Protein-Ligand Scoring with Convolutional Neural Networks

David Koes

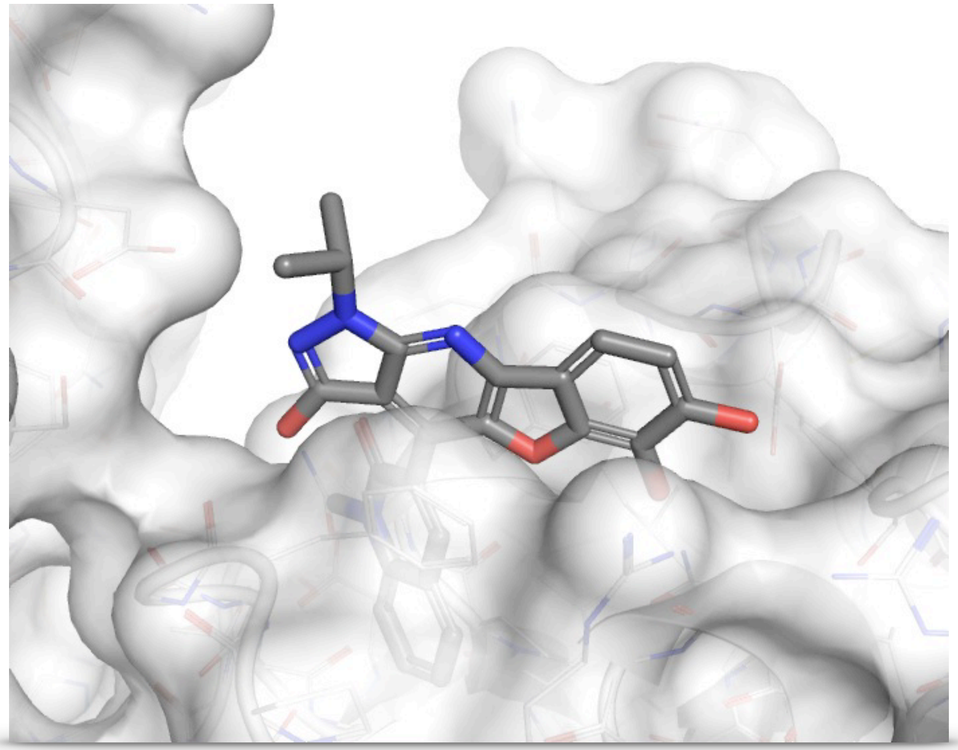
@david_koes

ACS

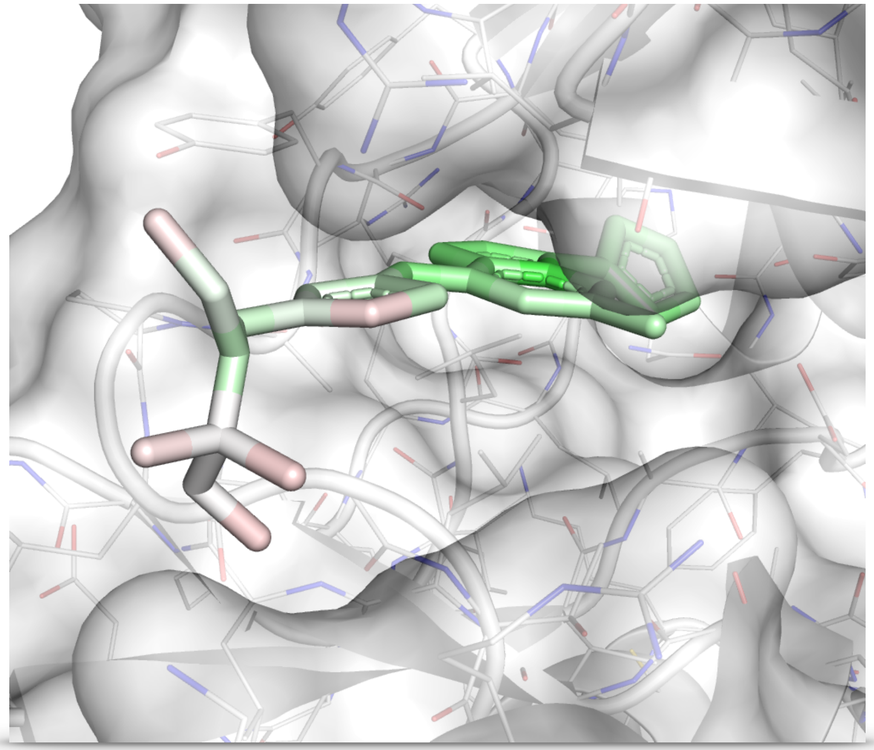
August 23, 2017

Structure Based Drug Design

Virtual Screening



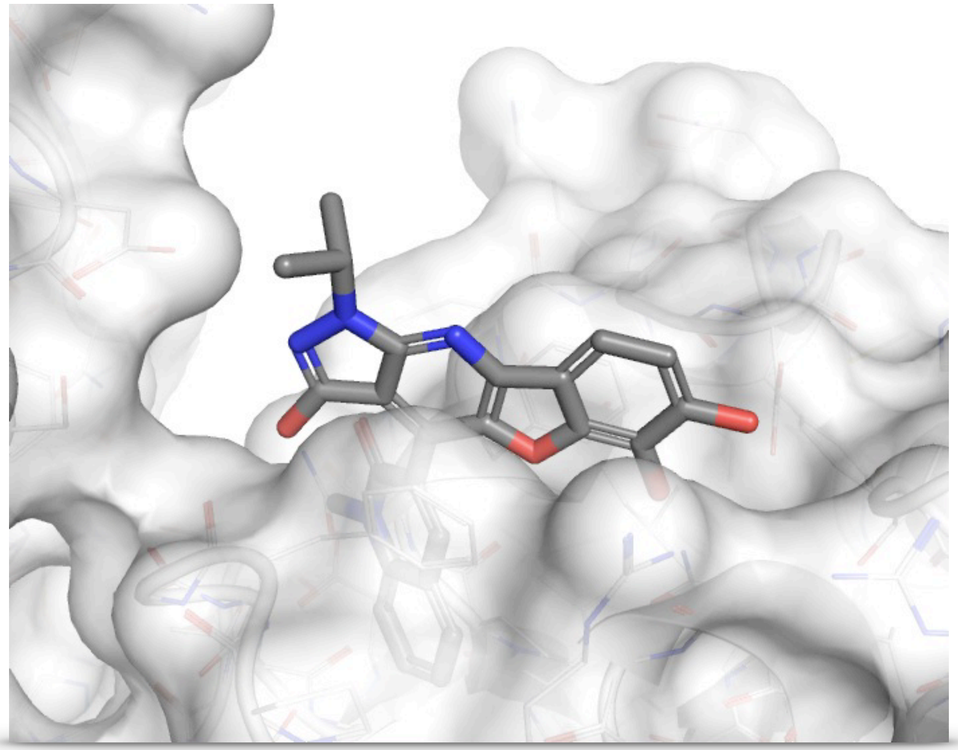
Lead Optimization



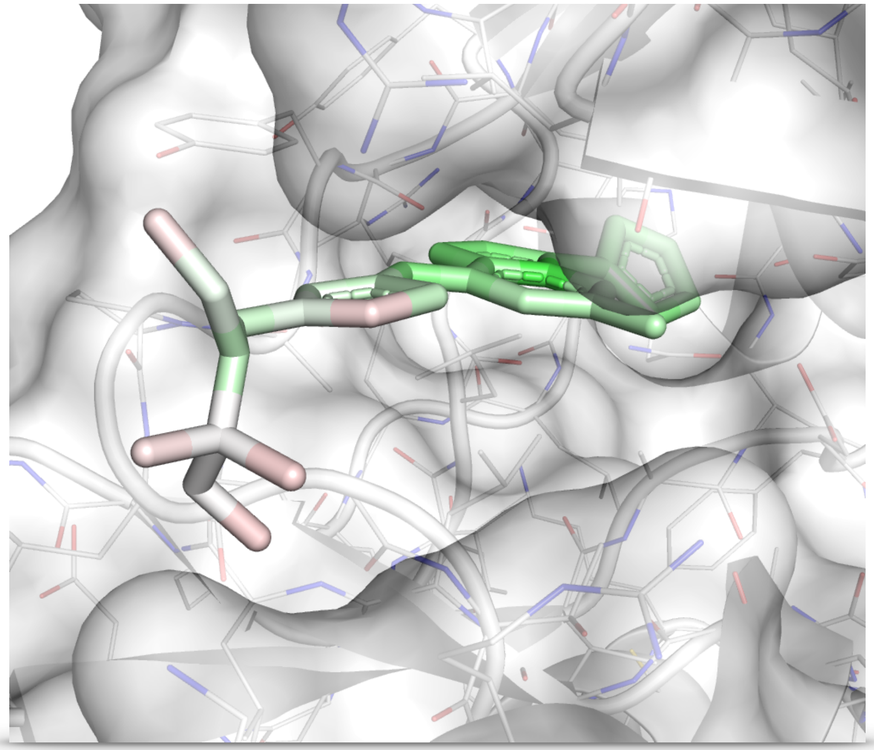
Pose Prediction → Binding Discrimination → Affinity Prediction

Structure Based Drug Design

Virtual Screening

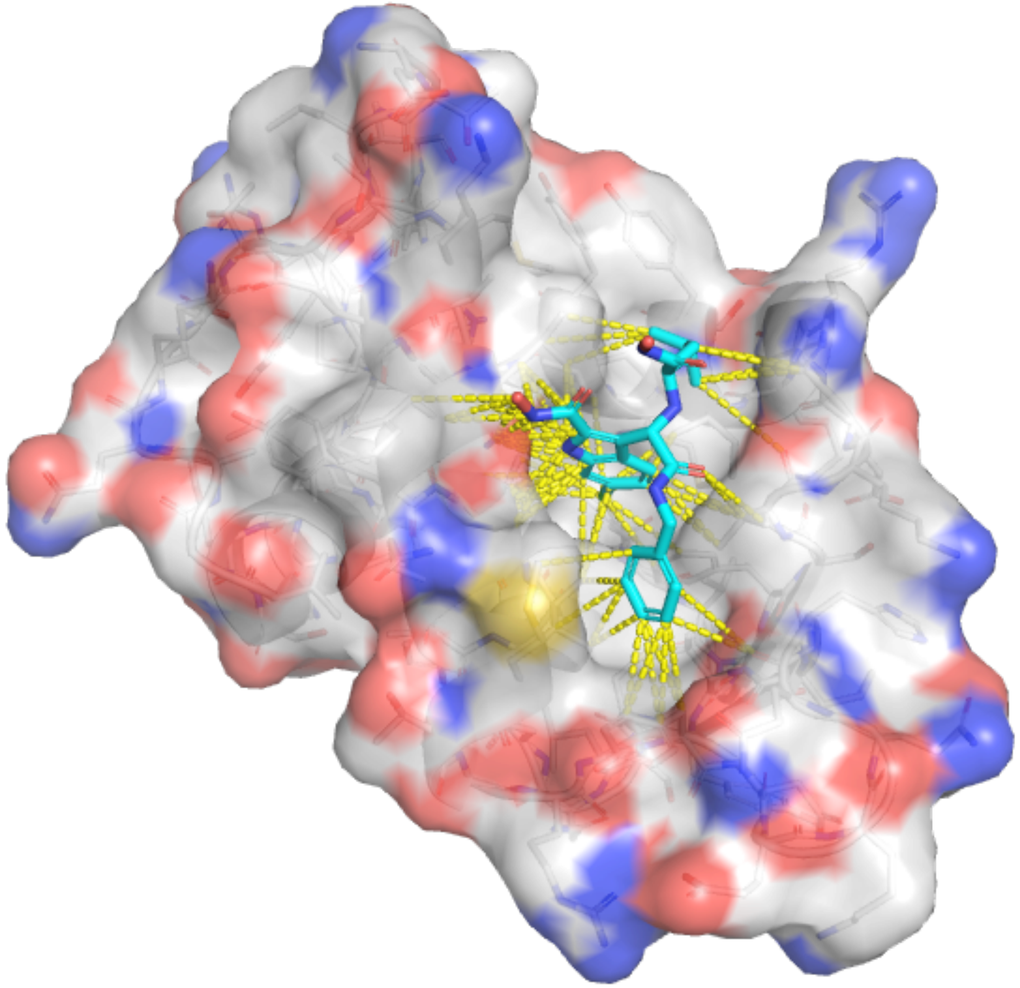


Lead Optimization



Pose Prediction → Binding Discrimination → Affinity Prediction

Protein-Ligand Scoring



AutoDock Vina

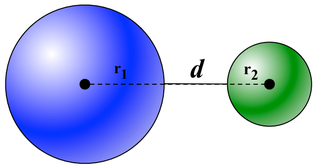
$$\text{gauss}_1(d) = w_{\text{guass}_1} e^{-(d/0.5)^2}$$

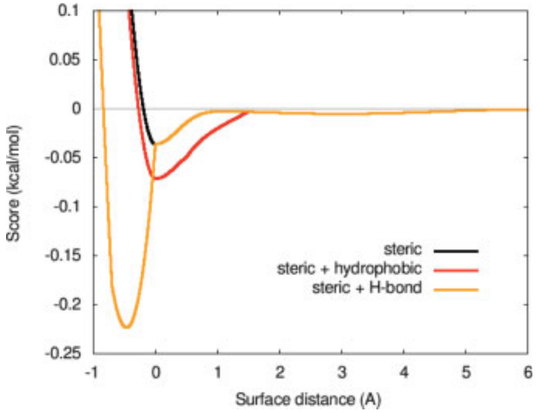
$$\text{gauss}_2(d) = w_{\text{guass}_2} e^{-((d-3)/2)^2}$$

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

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

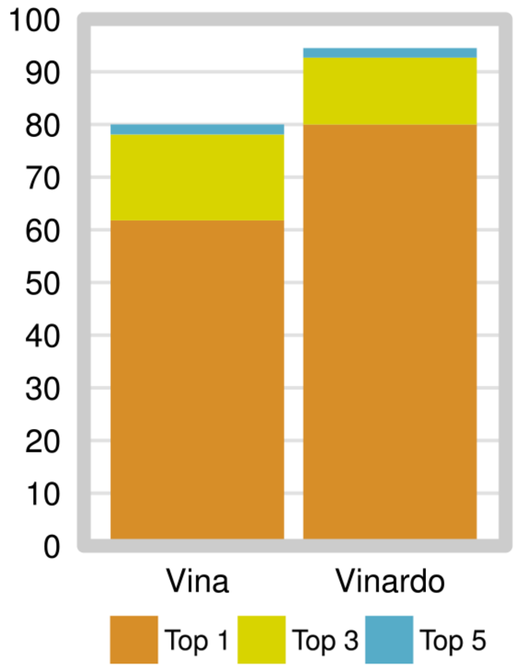
$$\text{hbond}(d) = \begin{cases} w_{\text{hbond}} & d < -0.7 \\ 0 & d > 0 \\ w_{\text{hbond}}(-\frac{10}{7} d) & otherwise \end{cases}$$

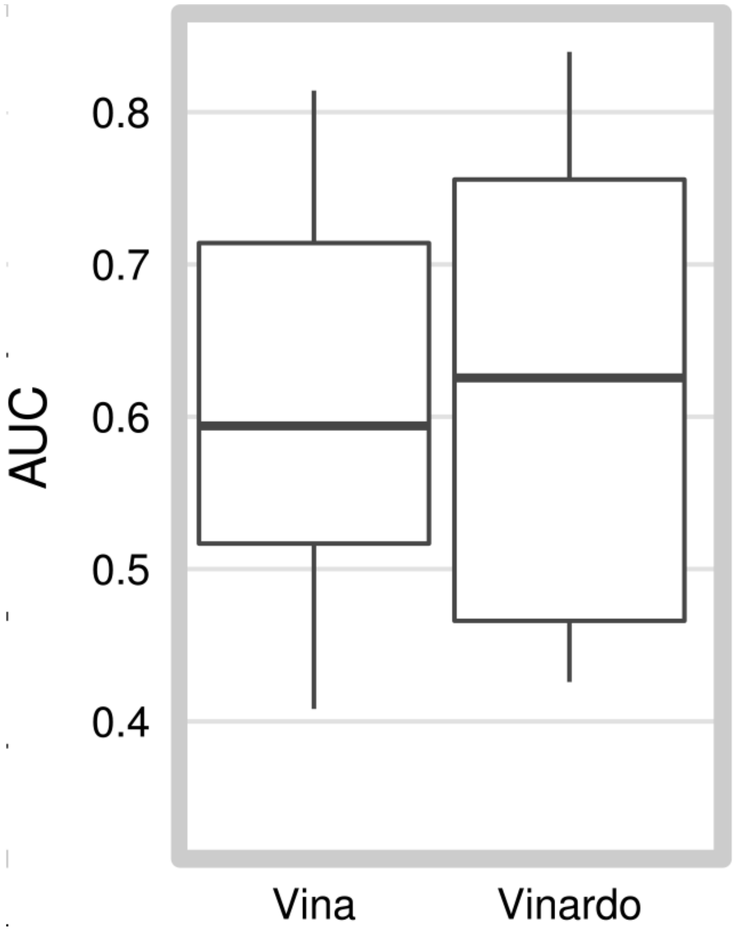


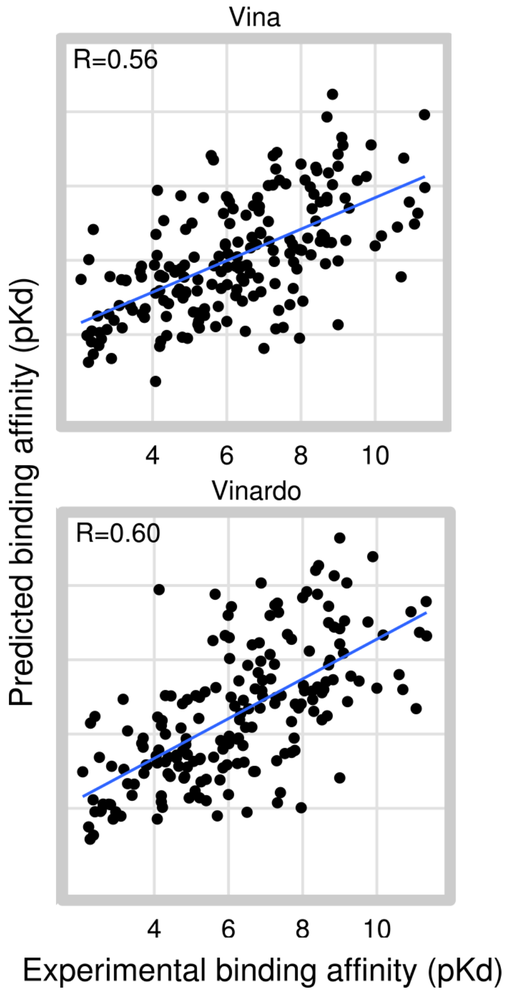


O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry* 31 (2010) 455-461

State of the Art







Pose Prediction Binding Discrimination Affinity Prediction

Quiroga R, Villarreal MA (2016) Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PLoS ONE* 11(5): e0155183. doi:10.1371/journal.pone.0155183

Can we do better?

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

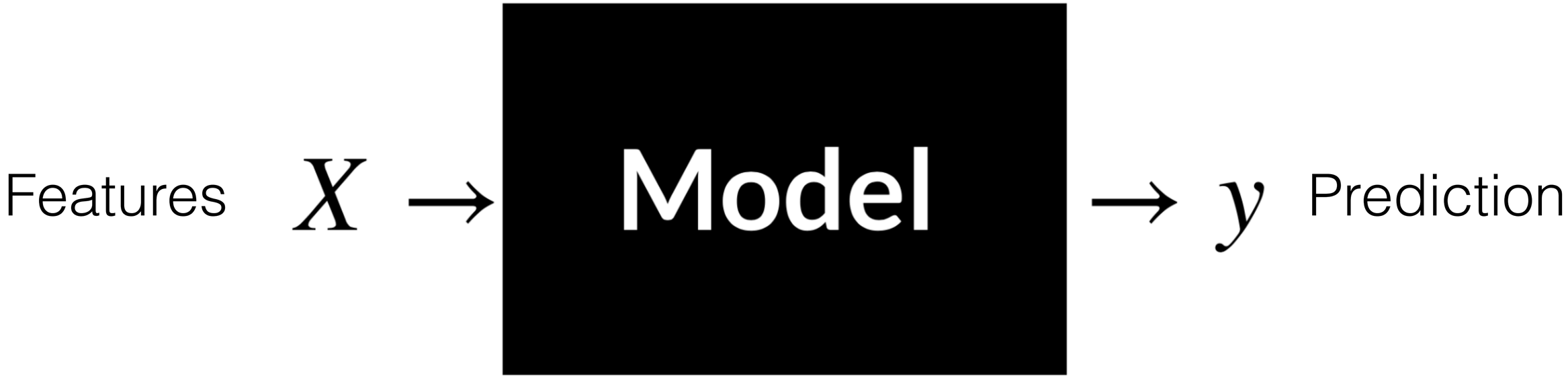
Can we do better?

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

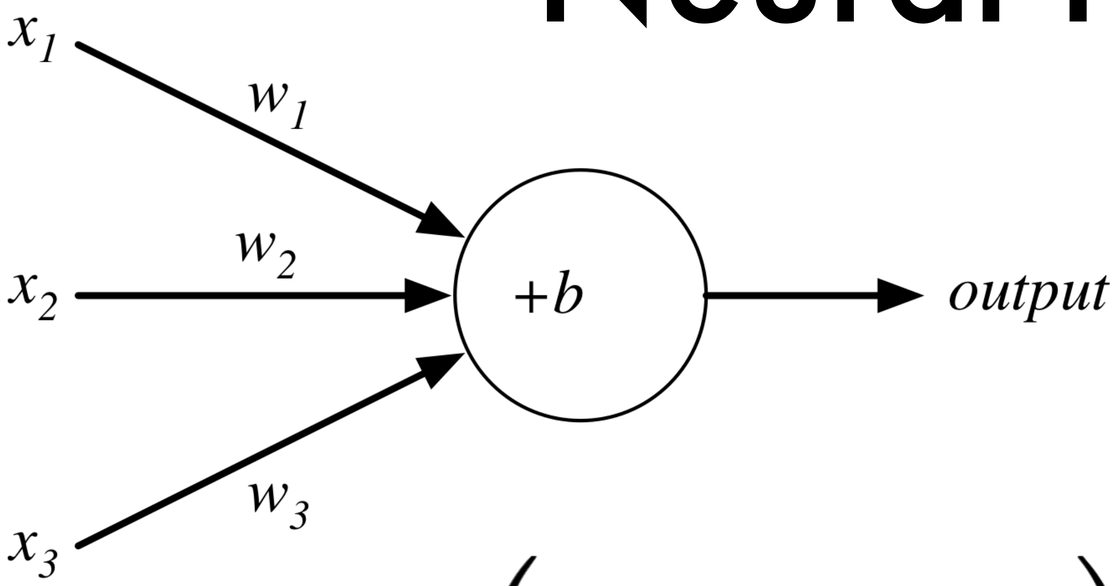
Key Idea: Leverage "big data"

