Computational Methods for Biological Modeling and Simulation

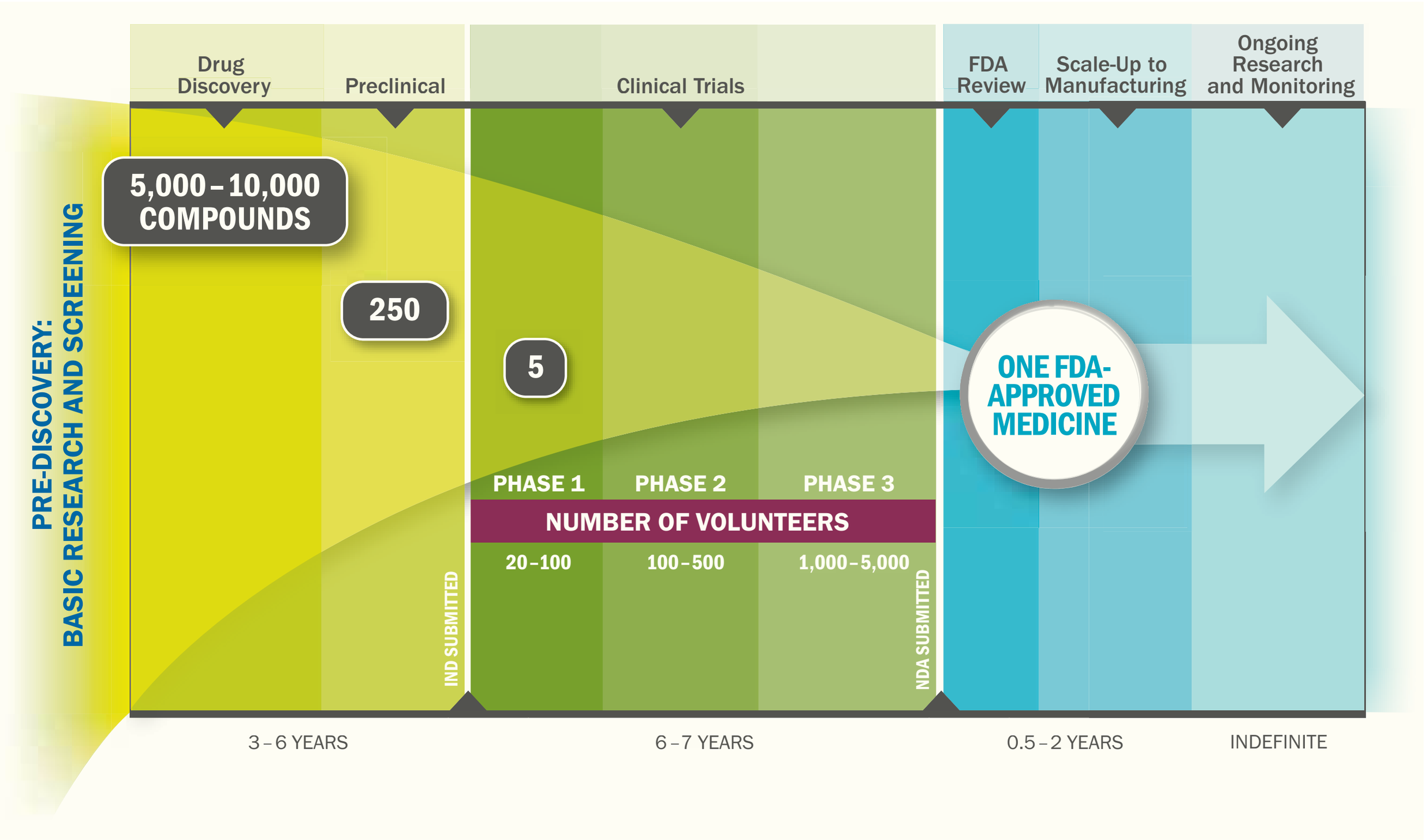
Guest Lecture

Computational Drug Discovery

David Koes

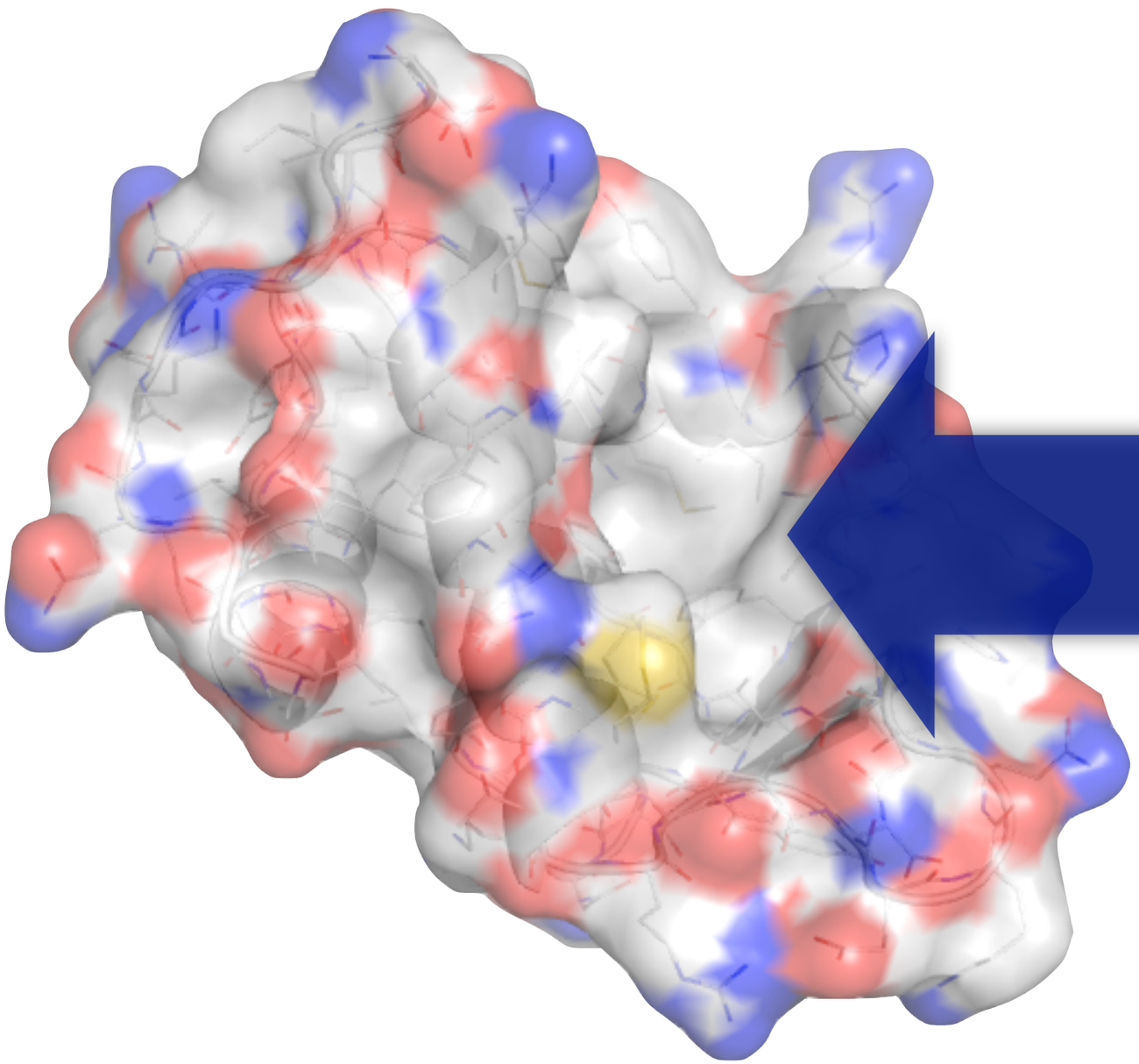
10/5/2017

Drug Development



http://phrma.org/sites/default/files/pdf/PhRMA%20Profile%202013.pdf

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?



?

