Protein-Ligand Scoring with Convolutional Neural Networks

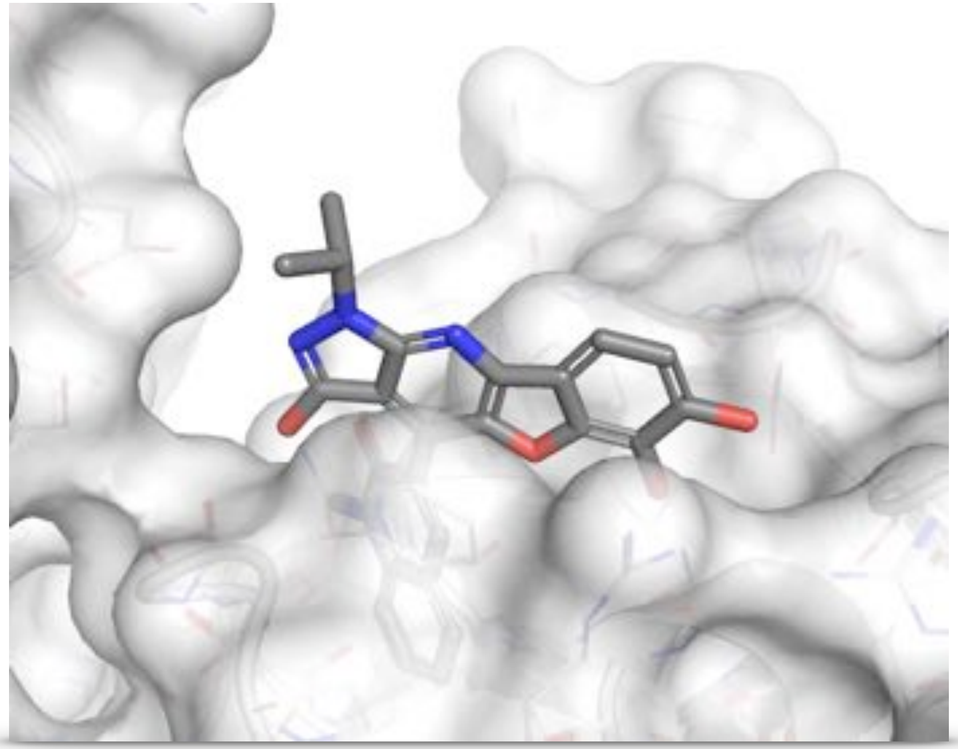
David Koes

@david_koes

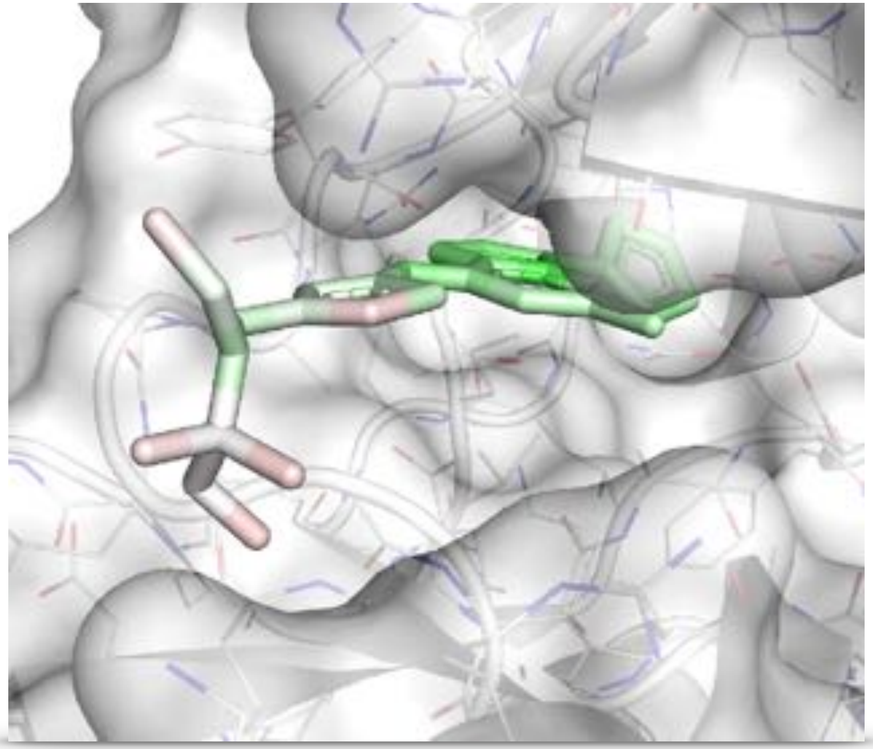
American Chemical Society Meeting
April 3, 2017

Structure Based Drug Design

Virtual Screening



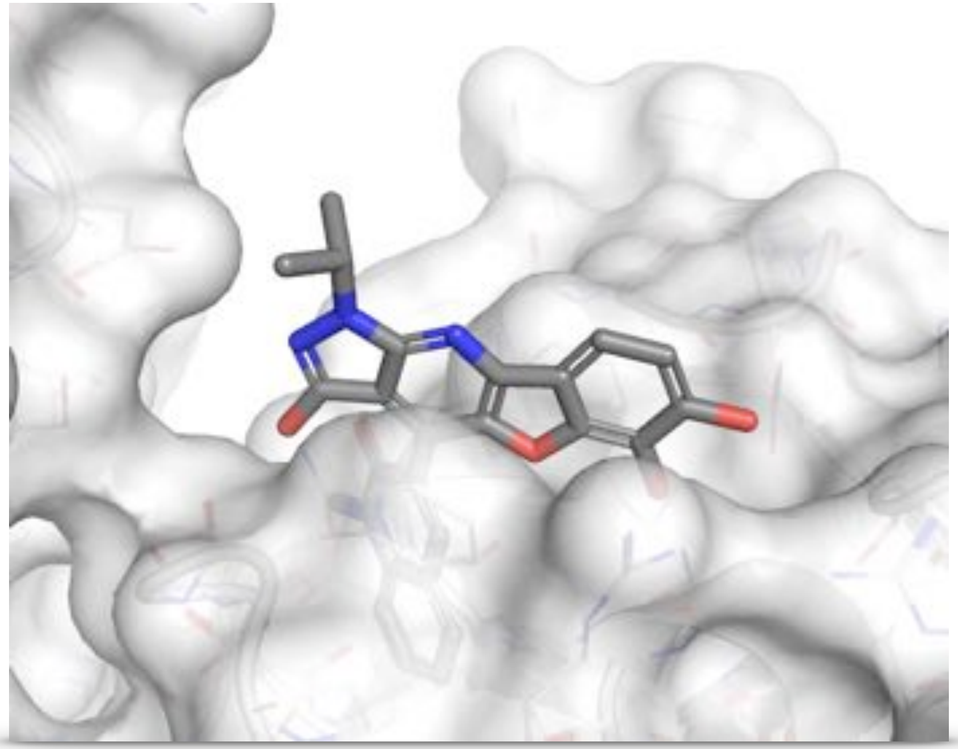
Lead Optimization



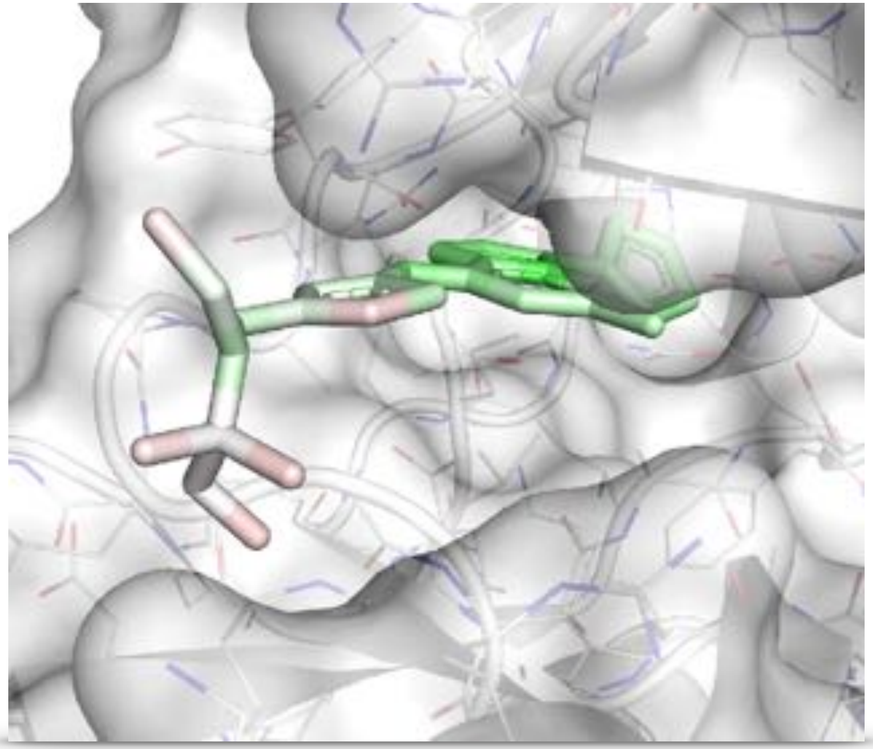
Pose Prediction → Binding Discrimination → Affinity Prediction

Structure Based Drug Design

Virtual Screening

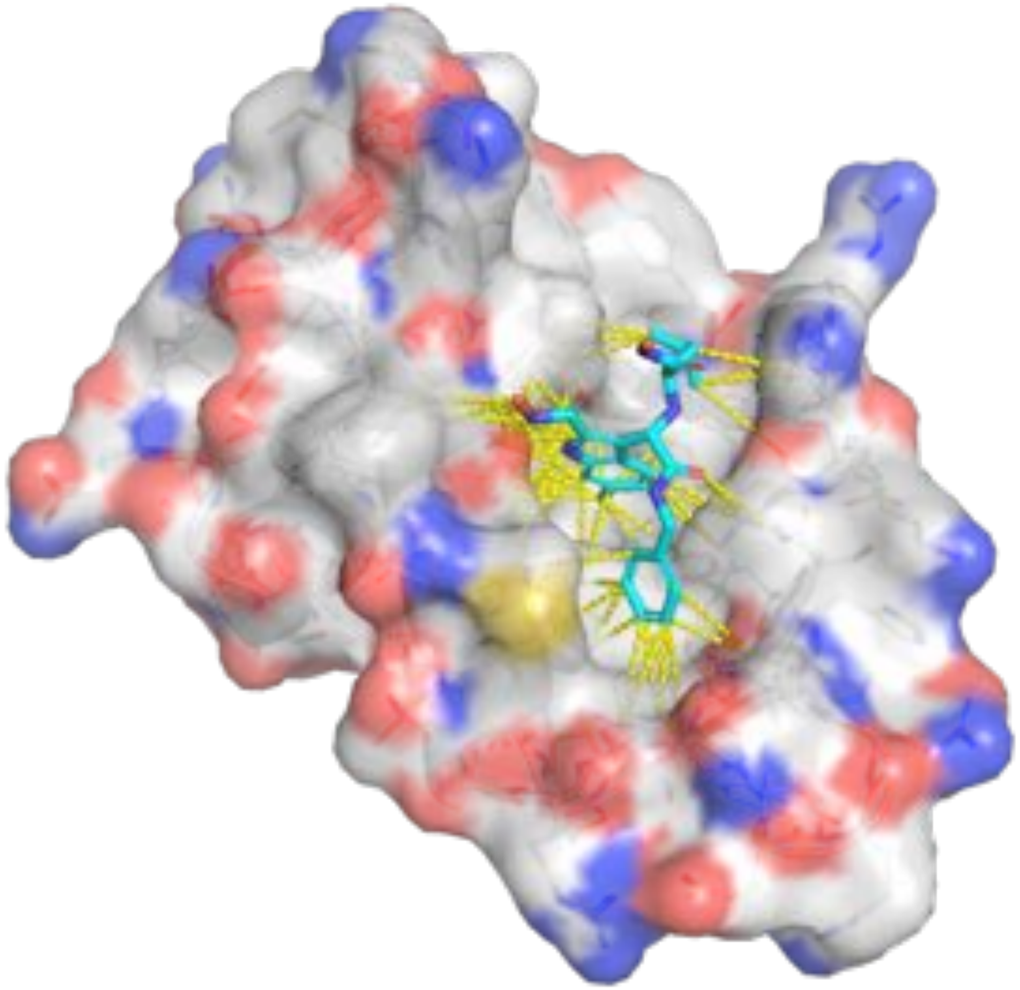


Lead Optimization



Pose Prediction → Binding Discrimination → Affinity Prediction

Protein-Ligand Scoring



AutoDock Vina

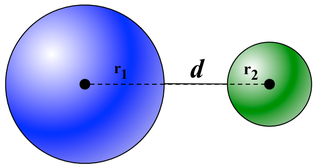
$$\text{gauss}_1(d) = w_{\text{guass}_1} e^{-(d/0.5)^2}$$

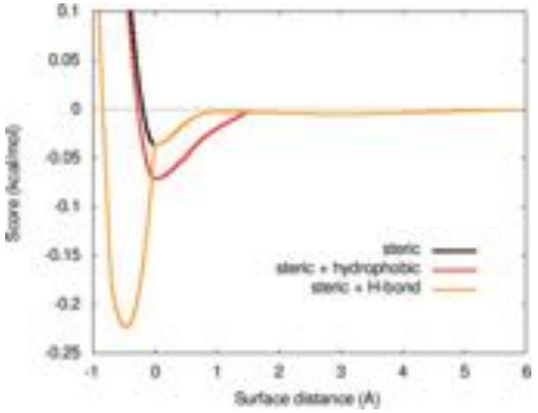
$$\text{gauss}_2(d) = w_{\text{guass}_2} e^{-((d-3)/2)^2}$$

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

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

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





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?

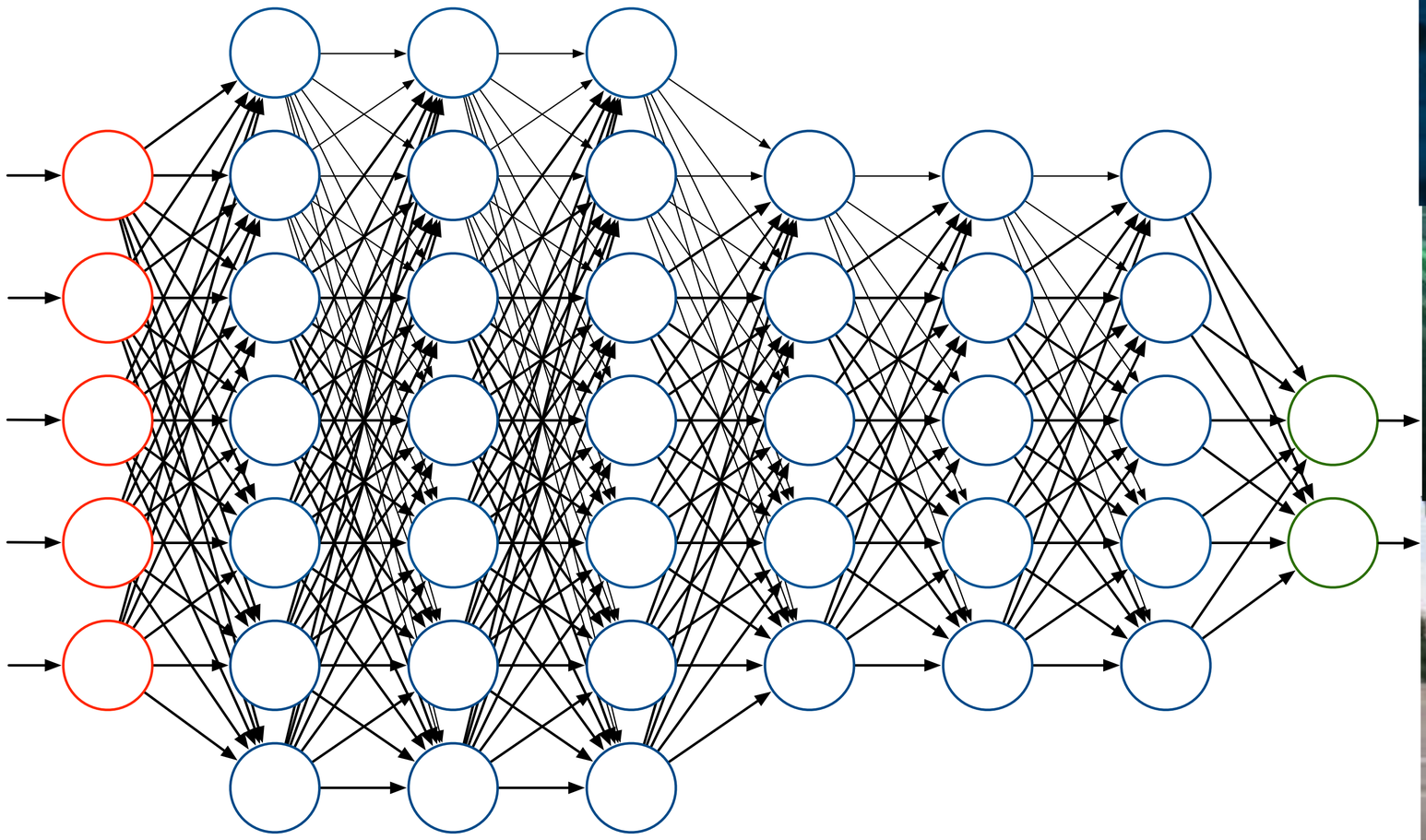
Can we do better?