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

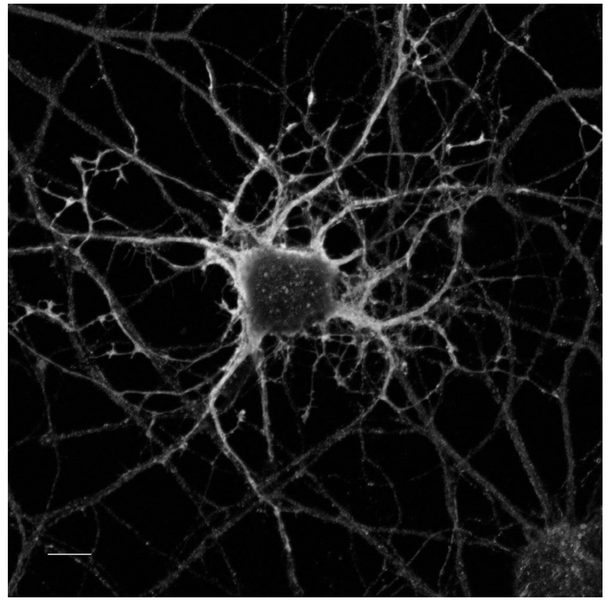
Machine Learning

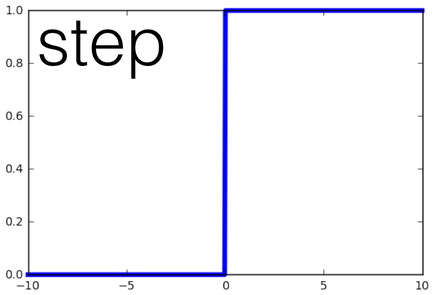


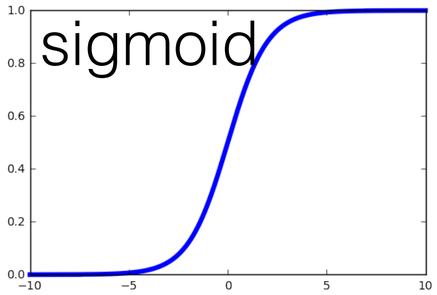
Neural Networks

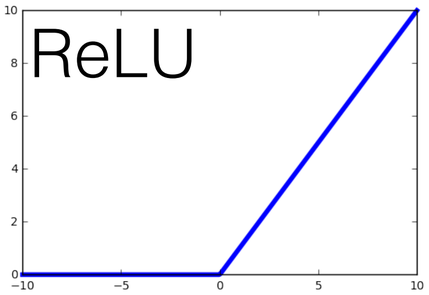


$$output = \sigma\left(\sum_i w_i x_i + b\right)$$

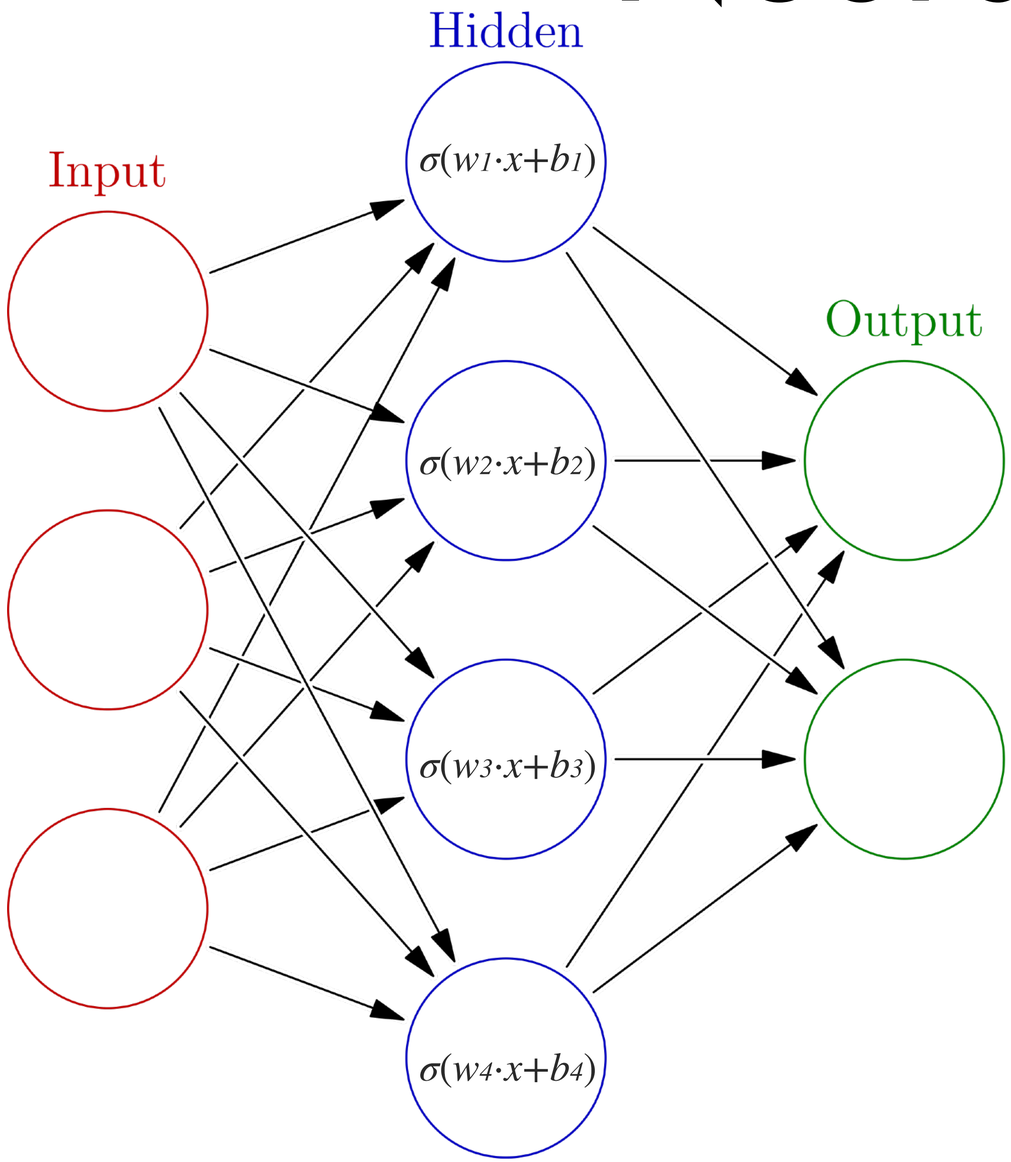


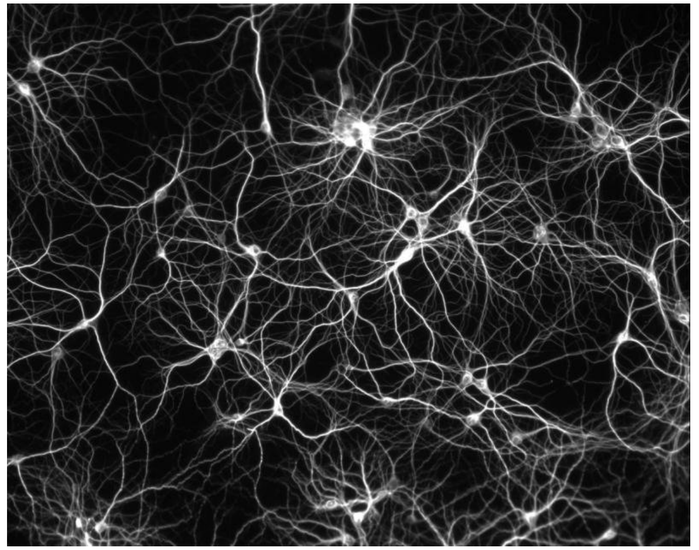






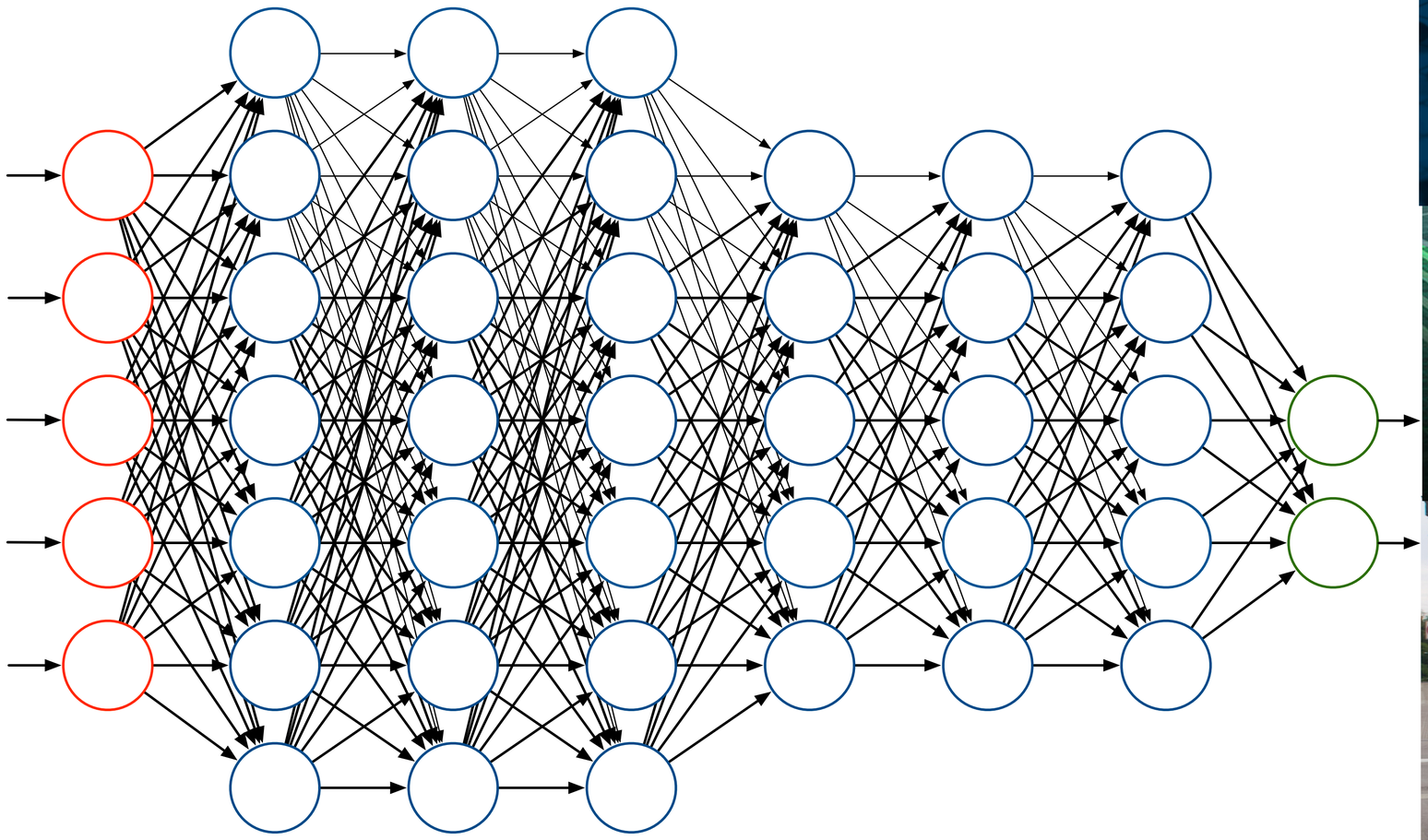
Neural Networks

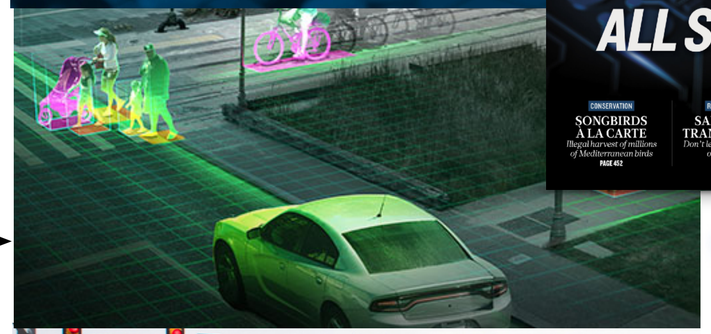




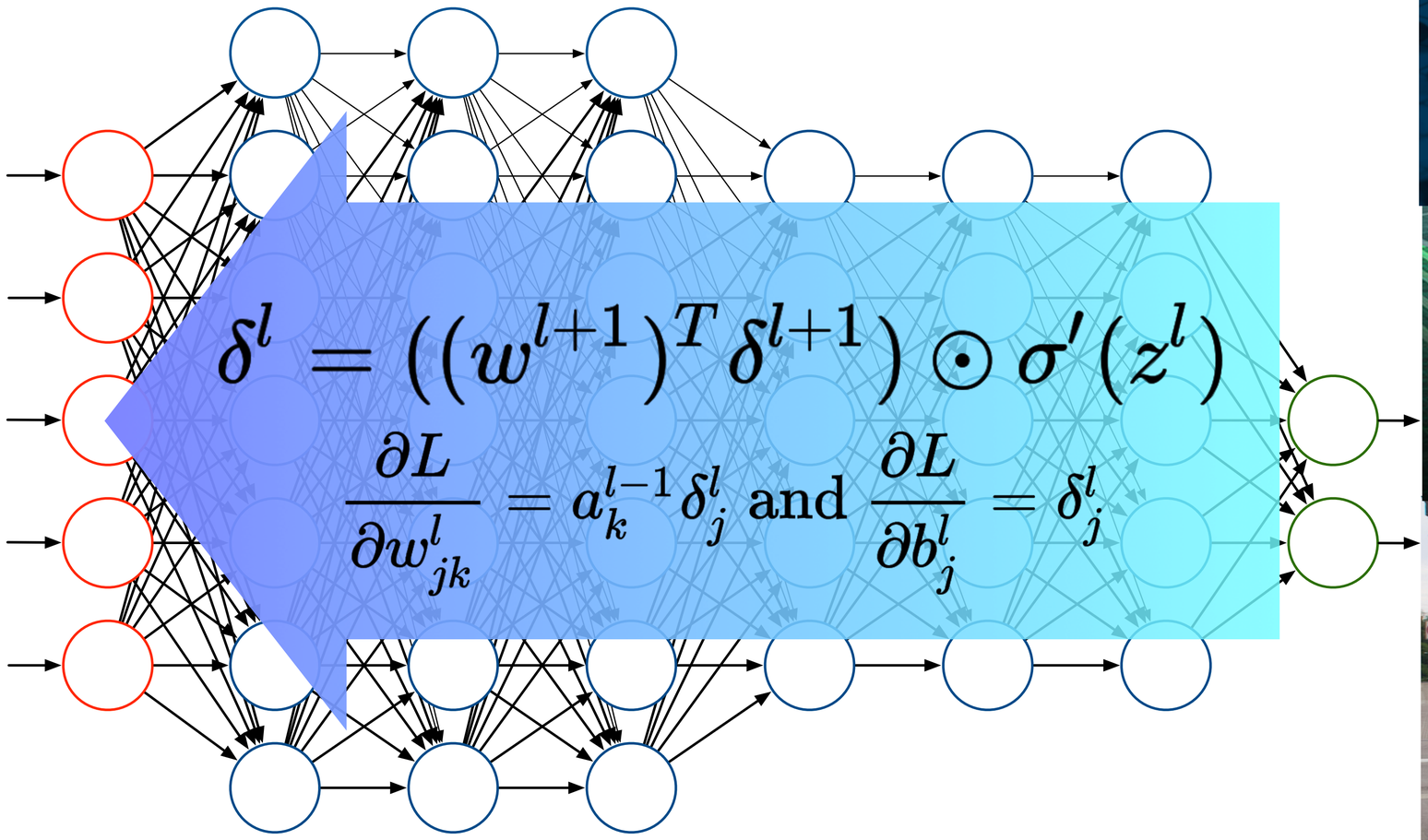
The **universal approximation theorem** states that, under reasonable assumptions, a feedforward **neural network** with a finite number of nodes **can approximate any continuous** function to within a given error over a bounded input domain.

Deep Learning





Deep Learning



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

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

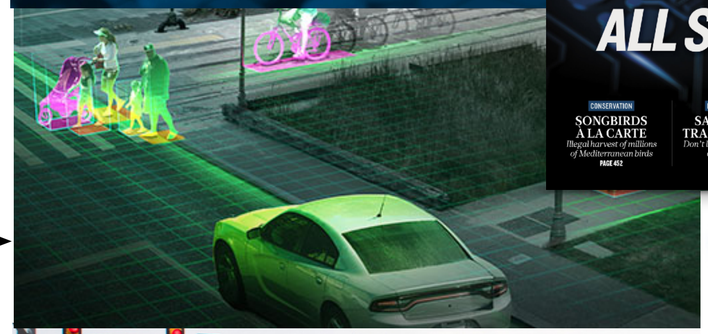
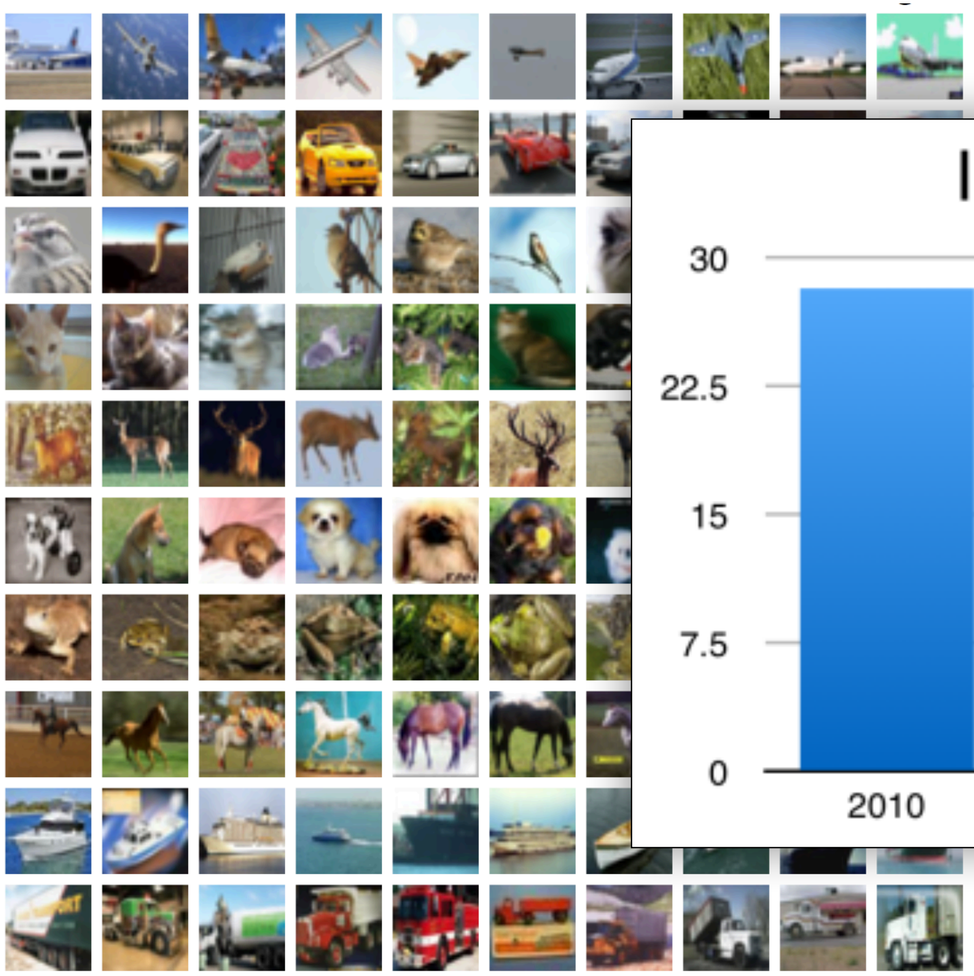
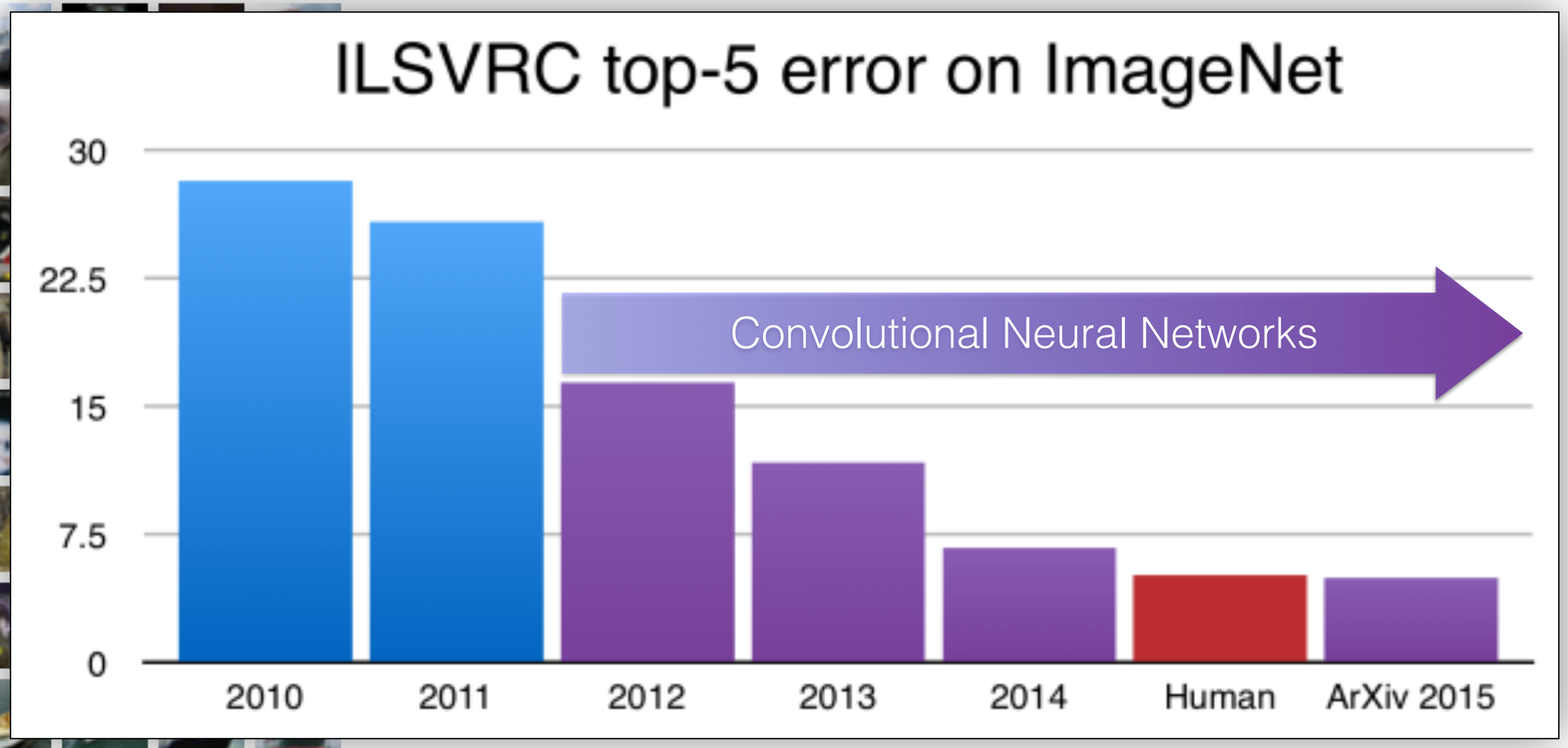


Image Recognition

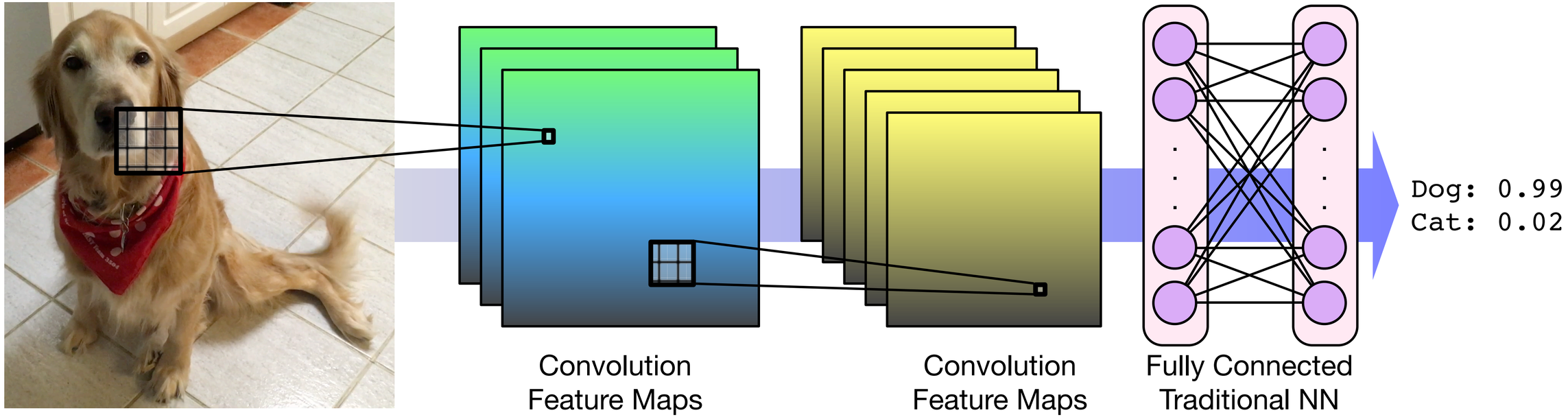
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

