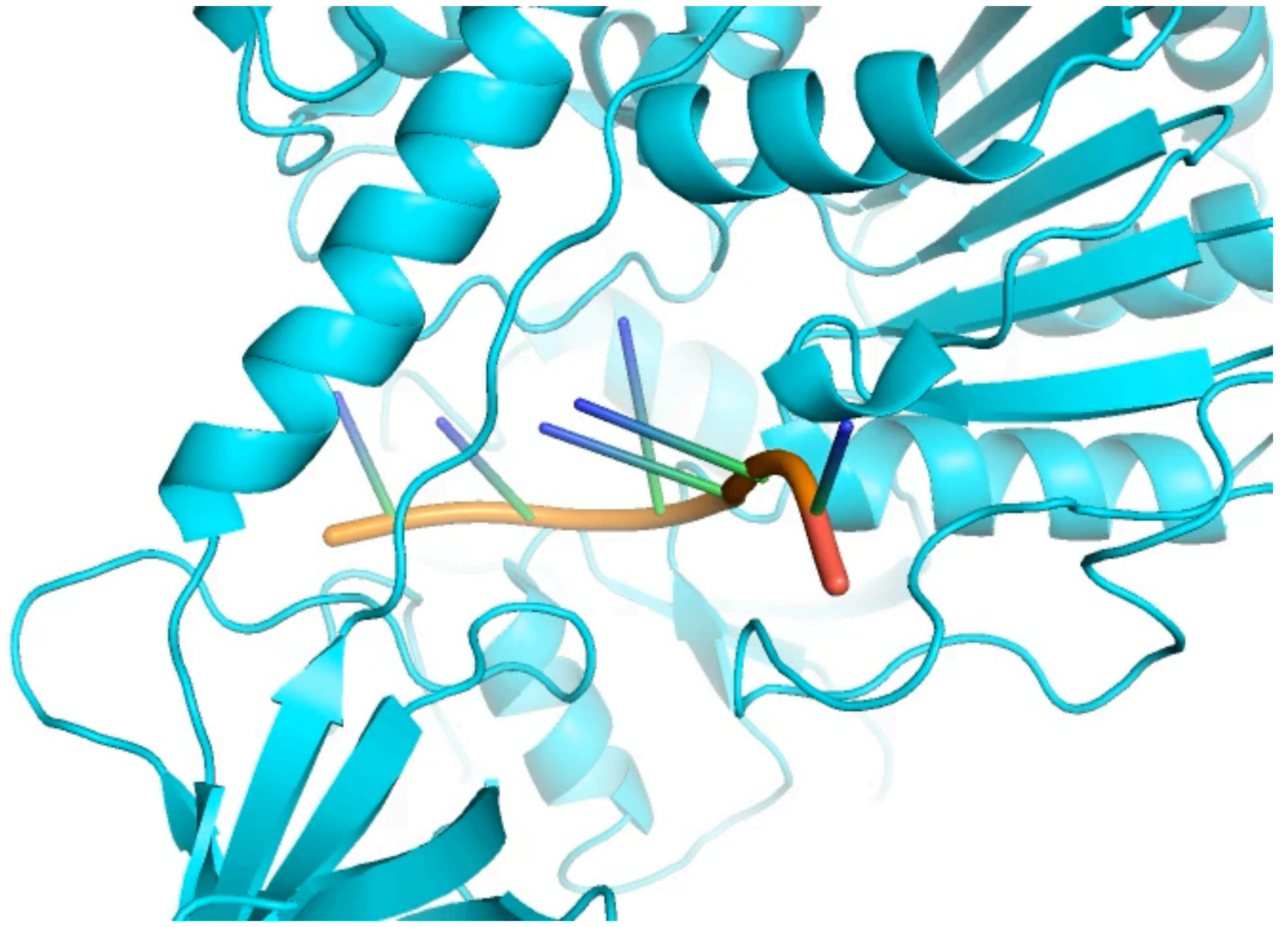
Randomly sample 100k ligands → Dock 100k ligands → Train XGBoost model to predict docking scores from ligand fingerprint → Predict docking scores for entire database → Select 100k molecules by highest predicted score → Dock 100k ligands

