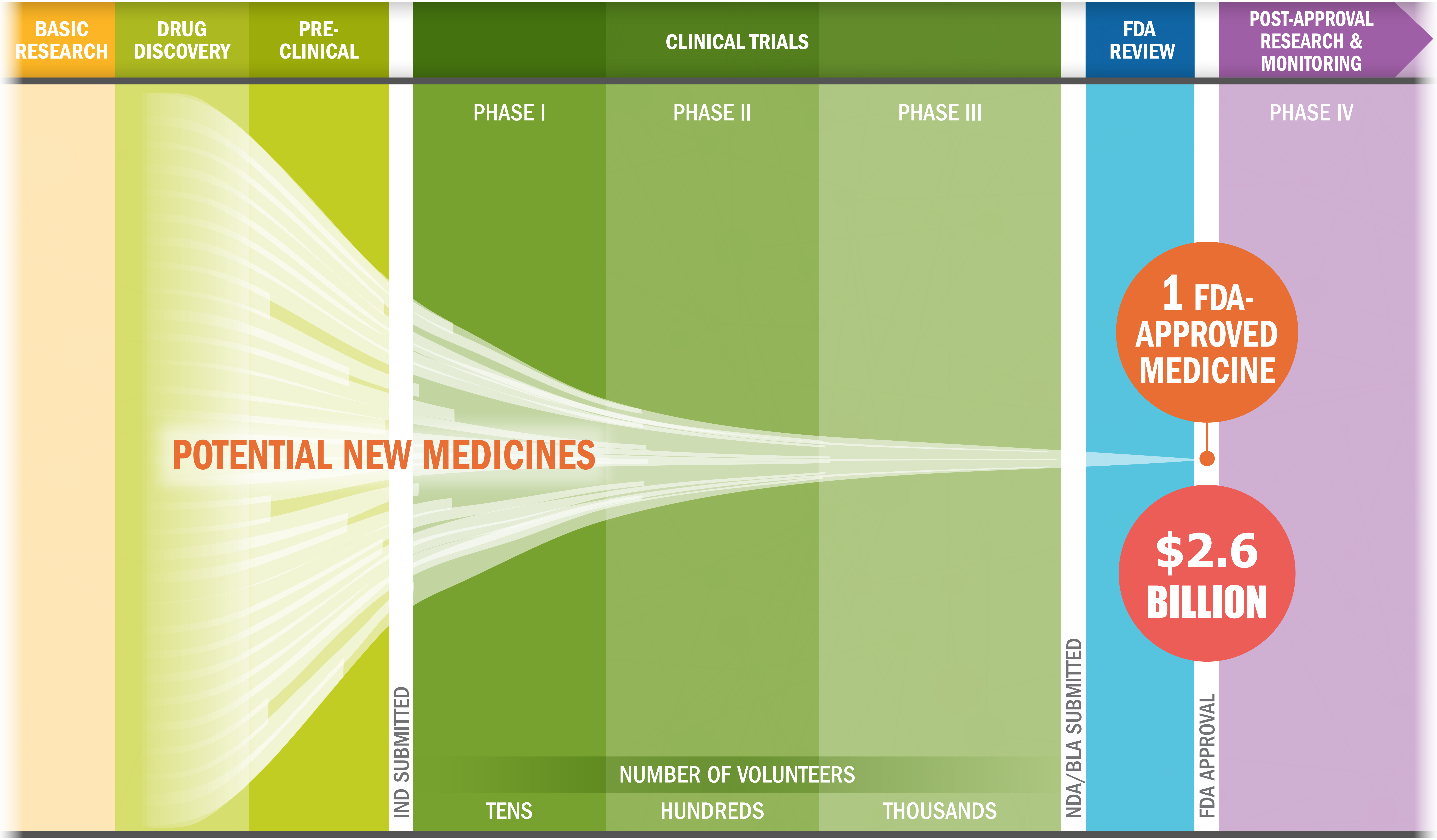
GPU-Accelerated Convolutional Neural Networks For Protein-Ligand Scoring

David Koes

@david_koes

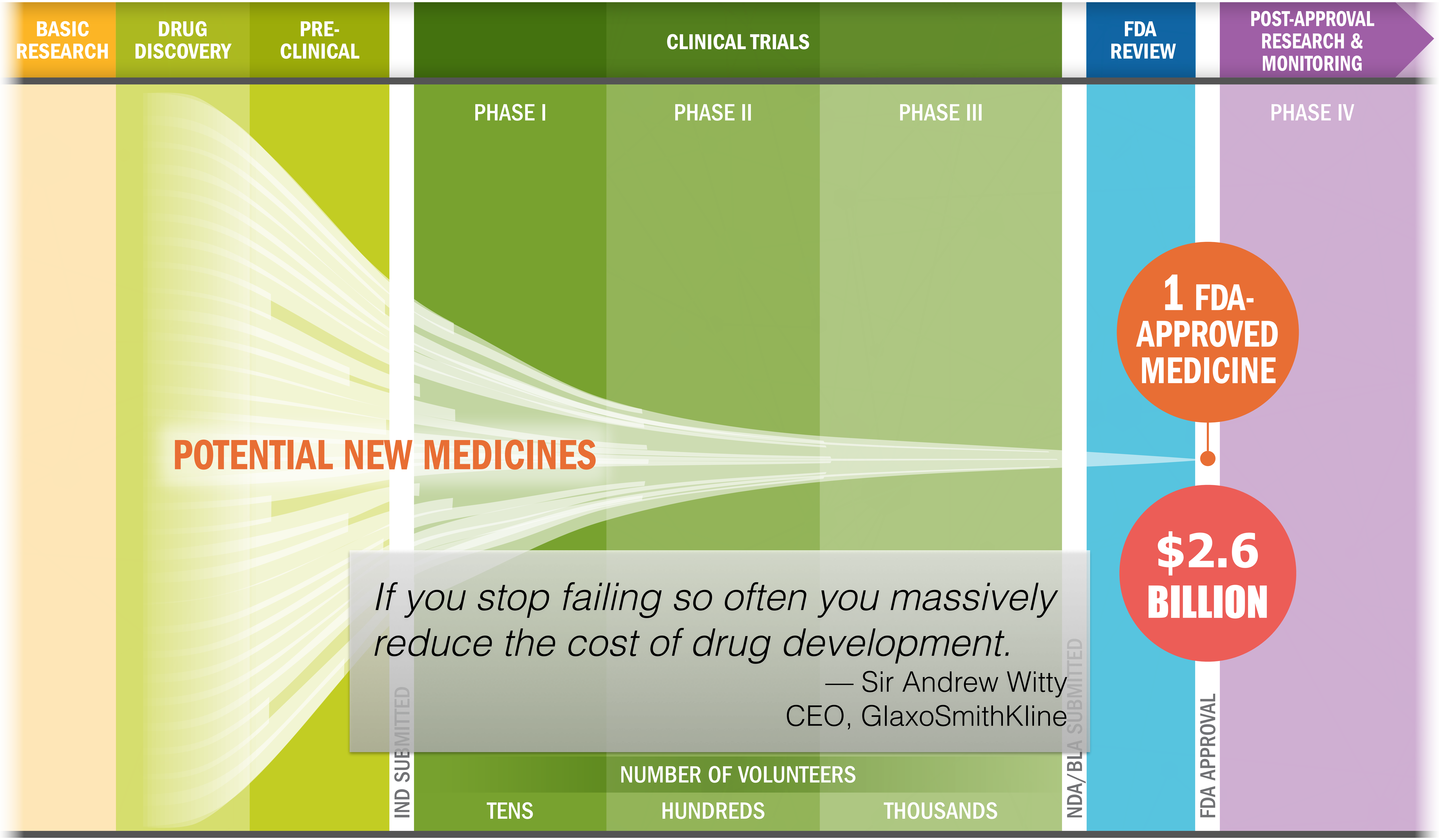
GPU Technology Conference
May 8, 2017

THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS



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

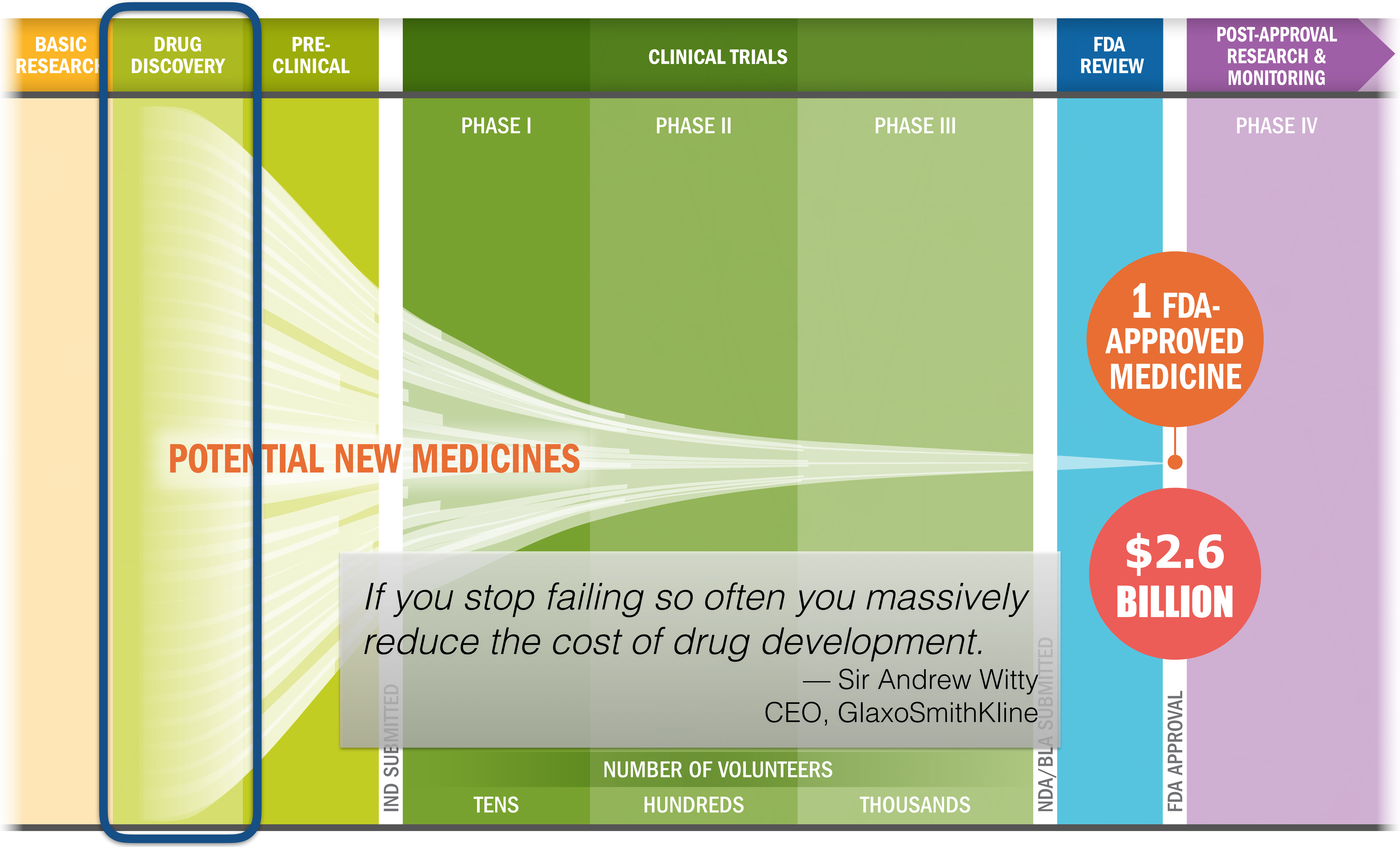
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS



> *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



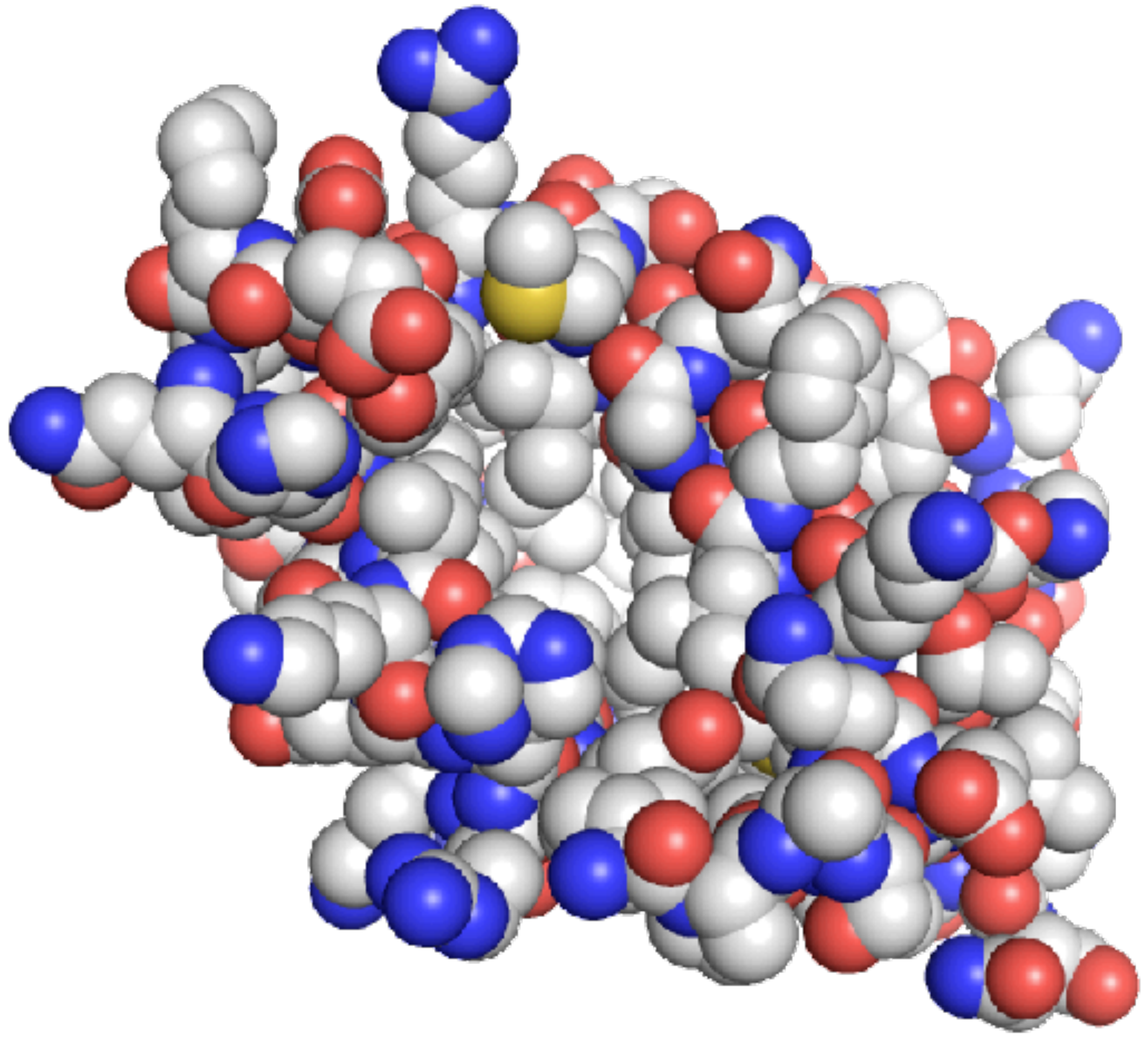
> *If you stop failing so often you massively reduce the cost of drug development.*
>
> — Sir Andrew Witty
> CEO, GlaxoSmithKline

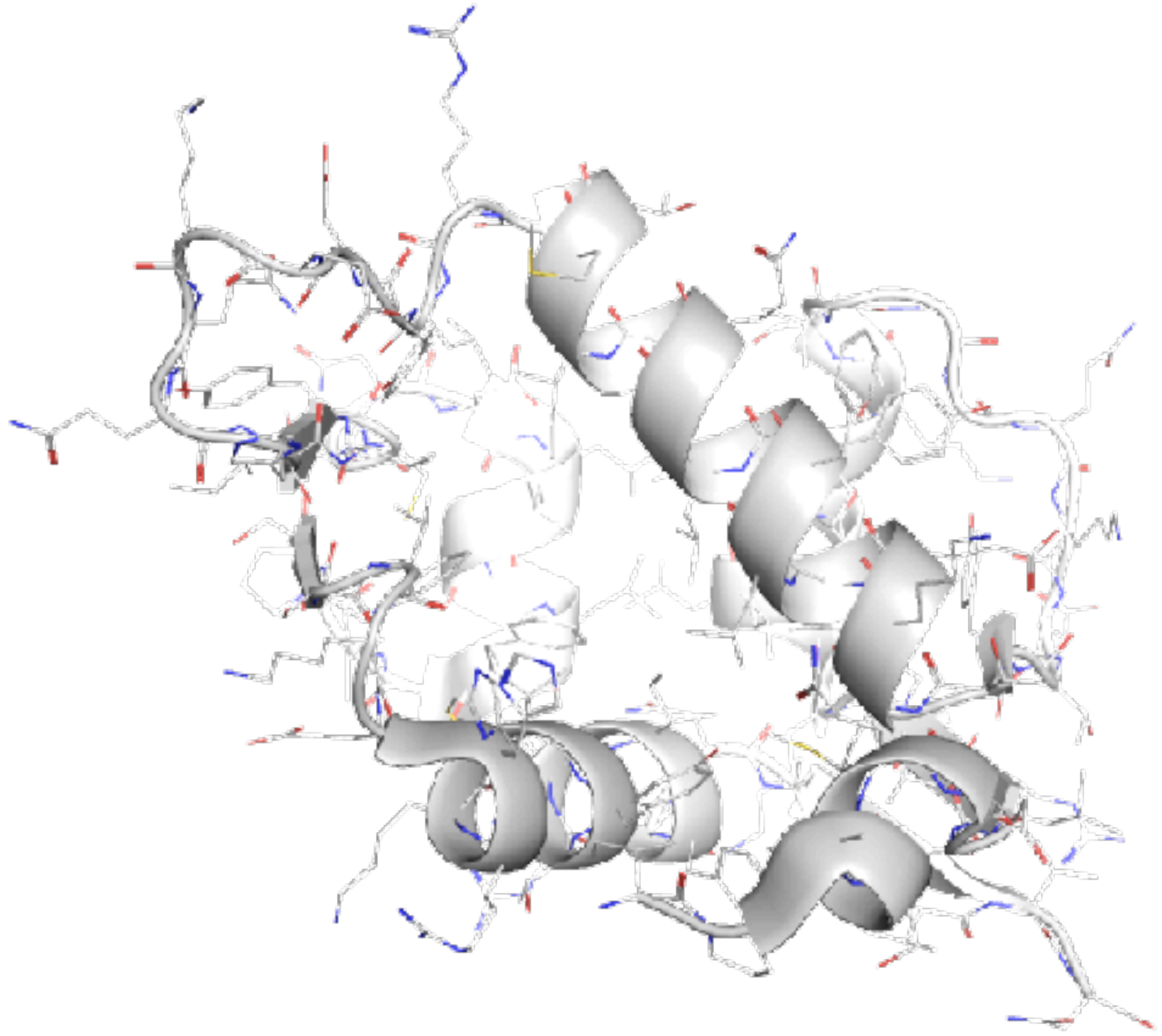
Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)

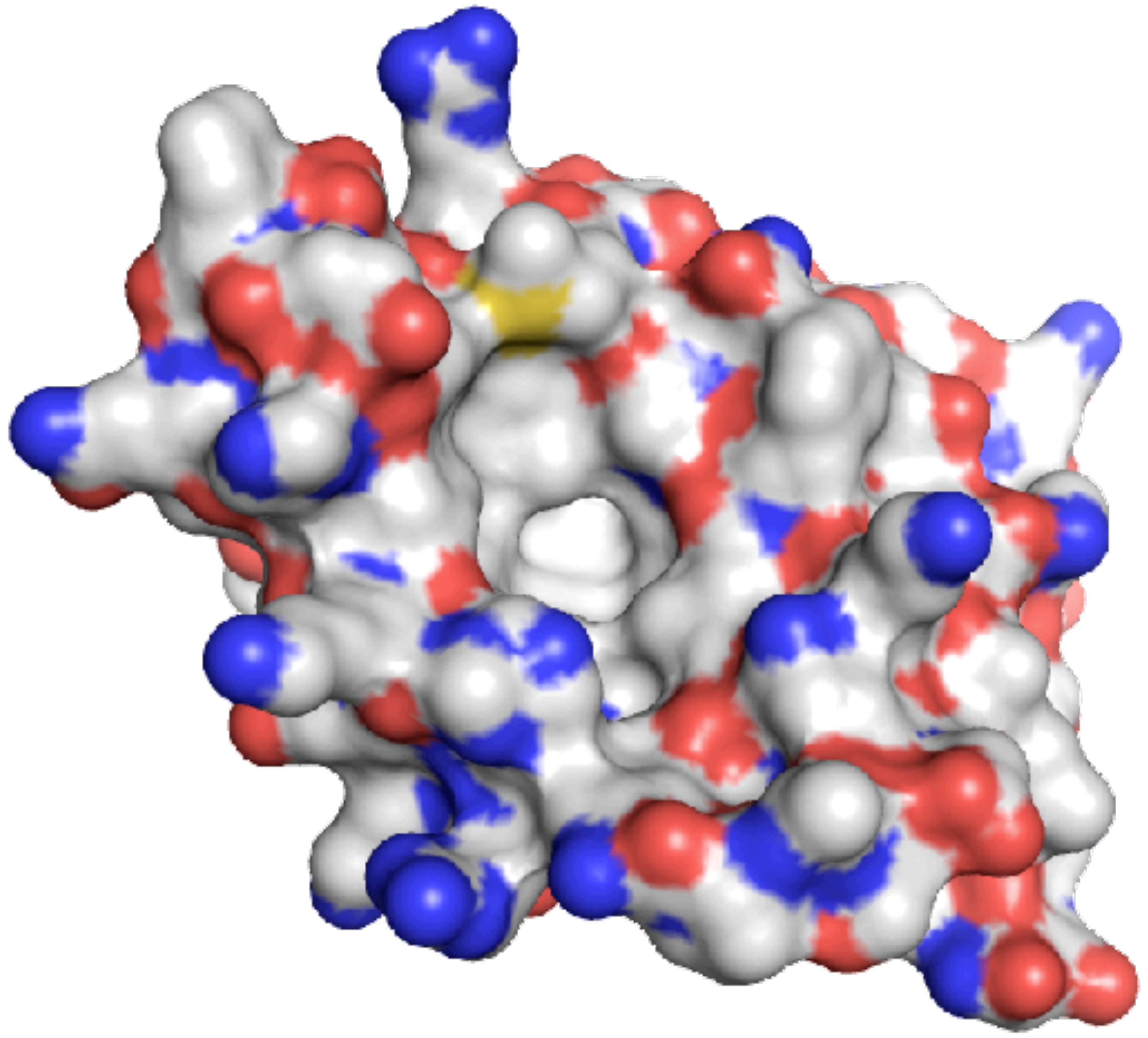
1. Does the compound do what you want it to?
2. Does the compound **not** do what you **don't** want it to?
3. Is what you want it to do the right thing?