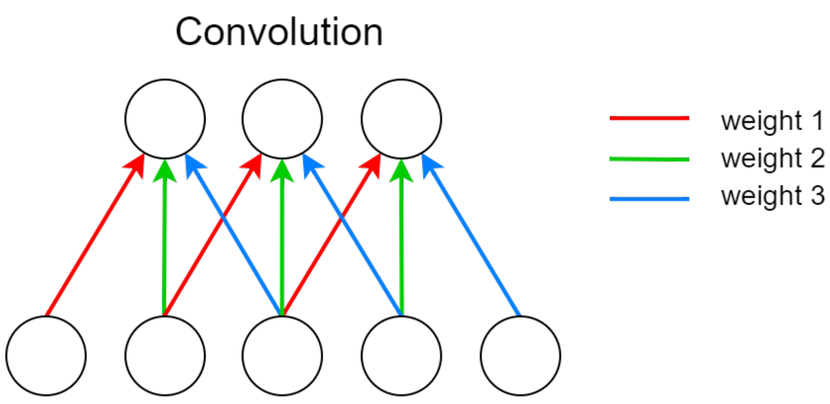


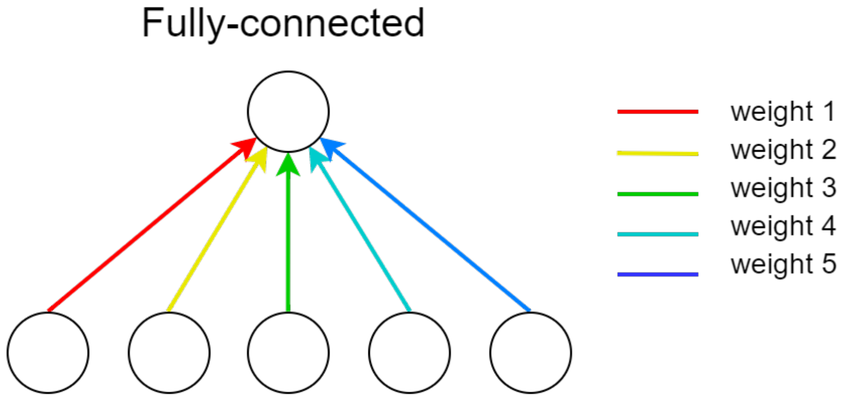


https://devblogs.nvidia.com

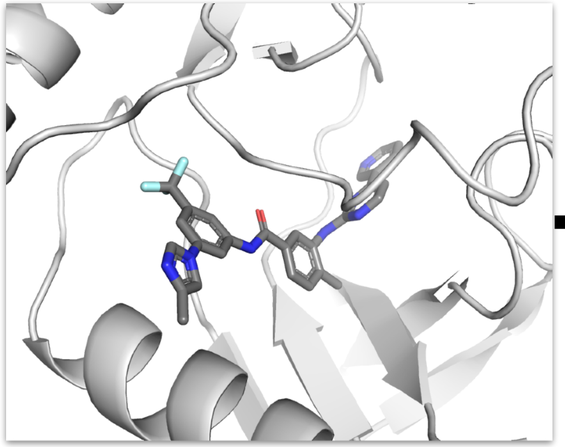
Convolutional Neural Networks







CNNs for Protein-Ligand Scoring



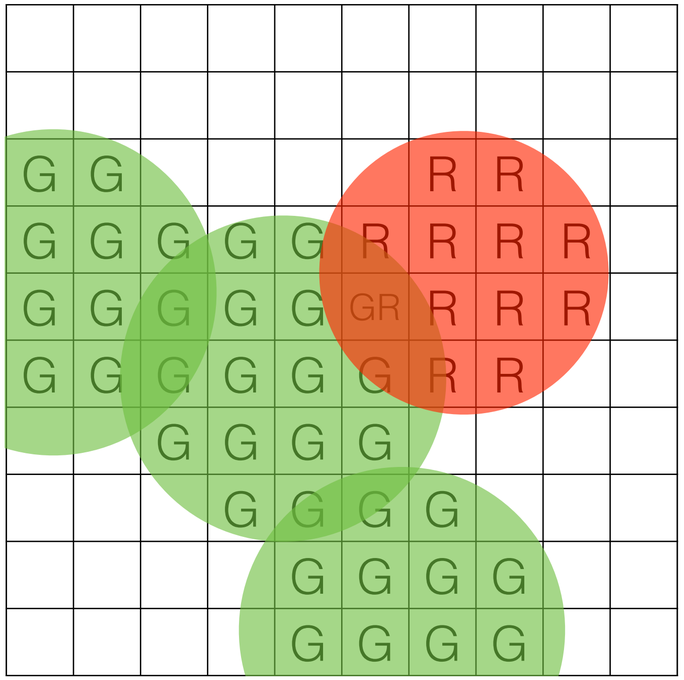
CNN

Pose Prediction

Binding Discrimination

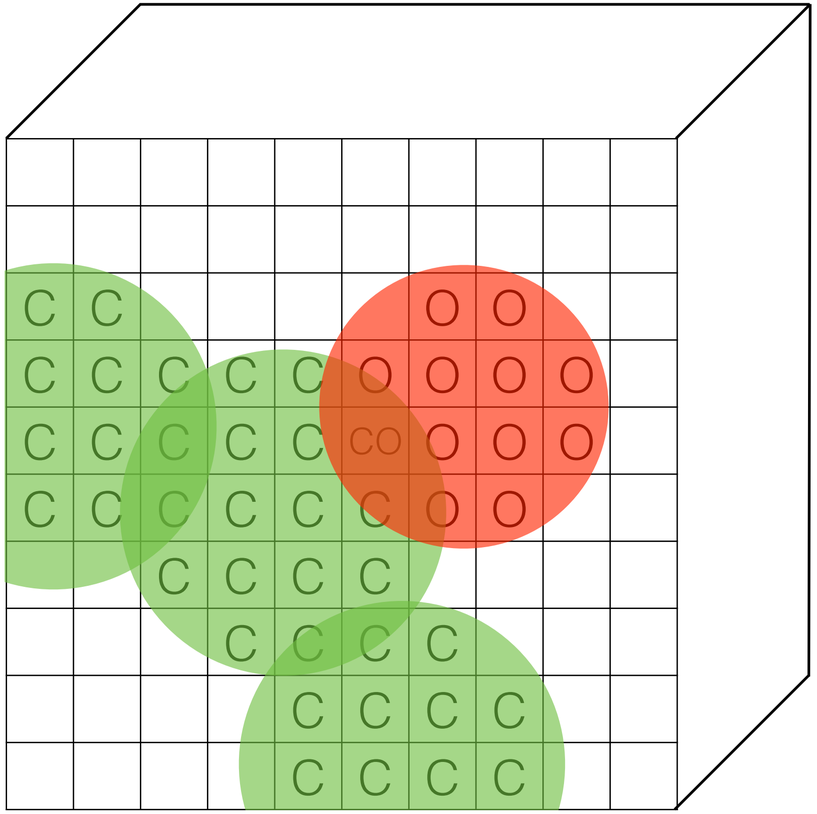
Affinity Prediction

Protein-Ligand Representation



(R,G,B) pixel

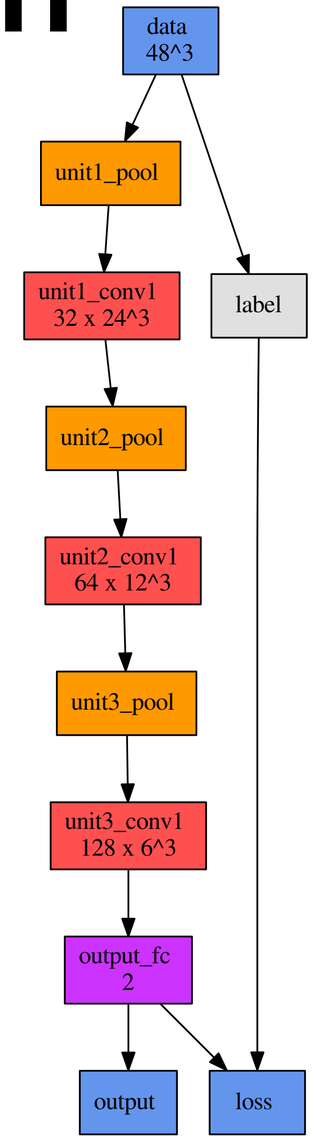
Protein-Ligand Representation



(R,G,B) pixel →

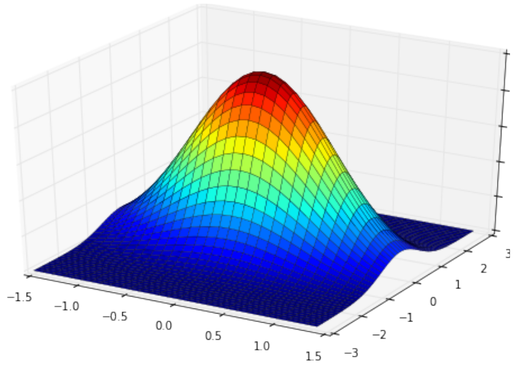
(Carbon, Nitrogen, Oxygen,…) **voxel**

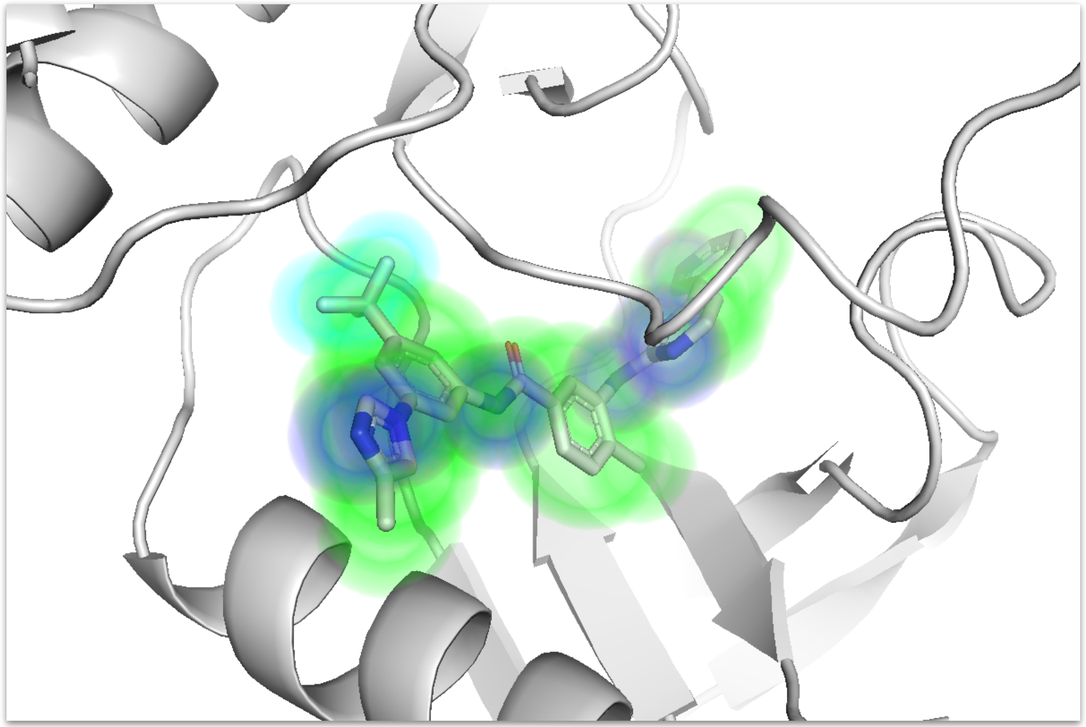
The only parameters for this representation are the choice of **grid resolution**, **atom density**, and **atom types**.



Atom Density

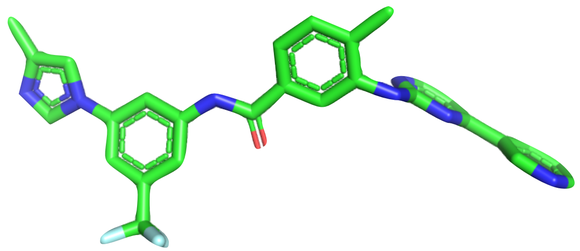
$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}} & 0 \leq d < r \\ \frac{4}{e^2 r^2} d^2 - \frac{12}{e^2 r} d + \frac{9}{e^2} & r \leq d < 1.5r \\ 0 & d \geq 1.5r \end{cases}$$





Gaussian

Atom Types

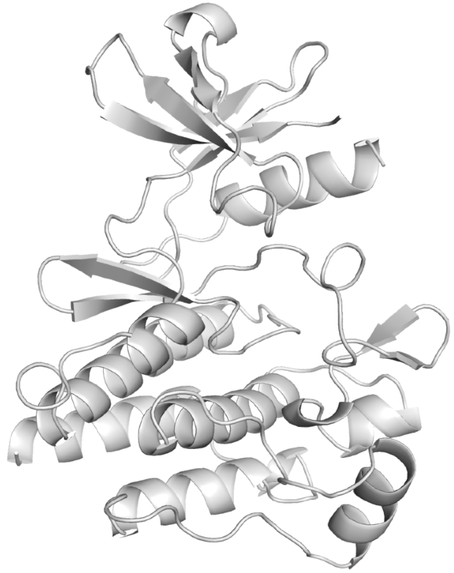


Ligand

AliphaticCarbonXSHydrophobe
AliphaticCarbonXSNonHydrophobe
AromaticCarbonXSHydrophobe
AromaticCarbonXSNonHydrophobe
Bromine
Chlorine
Fluorine
Iodine
Nitrogen
NitrogenXSAcceptor
NitrogenXSDonor
NitrogenXSDonorAcceptor
Oxygen
OxygenXSAcceptor
OxygenXSDonorAcceptor
Phosphorus
Sulfur
SulfurAcceptor

Receptor

AliphaticCarbonXSHydrophobe
AliphaticCarbonXSNonHydrophobe
AromaticCarbonXSHydrophobe
AromaticCarbonXSNonHydrophobe
Calcium
Iron
Magnesium
Nitrogen
NitrogenXSAcceptor
NitrogenXSDonor
NitrogenXSDonorAcceptor
OxygenXSAcceptor
OxygenXSDonorAcceptor
Phosphorus
Sulfur
Zinc



Training Data

Pose Prediction

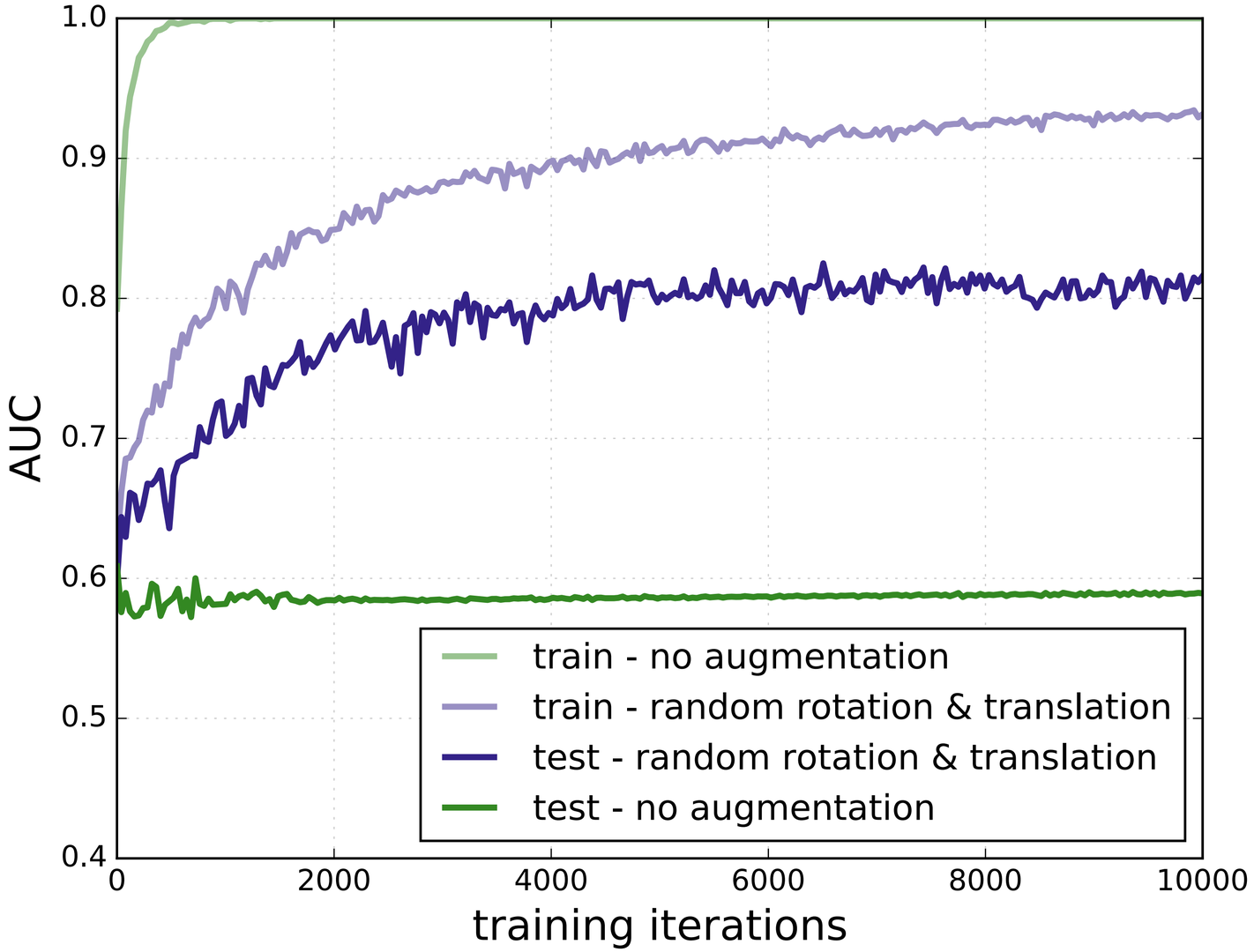
337 protein-ligand complexes

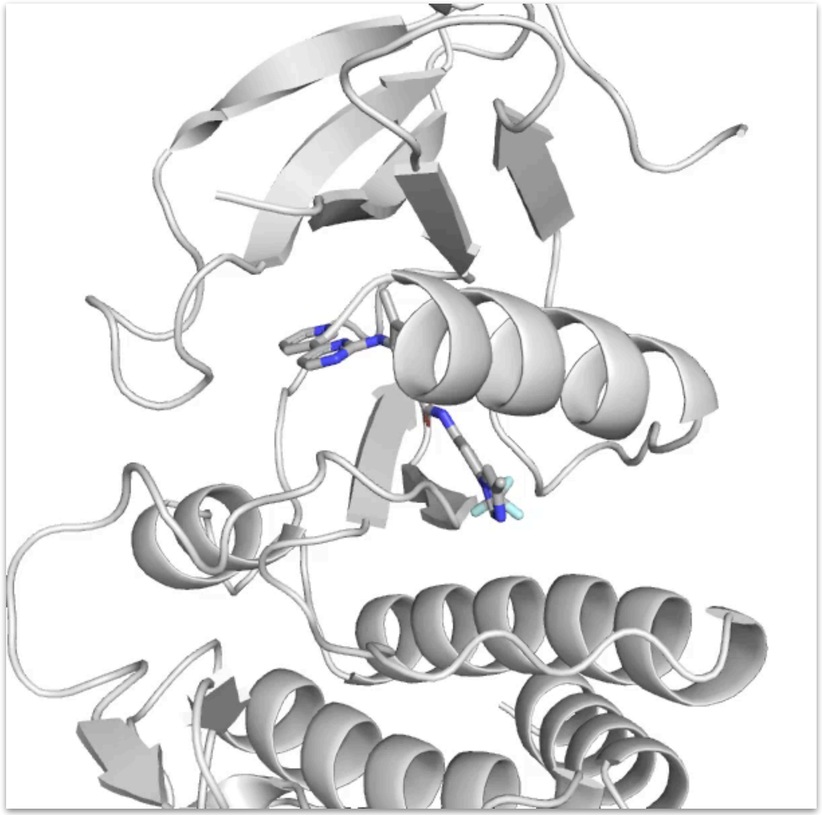
- curated for electron density
- diverse targets
- <10μM affinity
- **generate poses** with Vina
  - 745 <2Å RMSD (actives)
  - 3251 >4Å RMSD (decoys)

4056 protein-ligand complexes

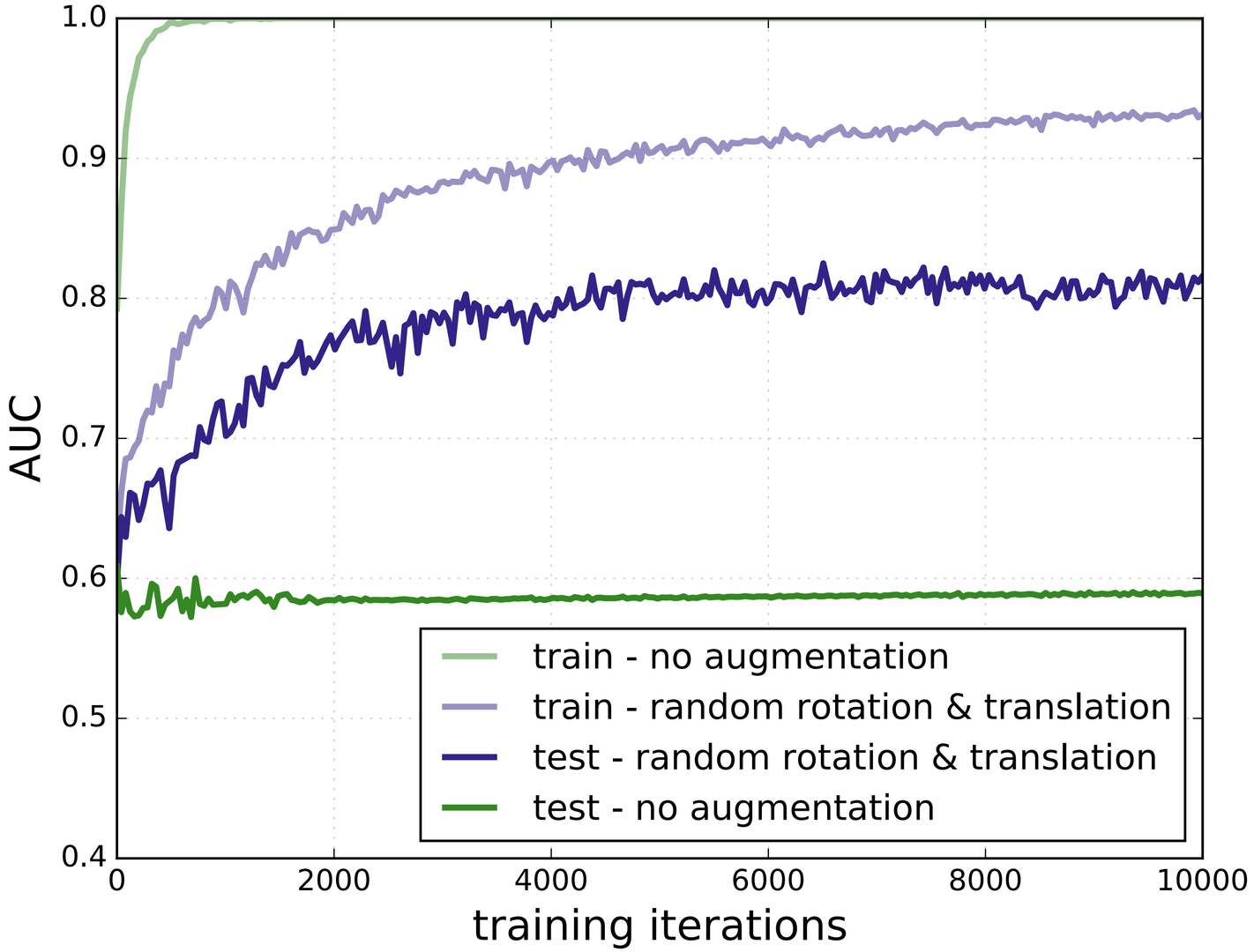
- diverse targets
- wide range of affinities
- **generate poses** with AutoDock Vina
- include minimized crystal pose
  - 8,688 <2Å RMSD (actives)
  - 76,743 >4Å RMSD (decoys)

