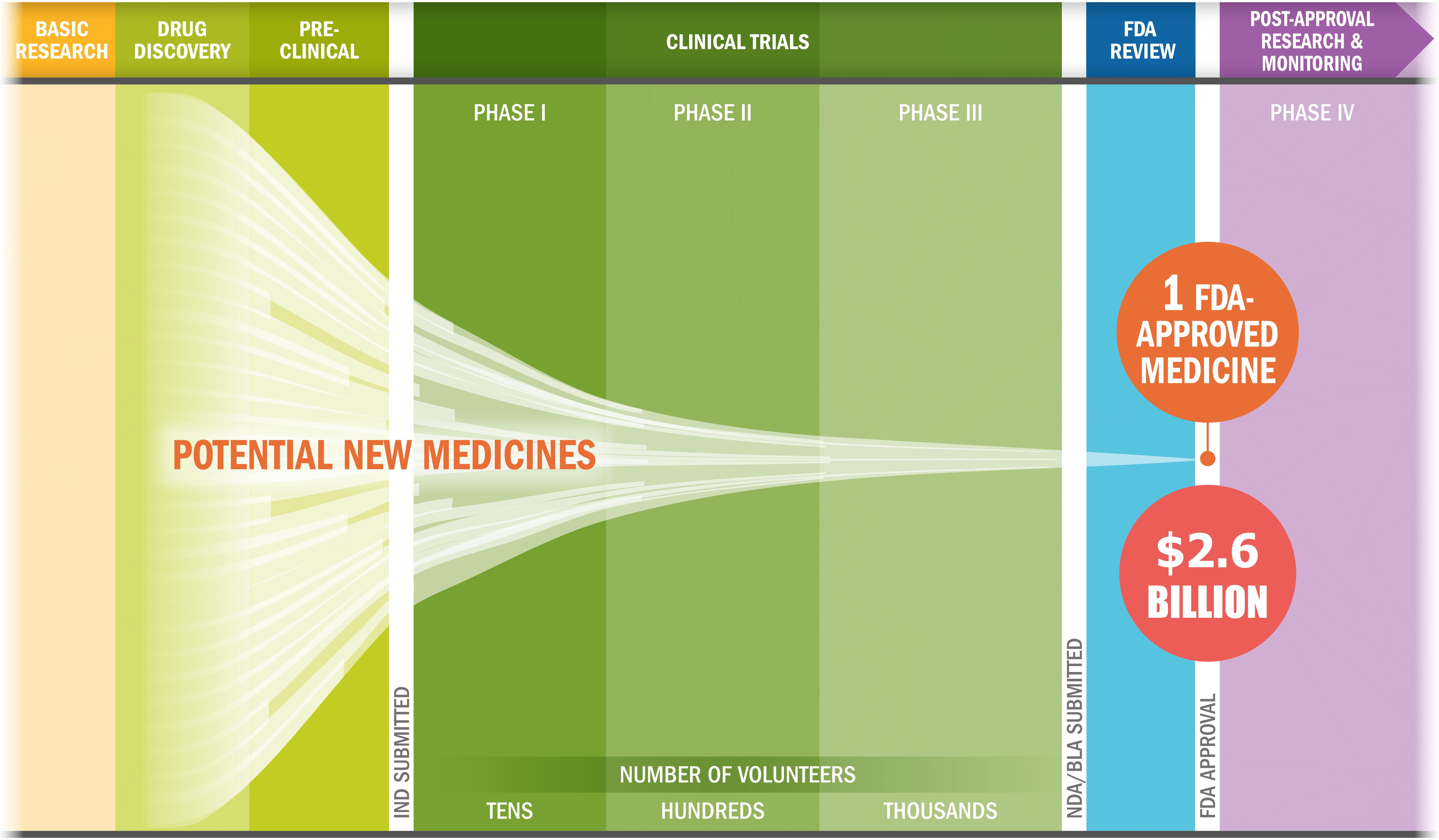
Deep Learning for Molecular Docking

David Koes

@david_koes

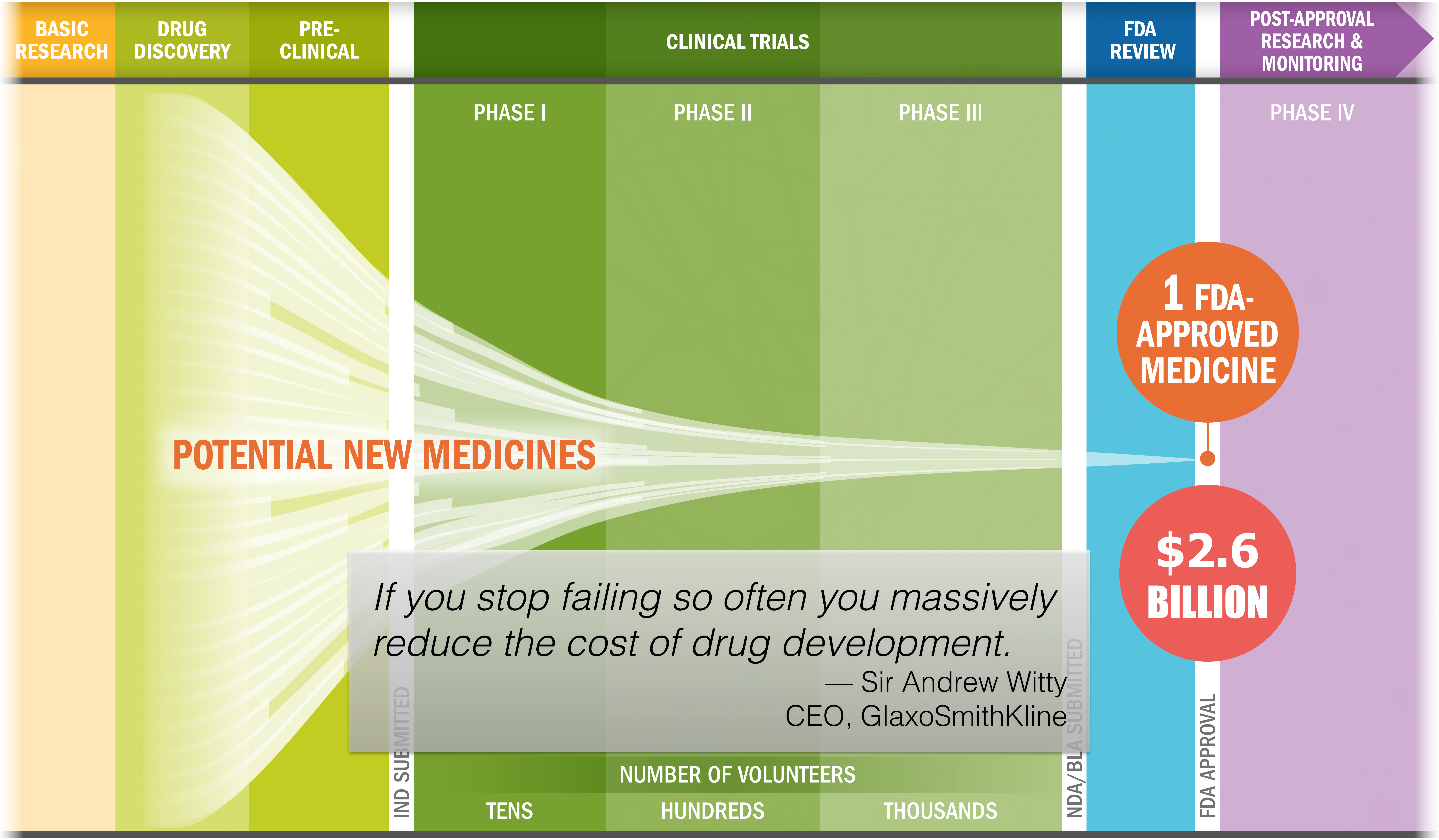
GPU Technology Conference
San Jose, CA
March 26, 2018

THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS



Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)

THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

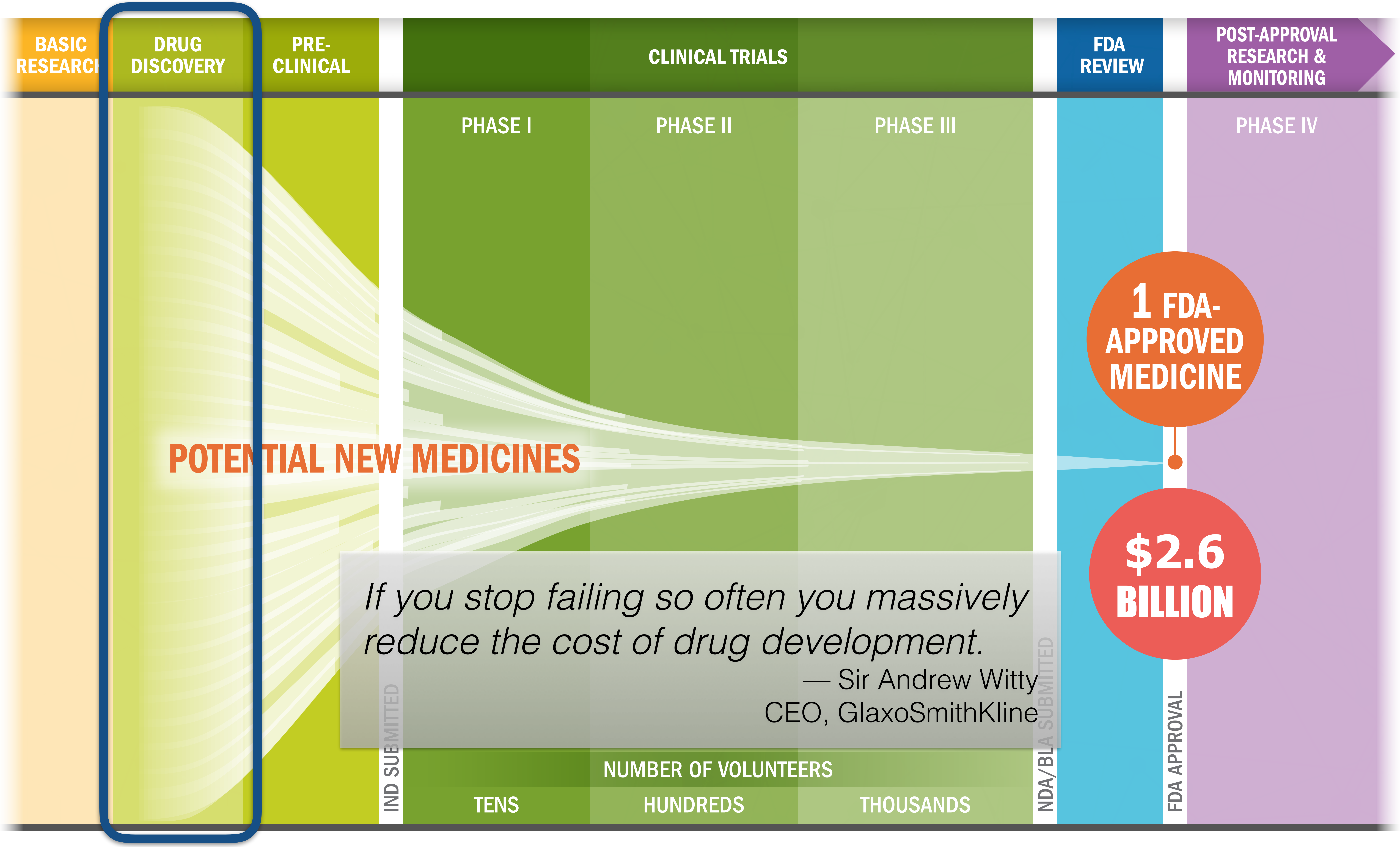


If you stop failing so often you massively reduce the cost of drug development.

— Sir Andrew Witty
CEO, GlaxoSmithKline

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)

THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

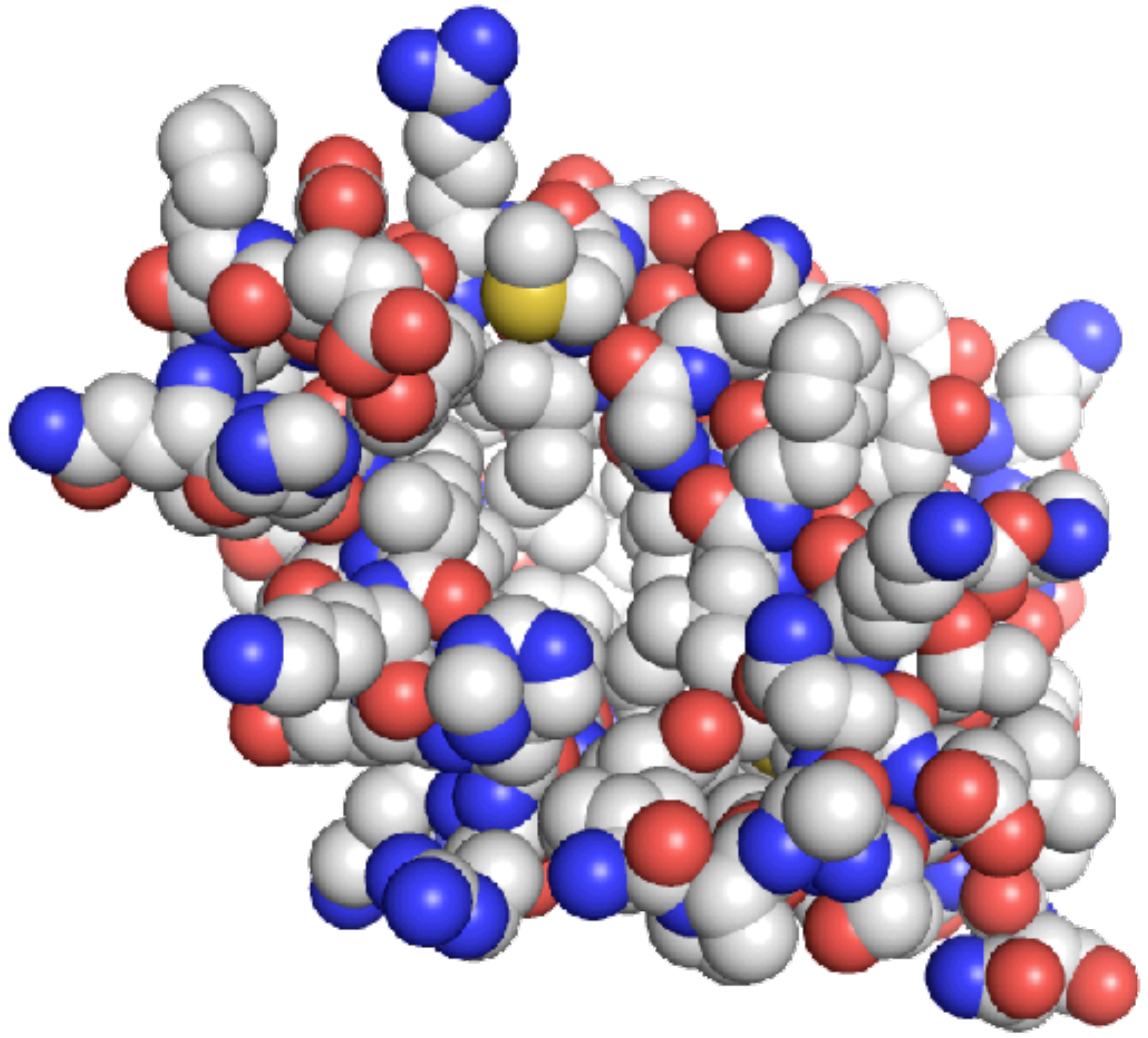


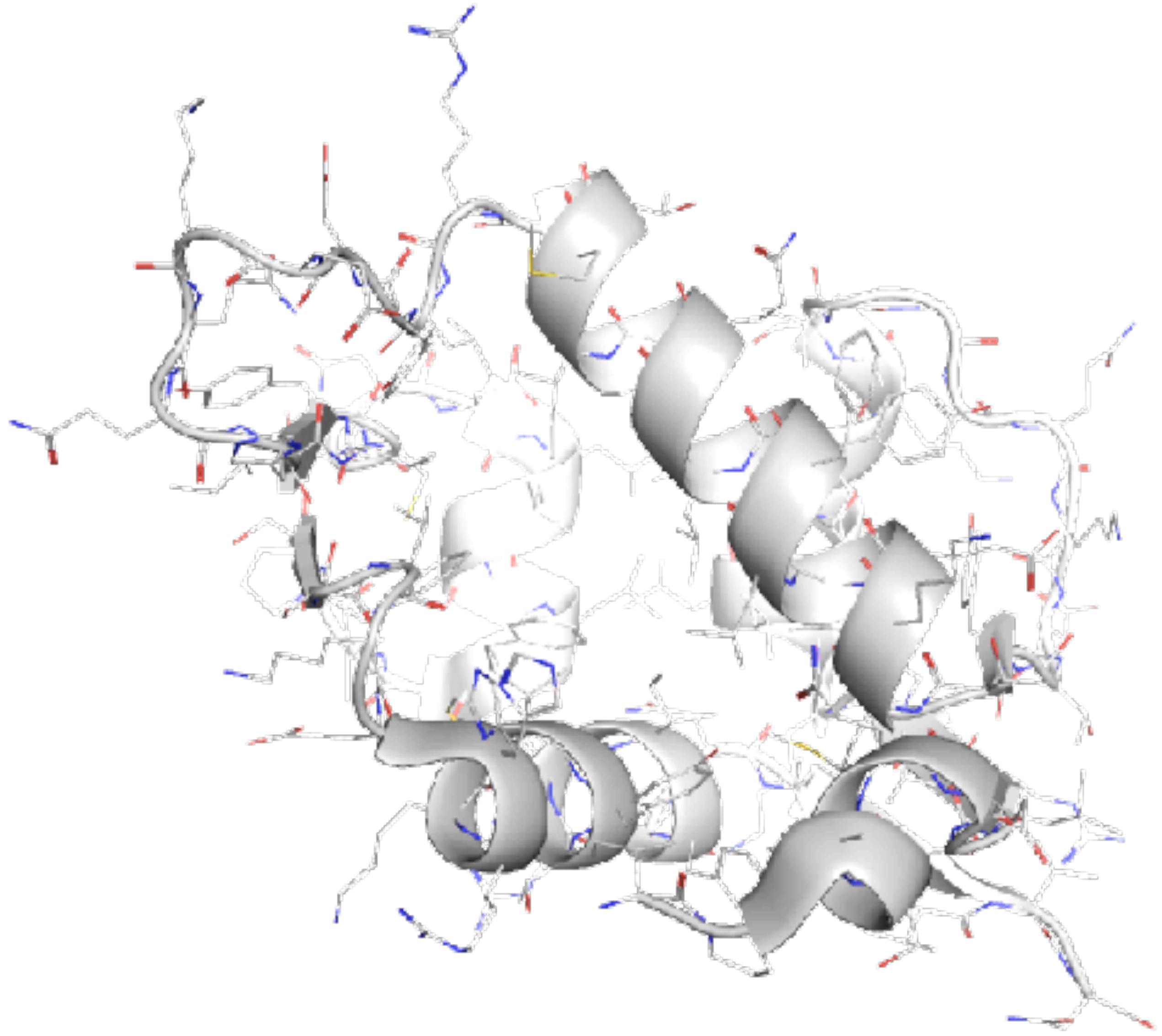
Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)

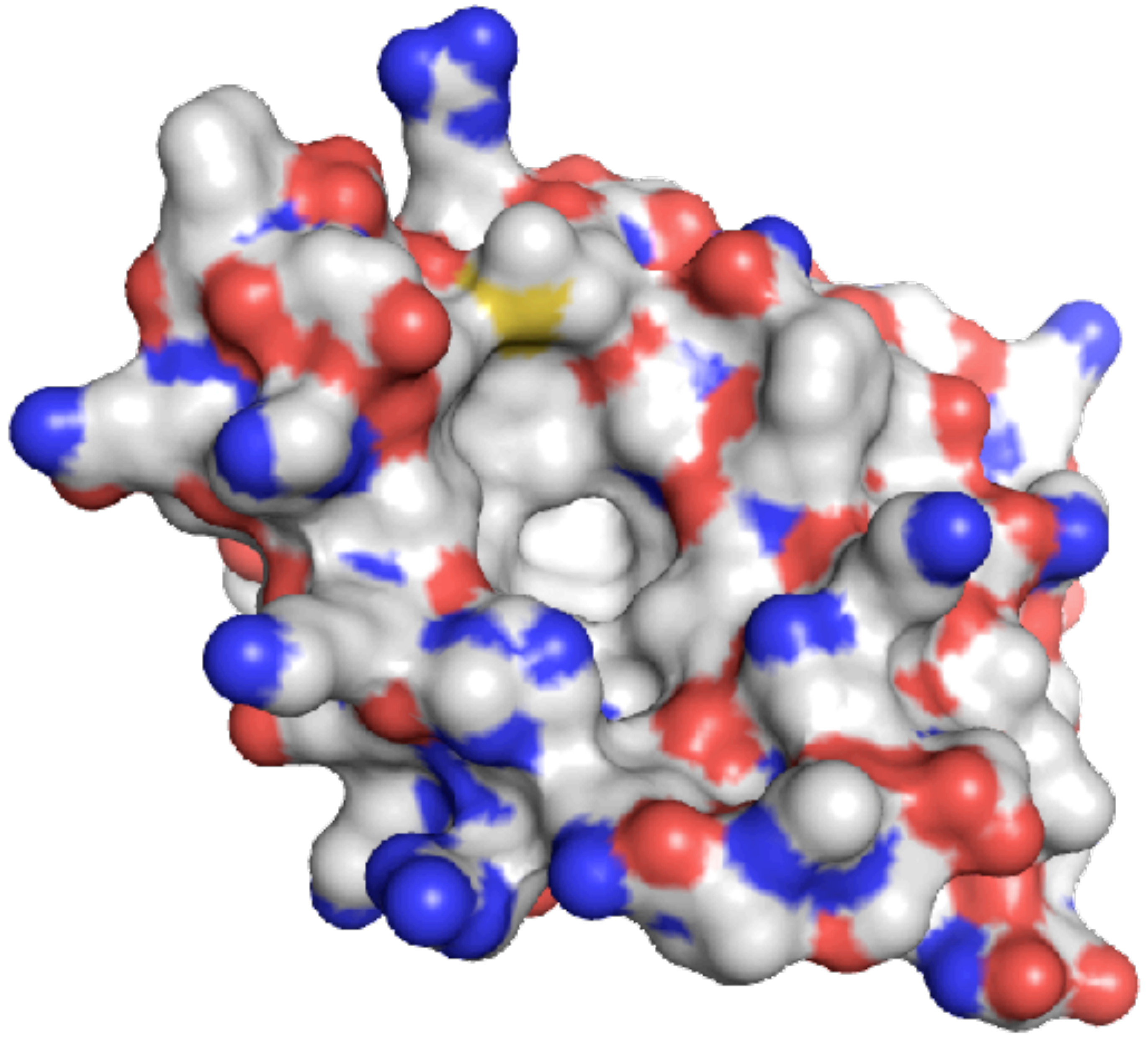
1. Does the compound do what you want it to?
2. Does the compound **not** do what you **don’t** want it to?
3. Is what you want it to do the right thing?