Ian Dunn



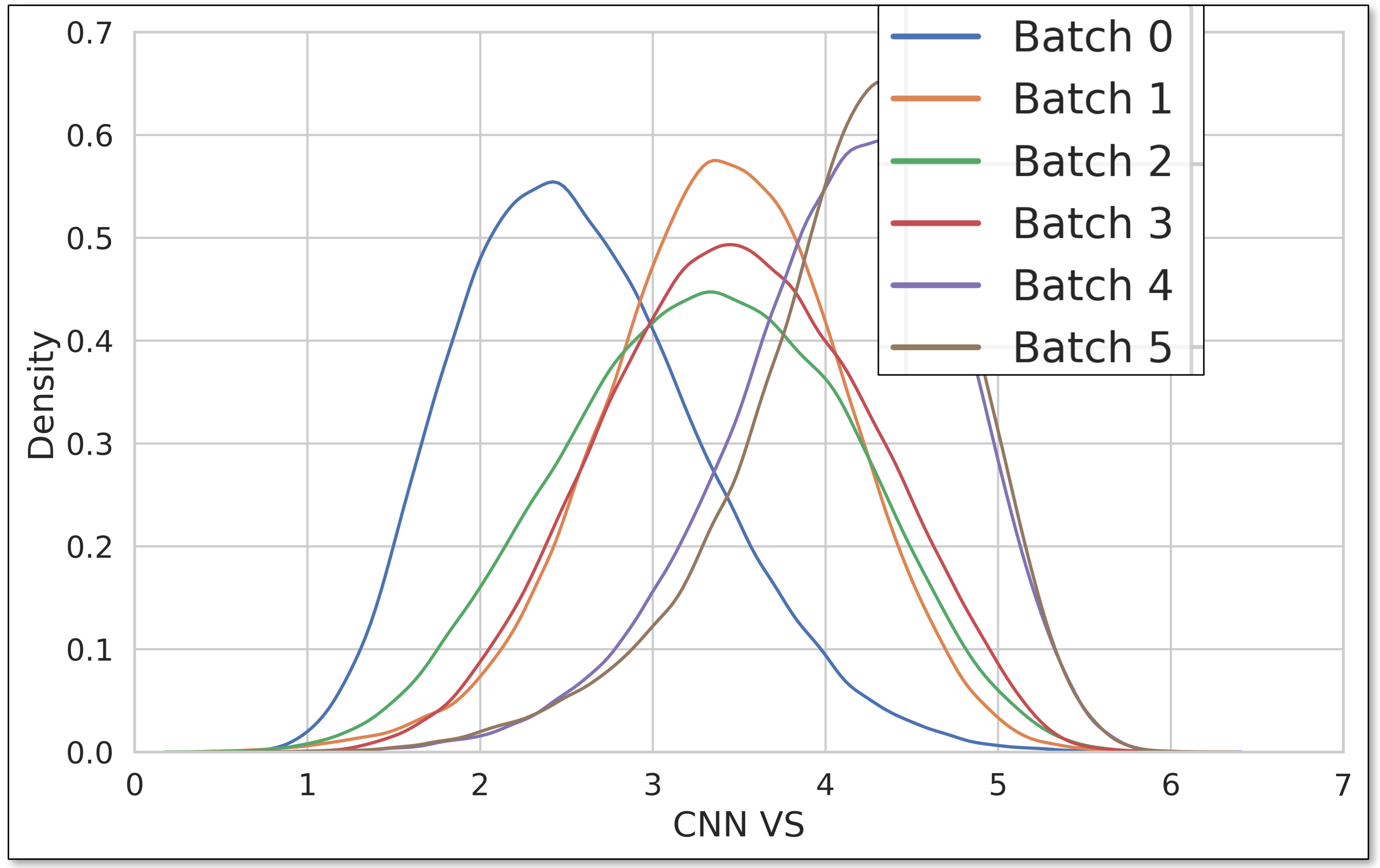
CACHE Challenge #2

- RNA binding site of SARS-COV2 NSP13
- “Deep Docking” of Enamine (4B)