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

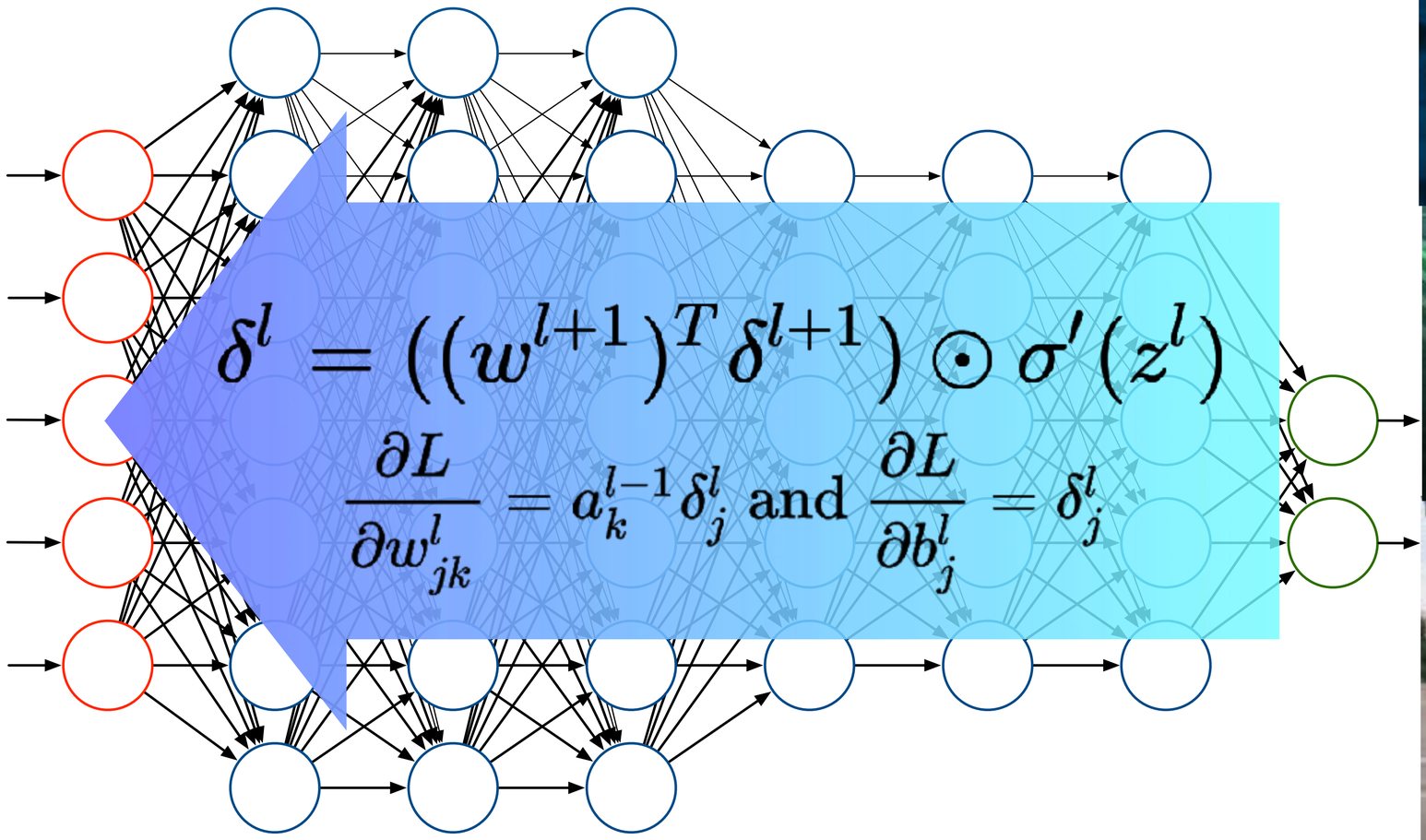
Key Idea: Leverage "big data"

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

Deep Learning



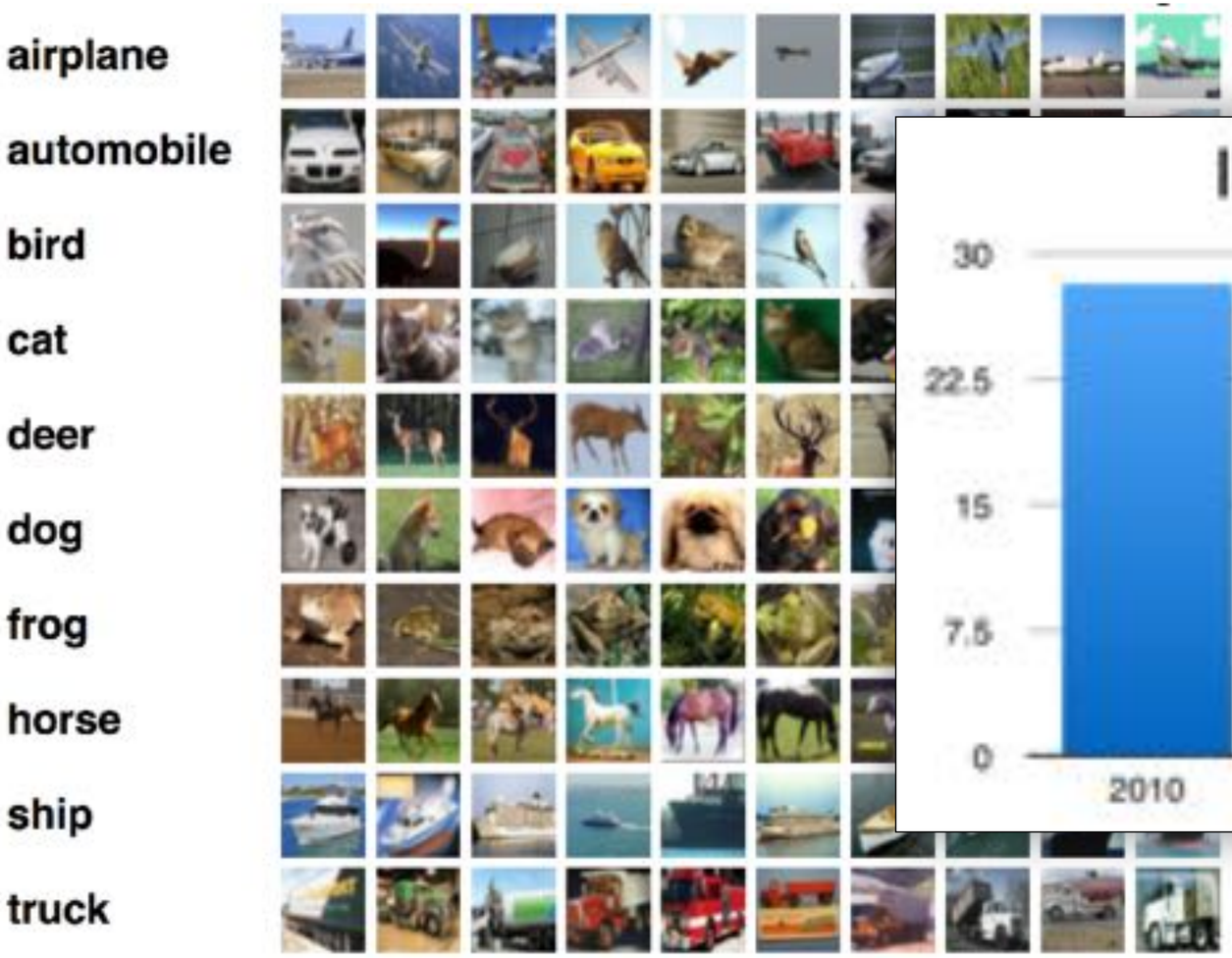
Deep Learning

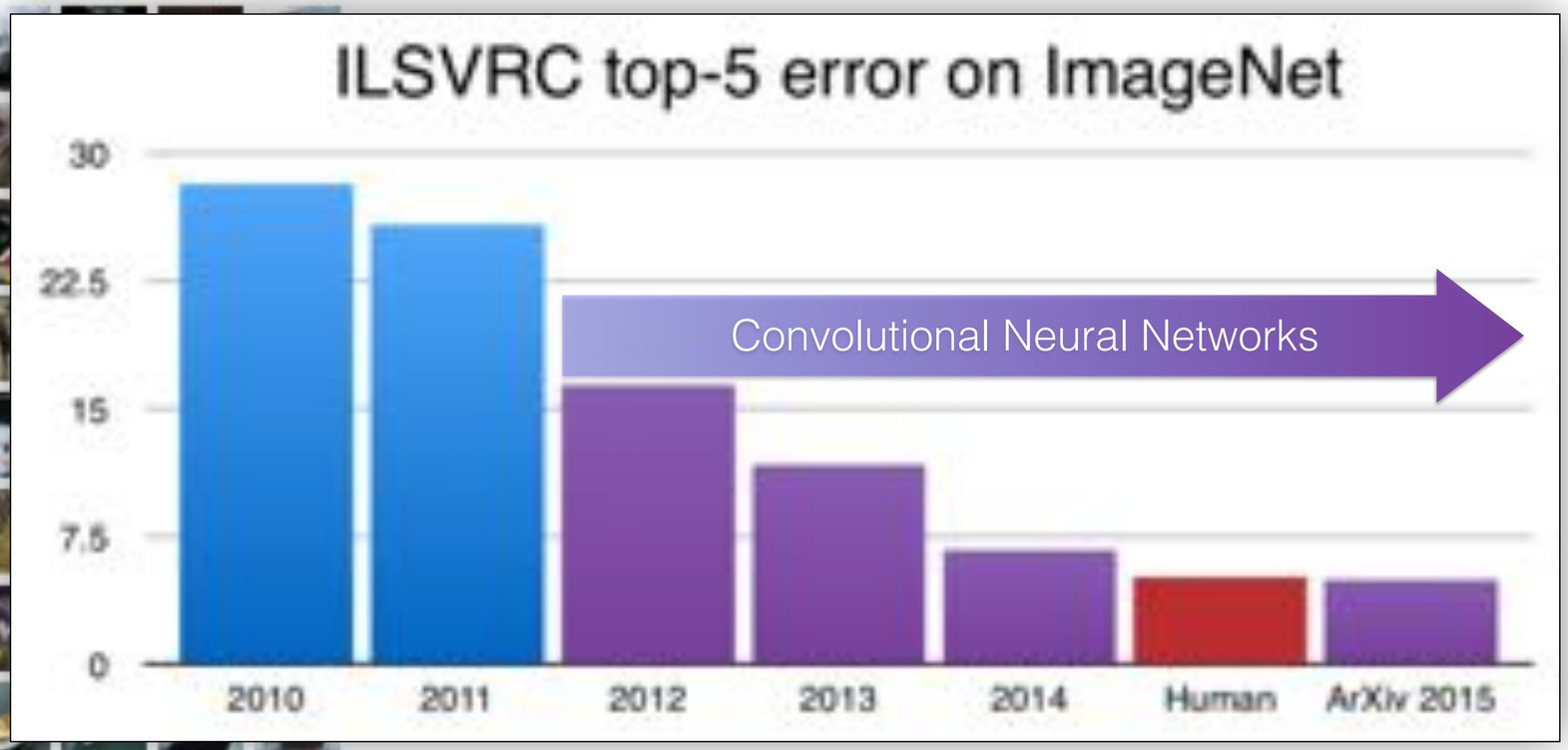


$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

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

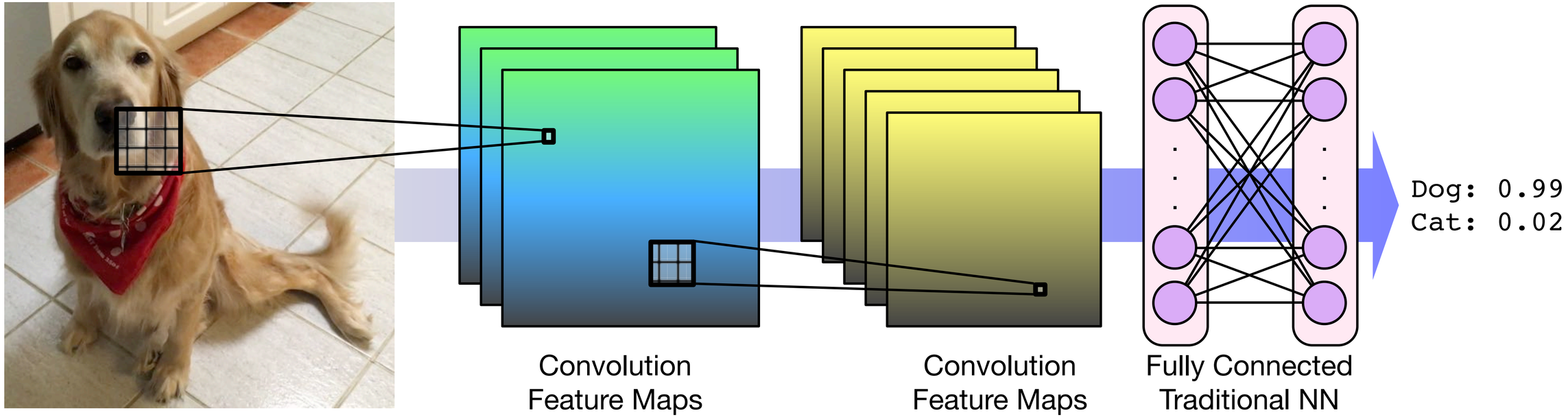
Image Recognition

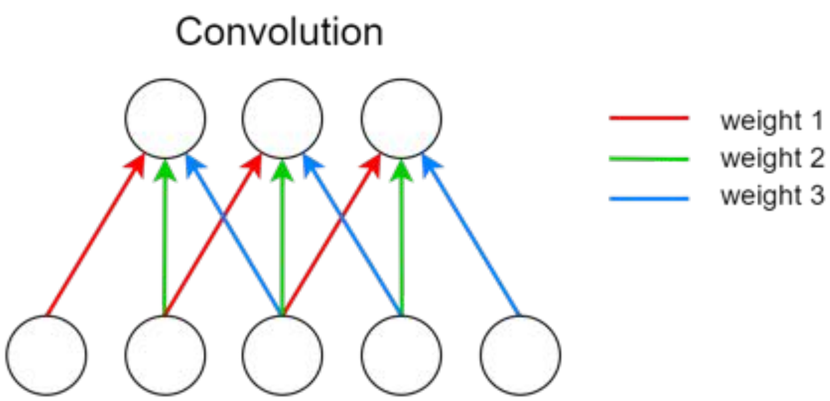


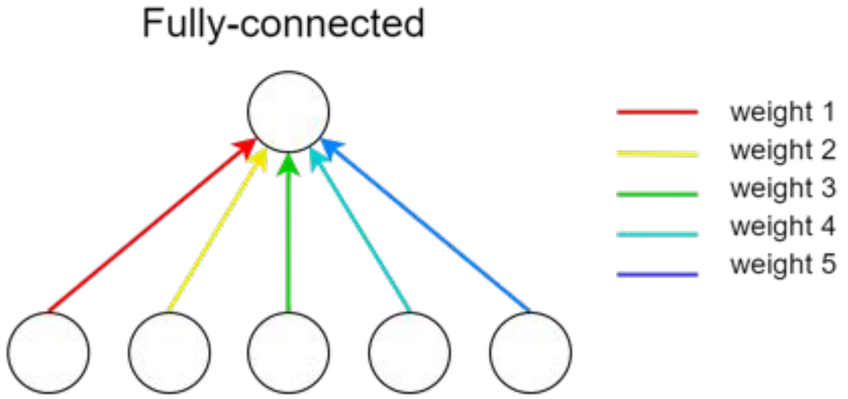


https://devblogs.nvidia.com

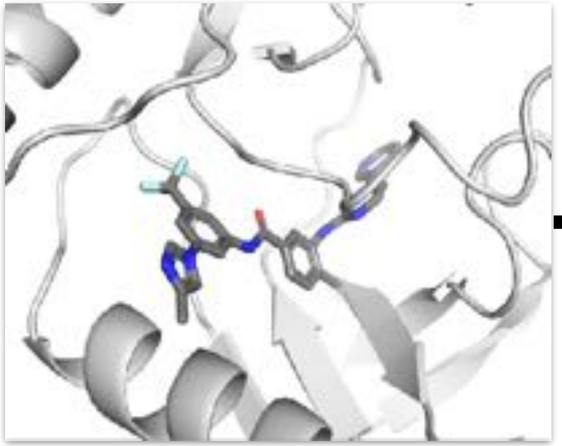
Convolutional Neural Networks







CNNs for Protein-Ligand Scoring



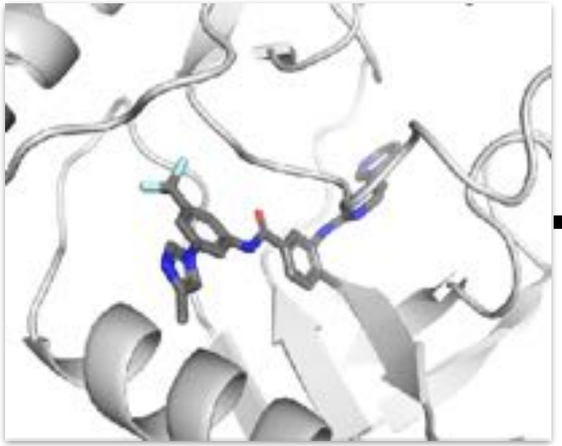
CNN

Pose Prediction

Binding Discrimination

Affinity Prediction

CNNs for Protein-Ligand Scoring



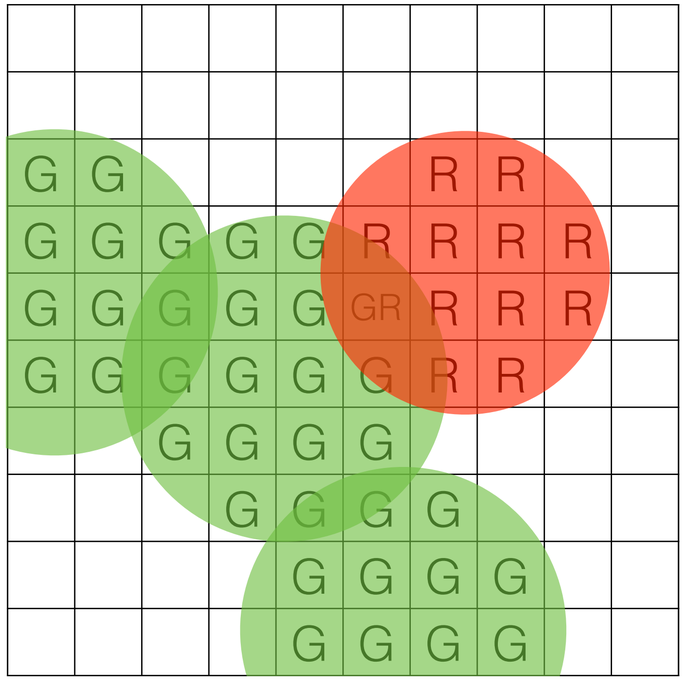
- Input representation
- Training
- Model optimization
- Visualize and Evaluation