Protein Structures

sequence → **structure** → function

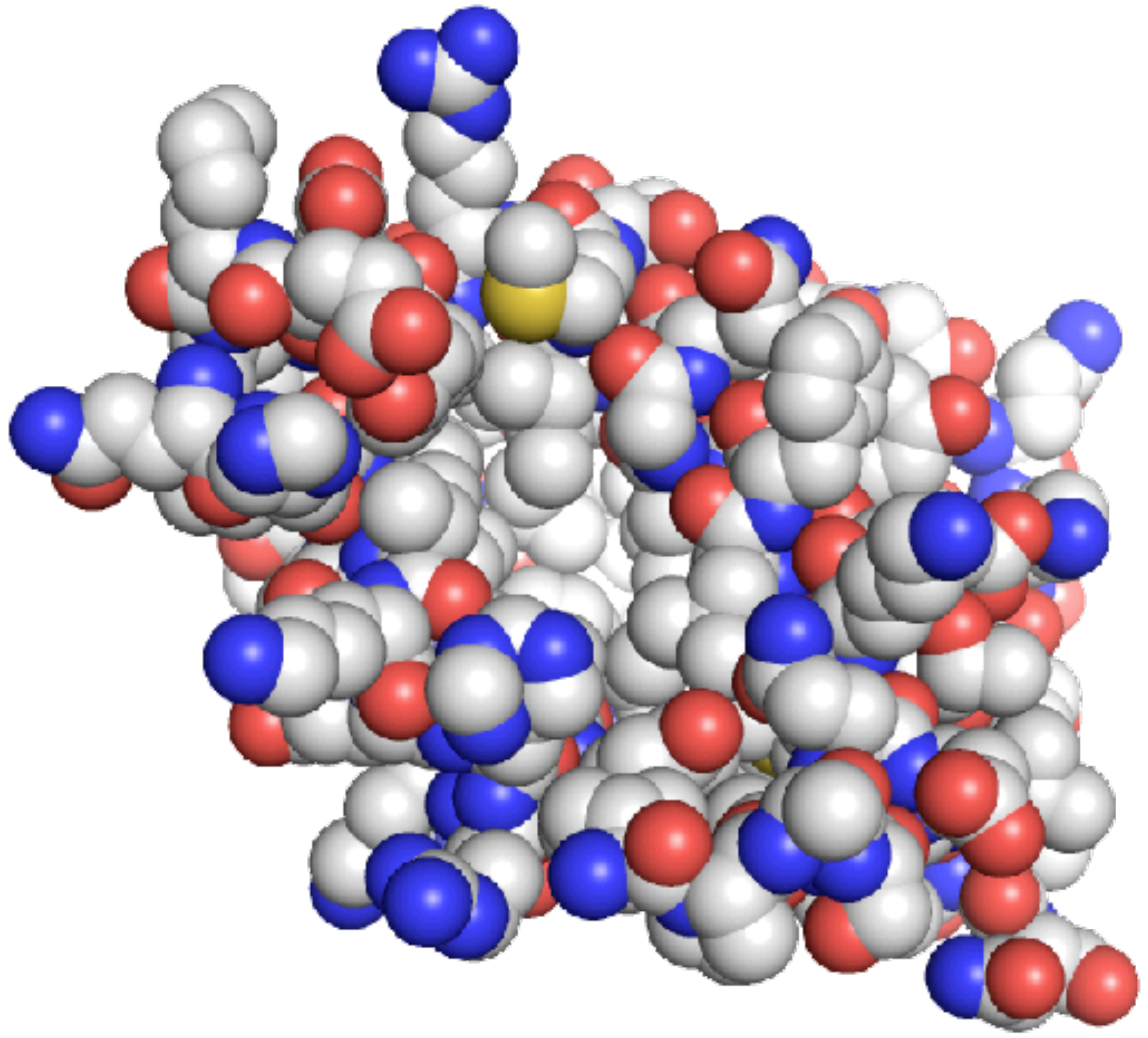


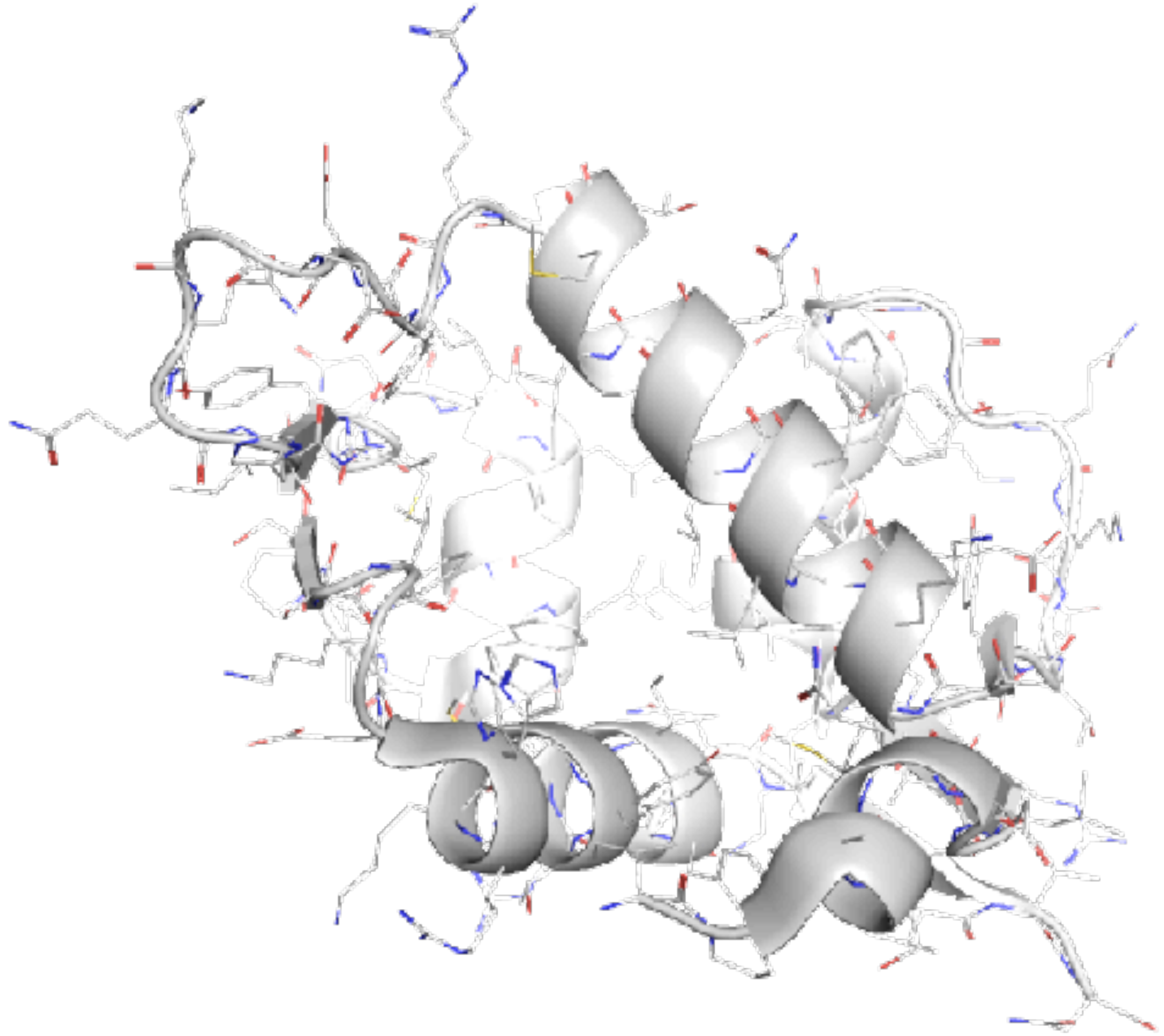


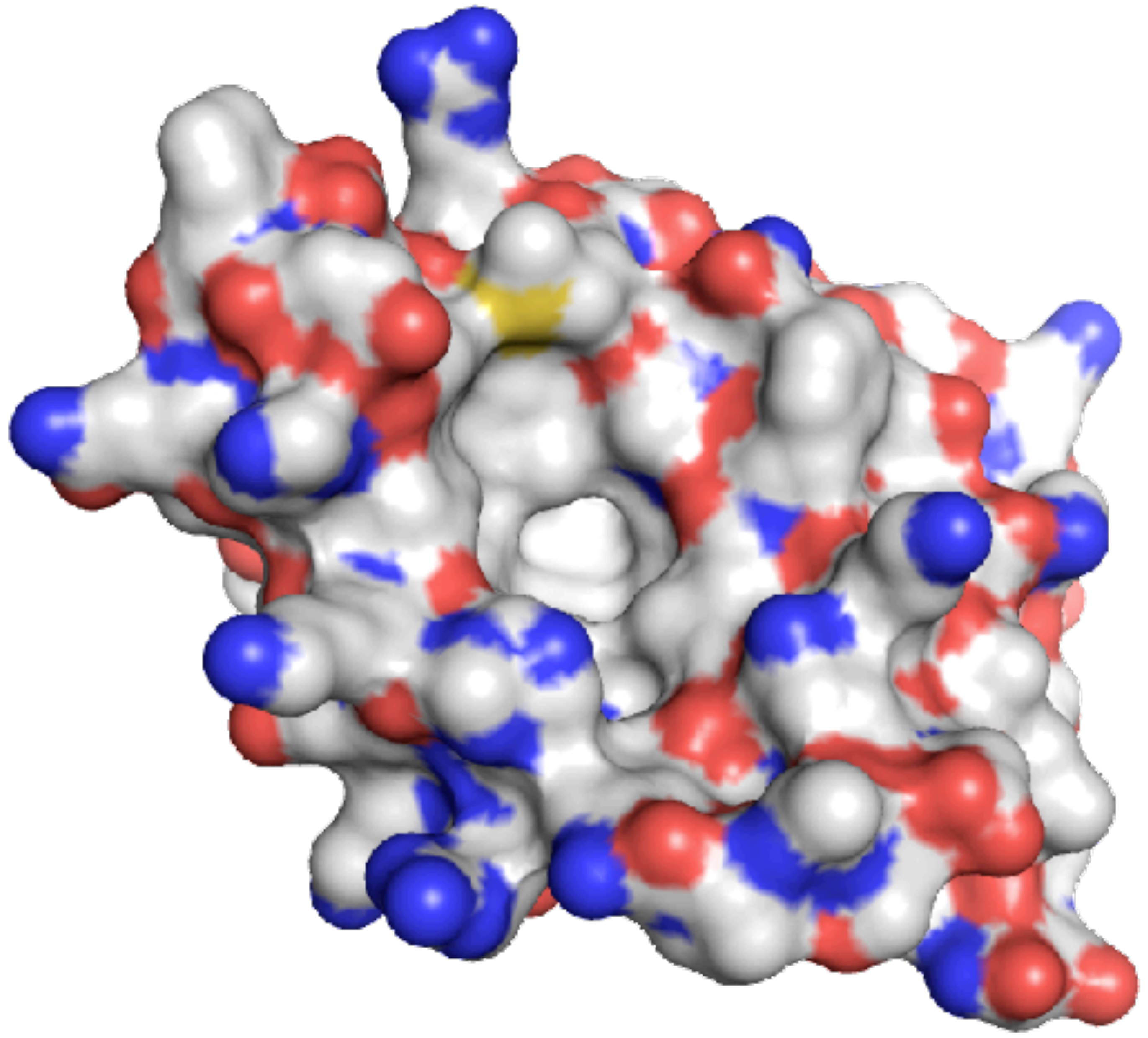


Protein Structures

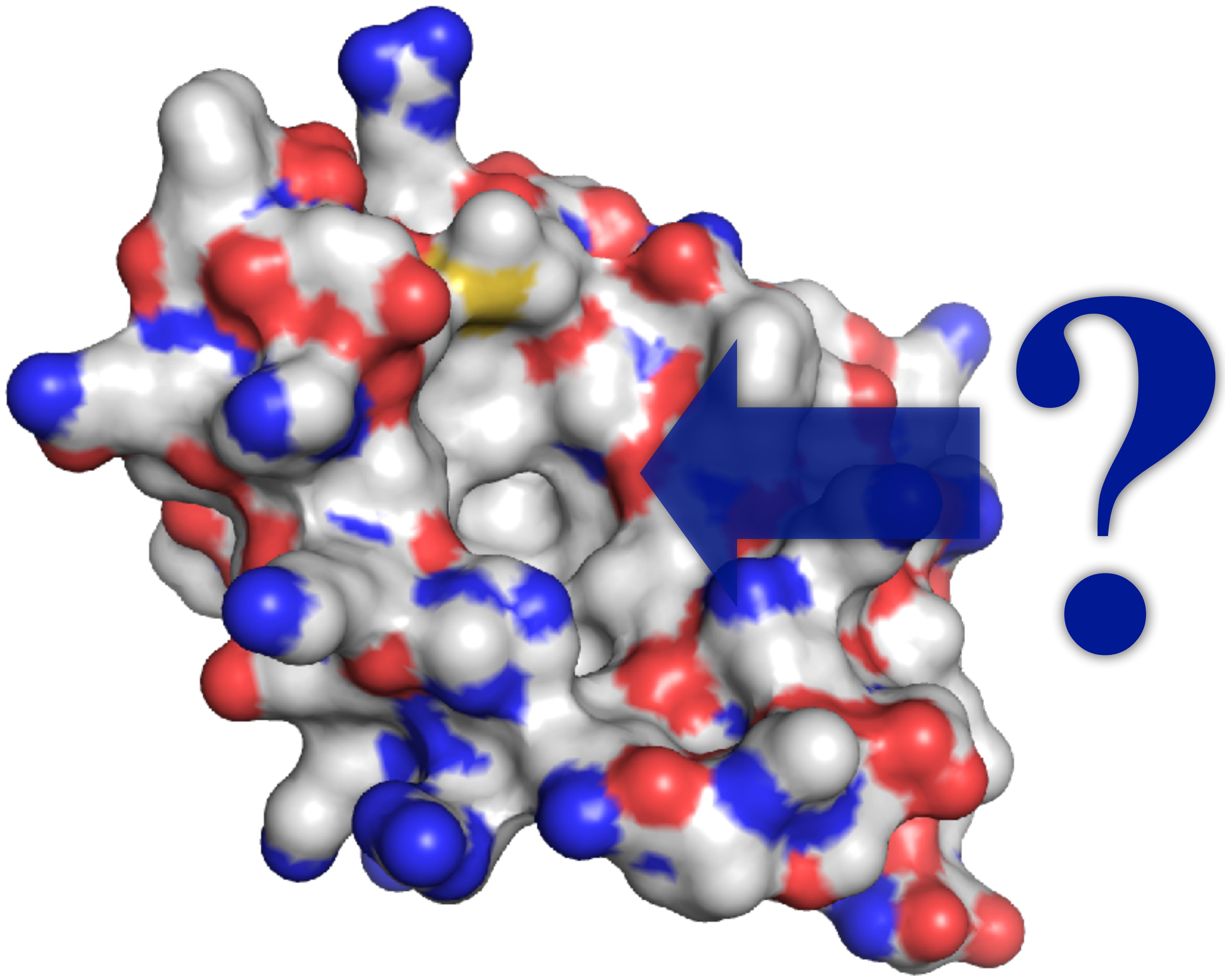
sequence → **structure** → function







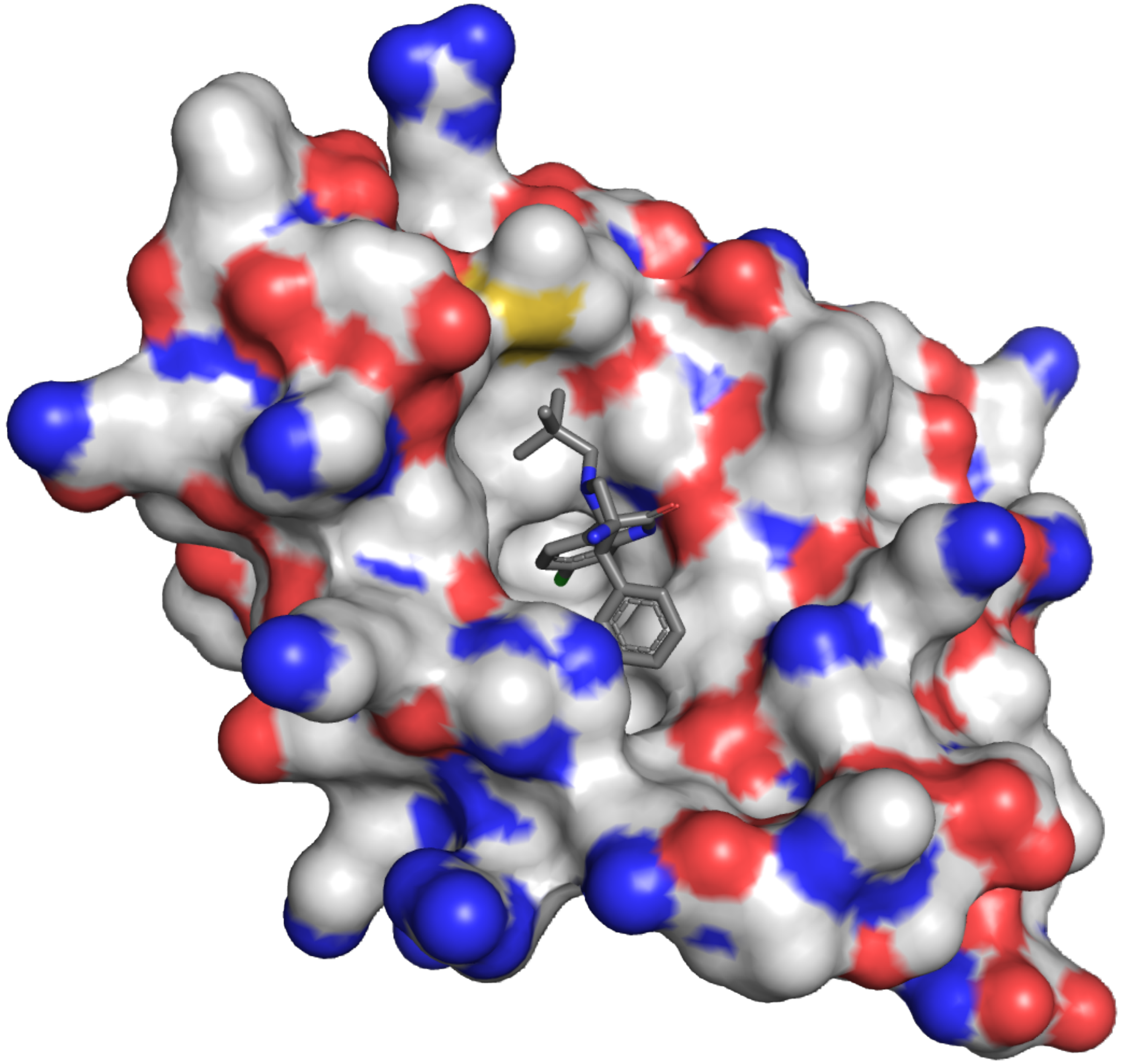
Structure Based Drug Design



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

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

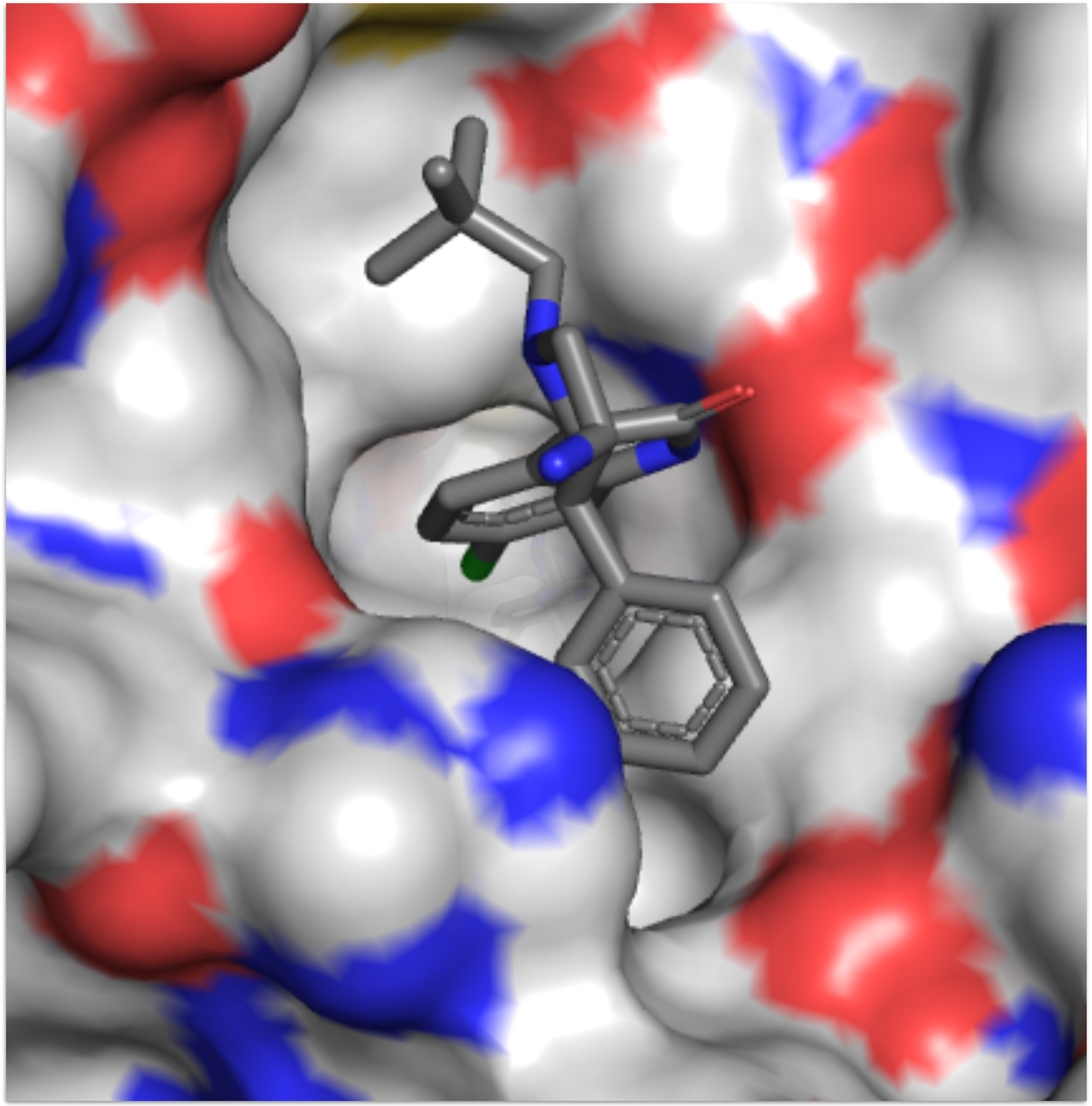
Structure Based Drug Design



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

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

Structure Based Drug Design

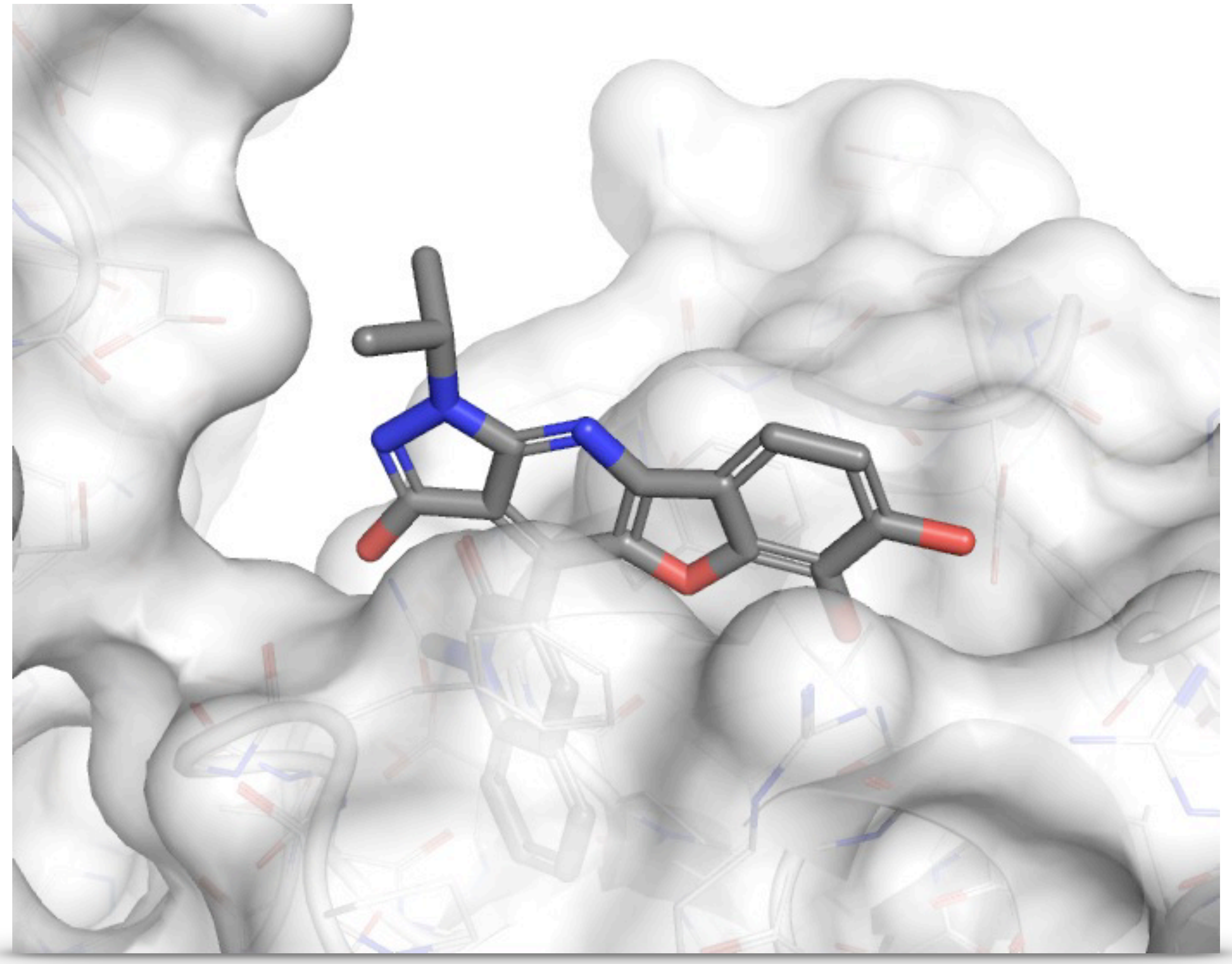


Unlike ligand based approaches, **generalizes to new targets**

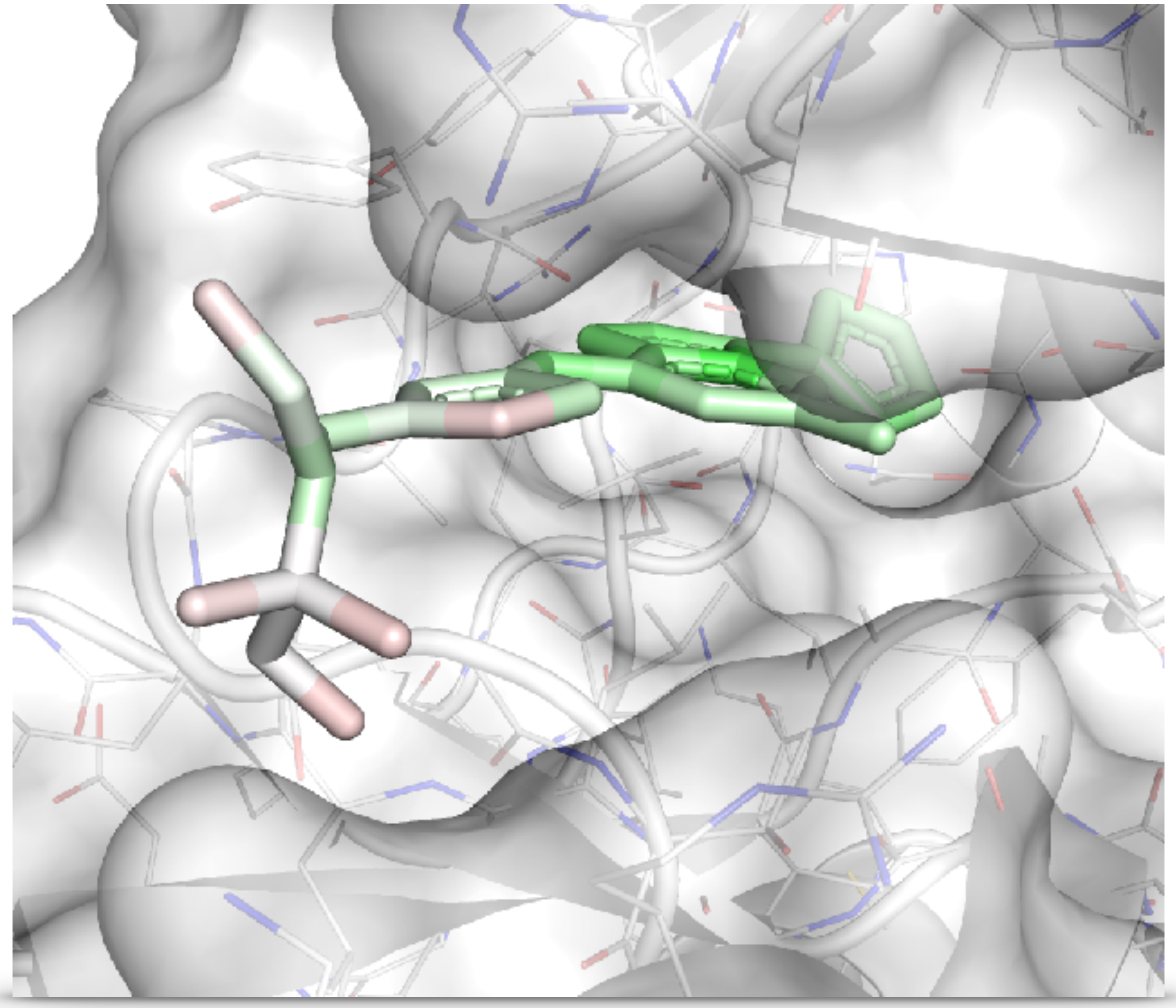
Requires **molecular target** with **known structure** and **binding site**

Structure Based Drug Design

Virtual Screening



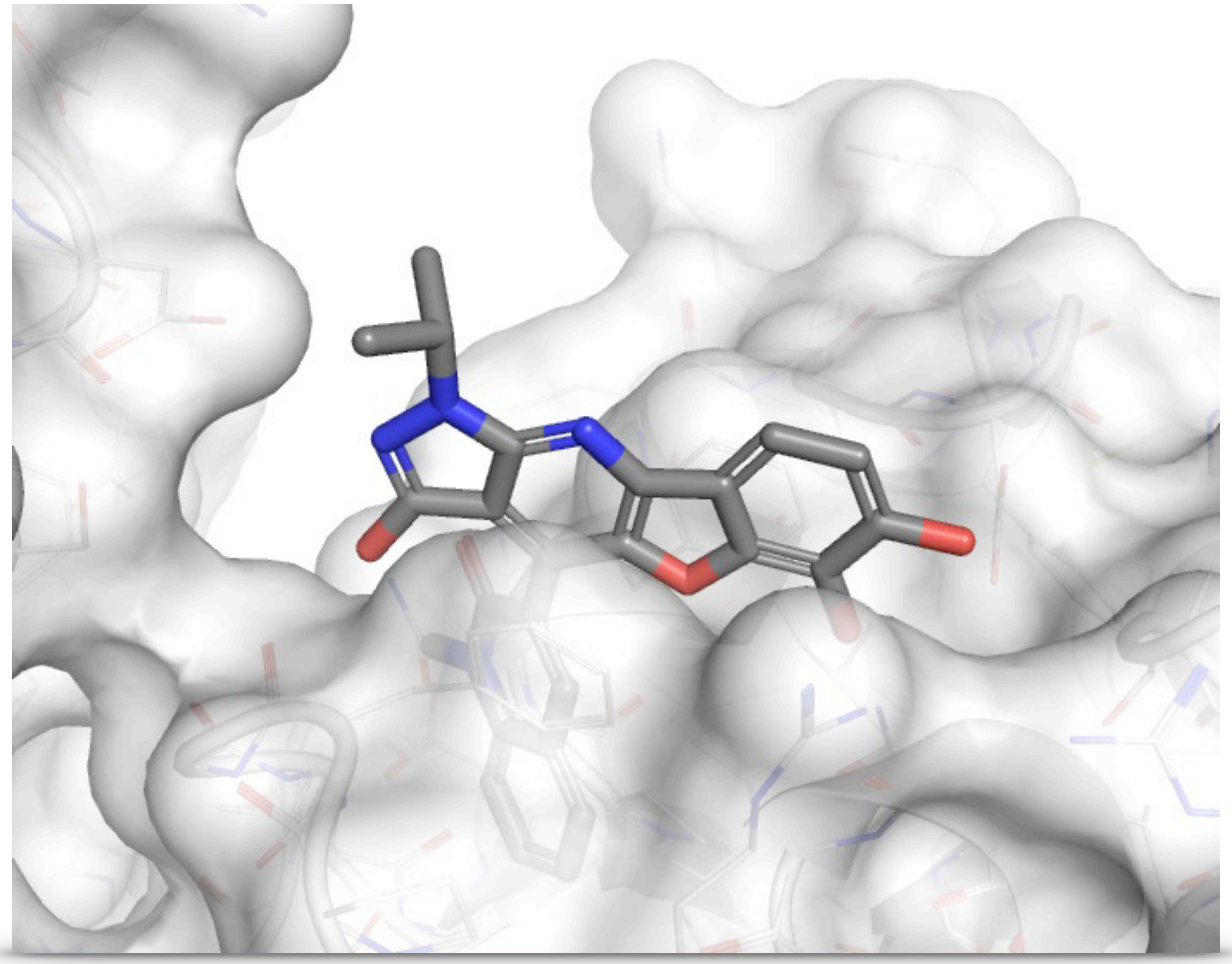
Lead Optimization



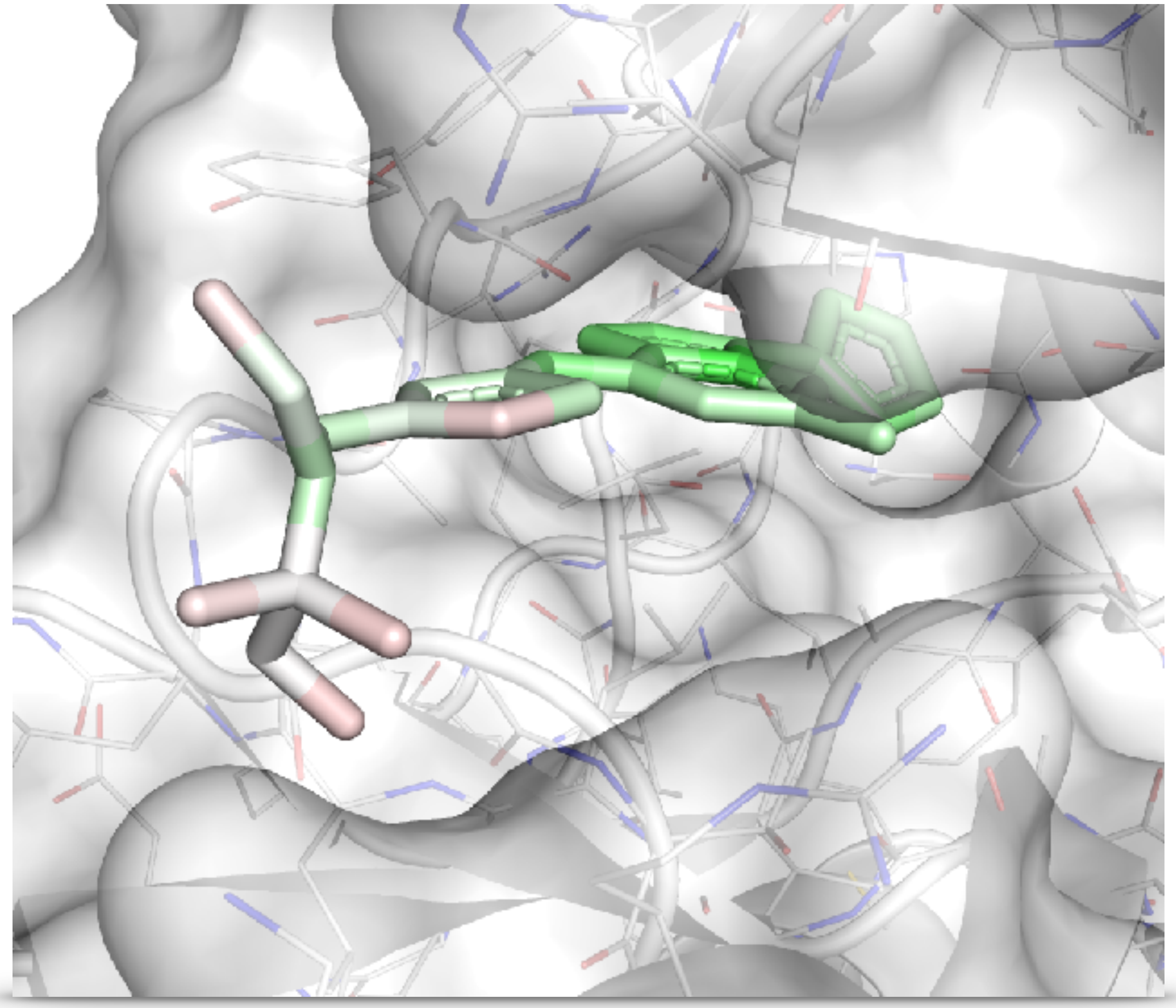
Pose Prediction → Binding Discrimination → Affinity Prediction

Structure Based Drug Design

Virtual Screening

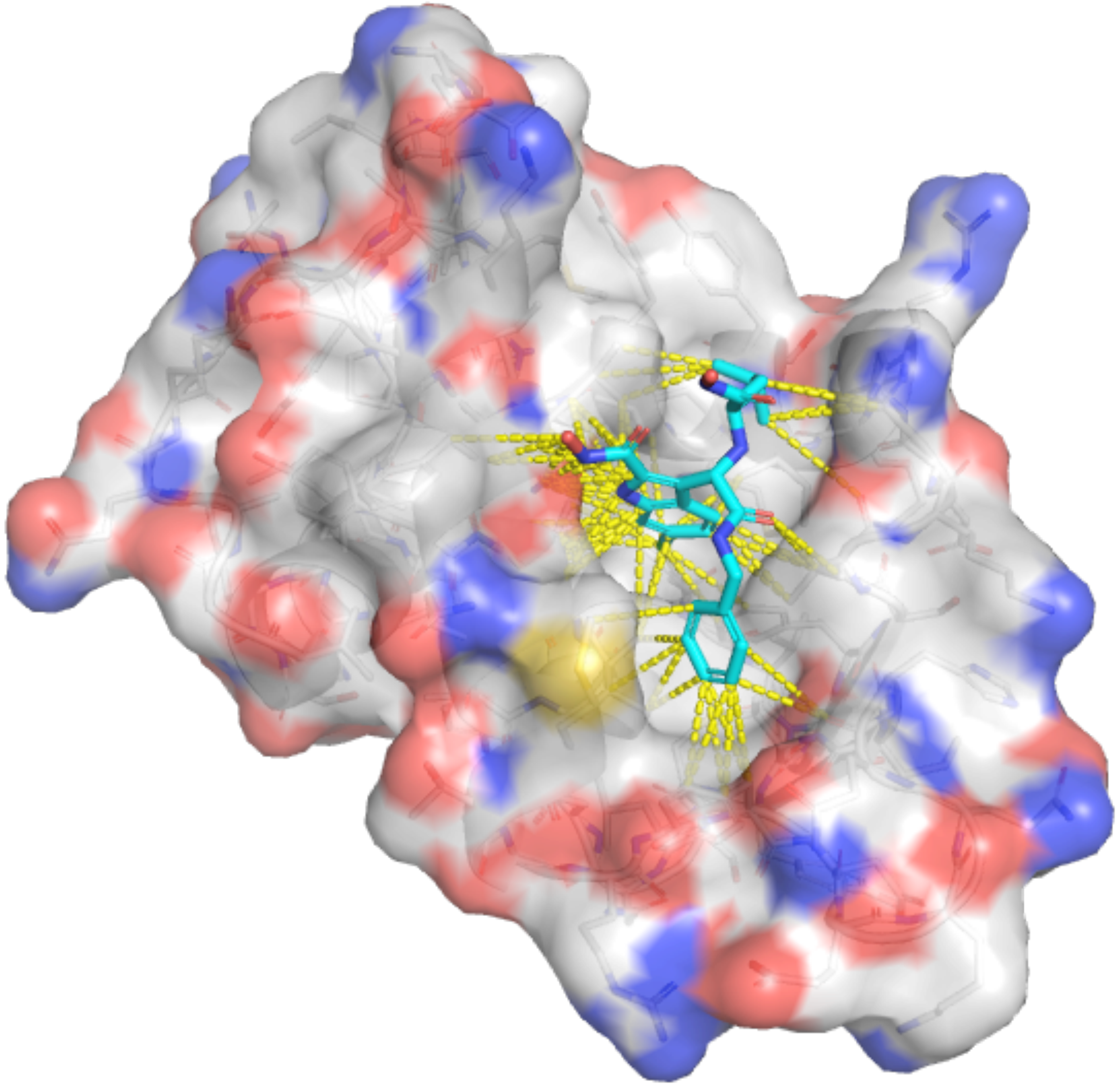


Lead Optimization

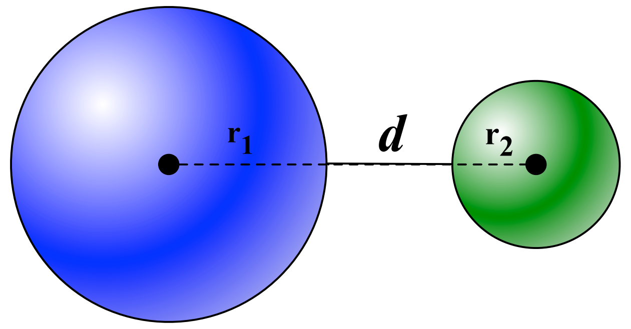


Pose Prediction Binding Discrimination Affinity Prediction

Protein-Ligand Scoring



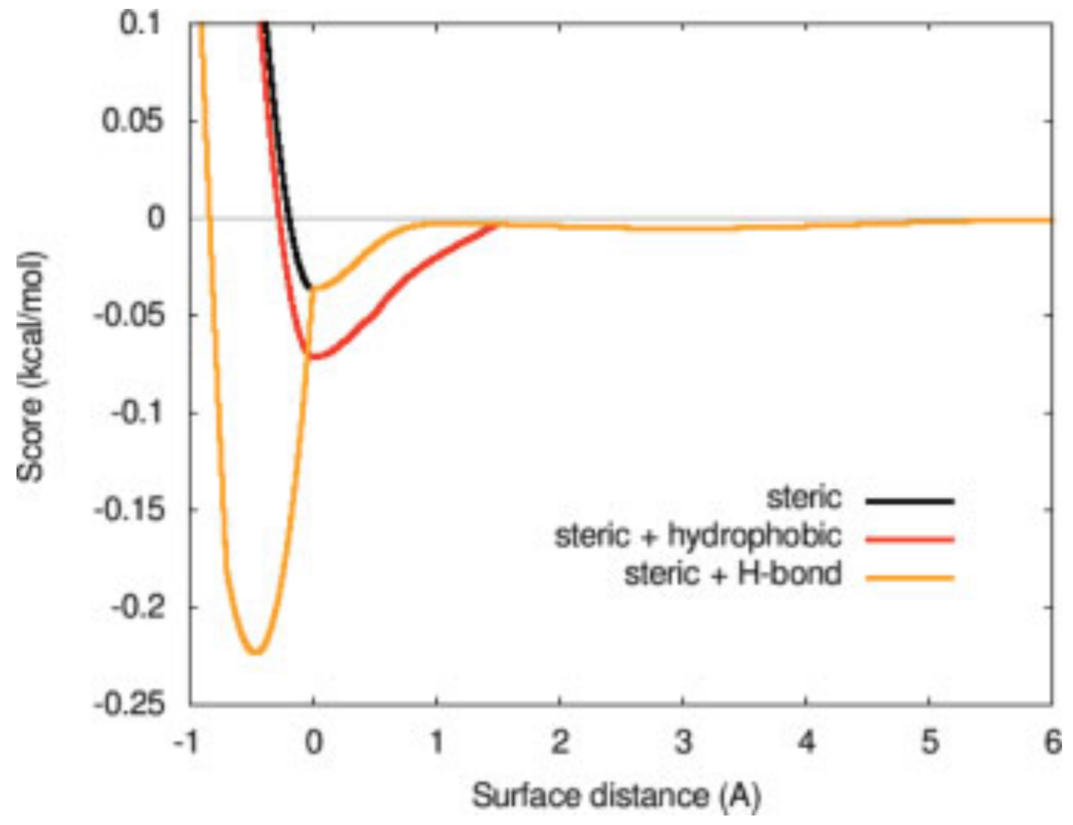
AutoDock Vina



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

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

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



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?

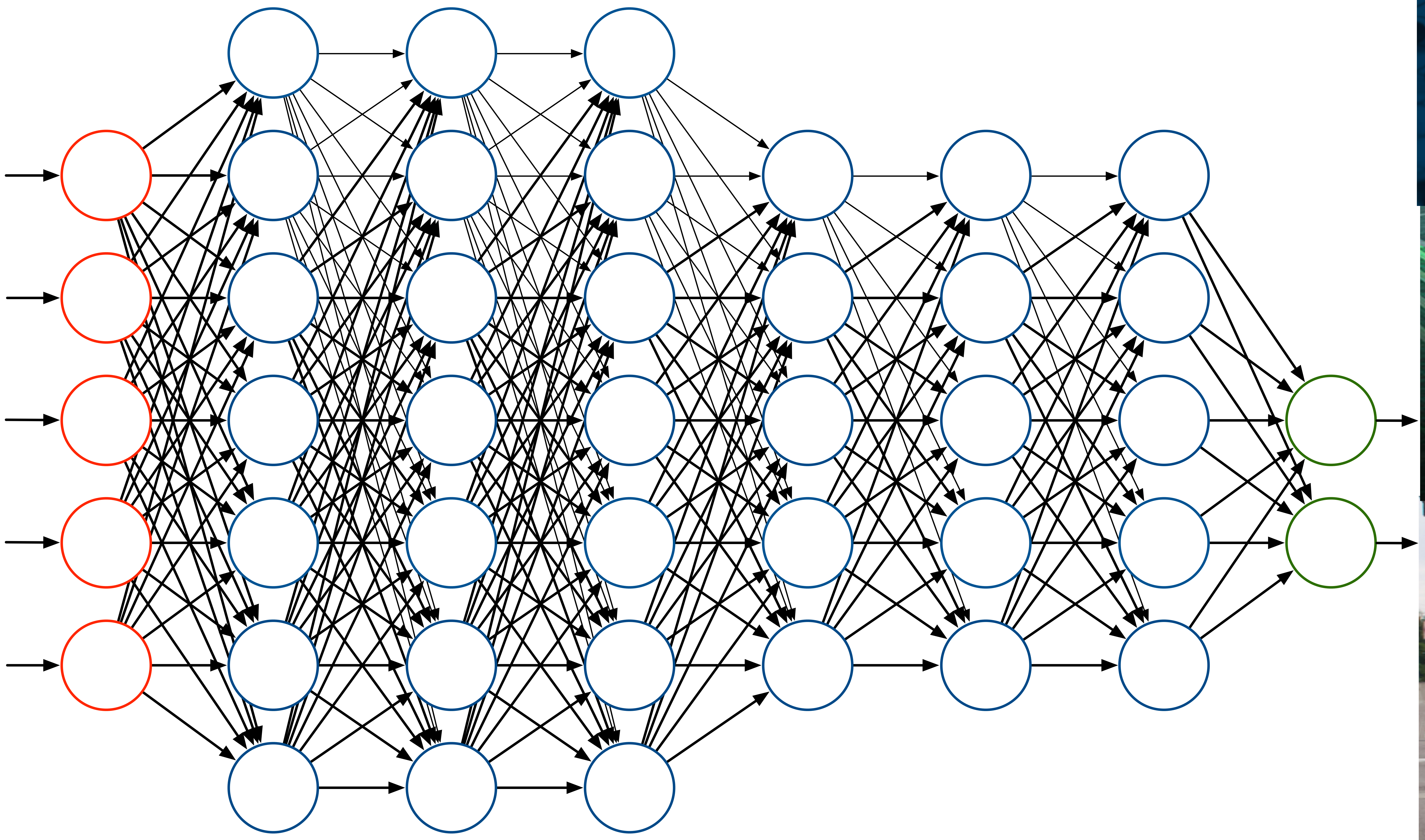
Can we do better?