Randomly sample 100k ligands → Dock 100k ligands → Train XGBoost model to predict docking scores from ligand fingerprint → Predict docking scores for entire database → Select 100k molecules by highest predicted score → Dock 100k ligands

Ian Dunn



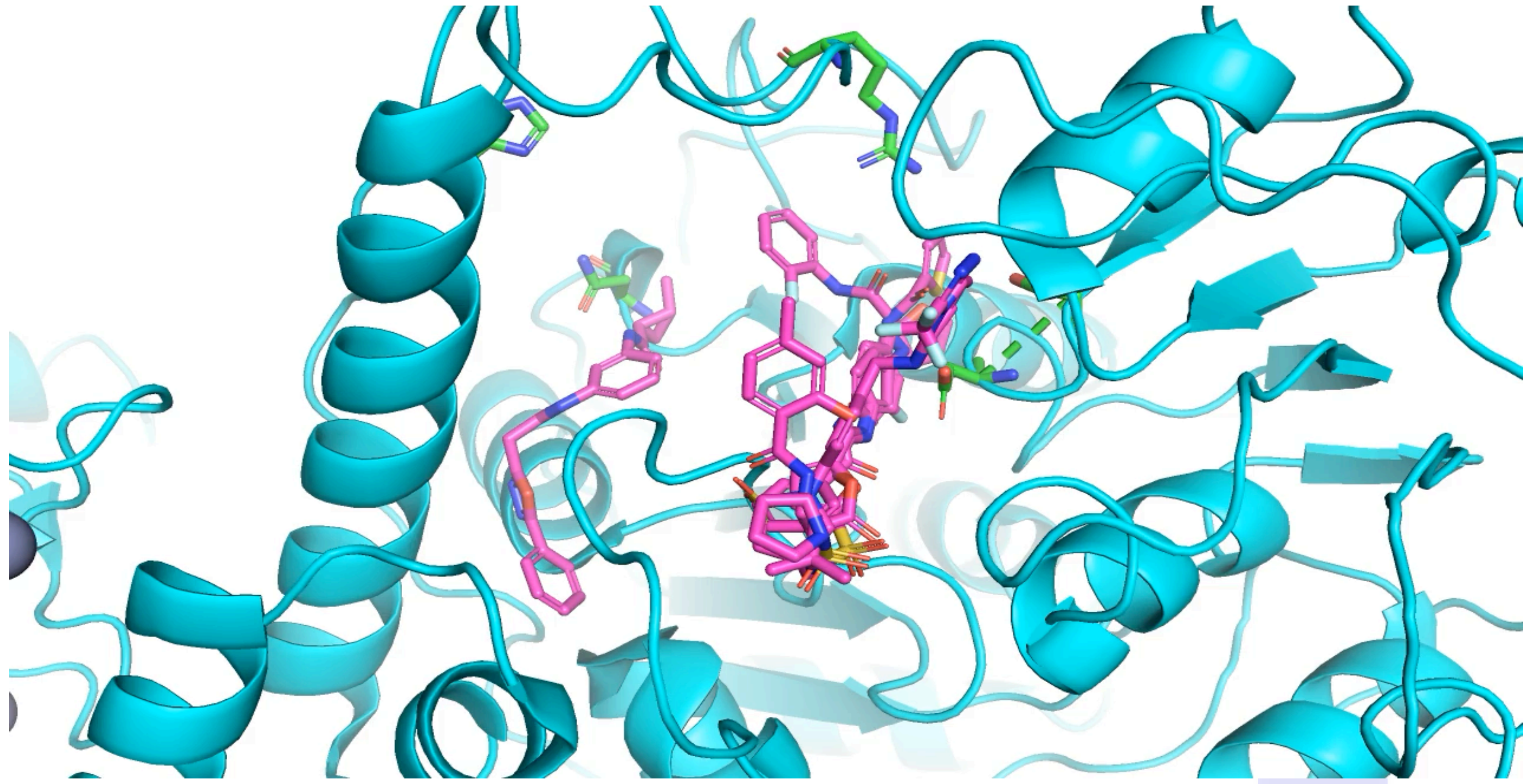
CACHE #2 Results

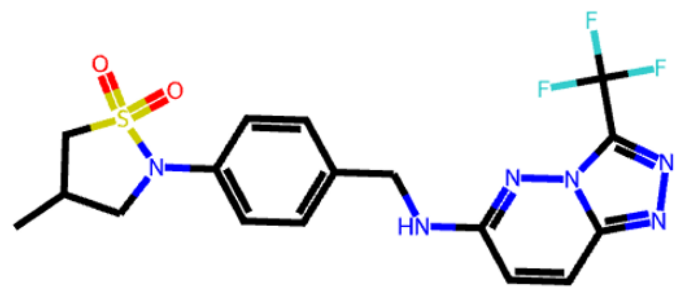
5/50 compounds identified as potential hits

>2x the average hit rate

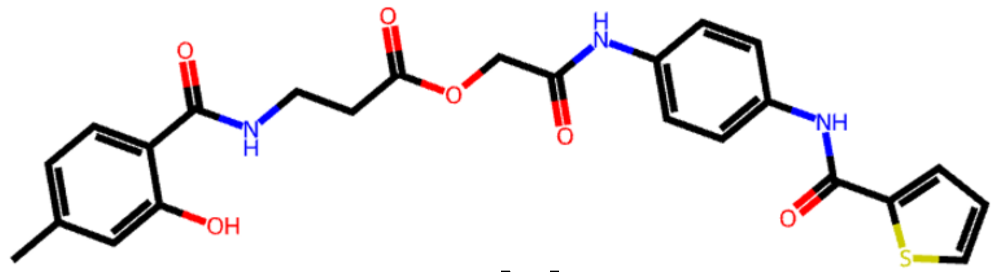
4/5 hits from last round of active learning