Protein Structures

sequence → **structure** → function

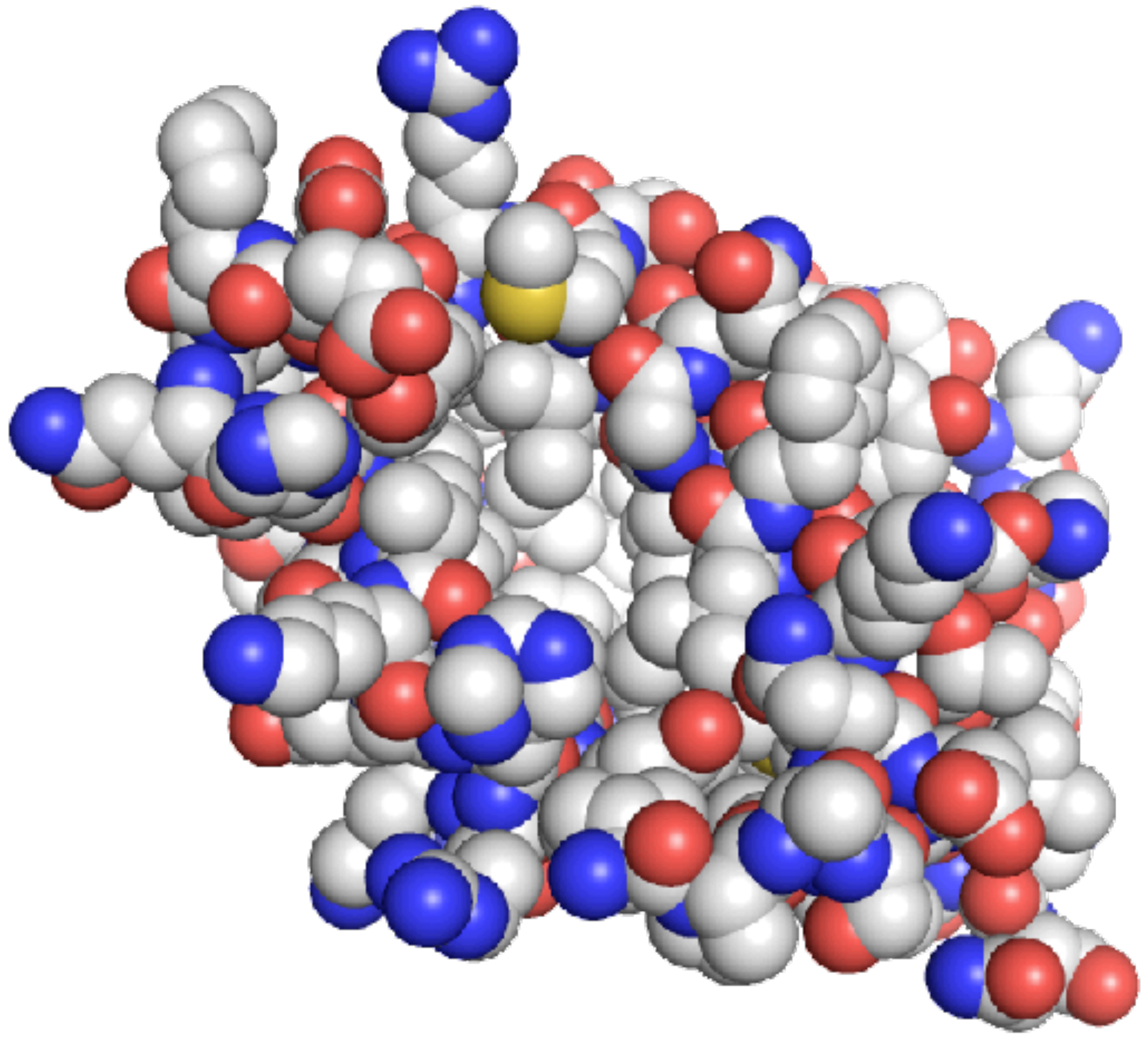


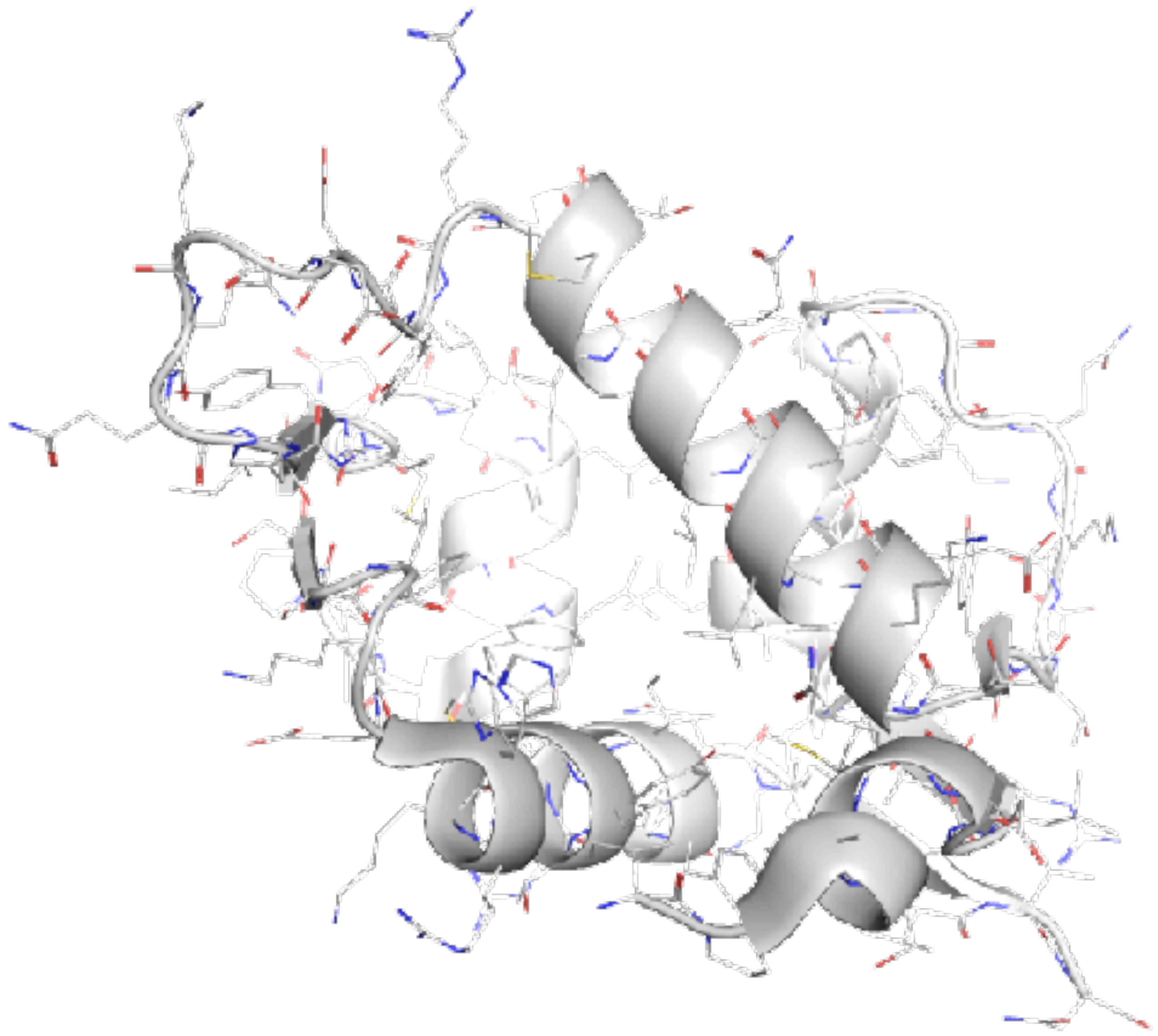


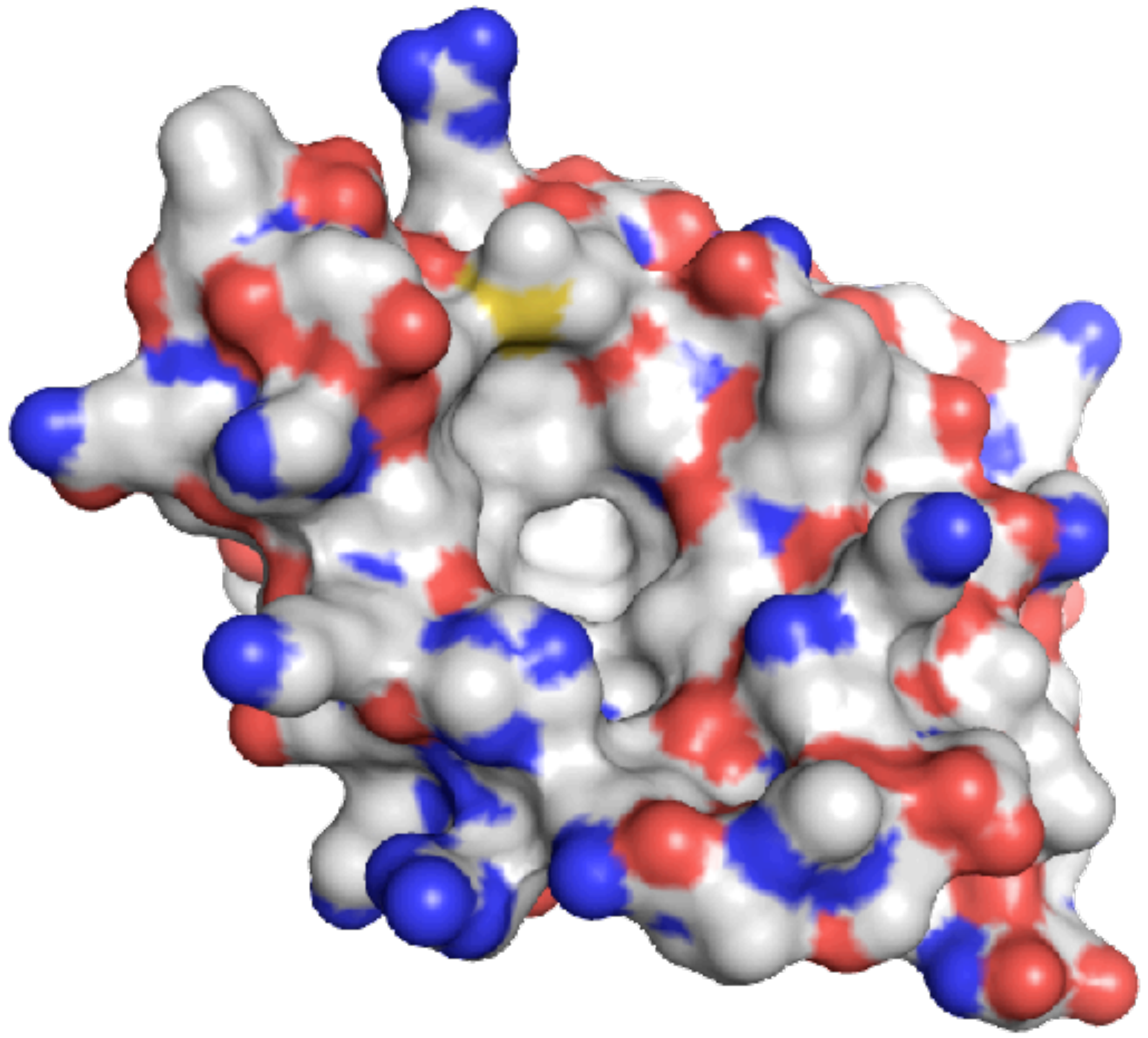


Protein Structures

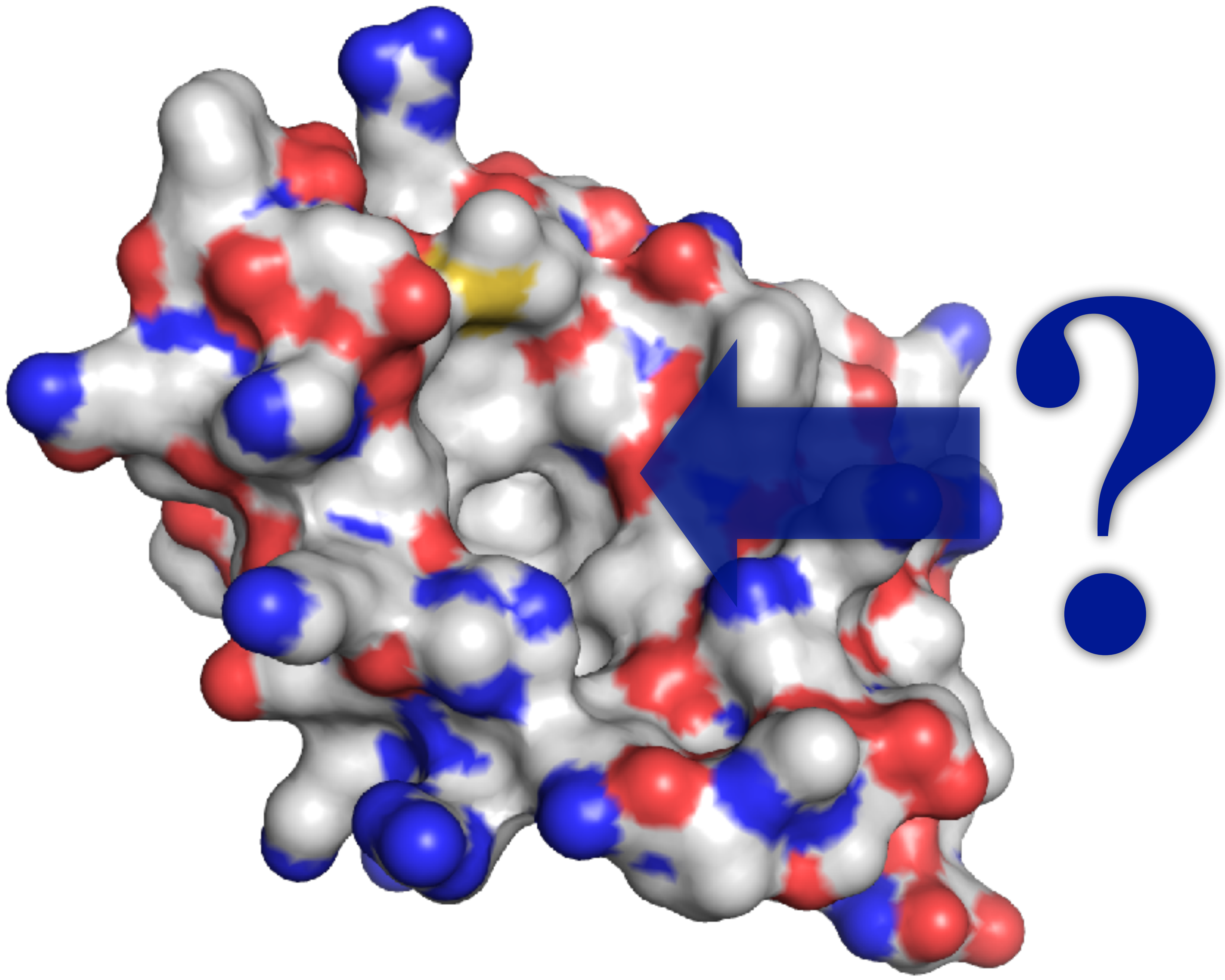
sequence → **structure** → function







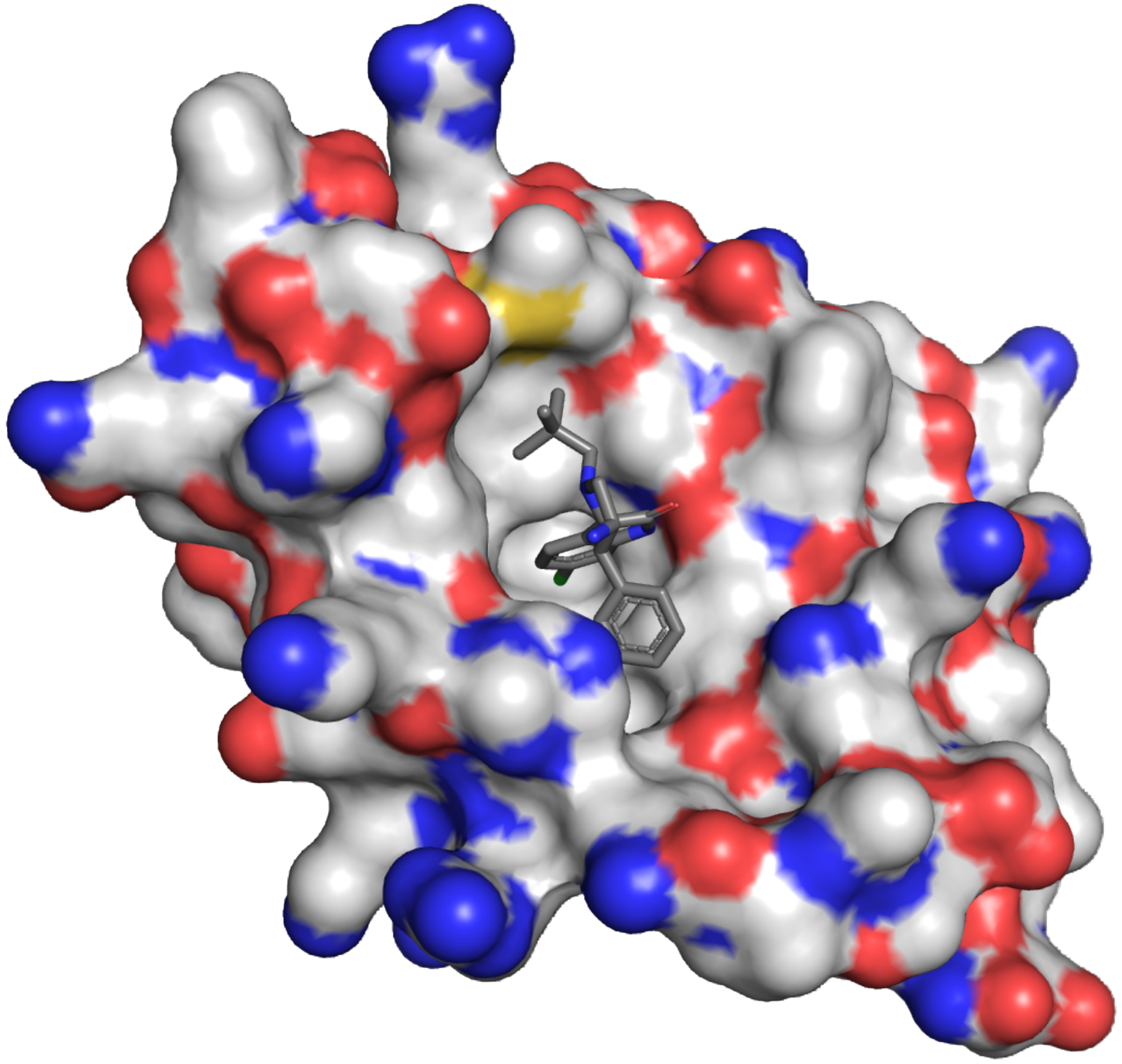
Structure Based Drug Design



Unlike ligand based approaches, **generalizes to new targets**

Requires **molecular target** with **known structure** and **binding site**

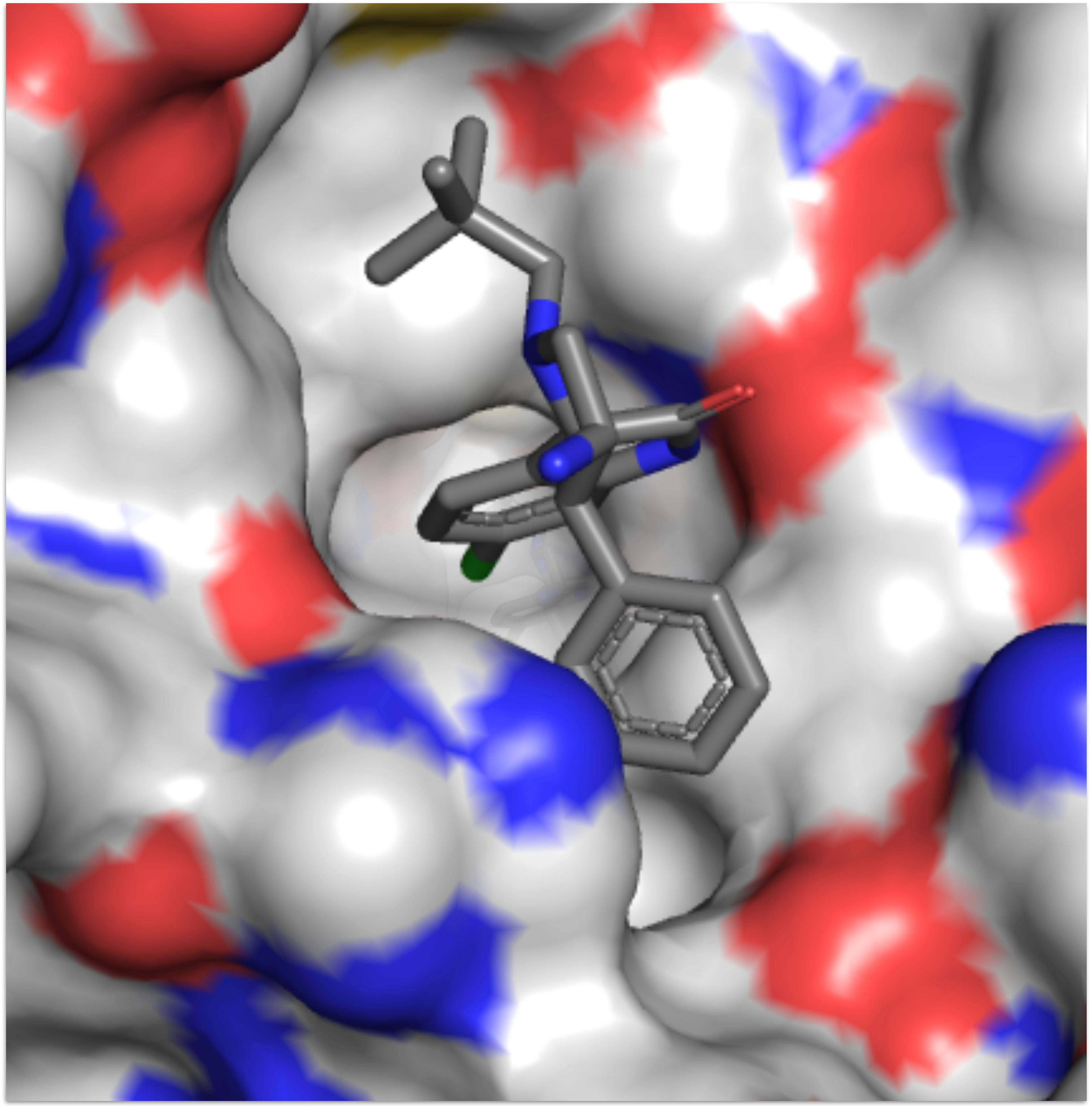
Structure Based Drug Design



Unlike ligand based approaches, **generalizes to new targets**

Requires **molecular target** with **known structure** and **binding site**

Structure Based Drug Design

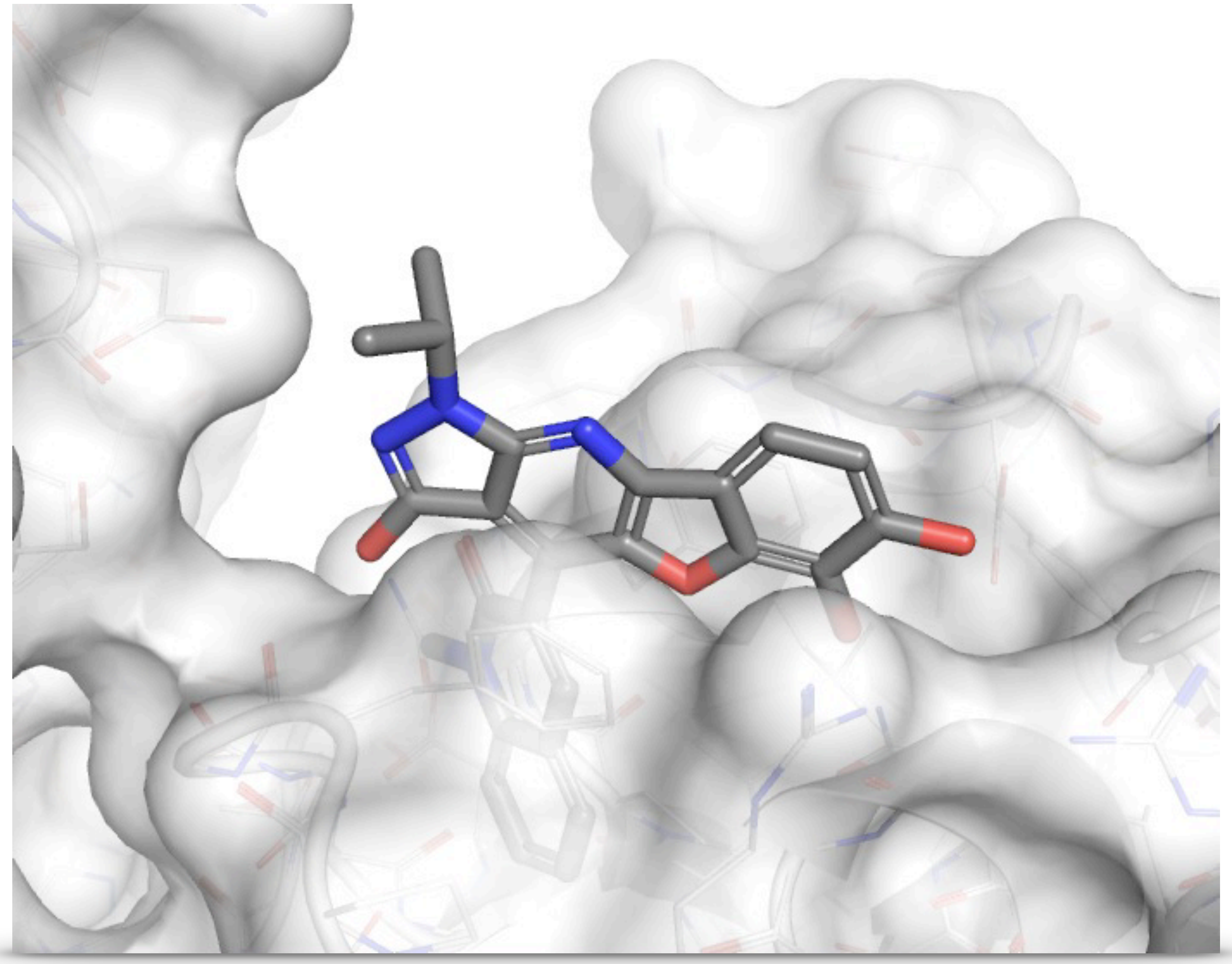


Unlike ligand based approaches, **generalizes to new targets**

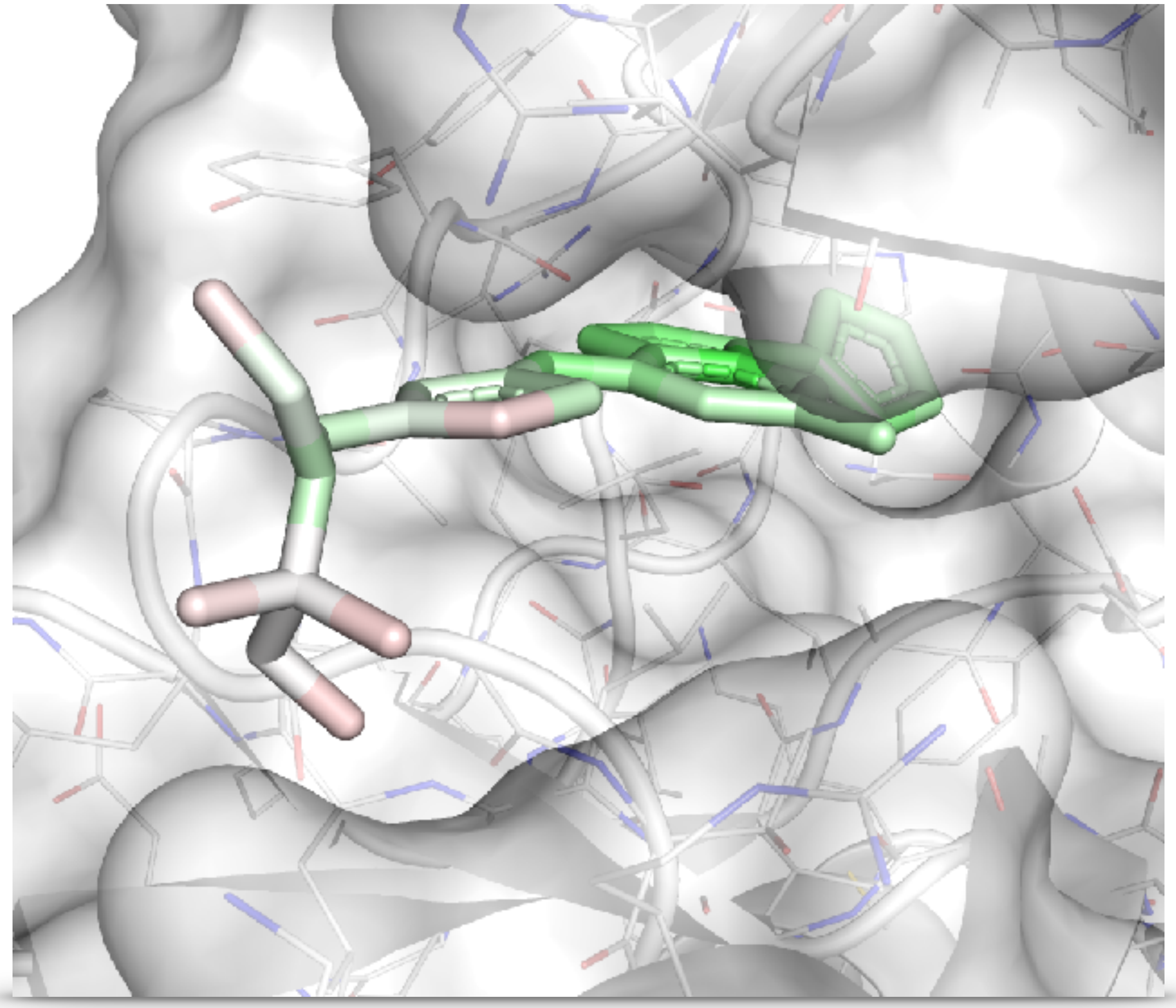
Requires **molecular target** with **known structure** and **binding site**

Structure Based Drug Design

Virtual Screening



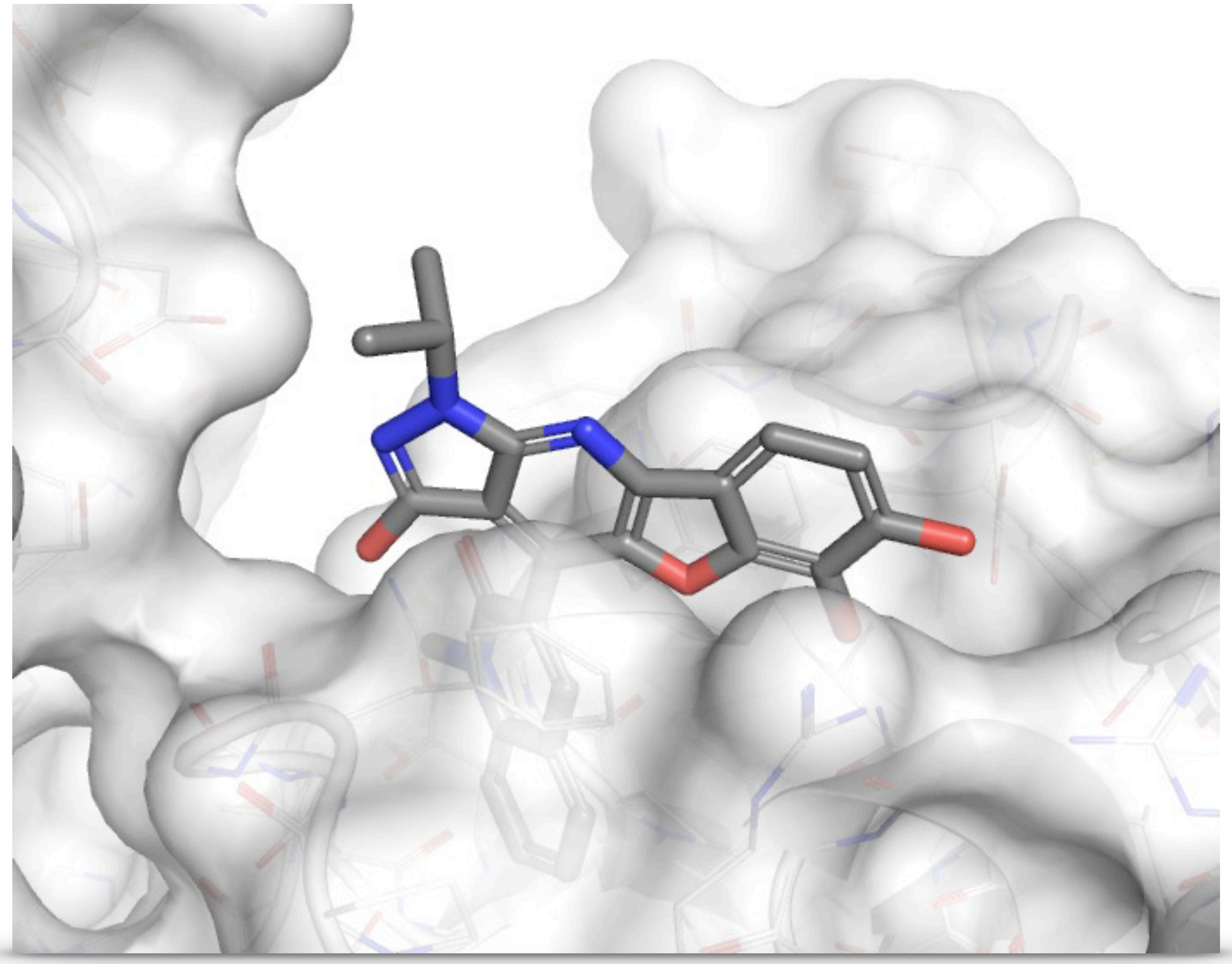
Lead Optimization



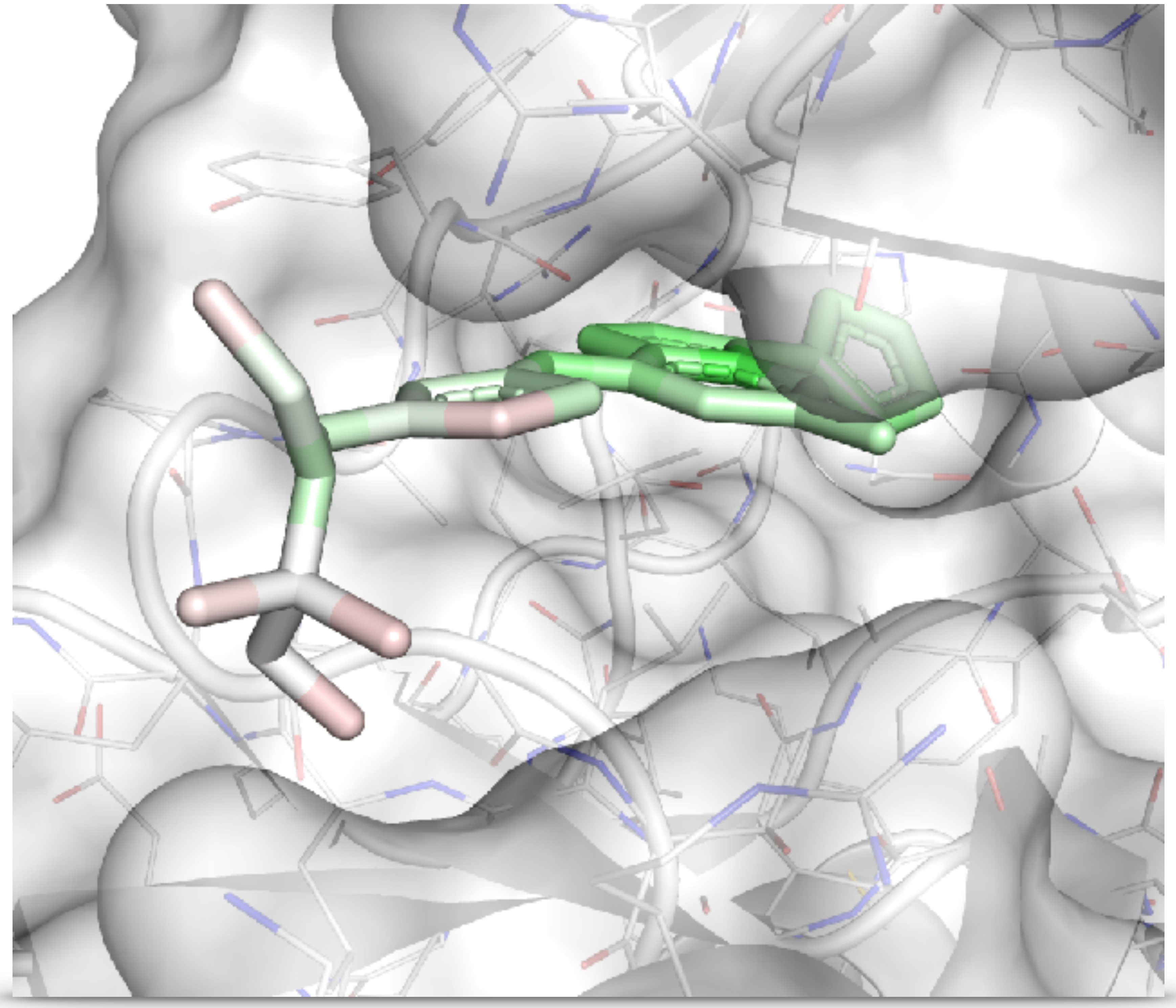
Pose Prediction → Binding Discrimination → Affinity Prediction