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

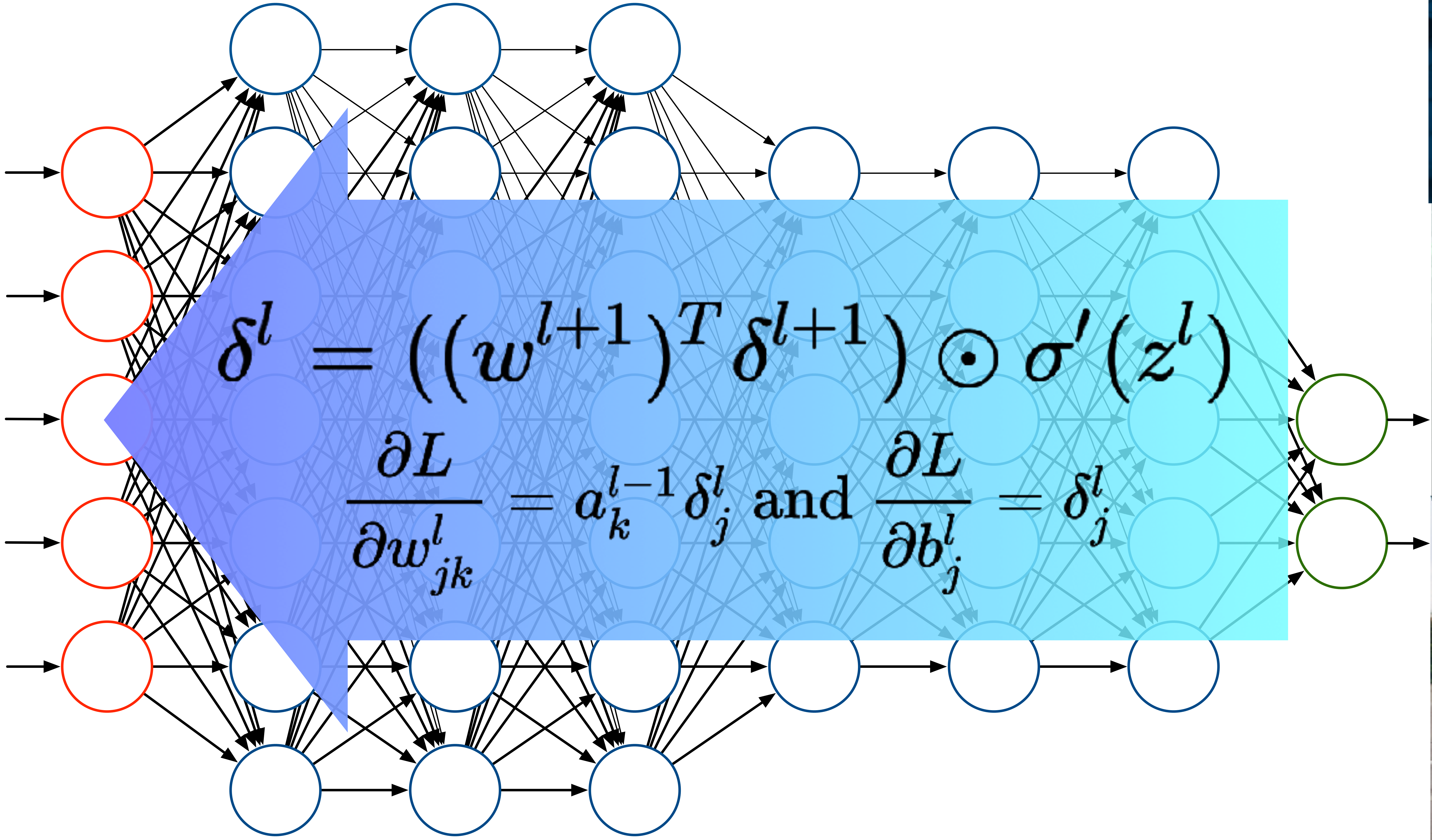
Key Idea: Leverage "big data"

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

Deep Learning



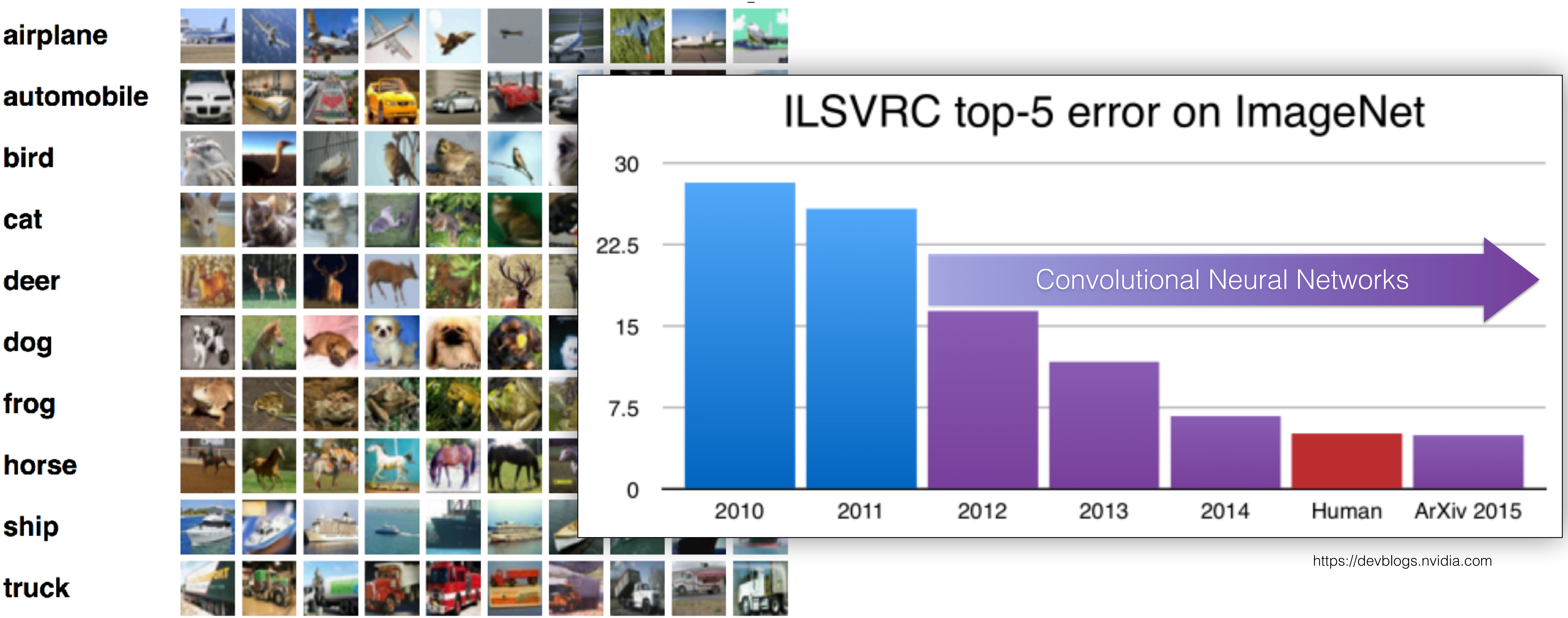
Deep Learning



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

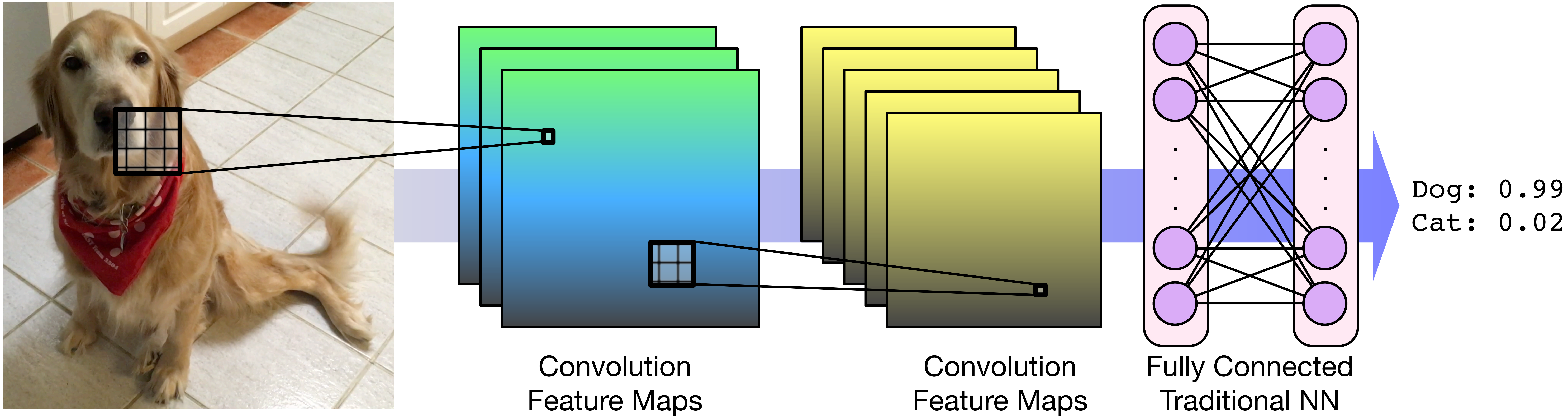
$$\frac{\partial L}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j \text{ and } \frac{\partial L}{\partial b^l_j} = \delta^l_j$$

Image Recognition

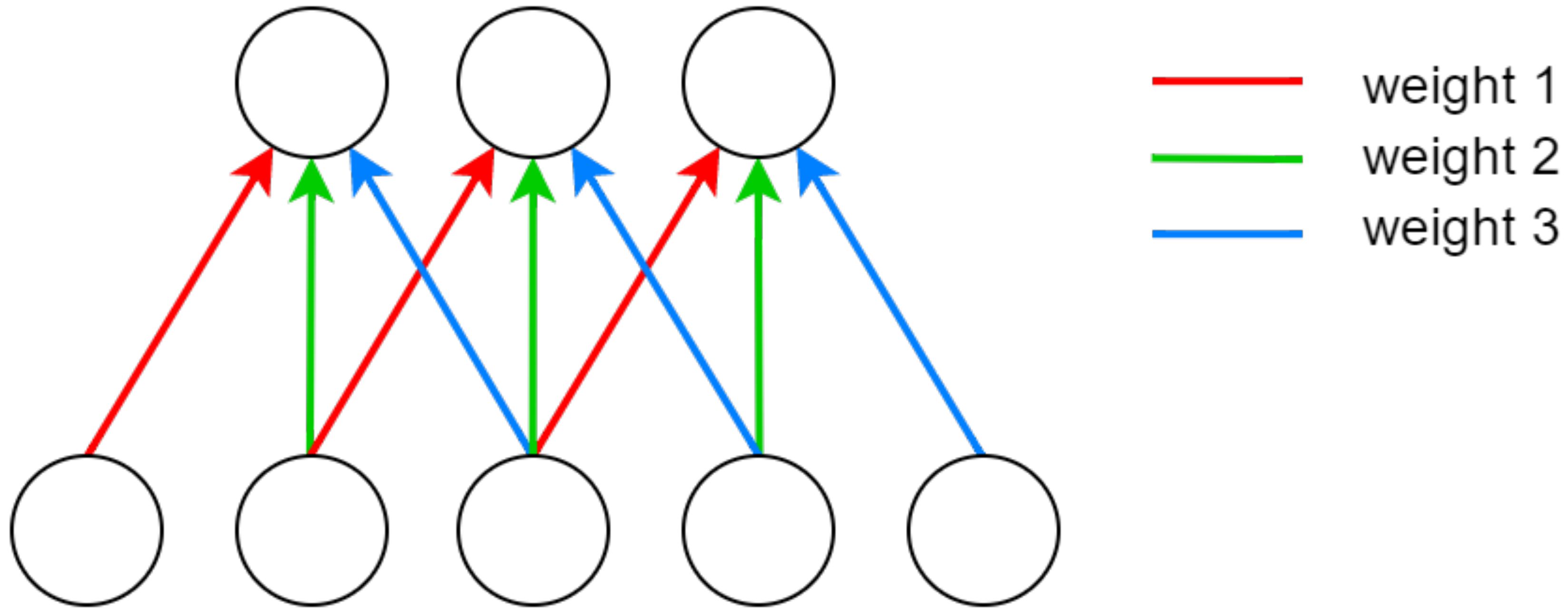


https://devblogs.nvidia.com

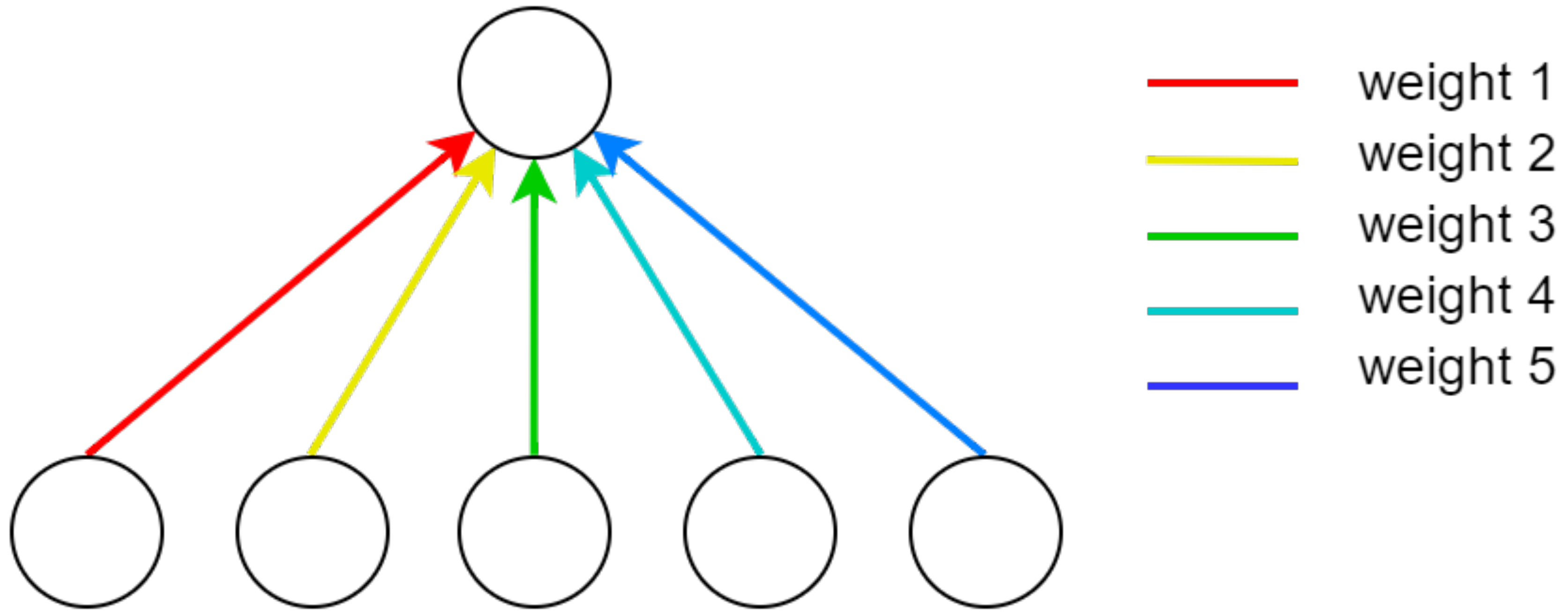
Convolutional Neural Networks



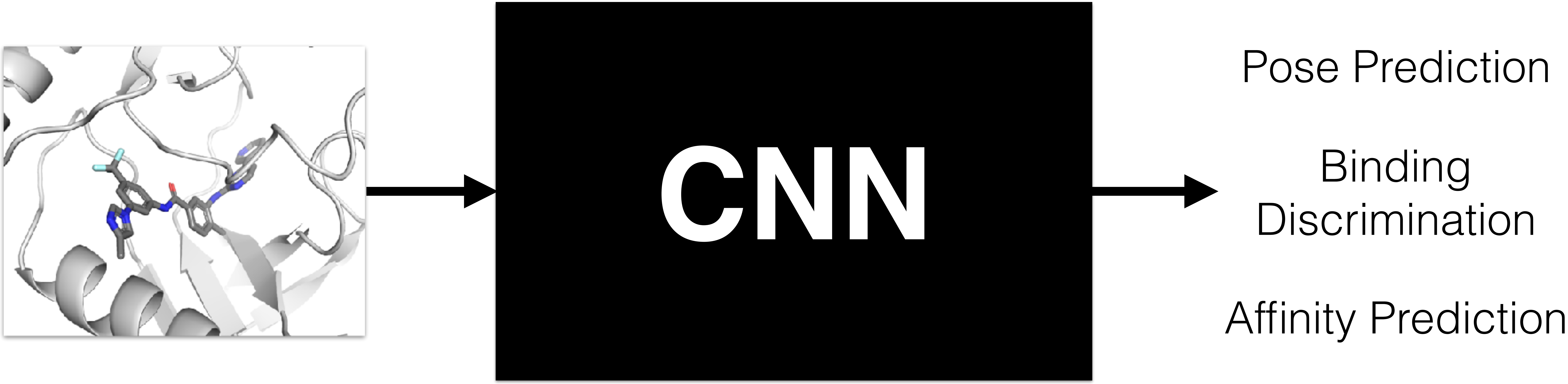
Convolution



Fully-connected



CNNs for Protein-Ligand Scoring



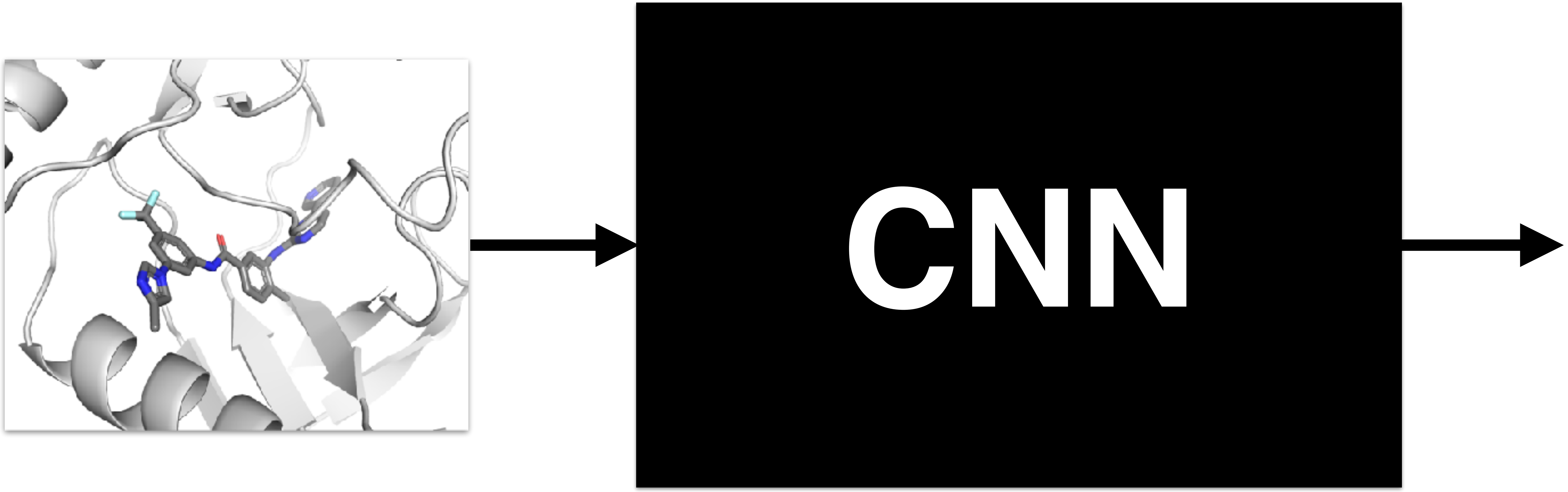
CNN

Pose Prediction

Binding Discrimination

Affinity Prediction

CNNs for Protein-Ligand Scoring



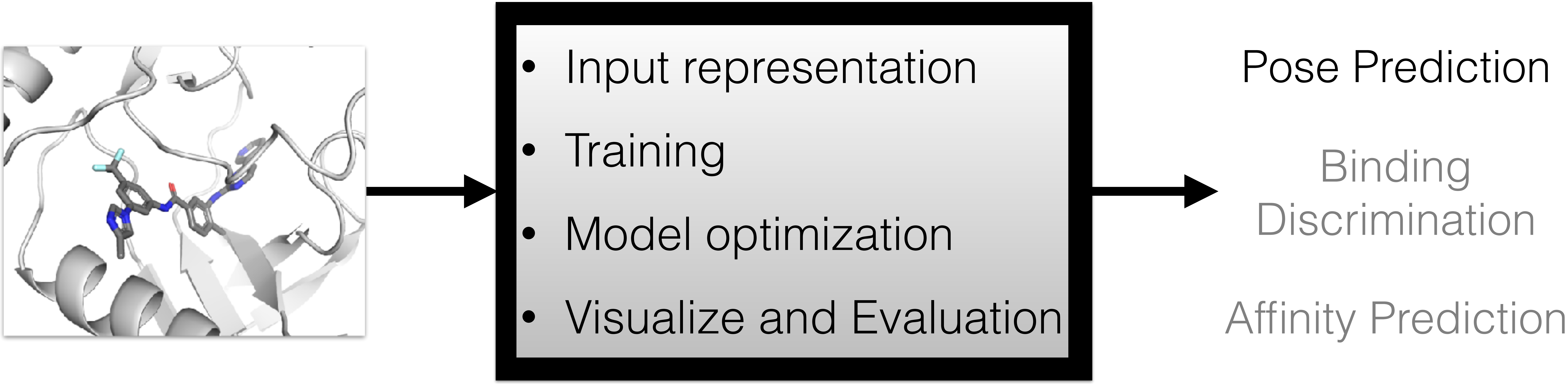
CNN

Pose Prediction

Binding Discrimination

Affinity Prediction

CNNs for Protein-Ligand Scoring



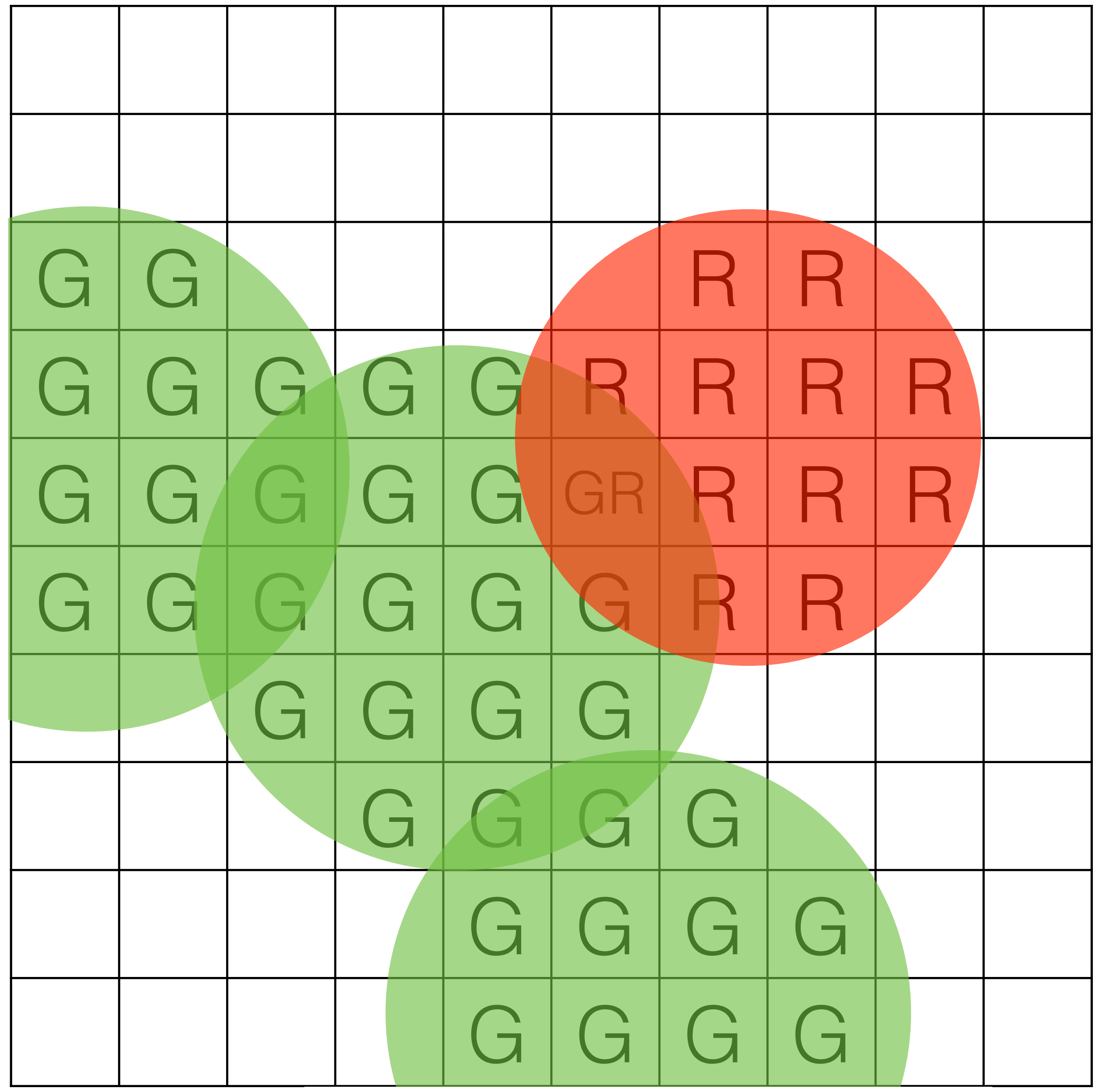
- Input representation
- Training
- Model optimization
- Visualize and Evaluation

Pose Prediction

Binding Discrimination

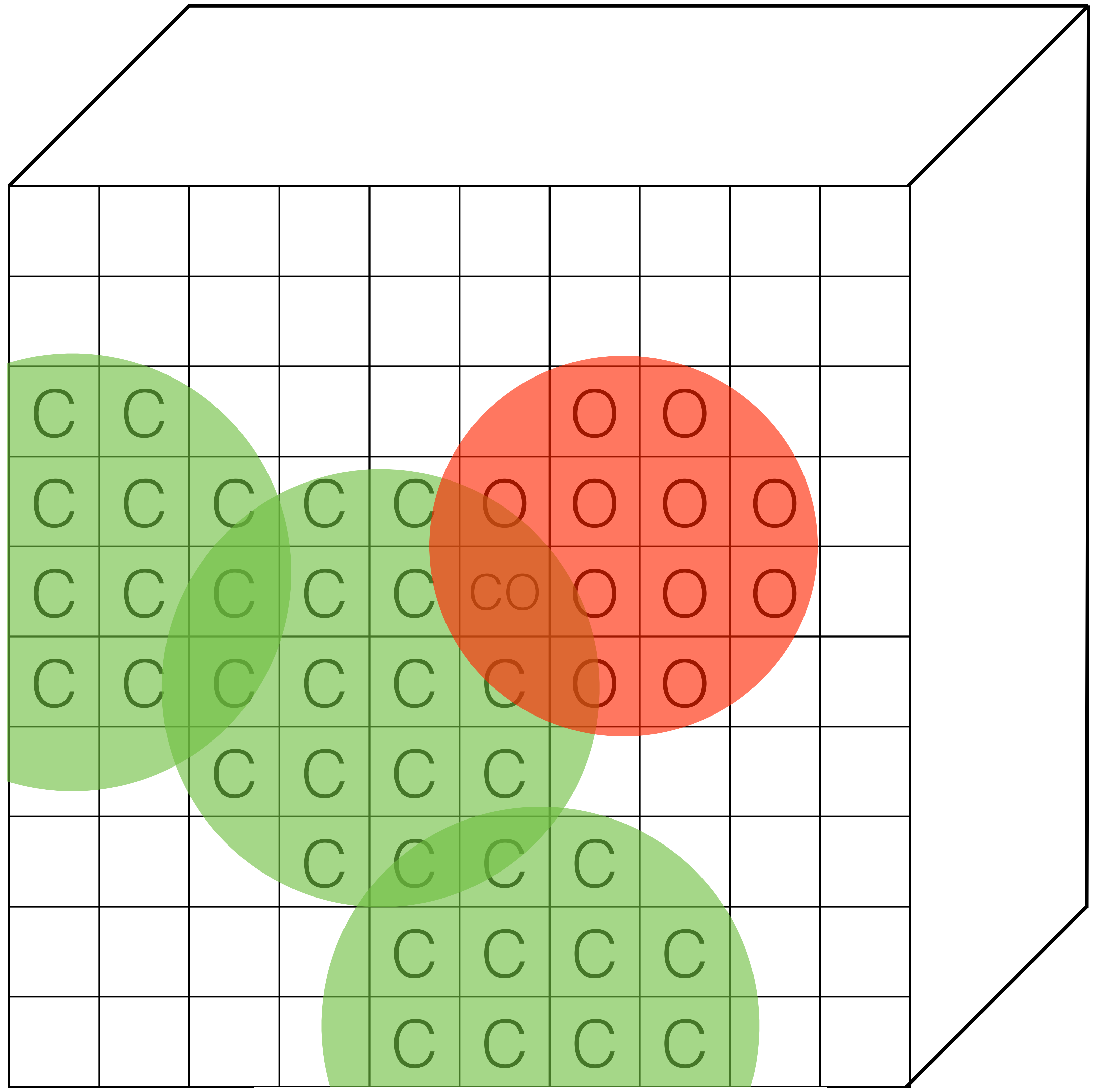
Affinity Prediction

Protein-Ligand Representation



(R,G,B) pixel

Protein-Ligand Representation



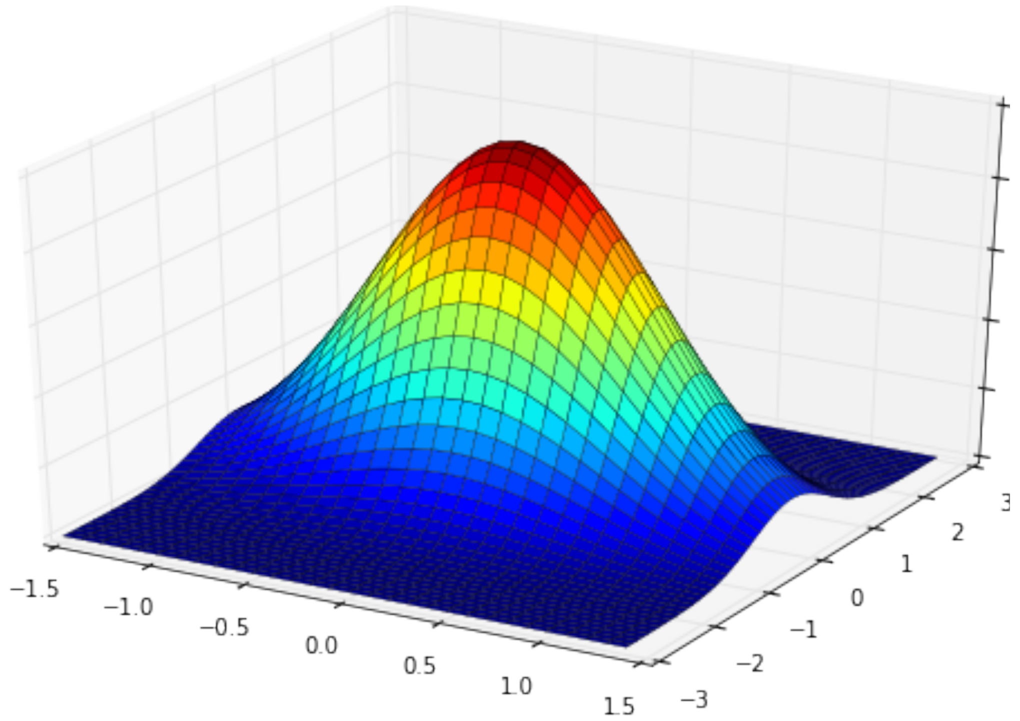
(R,G,B) pixel →

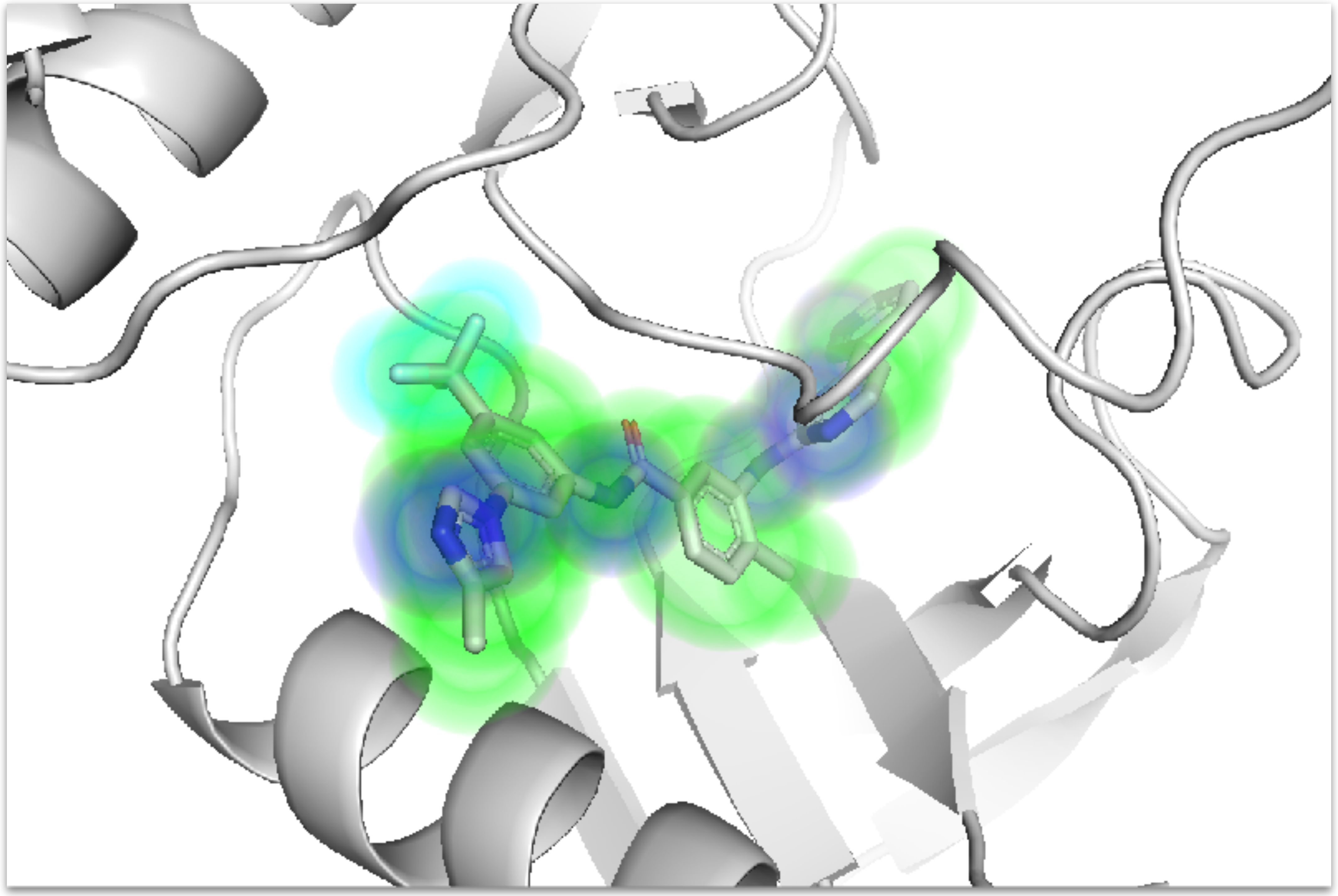
(Carbon, Nitrogen, Oxygen,…) **voxel**

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

Atom Density

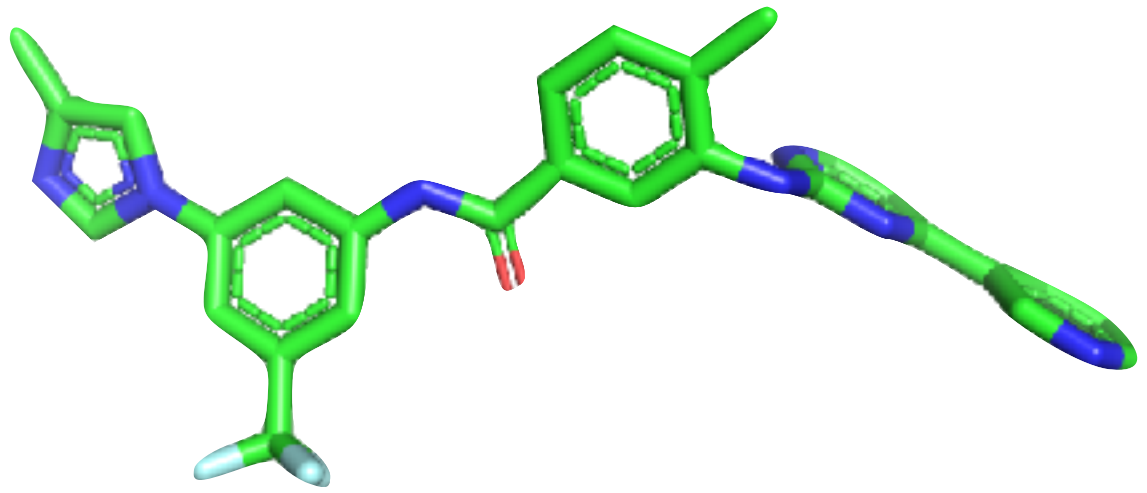
$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}} & 0 \leq d < r \\ \frac{4}{e^2 r^2} d^2 - \frac{12}{e^2 r} d + \frac{9}{e^2} & r \leq d < 1.5r \\ 0 & d \geq 1.5r \end{cases}$$