Highest affinity round 1 hit in the competition

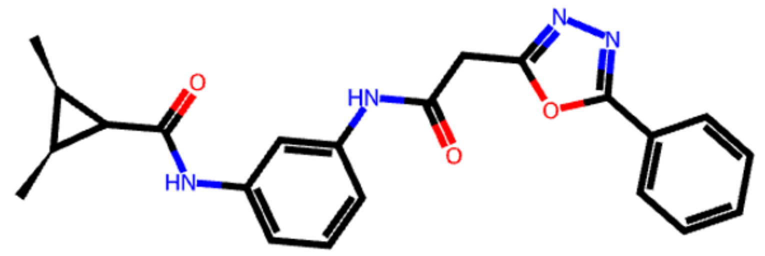




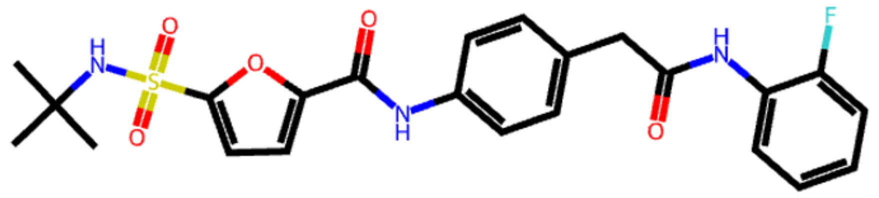
43μM



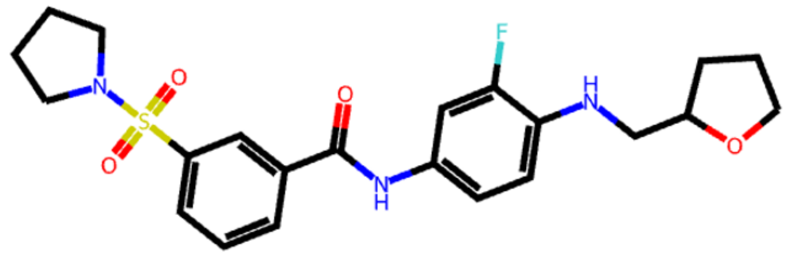
7μM