Approaches

Ligand Based

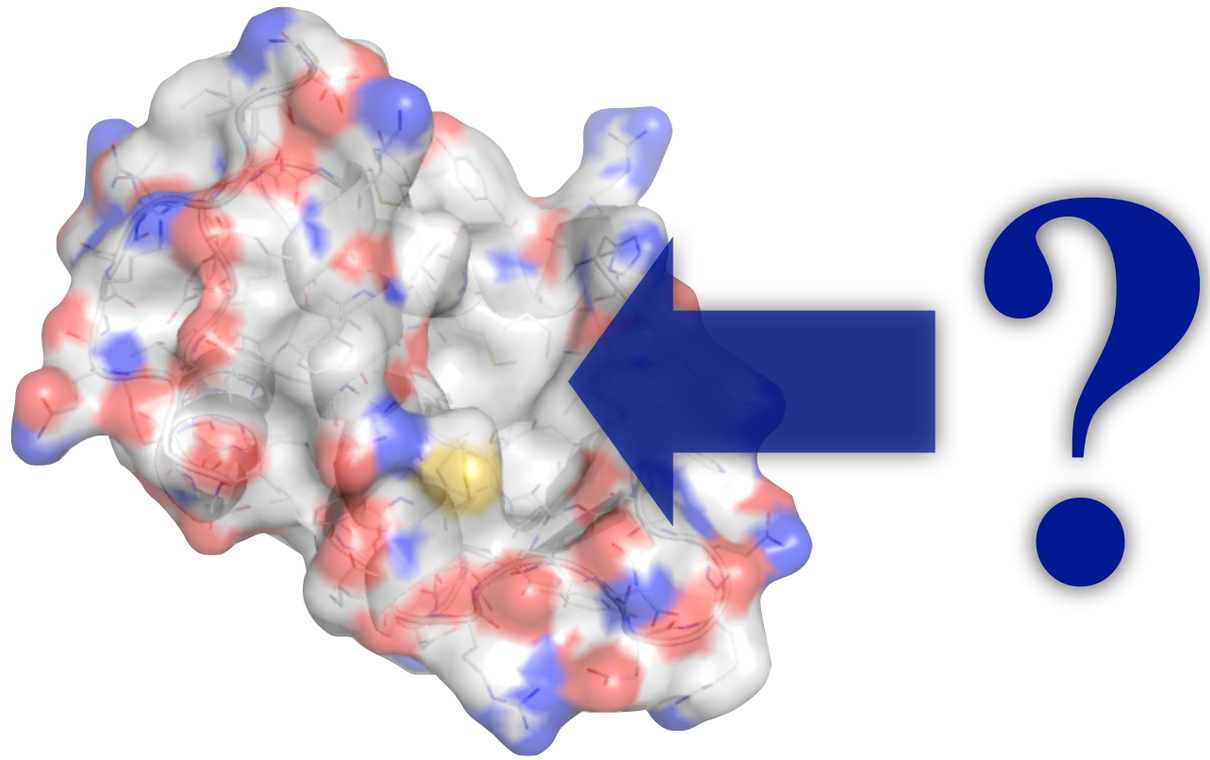
- similarity to known binder

Receptor Based

- dock and score

Interaction Based

- pharmacophore

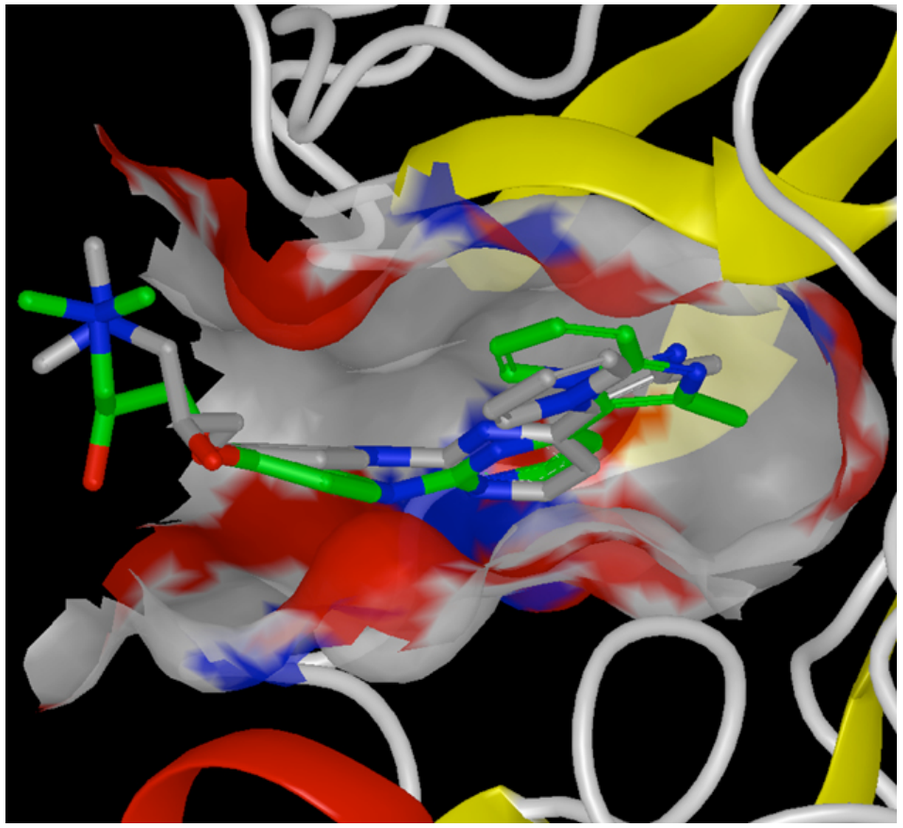