Pose Prediction

Binding Discrimination

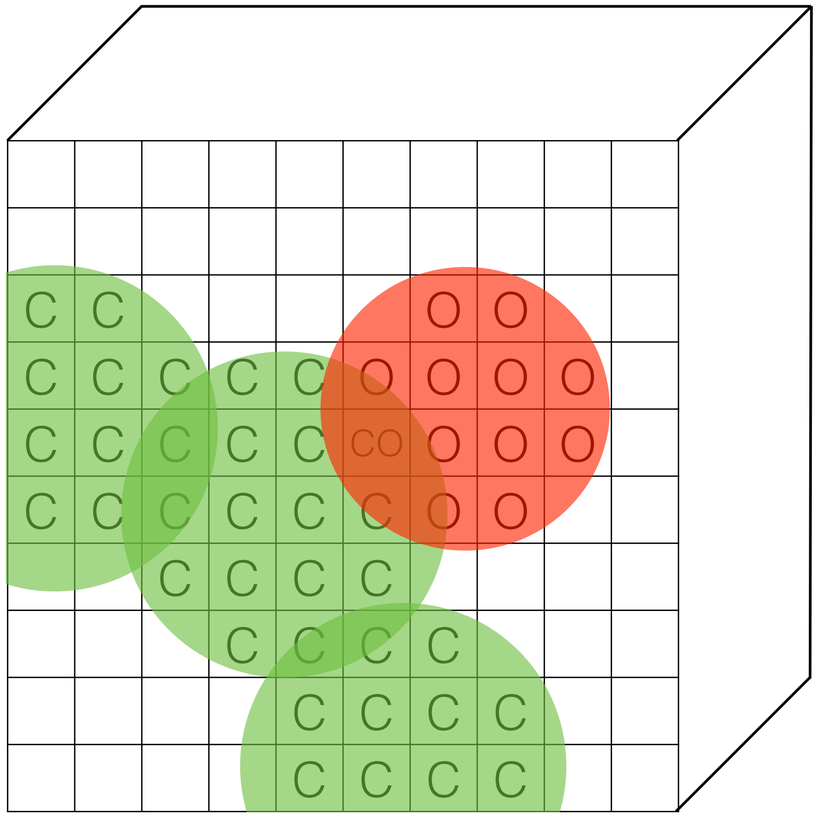
Affinity Prediction

Protein-Ligand Representation



(R,G,B) pixel

Protein-Ligand Representation



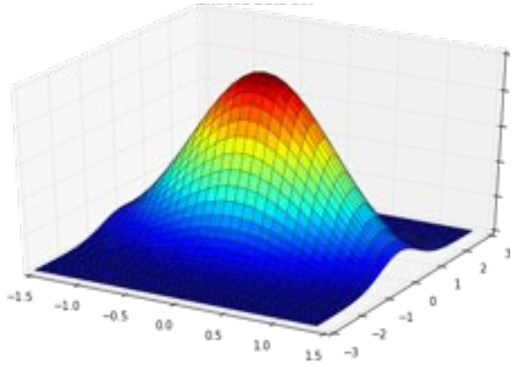
(R,G,B) pixel →

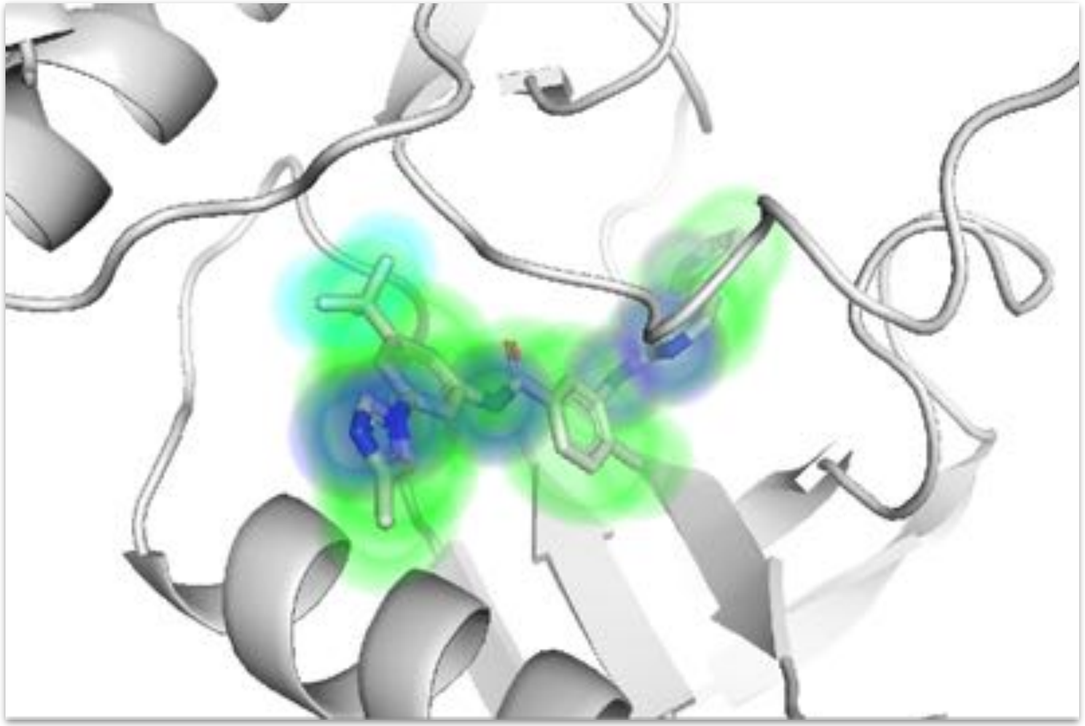
(Carbon, Nitrogen, Oxygen,…) **voxel**

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

Atom Density

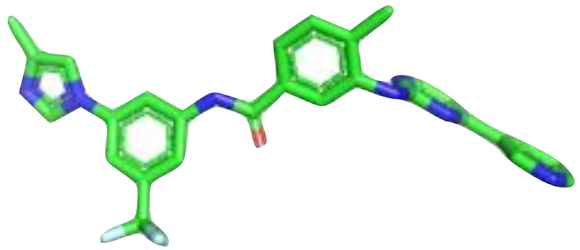
$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}} & 0 \leq d < r \\ \frac{4}{e^2 r^2} d^2 - \frac{12}{e^2 r} d + \frac{9}{e^2} & r \leq d < 1.5r \\ 0 & d \geq 1.5r \end{cases}$$





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

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)

Training Data

Binding Discrimination

D U D • E

102 targets

- 22,645 actives
- 1,407,145 decoys
- <10μM affinity
- **true poses unknown**
- **trust** docked poses

Affinity Prediction

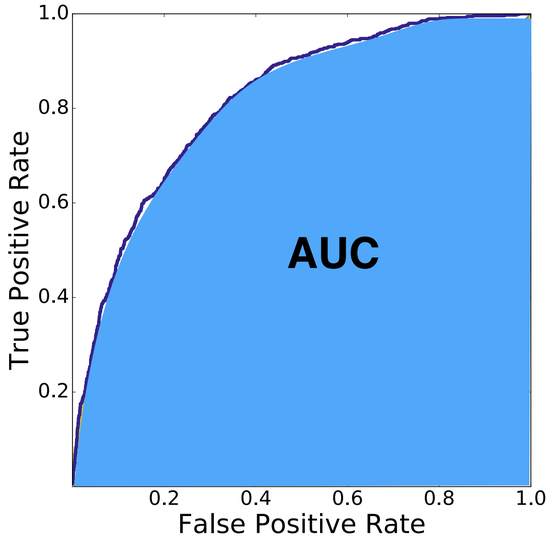
PDBbind

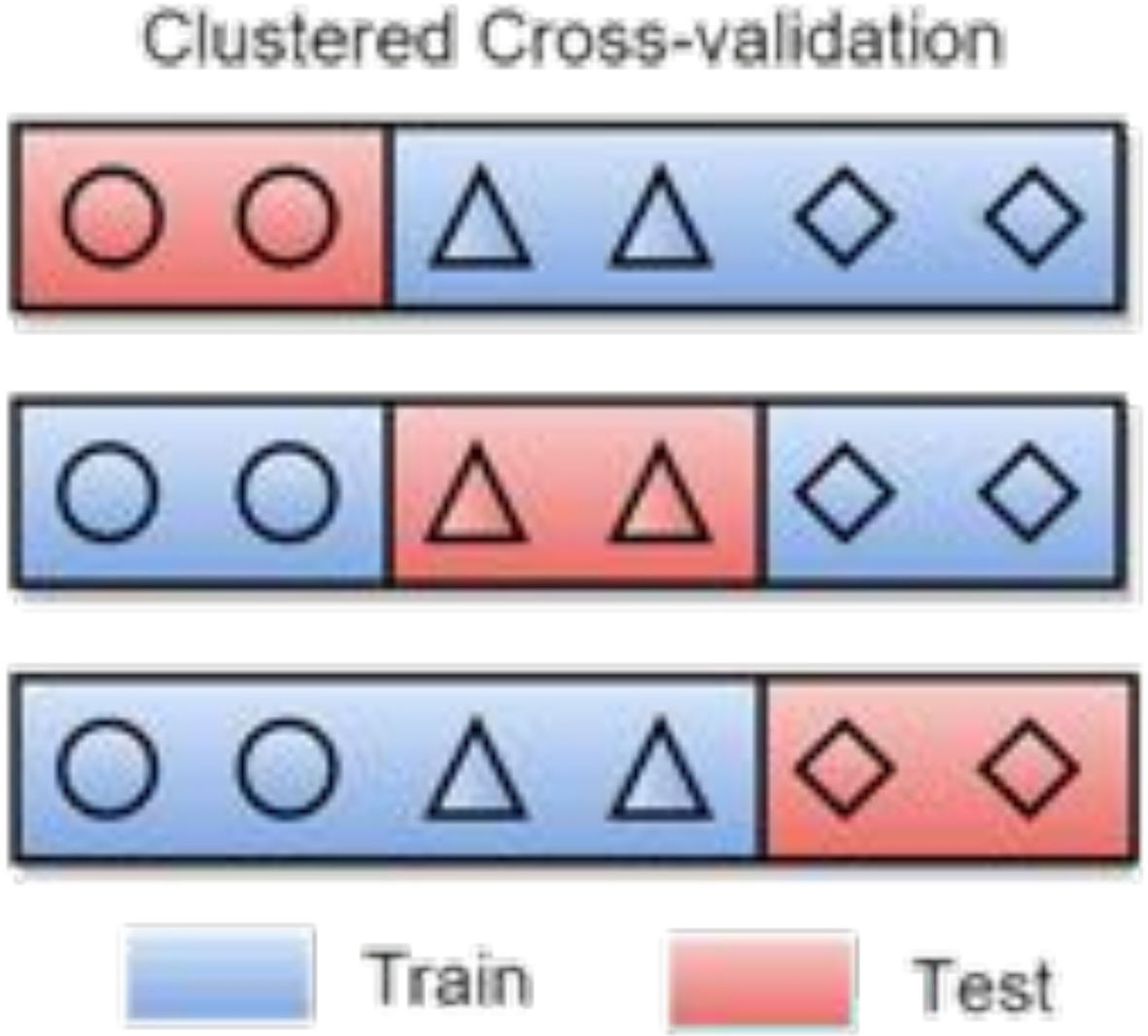
- 8,688 low RMSD poses
- assign known affinity
- **regression problem**

Model Evaluation

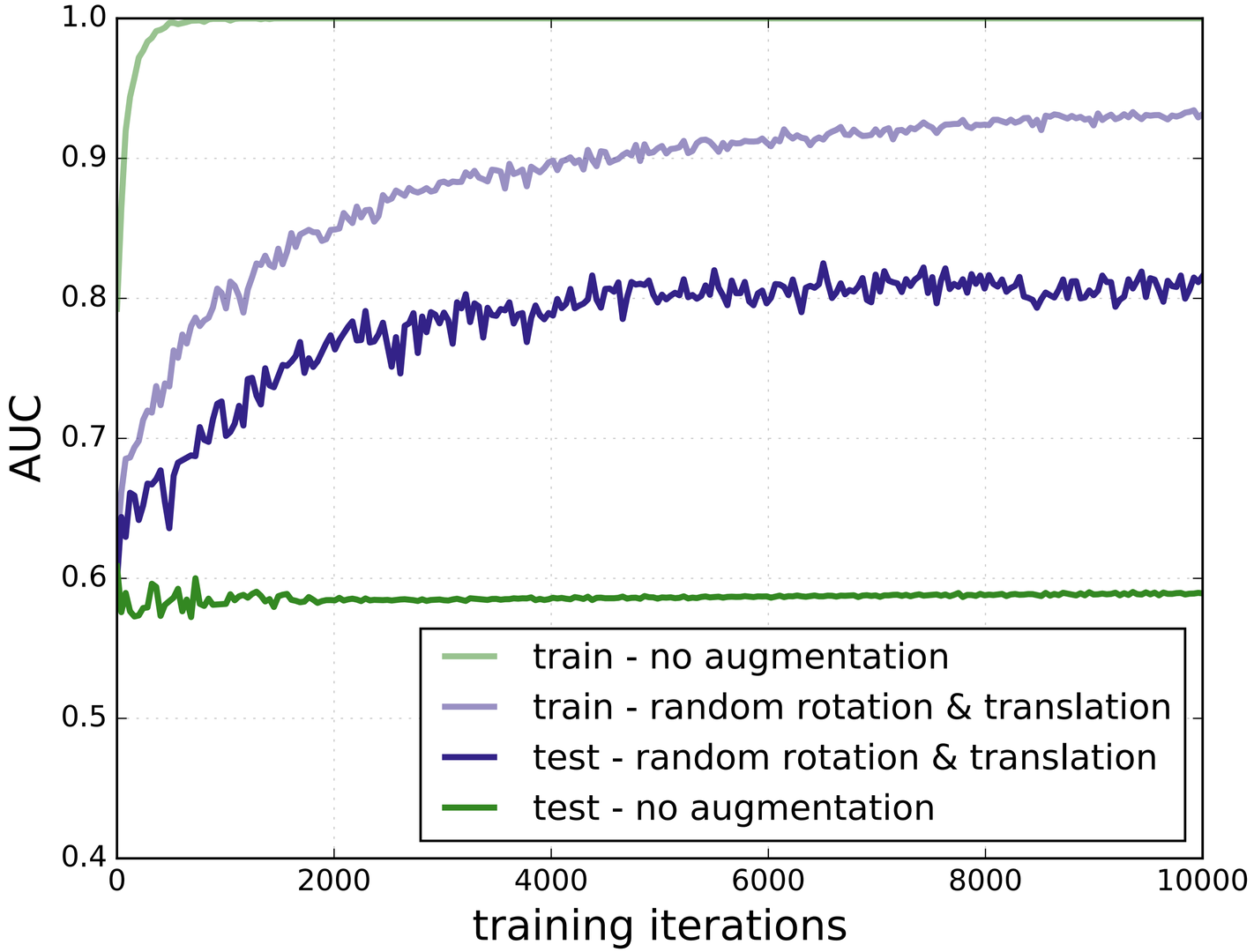
CSAR: >90% similar targets kept in same fold

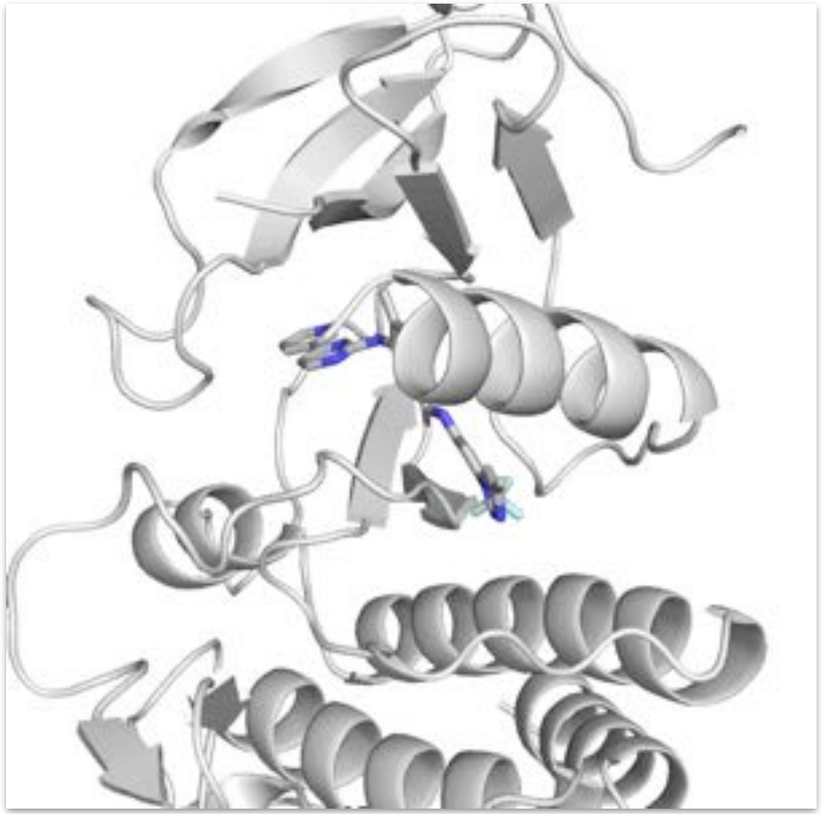
DUD-E & PDBbind: >80% similar targets kept in same fold