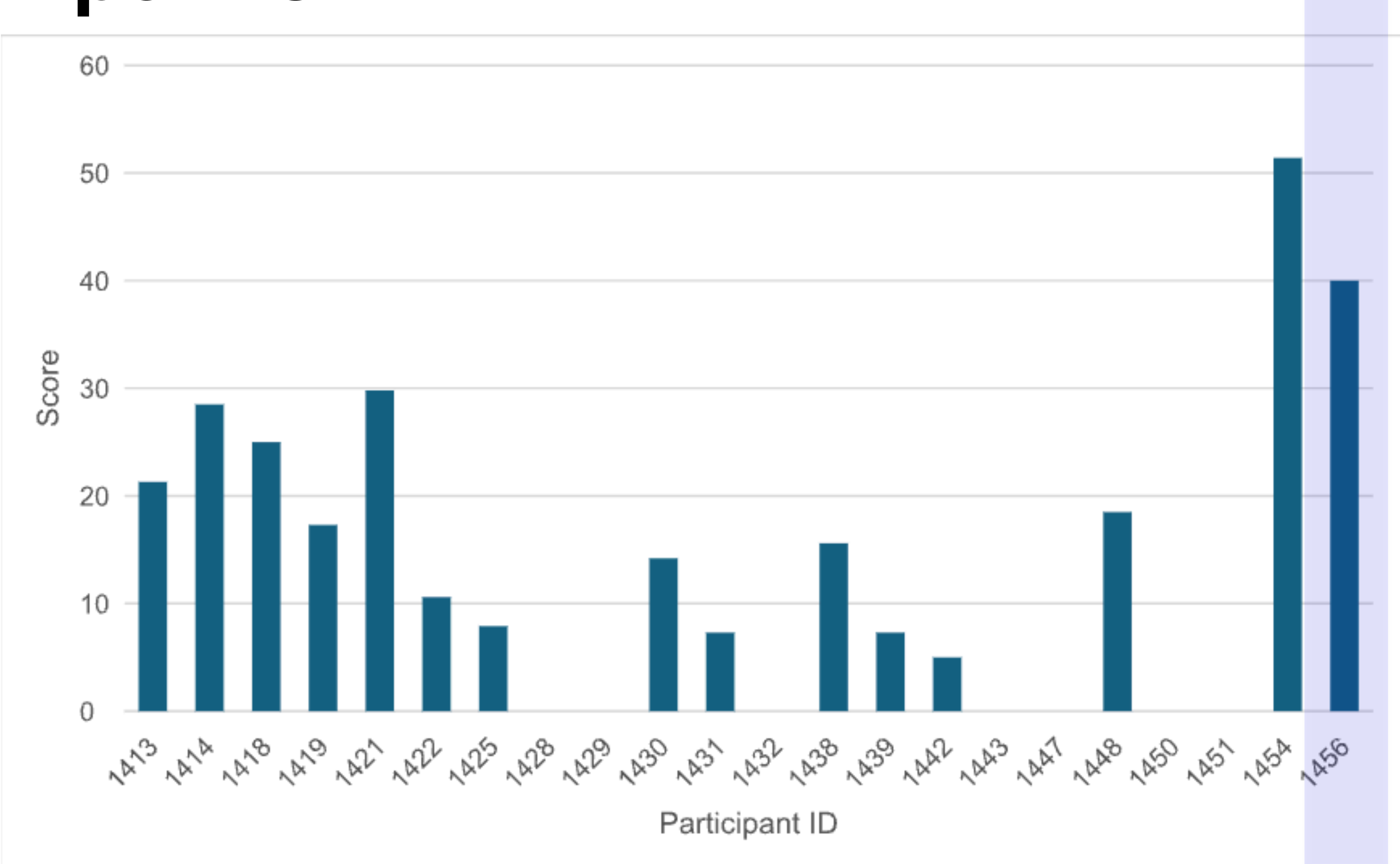
5μM



45μM



96μM



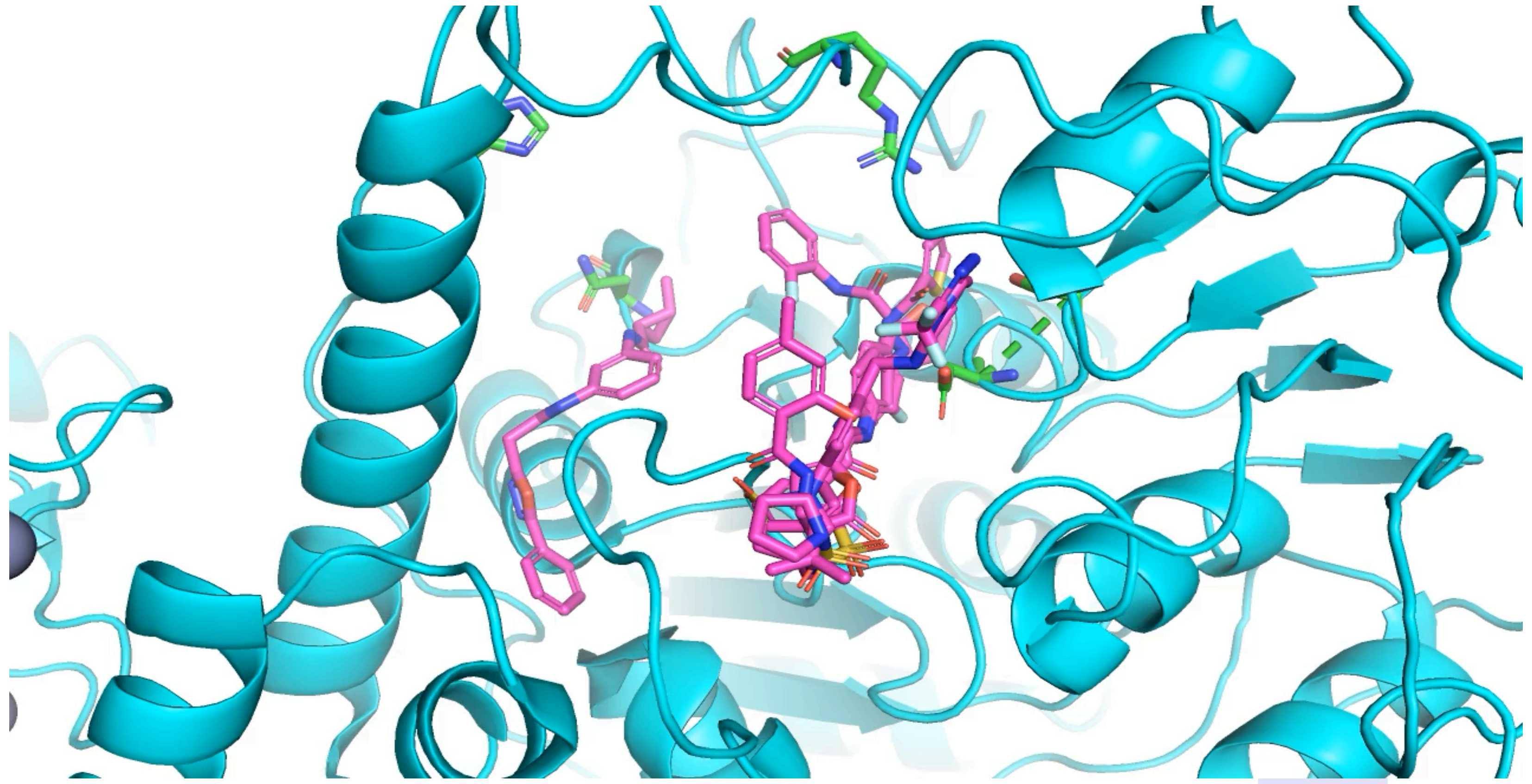