Gaussian

Atom Types

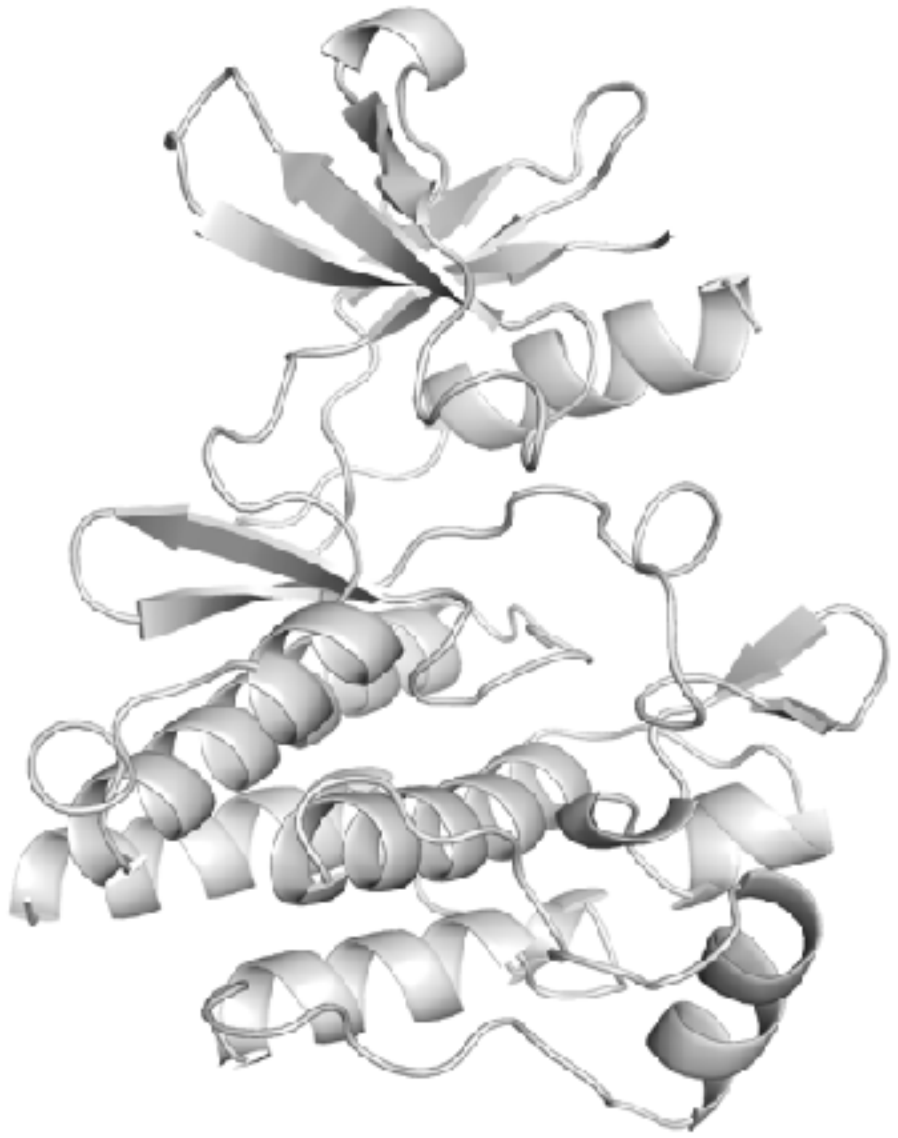


Ligand

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

CSAR

337 protein-ligand complexes

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

PDBbind

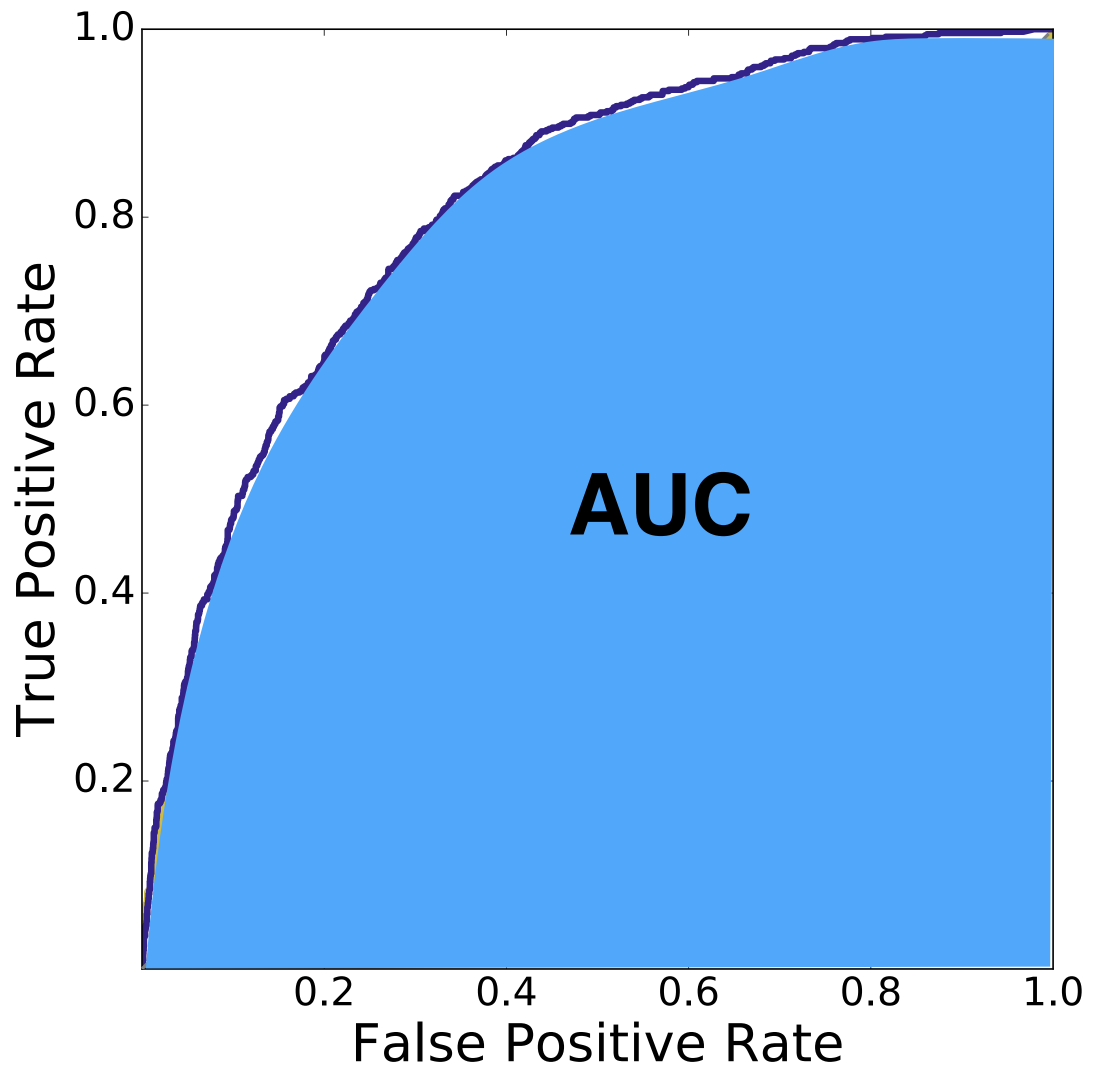
12,484 protein-ligand complexes

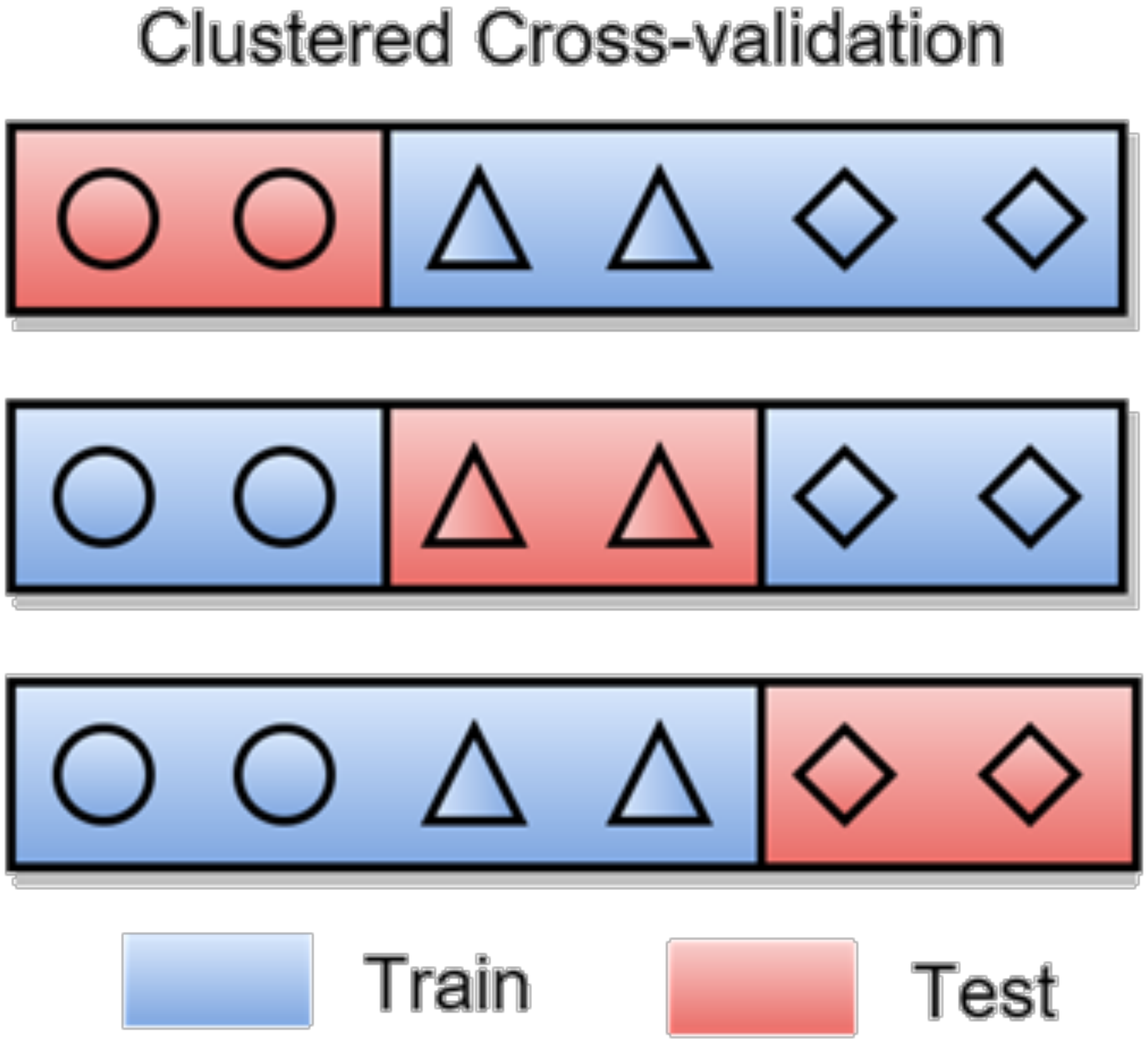
- diverse targets
- wide range of affinities
- **generate poses** with AutoDock Vina
- include minimized crystal pose
  - 24,727 <2Å RMSD (actives)
  - 244,192 >4Å RMSD (decoys)

Model Evaluation

CSAR: >90% similar targets kept in same fold

PDBbind: >80% similar targets kept in same fold

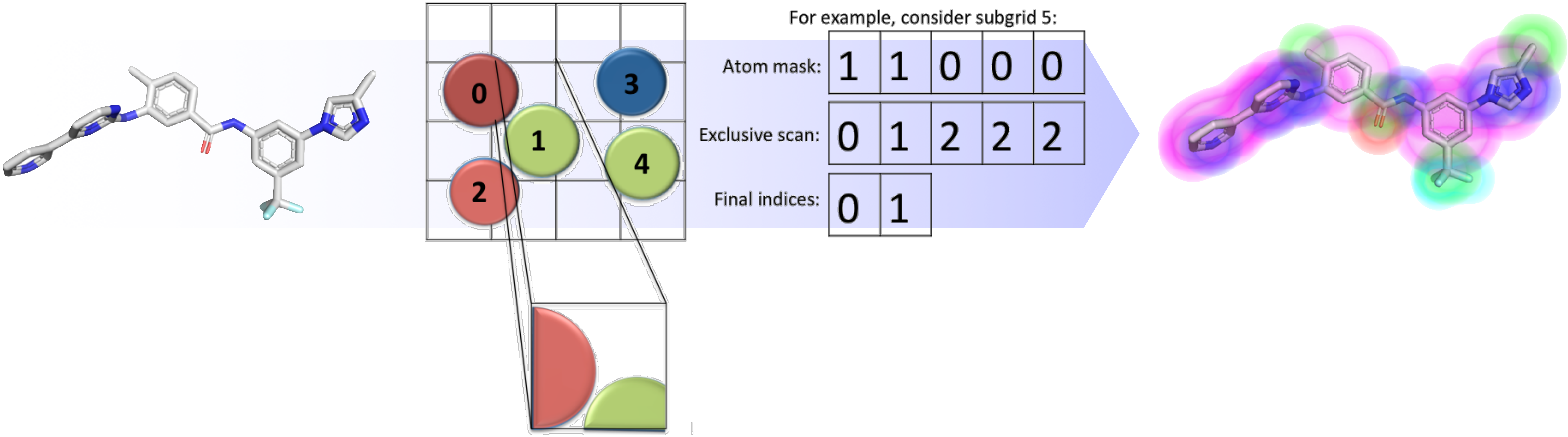




Model Training

Custom `MolGridDataLayer`

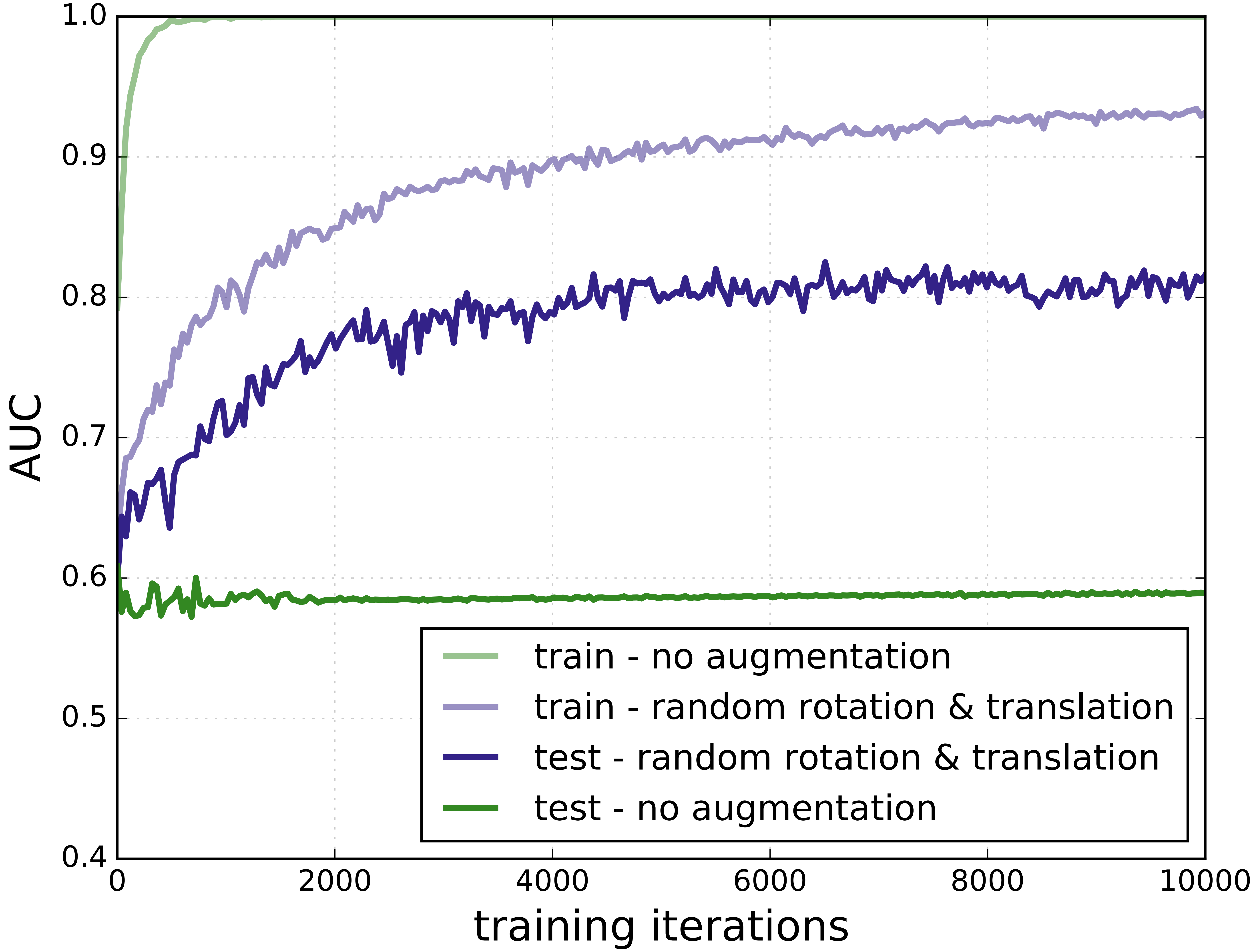
Parallelize over *atoms* to obtain a mask of atoms that overlap each grid region
Use exclusive scan to obtain a list of atom indices from the mask
Parallelize over *grid points*, using reduced atom list to avoid O(N_{atoms}) check

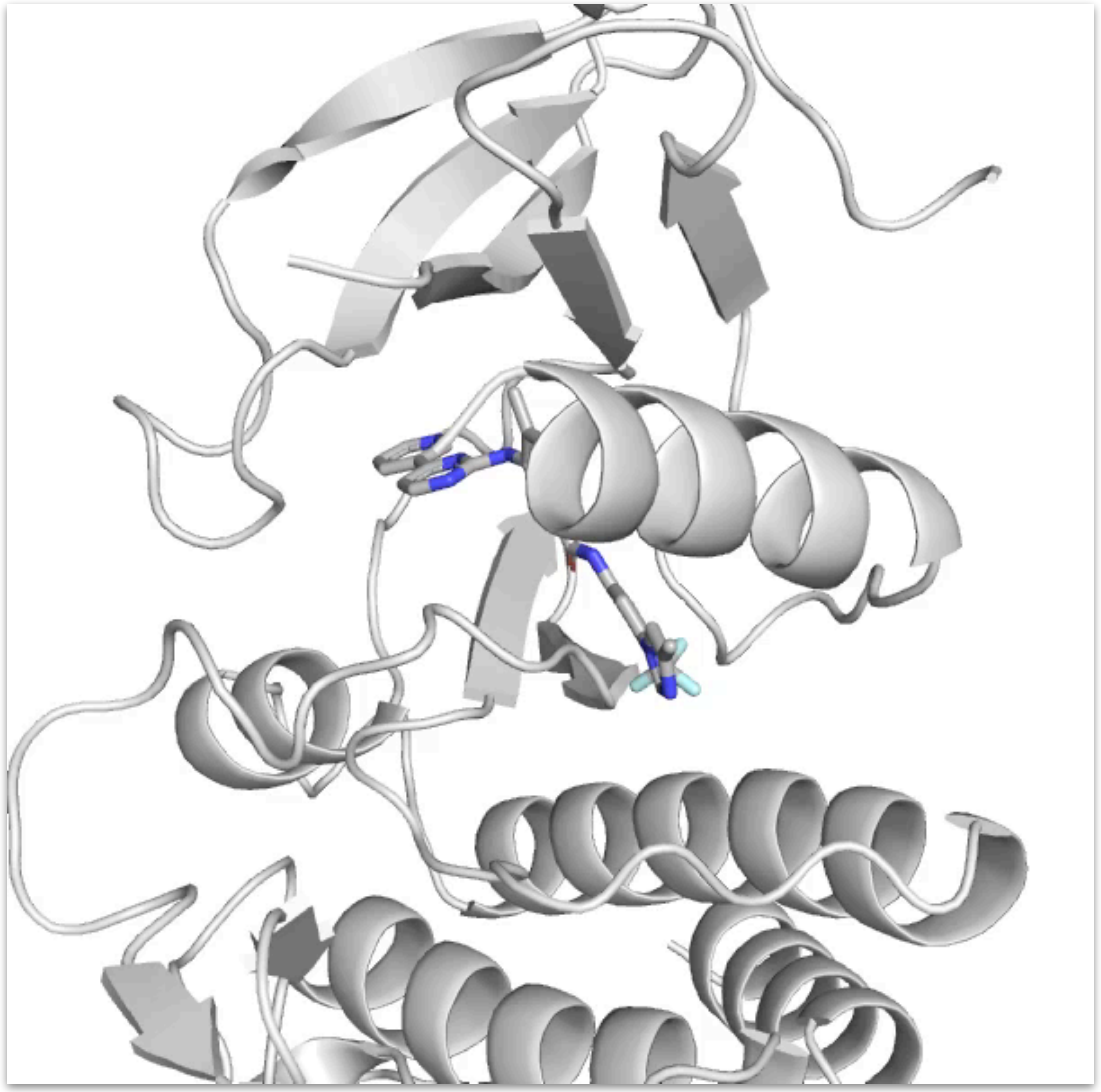


For example, consider subgrid 5:

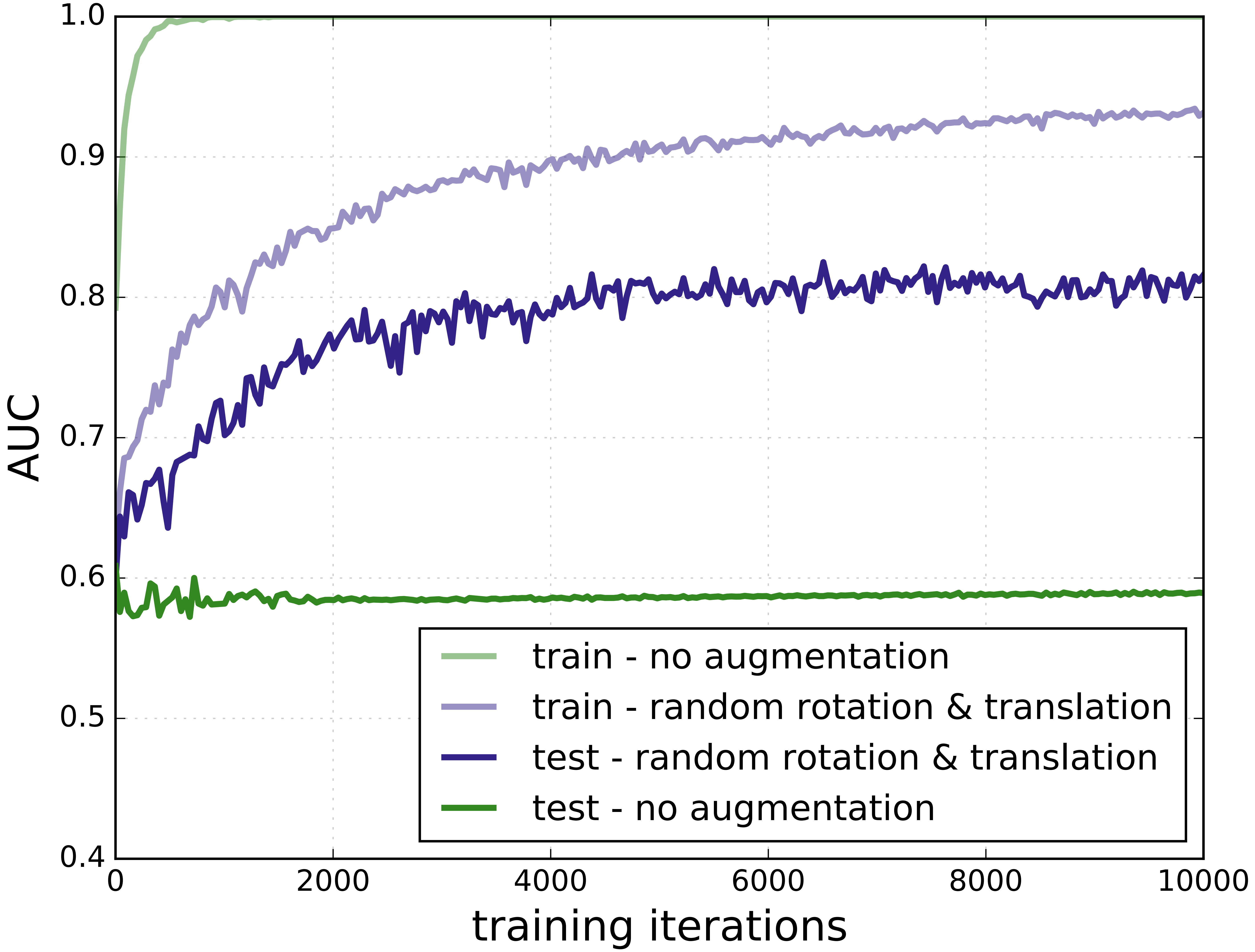
Atom mask:	1	1	0	0	0
Exclusive scan:	0	1	2	2	2
Final indices:	0	1			

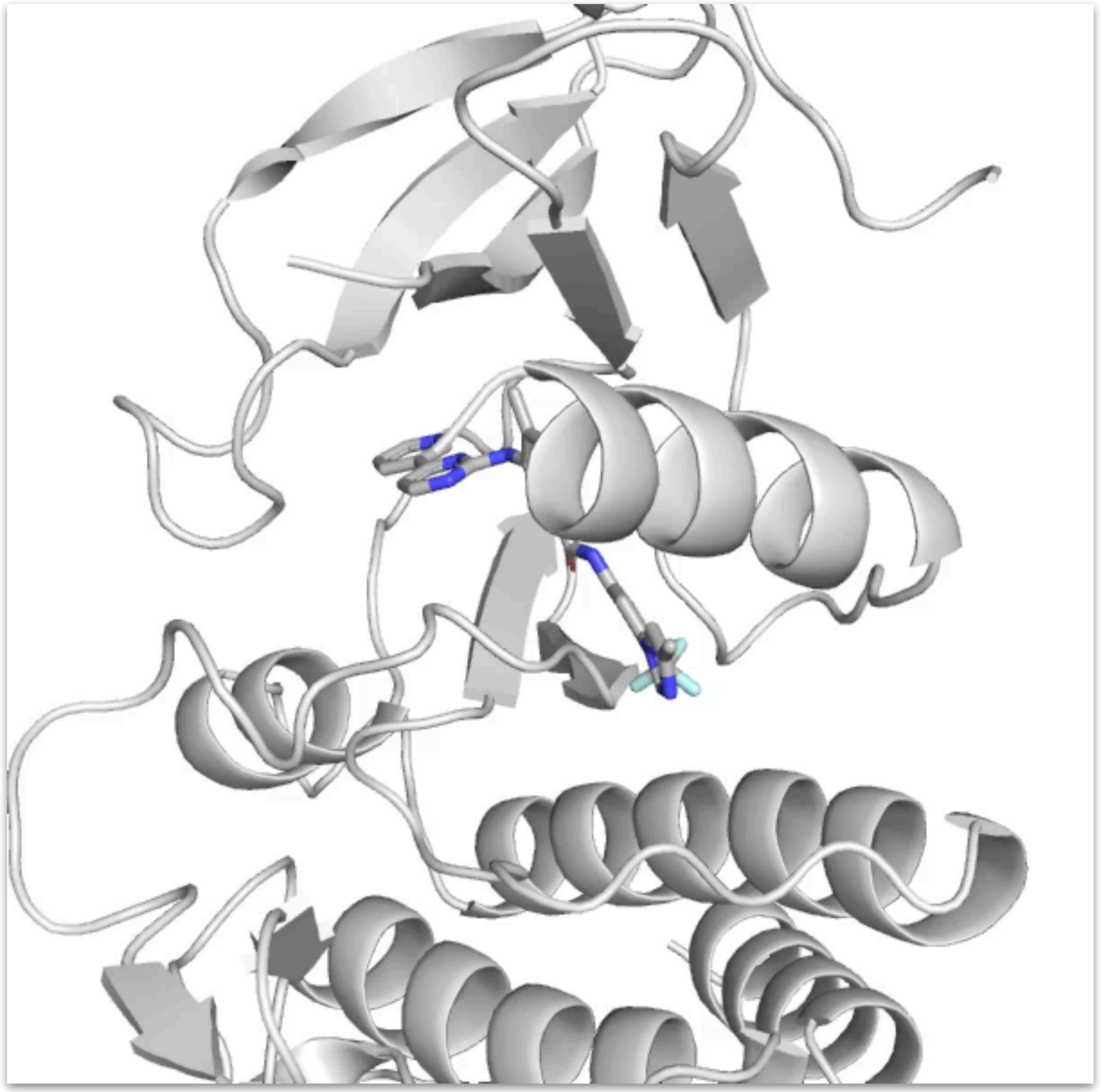
Data Augmentation





Data Augmentation





Model Optimization

Atom Types

- Vina (34)
- element-only (18)
- ligand-protein (2)