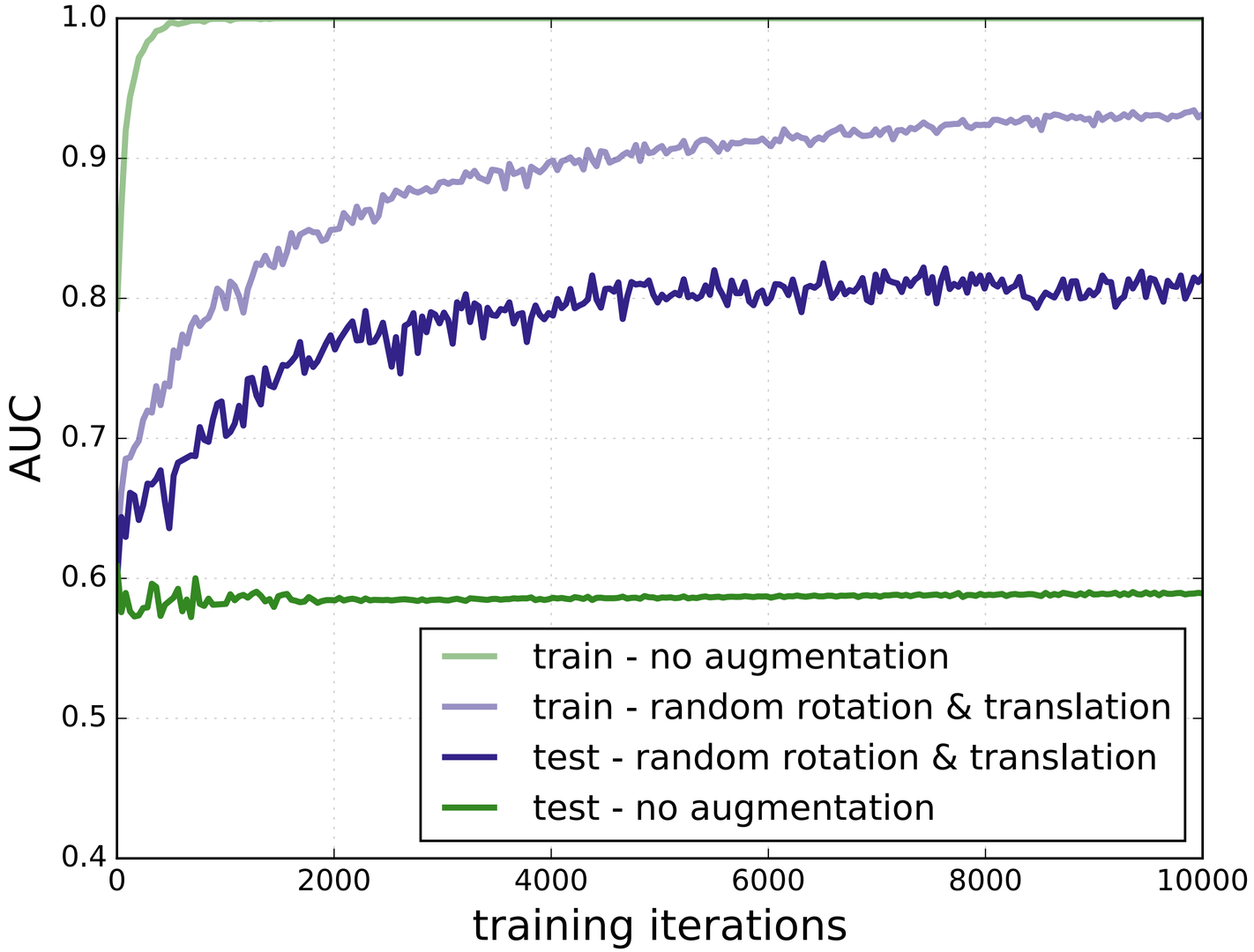


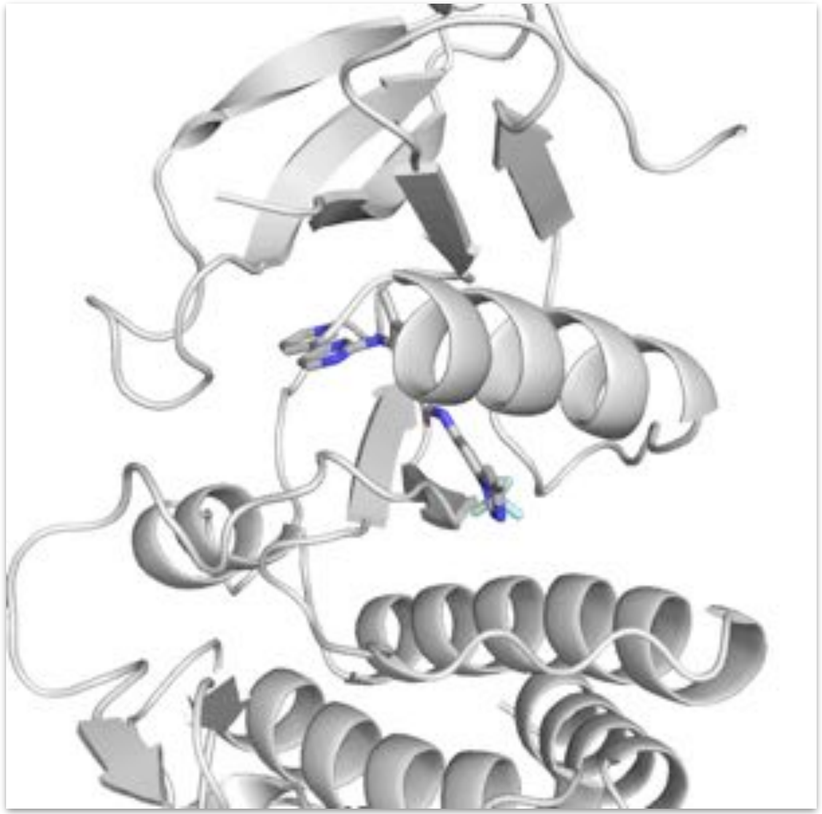
Data Augmentation

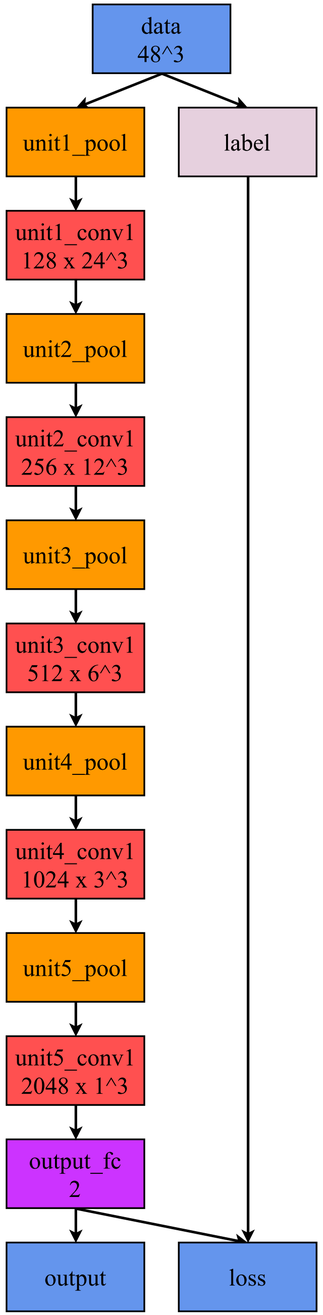




Data Augmentation







Model Optimization

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

Atom Density Type
- Boolean
- Gaussian

Radius Multiple

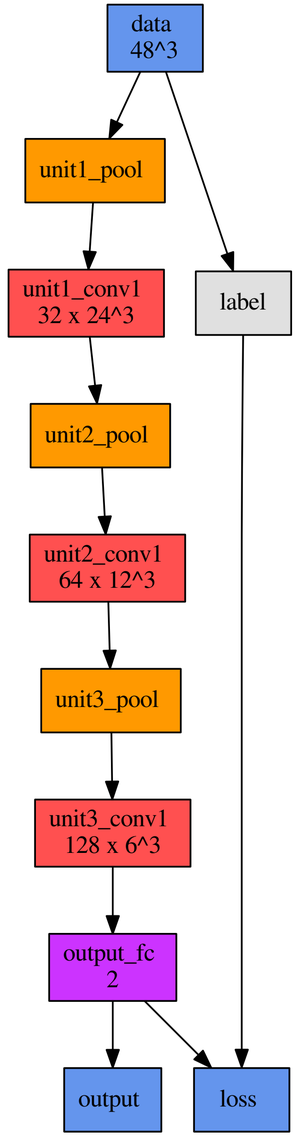
Resolution

Pooling

Depth

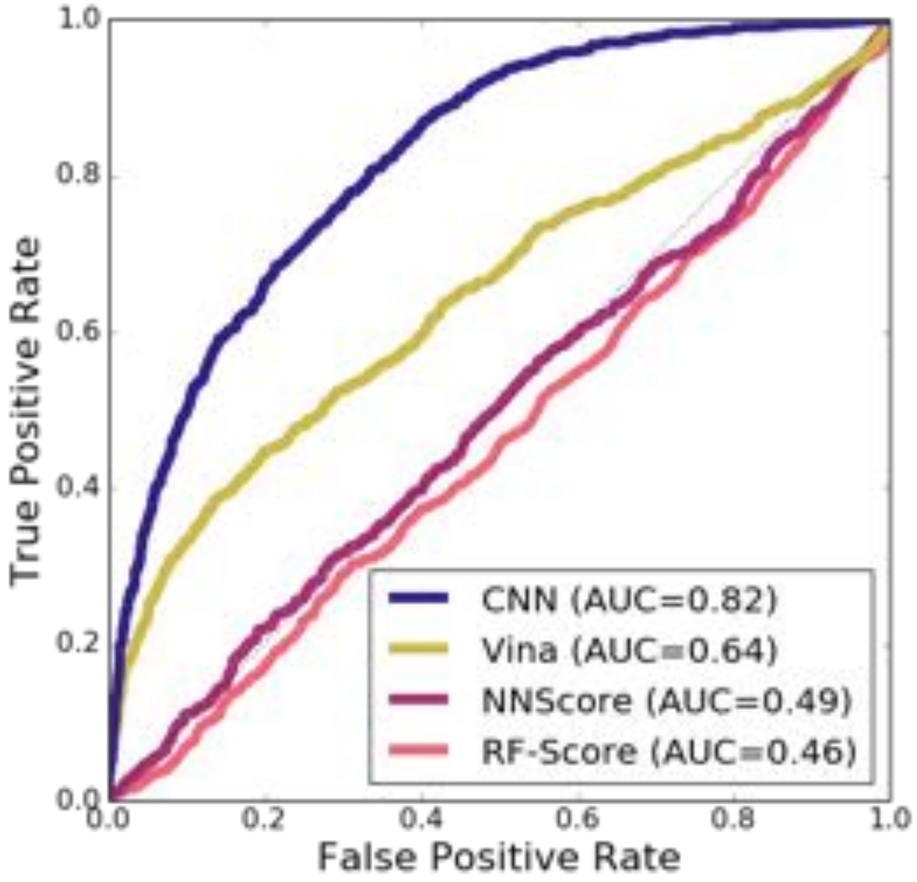
Width

Fully Connected Layers

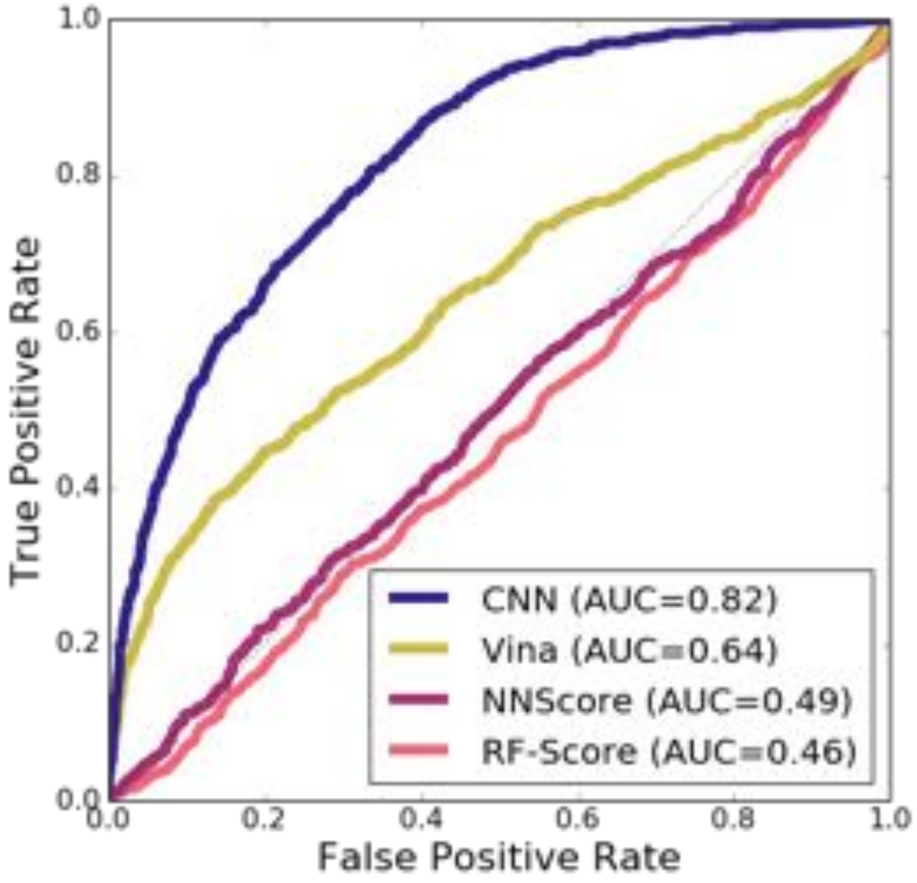


Cross-Validation Evaluation

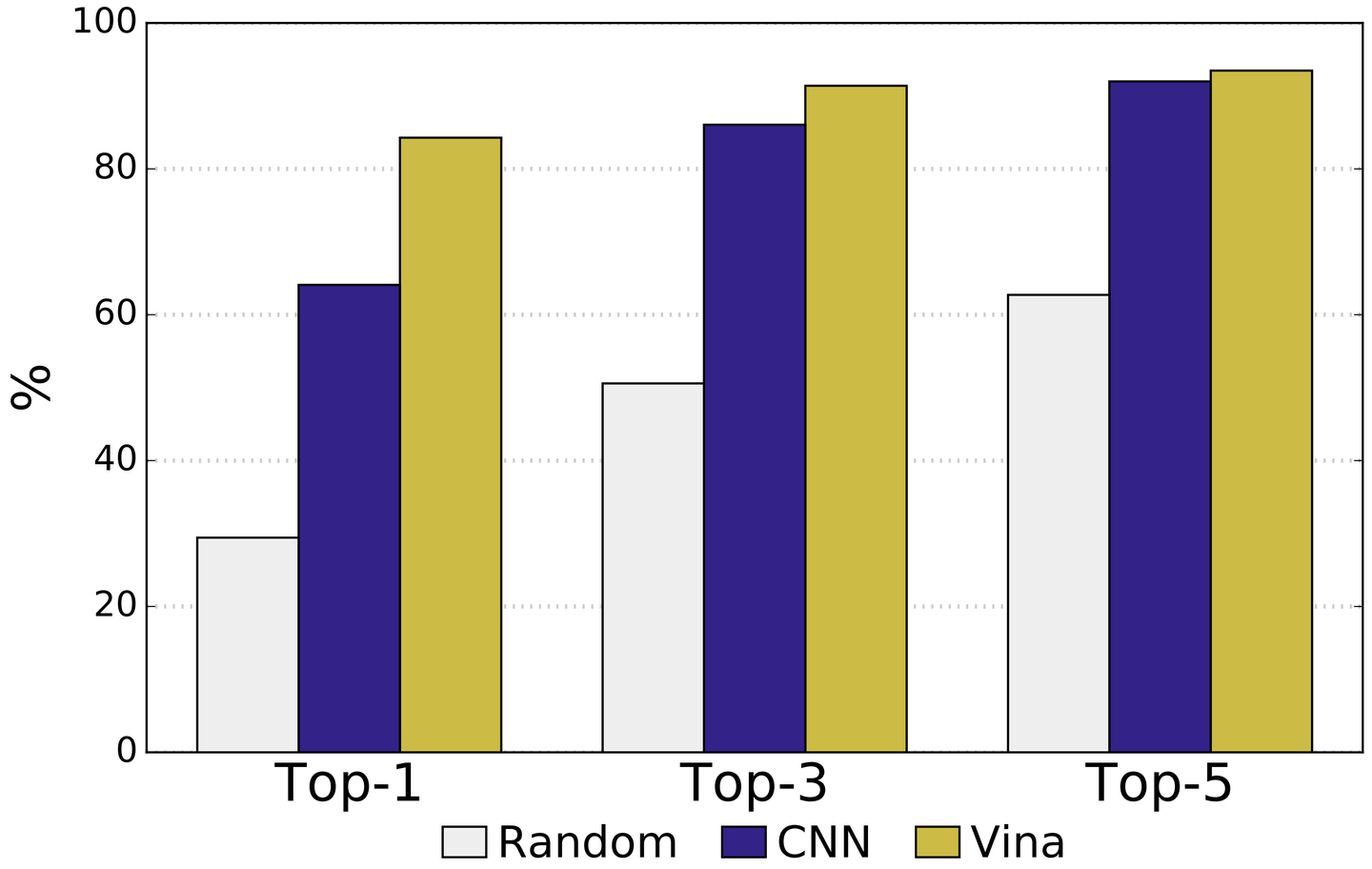
Pose Prediction (CSAR)



Pose Prediction (CSAR)