CACHE #2 Results

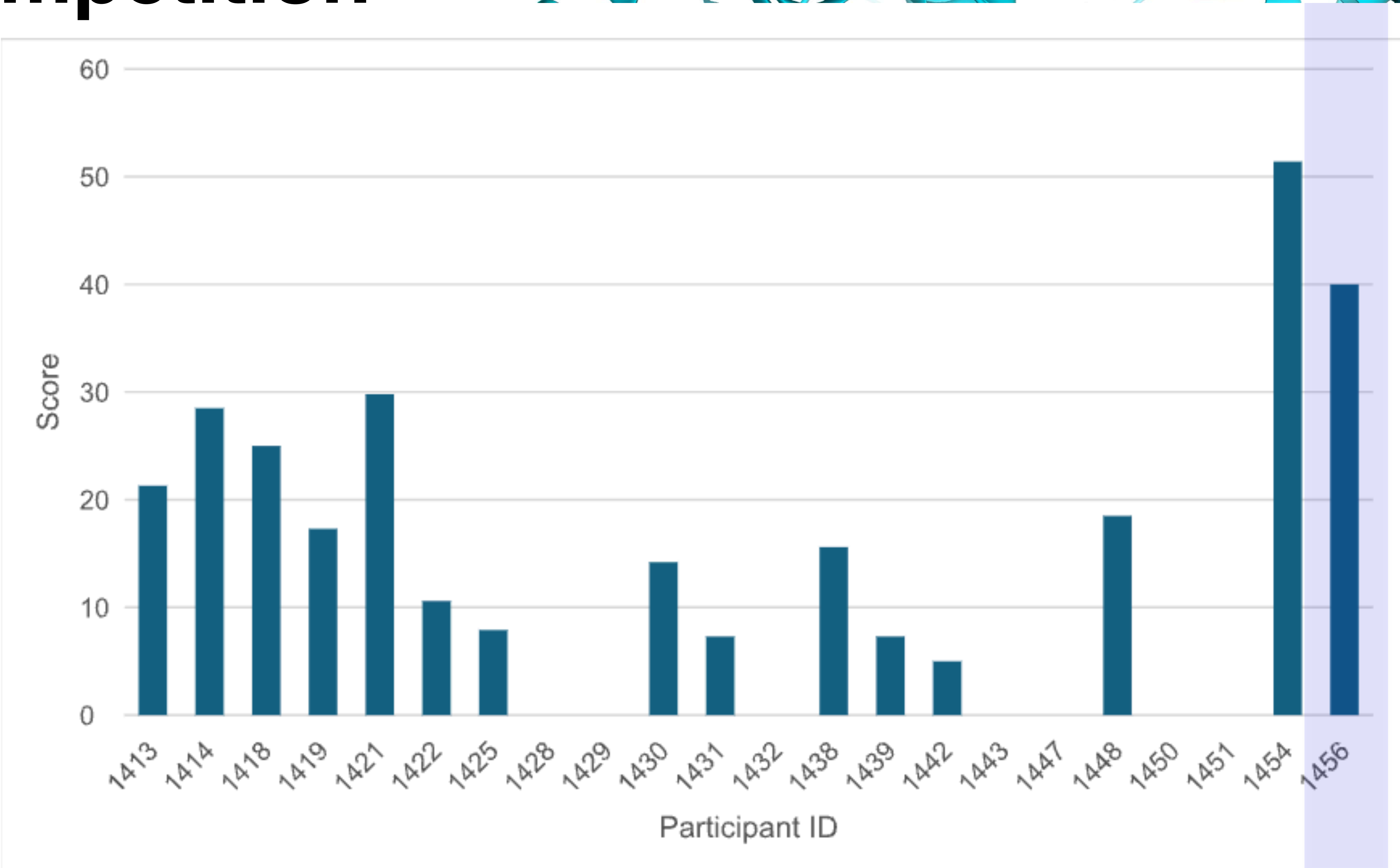
5/50 compounds identified as potential hits