Structure Based Drug Design

Virtual Screening

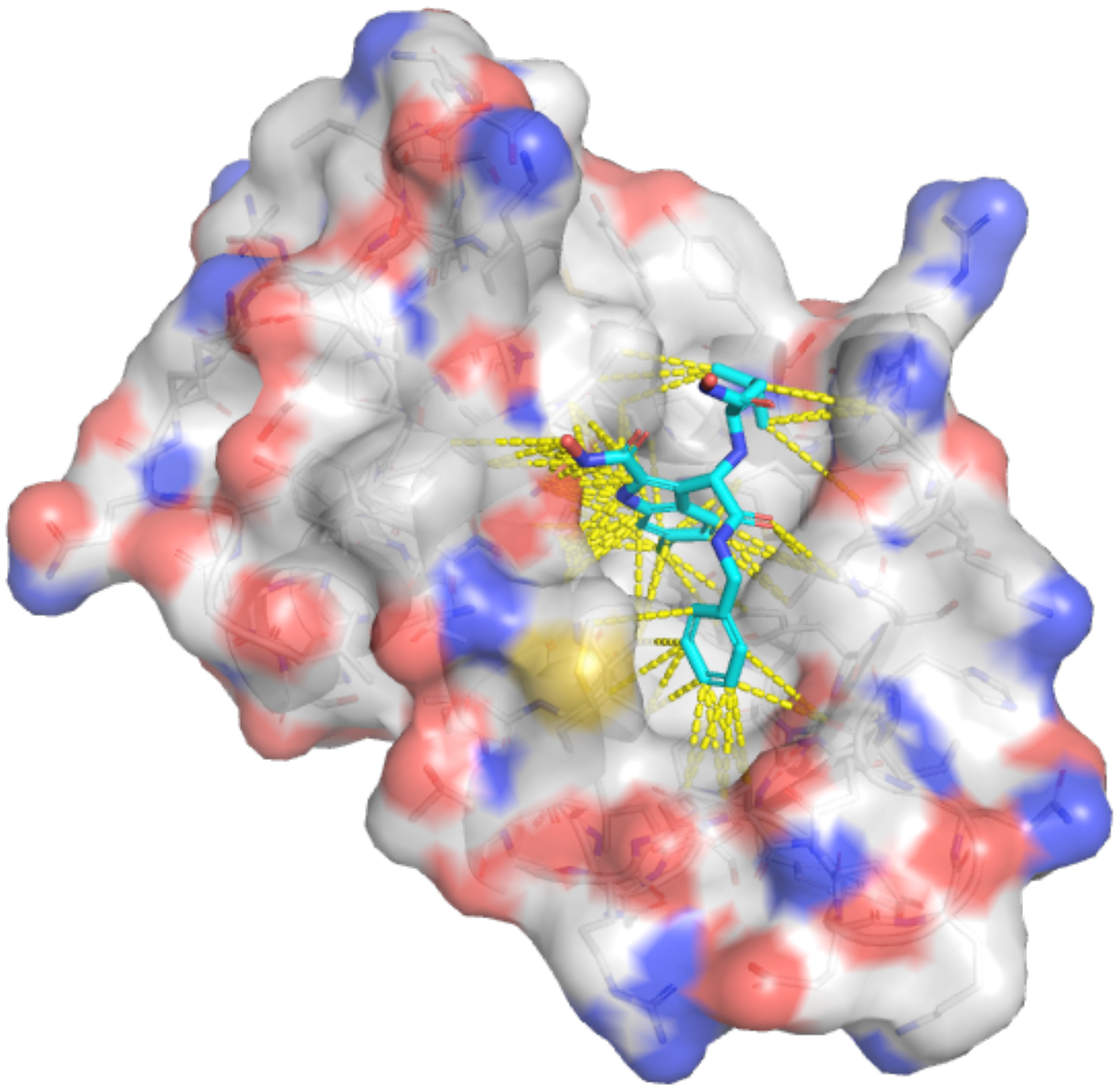


Lead Optimization



Pose Prediction → Binding Discrimination → Affinity Prediction

Protein-Ligand Scoring

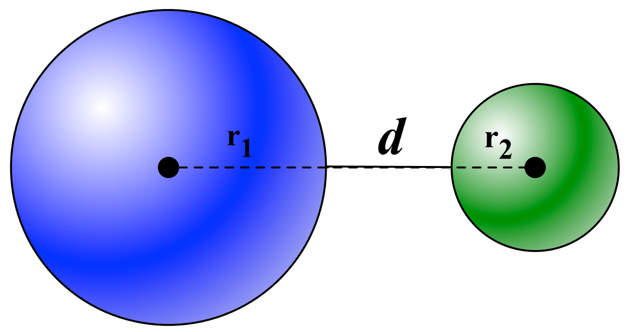


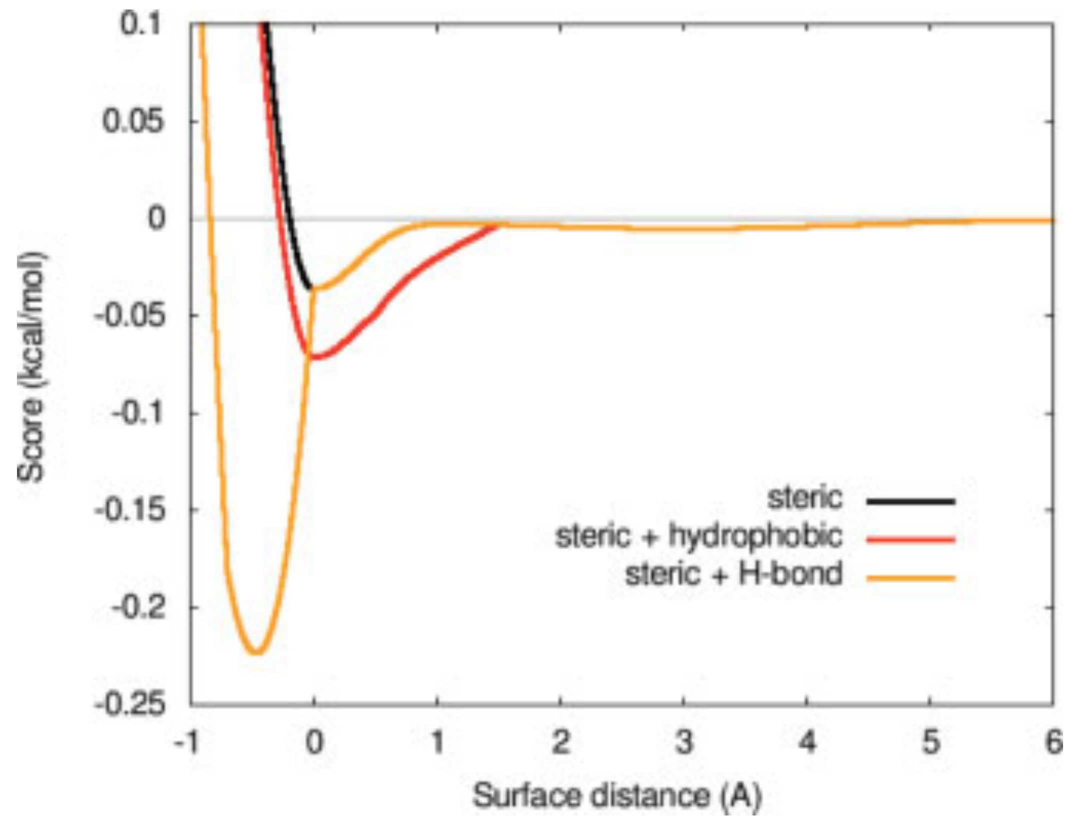
AutoDock Vina

$$\begin{aligned}\mathrm{gauss}_1(d) &= w_{\mathrm{guass}_1} e^{-(d/0.5)^2}\\ \mathrm{gauss}_2(d) &= w_{\mathrm{guass}_2} e^{-((d-3)/2)^2}\\ \mathrm{repulsion}(d) &= \begin{cases} w_{\mathrm{repulsion}} d^2 & d < 0 \\ 0 & d \geq 0 \end{cases}\end{aligned}$$

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

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





O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry* 31 (2010) 455-461

Can we do better?

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

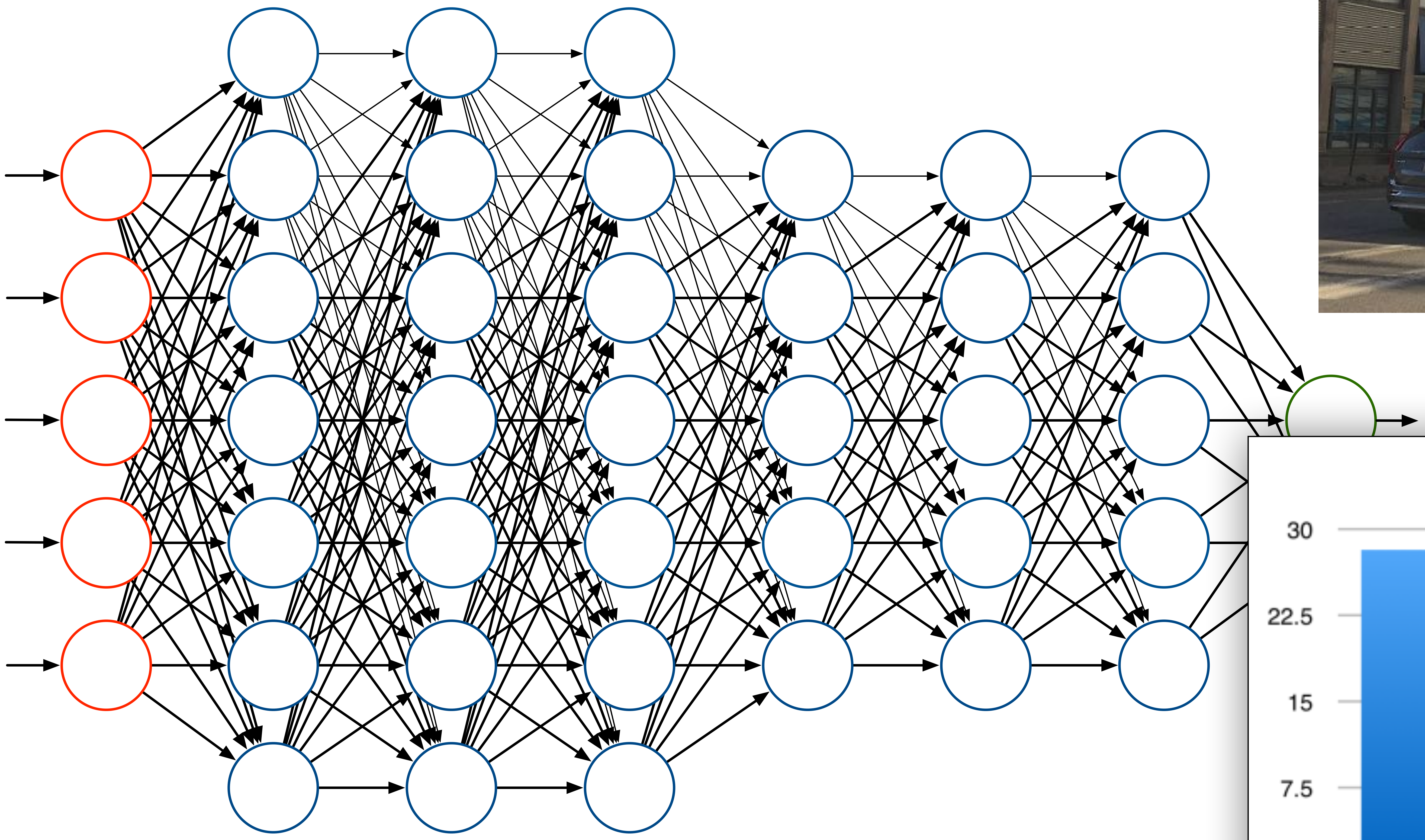
Can we do better?

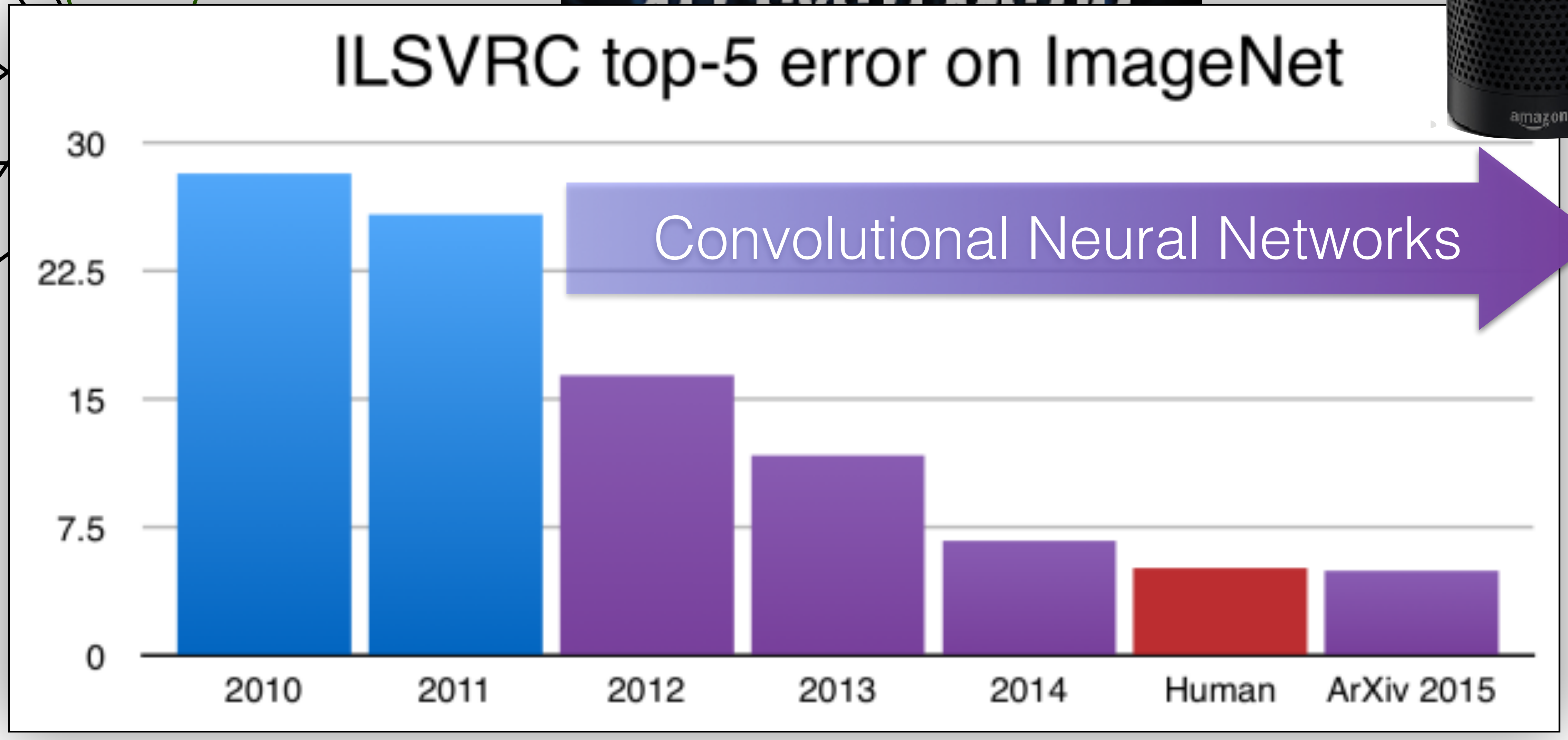
Accurate pose prediction, binding discrimination, **and** affinity prediction without sacrificing performance?

Key Idea: Leverage "big data"

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

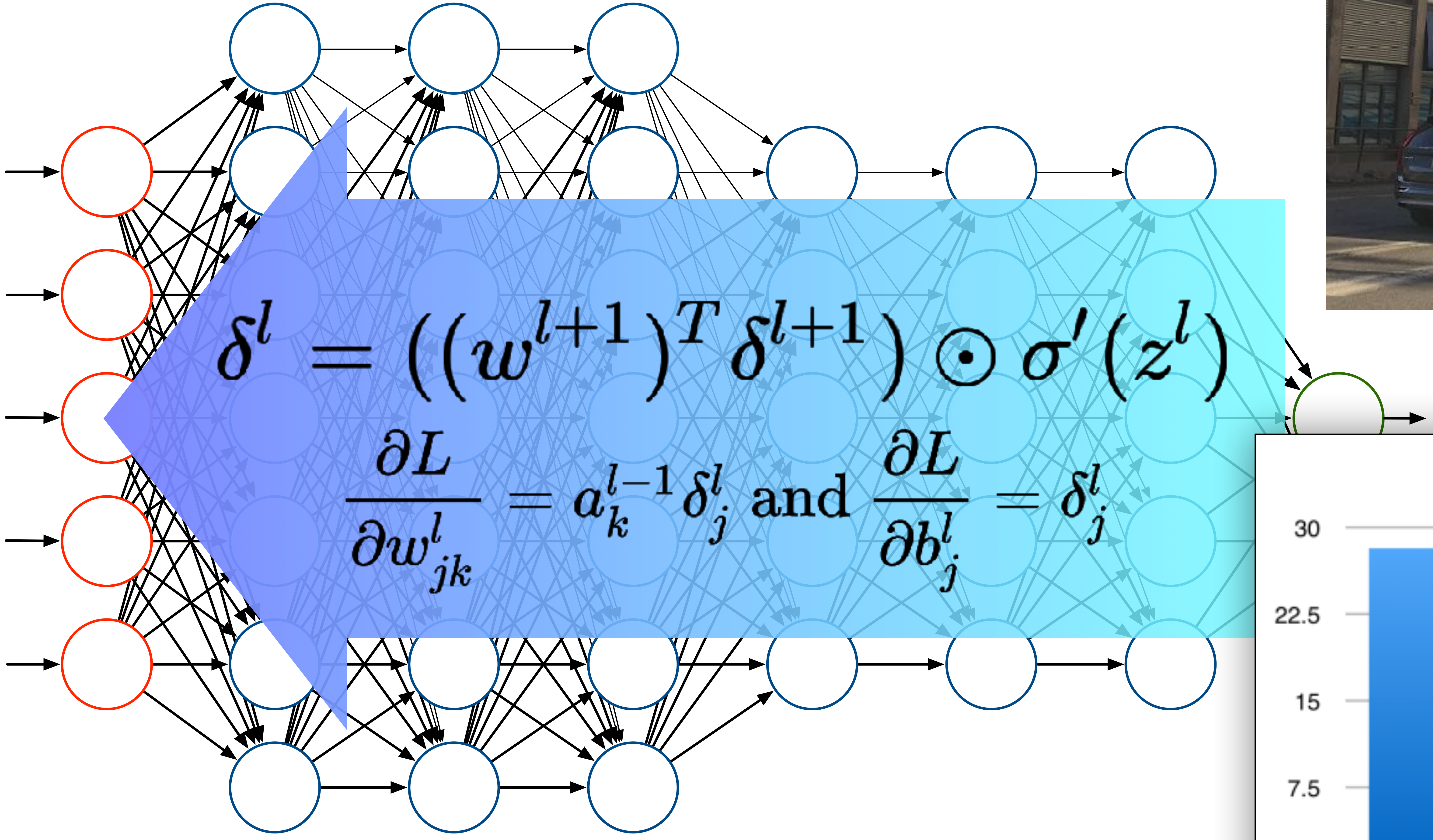
Deep Learning





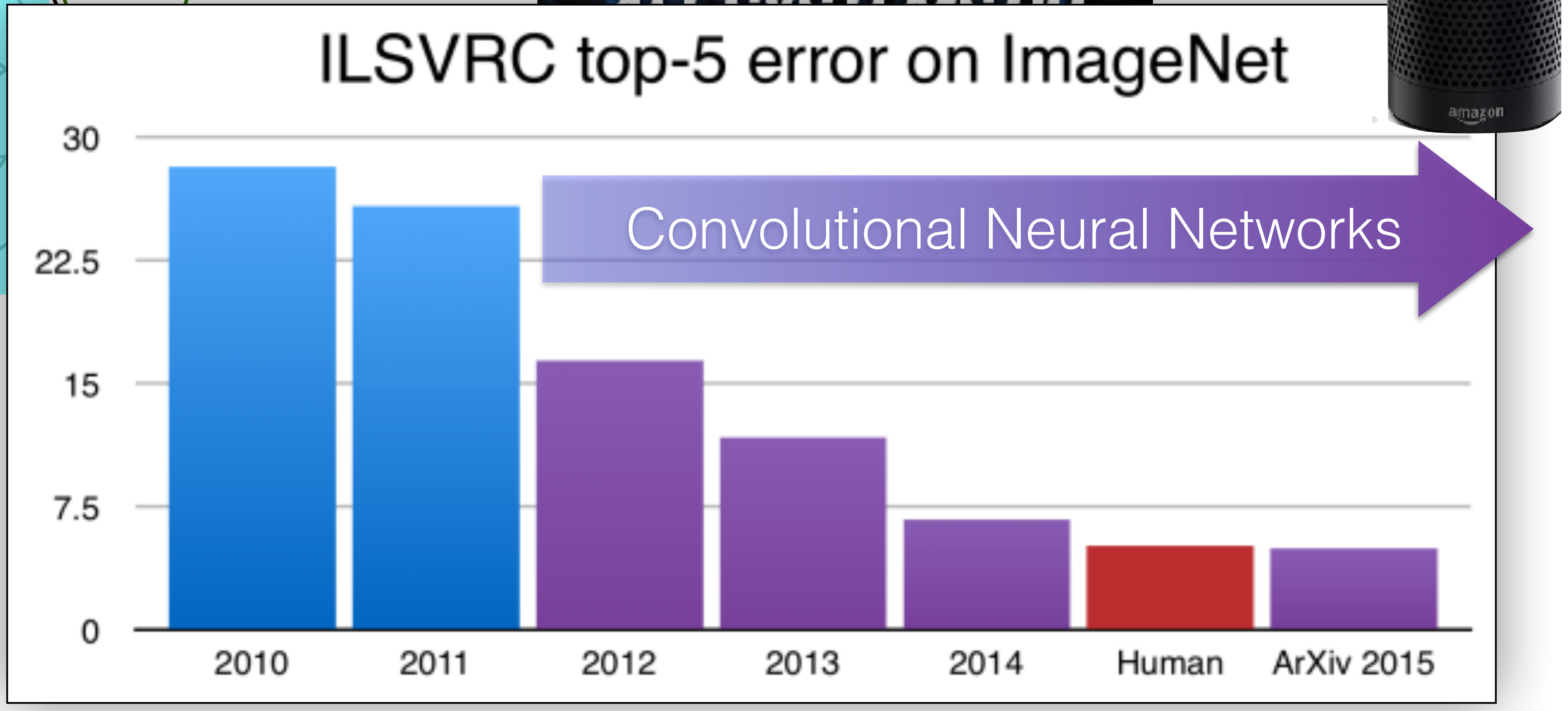
https://devblogs.nvidia.com

Deep Learning



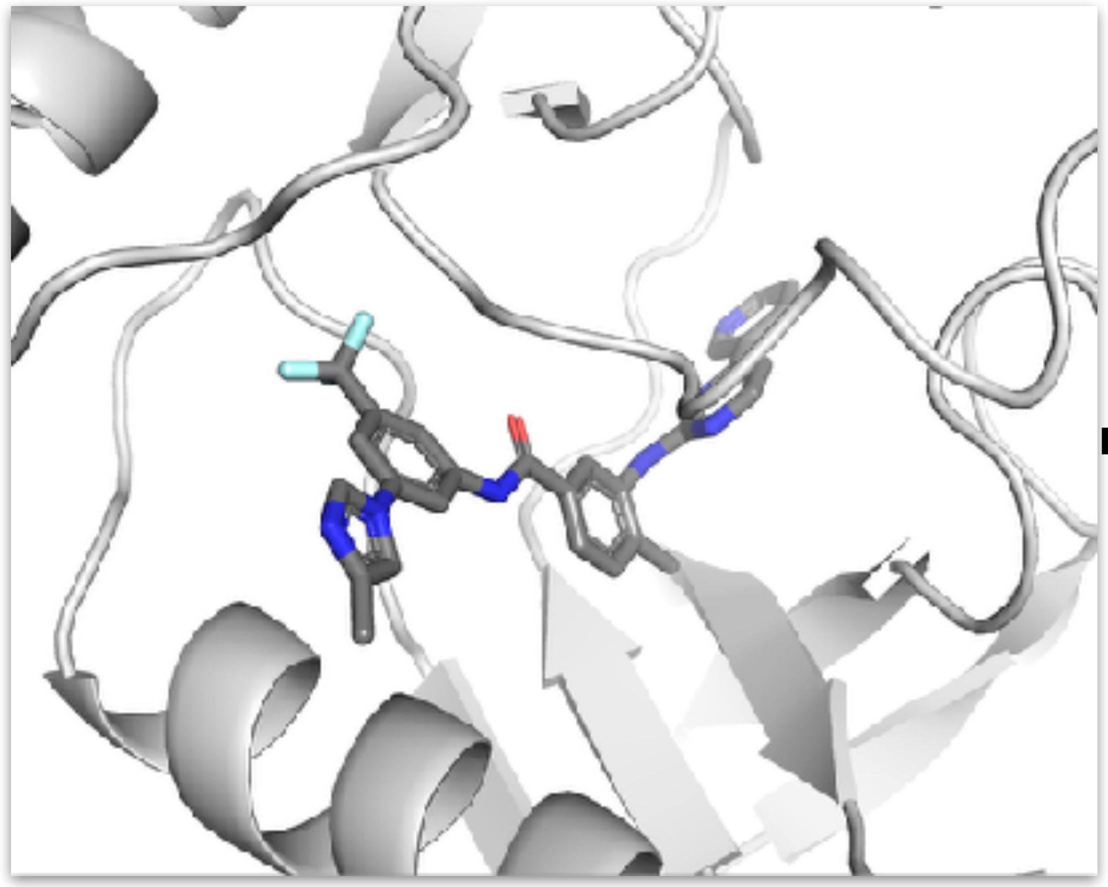
$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

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



https://devblogs.nvidia.com

CNNs for Protein-Ligand Scoring



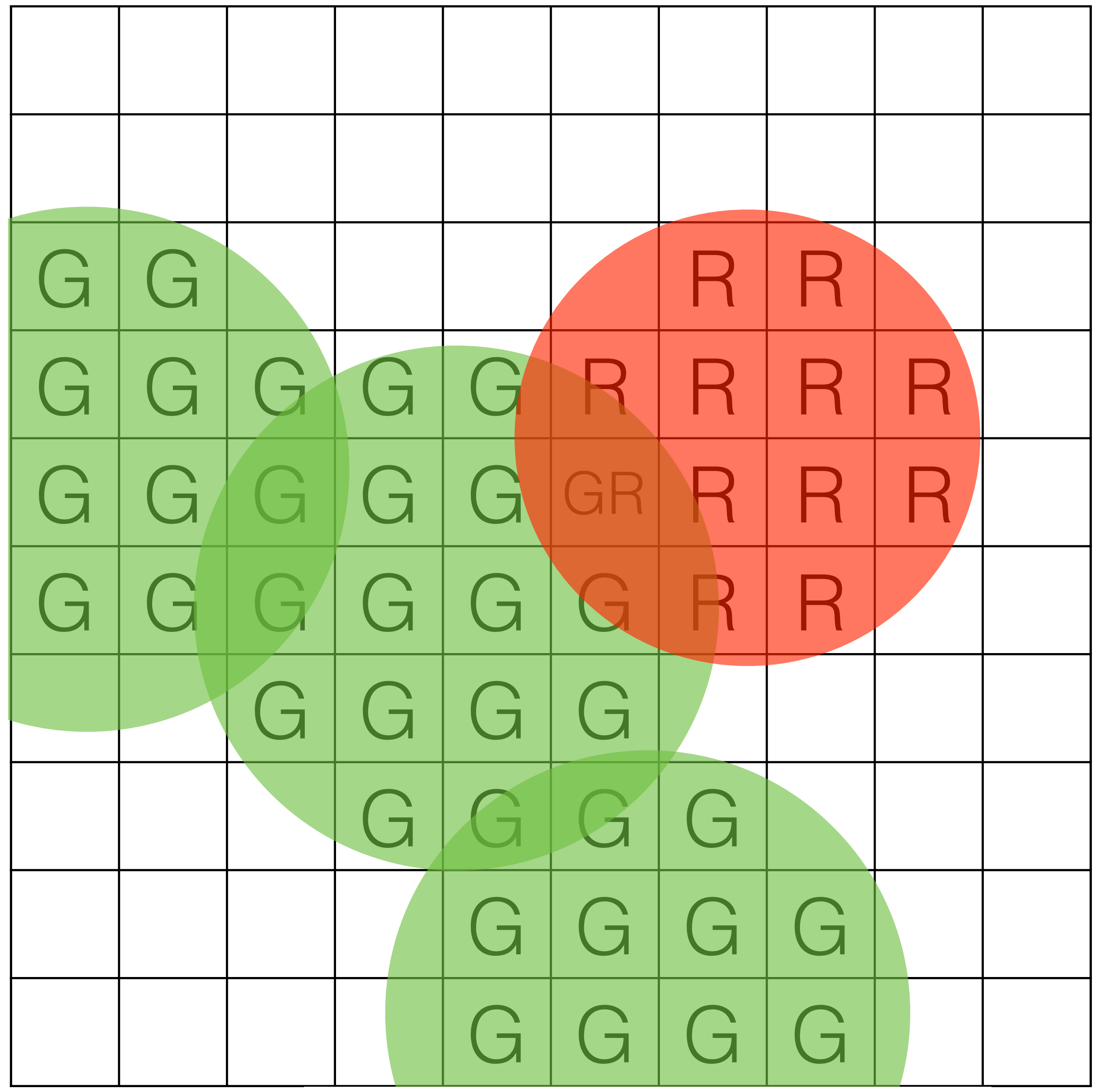
CNN

Pose Prediction

Binding Discrimination

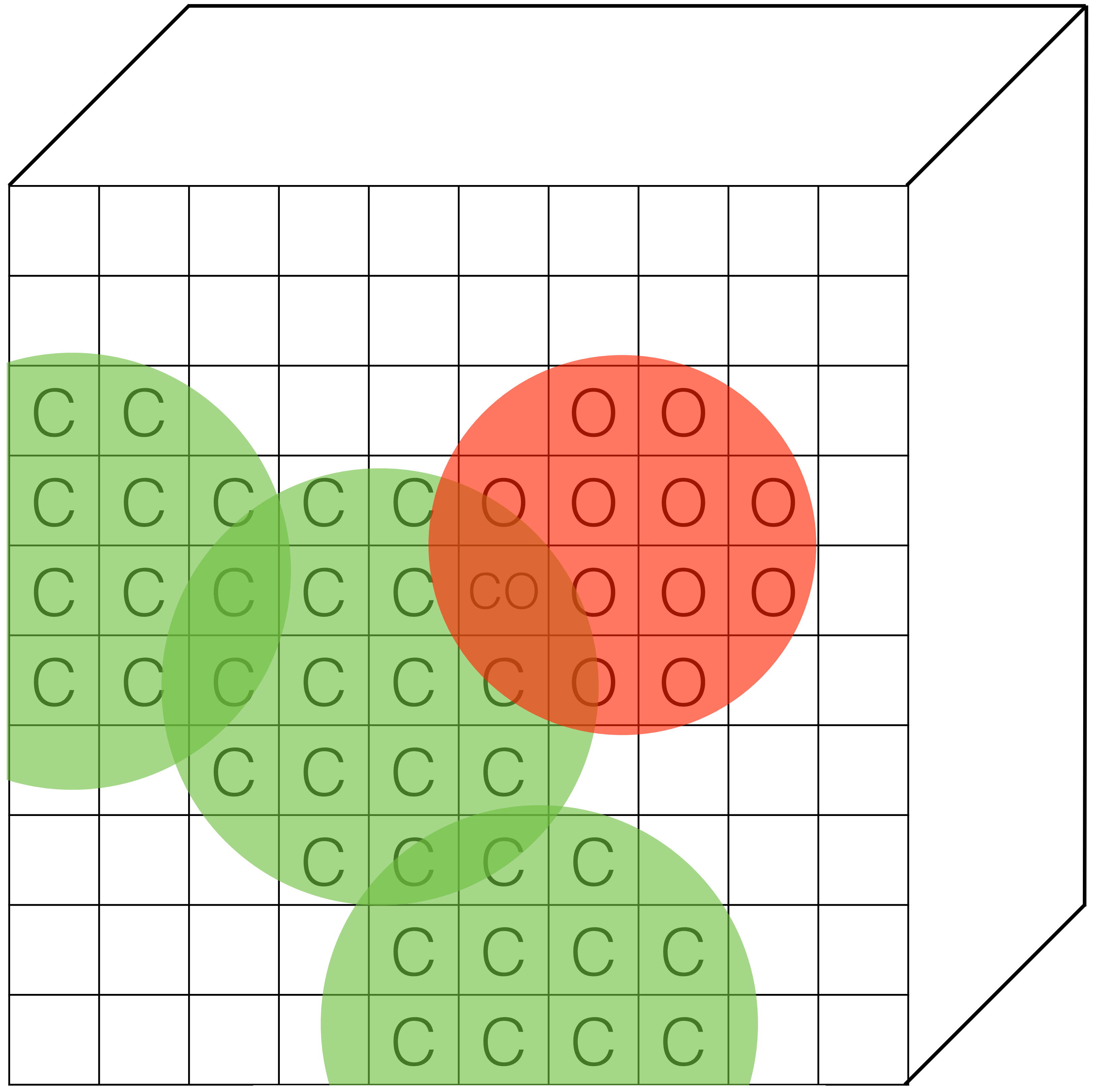
Affinity Prediction

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**.