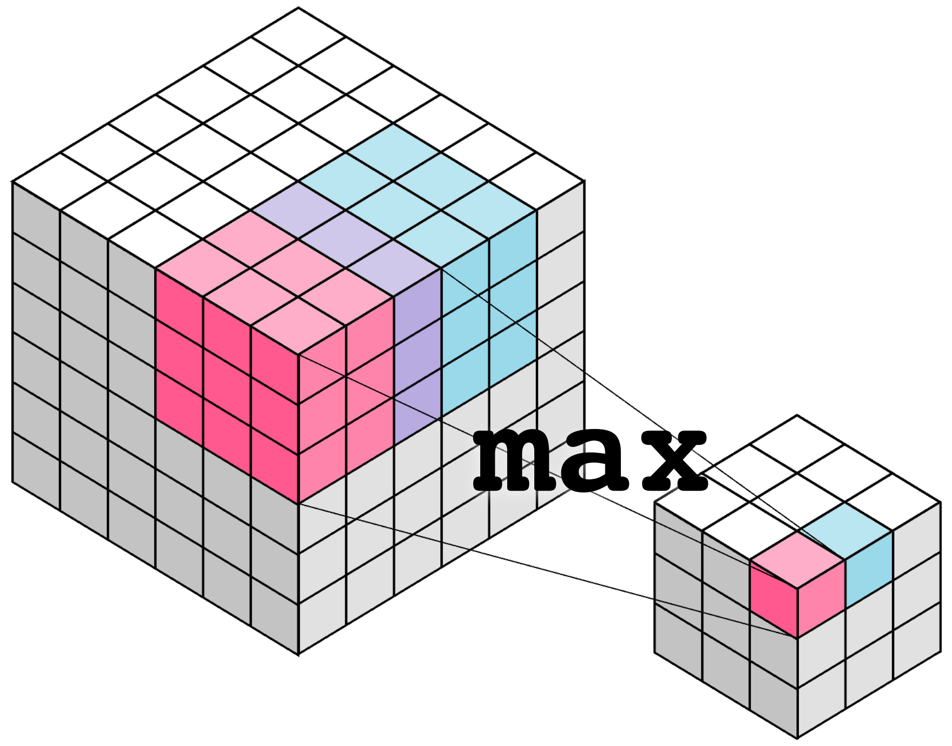
Atom Density Type

- Boolean
- Gaussian

Radius Multiple

Resolution

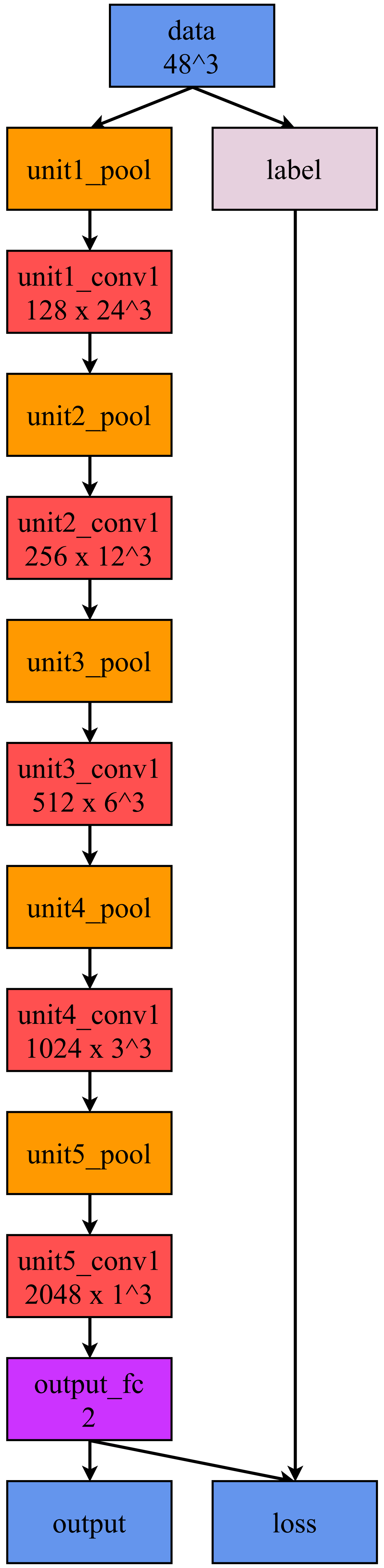
Pooling



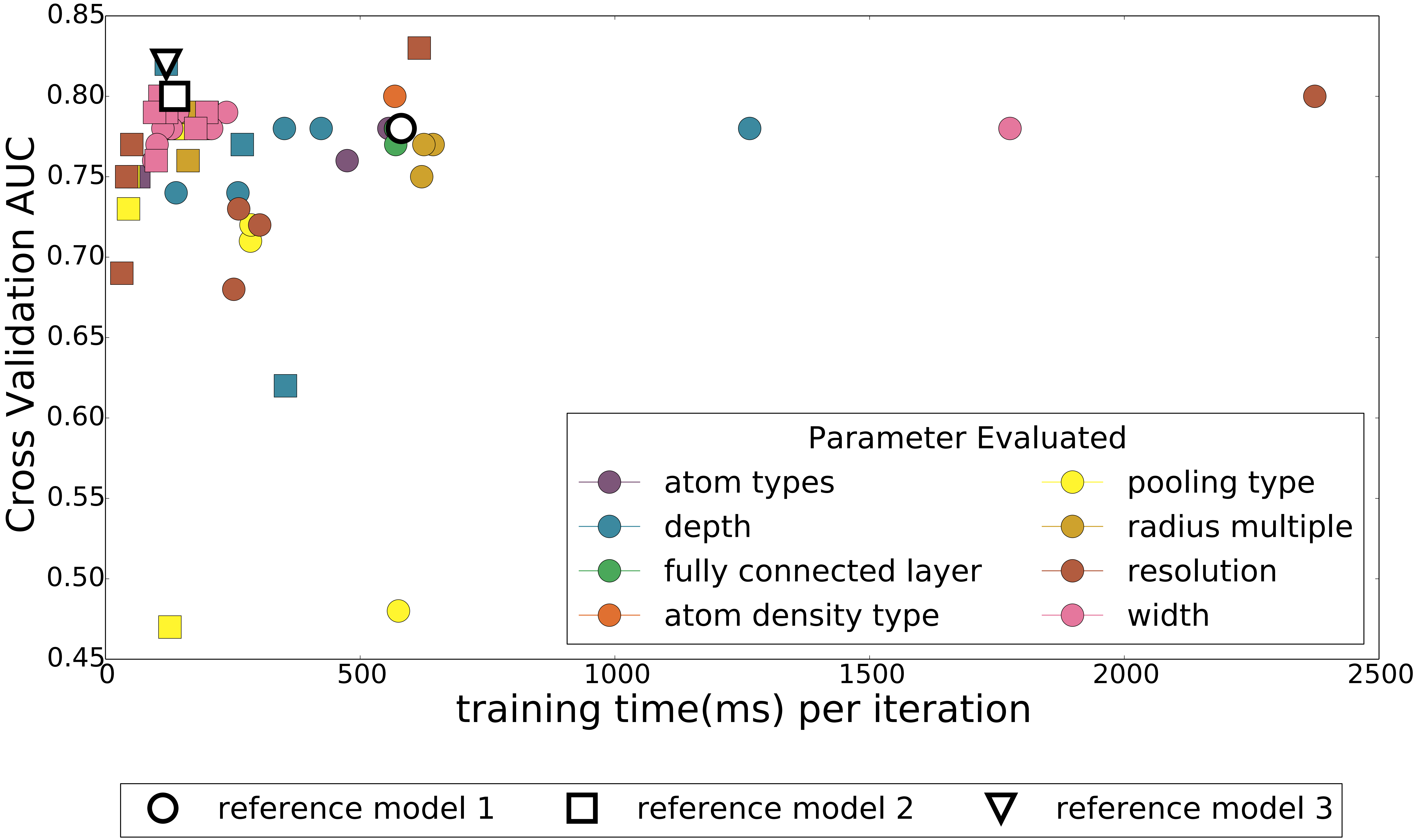
Depth

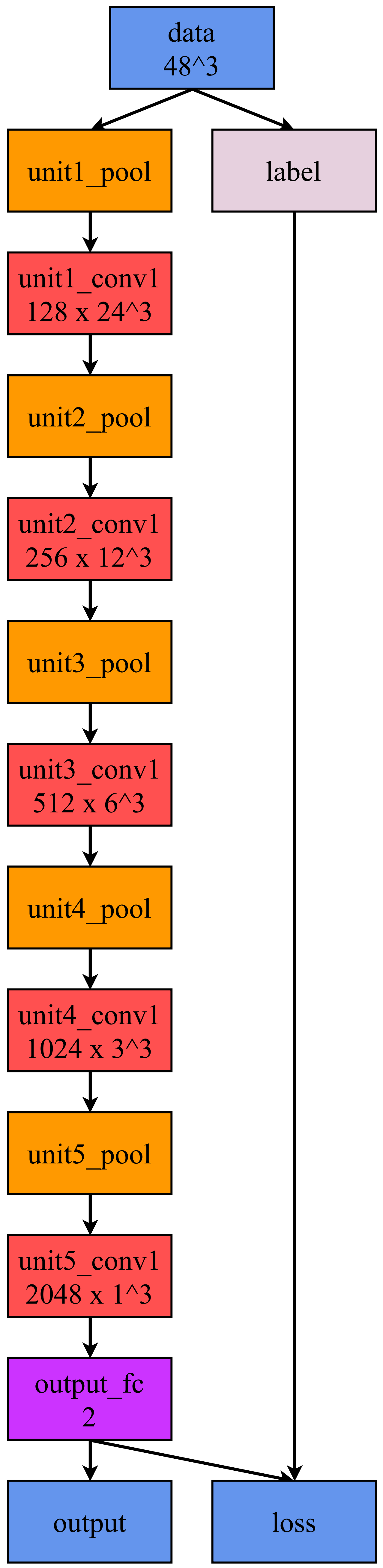
Width

Fully Connected Layers



Model Optimization





data
48^3

unit1_pool

label

unit1_conv1
128 x 24^3

unit2_pool

unit2_conv1
256 x 12^3

unit3_pool

unit3_conv1
512 x 6^3

unit4_pool

unit4_conv1
1024 x 3^3

unit5_pool

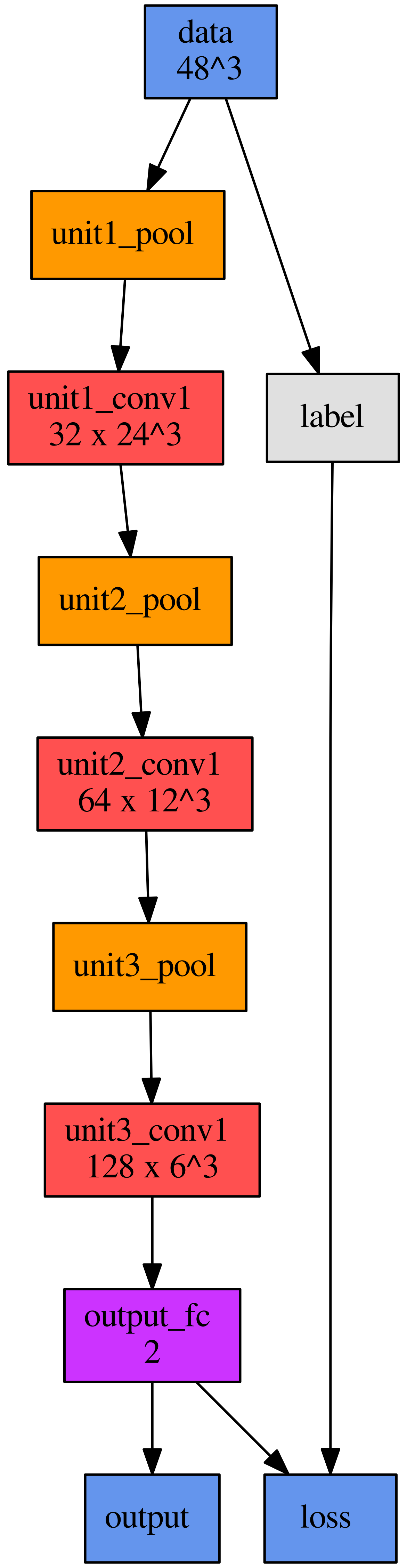
unit5_conv1
2048 x 1^3

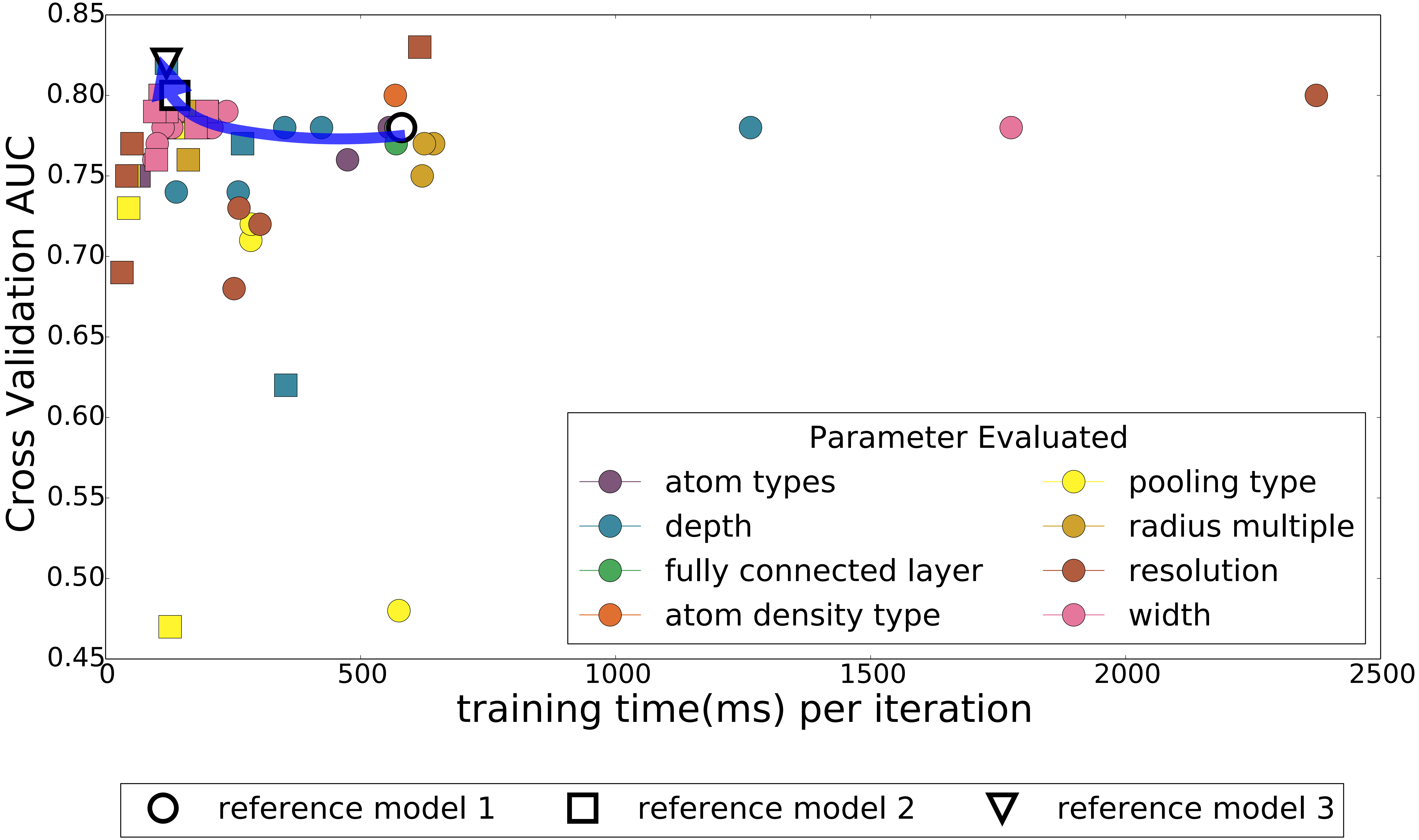
output_fc
2

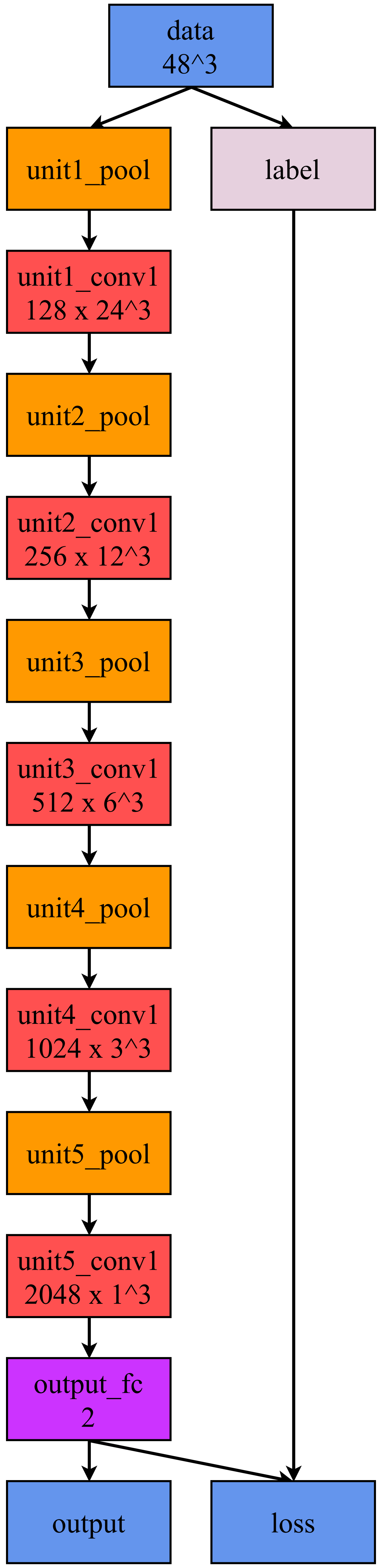
output

loss

Model Optimization

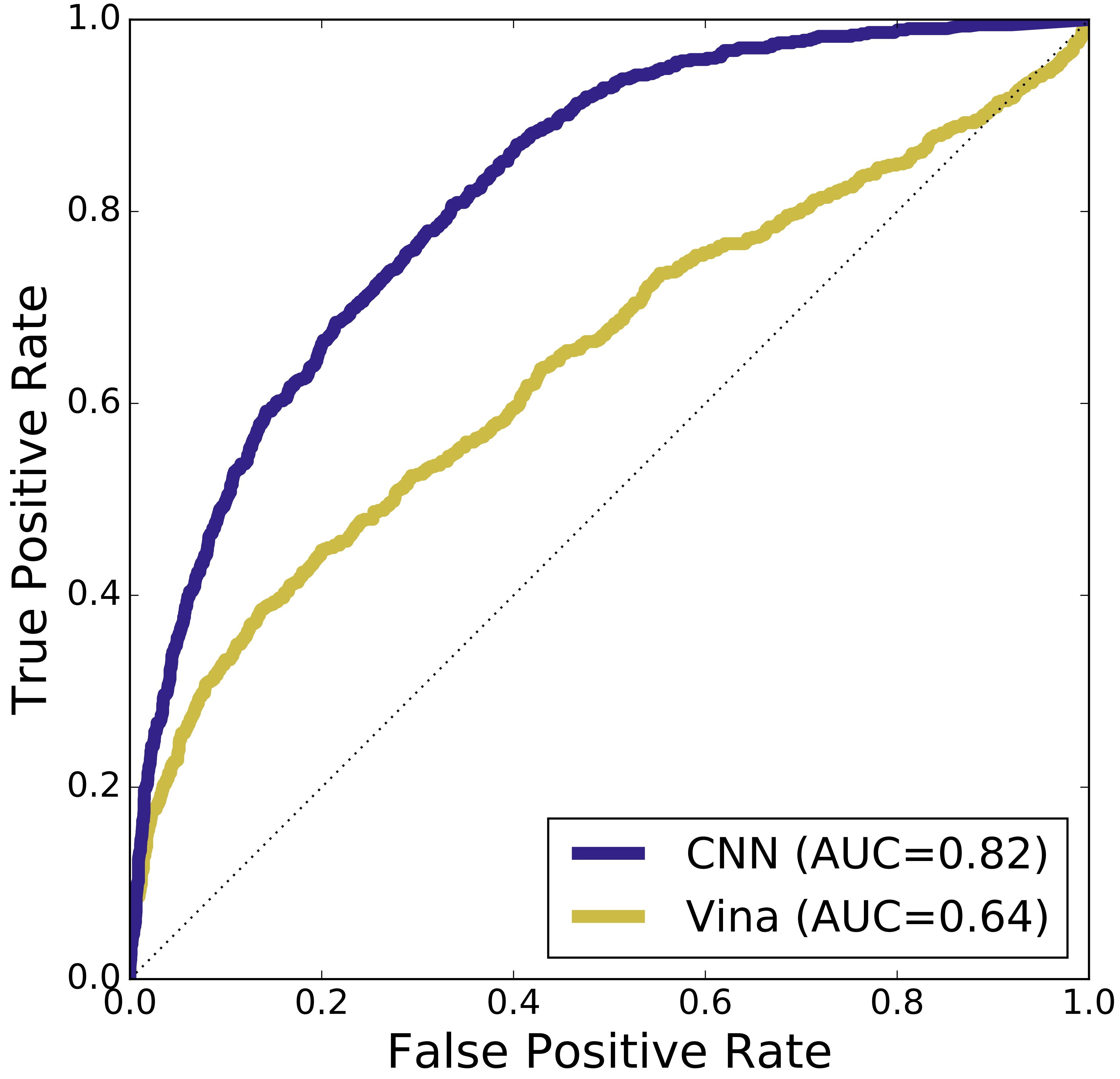




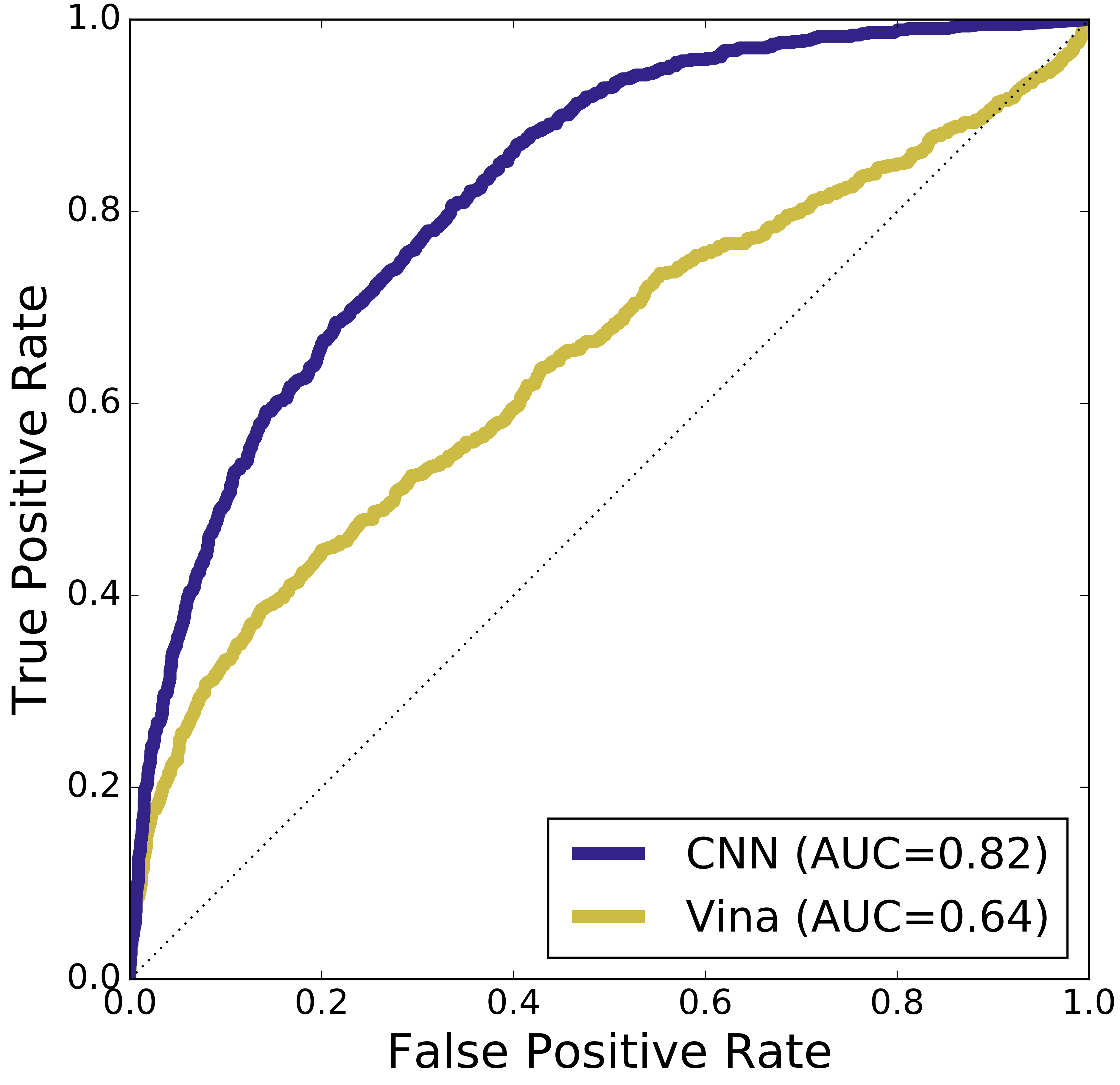


Cross-Validation Evaluation

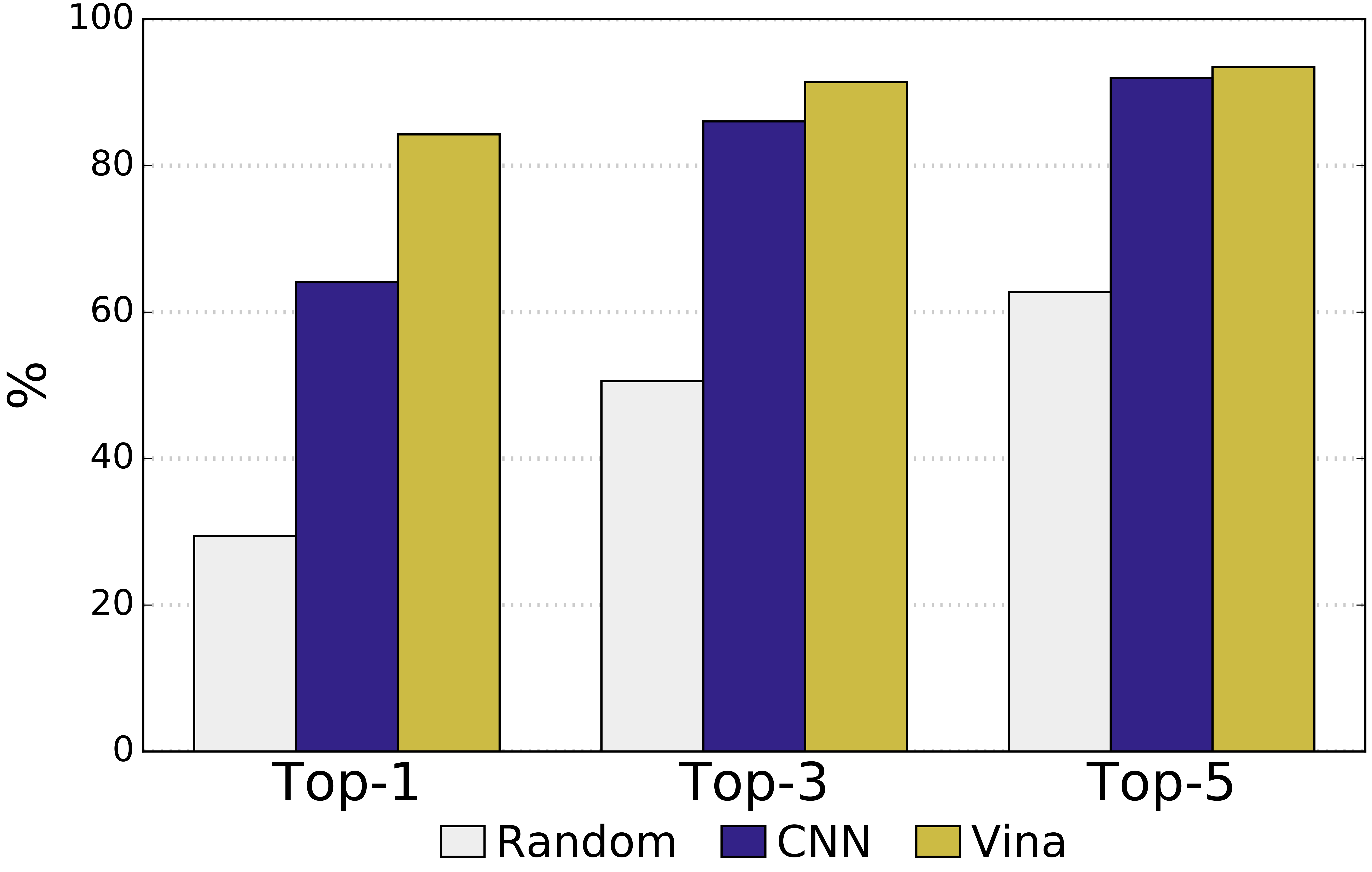
Pose Prediction (CSAR)



Pose Prediction (CSAR)

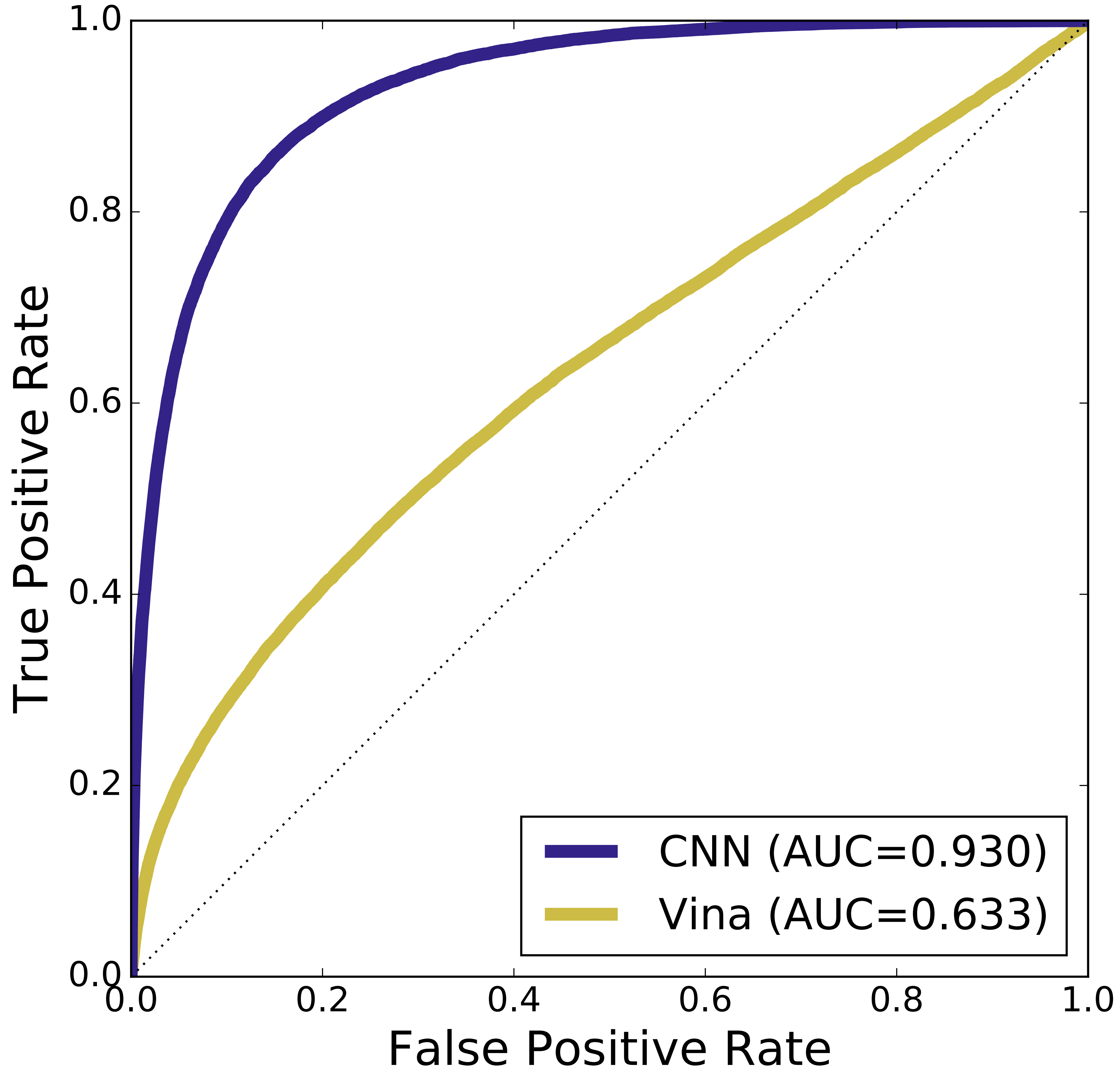


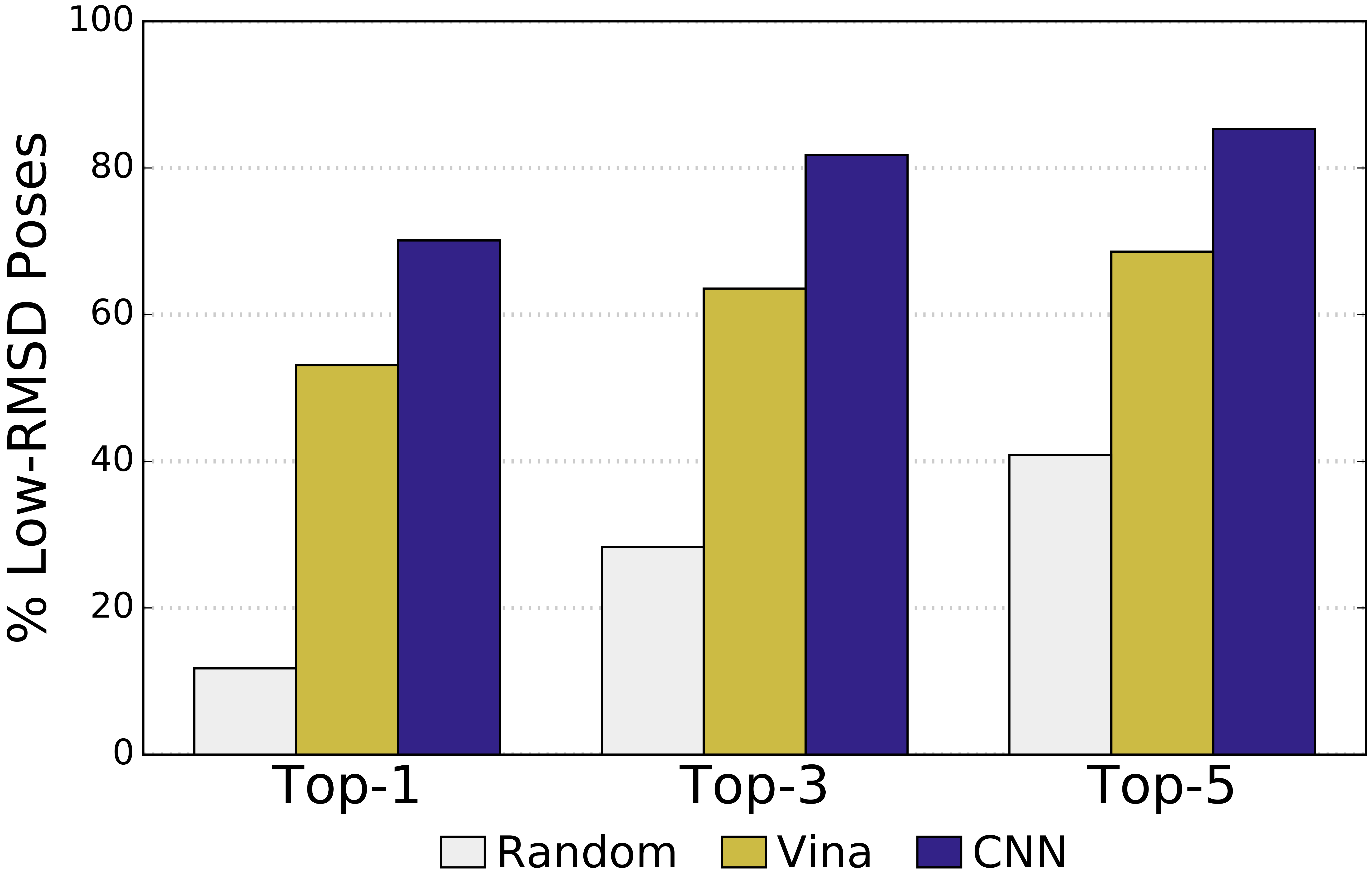
inter-target ranking



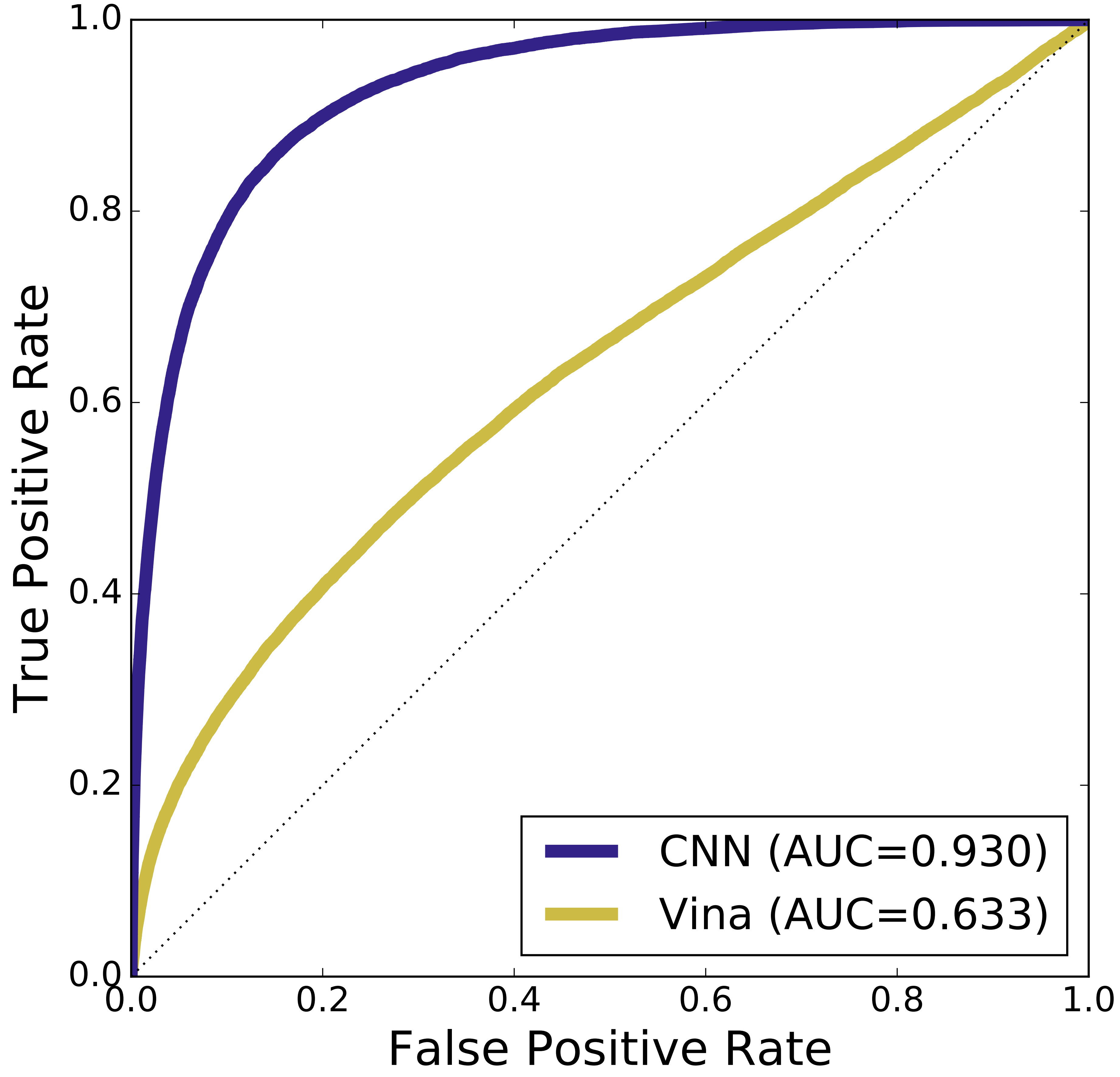
intra-target ranking

Pose Prediction (PDBbind)