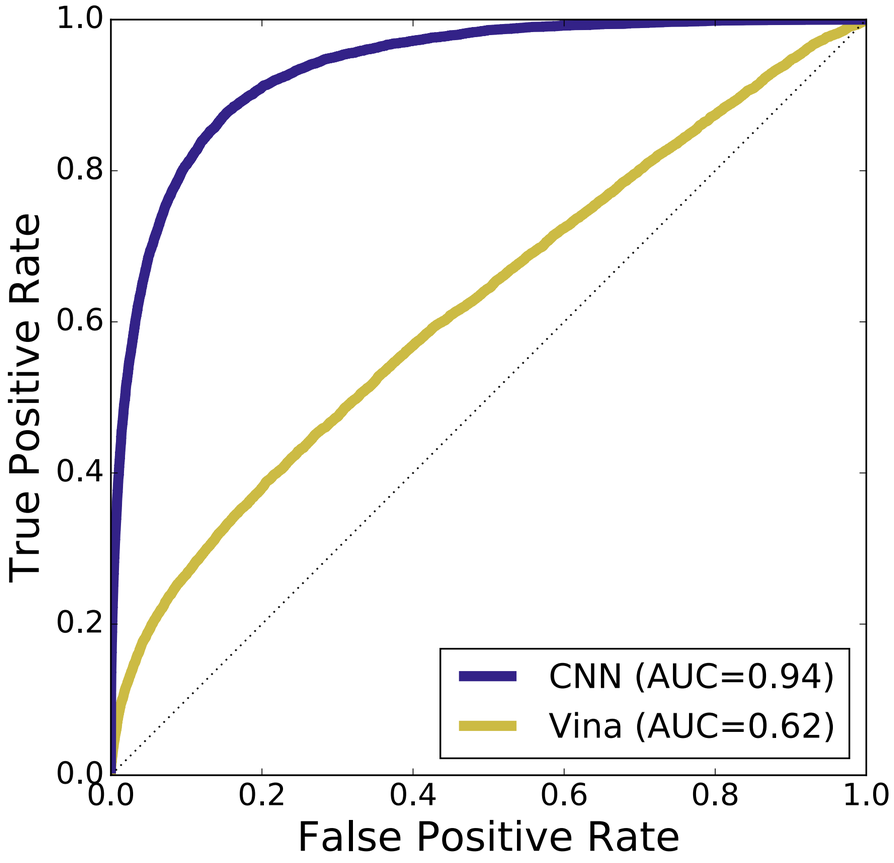


inter-target ranking

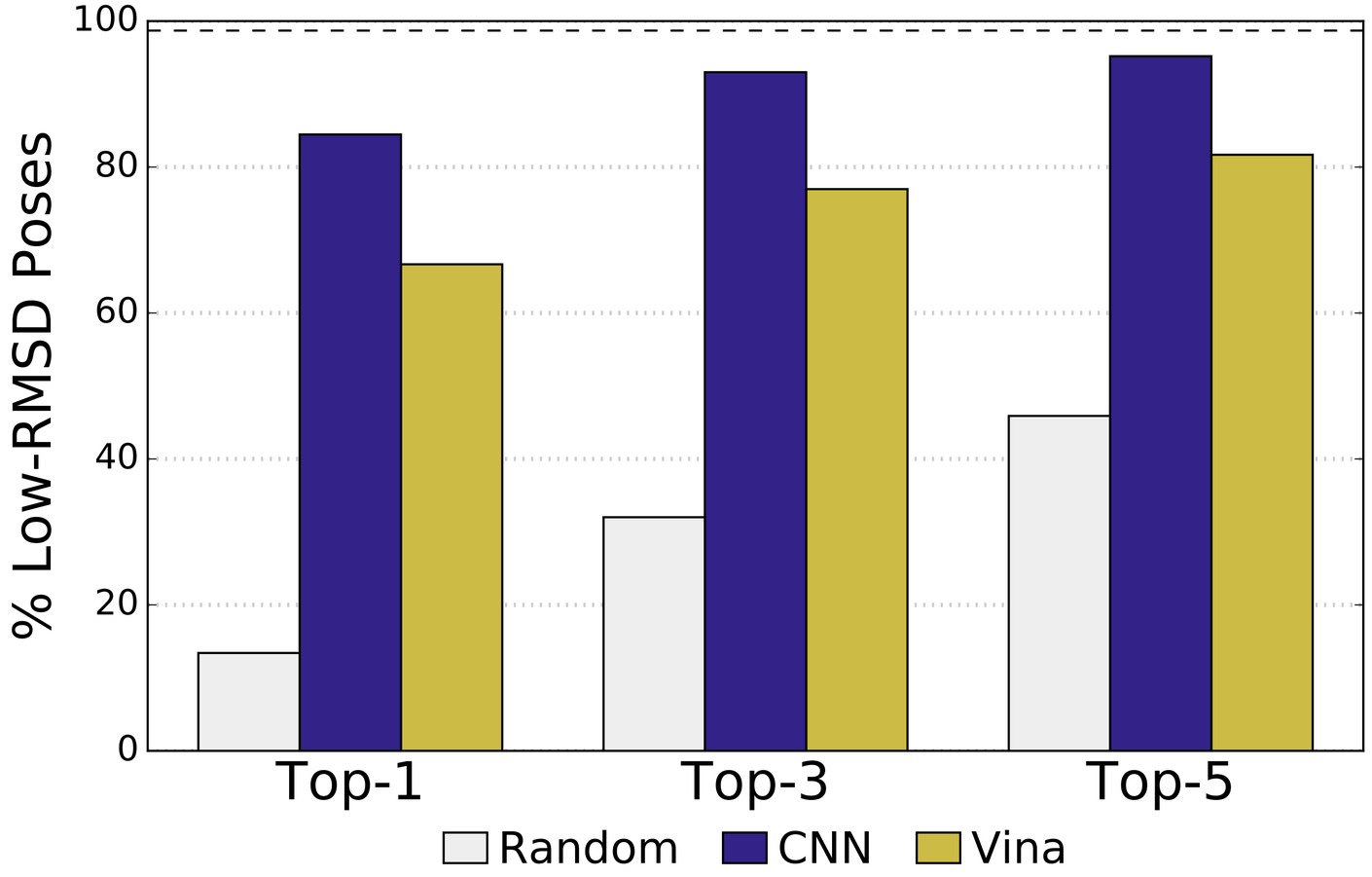


intra-target ranking

Pose Prediction (PDBbind)

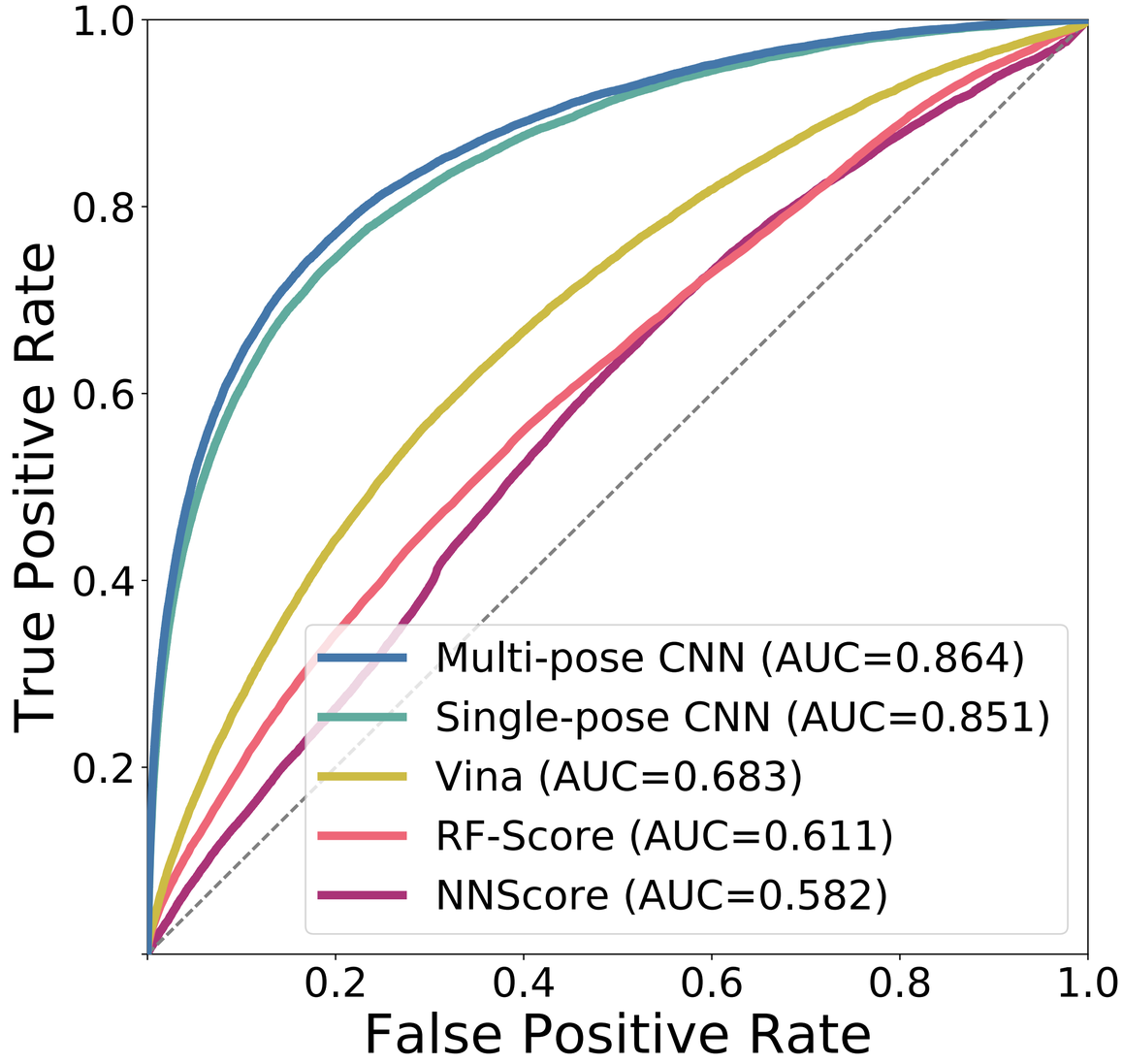


inter-target ranking



intra-target ranking

Binding Determination

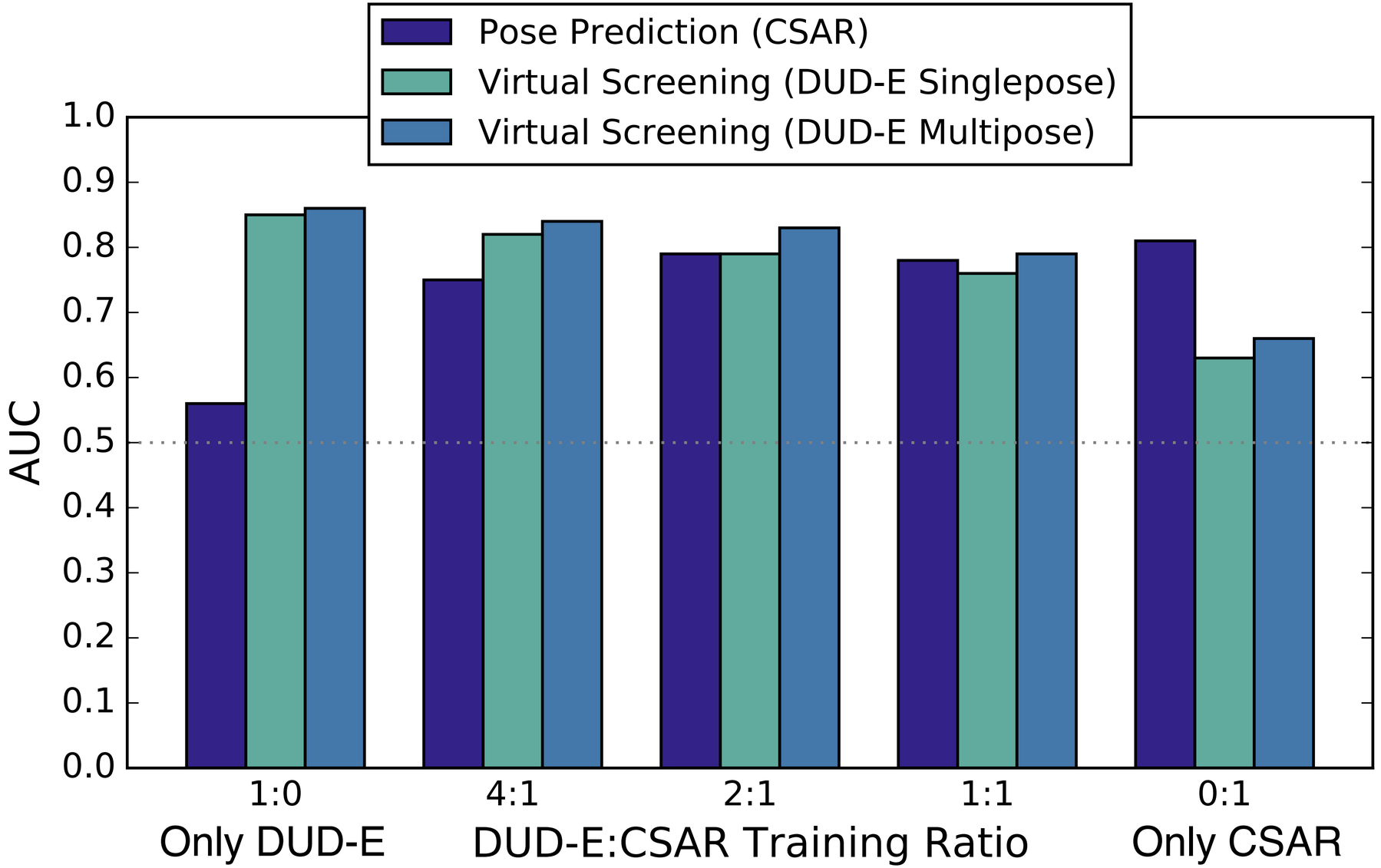


D U D • E

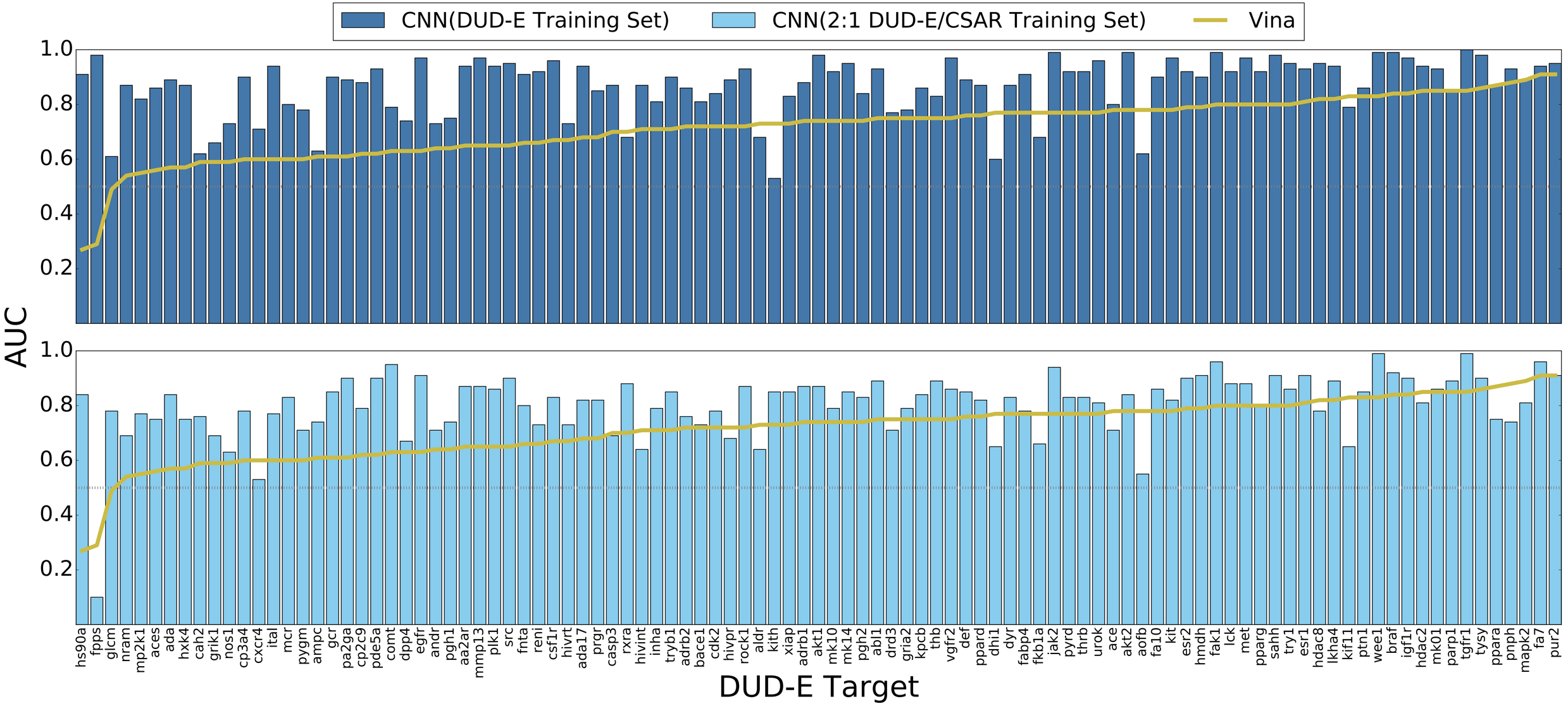
102 targets

- 22,645 actives
- 1,407,145 decoys
- <10μM affinity
- **true poses unknown**
- use top docked pose