Training Data

PDBbind

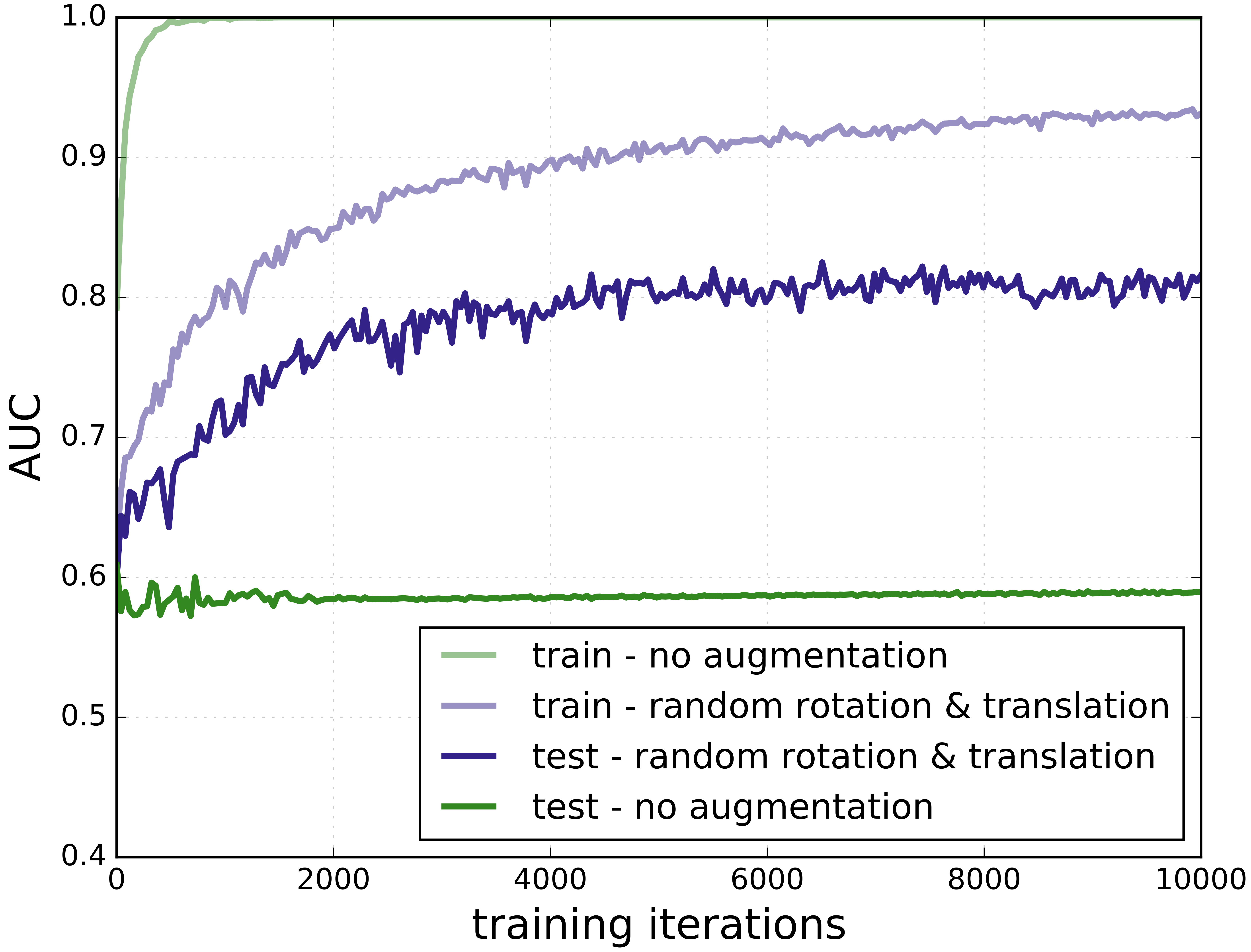
Pose Prediction

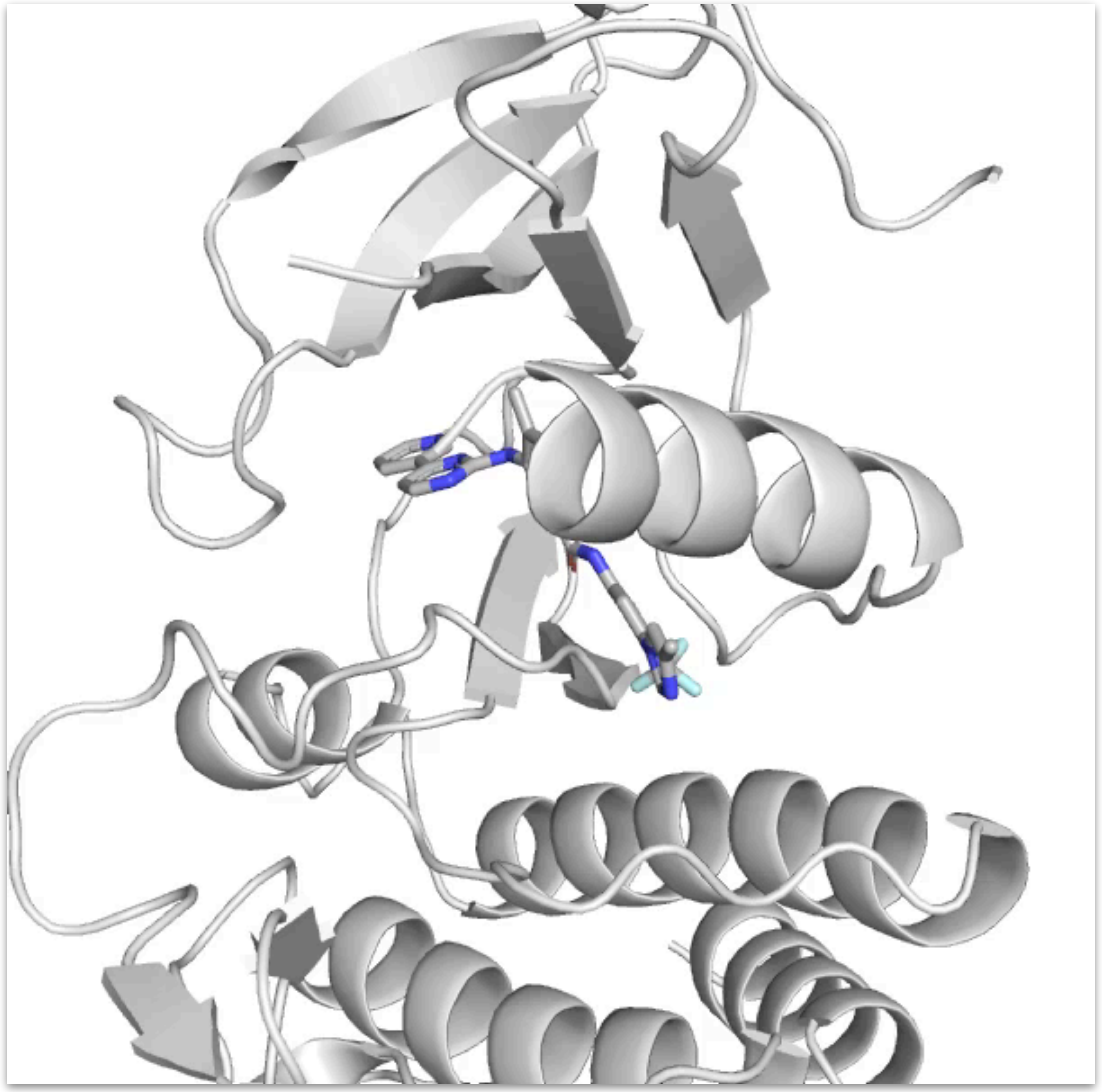
4056 protein-ligand complexes
- diverse targets
- wide range of affinities
- generate poses with AutoDock Vina
- include minimized crystal pose

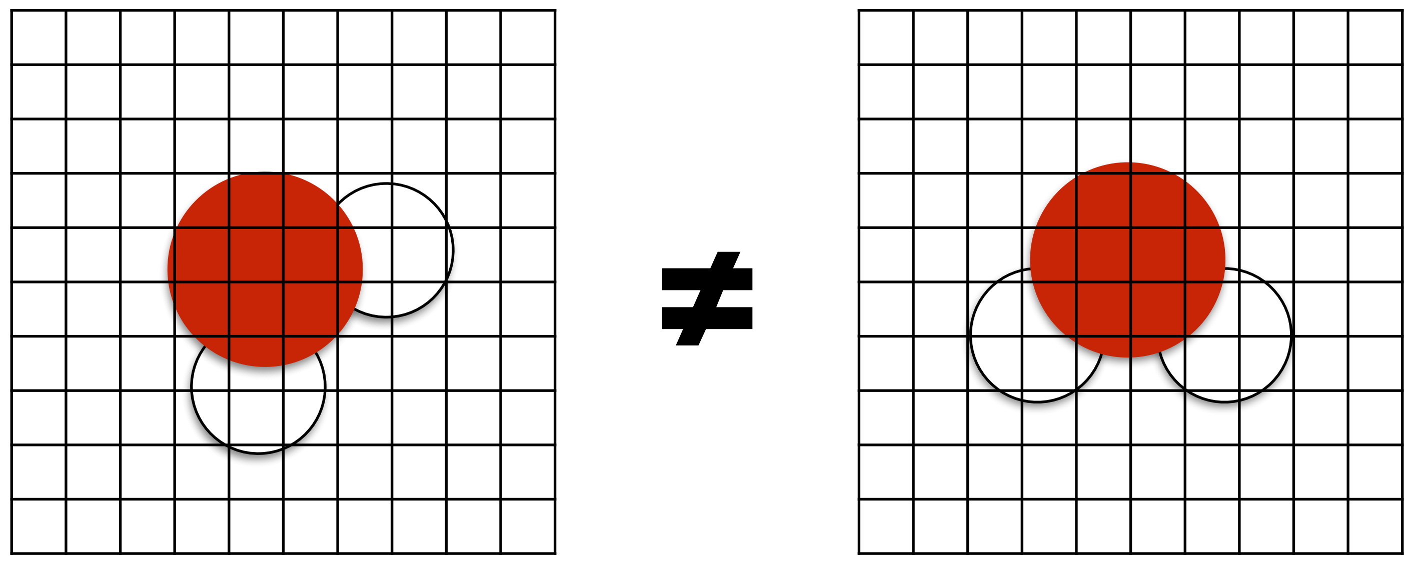
Affinity Prediction

- 8,688 low RMSD poses
- assign known affinity
- **regression problem**

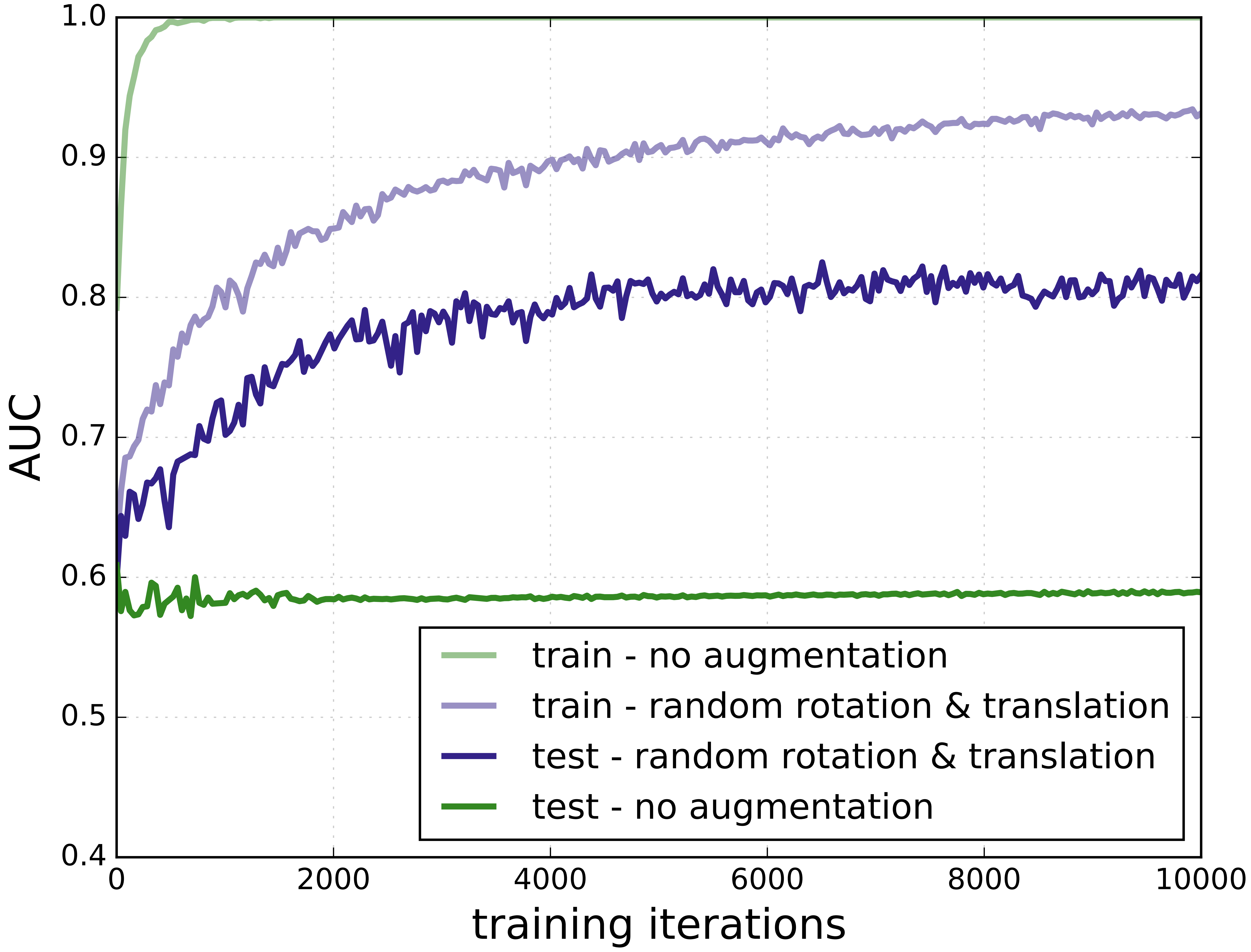
Data Augmentation

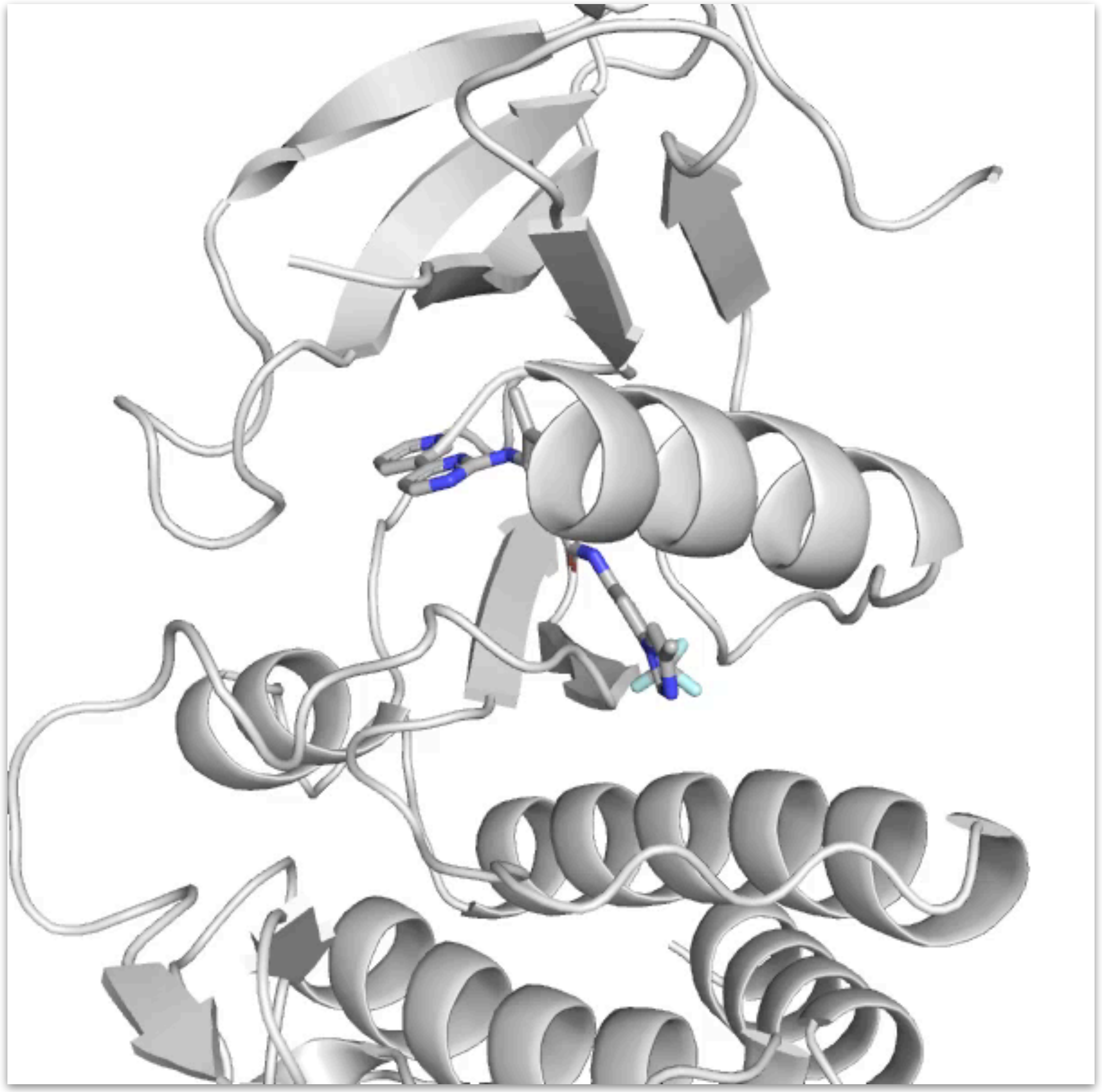


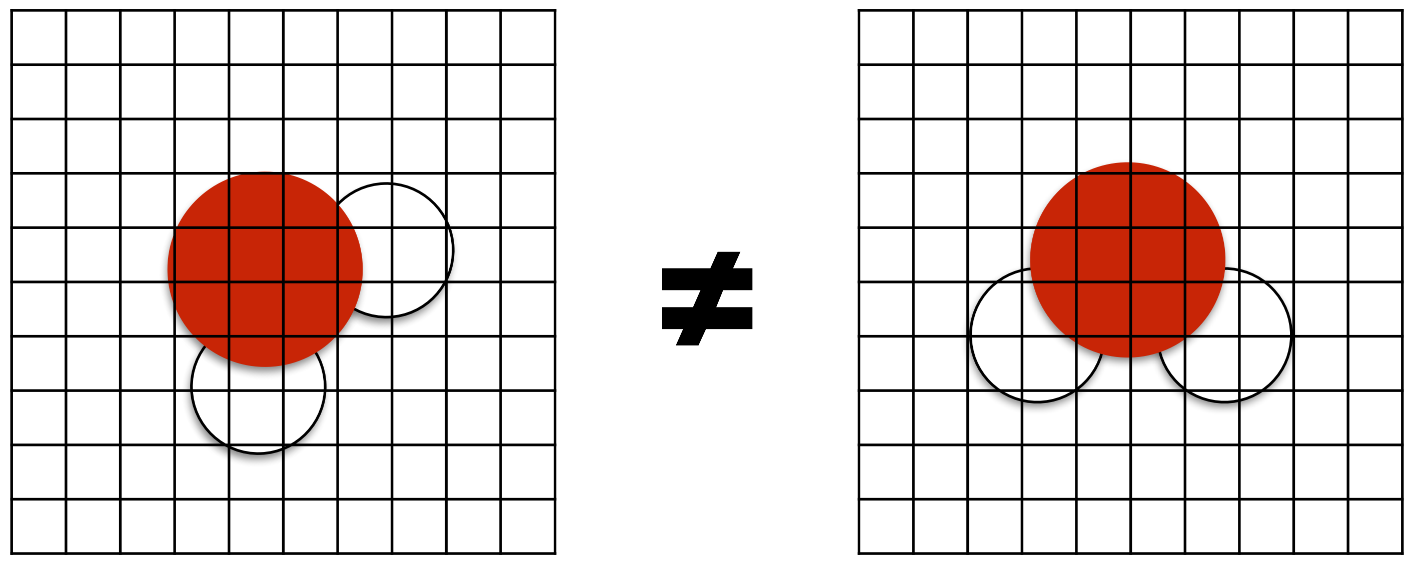




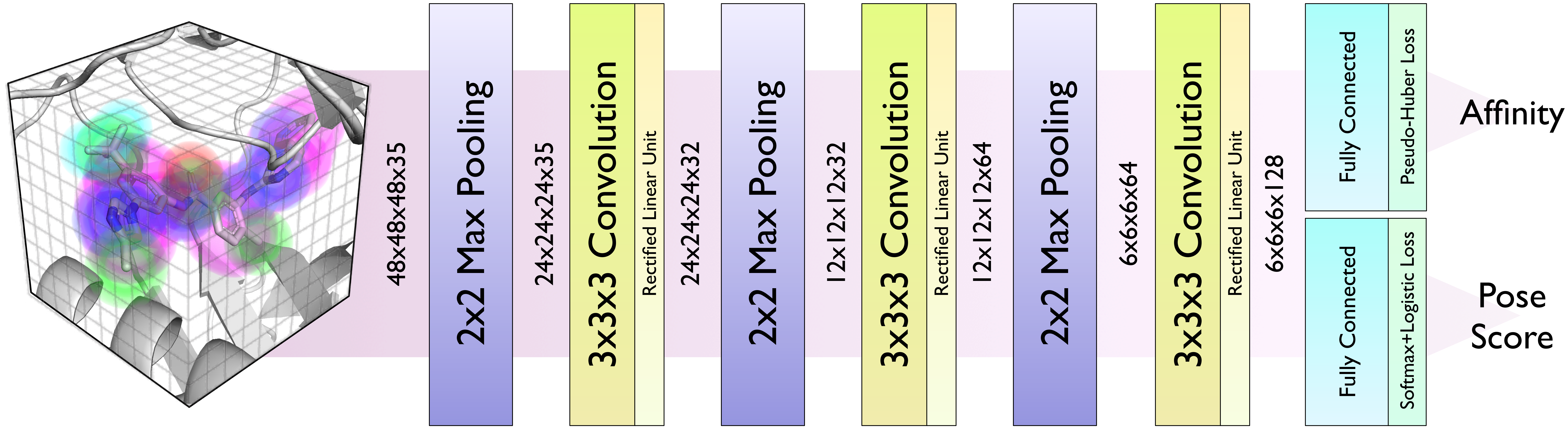
Data Augmentation



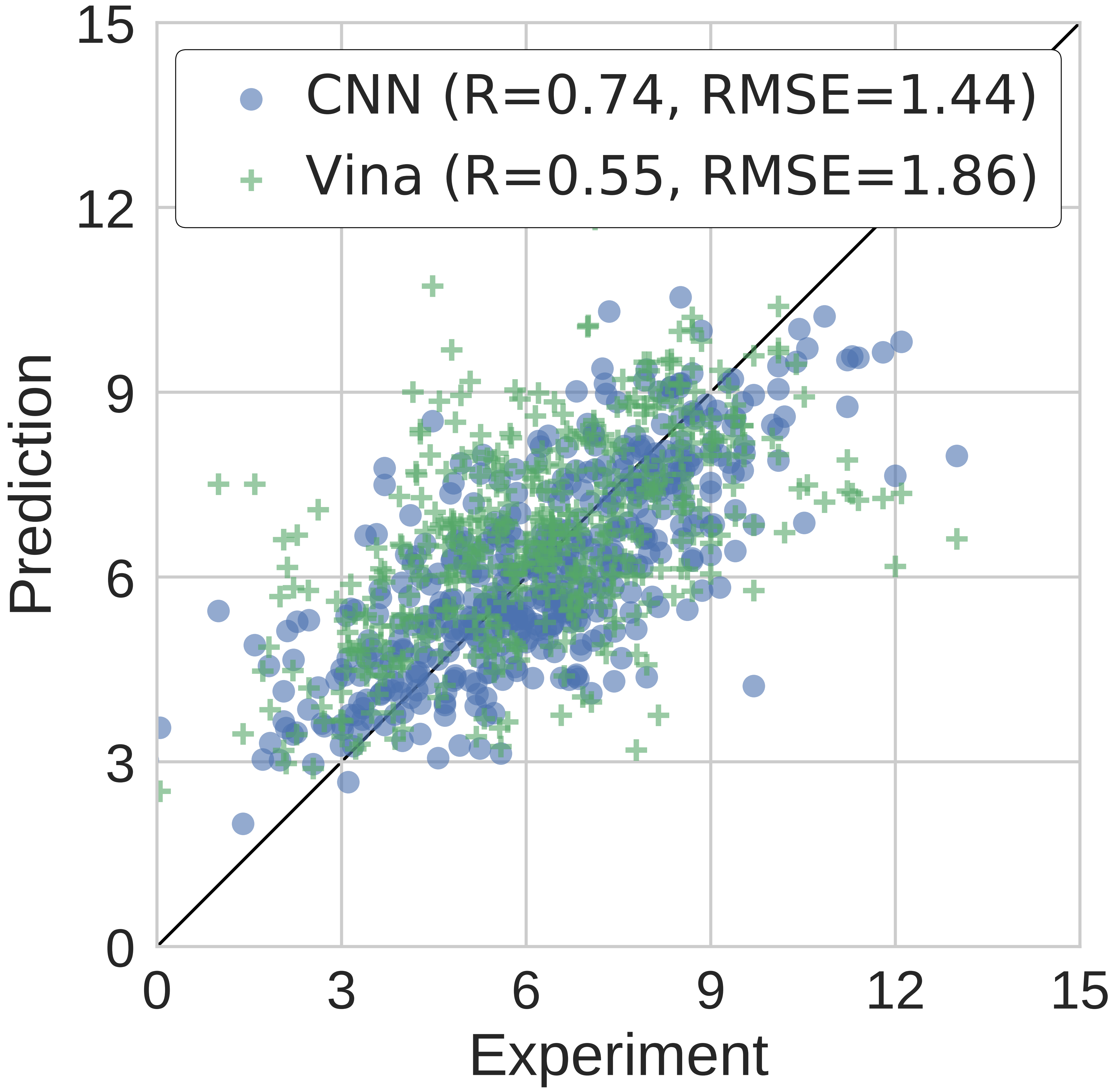


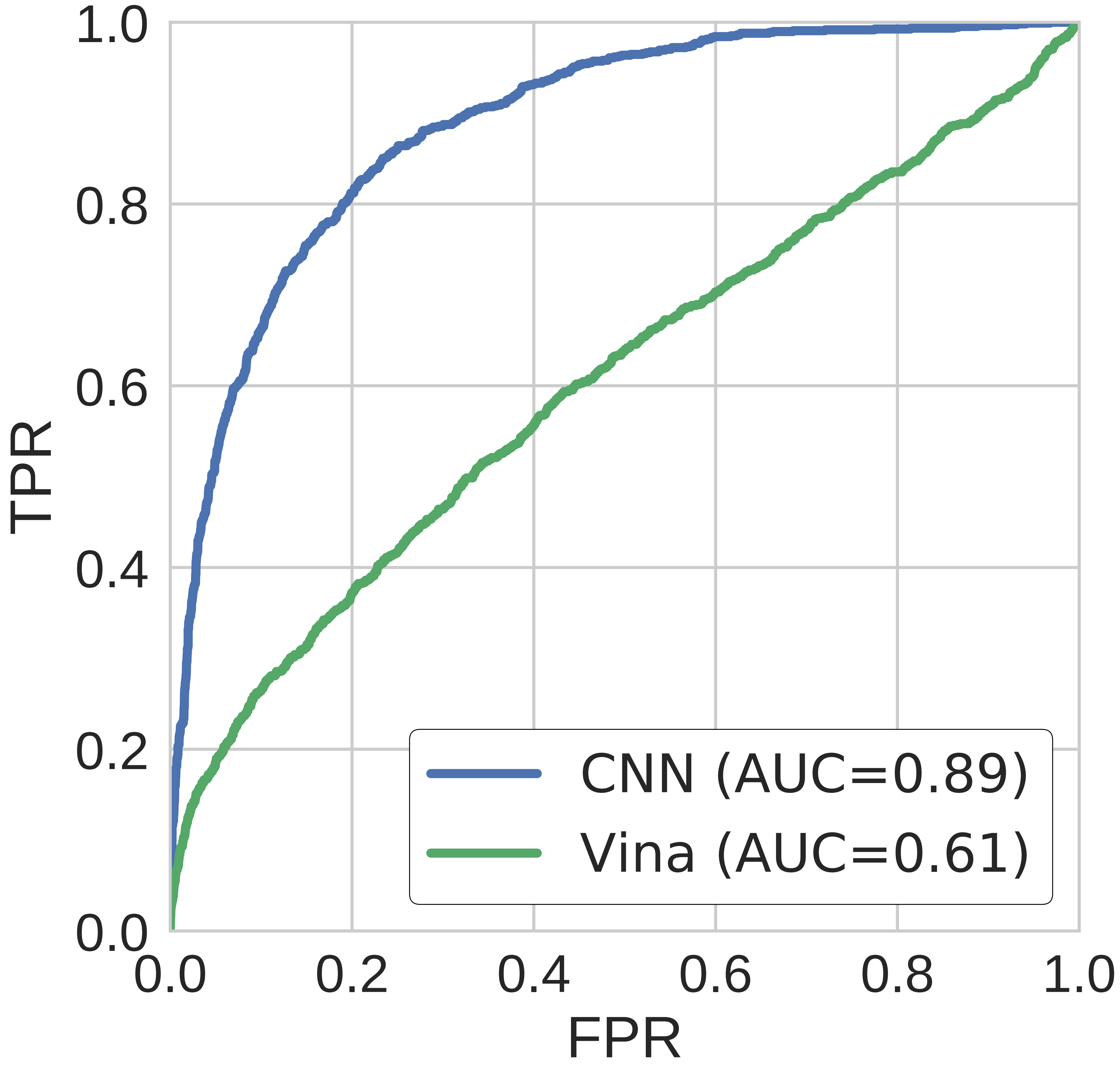


Model


48x48x48x35
2x2 Max Pooling
24x24x24x35
3x3x3 Convolution
Rectified Linear Unit
24x24x24x32
2x2 Max Pooling
12x12x12x32
3x3x3 Convolution
Rectified Linear Unit
12x12x12x64
2x2 Max Pooling
6x6x6x64
3x3x3 Convolution
Rectified Linear Unit
6x6x6x128
Fully Connected
Pseudo-Huber Loss
Affinity
Fully Connected
Softmax+Logistic Loss
Pose Score