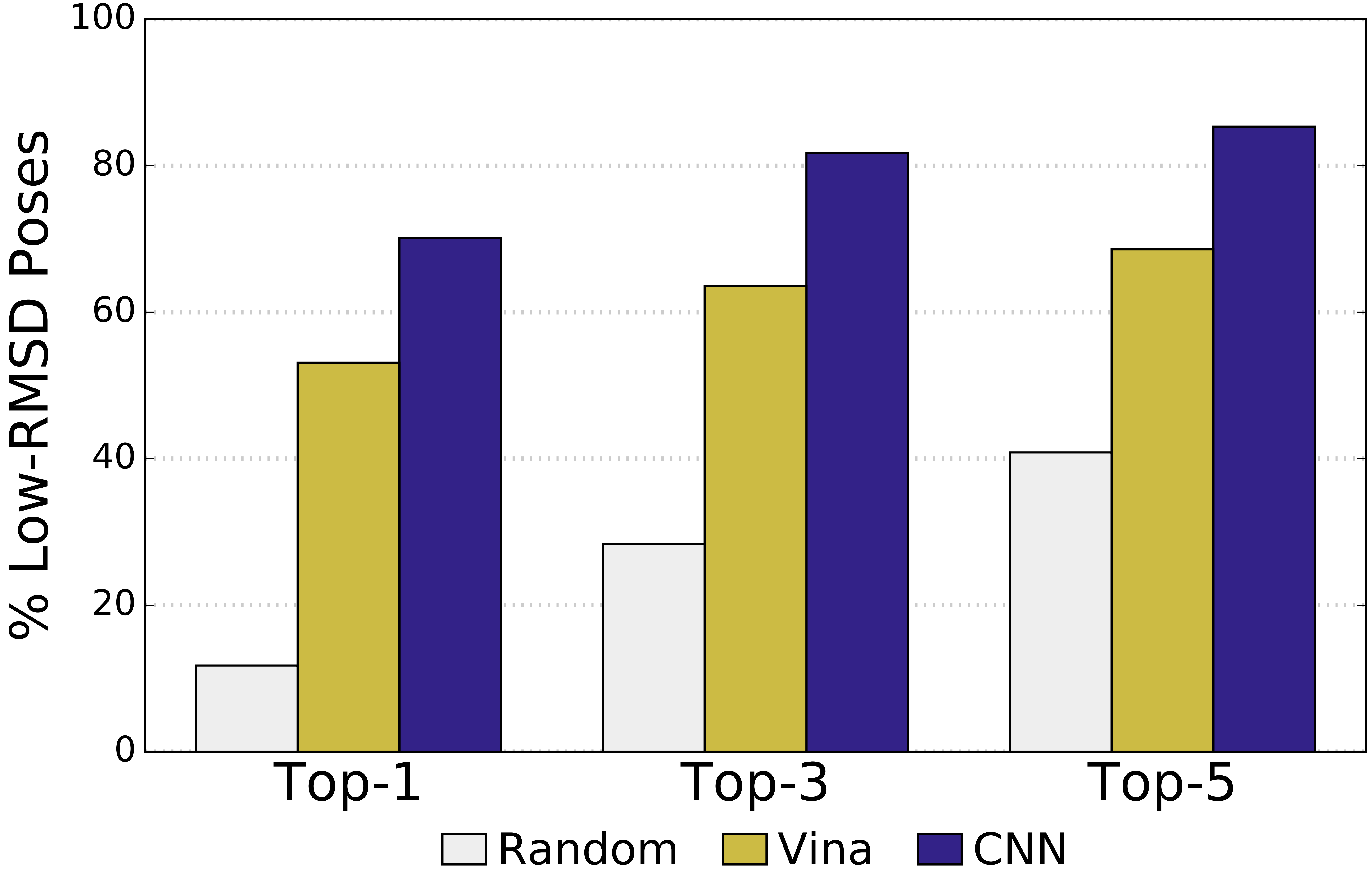




Pose Prediction (PDBbind)