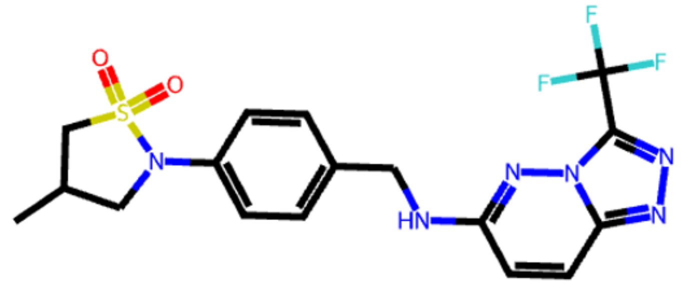
>2x the average hit rate

4/5 hits from last round of active learning

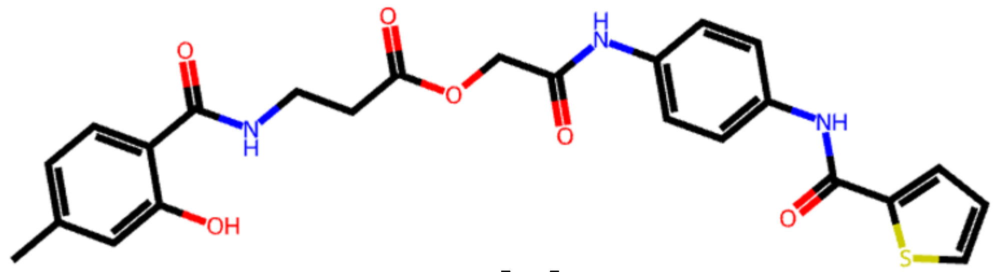
Highest affinity round 1 hit in the competition



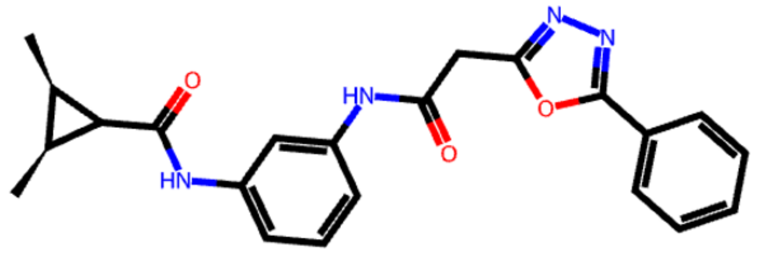




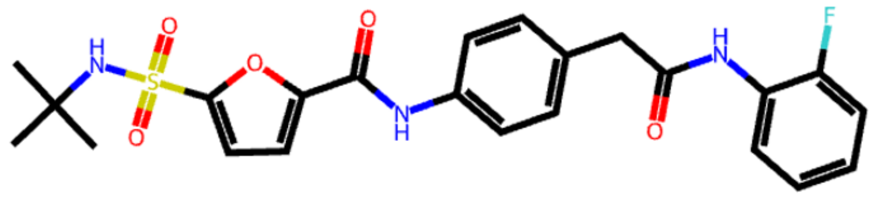
43µM



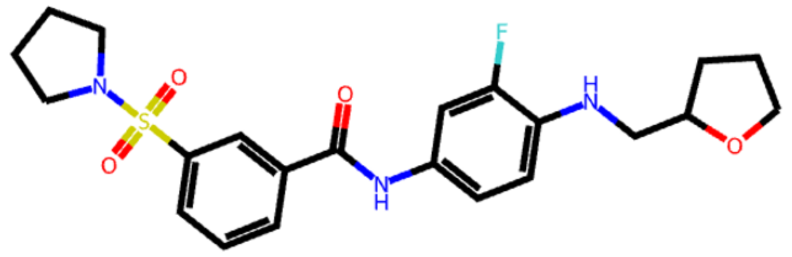
7µM



5µM



45µM



96µM

Key Points

- Ligand-Based vs Receptor Based
- Fingerprints
- Pharmacophores
- Docking
- Rise of the Machines!