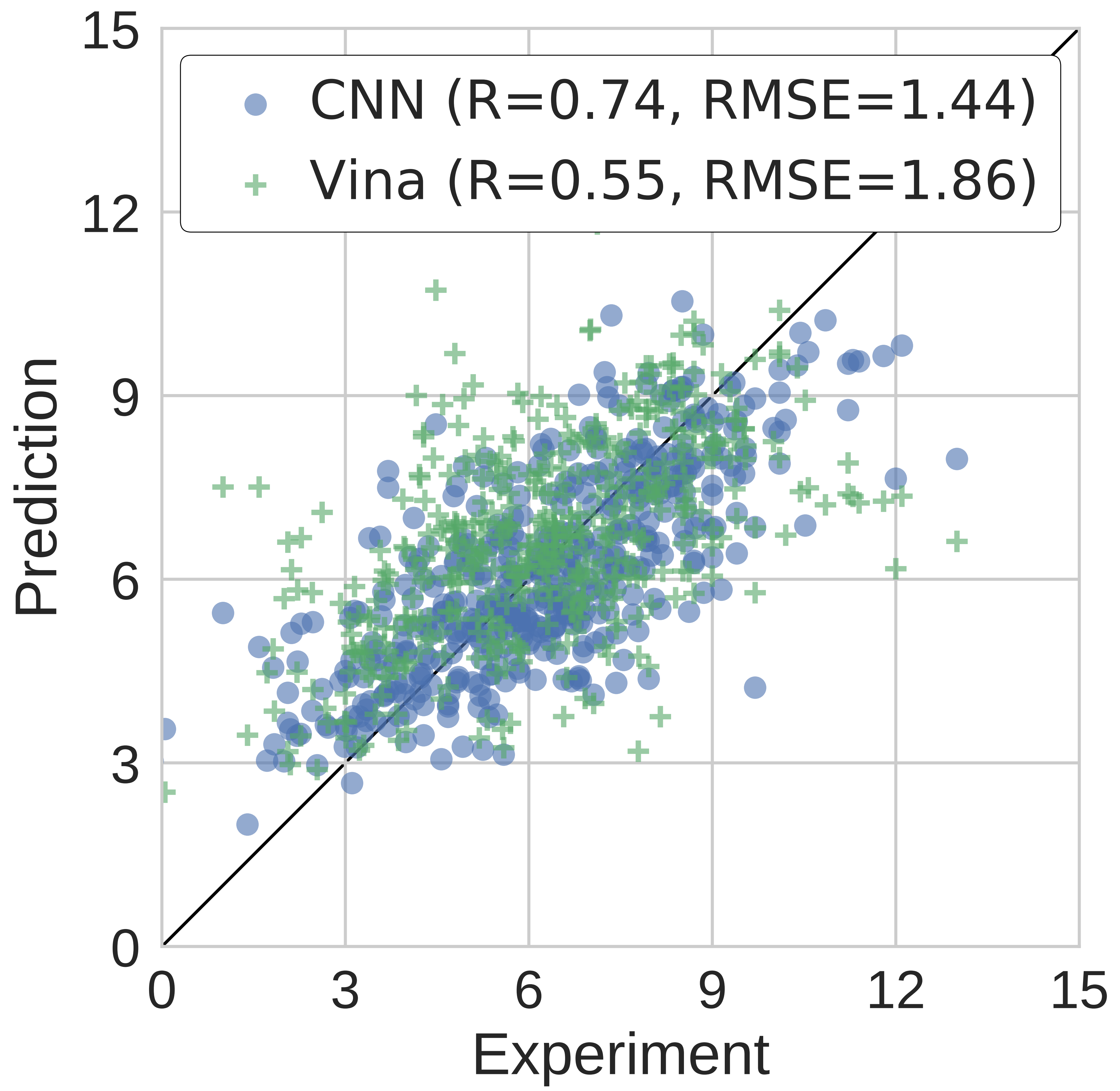
Results

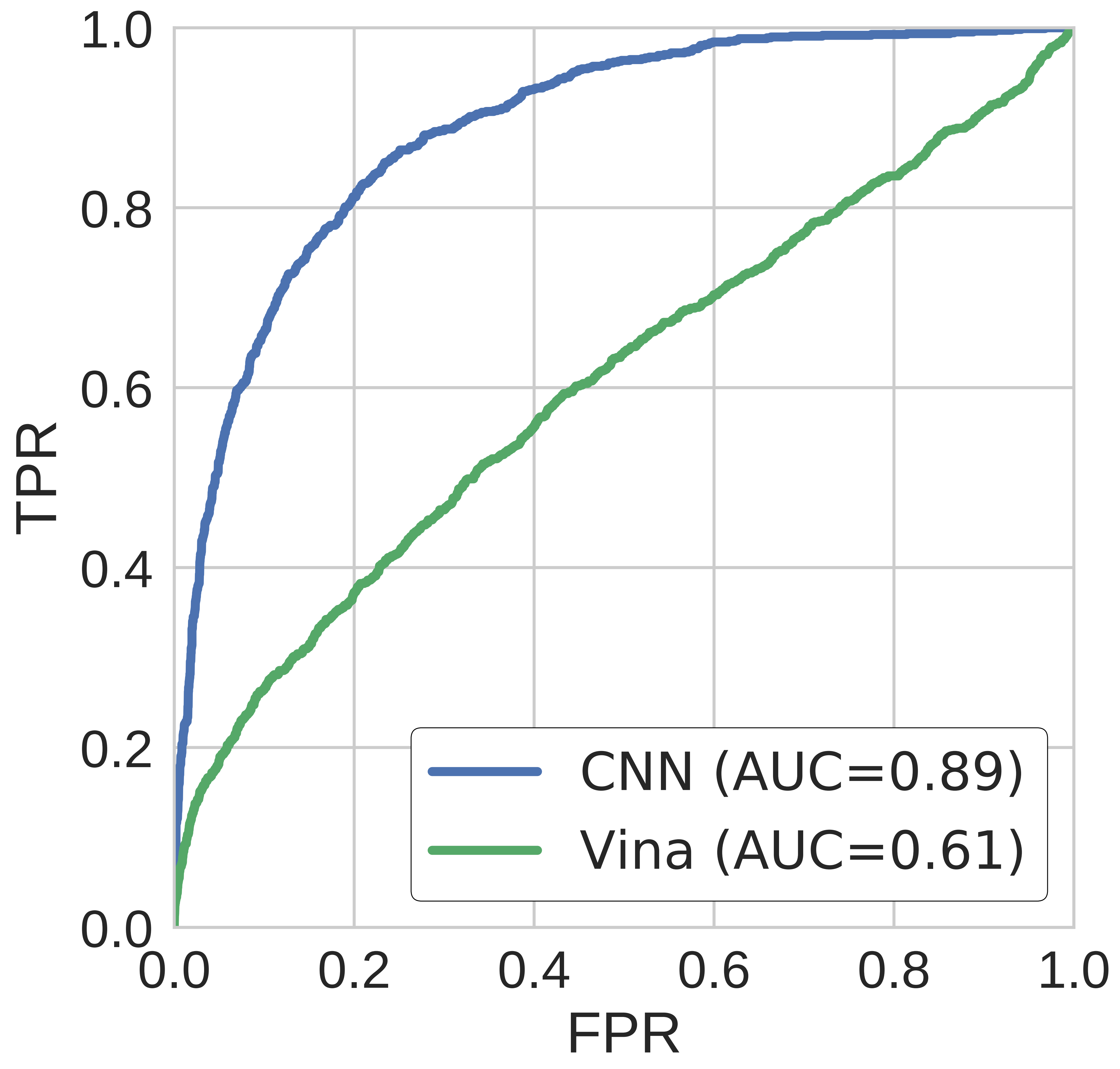




Trained on PDBbind refined; tested on CSAR

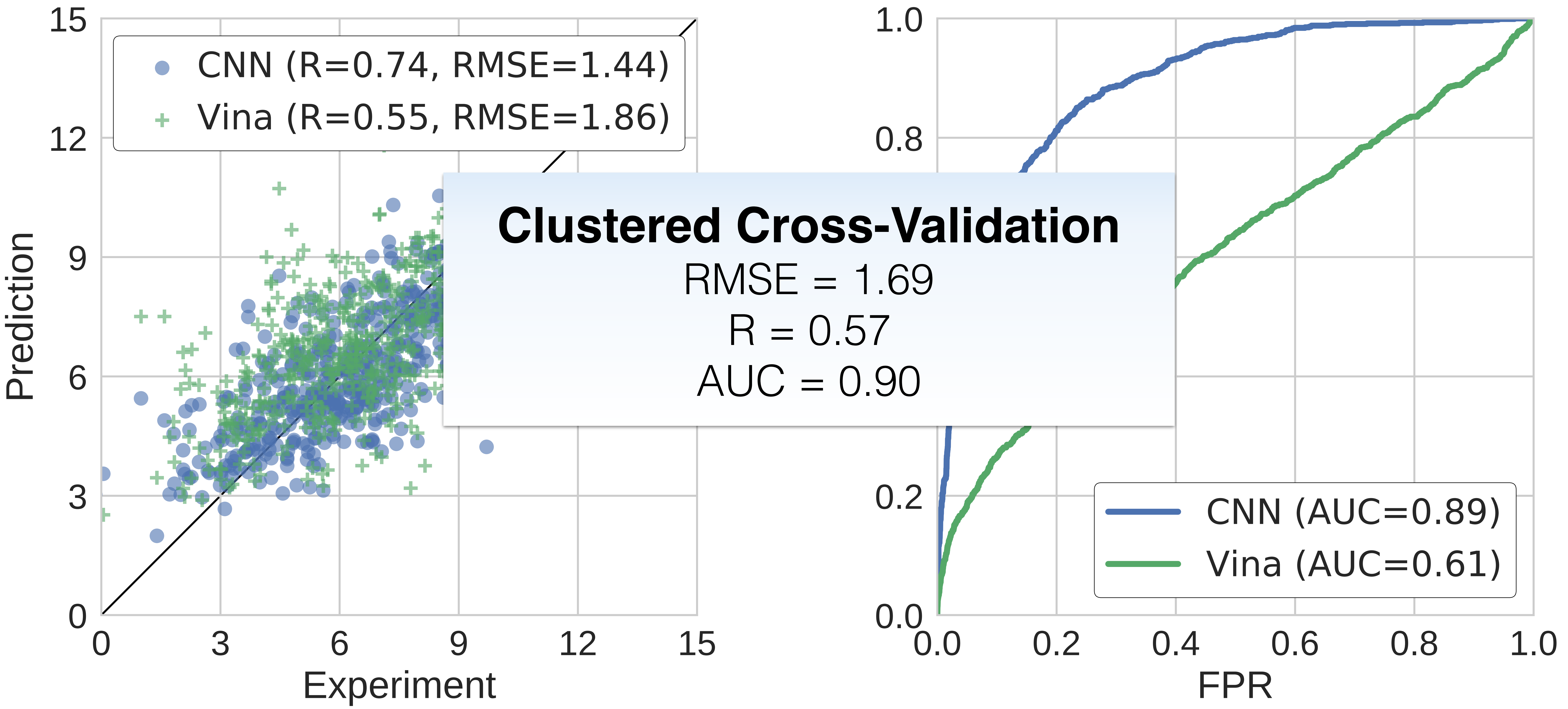
Results





Trained on PDBbind refined; tested on CSAR

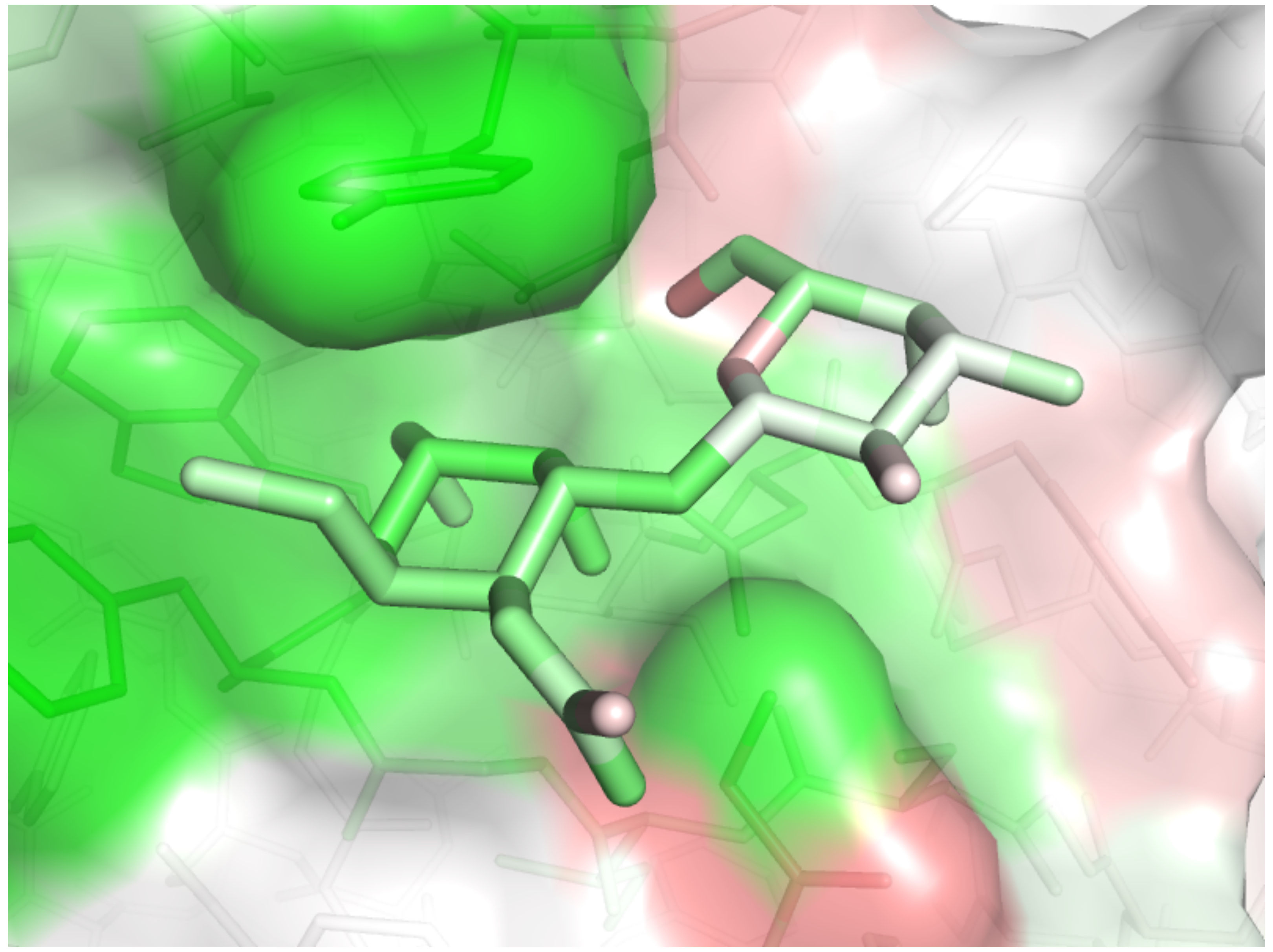
Results



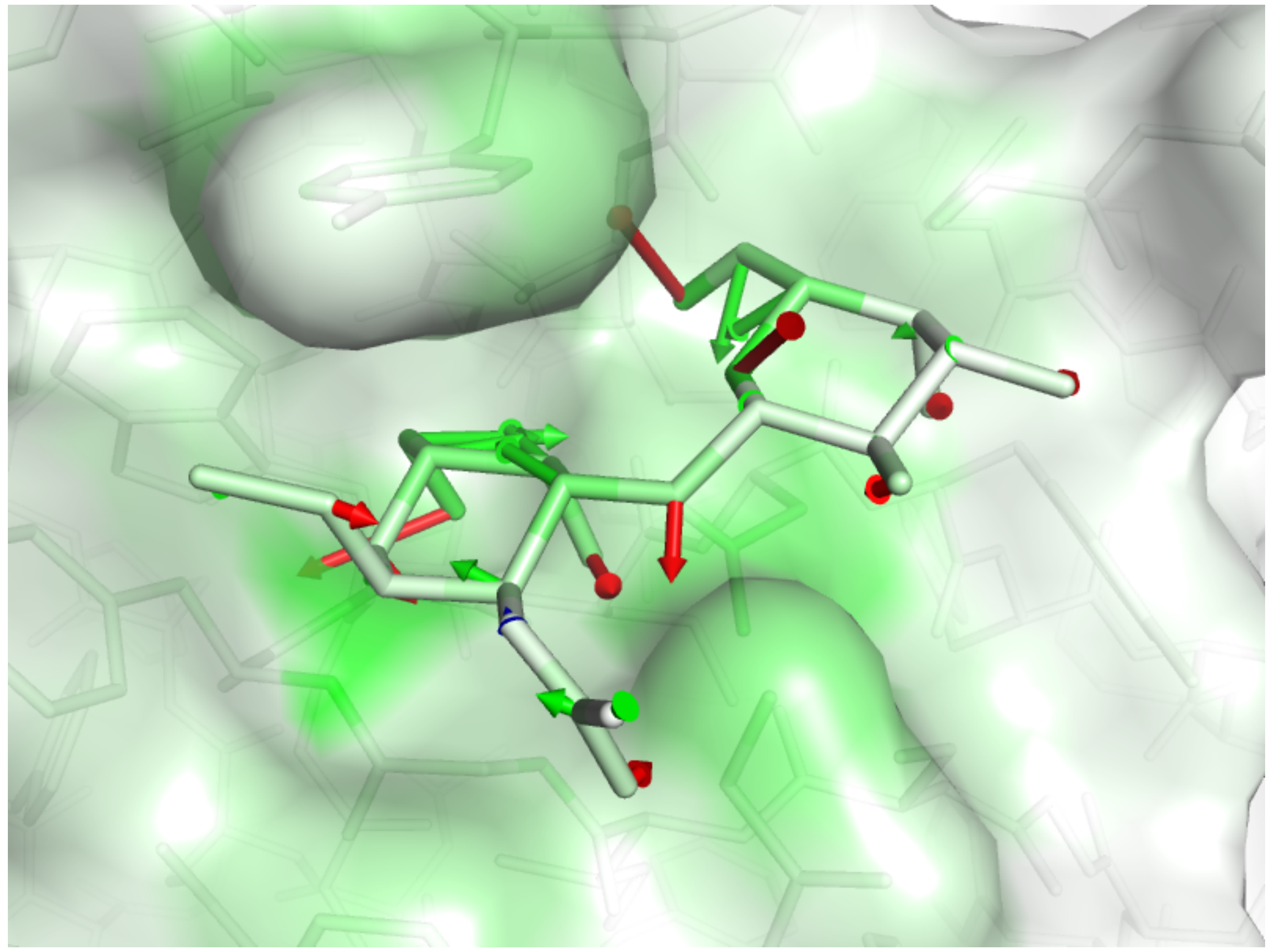
Clustered Cross-Validation
RMSE = 1.69
R = 0.57
AUC = 0.90