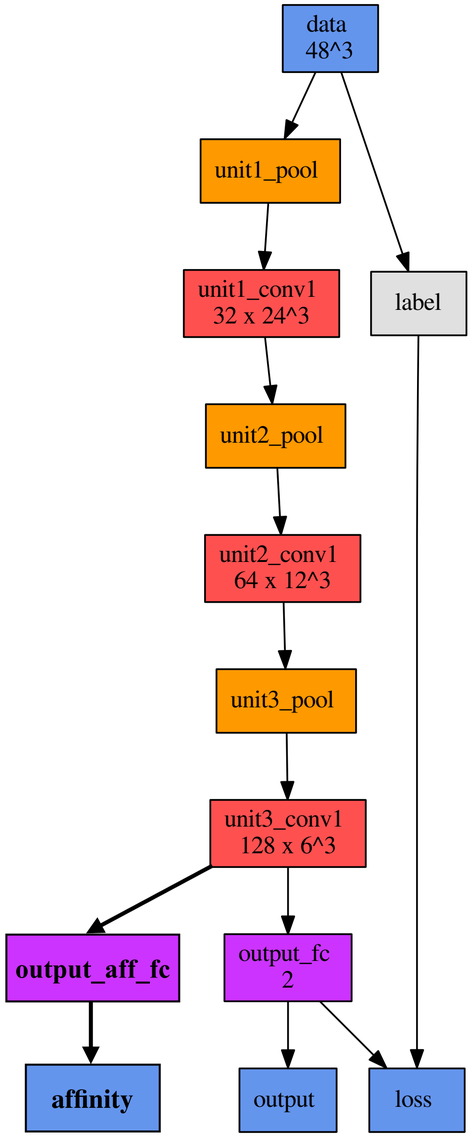
Combined Training

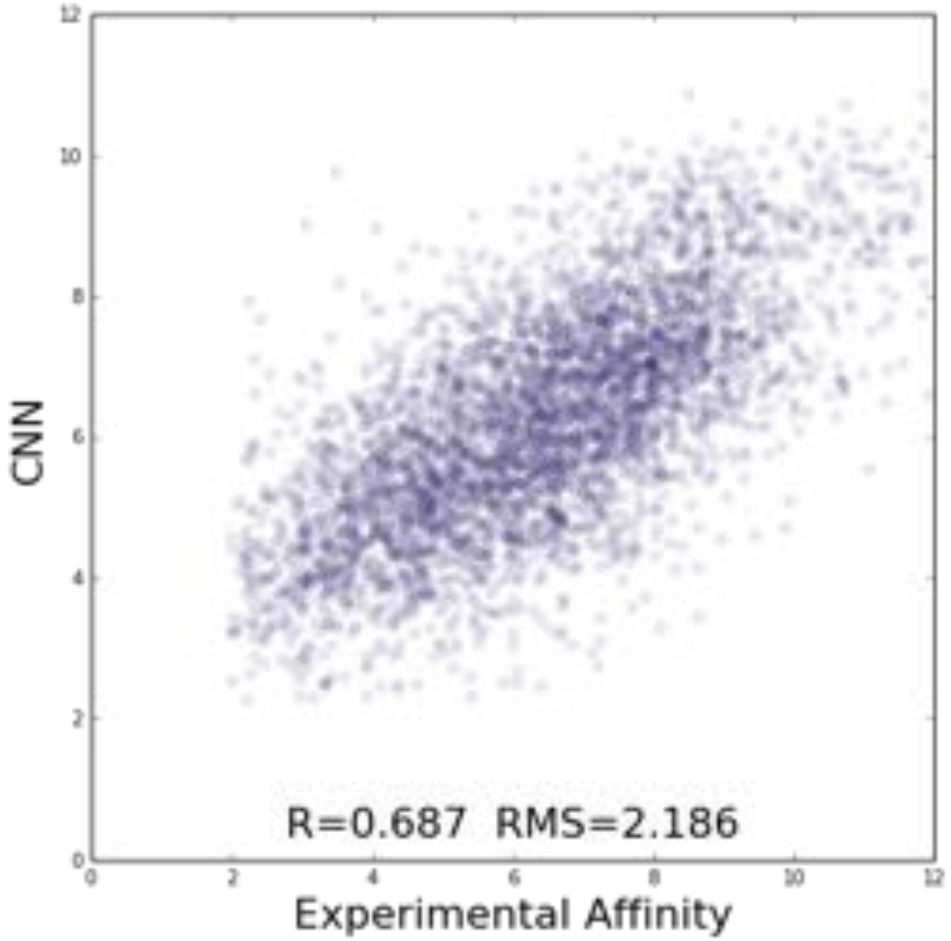


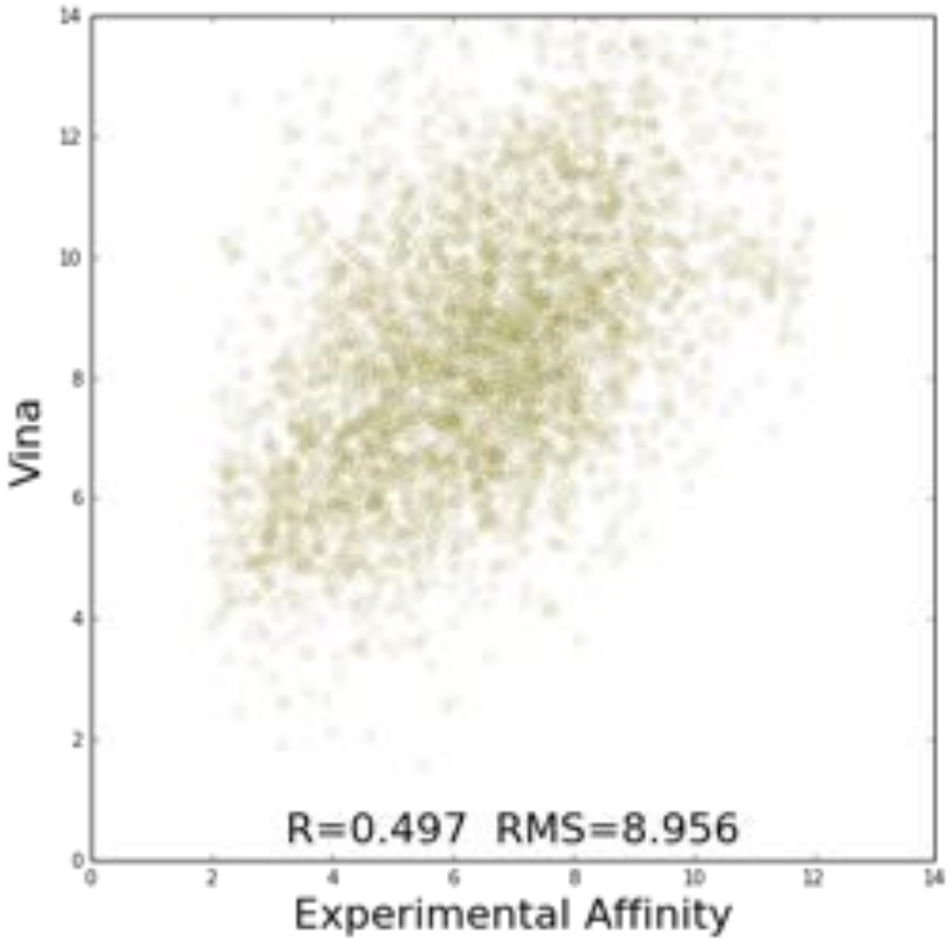
Binding Determination



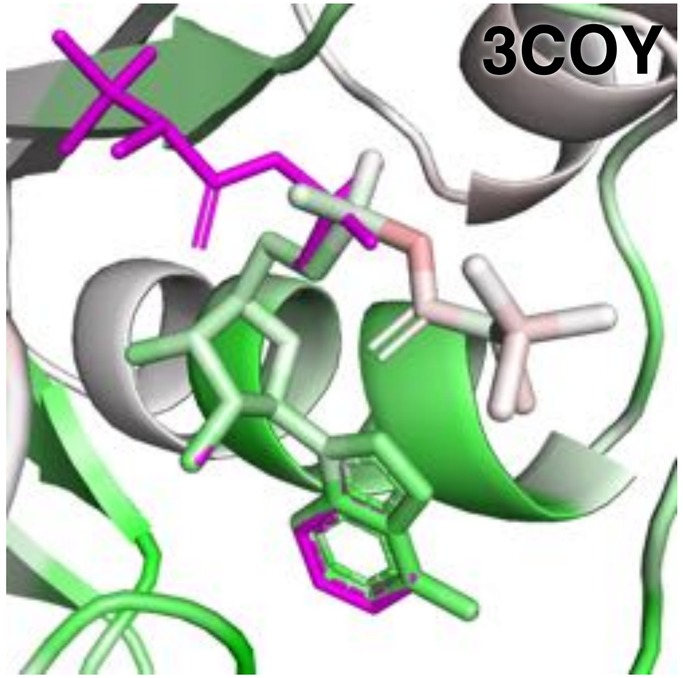
Affinity Prediction

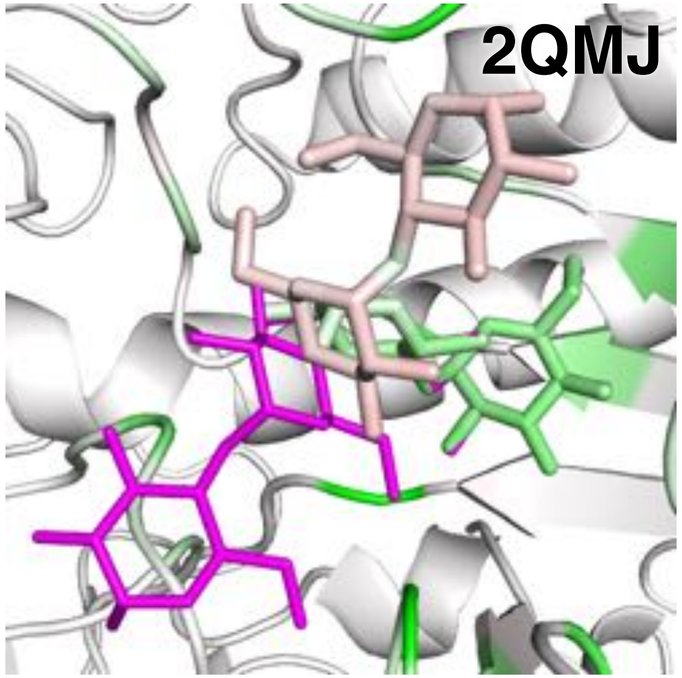


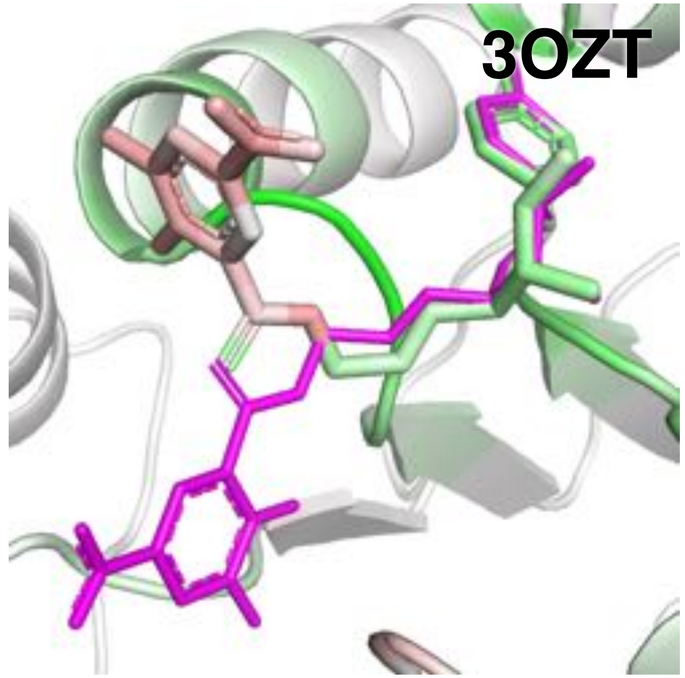




Visual Examples

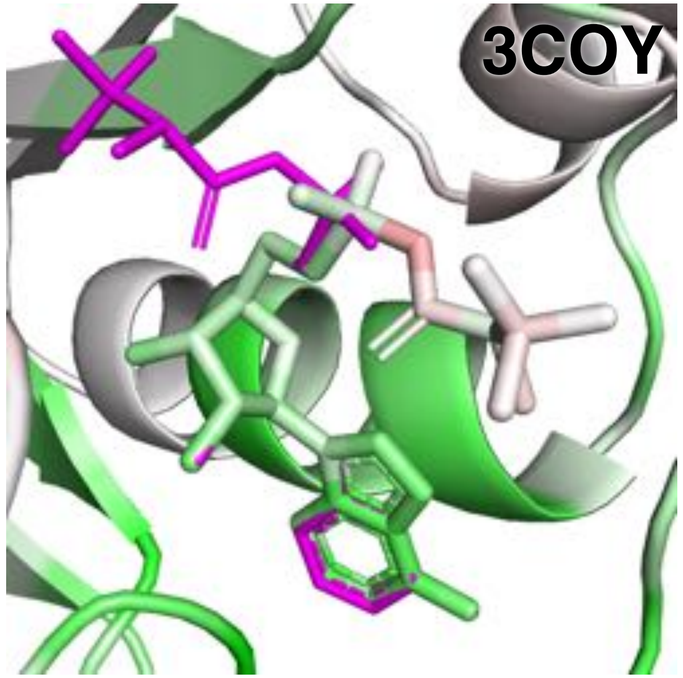


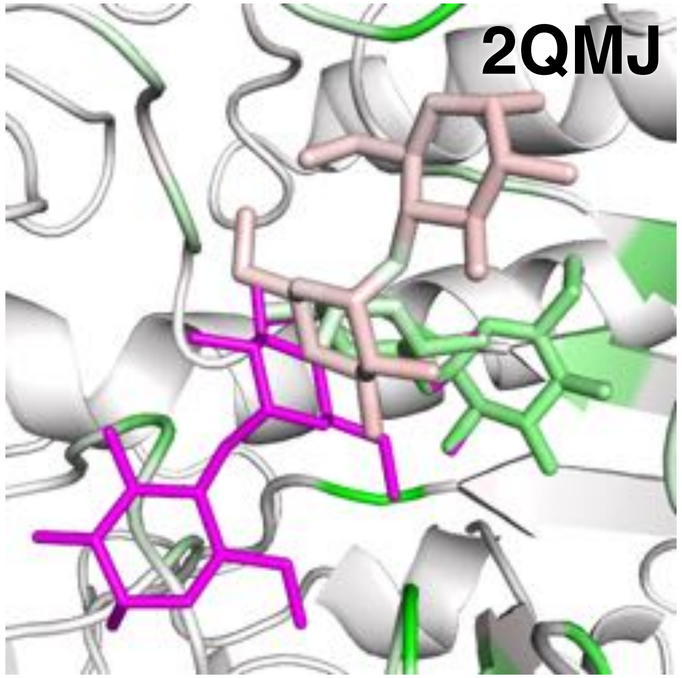


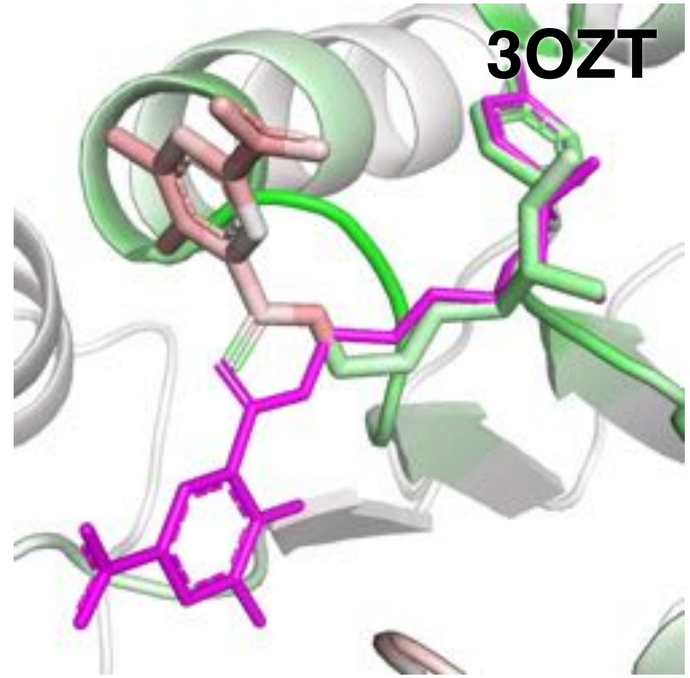


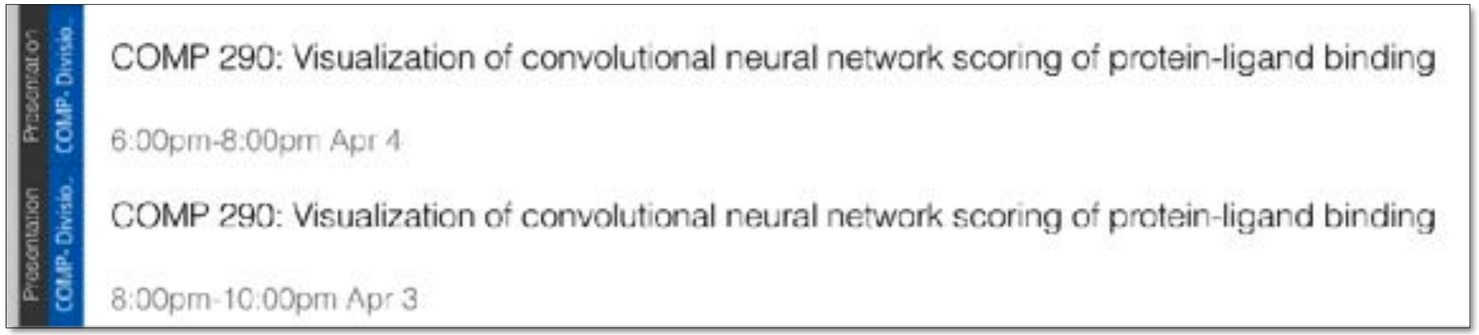
Partially Aligned Poses
Combined 2:1 Training Set