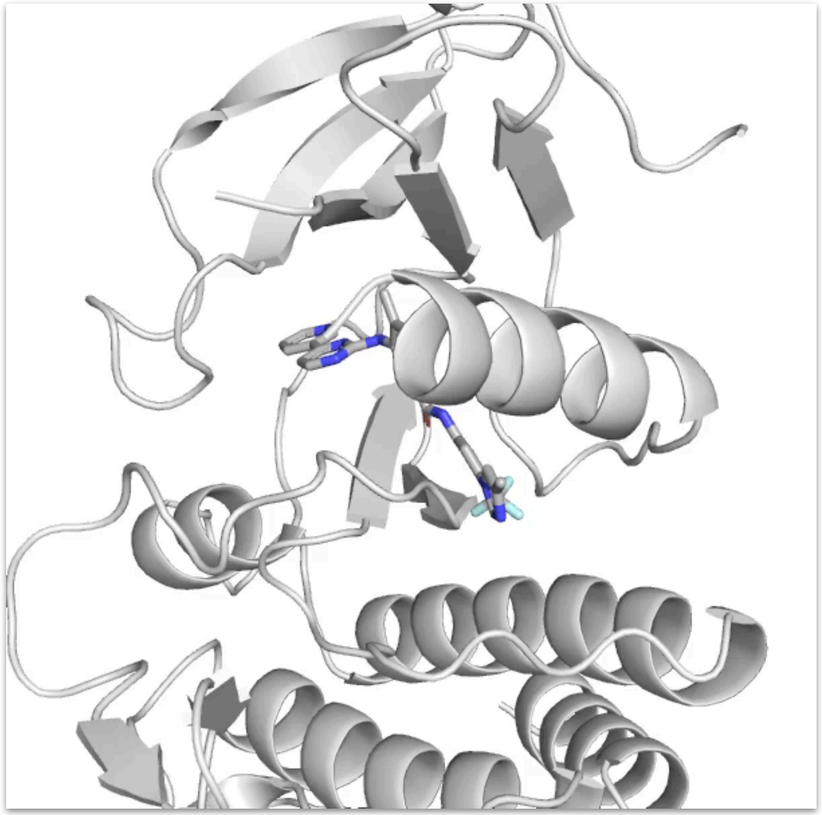
Data Augmentation





Data Augmentation

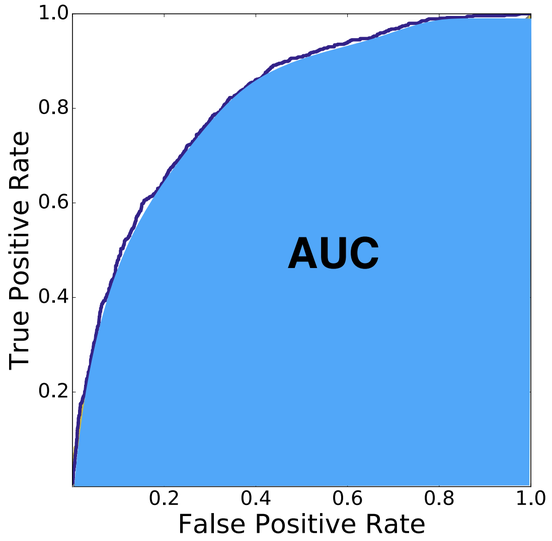


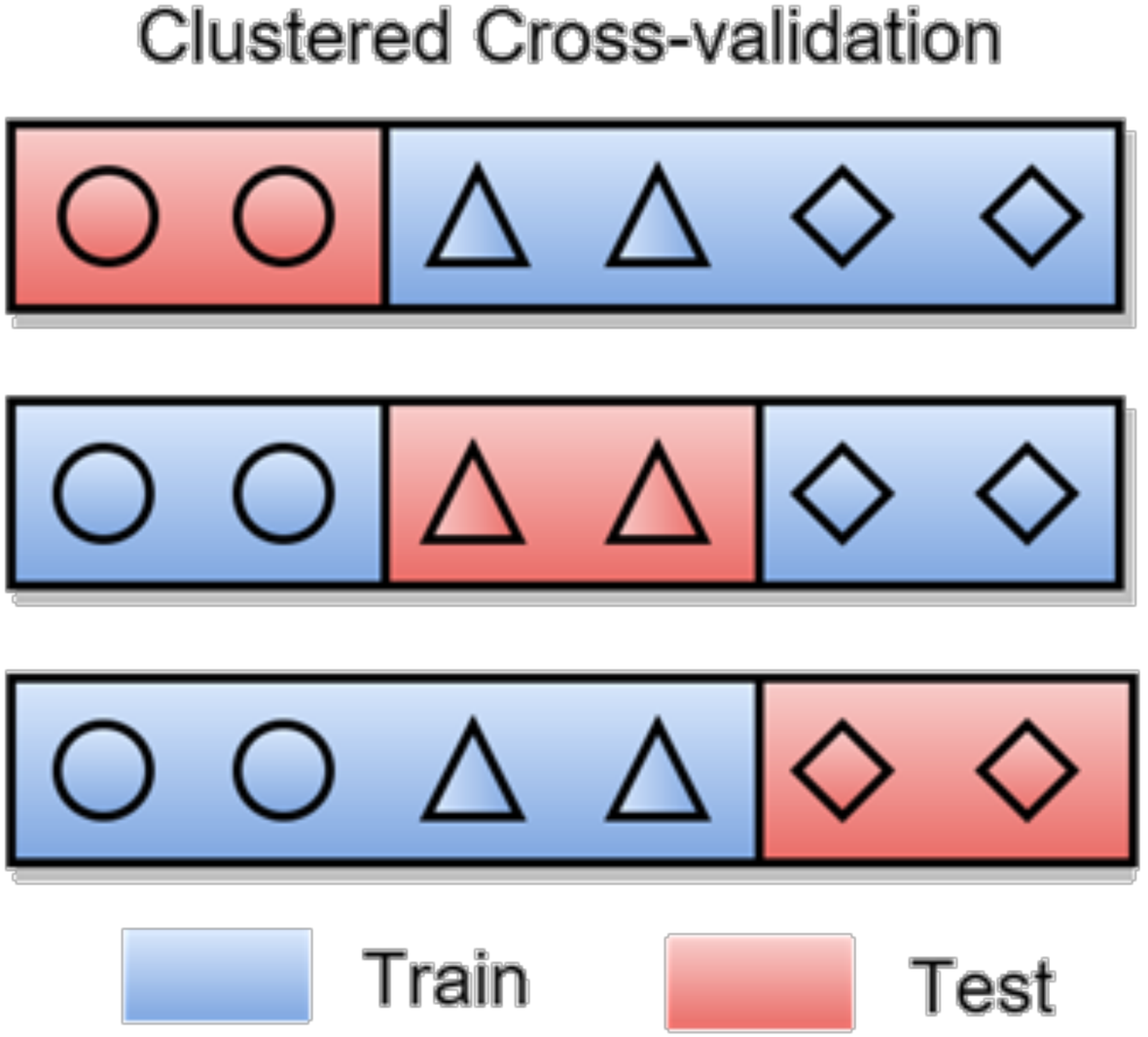


Model Evaluation

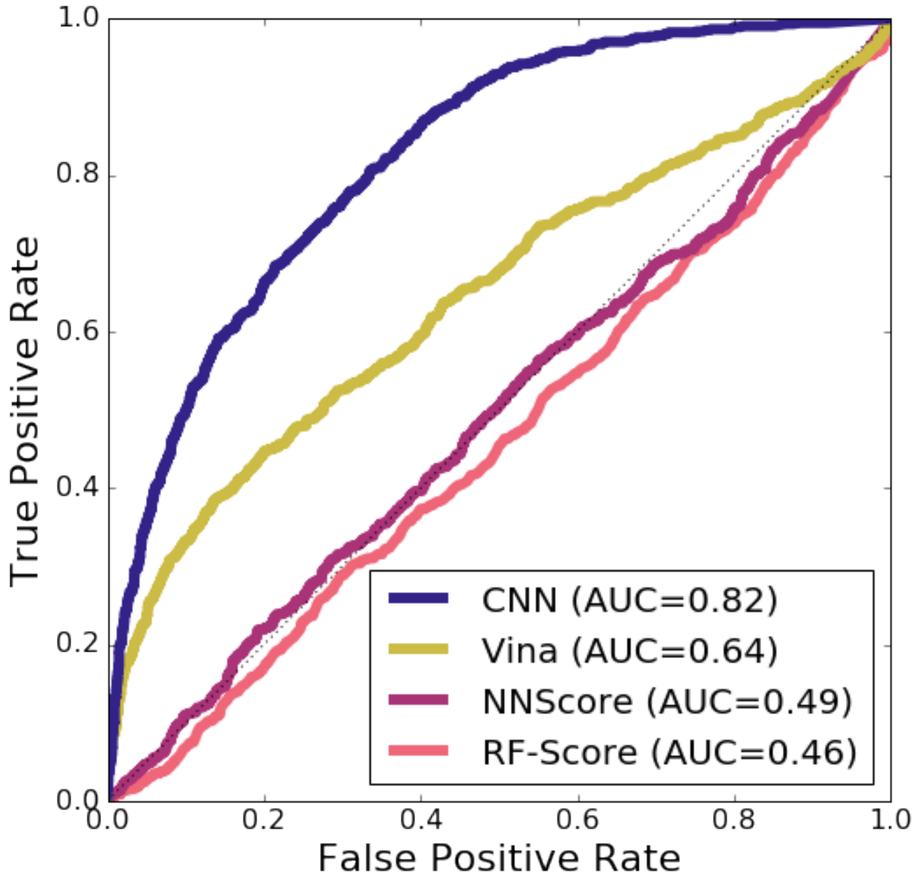
CSAR: >90% similar targets kept in same fold

DUD-E & PDBbind: >80% similar targets kept in same fold

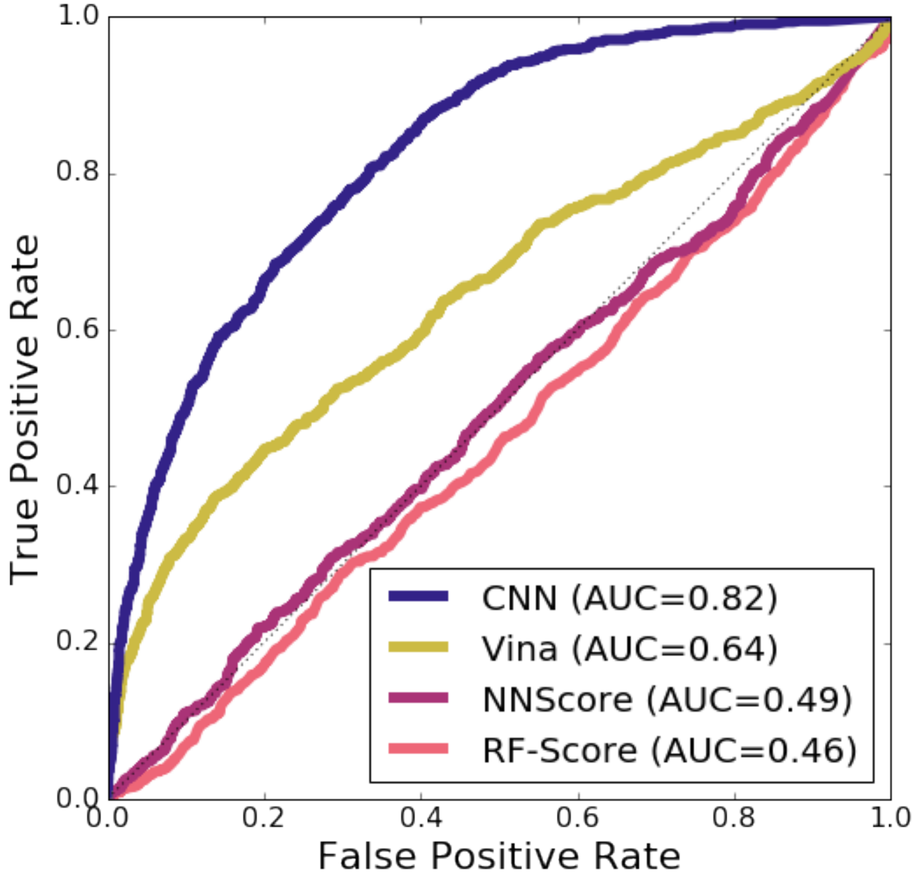




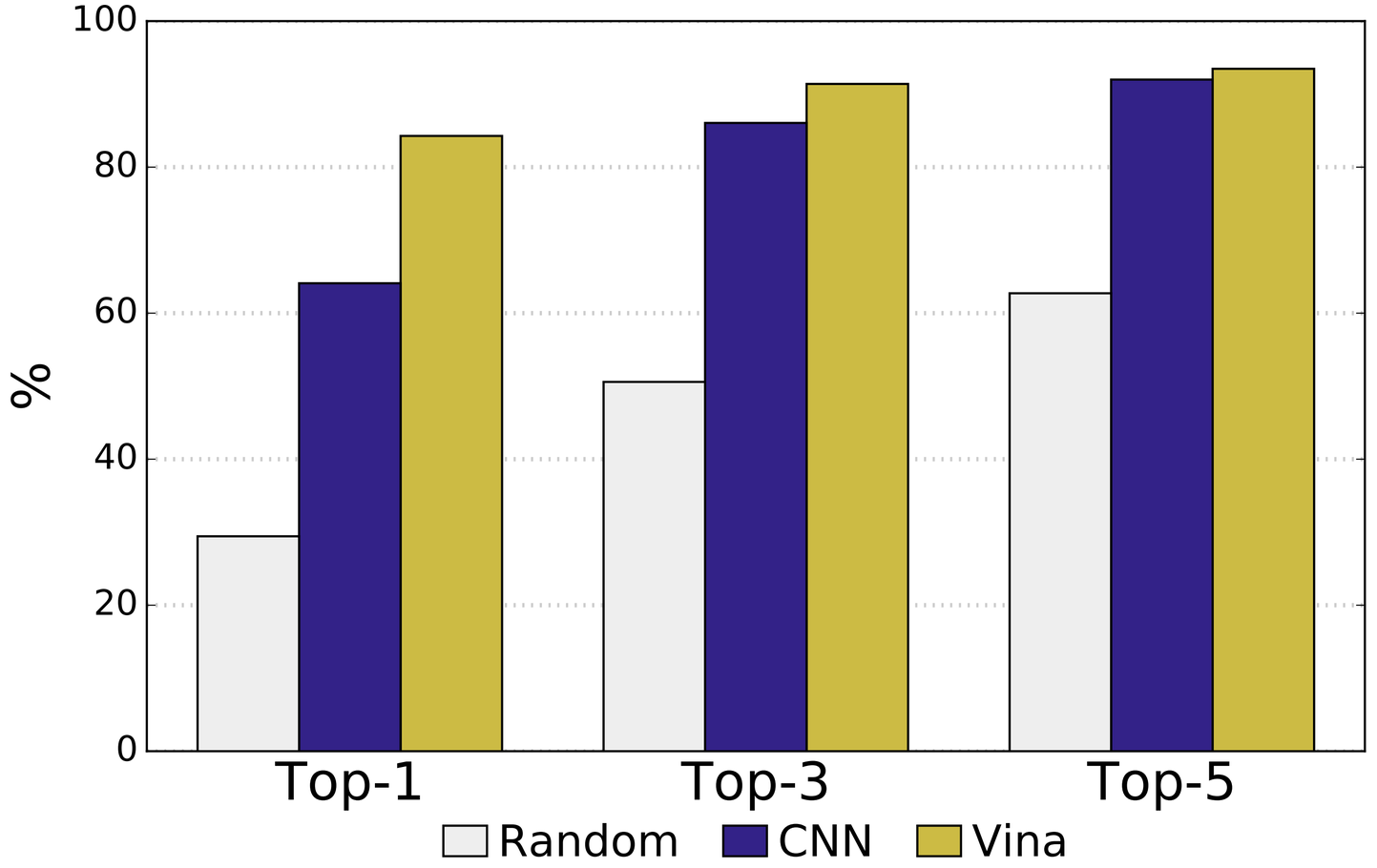
Pose Prediction (CSAR)



Pose Prediction (CSAR)

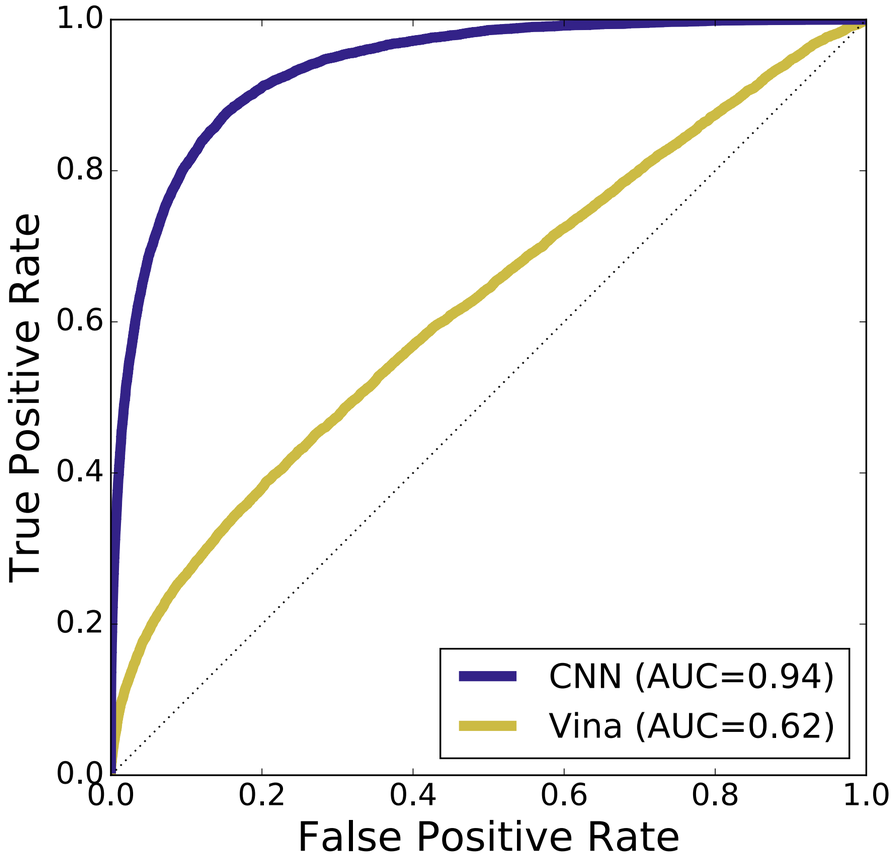


inter-target ranking

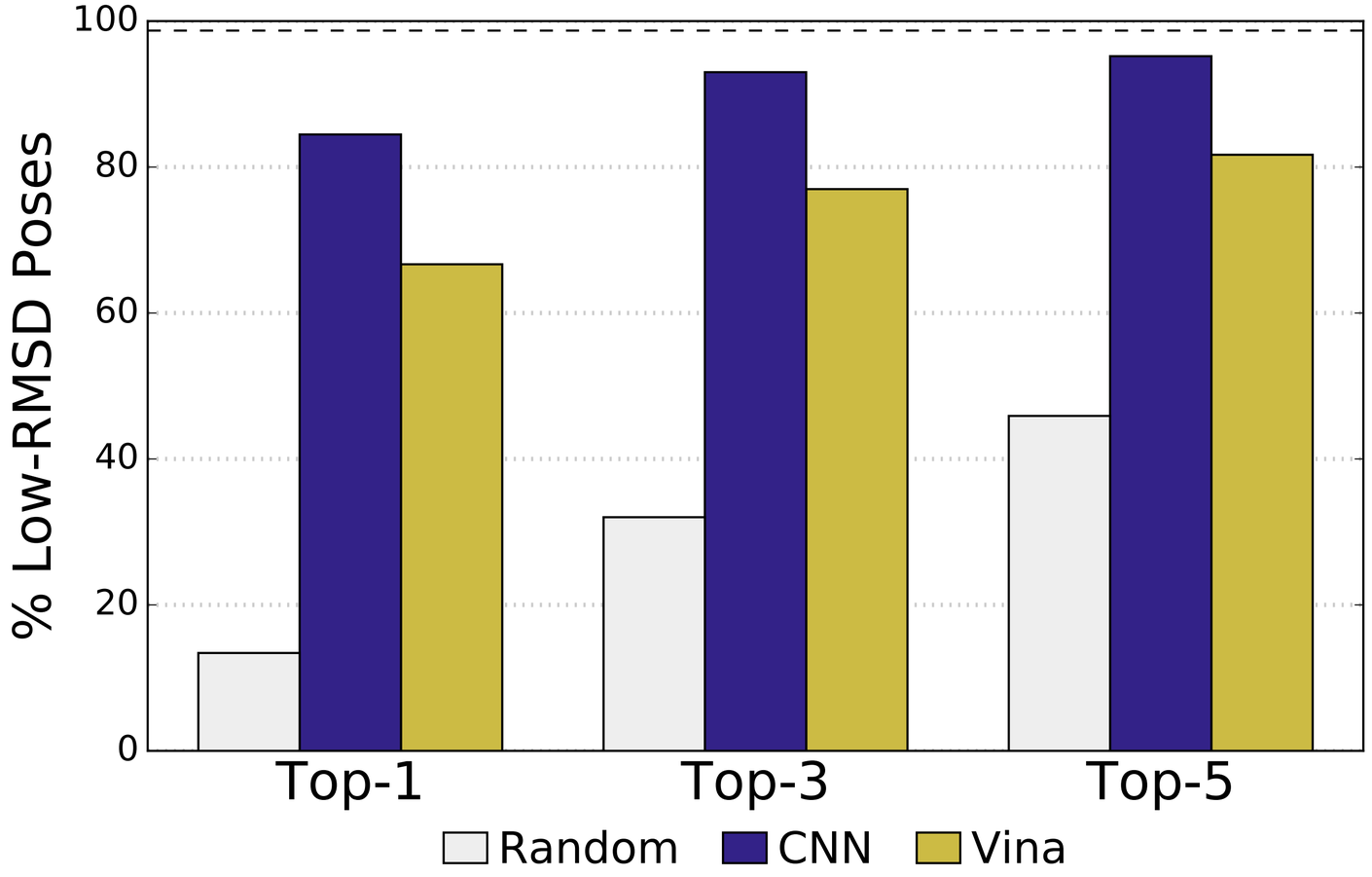


intra-target ranking

Pose Prediction (PDBbind)