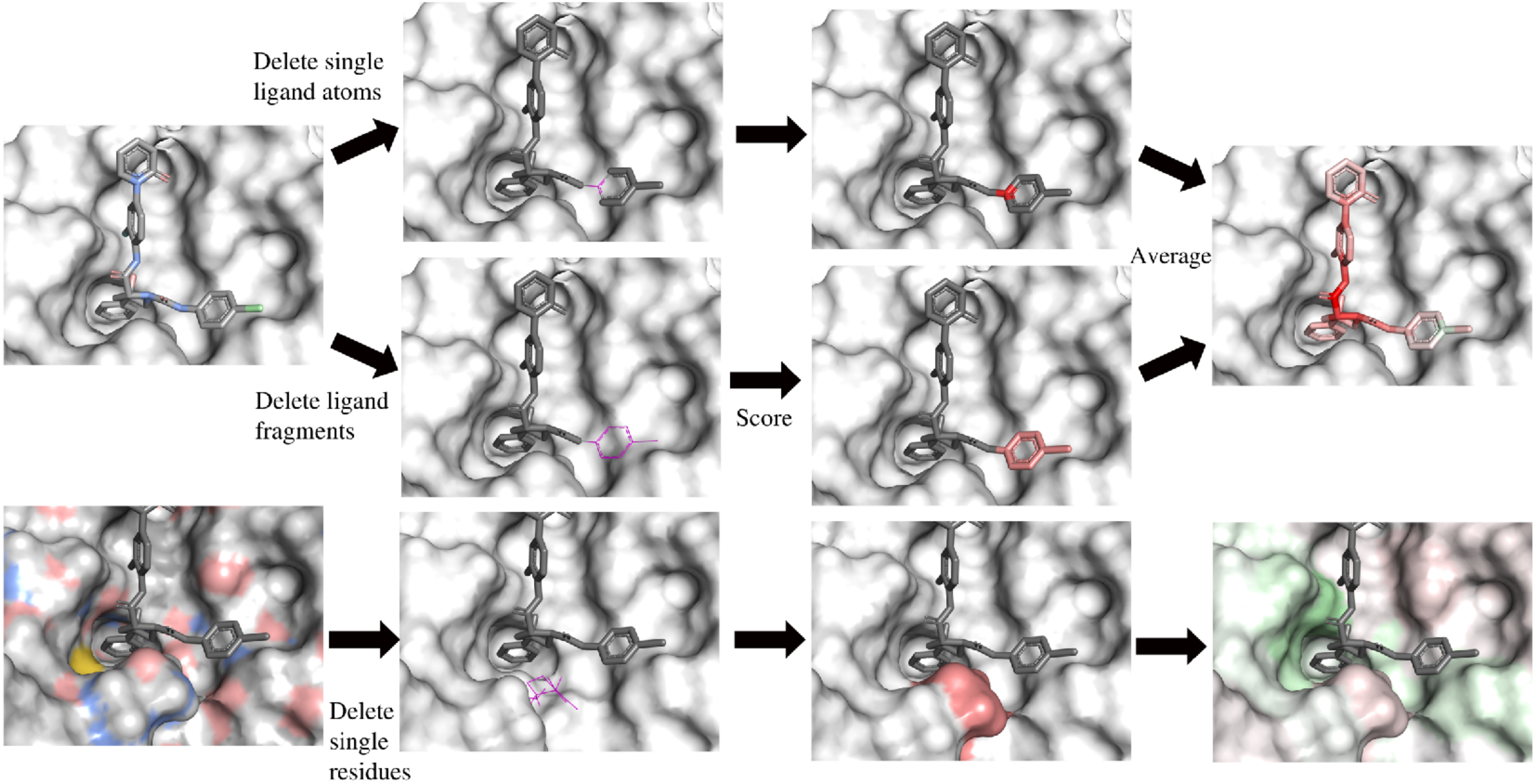


inter-target ranking

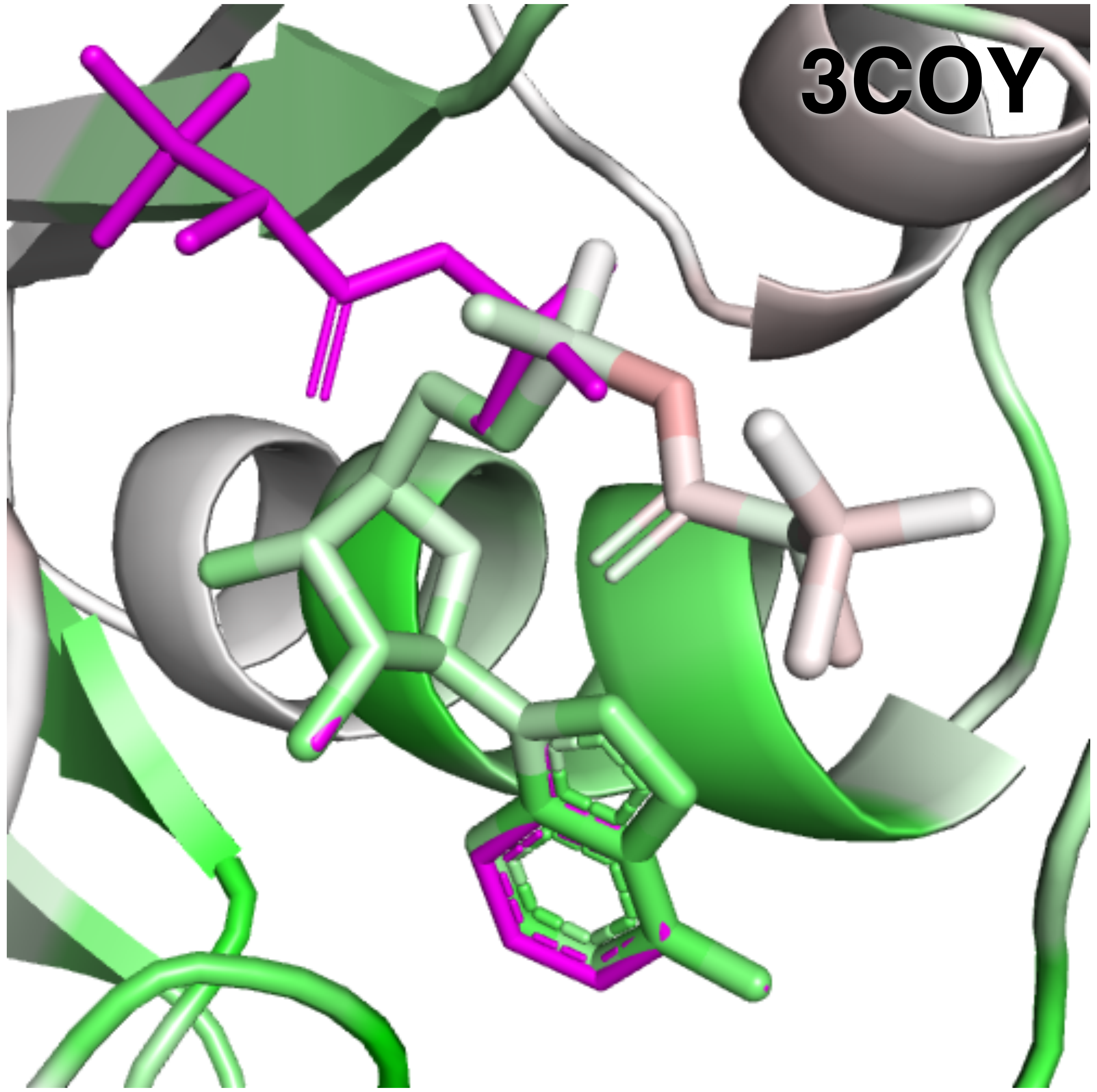


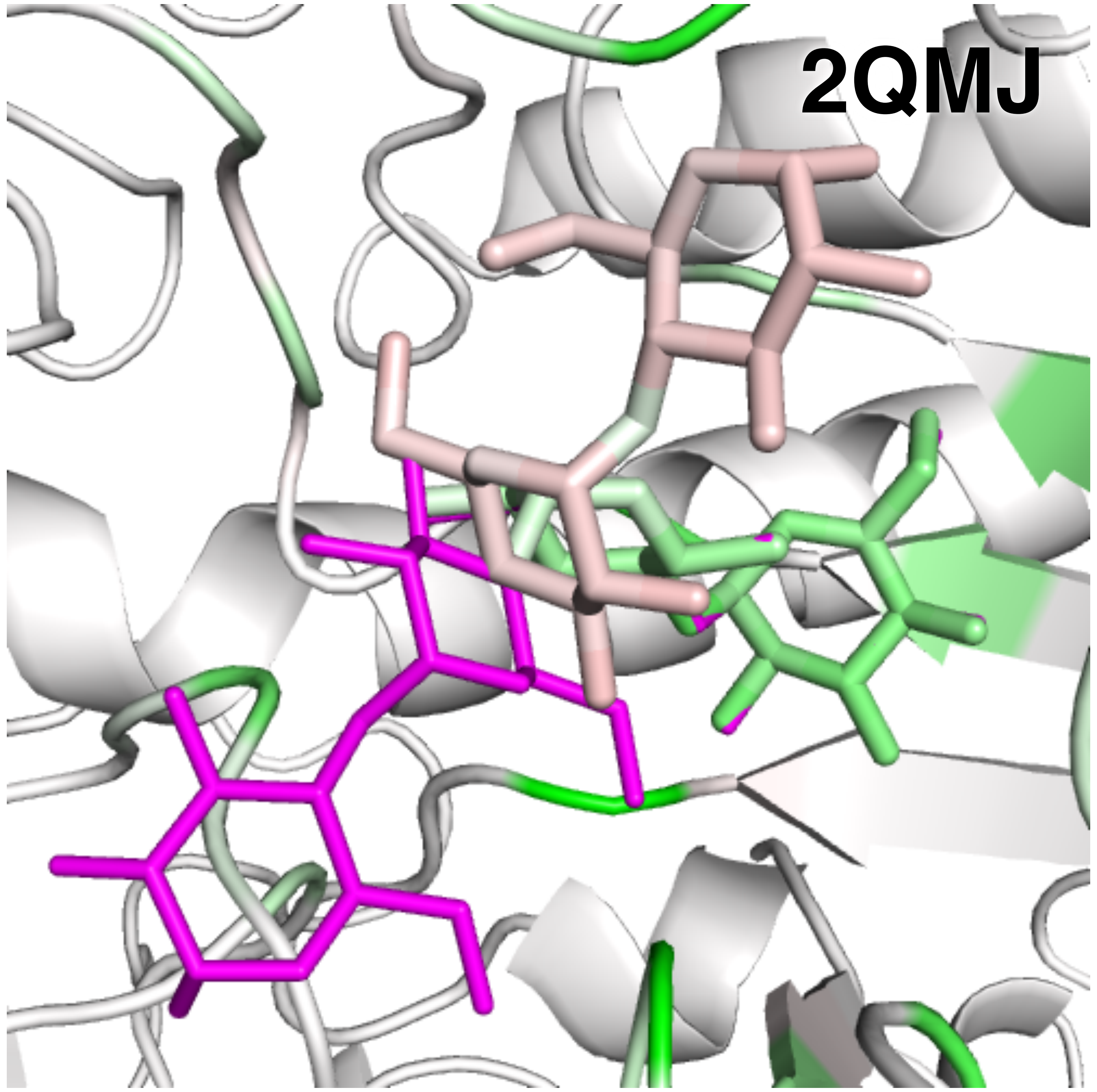
intra-target ranking

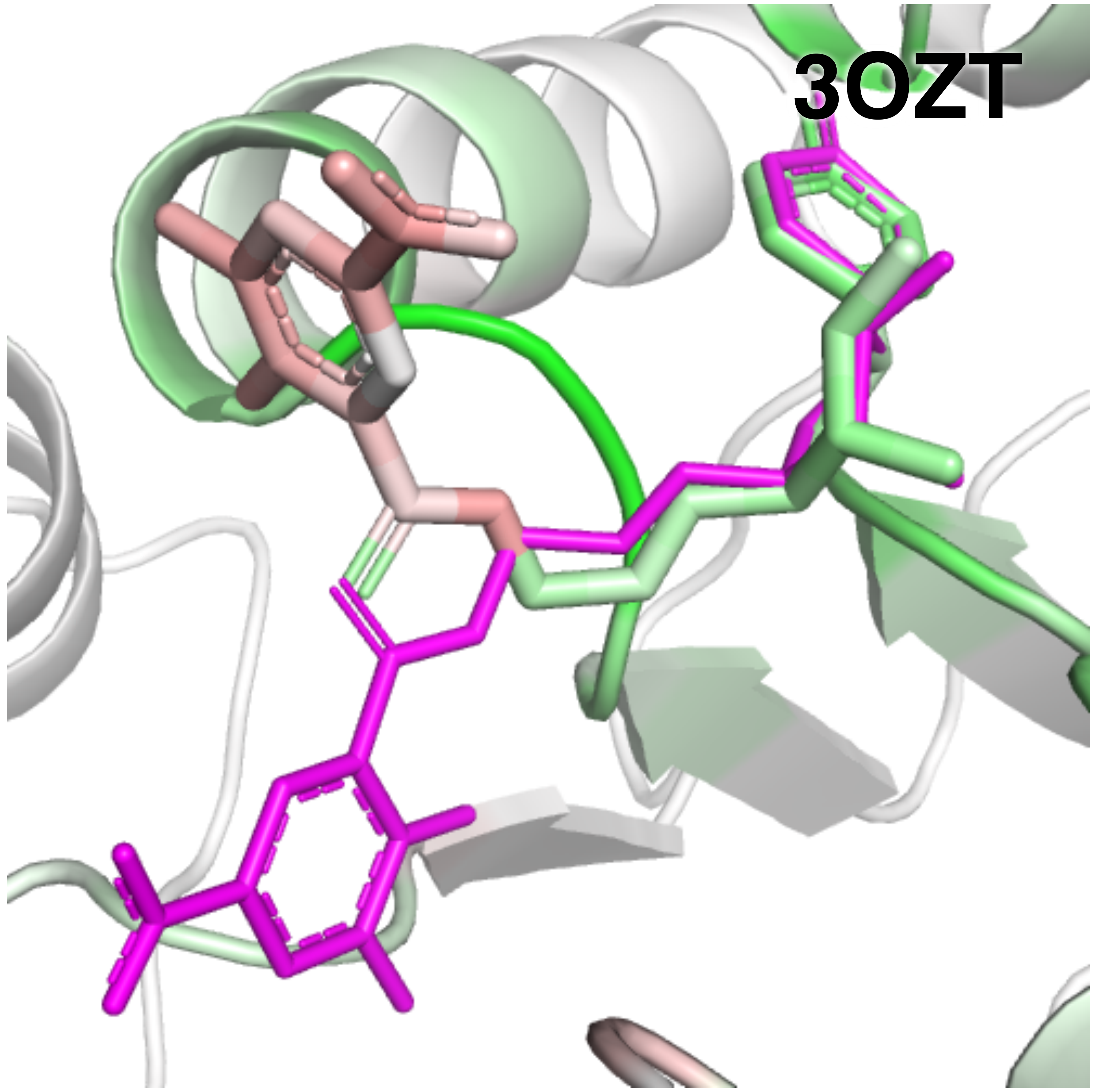
Visualization



Examples

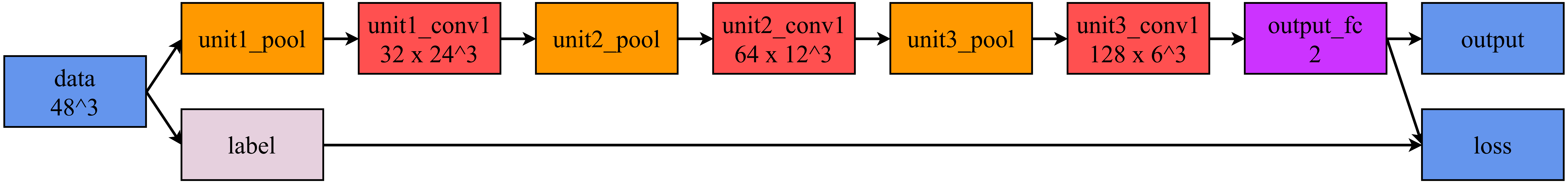




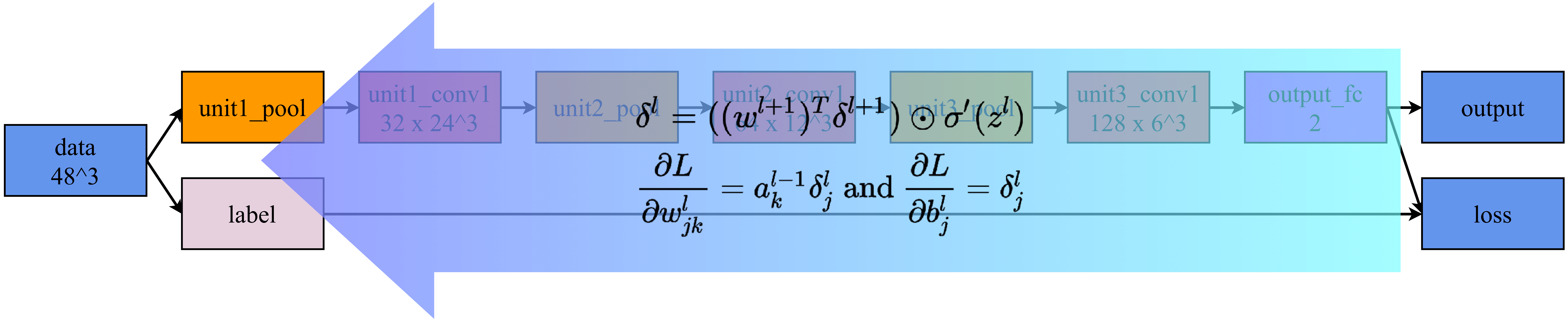


Partially Aligned Poses

Beyond Scoring



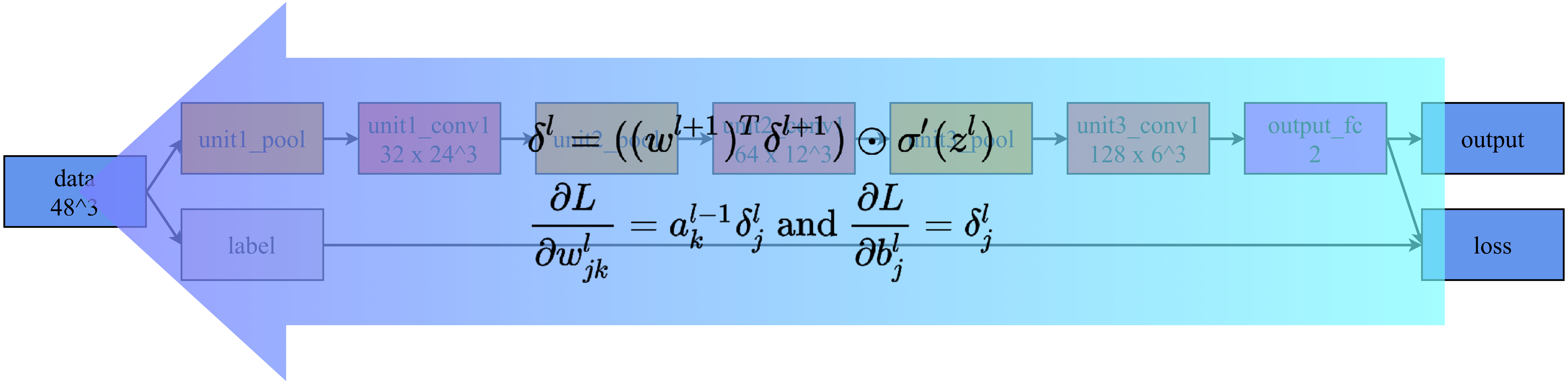
Beyond Scoring



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

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

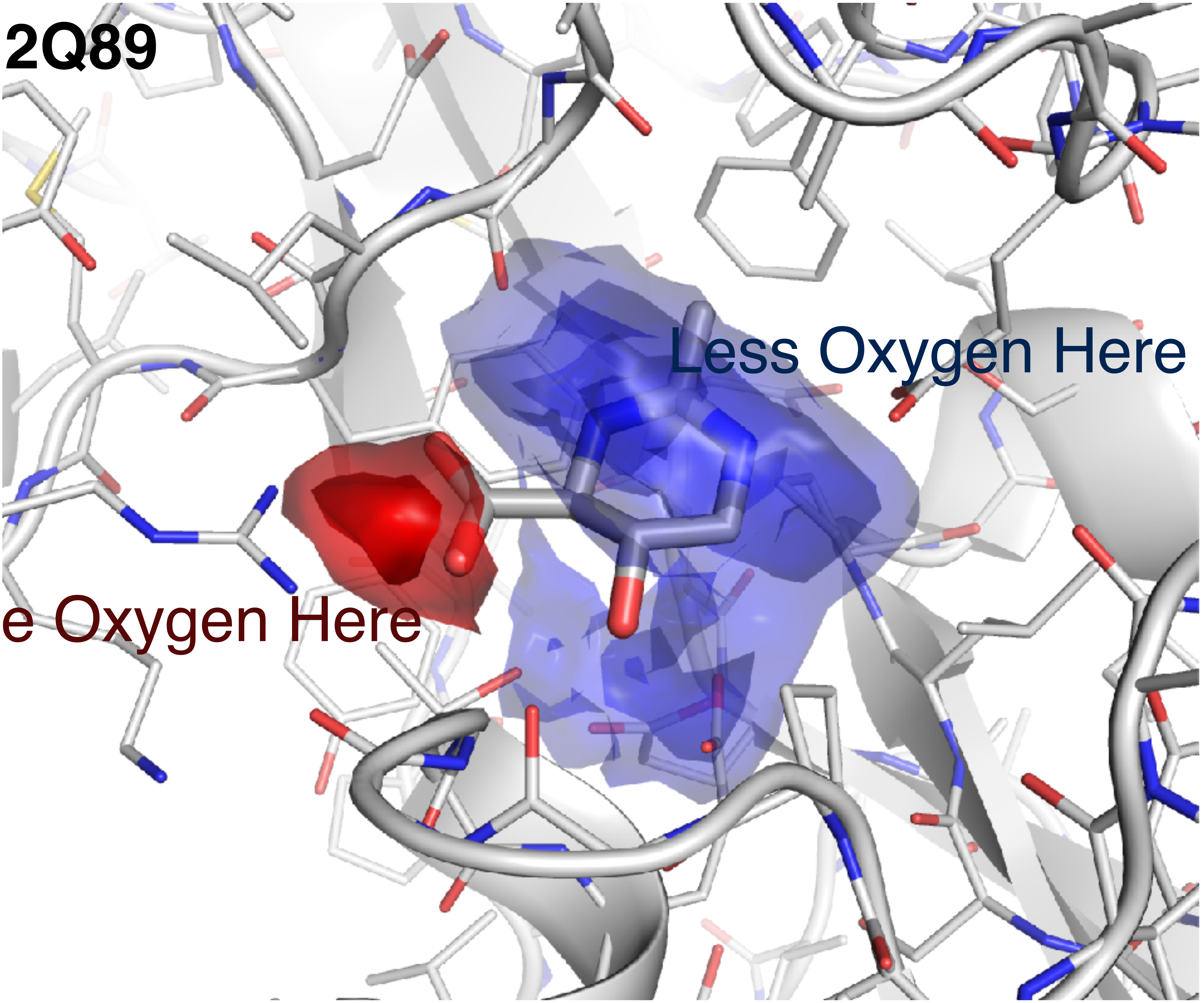
Beyond Scoring



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

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

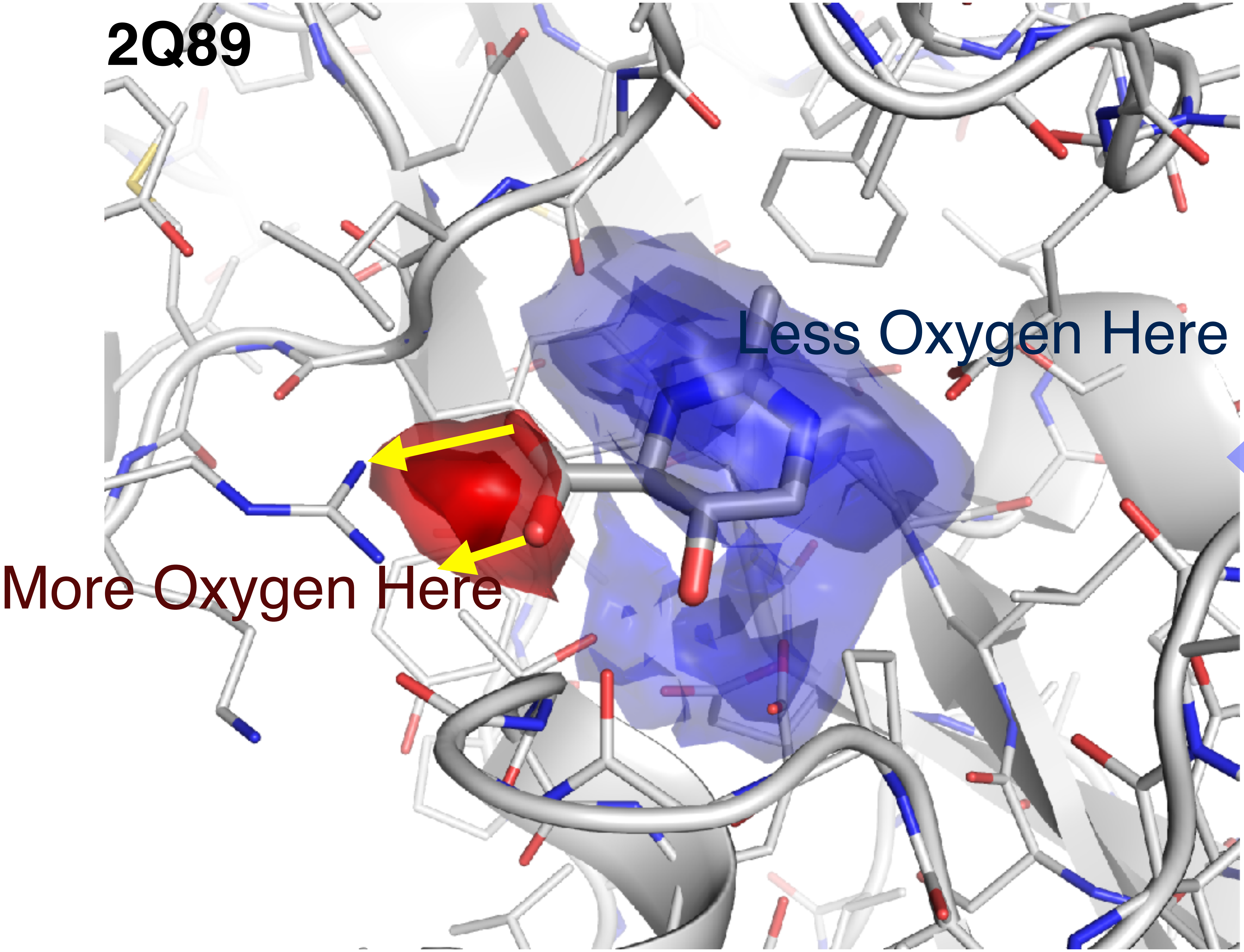
Beyond Scoring

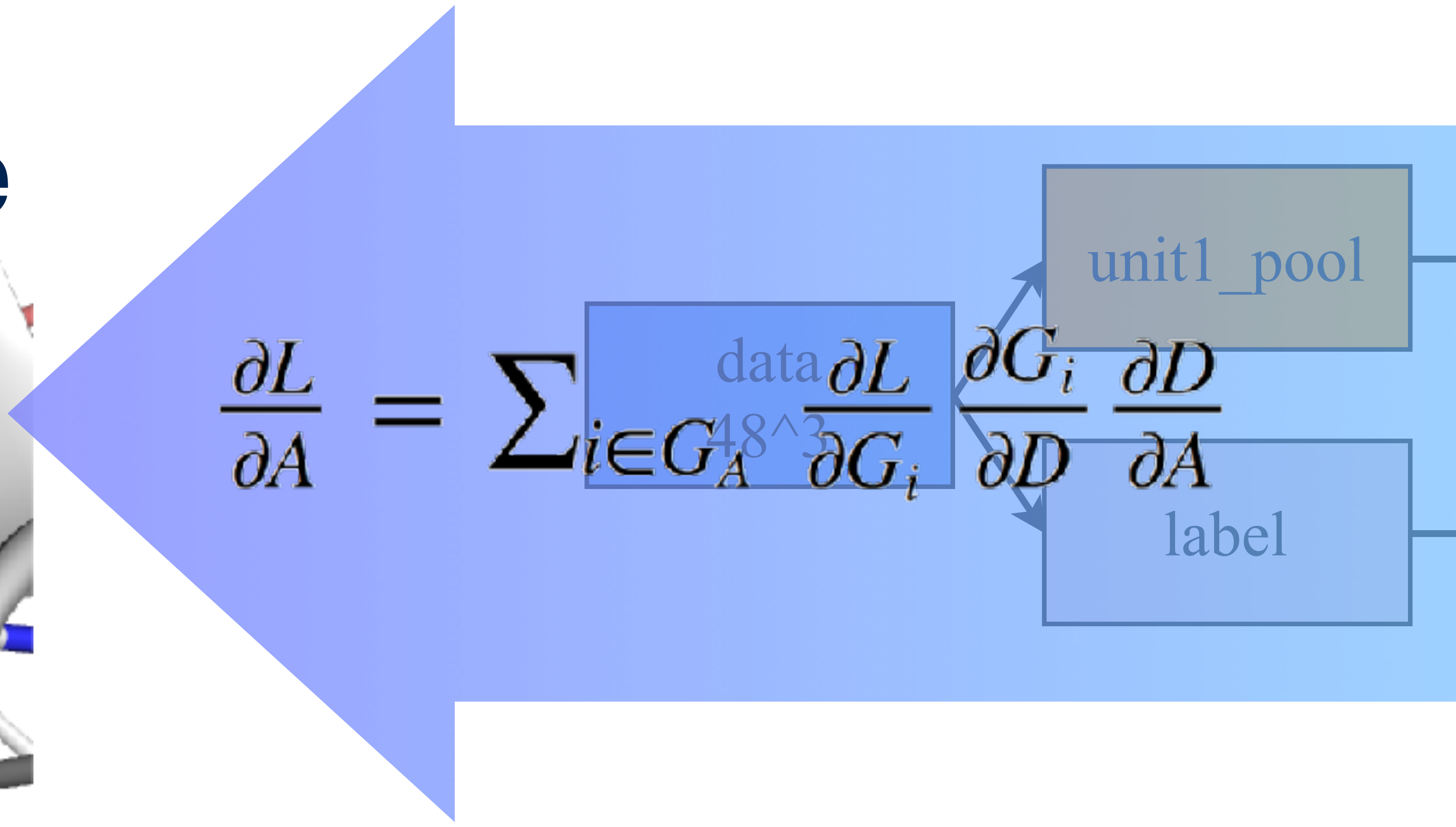


Less Oxygen Here

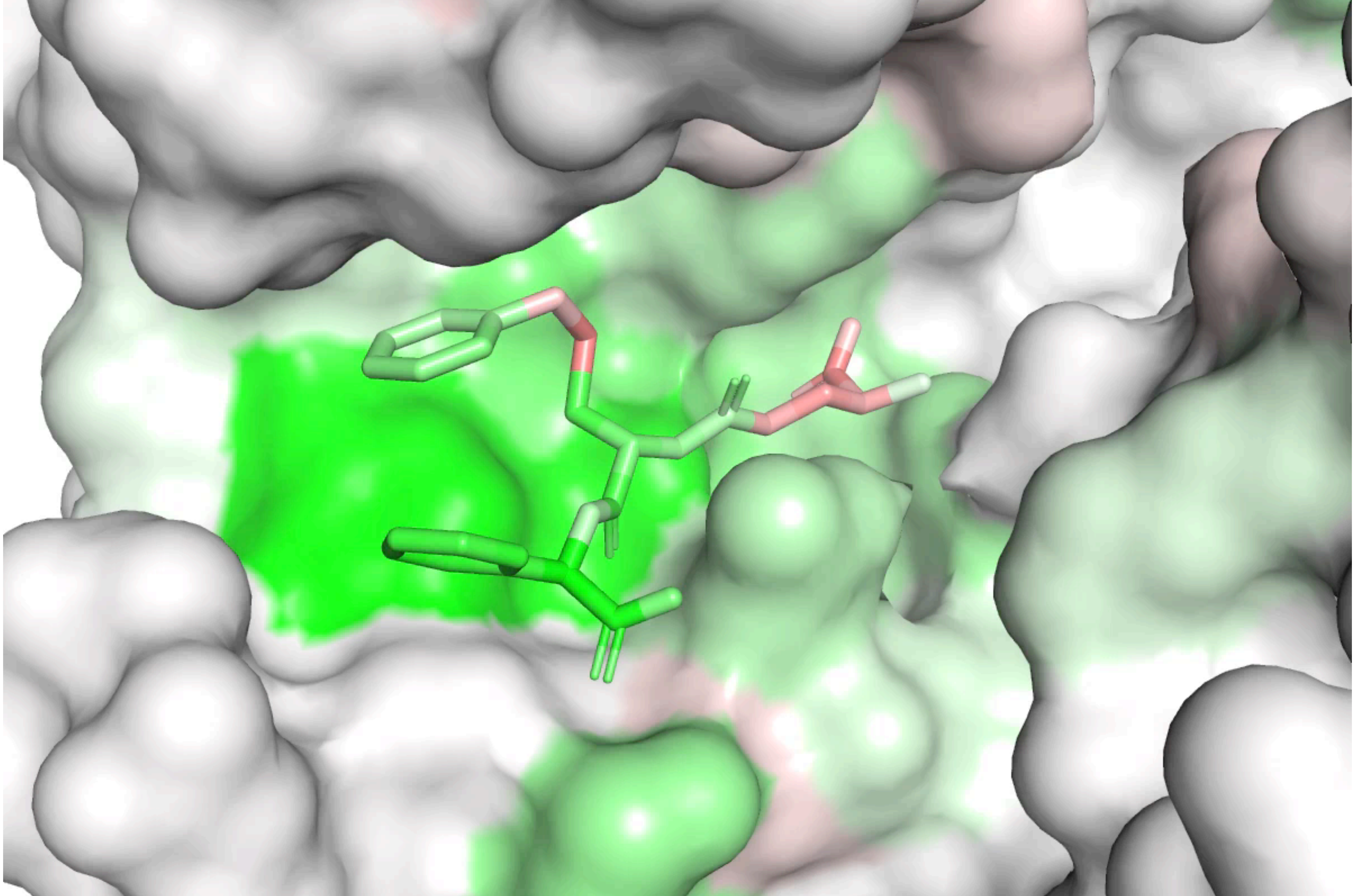
More Oxygen Here

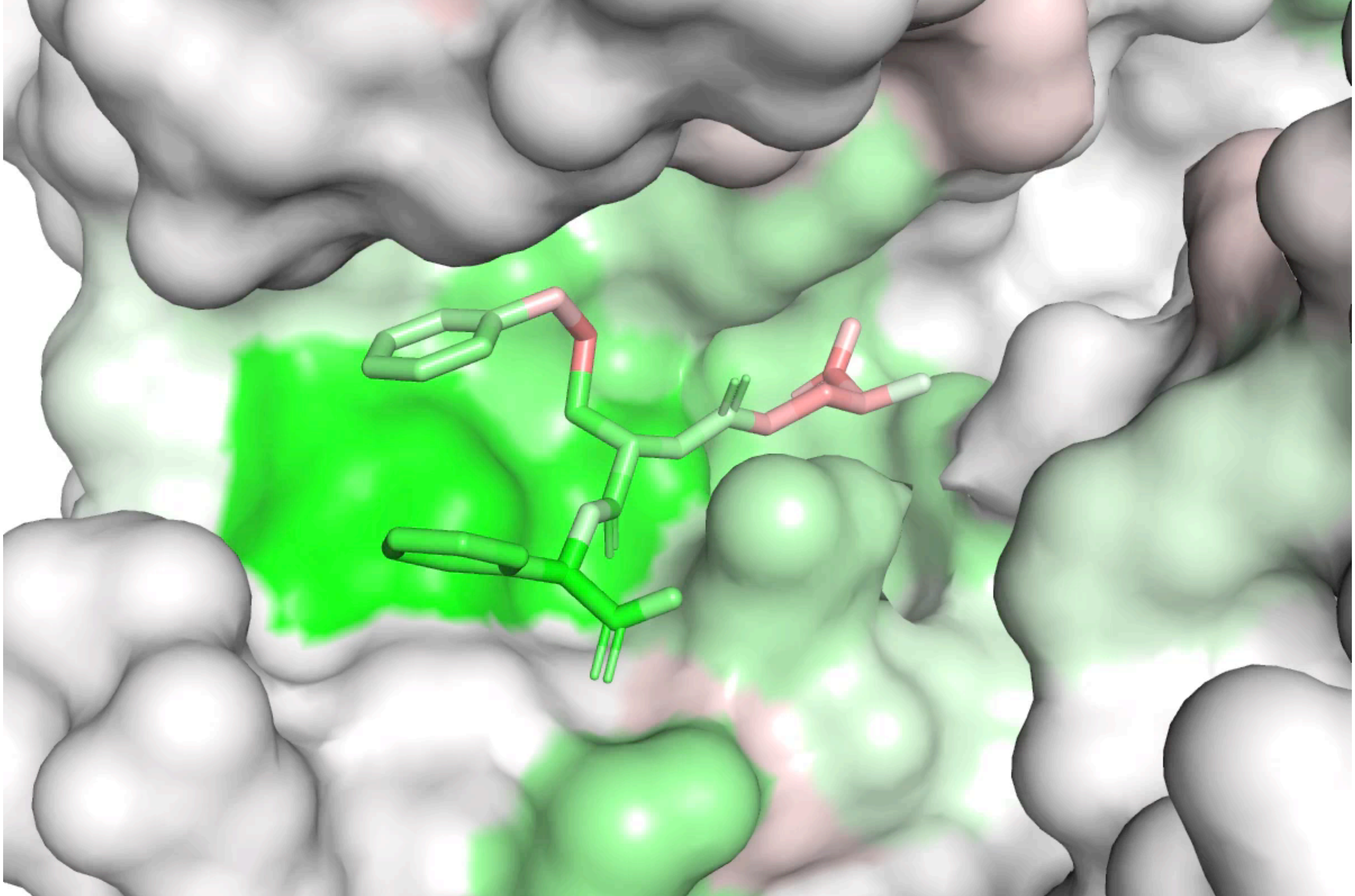
Beyond Scoring

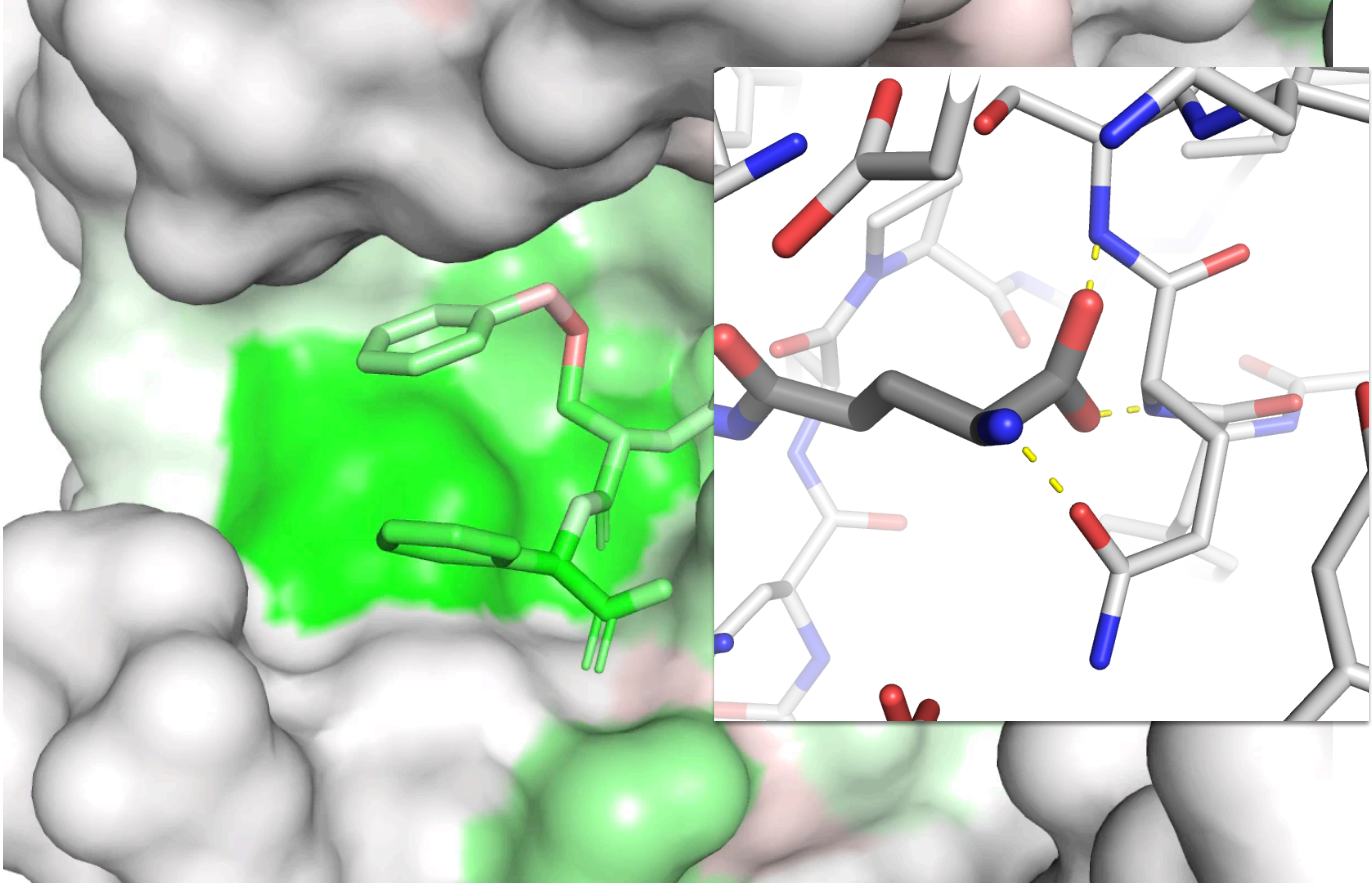


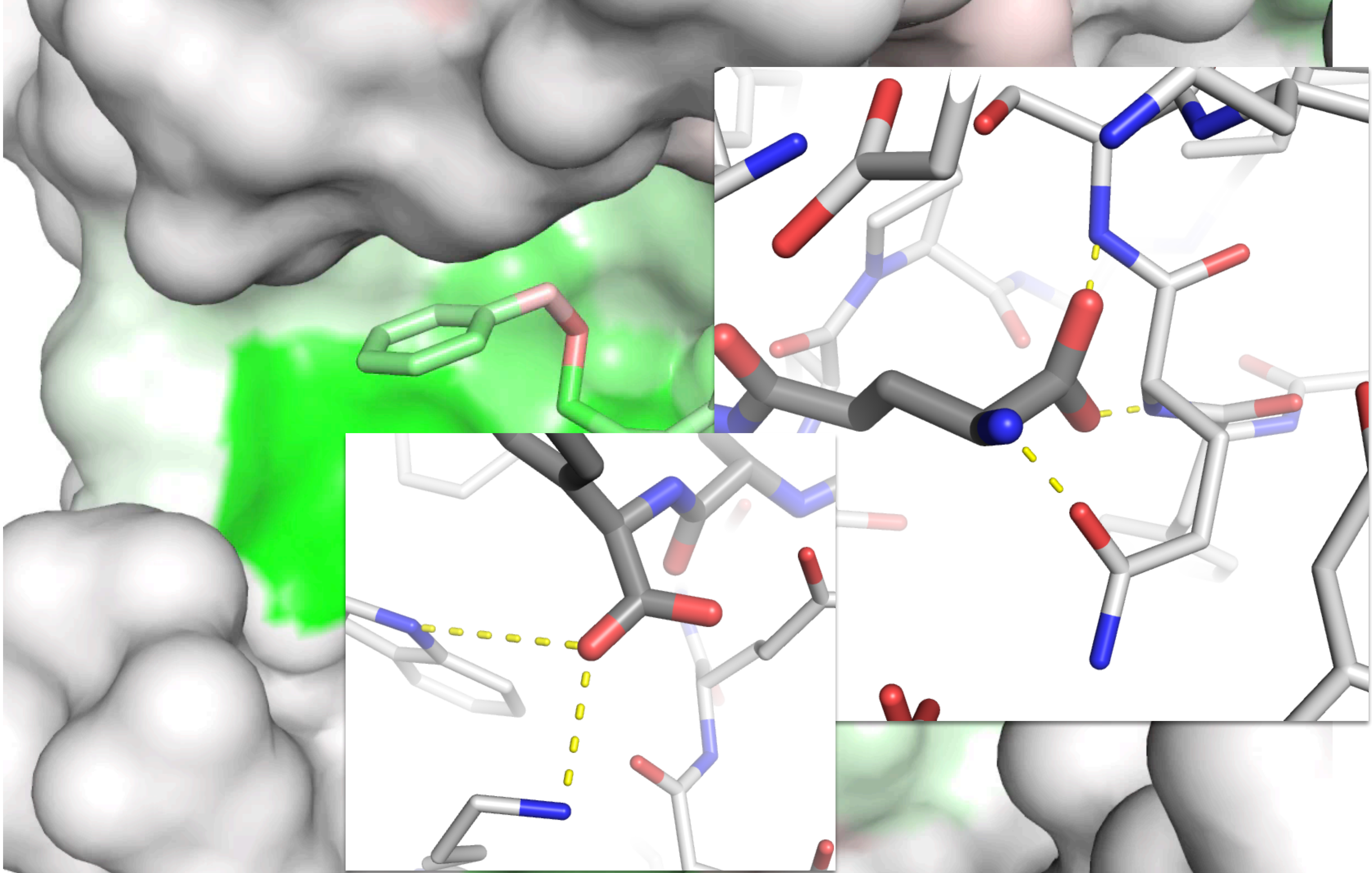


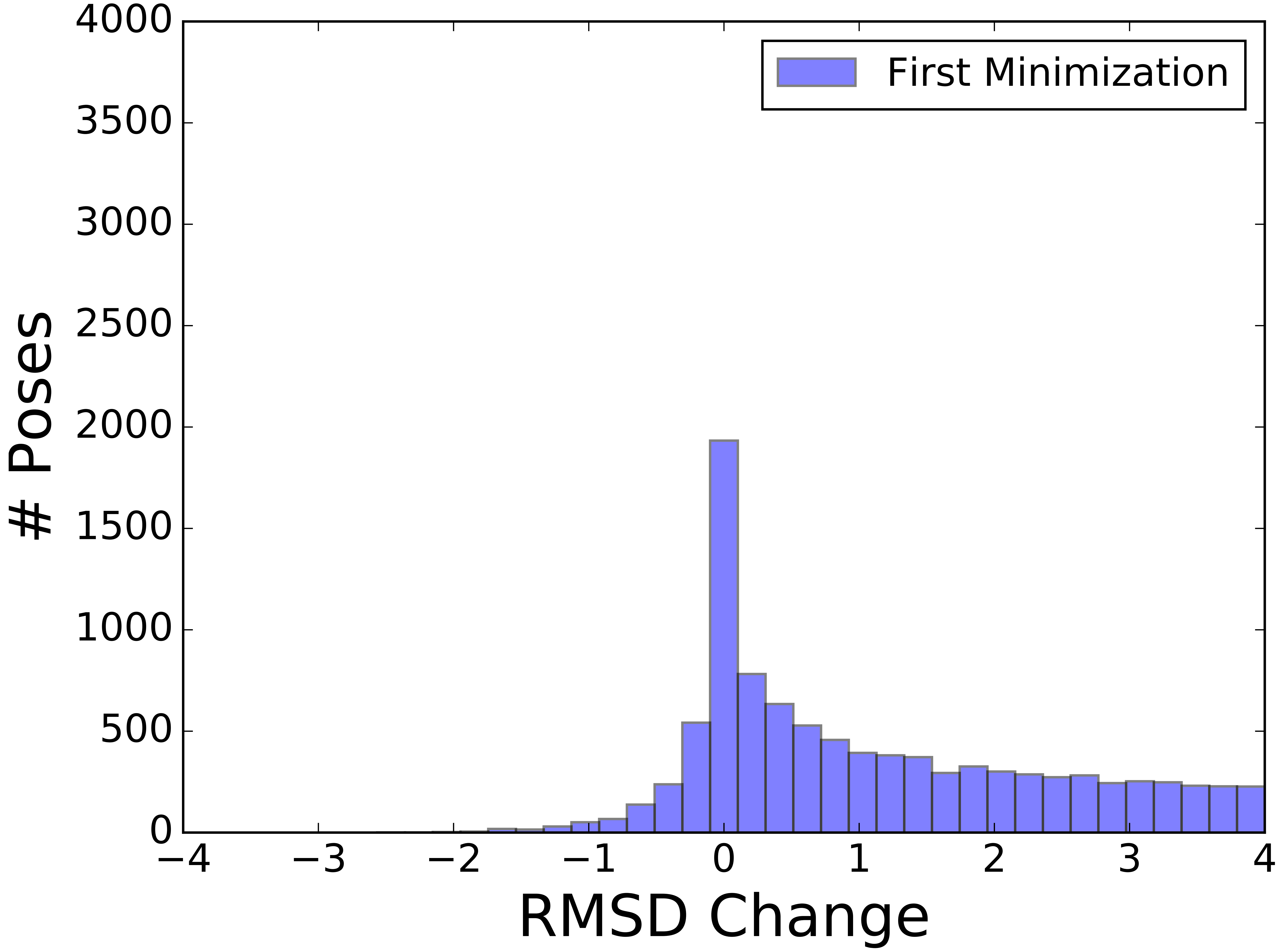
$$\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

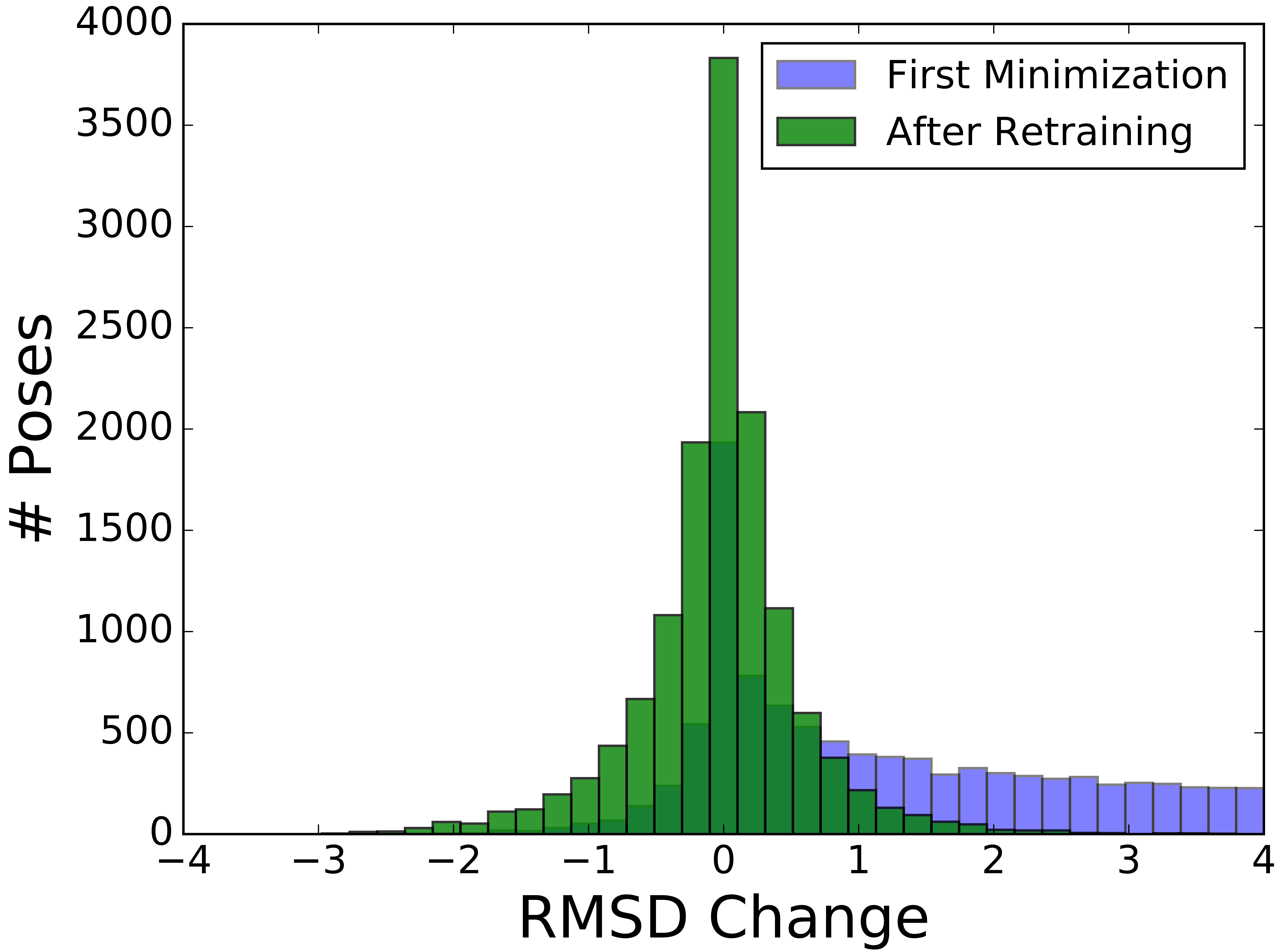






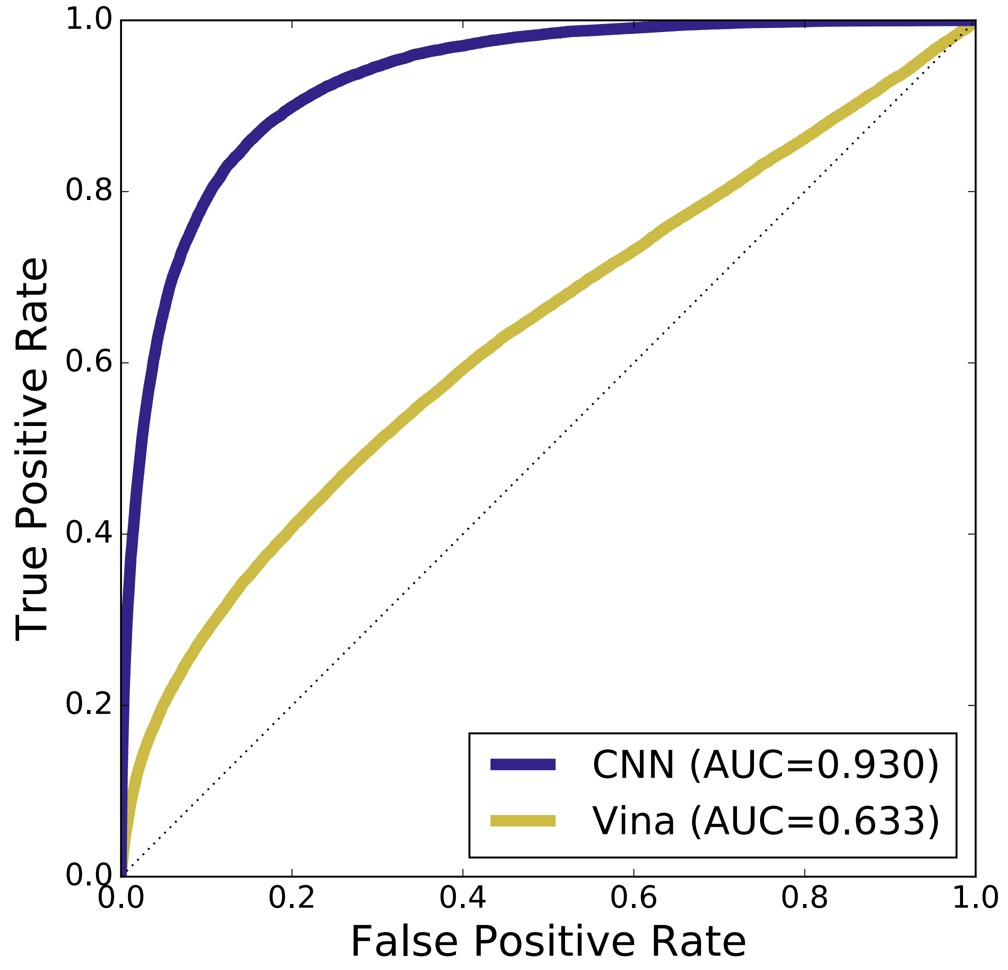



First Minimization
# Poses
RMSD Change
0
500
1000
1500
2000
2500
3000
3500
4000
−4
−3
−2
−1
0
1
2
3
4

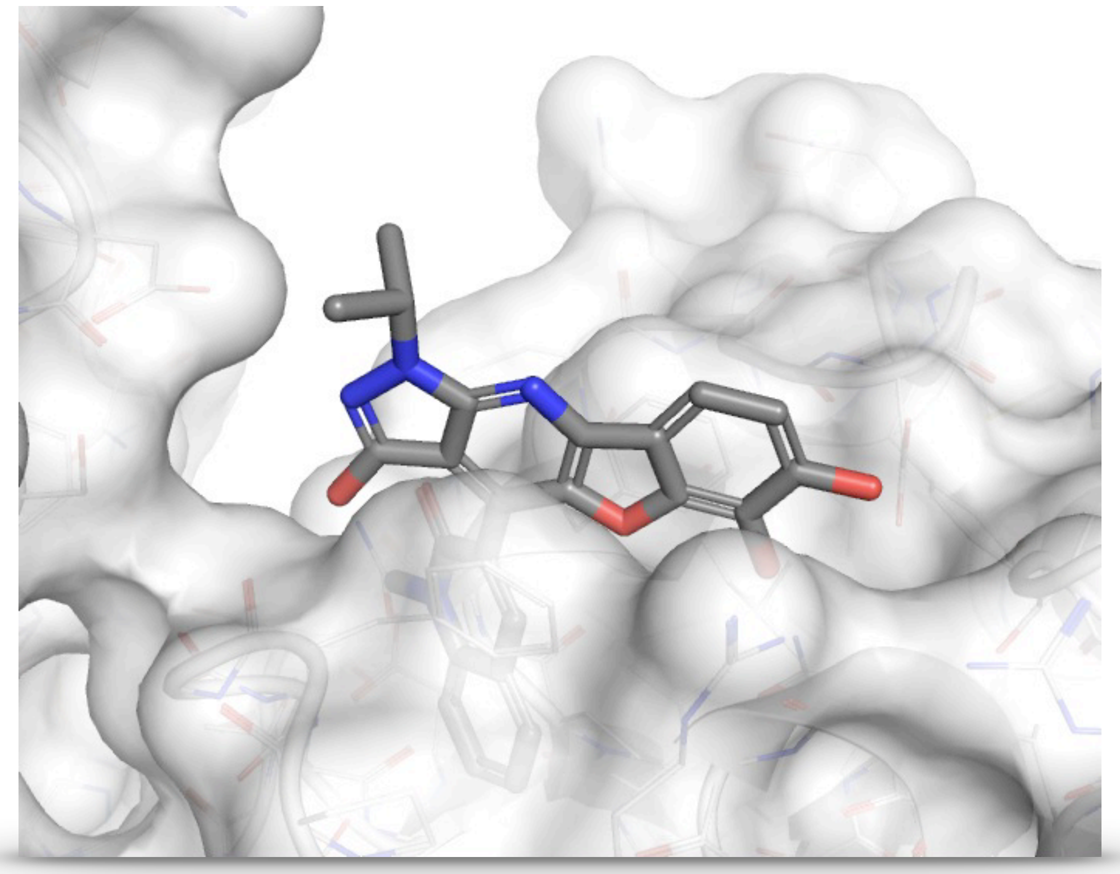

4000
3500
3000
2500
2000
1500
1000
500
0
# Poses
−4
−3
−2
−1
0
1
2
3
4
RMSD Change
First Minimization
After Retraining