Visual Examples

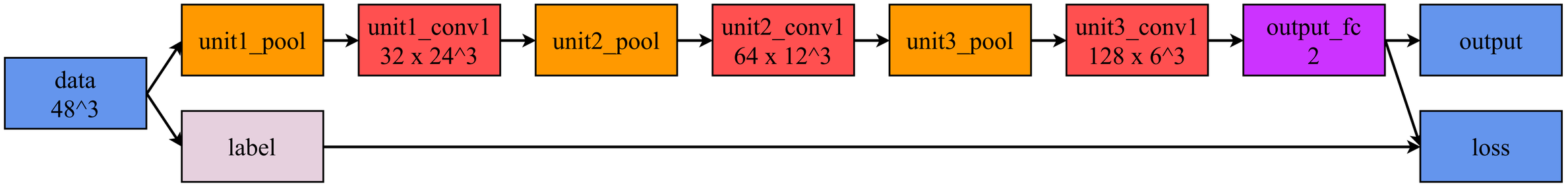




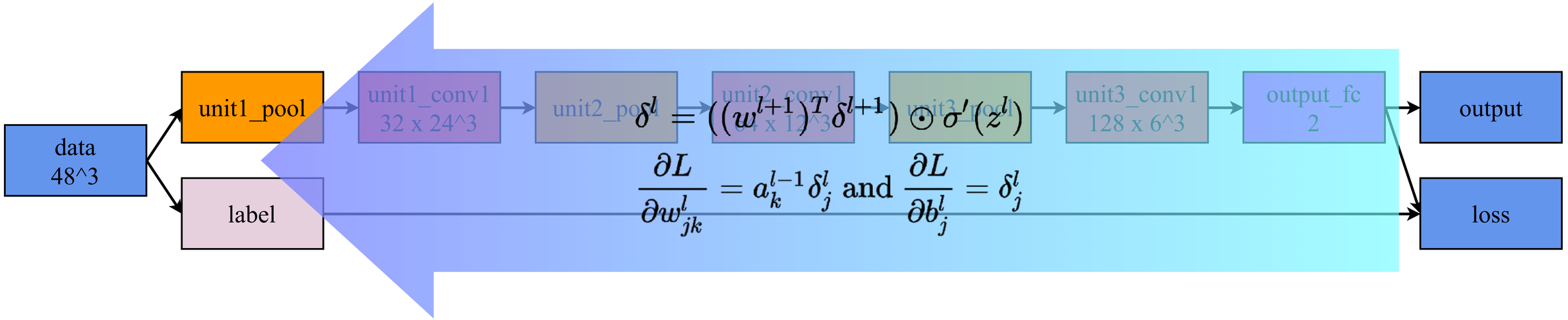




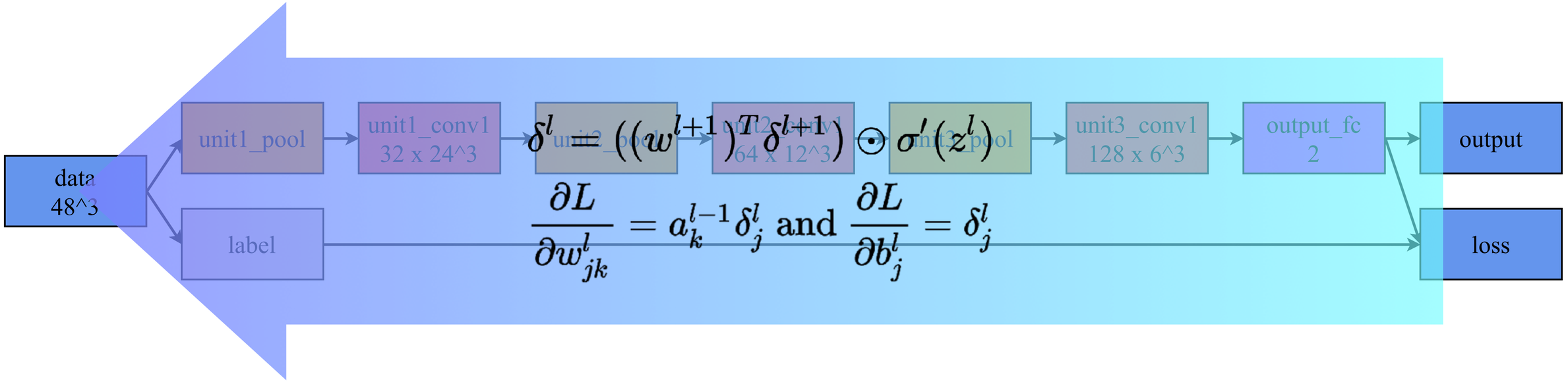
Beyond Scoring



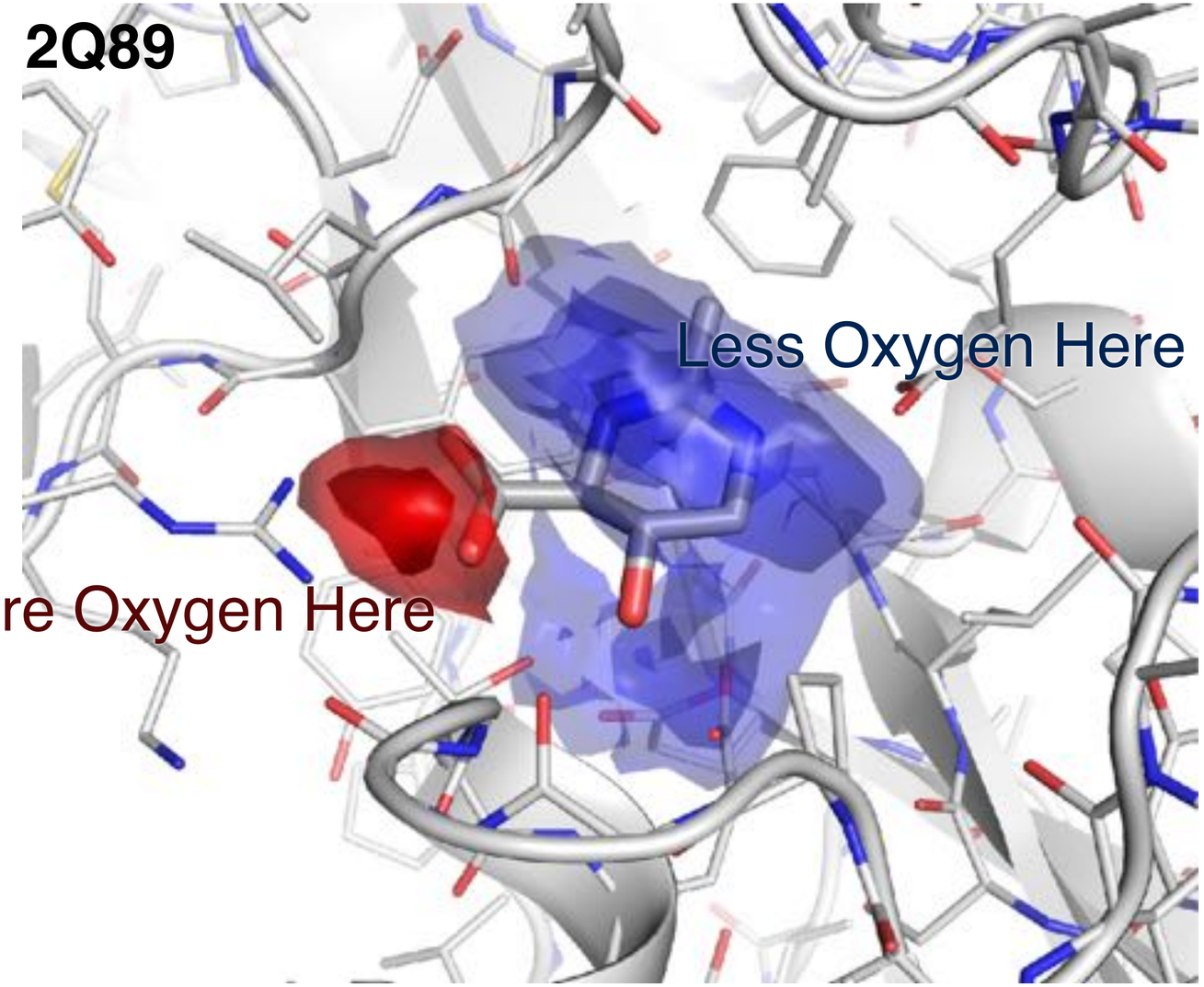
Beyond Scoring



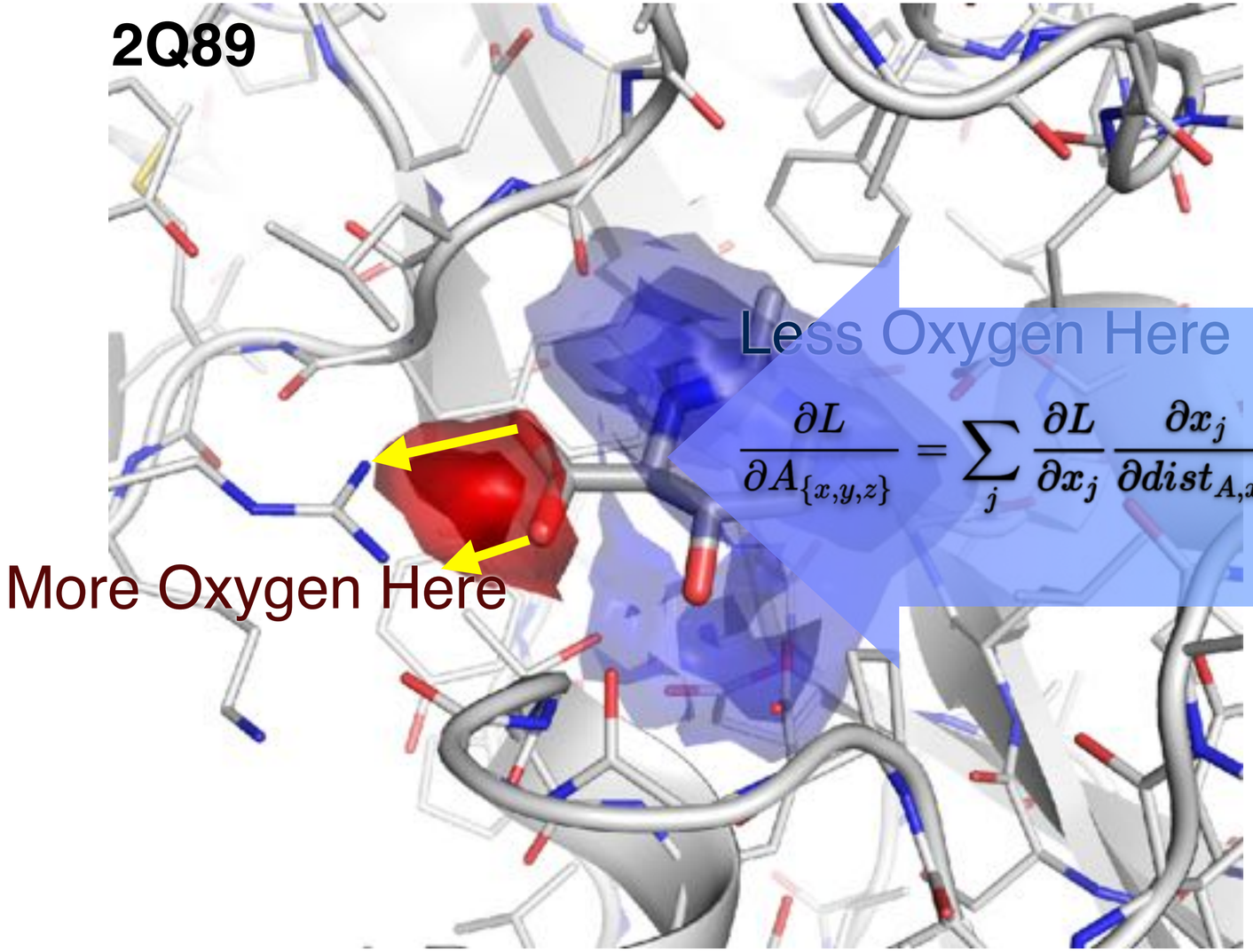
Beyond Scoring

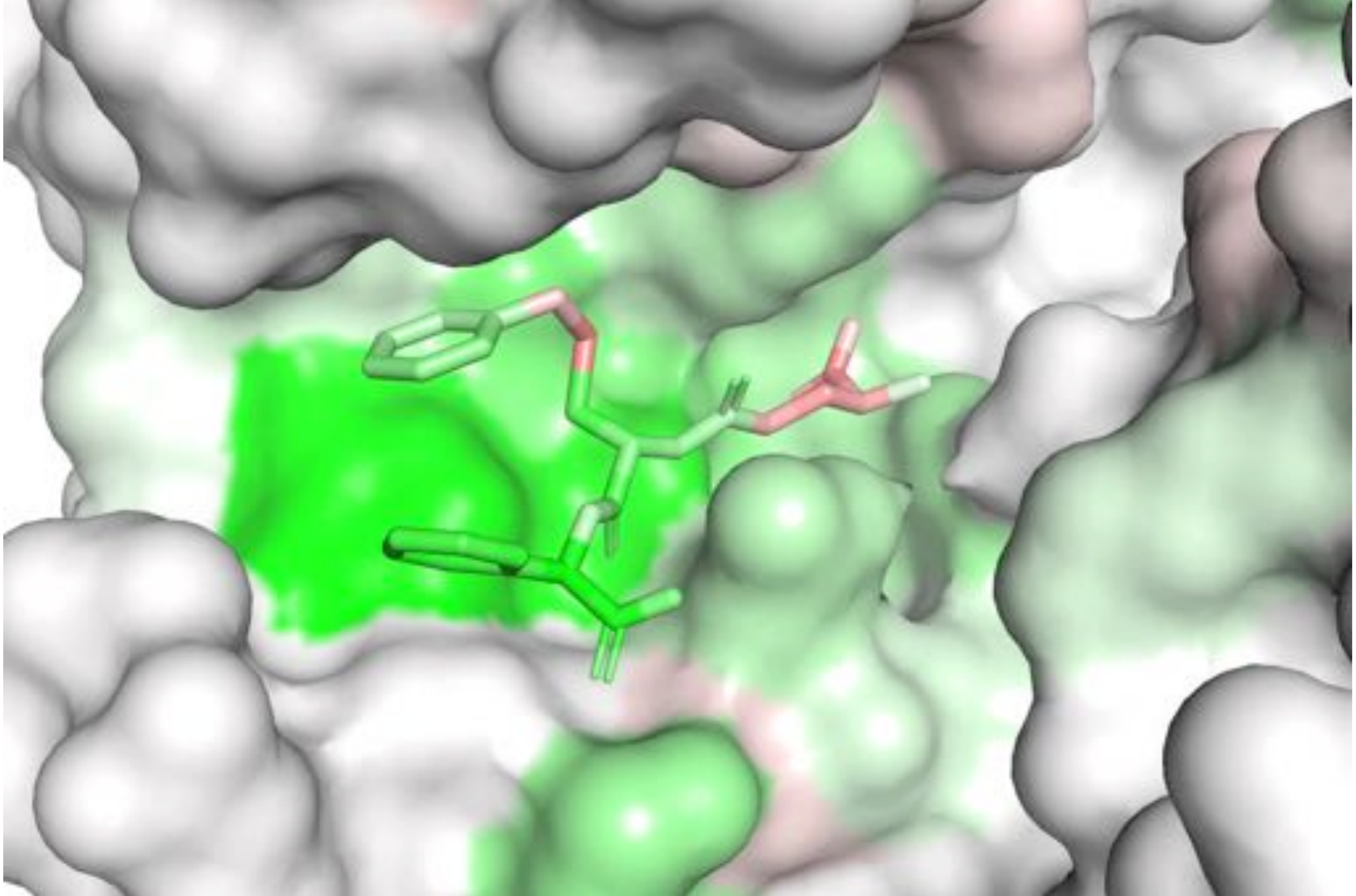


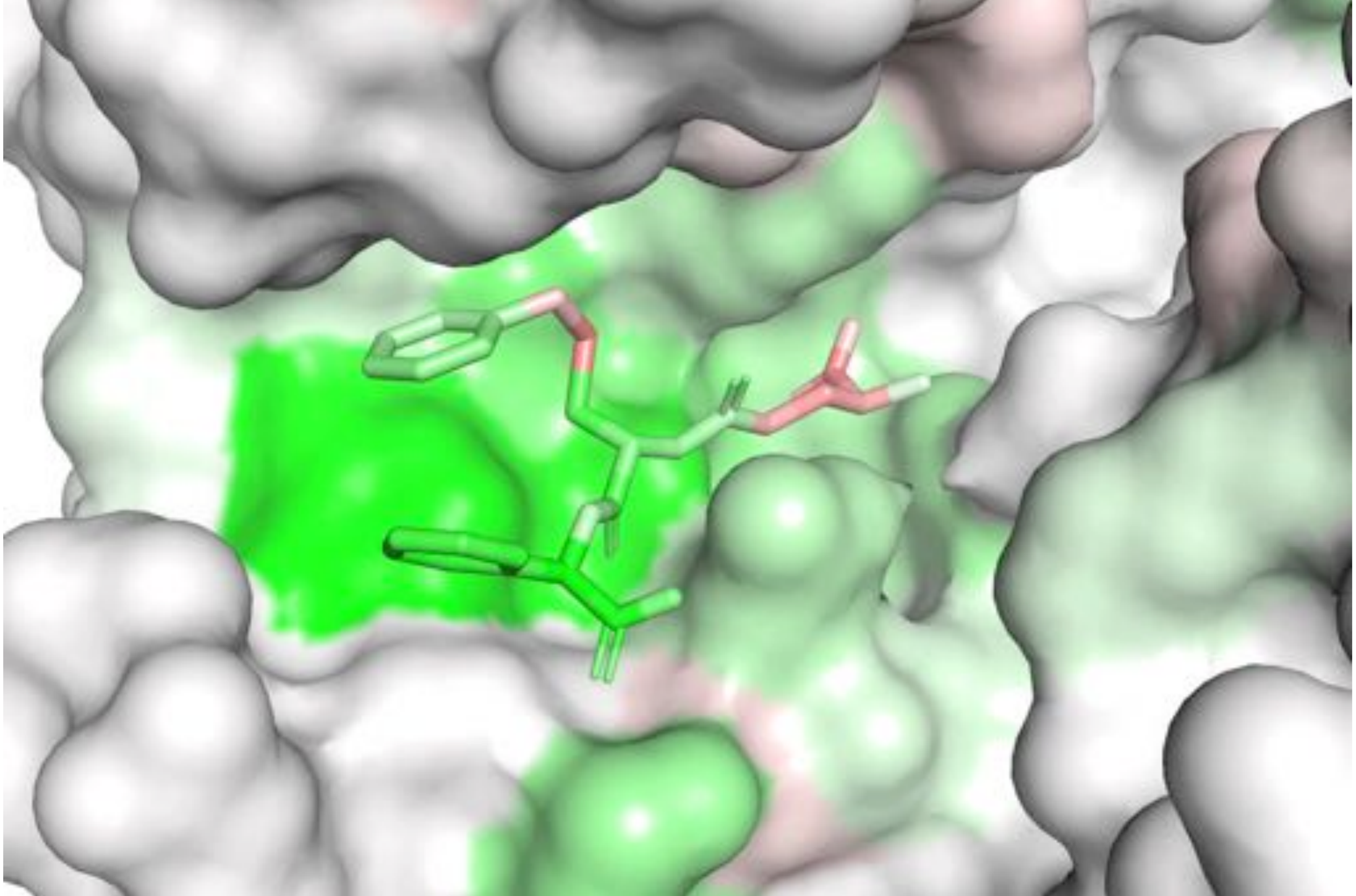
Beyond Scoring

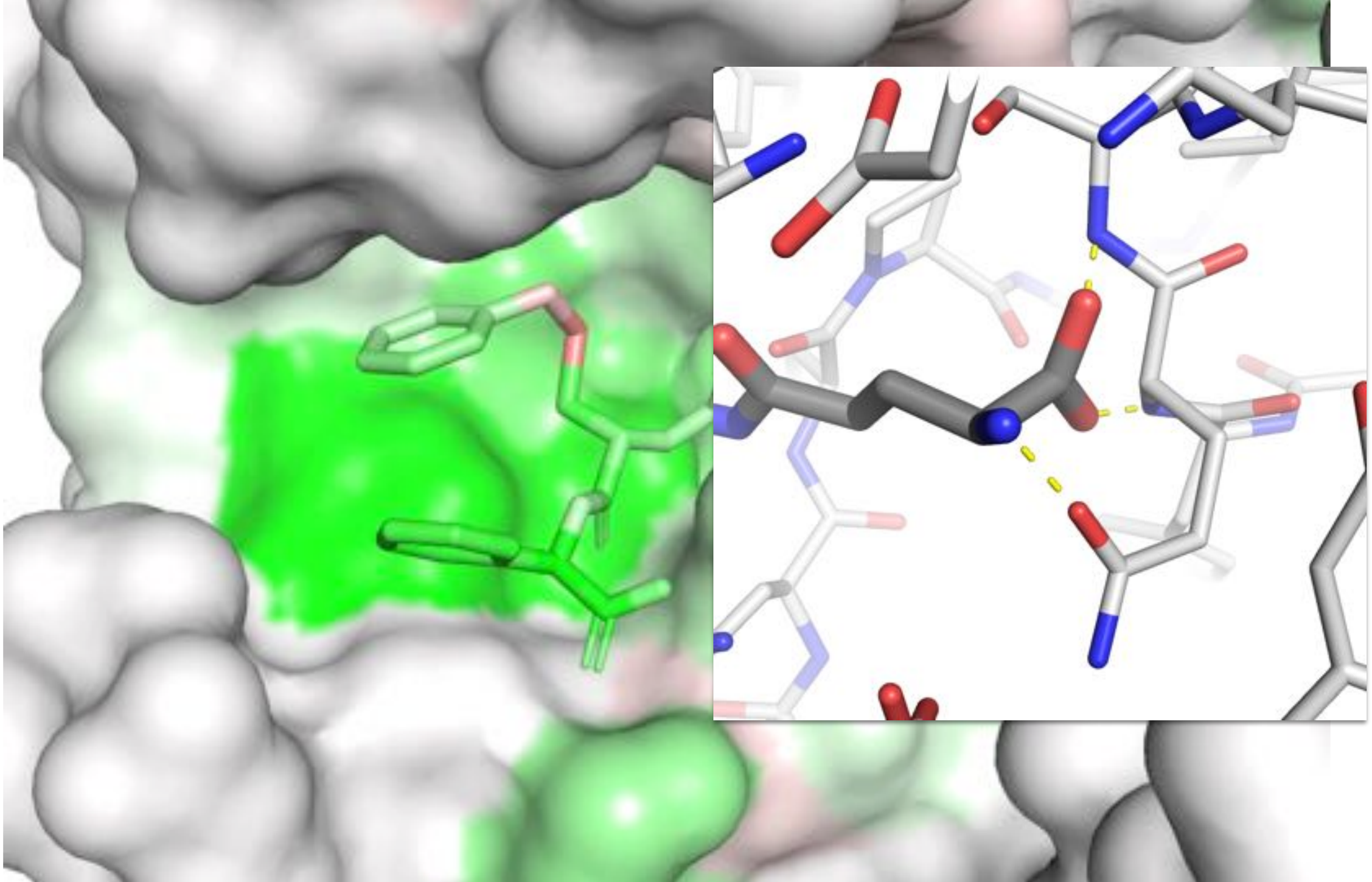


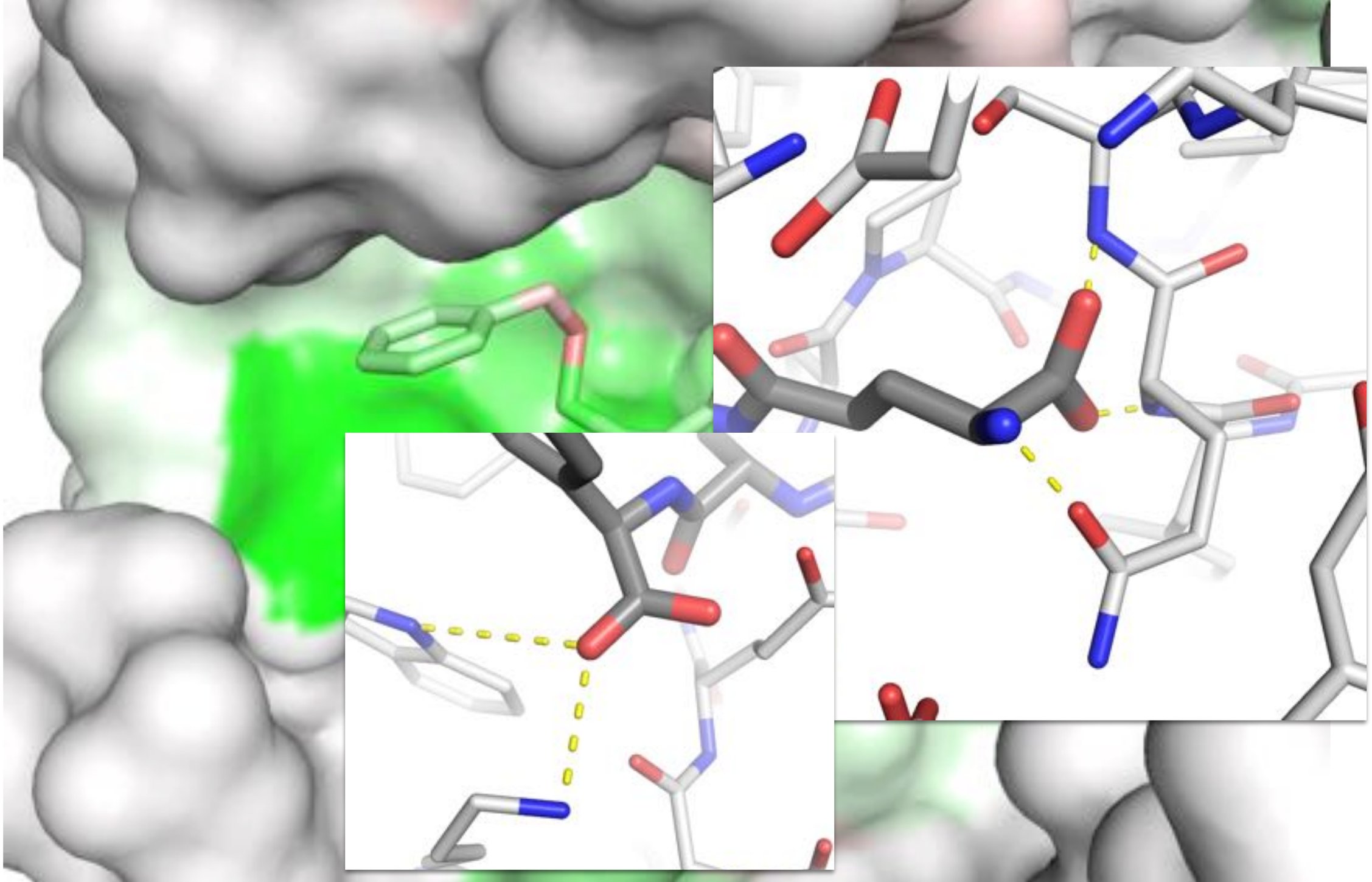
Beyond Scoring







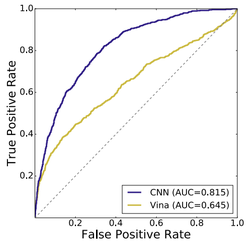


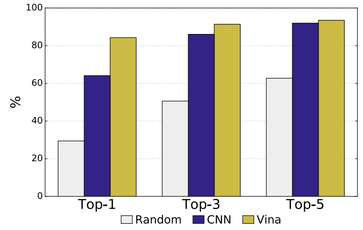


CNN Summary

Pose Prediction (Selection)

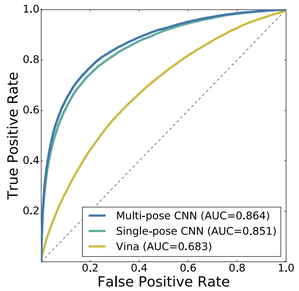
- consistently better than Vina at *inter*-target ranking
- consistently worse than Vina at *intra*-target ranking

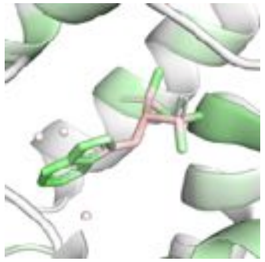




Binding Determination (Virtual Screening)

- Generally better than Vina, **but**
- the model is pose-insensitive

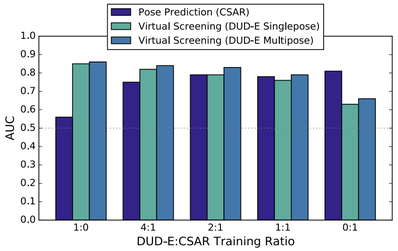




You get what you train for…

Combined Training

- Get (mostly) best of both worlds
- …including affinity prediction

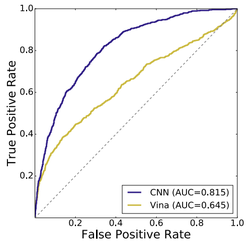