Trained on PDBbind refined; tested on CSAR

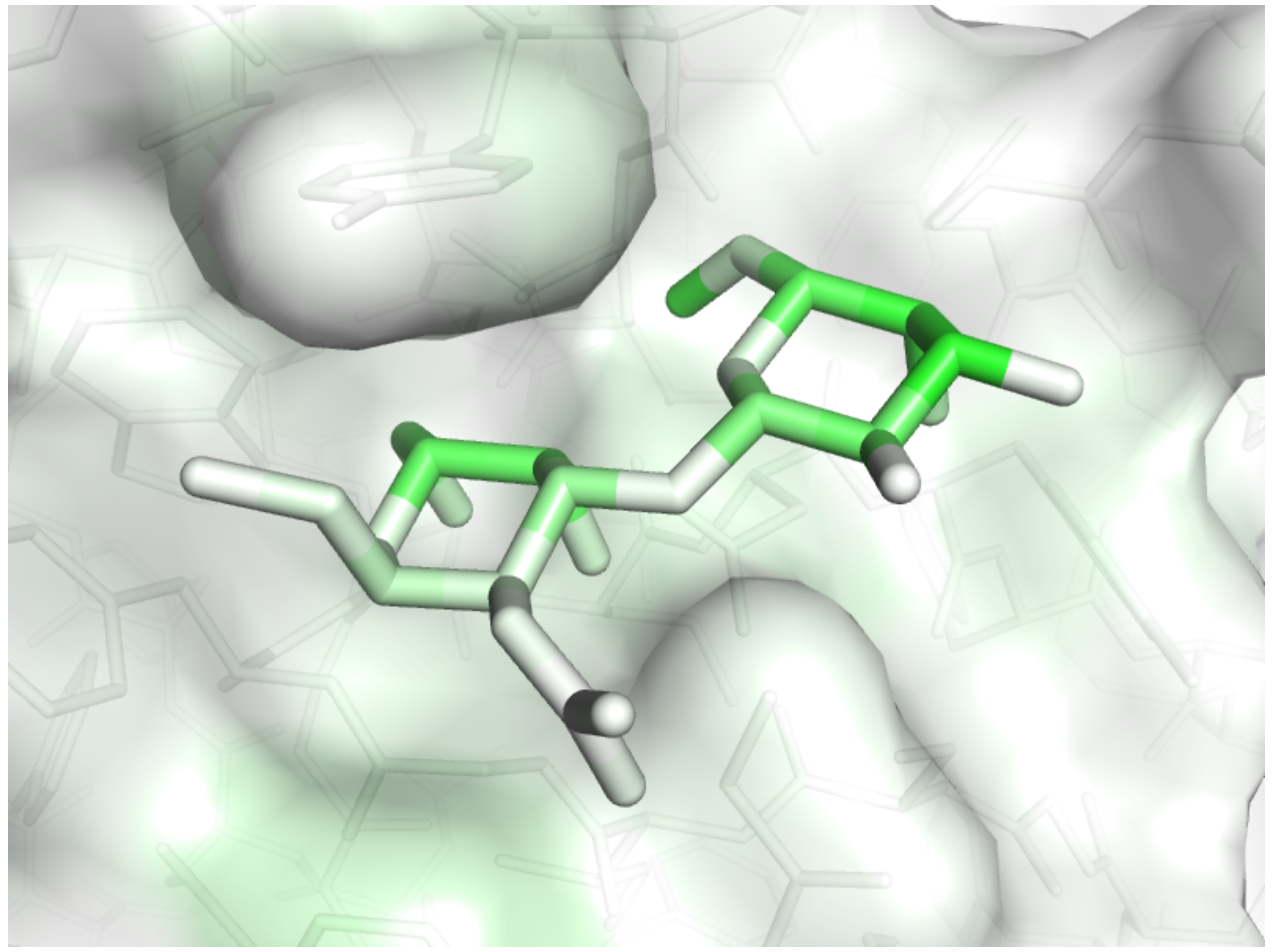
Visualization



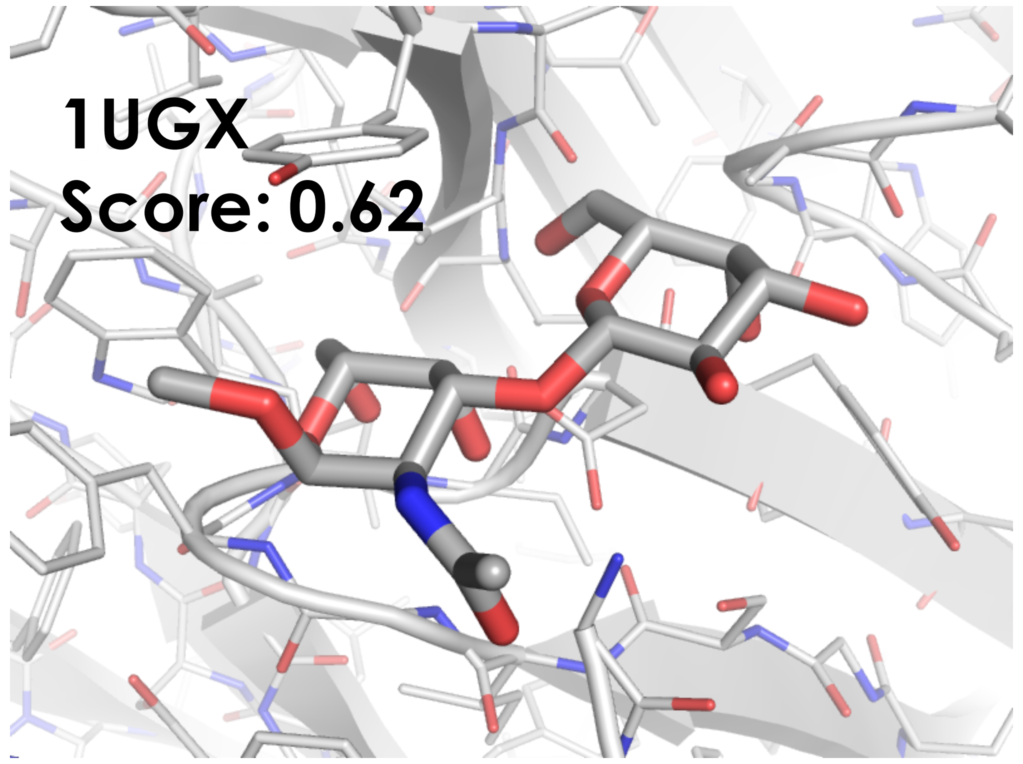
masking



gradients

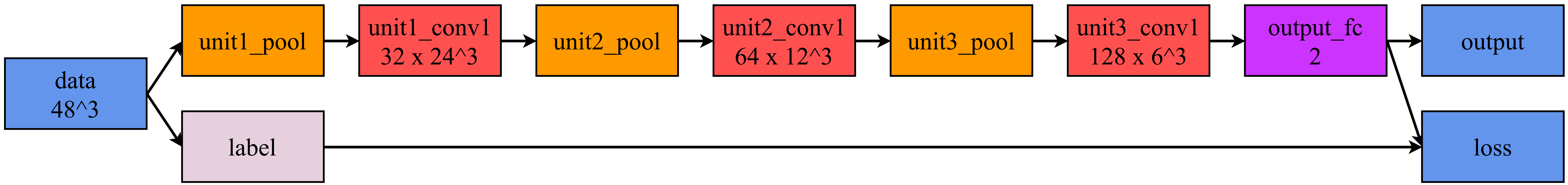


layer-wise relevance

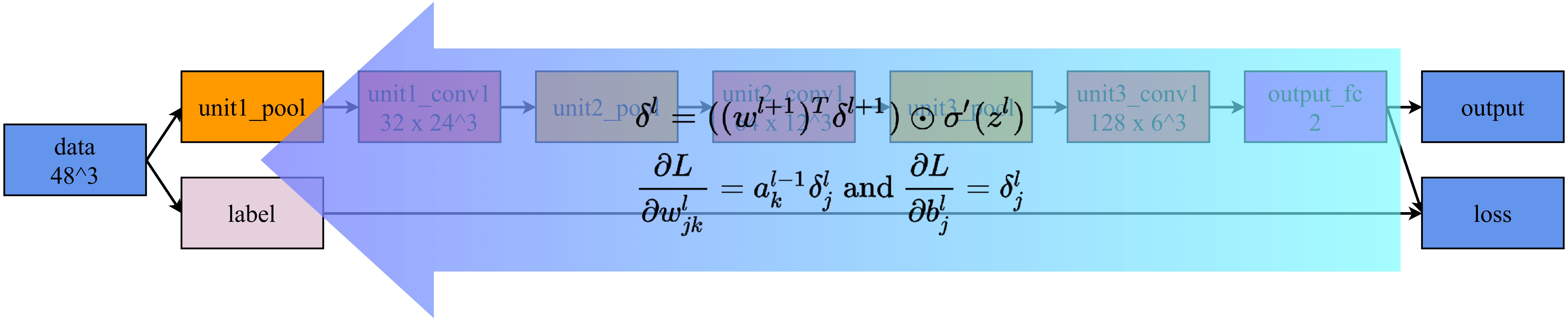


Visualizing Empty Space

Beyond Scoring



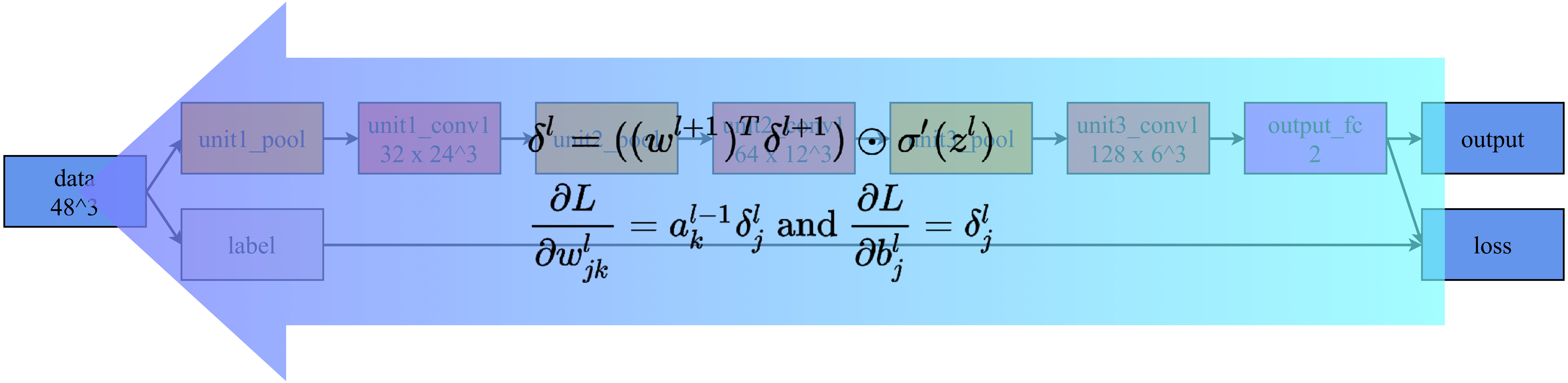
Beyond Scoring



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

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

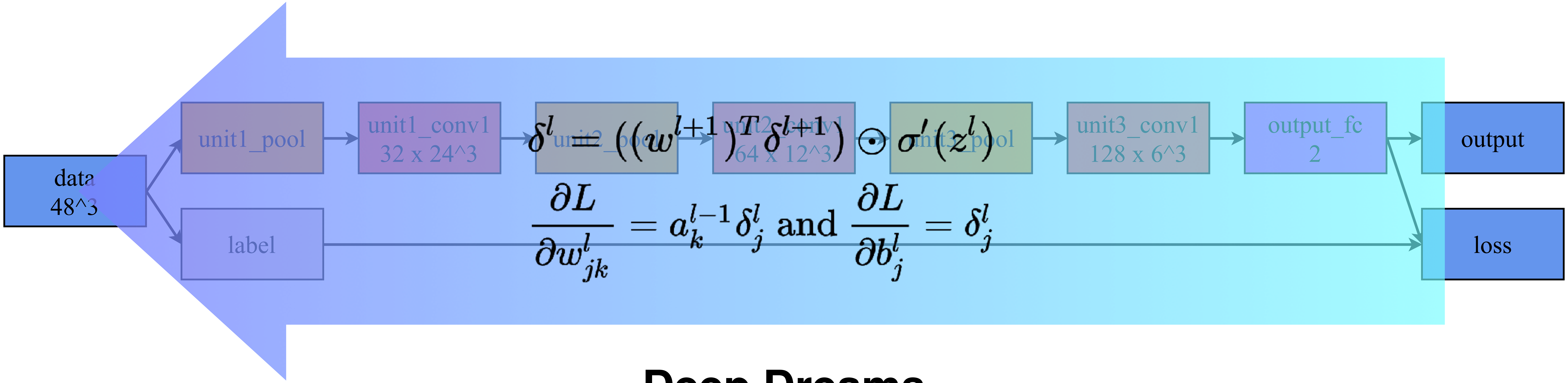
Beyond Scoring



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

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

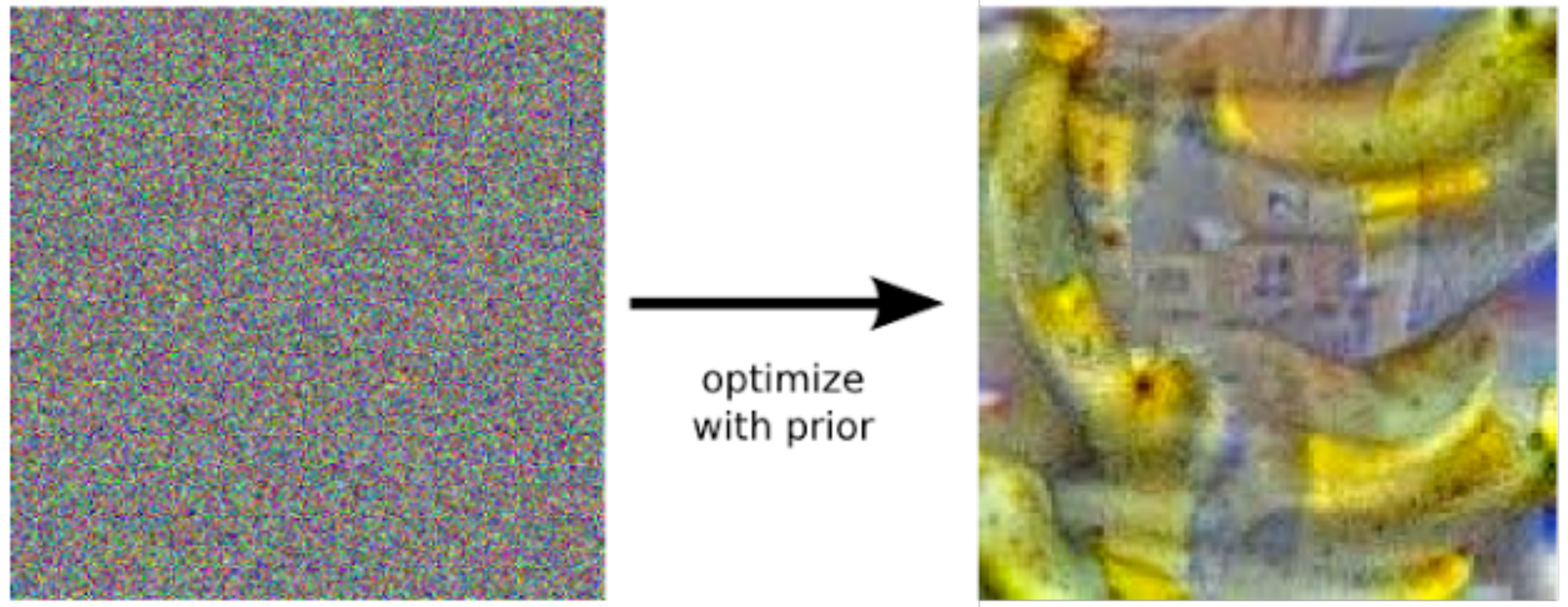
Beyond Scoring



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

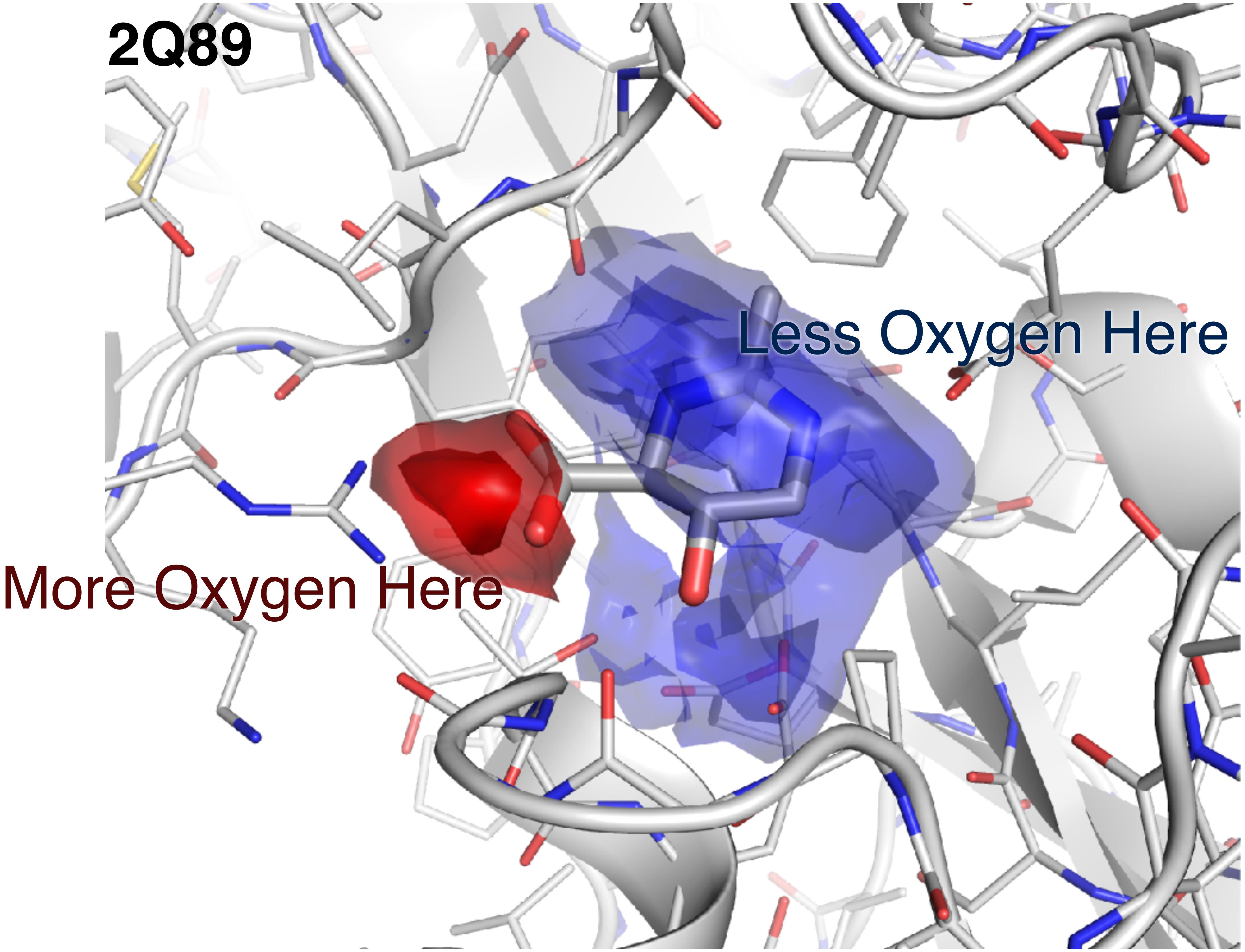
$$\frac{\partial L}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j \text{ and } \frac{\partial L}{\partial b^l_j} = \delta^l_j$$

Deep Dreams

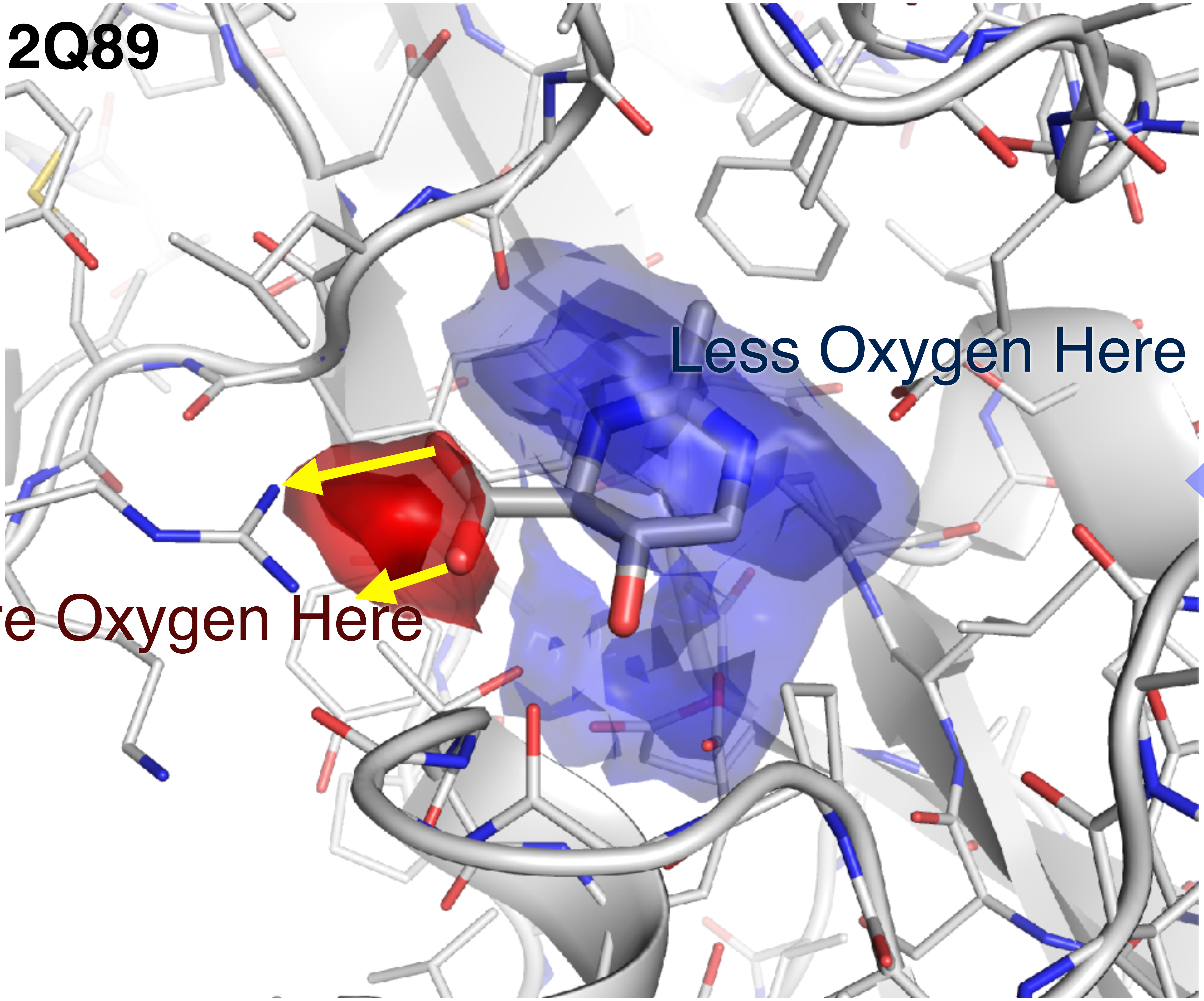


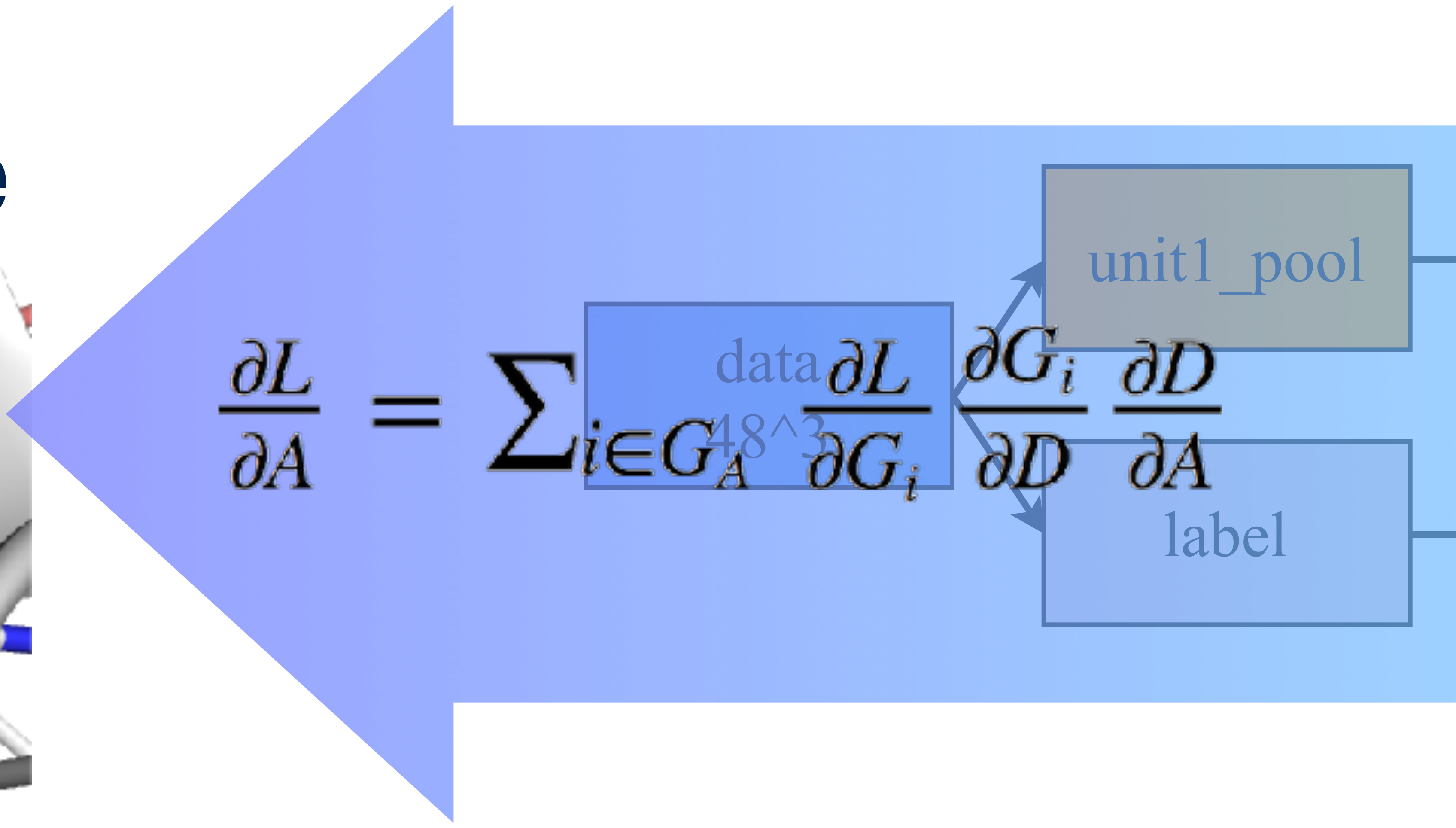
https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Beyond Scoring