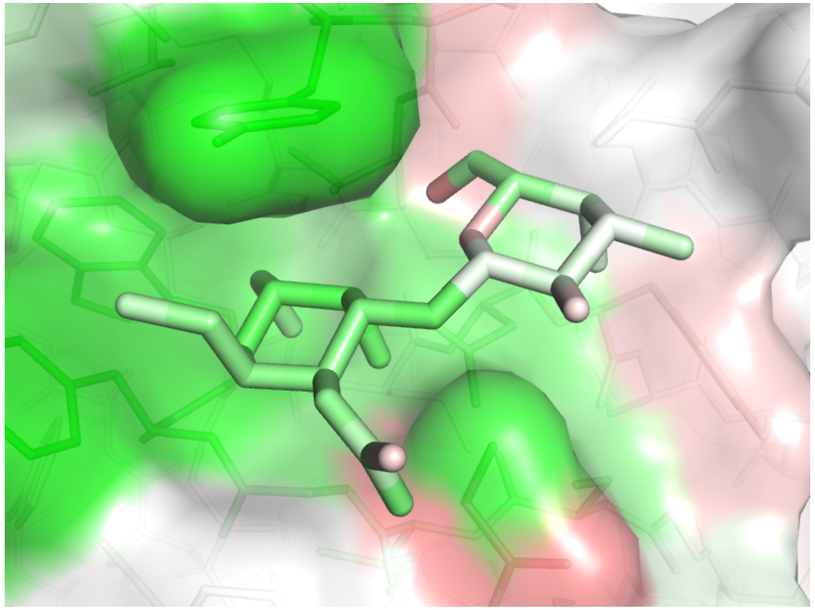


inter-target ranking

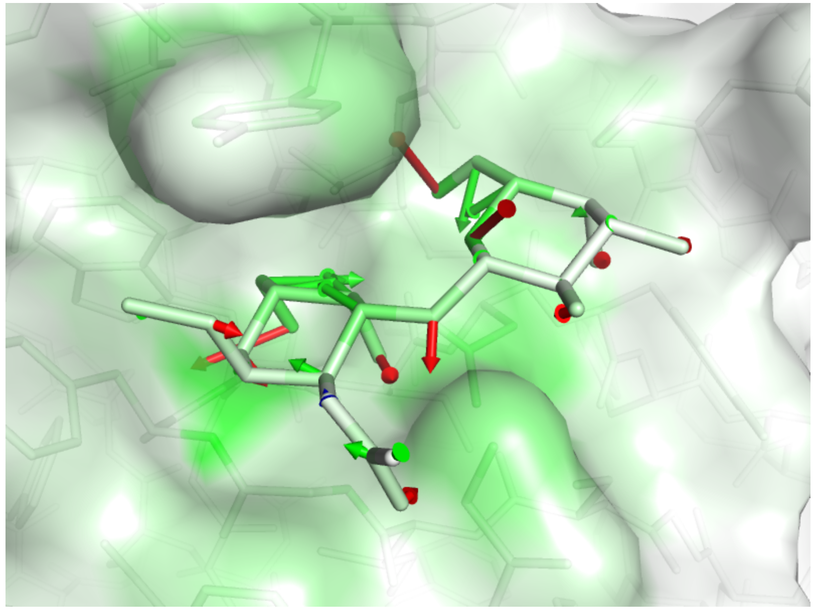


intra-target ranking

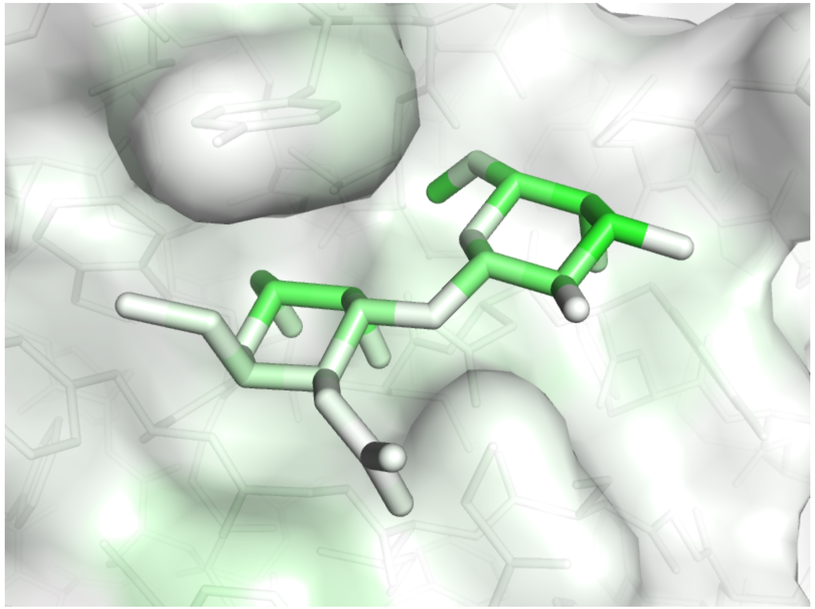
Visualization



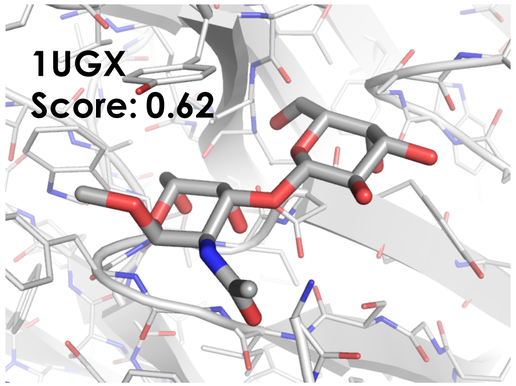
masking



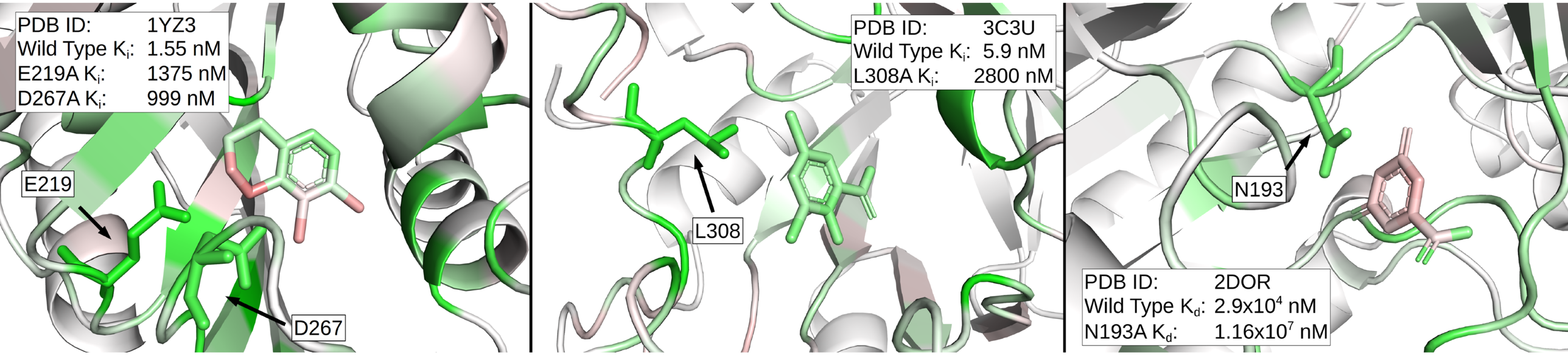
gradients



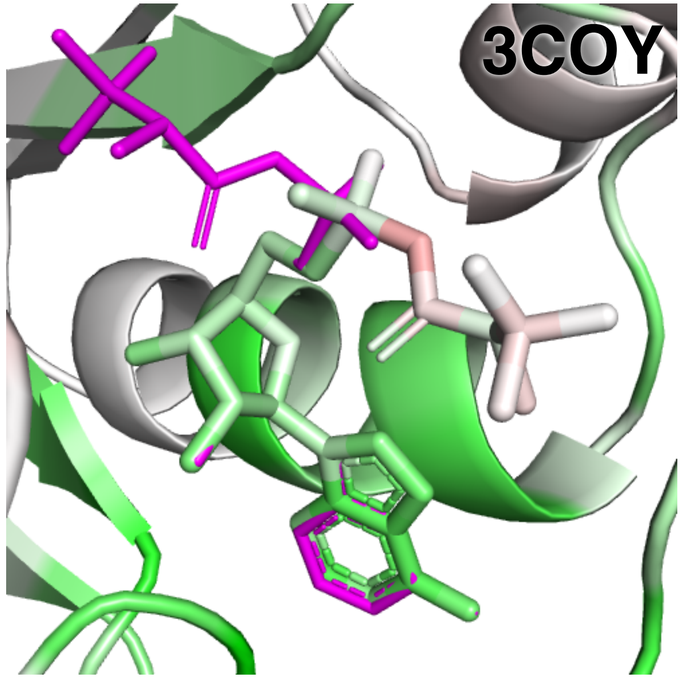
layer-wise relevance

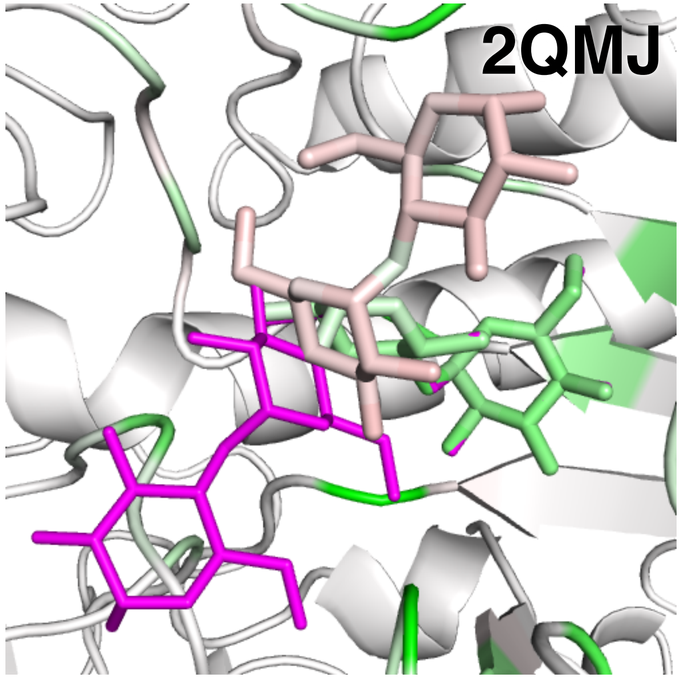


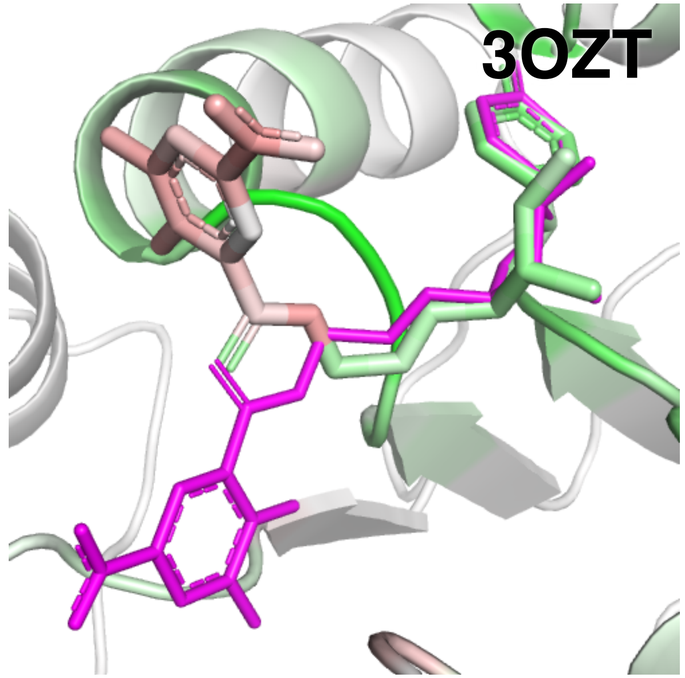
Visualizing Enzymes



Pose Sensitivity

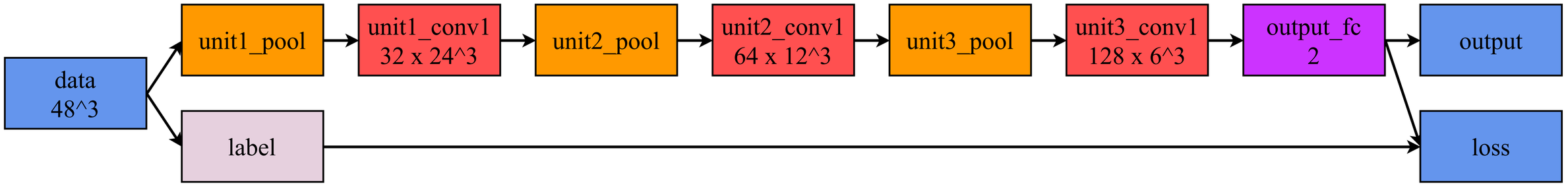




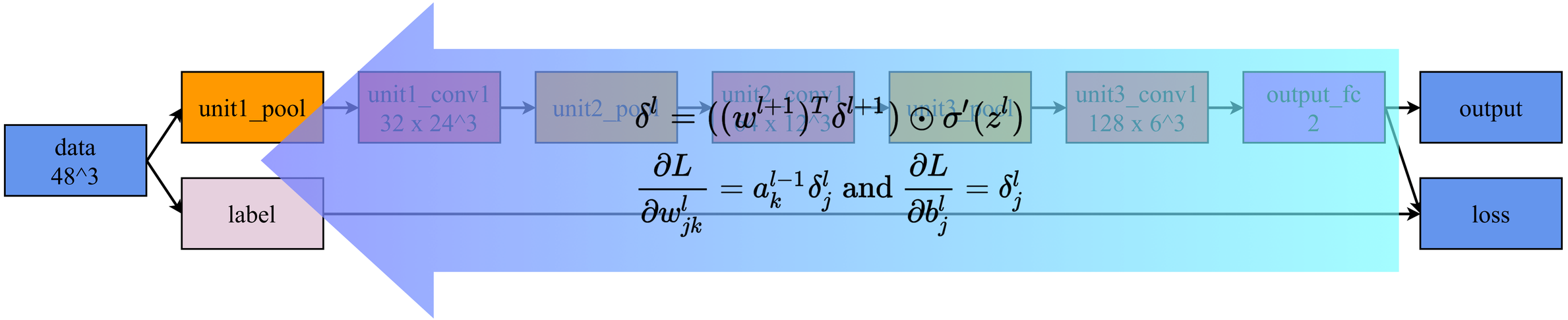


Partially Aligned Poses

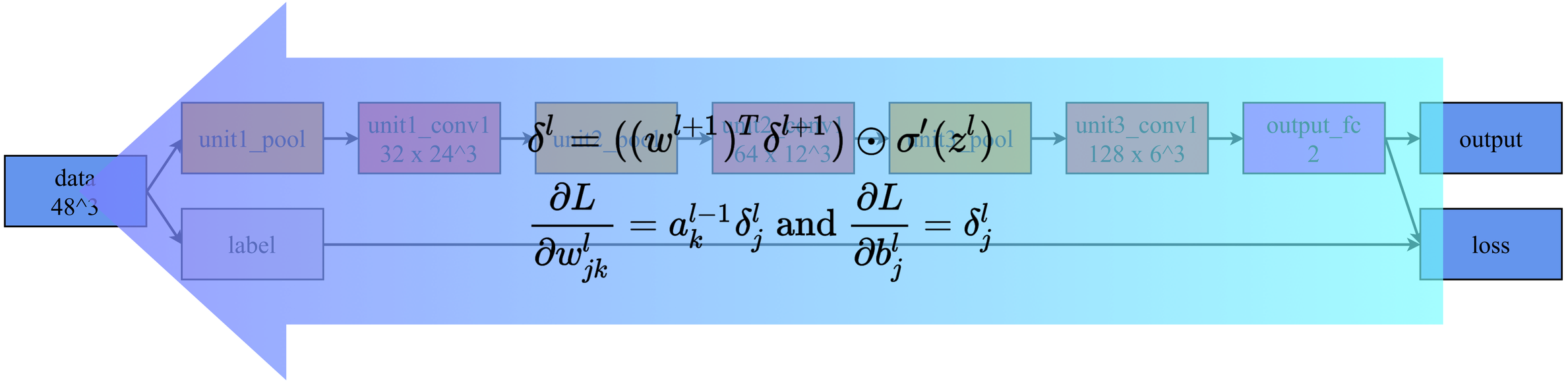
Beyond Scoring



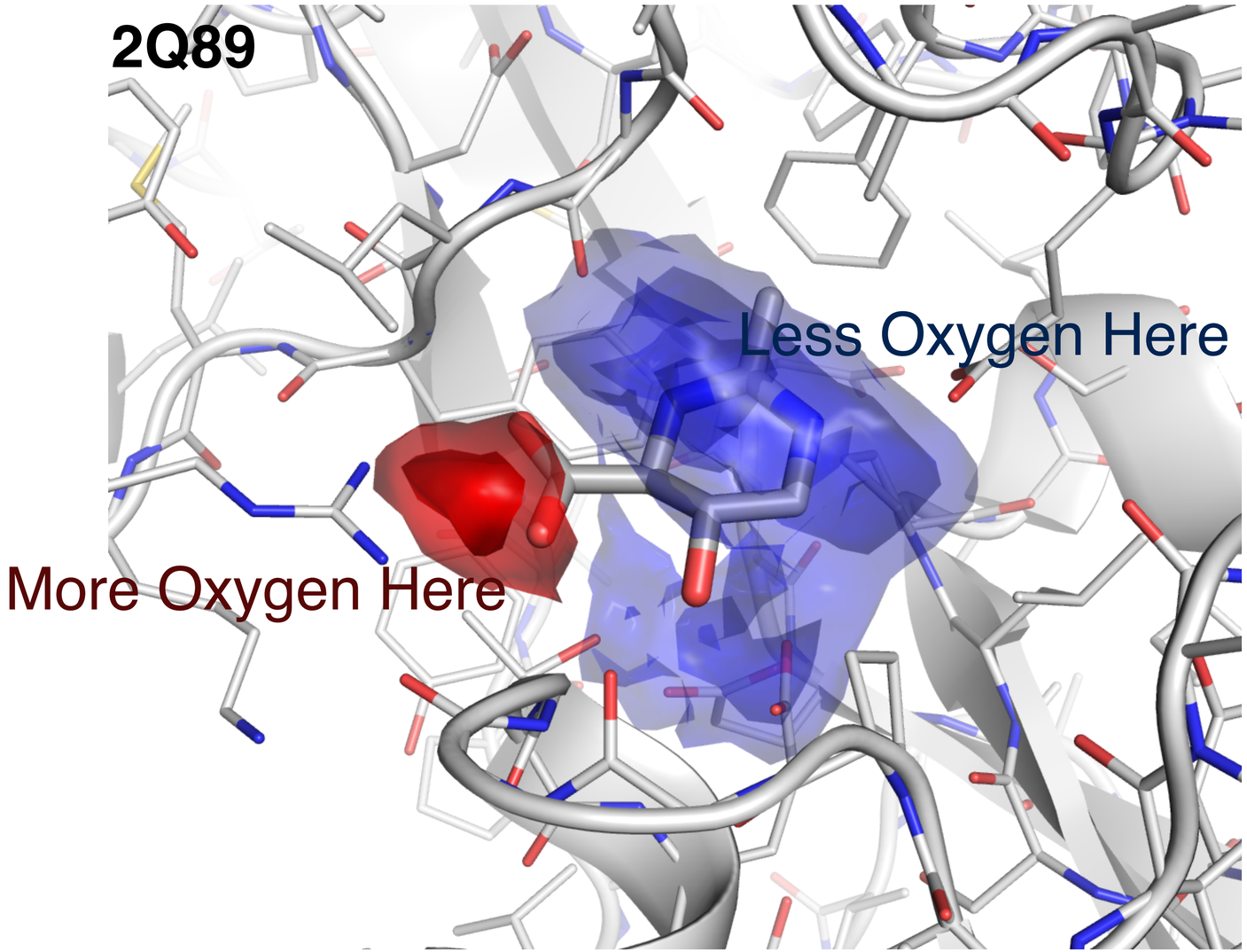
Beyond Scoring



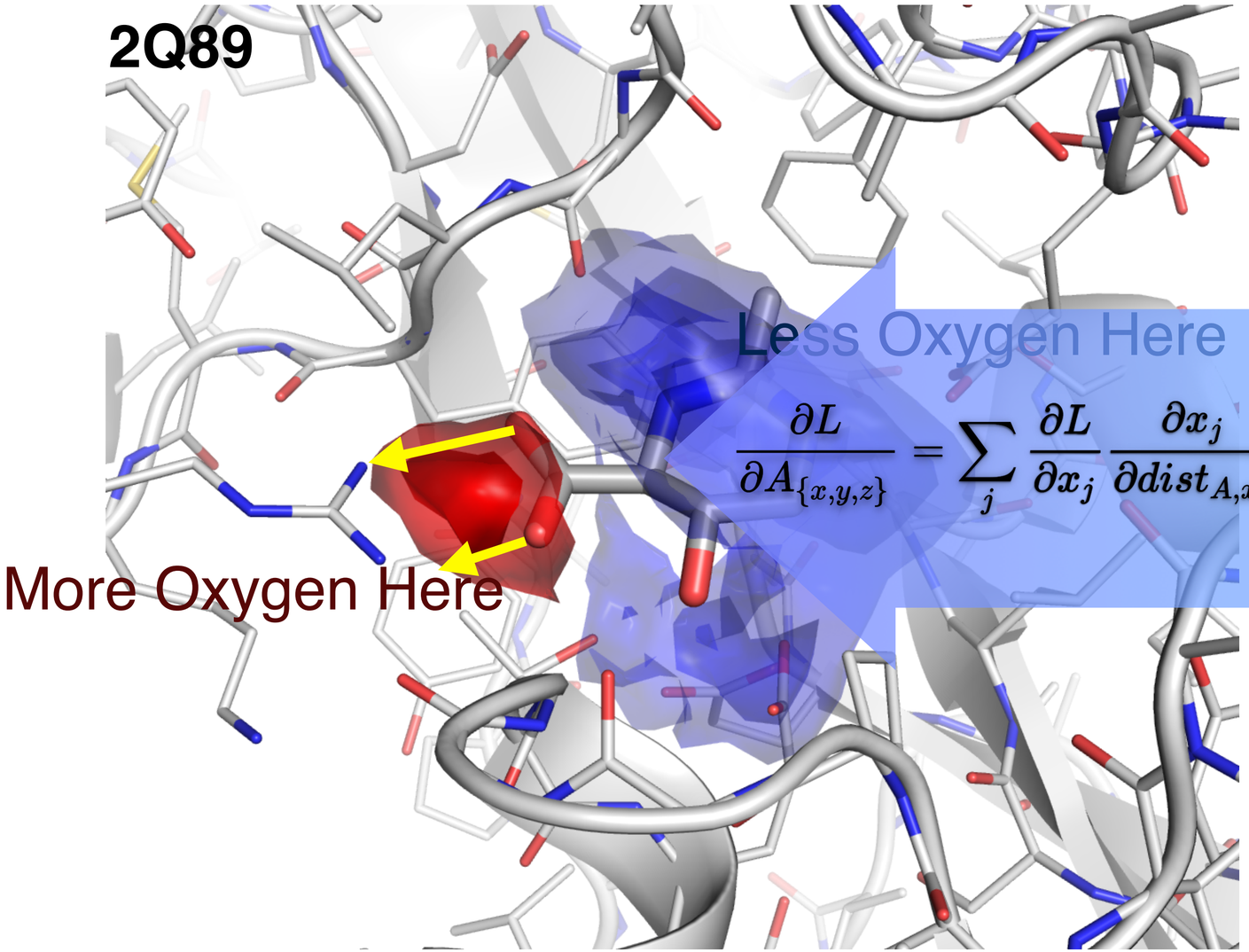
Beyond Scoring

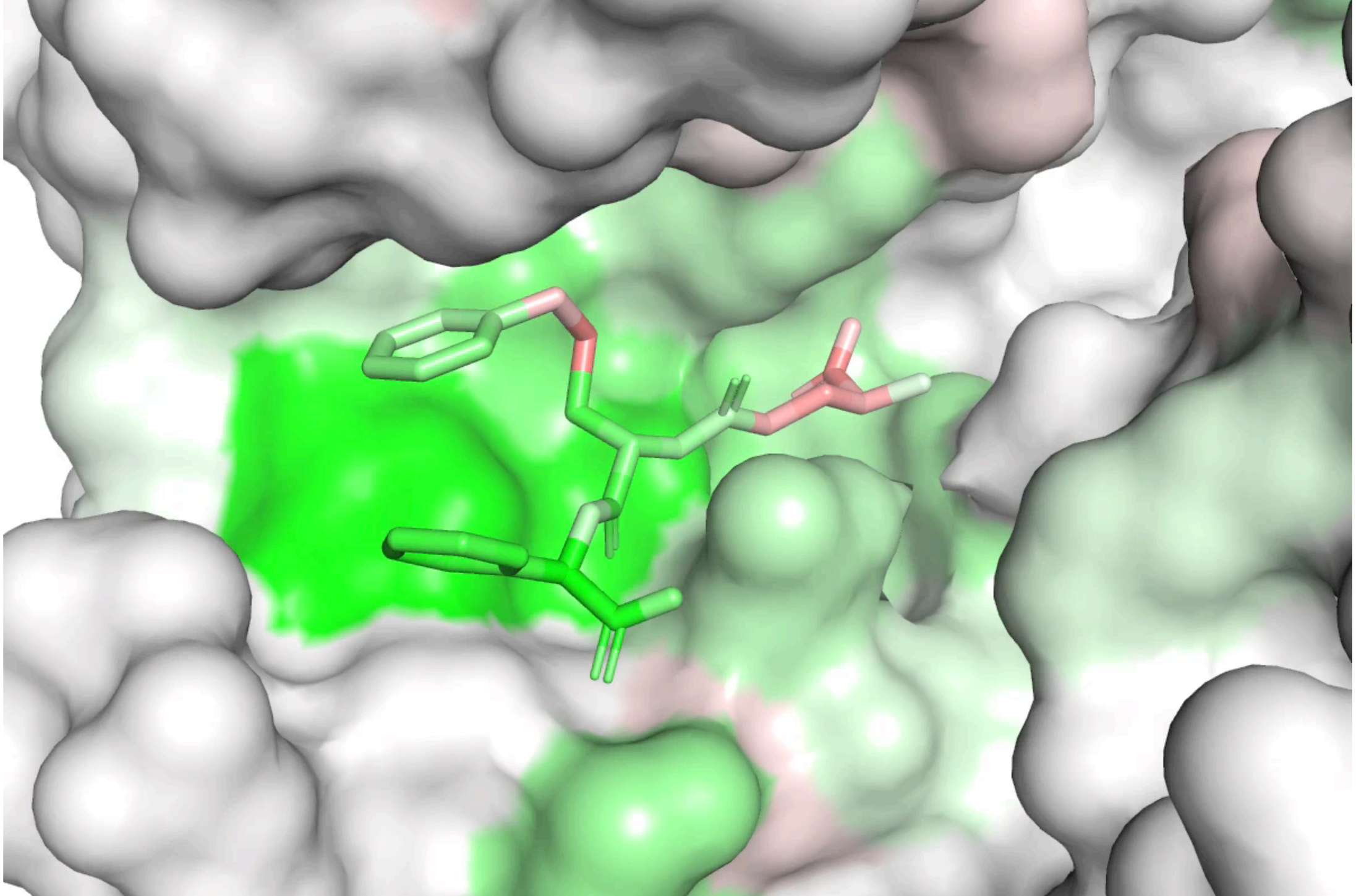


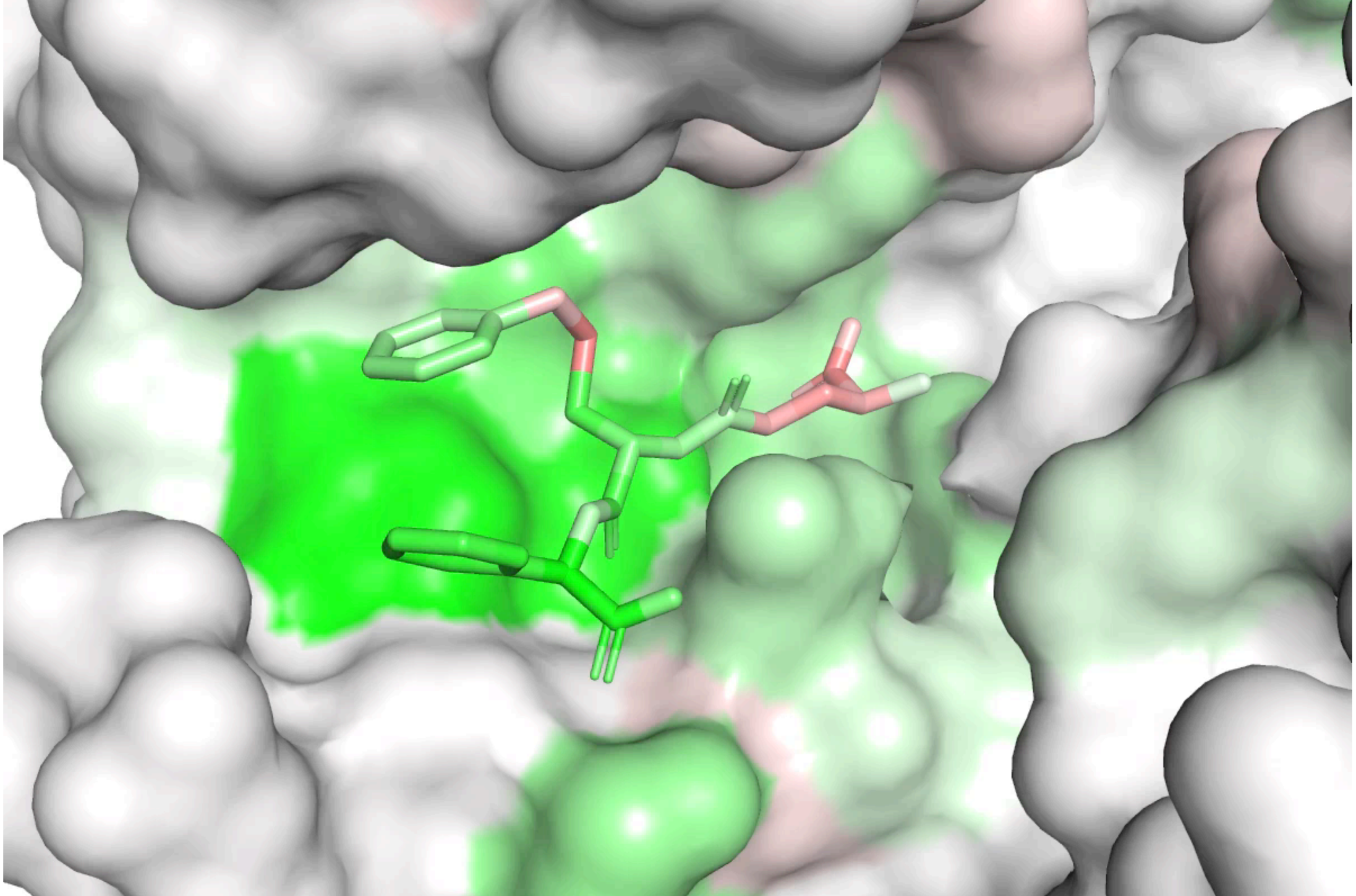
Beyond Scoring

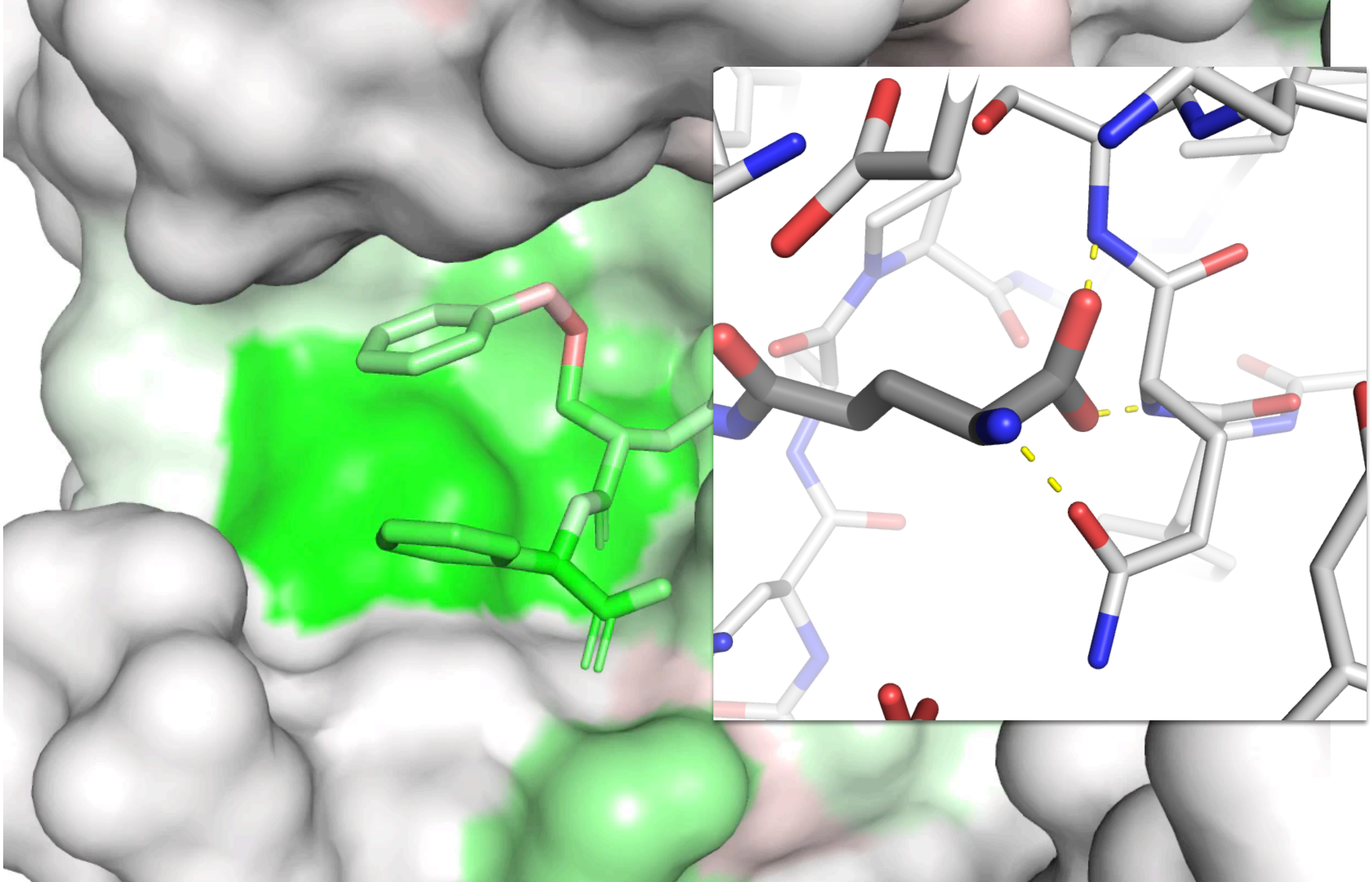


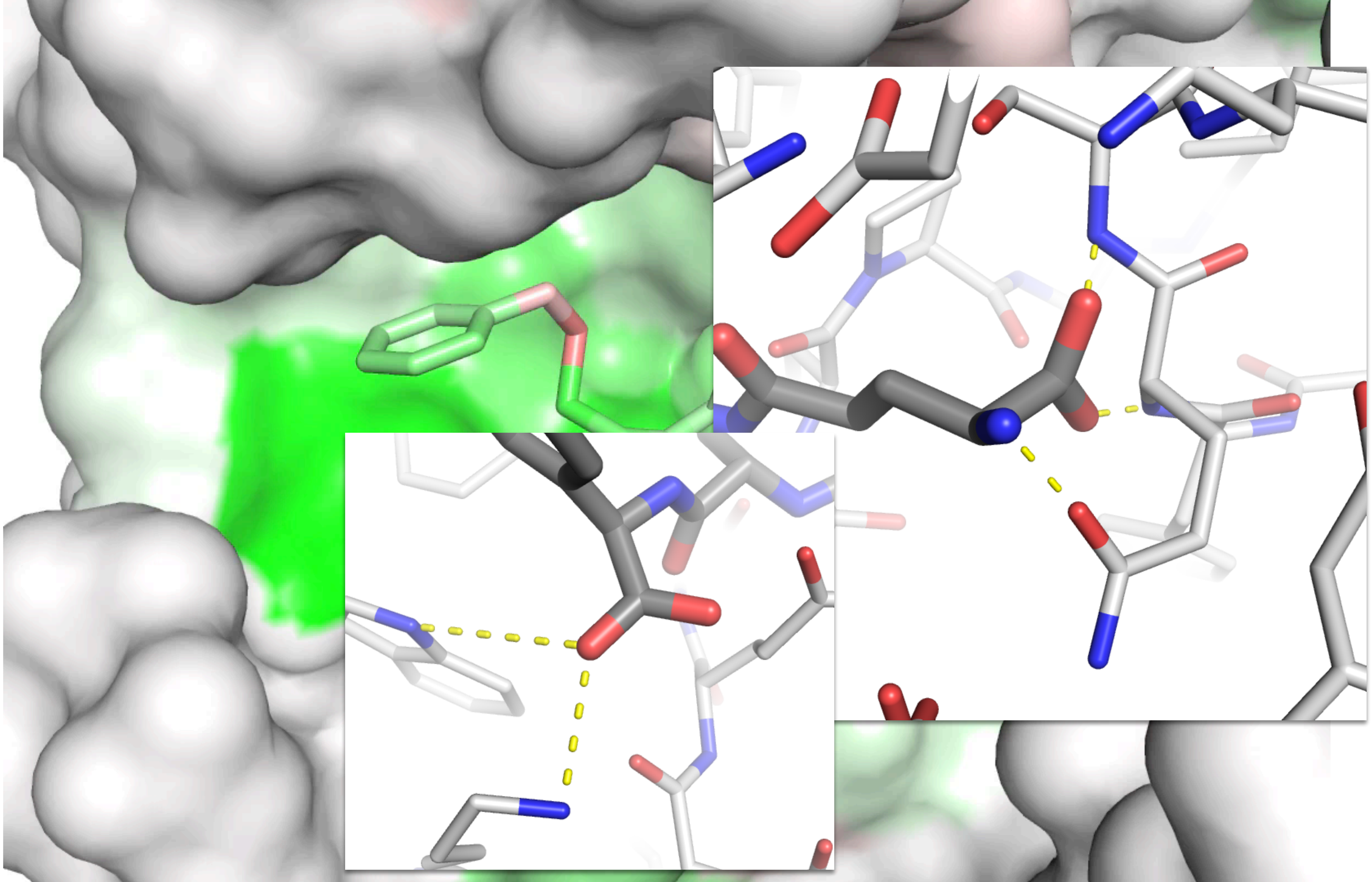
Beyond Scoring



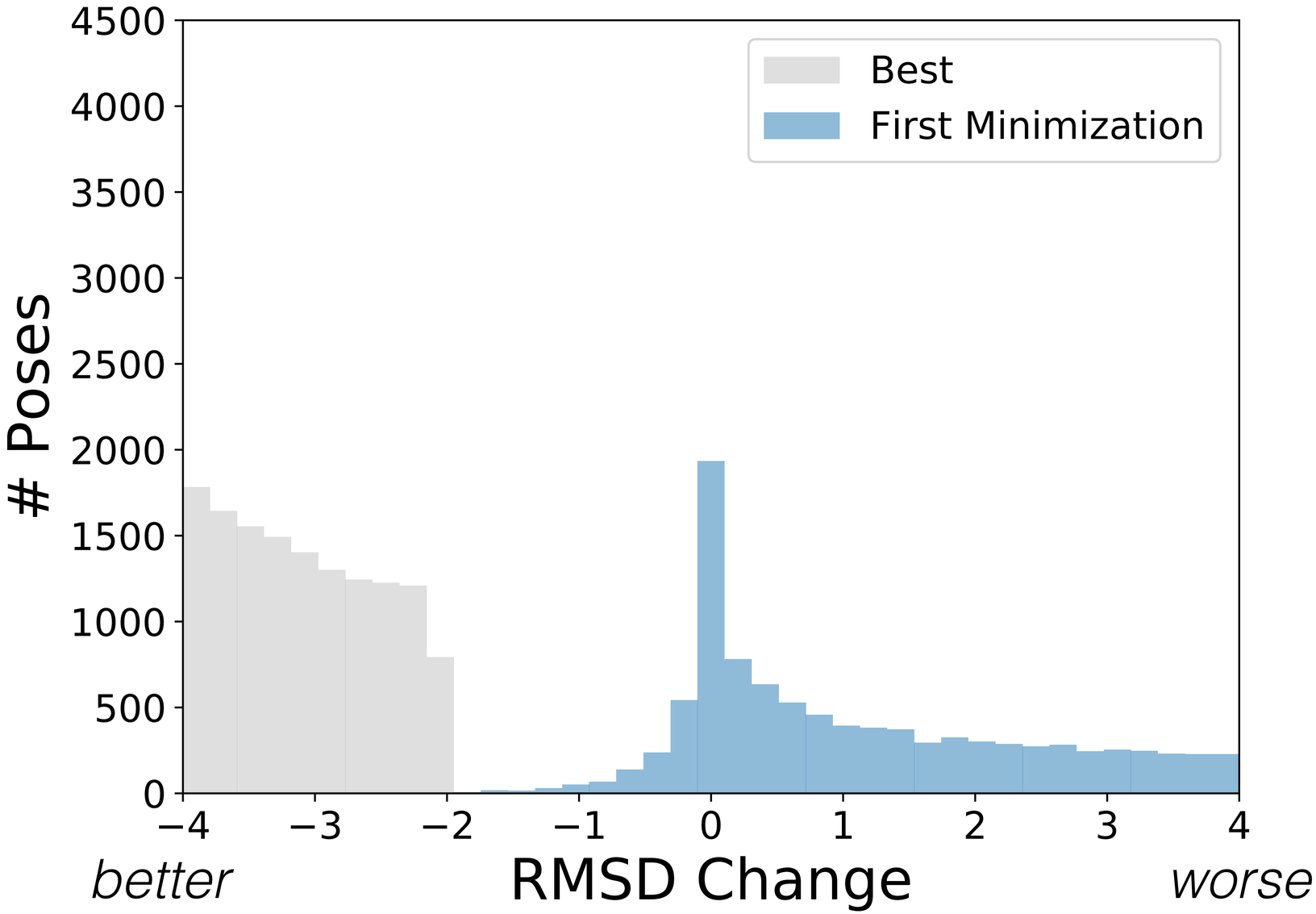


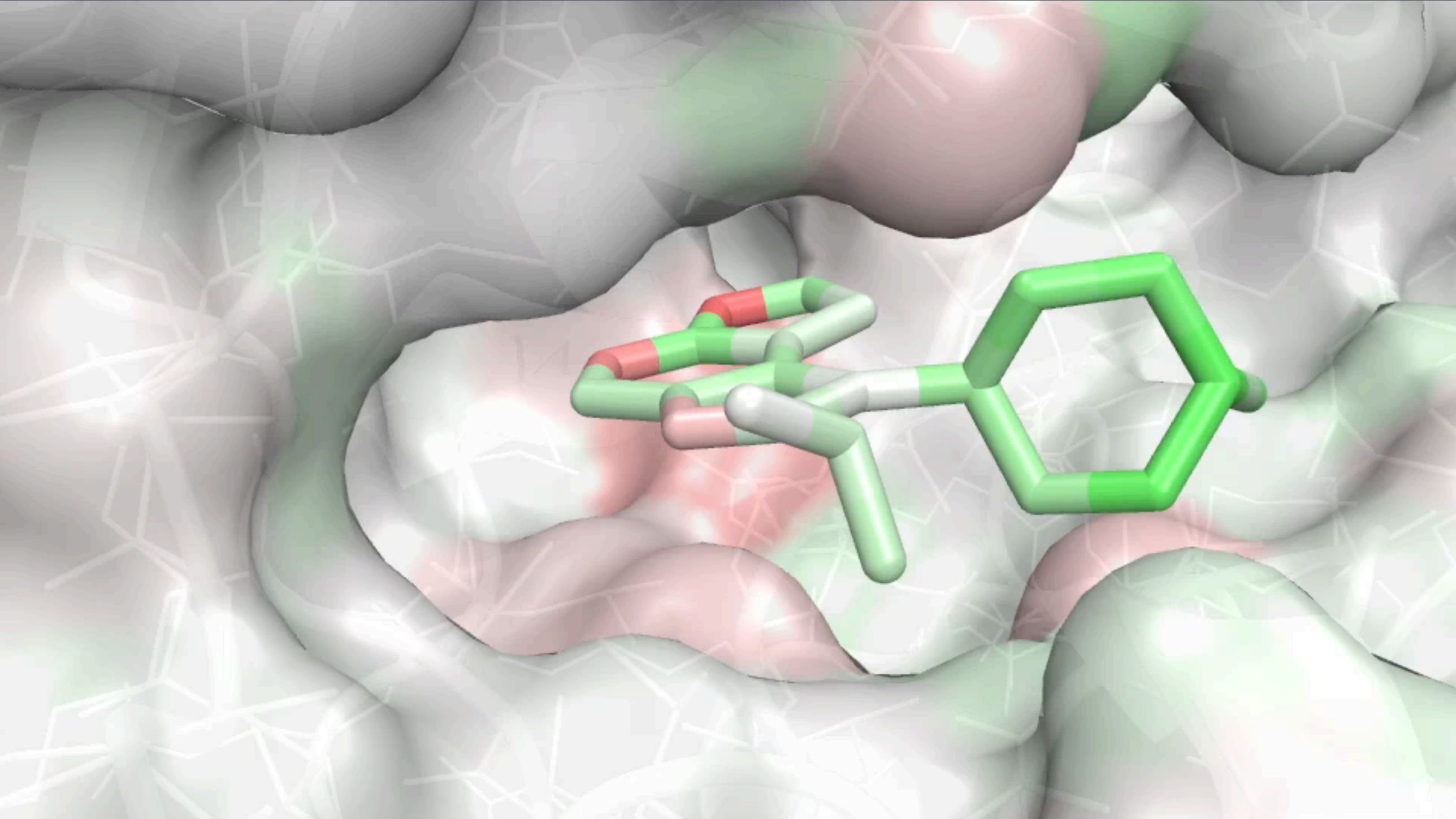


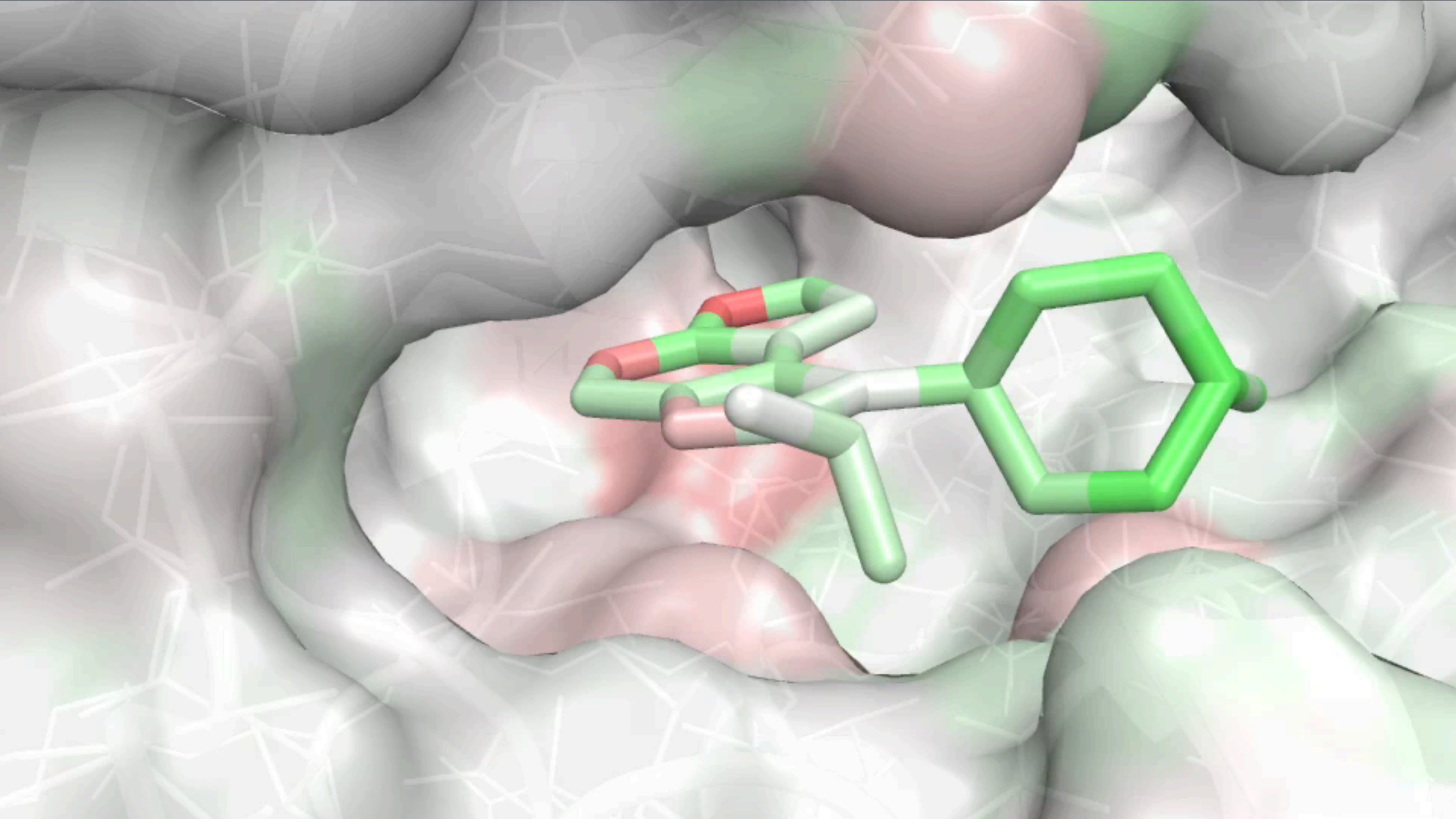


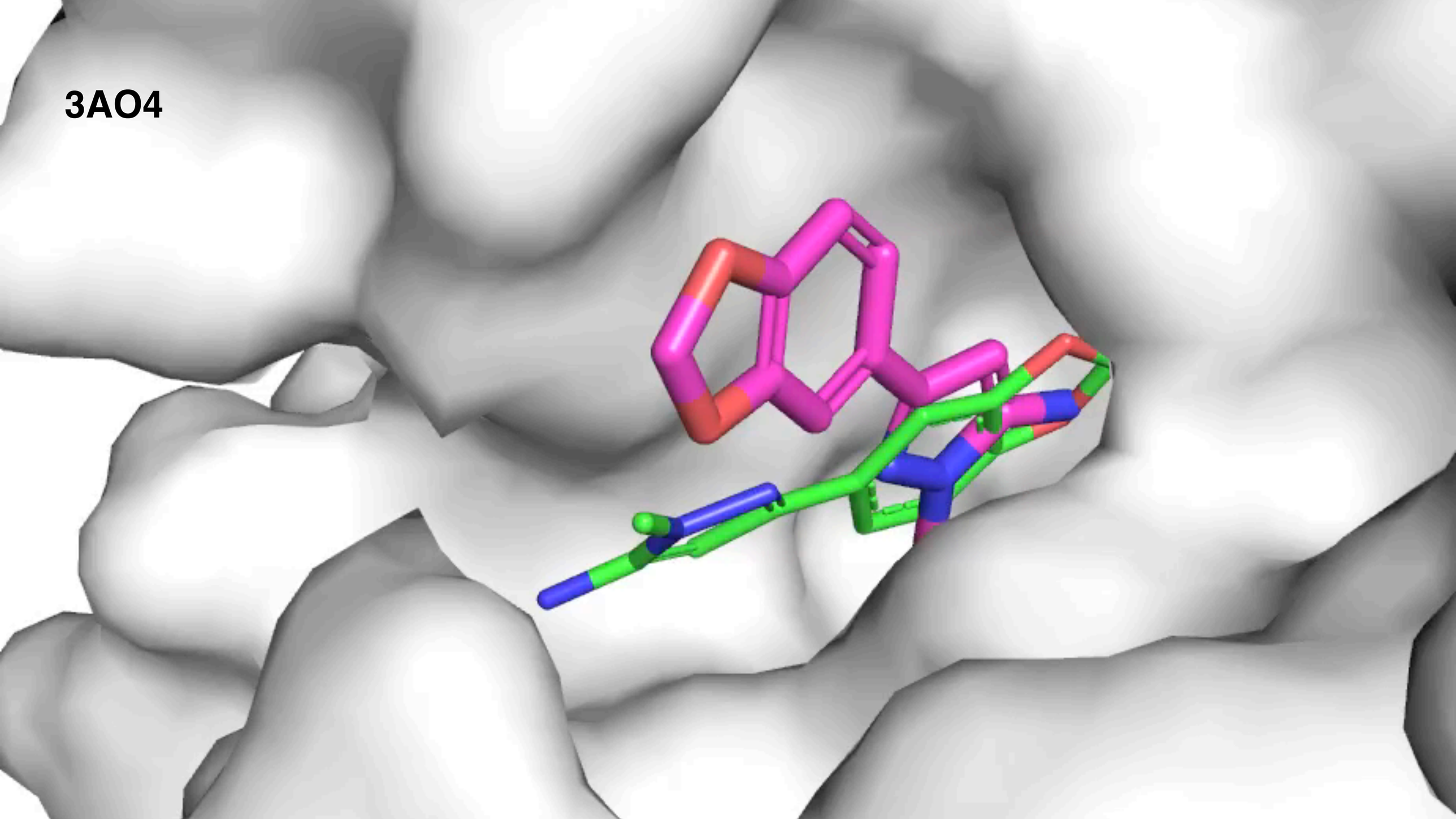


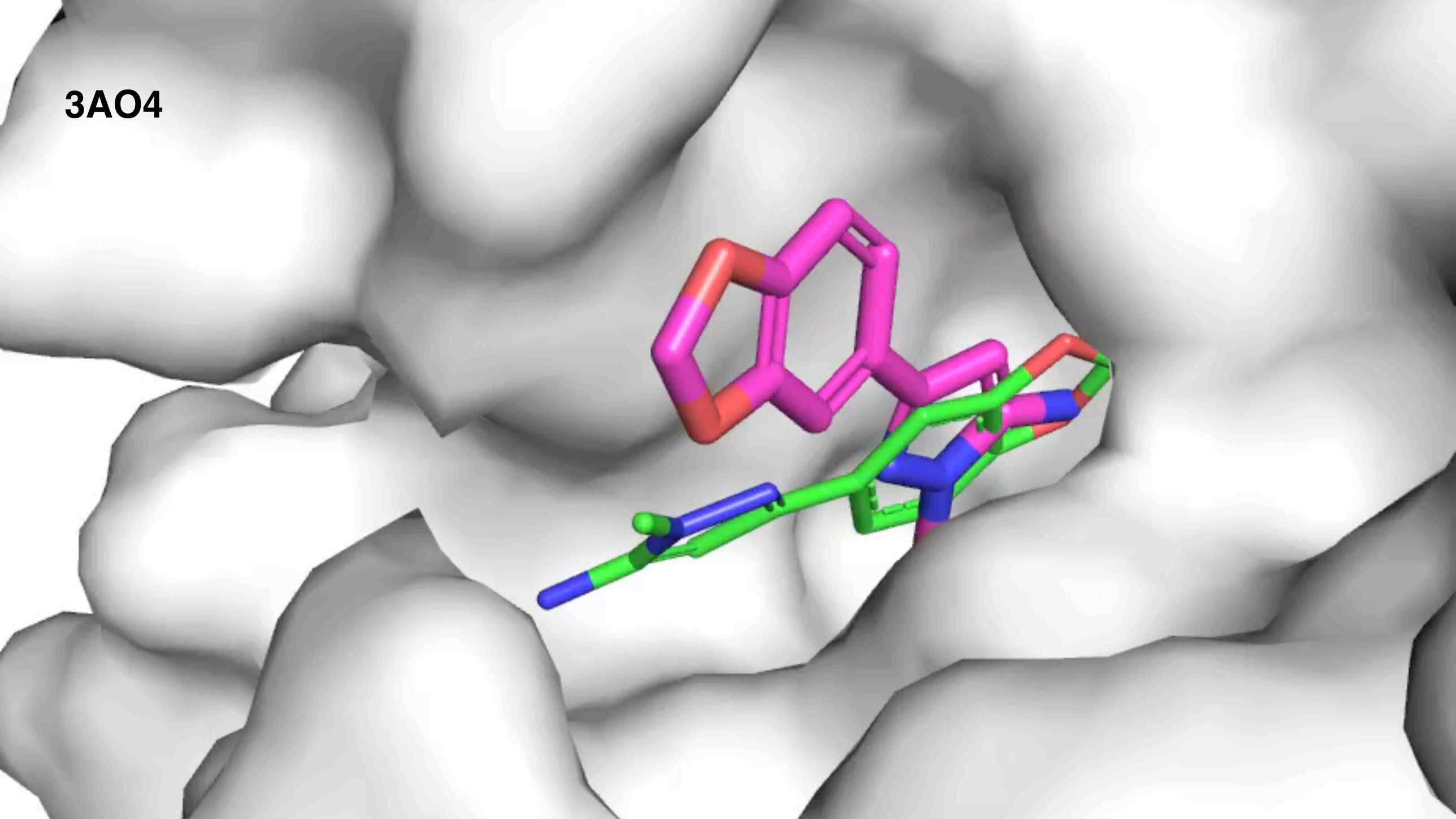
Minimizing Low RMSD Poses



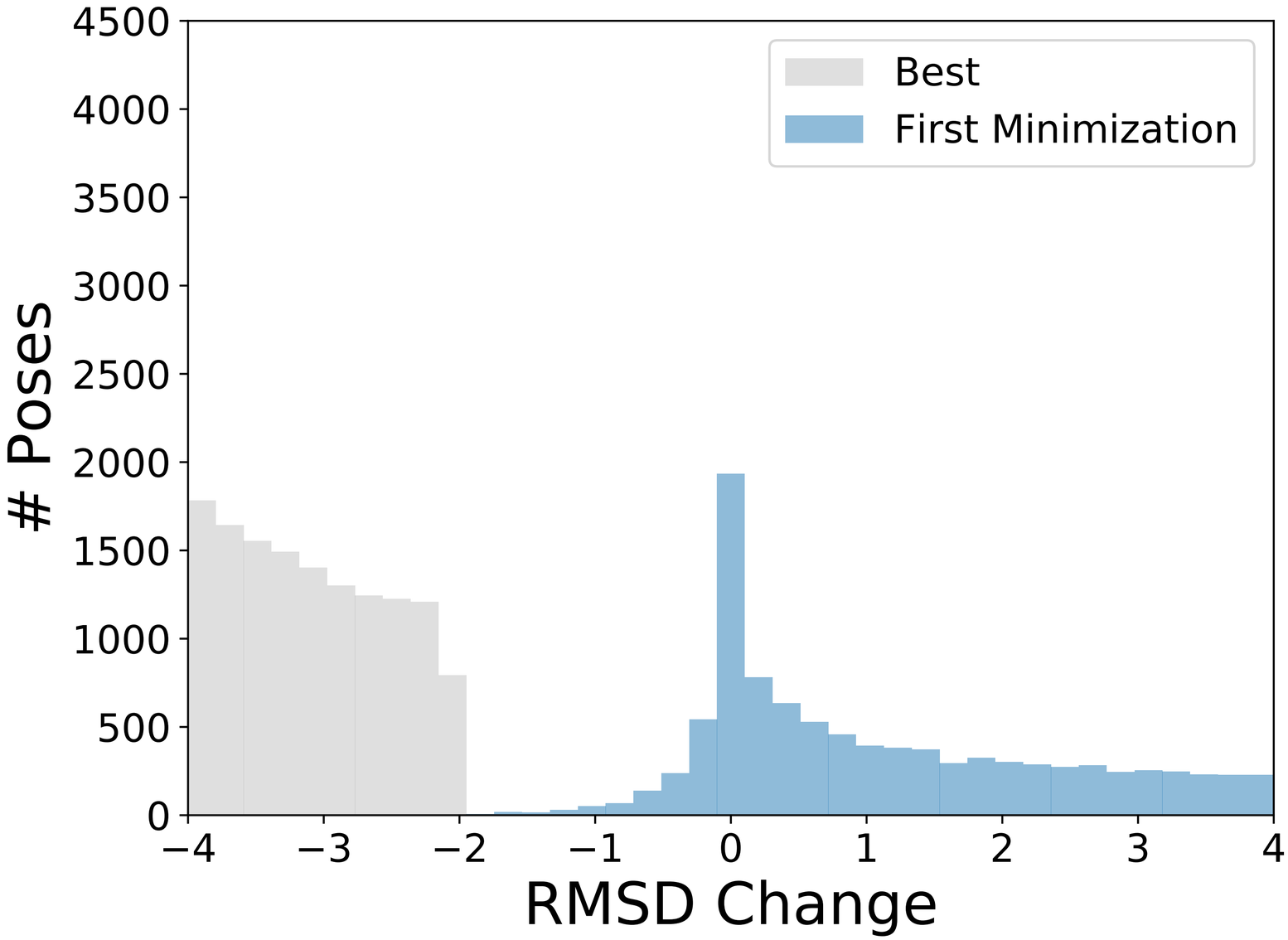