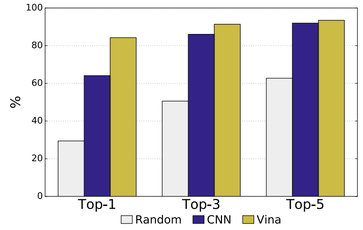
…but you can train for what you want

CNN Summary

Pose Prediction (Selection)

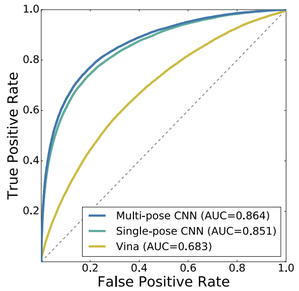
- consistently better than Vina at *inter*-target ranking
- consistently worse than Vina at *intra*-target ranking

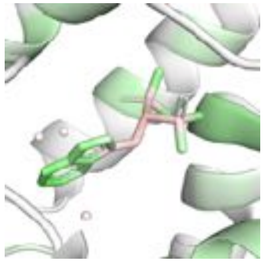




Binding Determination (Virtual Screening)

- Generally better than Vina, **but**
- the model is pose-insensitive

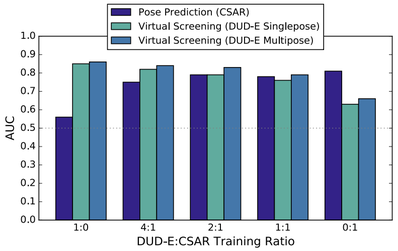




You get what you train for…

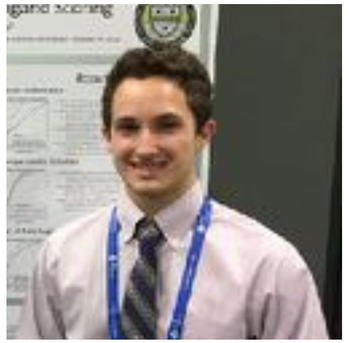
Combined Training

- Get (mostly) best of both worlds
- …including affinity prediction

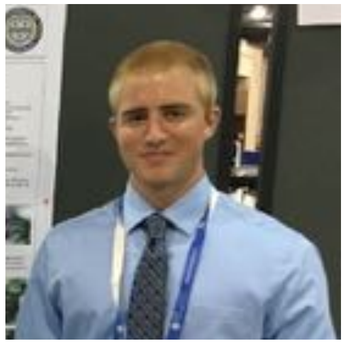


…but you can train for what you want

Acknowledgements

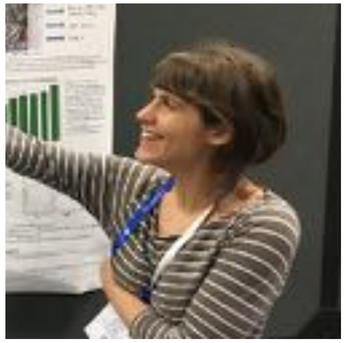


Matt Ragoza



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

Department of Computational and Systems Biology

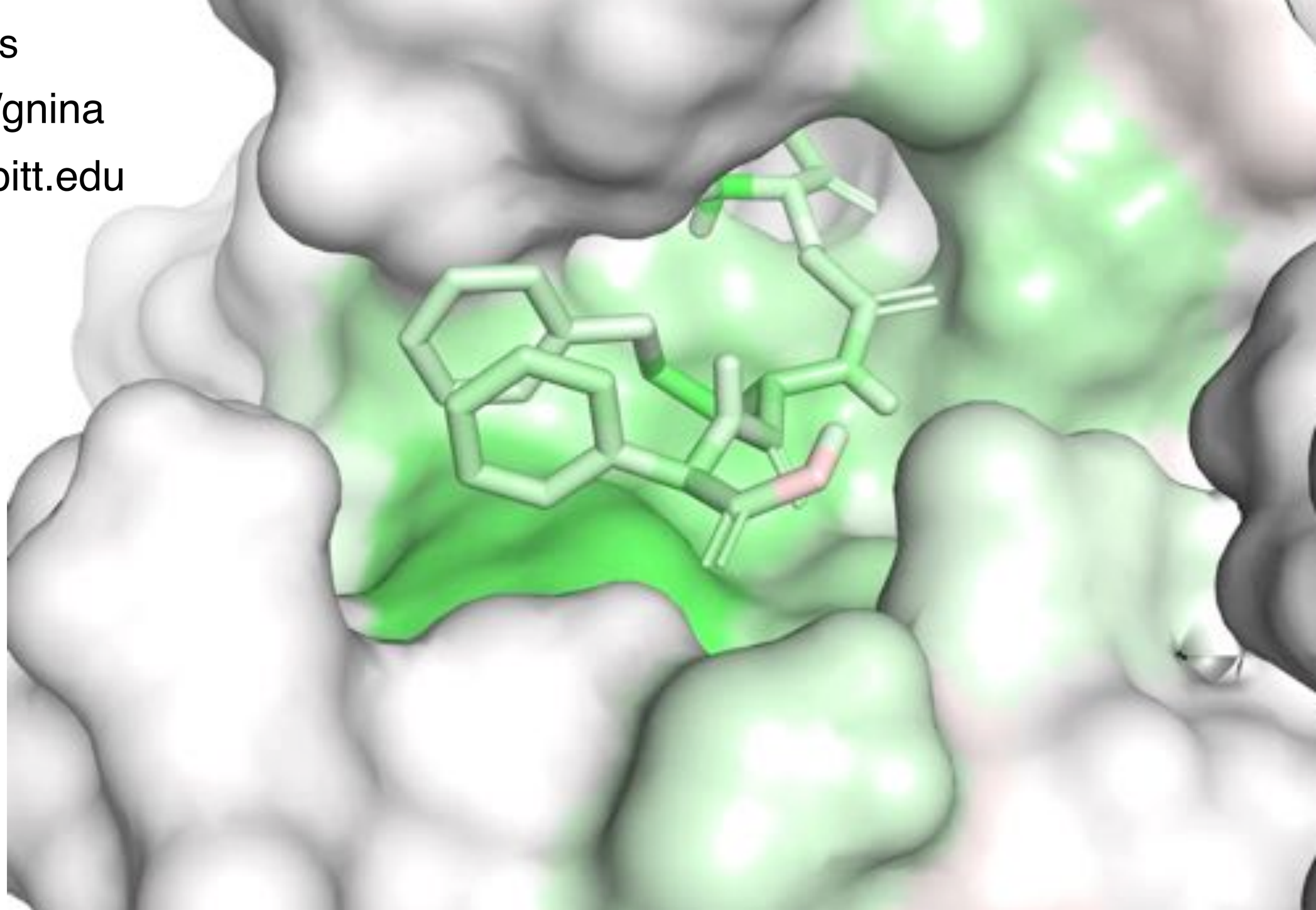
TECBIO REU

National Institute of General Medical Sciences
R01GM108340

@david_koes

github.com/gnina

http://bits.csb.pitt.edu



@david_koes

github.com/gnina

http://bits.csb.pitt.edu