The Future

Pose Selection

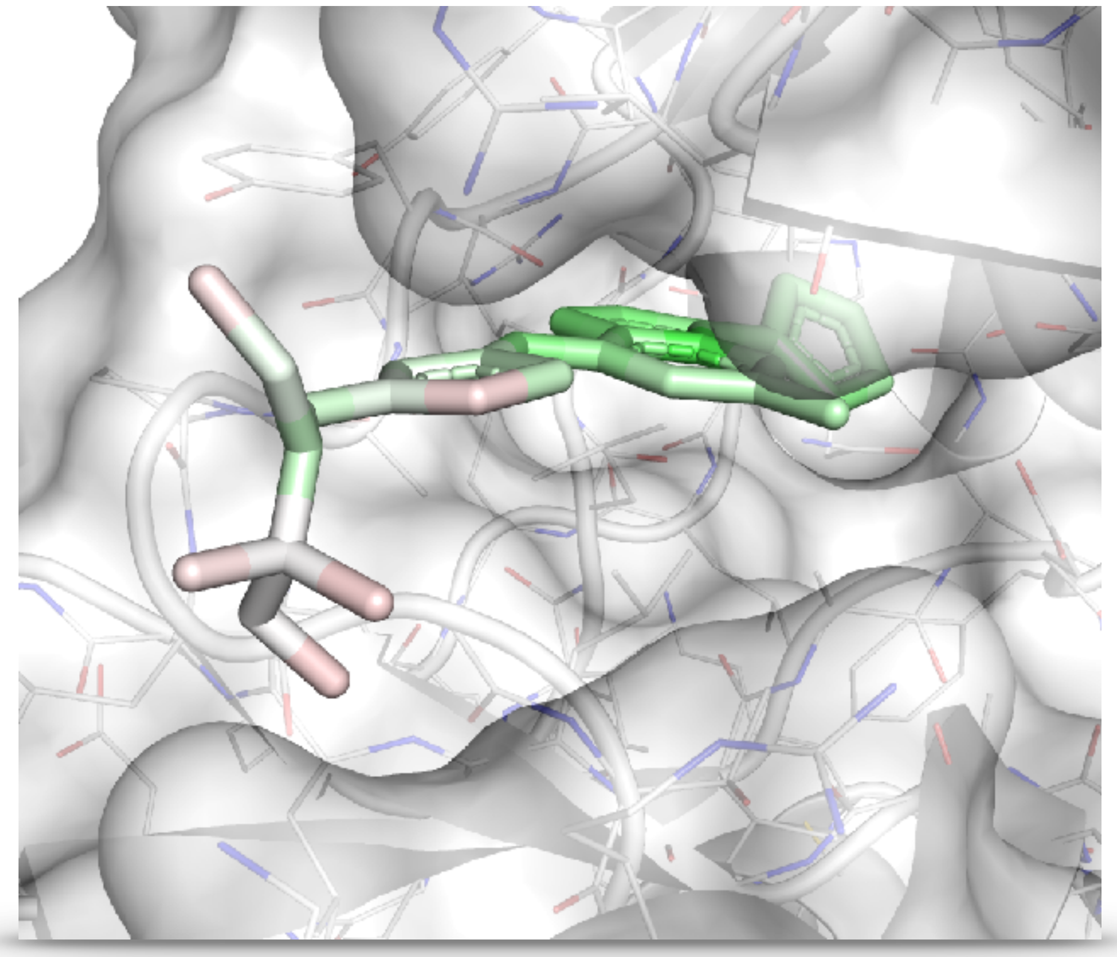


Pose *Generation*



Virtual Screening

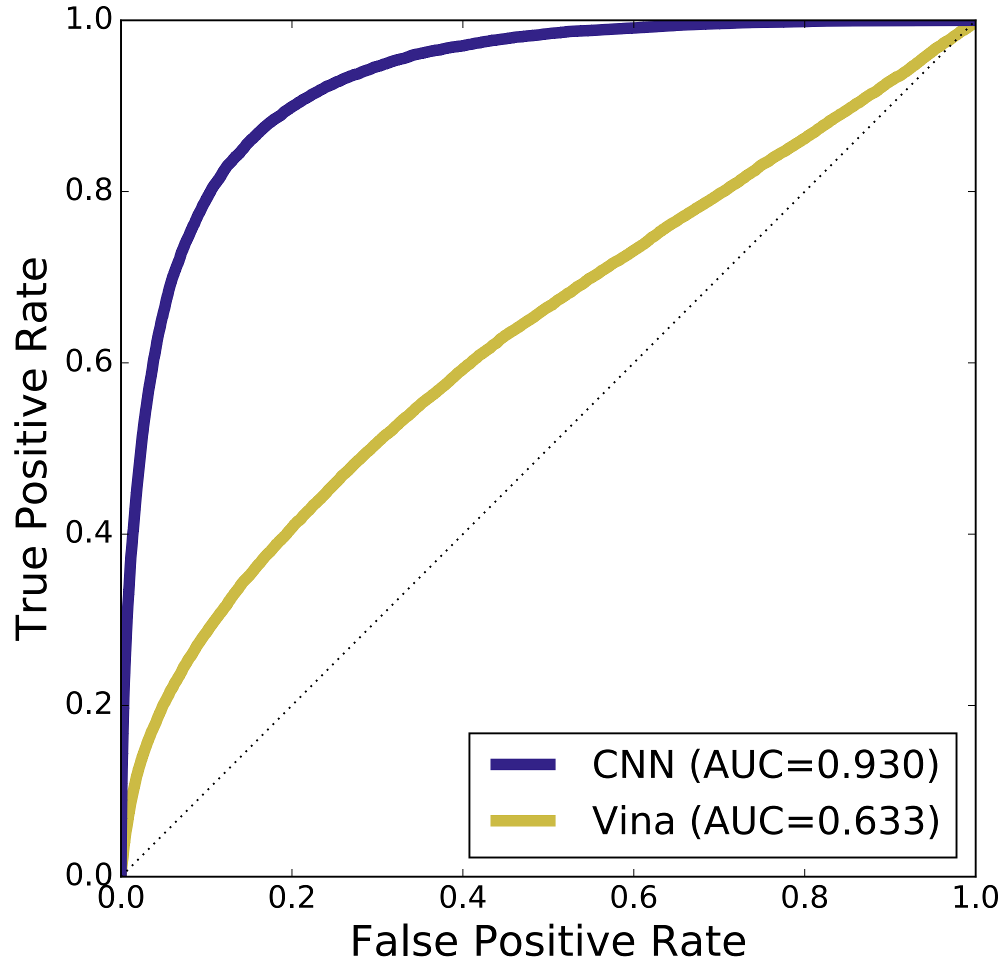
Compound Generation



Lead Optimization

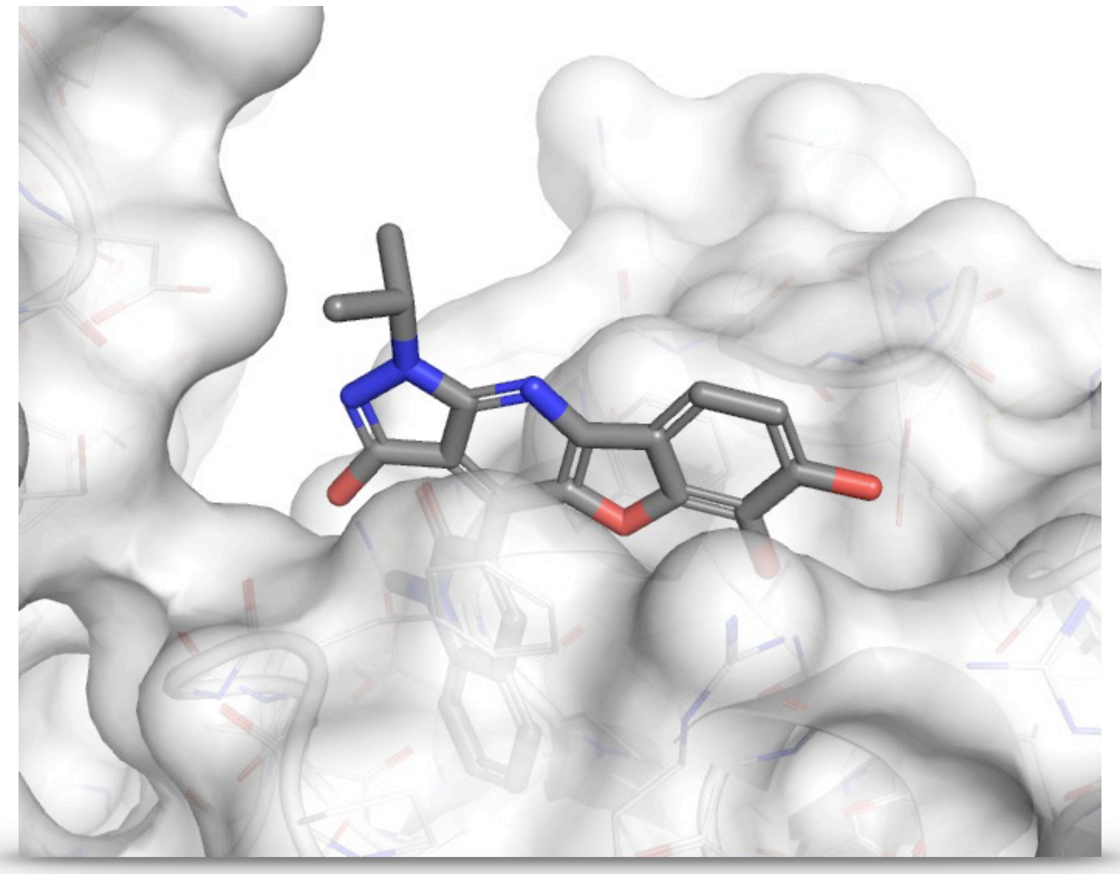
The Future

Pose Selection



Iterative Training

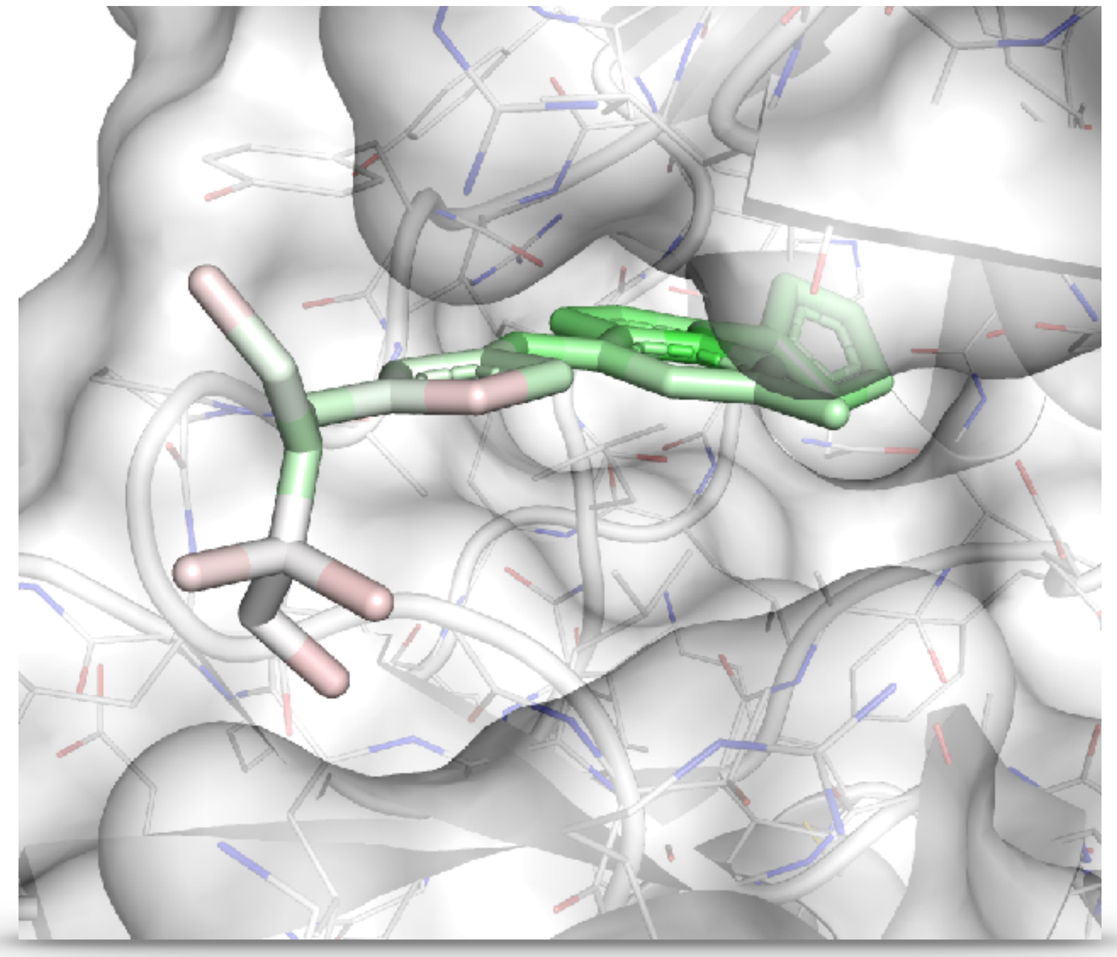
Pose *Generation*



Virtual Screening

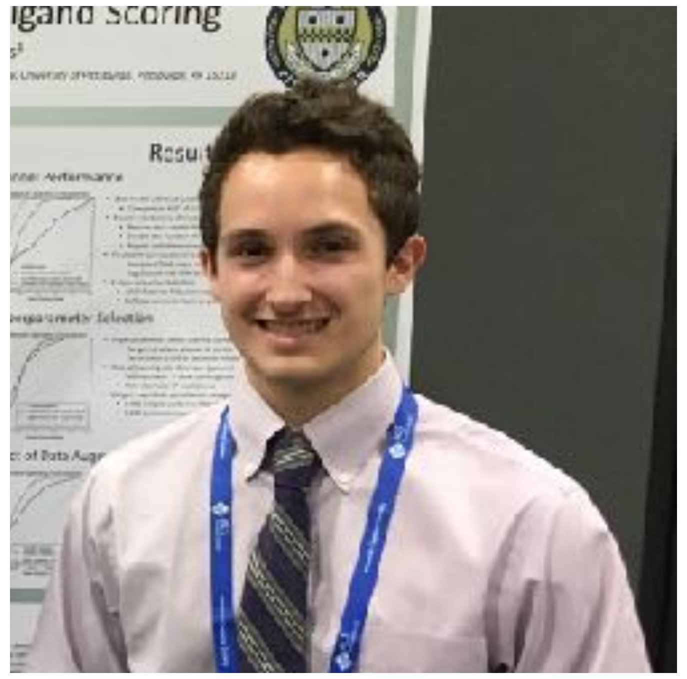
Iterative Training

Compound Generation

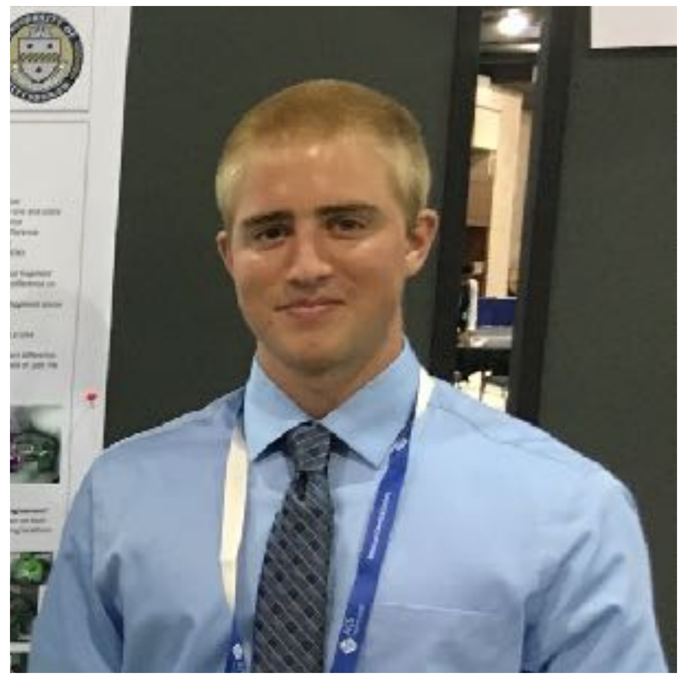


Lead Optimization

Acknowledgements

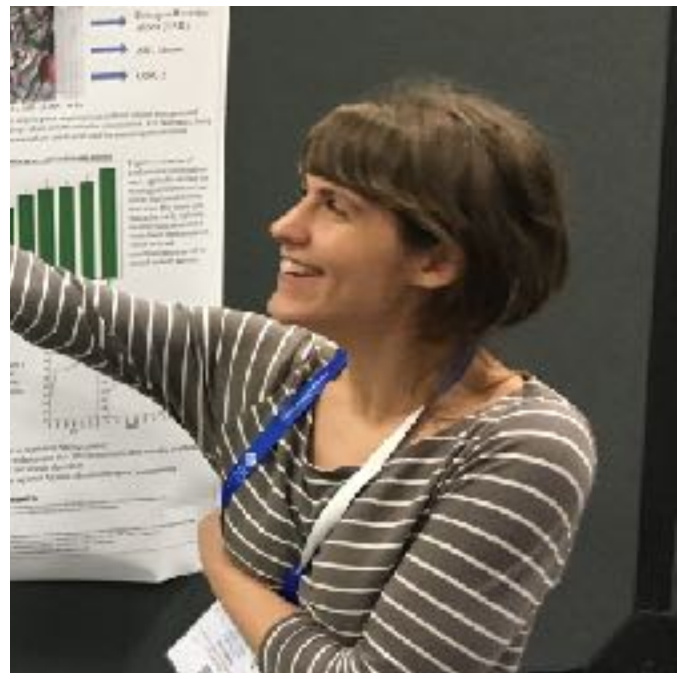


Matt Ragoza



Josh Hochuli

Elisa Idrobo



Jocelyn Sunseri

Group Members

Jocelyn Sunseri
Matt Ragoza
Josh Hochuli
Roosha Mandal
Alec Helbling
Lily Turner
Aaron Zheng
Sara Amato
Lily Turner
Aaron Zheng
Gibran Biswas

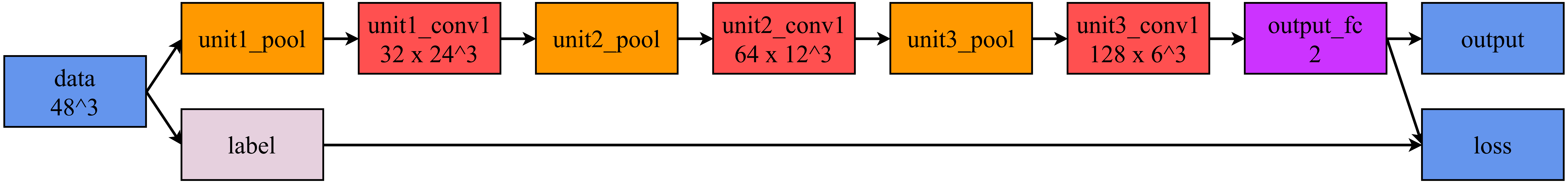
University of Pittsburgh Veritas Virtus 1787

Department of Computational and Systems Biology

TECBio REU

National Institute of General Medical Sciences
R01GM108340

Questions?



data 48^3

unit1_pool

unit1_conv1 32 x 24^3

unit2_pool

unit2_conv1 64 x 12^3

unit3_pool

unit3_conv1 128 x 6^3

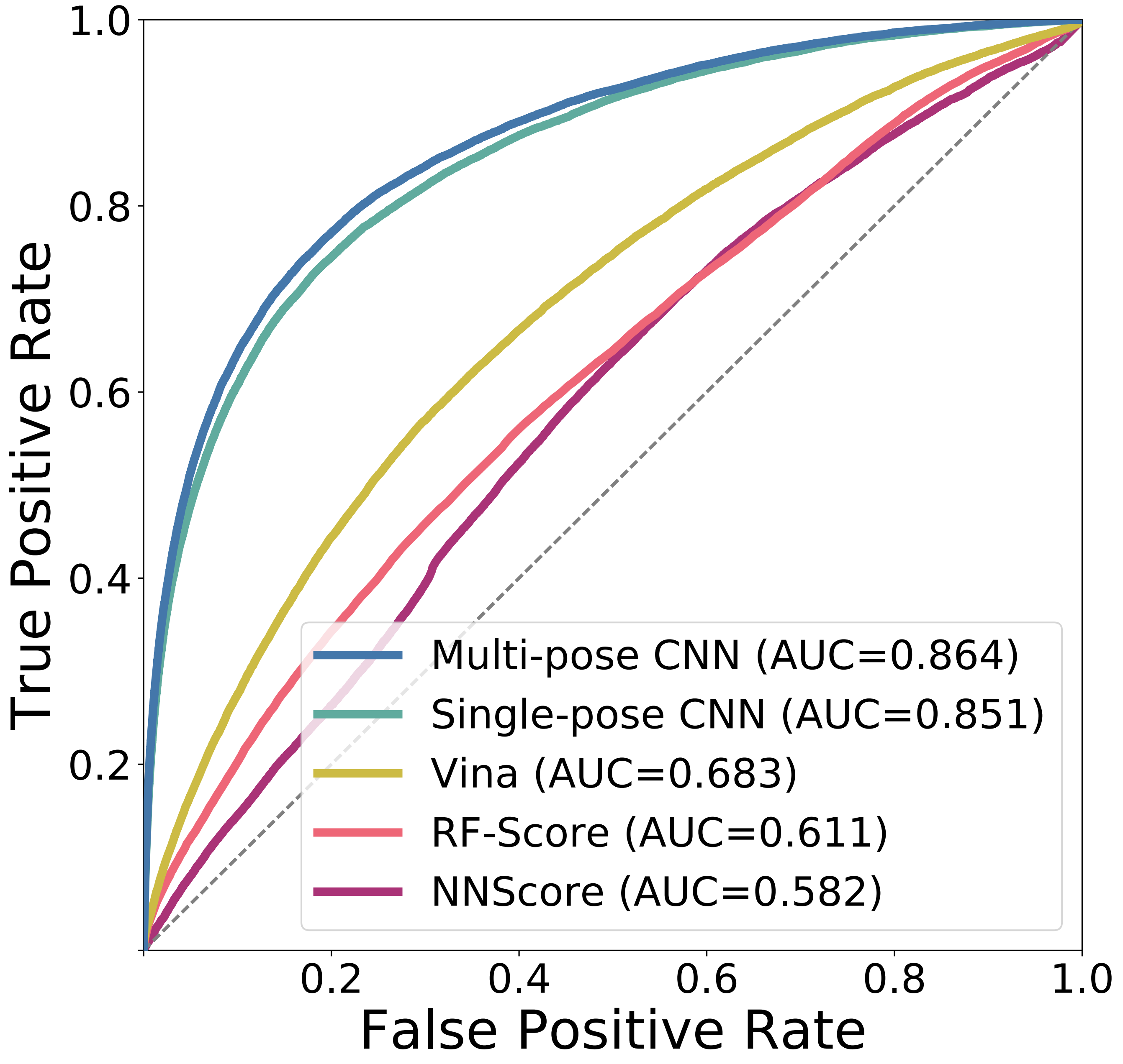
output_fc 2

output

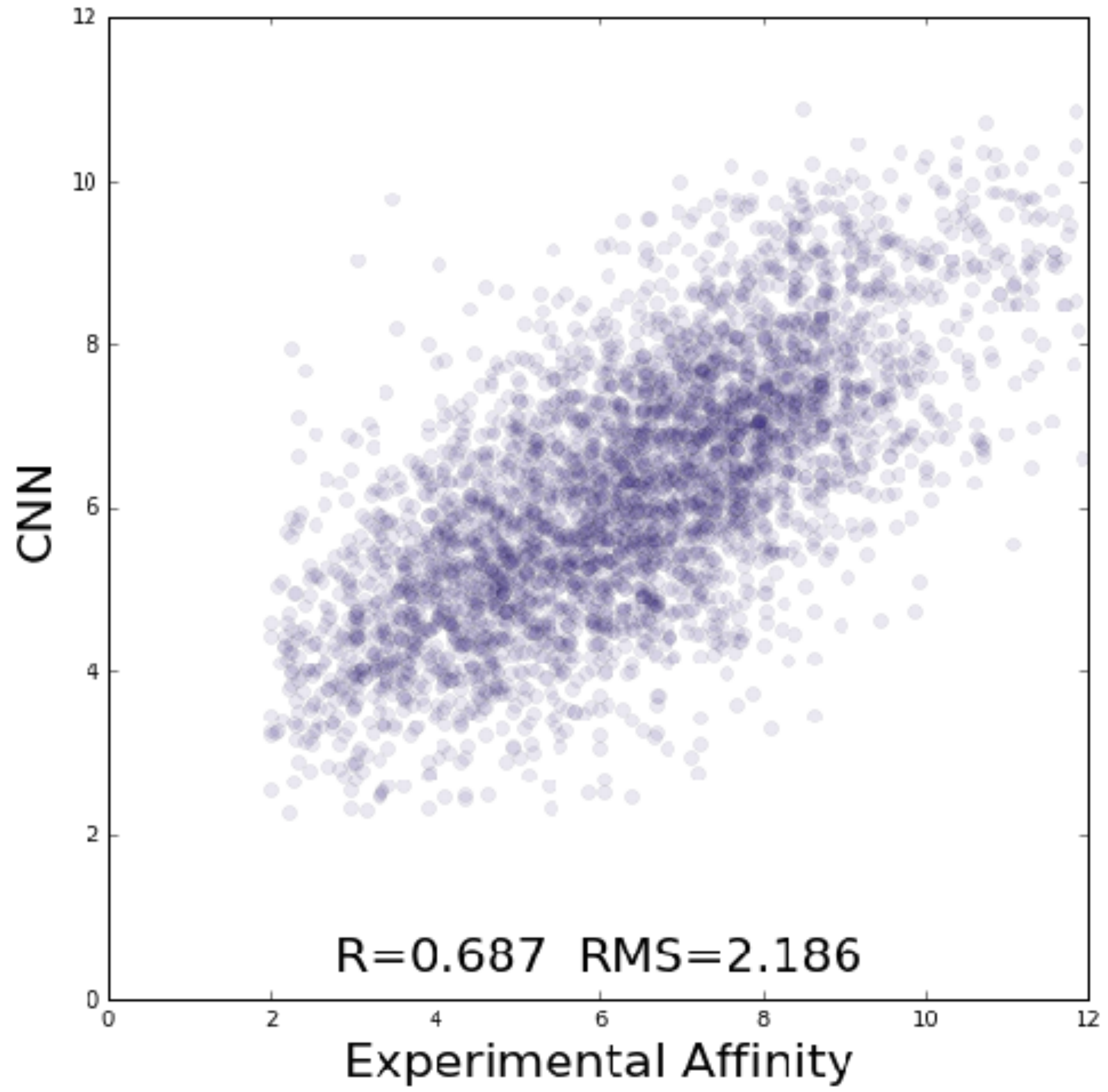
label

loss

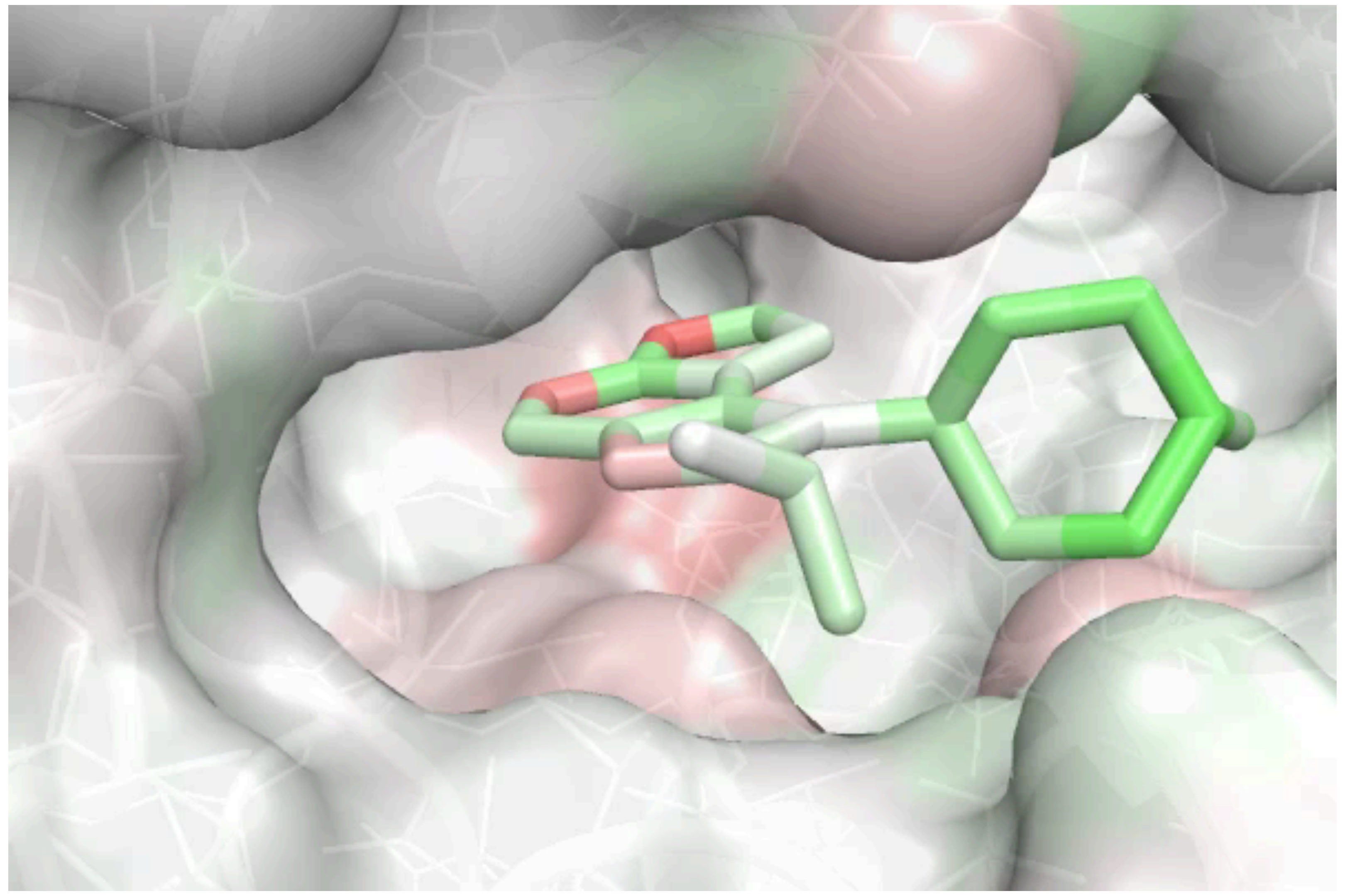
Binding Determination



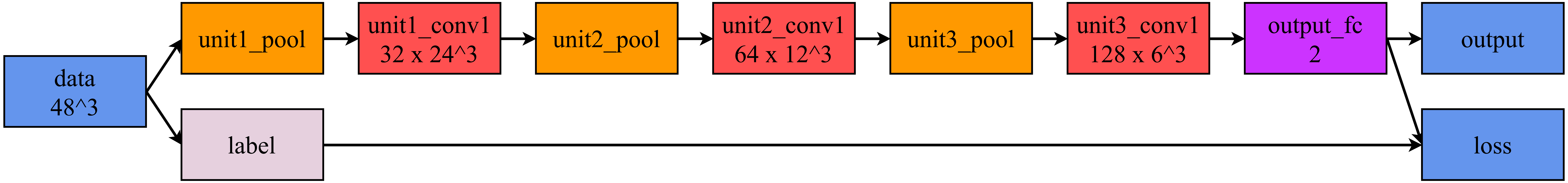
Affinity Prediction



Relevance Propagation



Questions?



data 48^3

unit1_pool

unit1_conv1 32 x 24^3

unit2_pool

unit2_conv1 64 x 12^3

unit3_pool

unit3_conv1 128 x 6^3

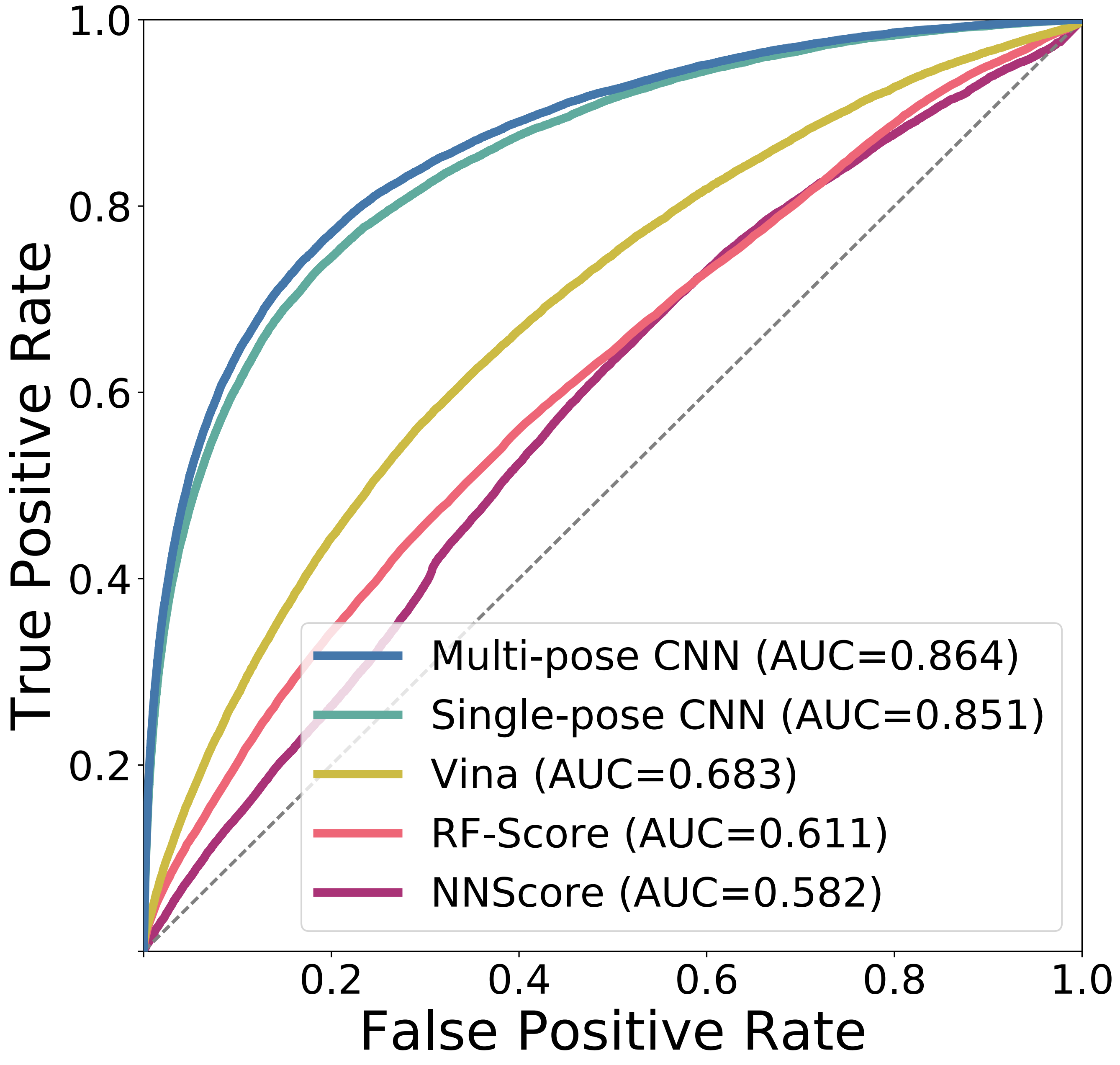
output_fc 2

output

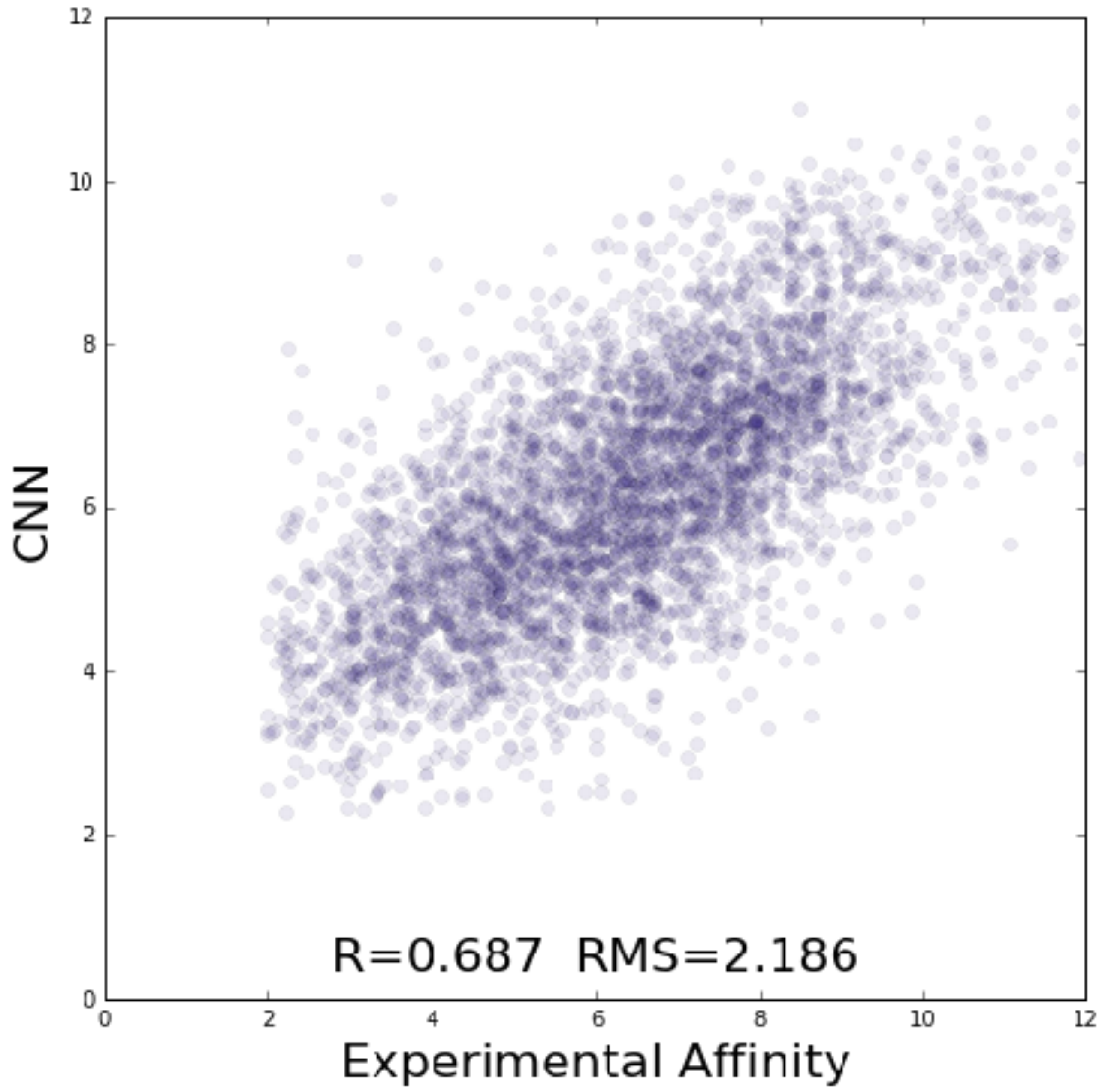
label

loss

Binding Determination



Affinity Prediction



Relevance Propagation