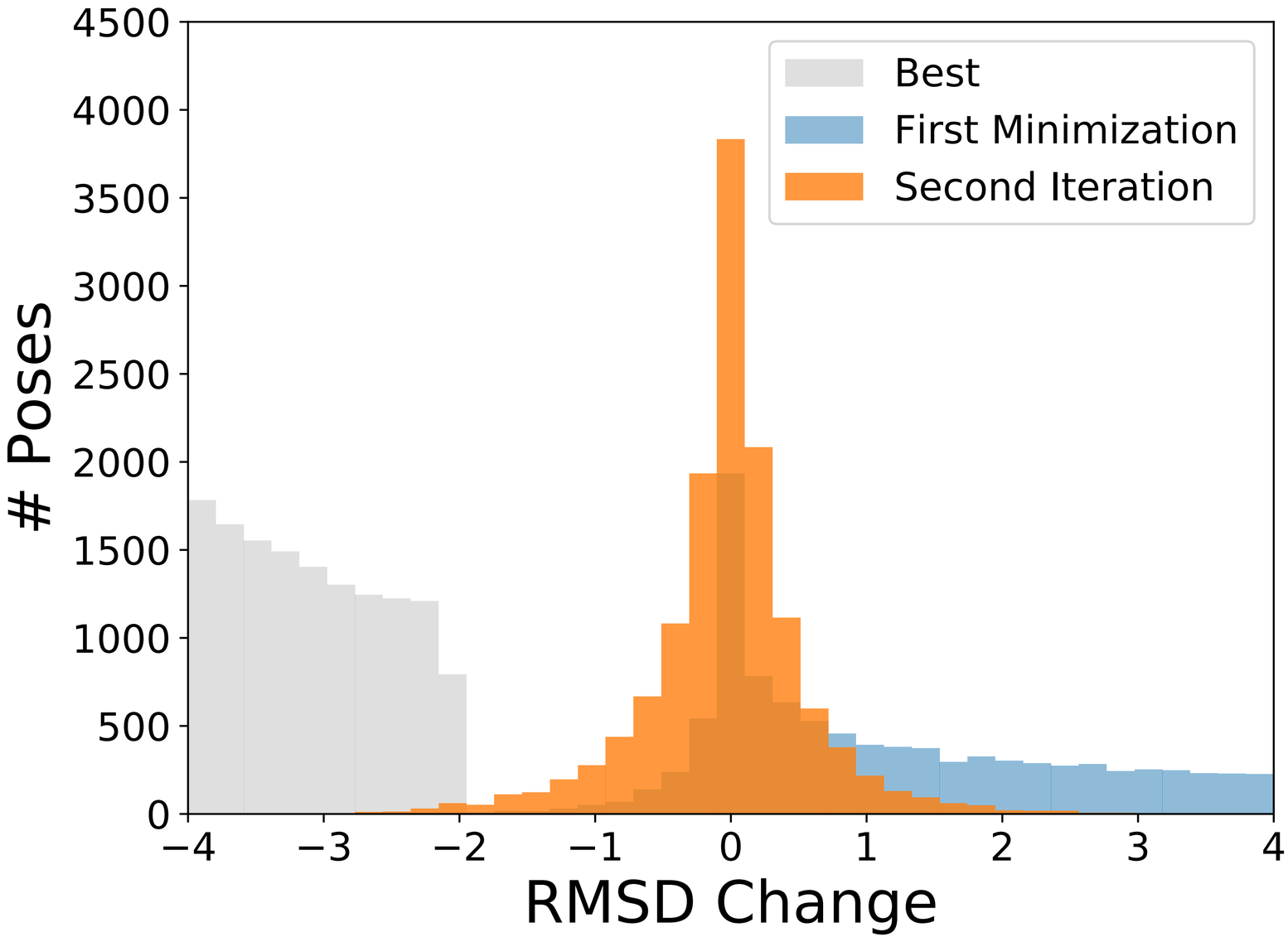

3AO4


3AO4

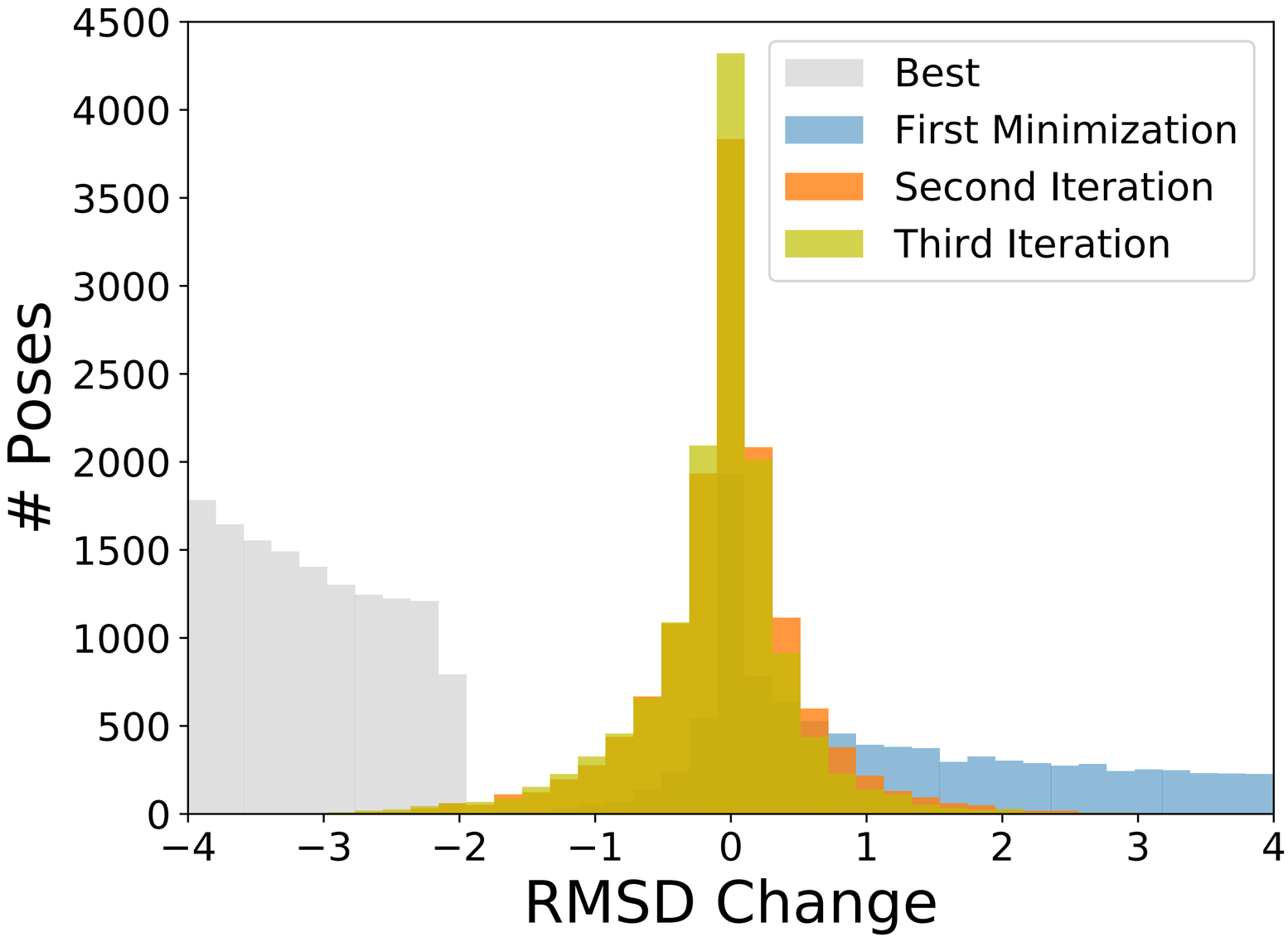
Iterative Refinement

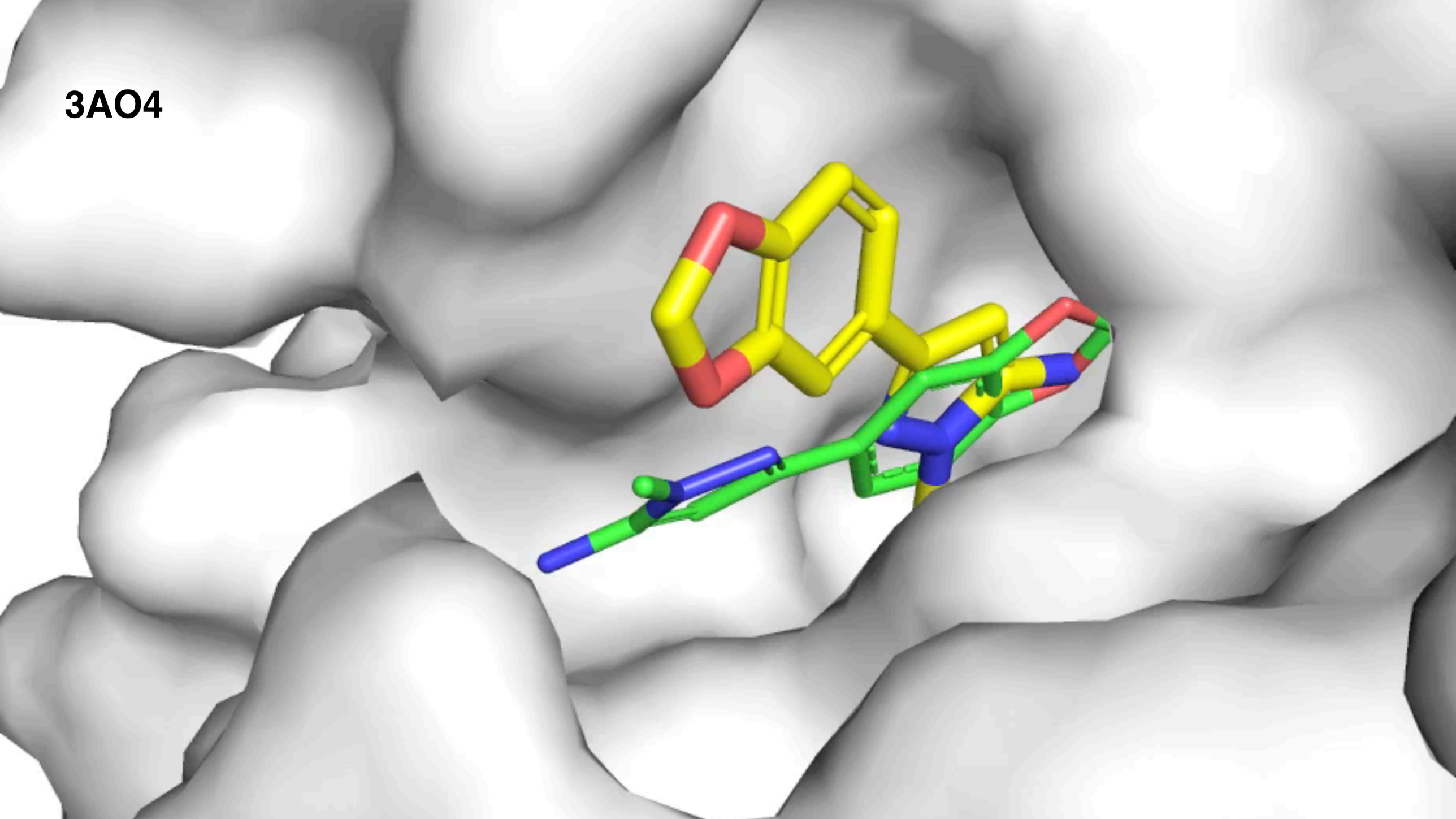


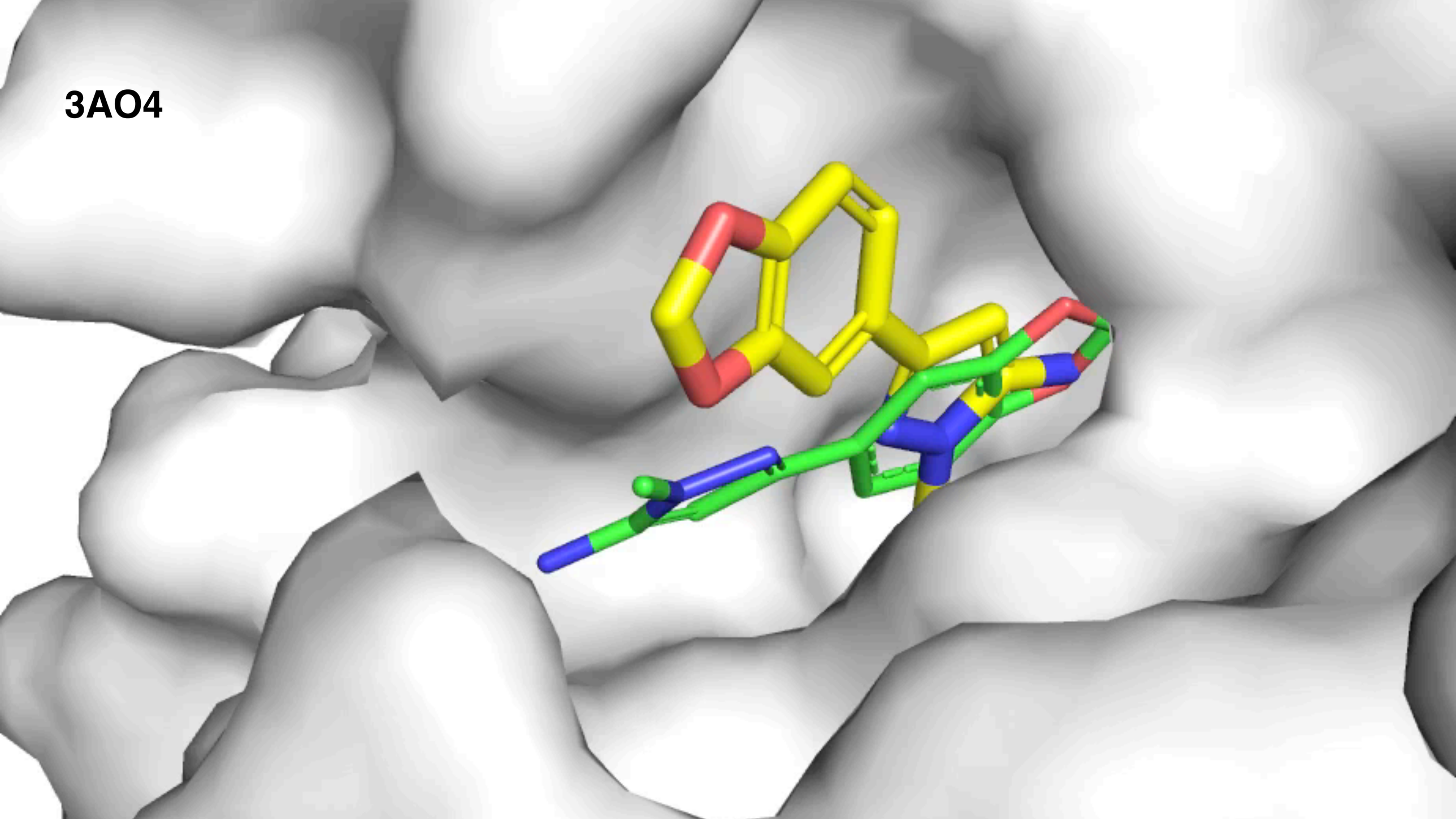
Iterative Refinement



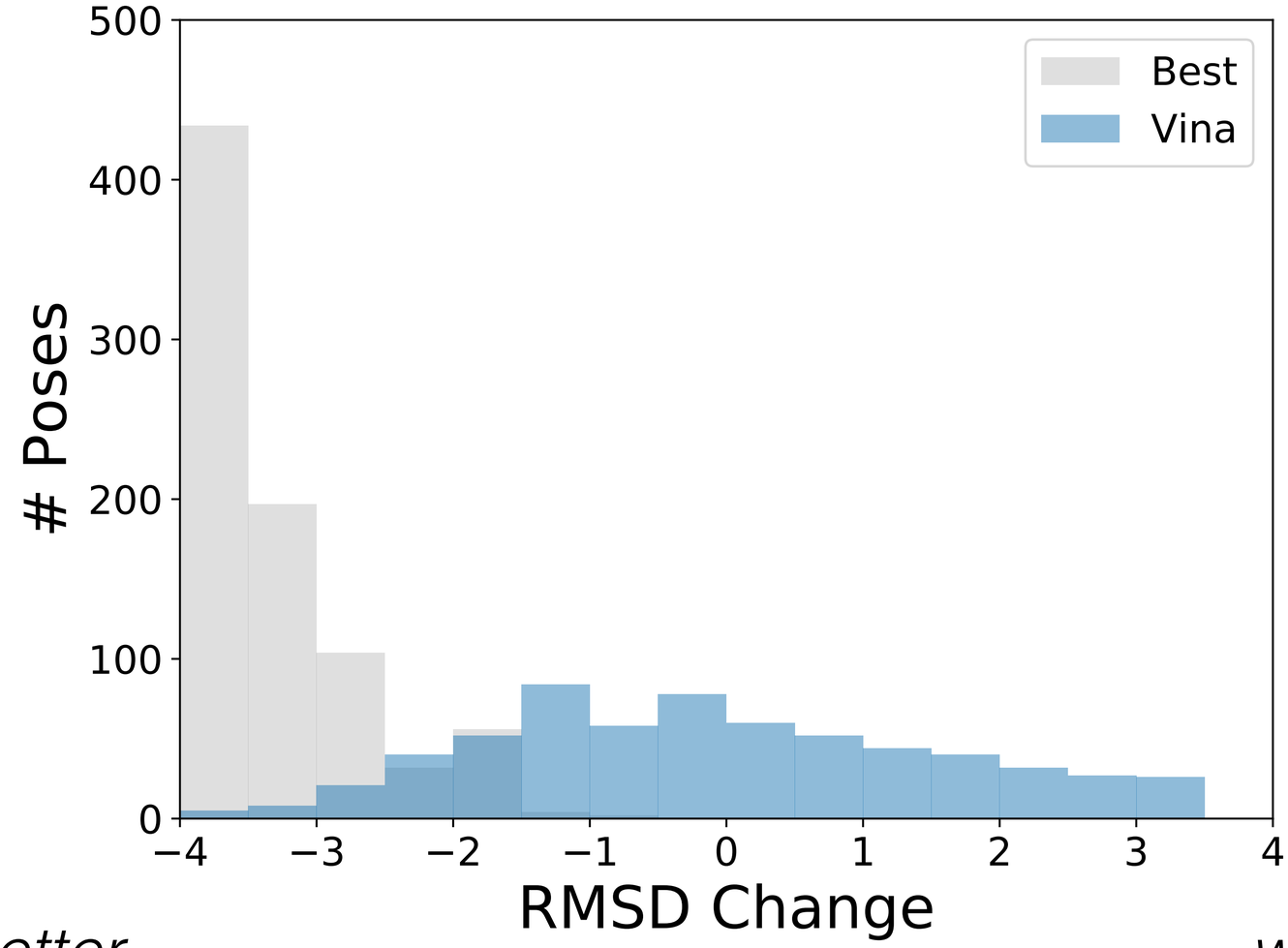
Iterative Refinement




3AO4

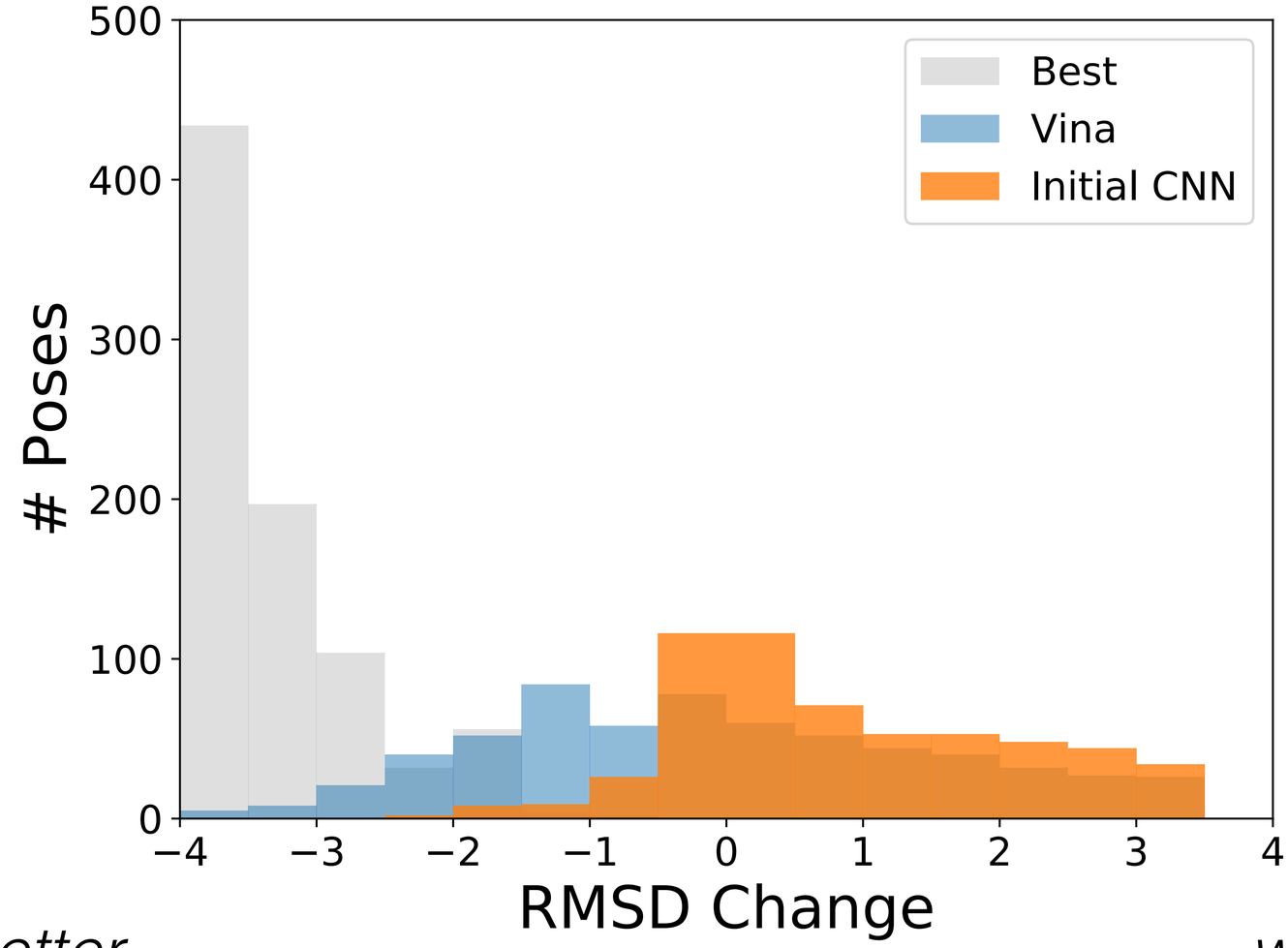

3AO4

Minimizing Random Poses



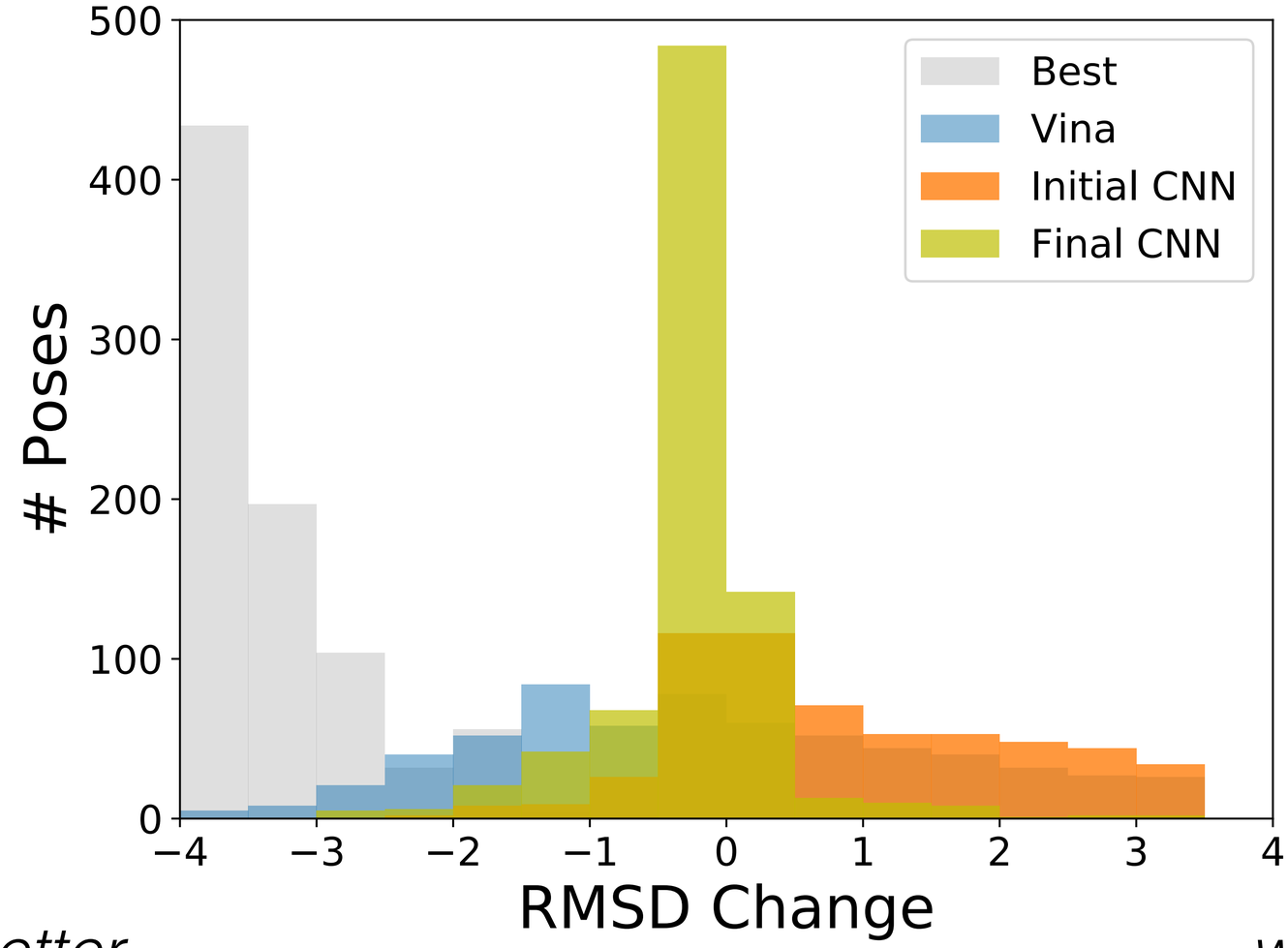
better *worse*

Minimizing Random Poses



better *worse*

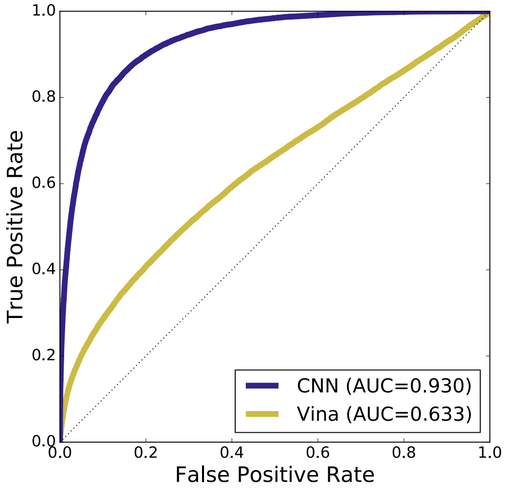
Minimizing Random Poses



better *worse*

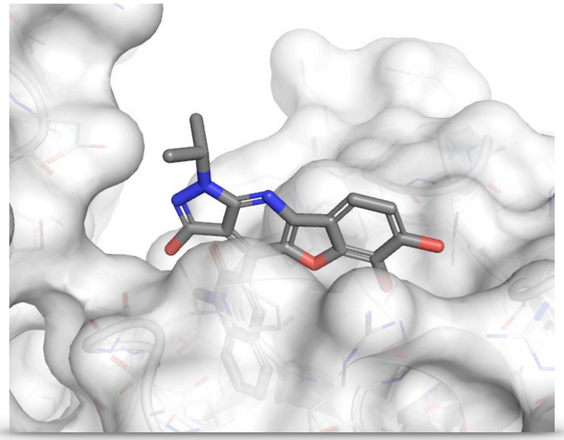
The Future

Pose Selection



Iterative Training

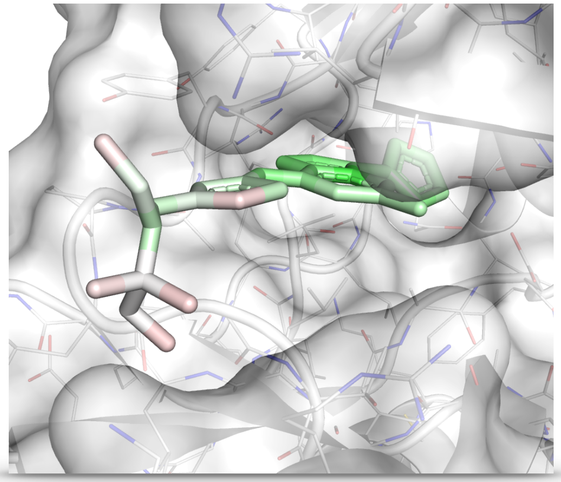
Pose *Generation*



Virtual Screening

Iterative Training

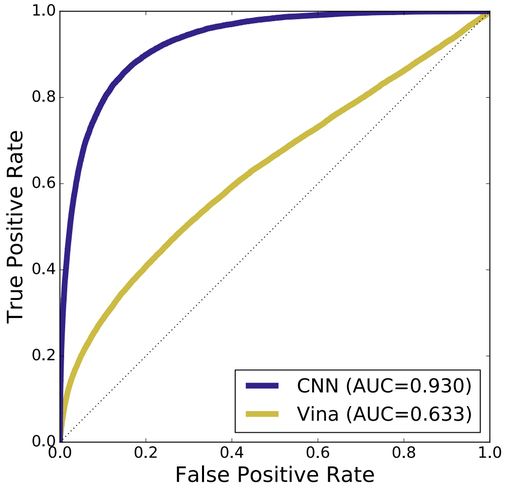
Compound Generation



Lead Optimization

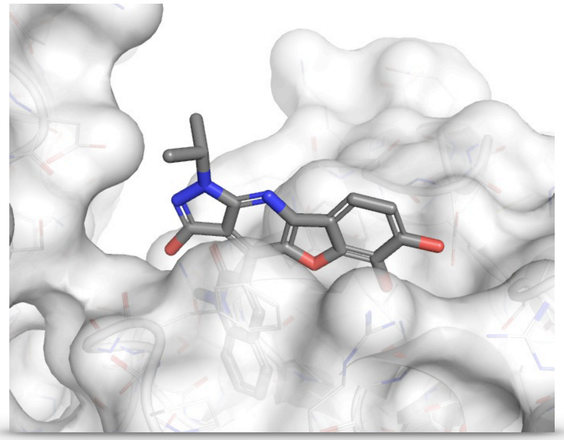
The Future

Pose Selection



Iterative Training

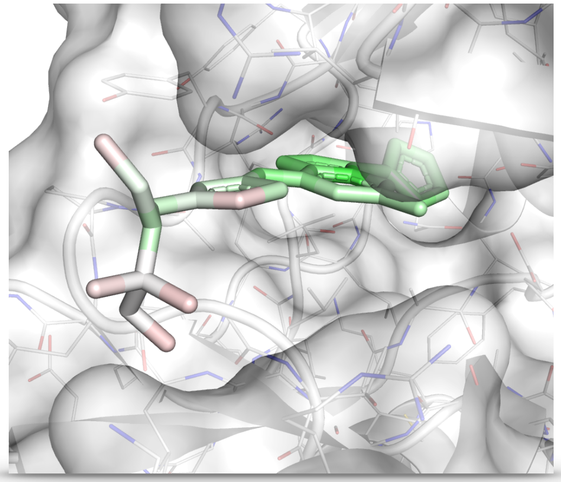
Pose *Generation*



Virtual Screening

Iterative Training

Compound Generation



Lead Optimization

Related Work

MolecuLeNet: A continuous-filter convolutional neural network for modeling quantum interactions

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, Klaus-Robert Müller