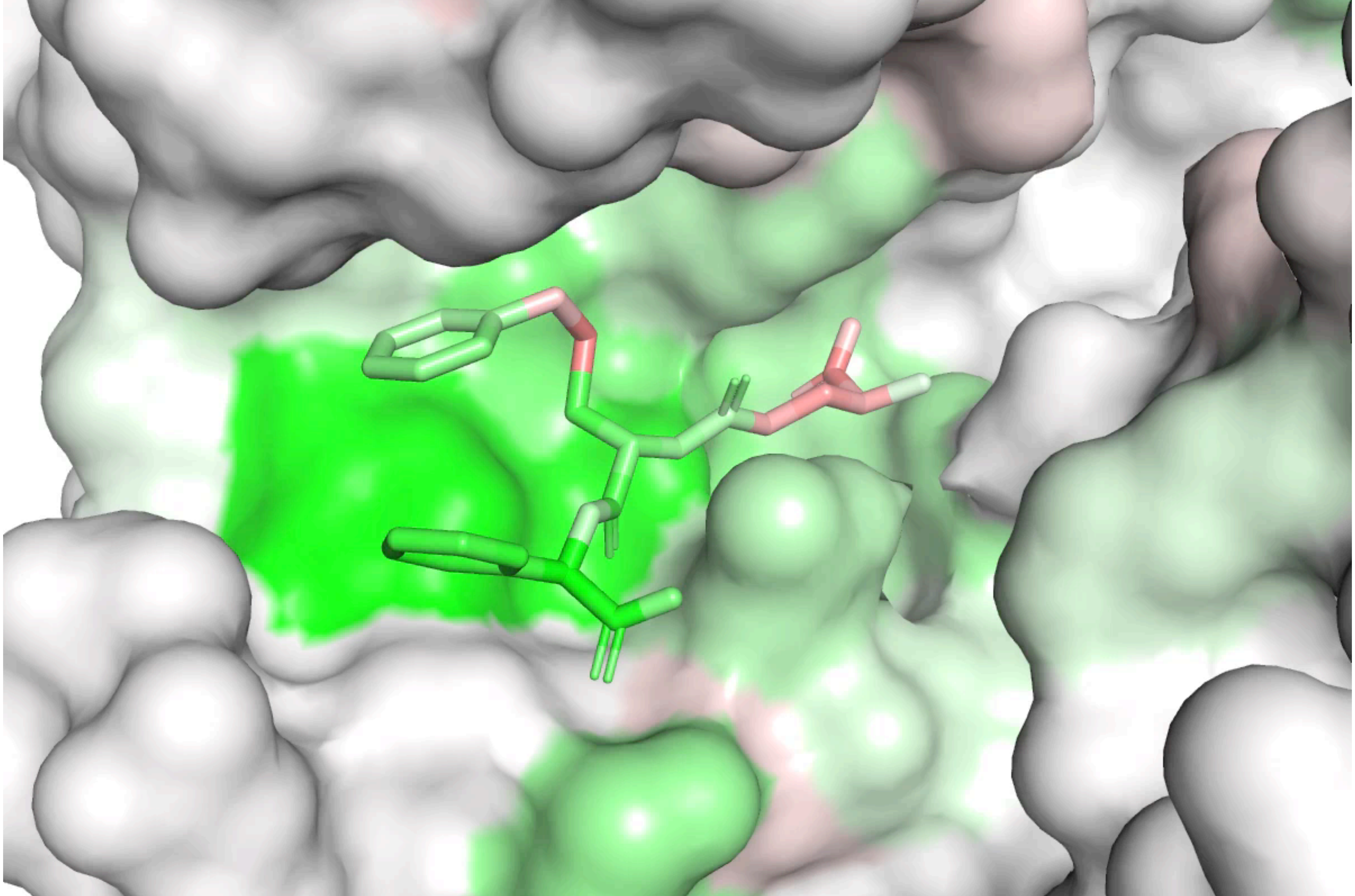


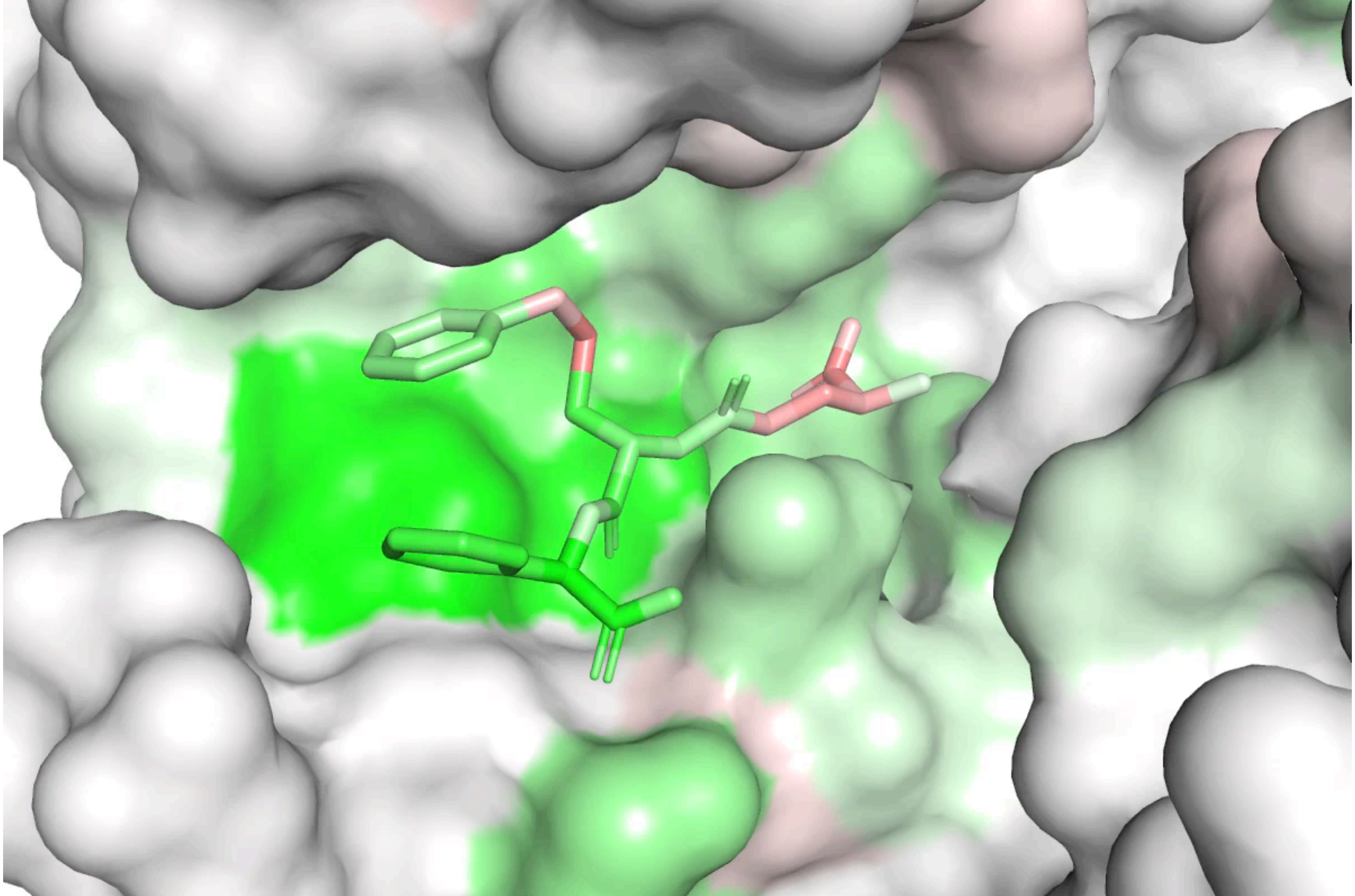
Beyond Scoring

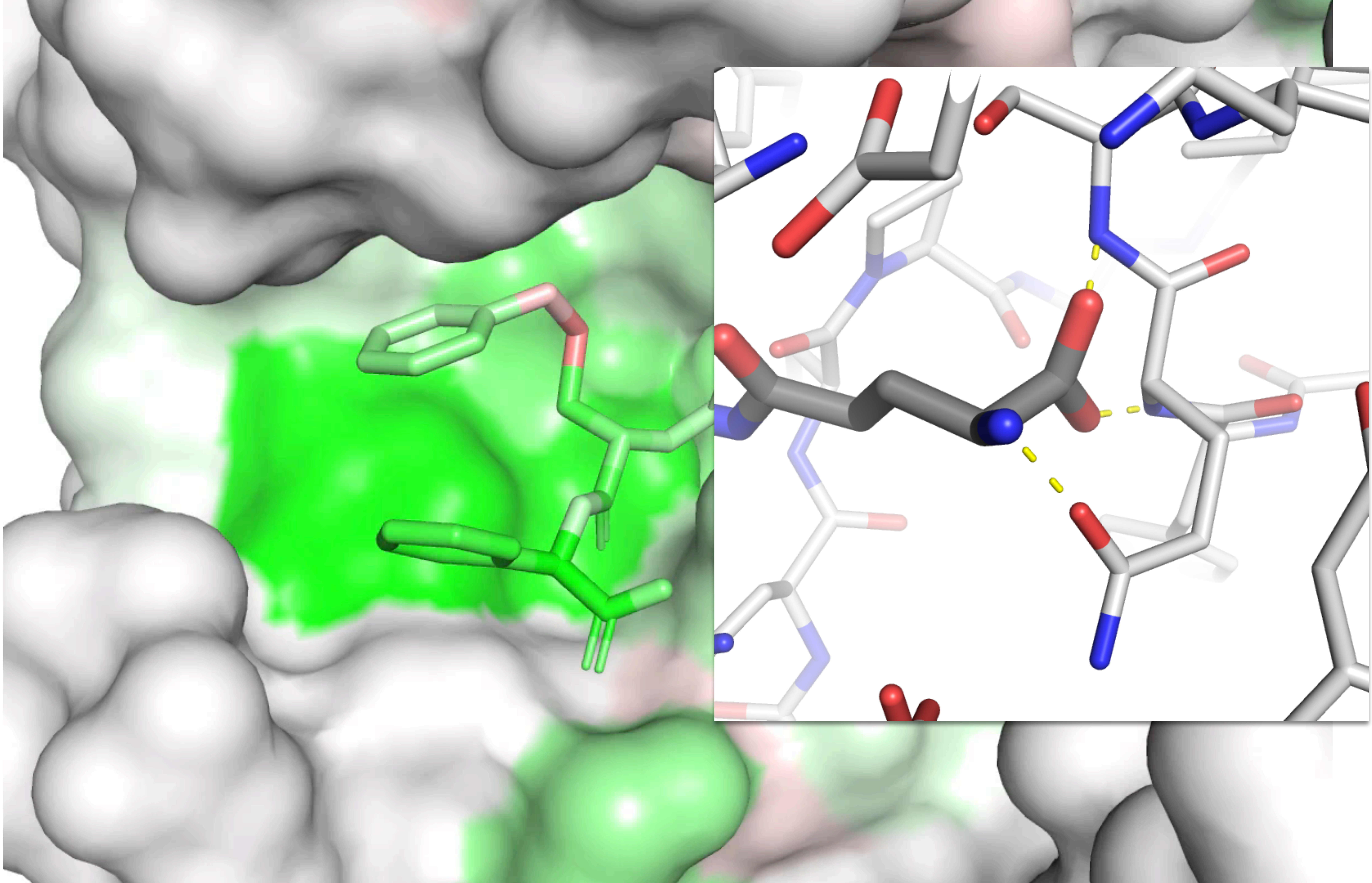


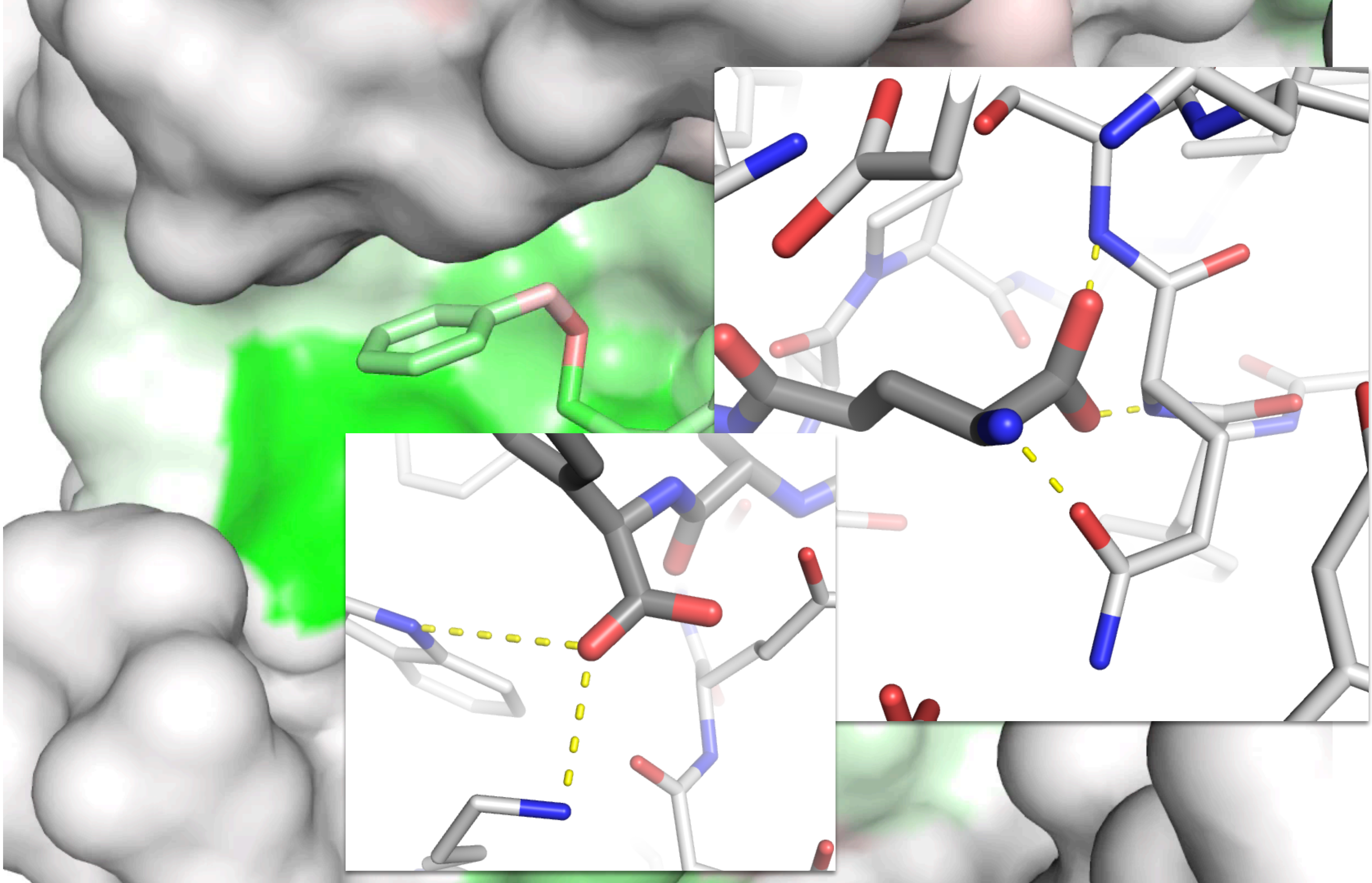


$$\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

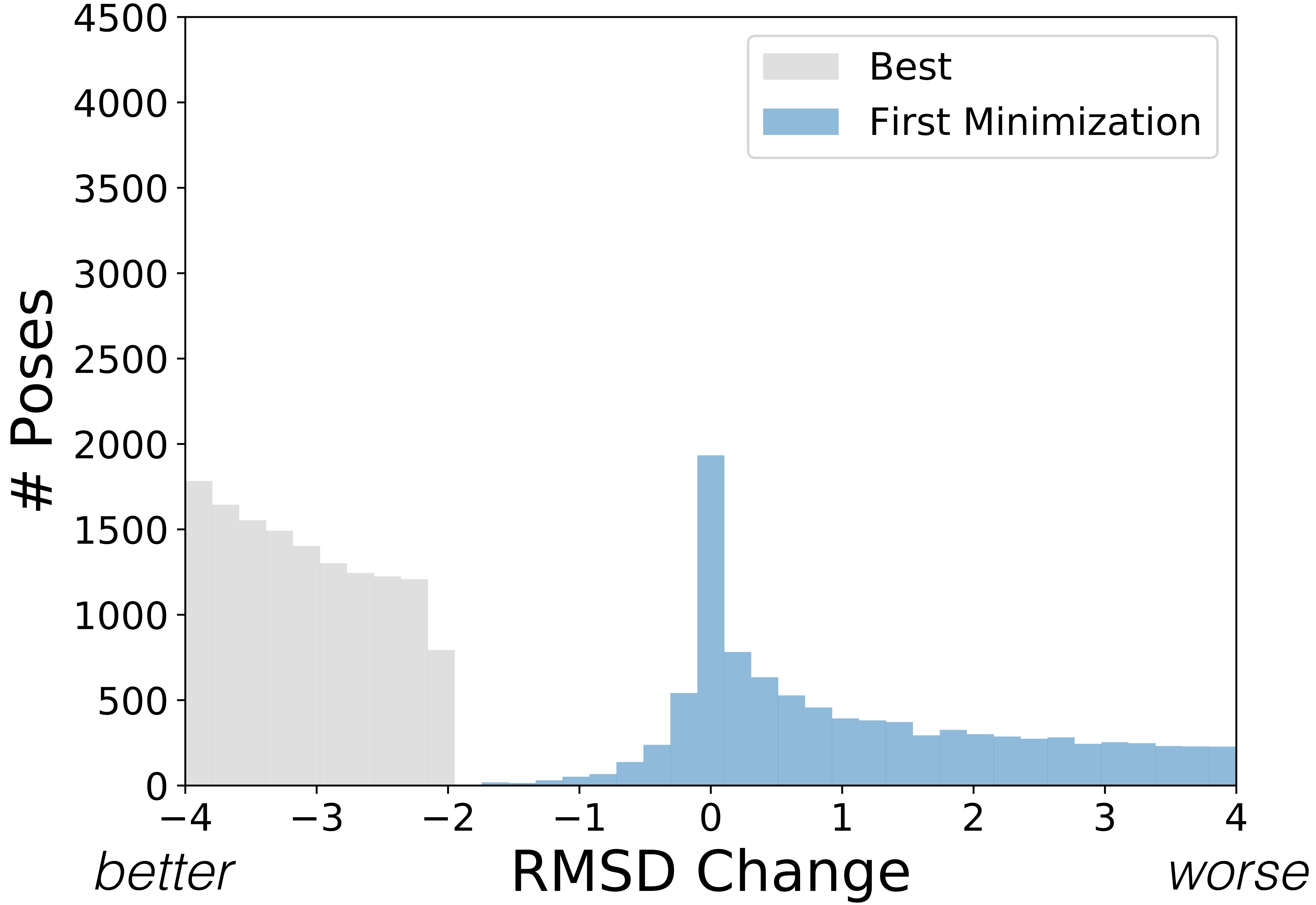




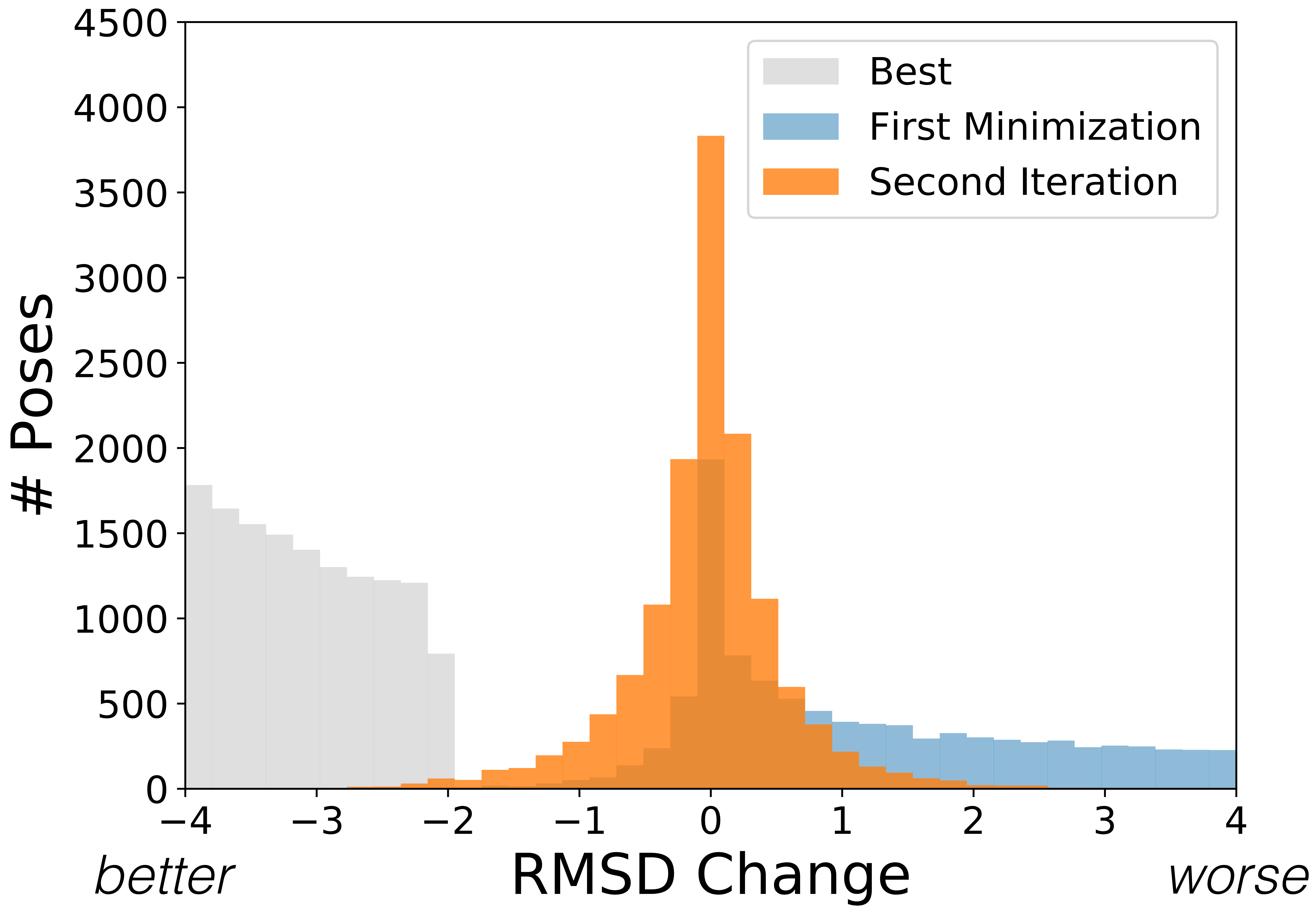




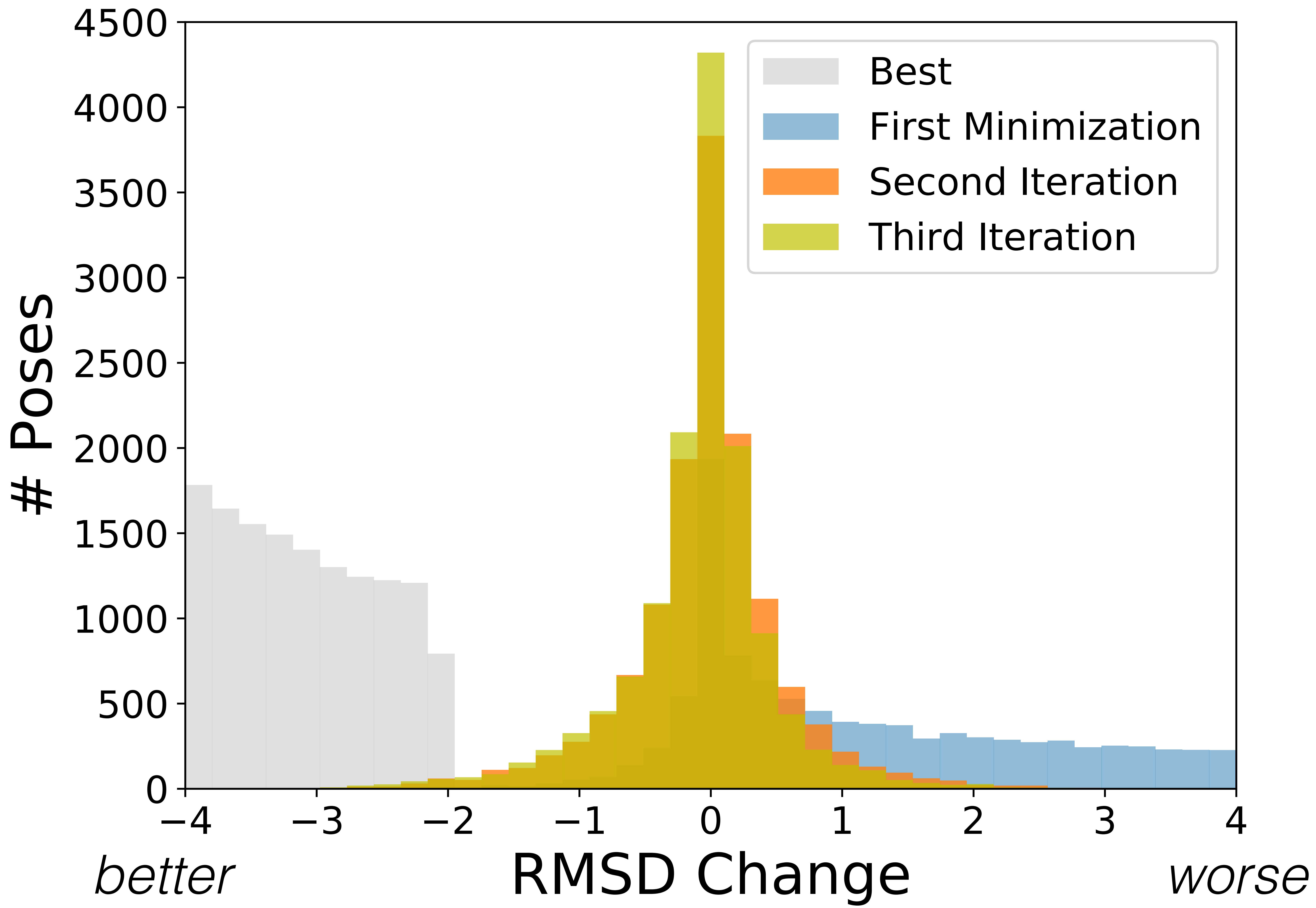
Optimizing Low RMSD Poses



Iterative Refinement



Iterative Refinement



Docking

vina/smina/gnina

Sampling

MCMC

MCMC

MCMC

MCMC

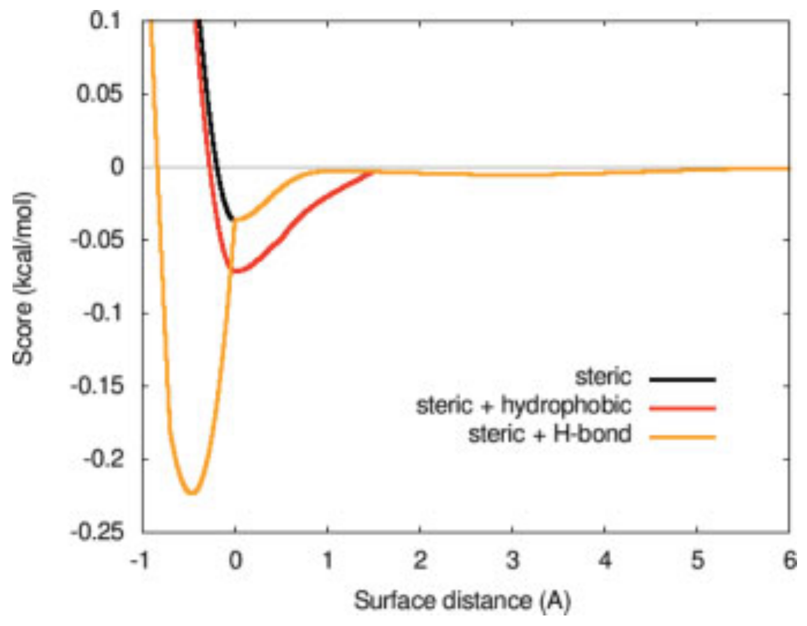
MCMC

⋮

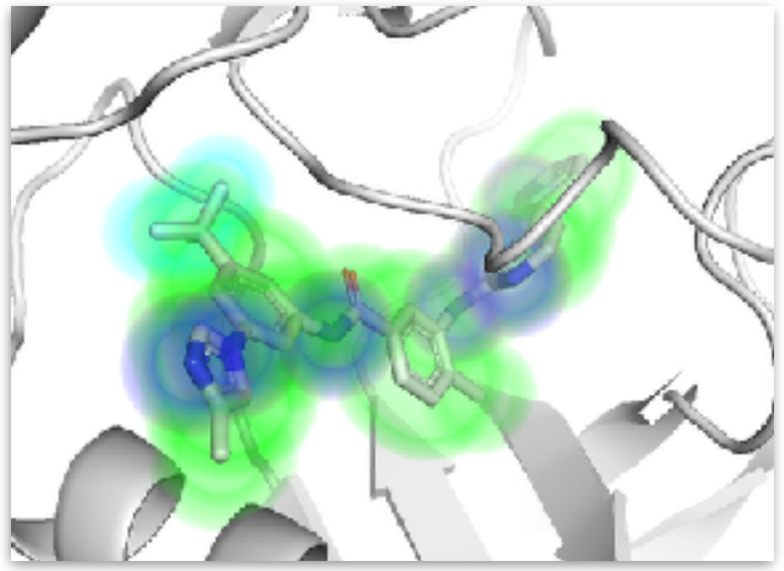
N (50) independent Monte Carlo chains
Scored with grid-accelerated Vina
Best identified pose retained

best poses

Refinement



Vina



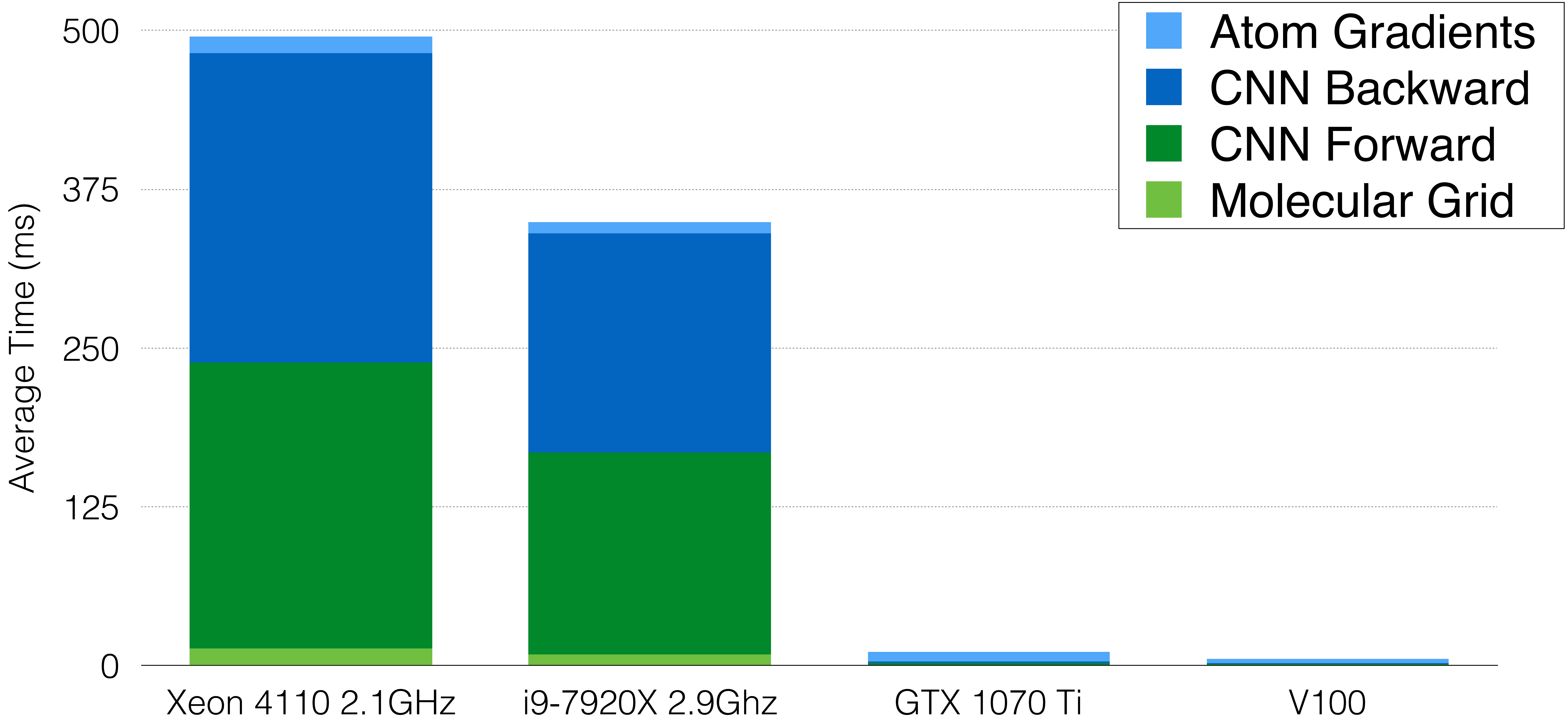
CNN

Rescoring

CNN
pose
affinity

Full CNN Docking

GPU Performance



Prospective Evaluation: D3R

Grand Challenge 3

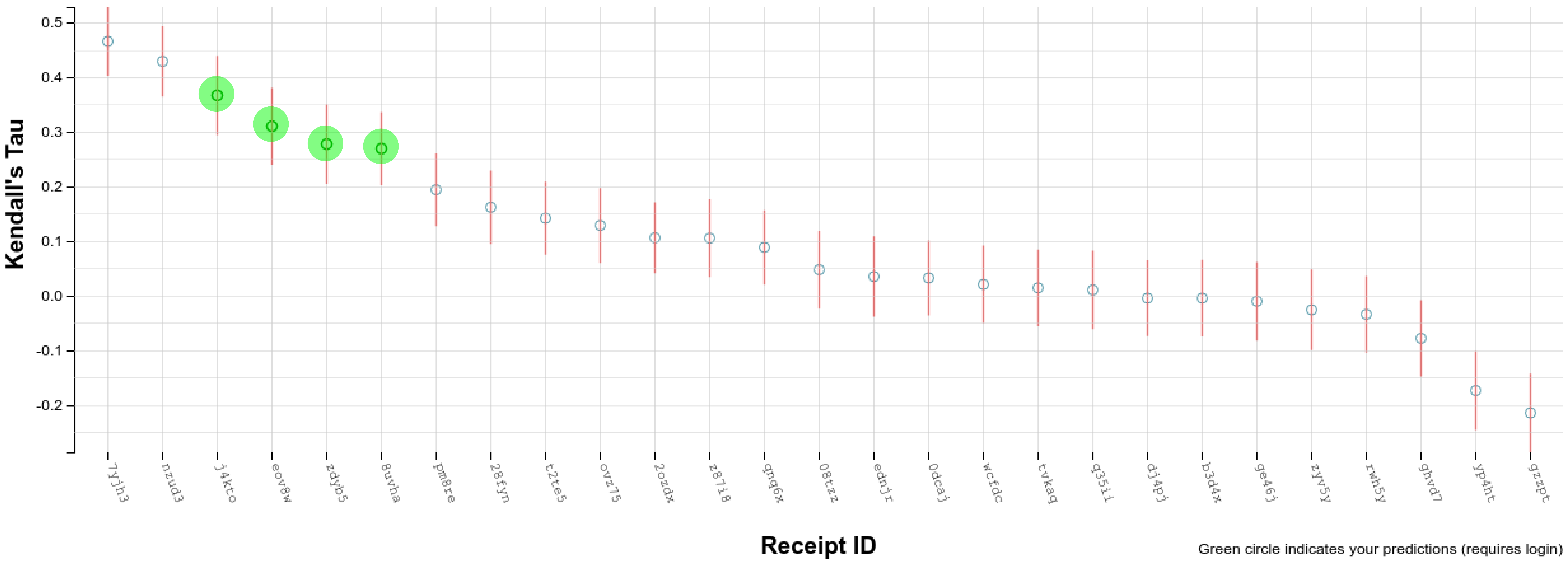
Spearman Correlation

	cnn_docked_affinity	cnn_rescore_affinity	cnn_docked_scoring	cnn_rescore_scoring	vina
cat	0.0701	0.154	-0.0351	0.178	**0.179**
p38a	-0.0784	-0.116	-0.329	-0.305	**-0.0631**
vegfr2	0.366	**0.484**	0.434	0.448	0.414
jak2	**0.428**	0.338	0.39	0.27	0.106
jak2_sub3	**0.68**	0.369	-0.372	0.159	-0.633
tie2	0.648	**0.835**	0.136	-0.078	0.561
abl1	0.634	**0.745**	0.005	0.182	0.713

Grand Challenge 3: The Good

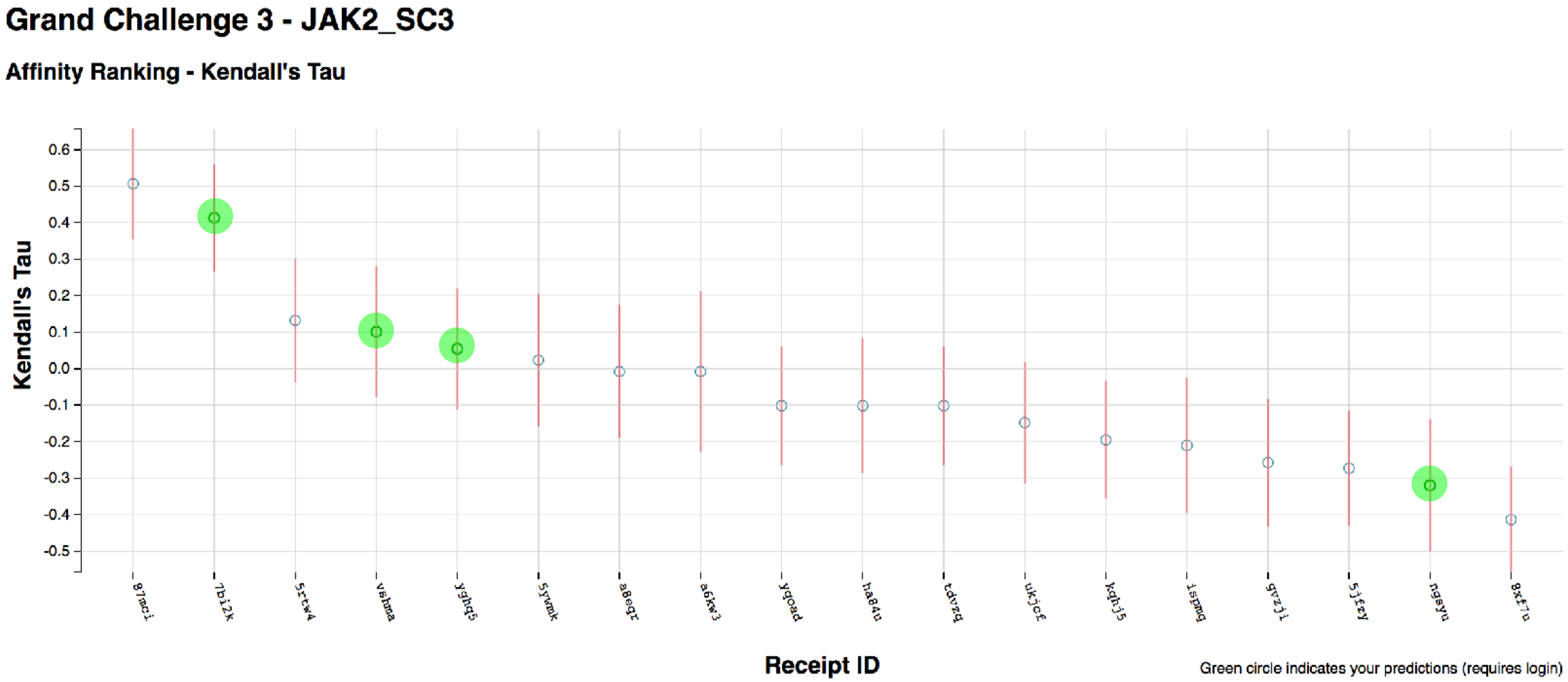
Grand Challenge 3 - JAK2_SC2

Affinity Ranking - Kendall's Tau



Green circle indicates your predictions (requires login)

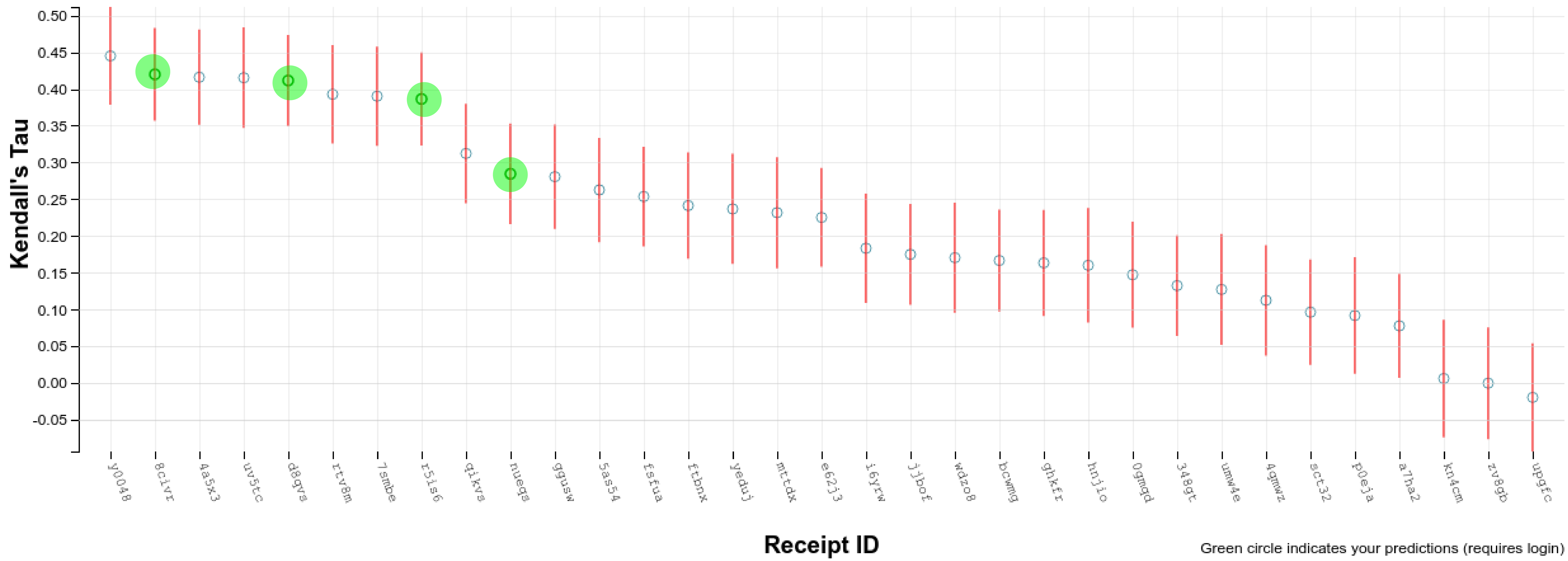
Grand Challenge 3: The Good



Grand Challenge 3: The Good

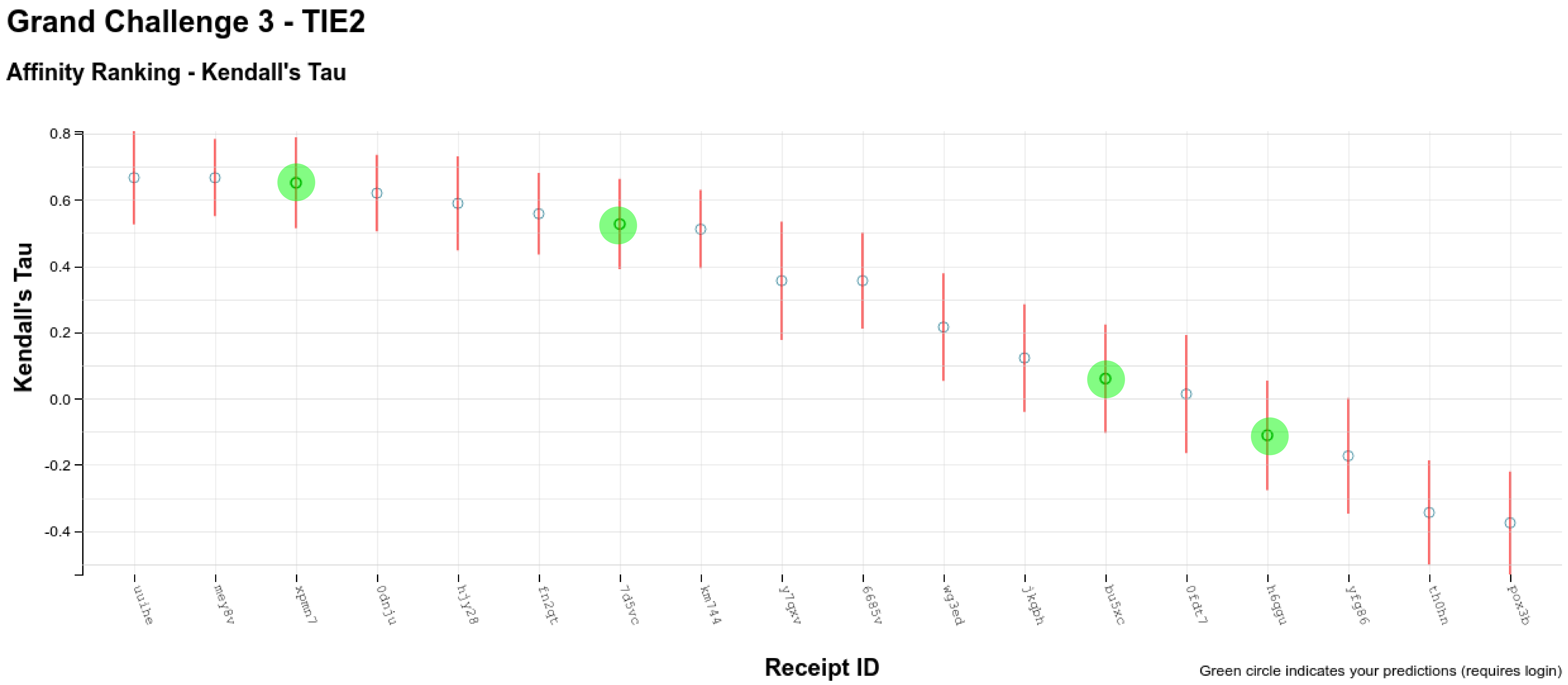
Grand Challenge 3 - VEGFR2

Affinity Ranking - Kendall's Tau



Green circle indicates your predictions (requires login)