(Submitted on 26 Jun 2017)

Automatic chemical design using a data-driven continuous representation of molecules

Rafael Gómez-Bombarelli, David Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, Alán Aspuru-Guzik

(Submitted on 7 Oct 2016 (v1), last revised 6 Jan 2017 (this version, v2))

AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery

Izhar Wallach, Michael Dzamba, Abraham Heifets

(Submitted on 10 Oct 2015)

ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost

Justin S. Smith, Olexandr Isayev, Adrian E. Roitberg

(Submitted on 27 Oct 2016 (v1), last revised 6 Feb 2017 (this version, v4))

Convolutional Networks on Graphs for Learning Molecular Fingerprints

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, Ryan P. Adams

(Submitted on 30 Sep 2015 (v1), last revised 3 Nov 2015 (this version, v2))

Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity

Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, Vijay S. Pande

(Submitted on 30 Mar 2017)

Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules

Alessandro Lusci*†, Gianluca Pollastri†, and Pierre Baldi*‡

† School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland

‡ Department of Computer Science, University of California, Irvine, Irvine, California 92697, United States

J. Chem. Inf. Model., **2013**, *53* (7), pp 1563–1575

DOI: 10.1021/ci400187y

Publication Date (Web): June 24, 2013

Low Data Drug Discovery with One-shot Learning

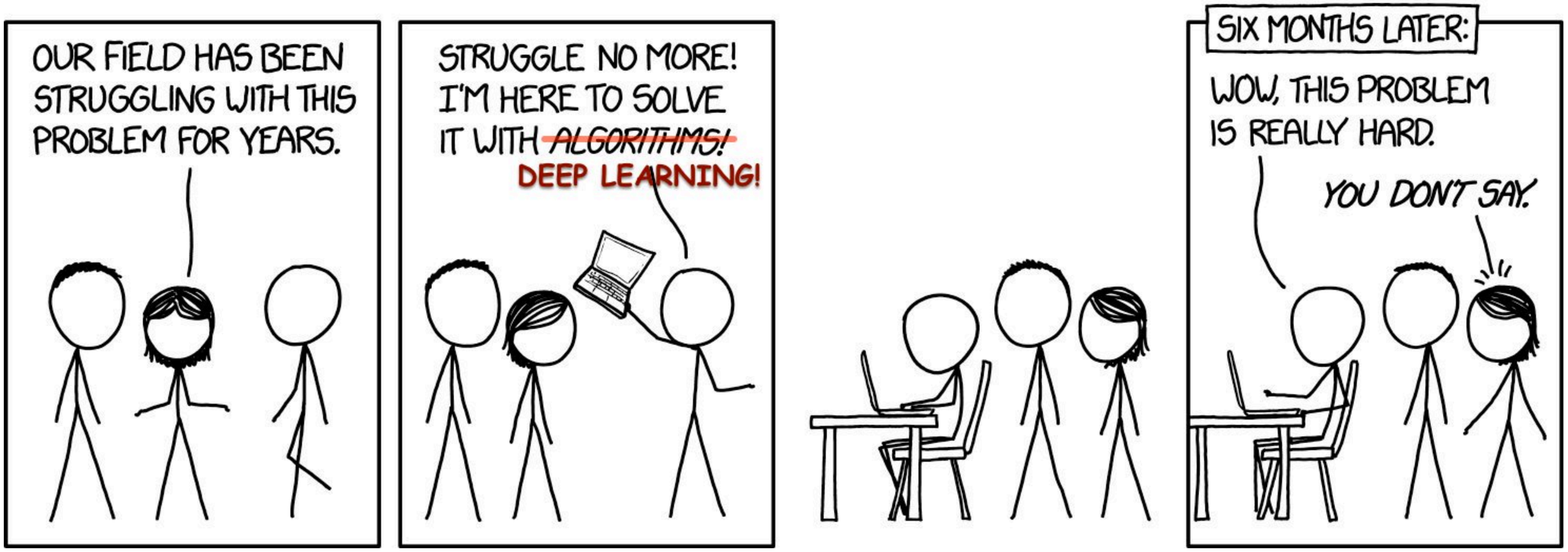
Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, Vijay Pande

(Submitted on 10 Nov 2016)

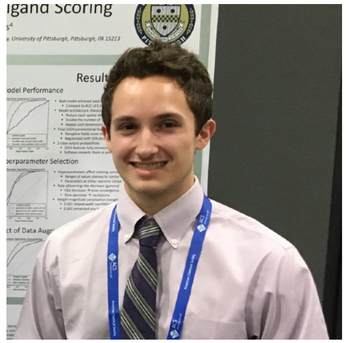
Massively Multitask Networks for Drug Discovery

Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, Vijay Pande

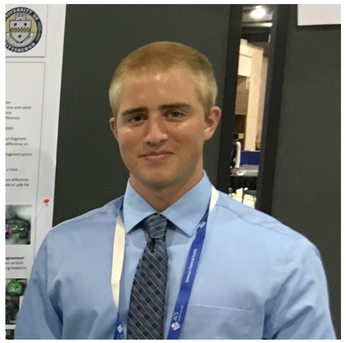
(Submitted on 6 Feb 2015)


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS! DEEP LEARNING!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

Acknowledgements

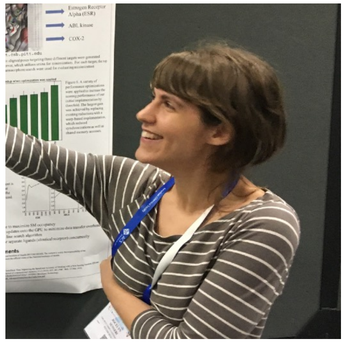


Matt Ragoza



Josh Hochuli

Elisa Idrobo



Jocelyn Sunseri

Group Members

Jocelyn Sunseri
Matt Ragoza
Josh Hochuli
Pulkit Mittal
Alec Helbling
Tamar Skaist
Christopher Dunstan

Department of Computational and Systems Biology

TECBIO REU

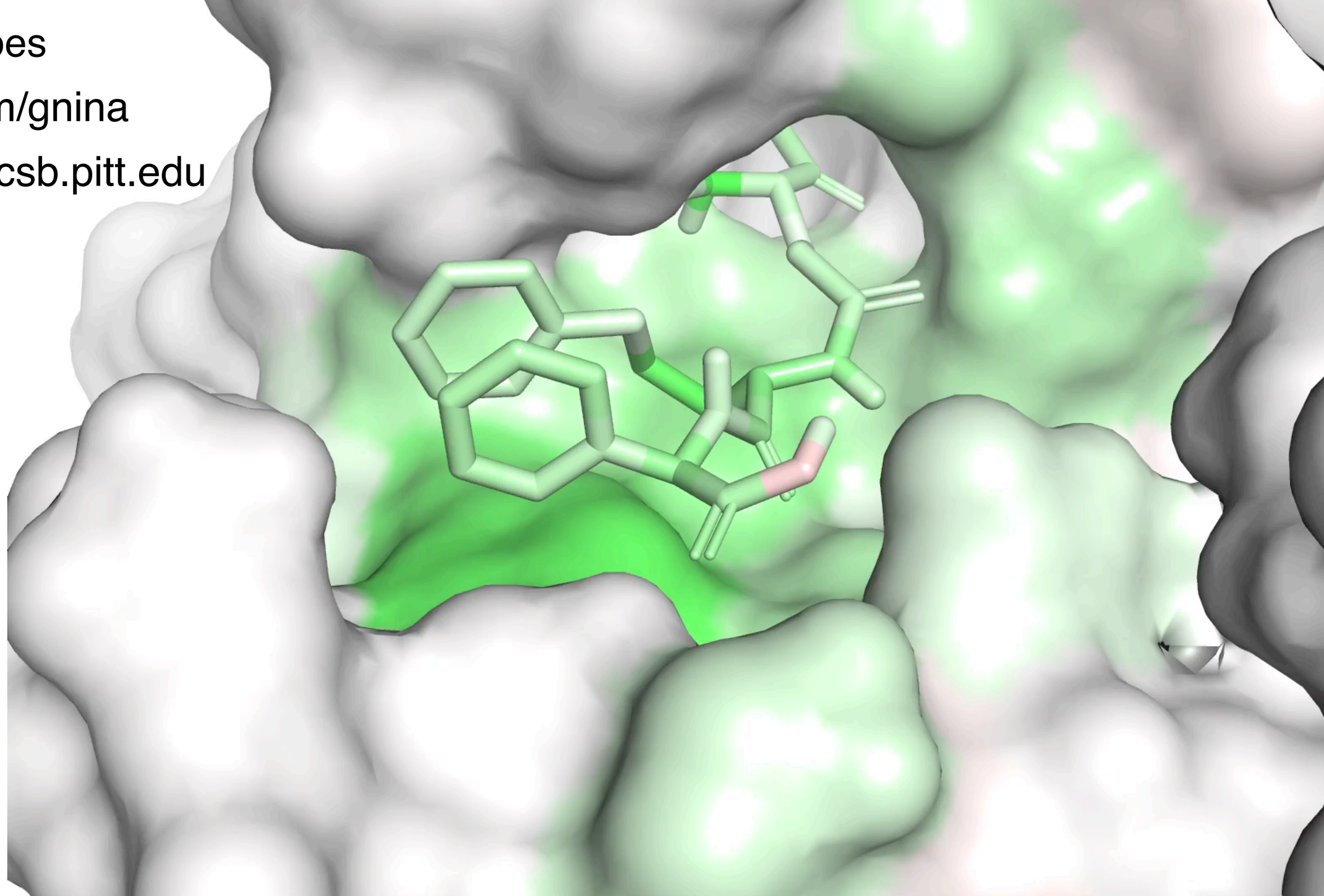
National Institute of General Medical Sciences
R01GM108340

NVIDIA

@david_koes

github.com/gnina

http://bits.csb.pitt.edu



@david_koes

github.com/gnina

http://bits.csb.pitt.edu