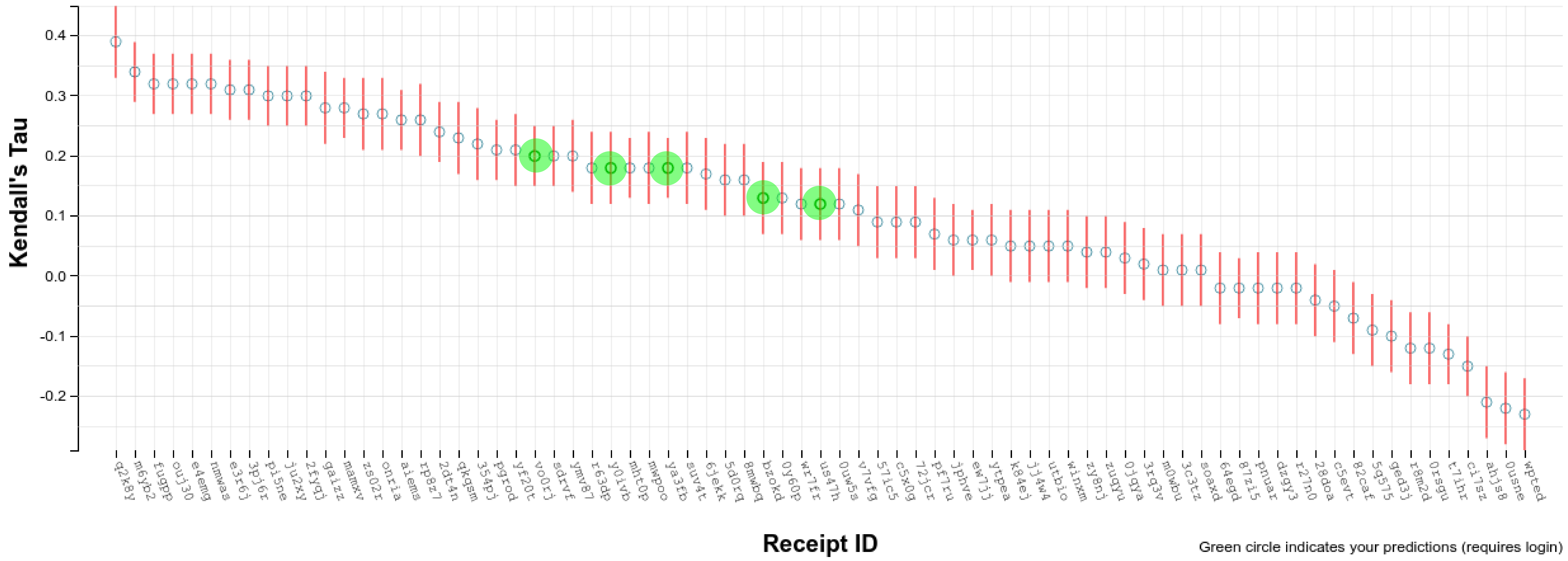
Grand Challenge 3: The Good



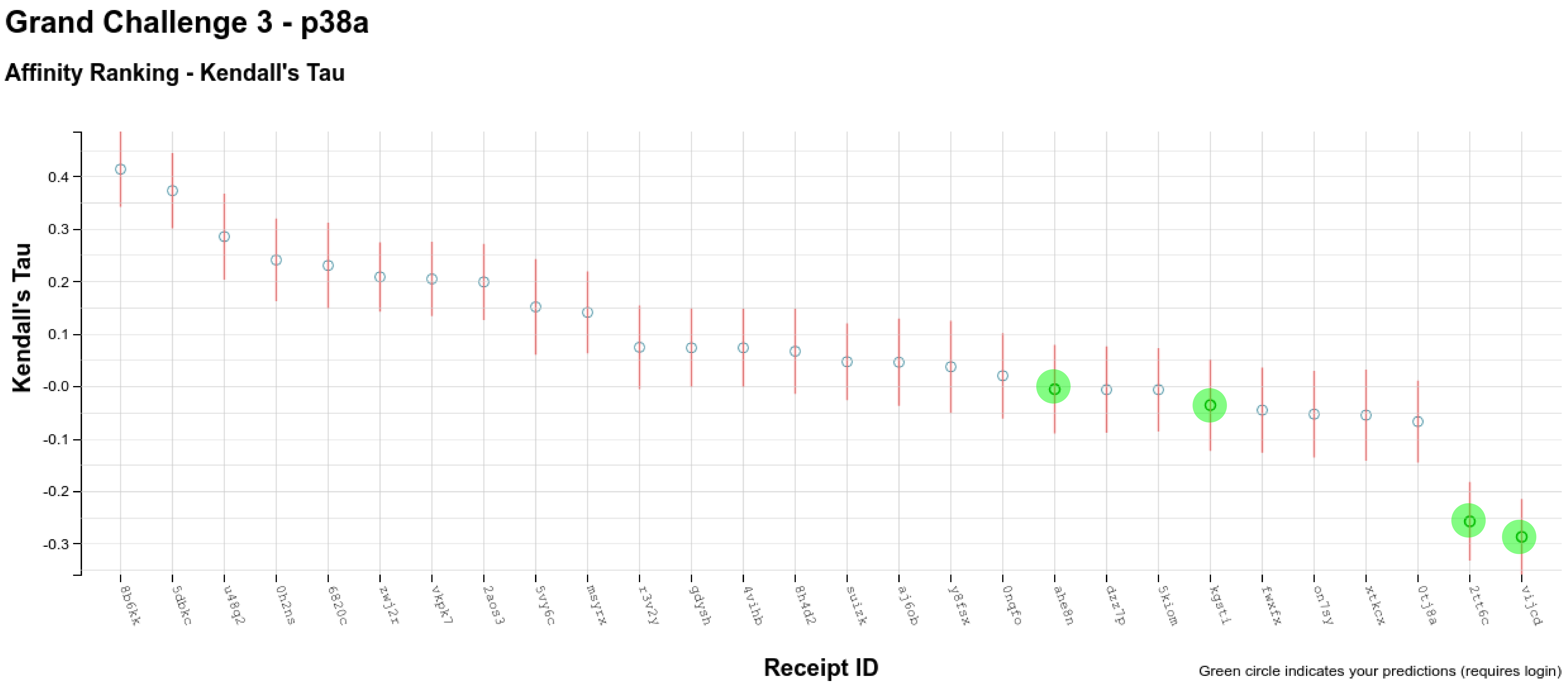
Grand Challenge 3: The Bad

Grand Challenge 3 - CatS_stage2

Affinity Ranking - Kendall's Tau

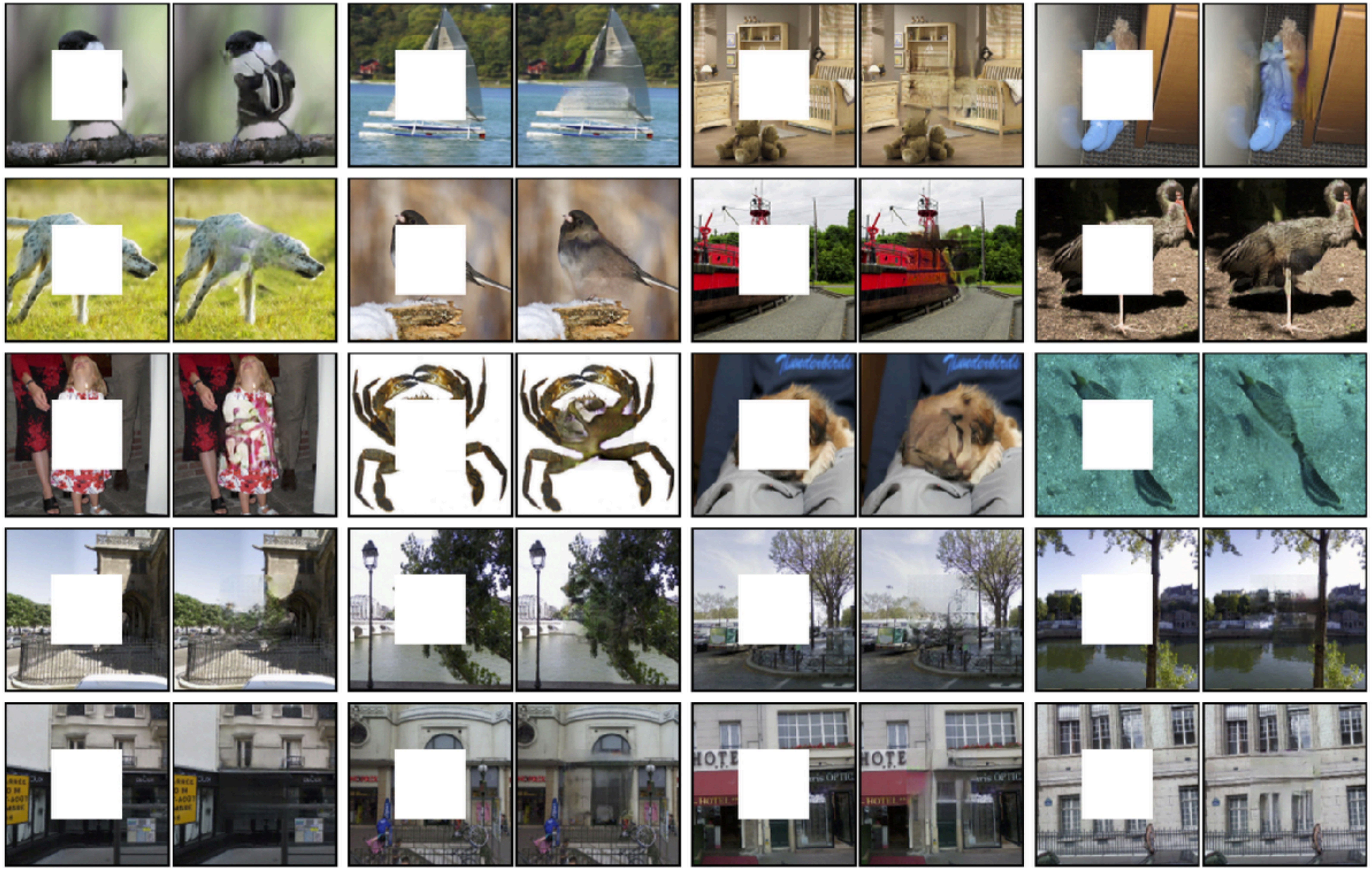


Grand Challenge 3: The Ugly



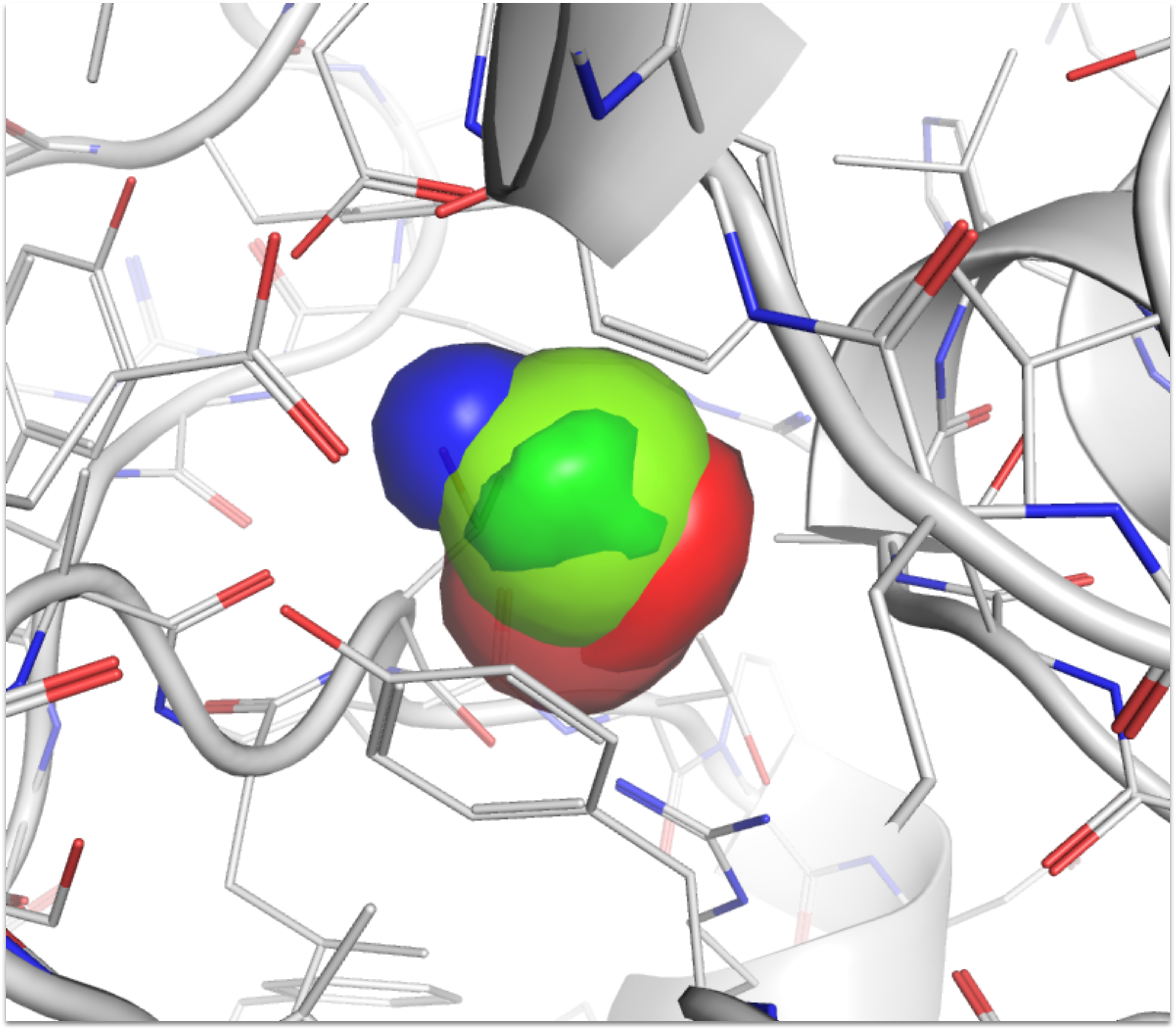
and now for something completely different…

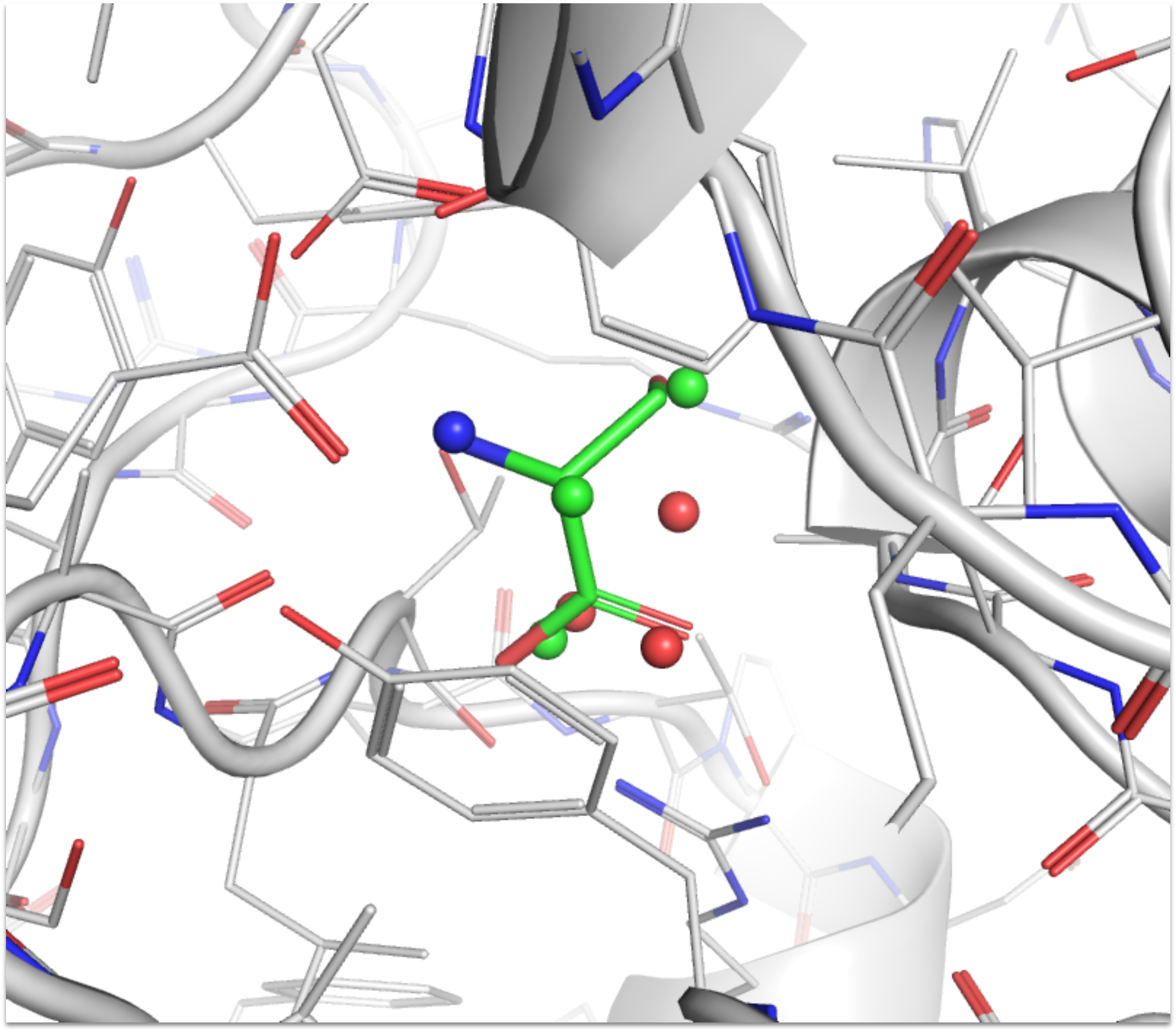
Context Encoding



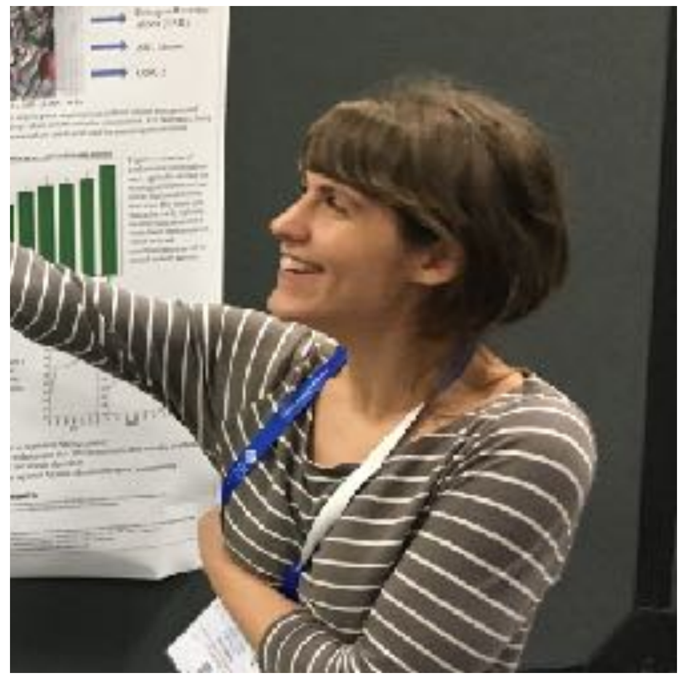
http://people.eecs.berkeley.edu/~pathak/context_encoder/

Molecular Context Encoding

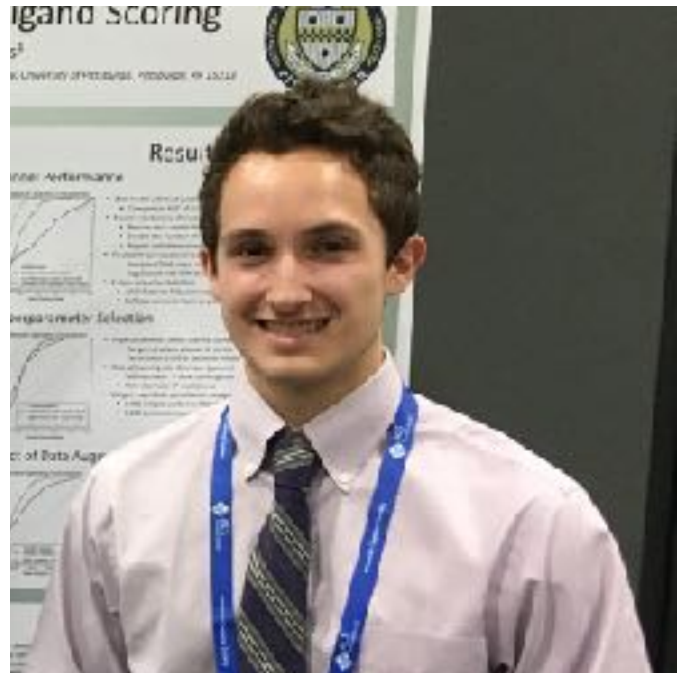




Acknowledgements



Jocelyn Sunseri



Matt Ragoza

Josh Hochuli

Lily Turner

Group Members
Jocelyn Sunseri
Jonathan King
Paul Francoeur
Matt Ragoza
Josh Hochuli
Lily Turner
Pulkit Mittal
Alec Helbling
Gibran Biswas
Sharanya Bandla
Faiha Khan

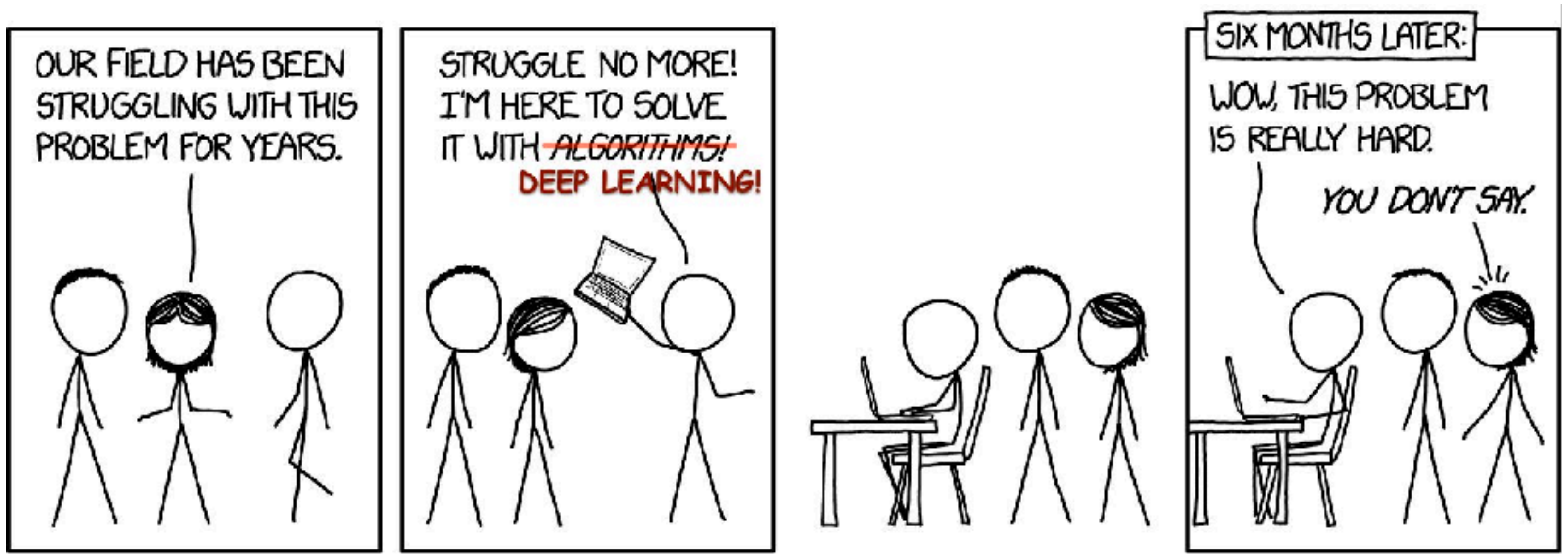
Department of Computational and Systems Biology

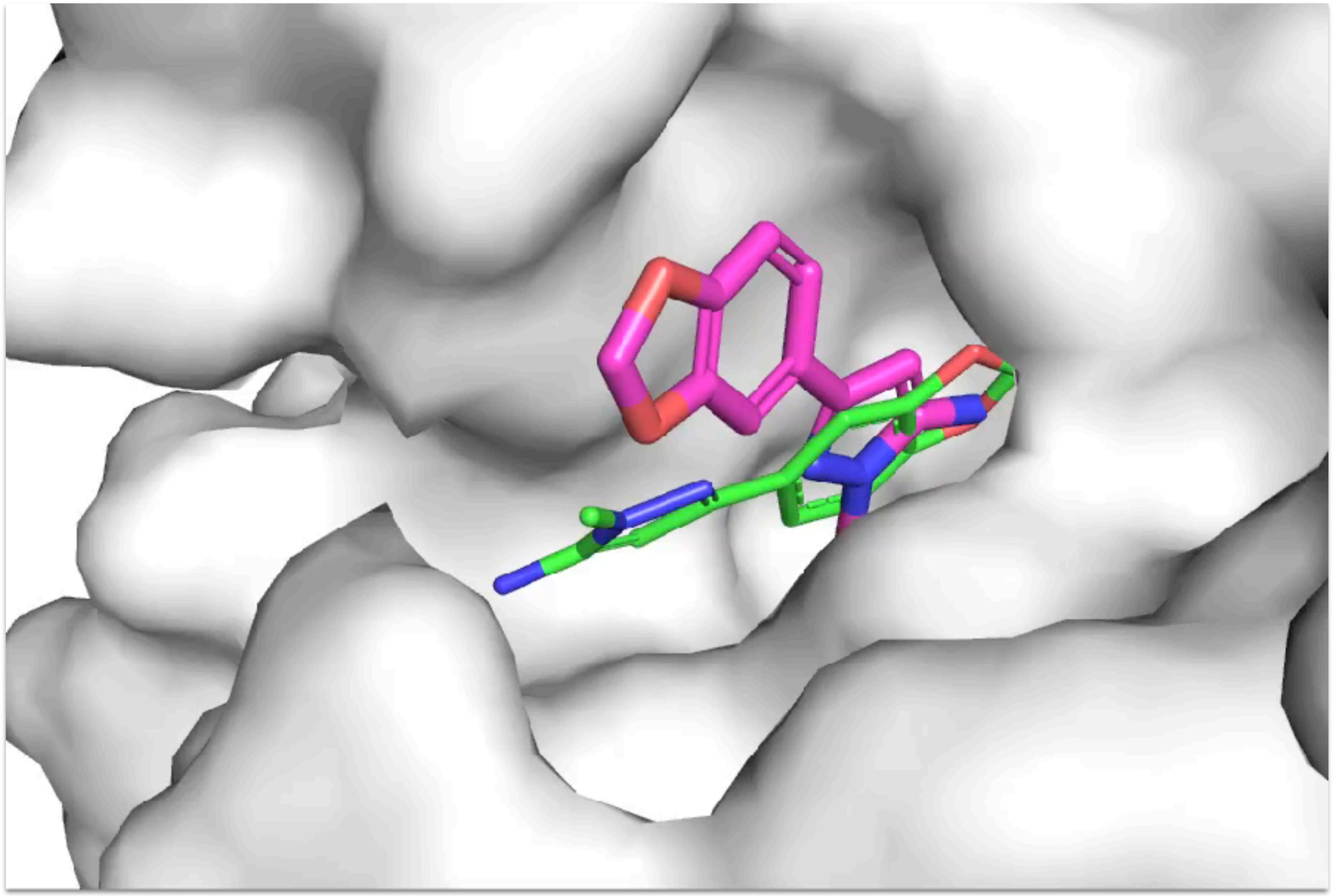
AI GRANT
National Institute of General Medical Sciences
R01GM108340
NVIDIA
TECBio REU

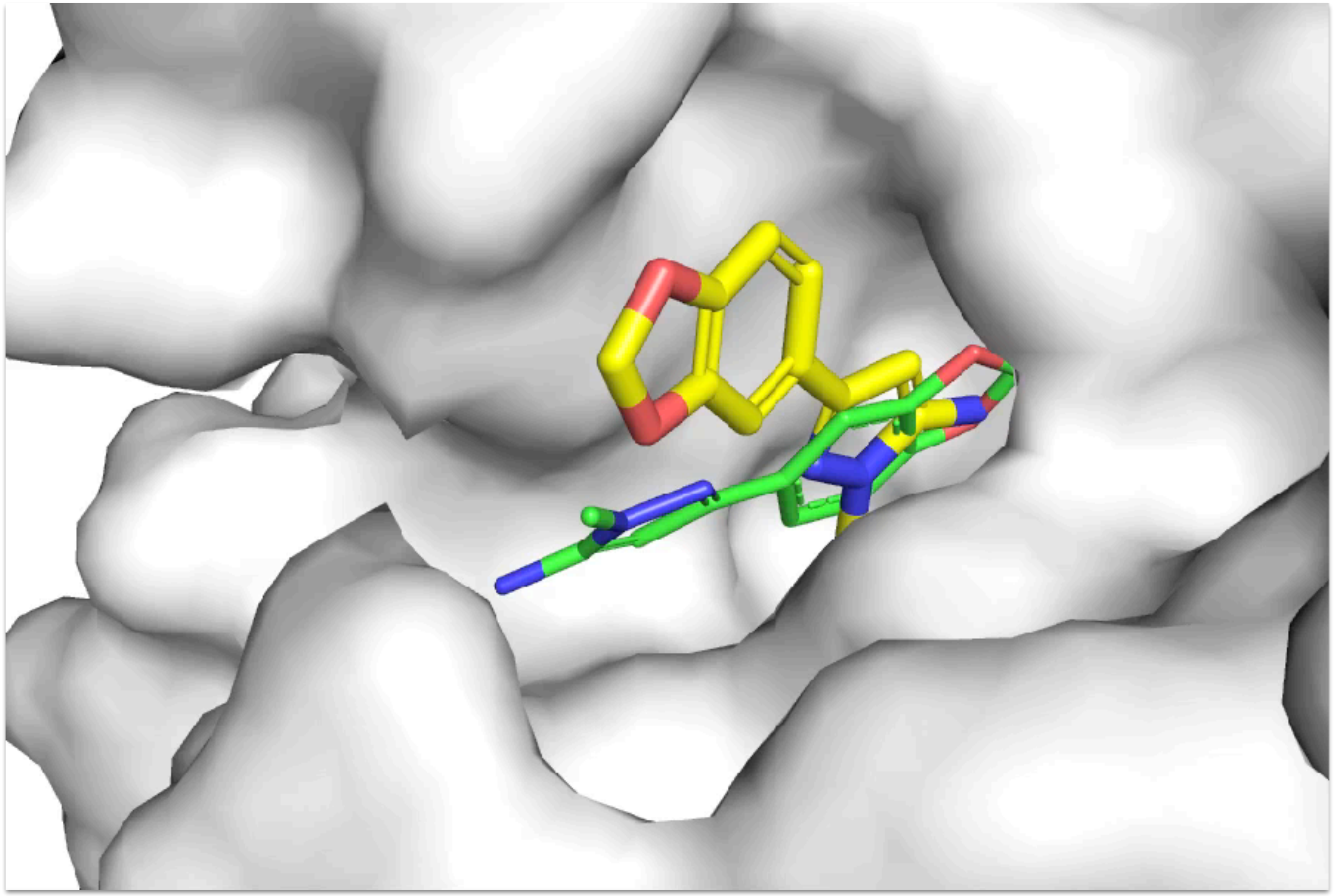
github.com/gnina

http://bits.csb.pitt.edu

@david_koes







github.com/gnina

http://bits.csb.pitt.edu

@david_koes