Goal: Find something that binds (potency)

Docking

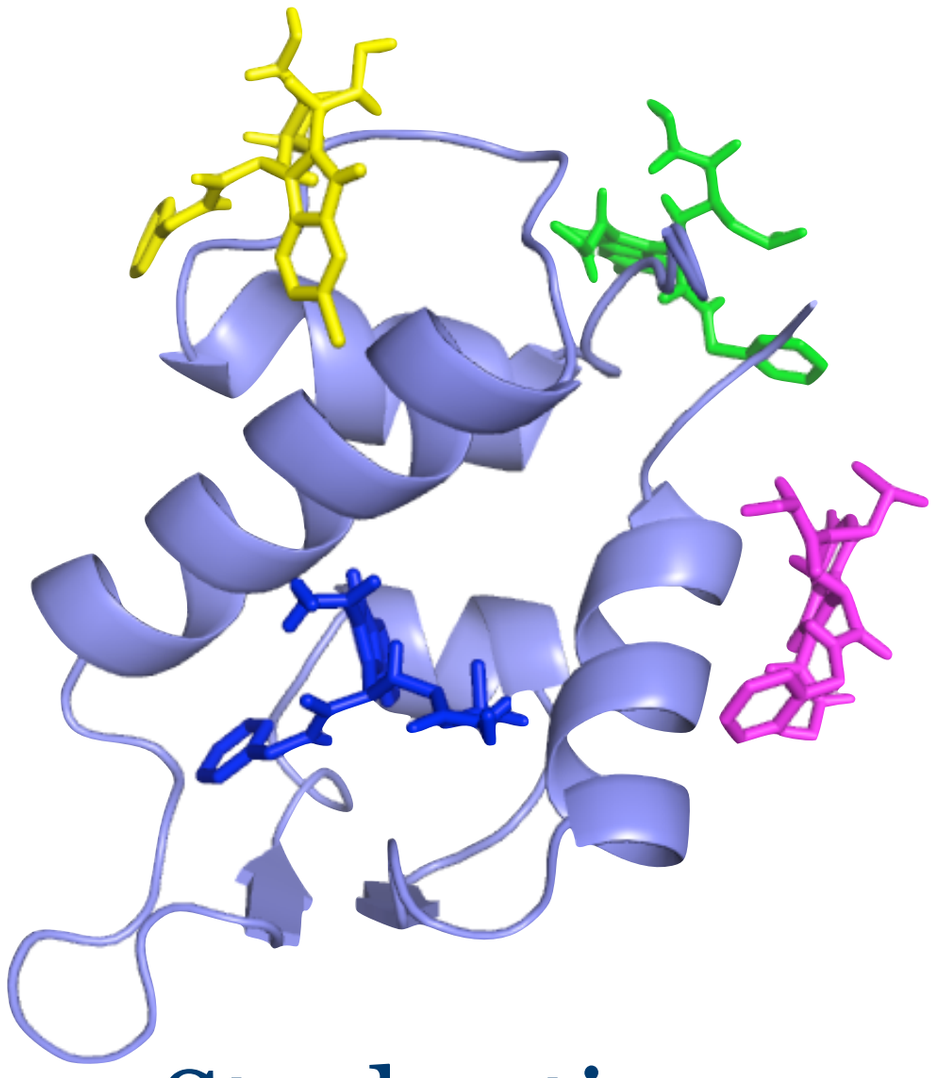
Determine the **conformation** and **pose** of a ligand at a docking site

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

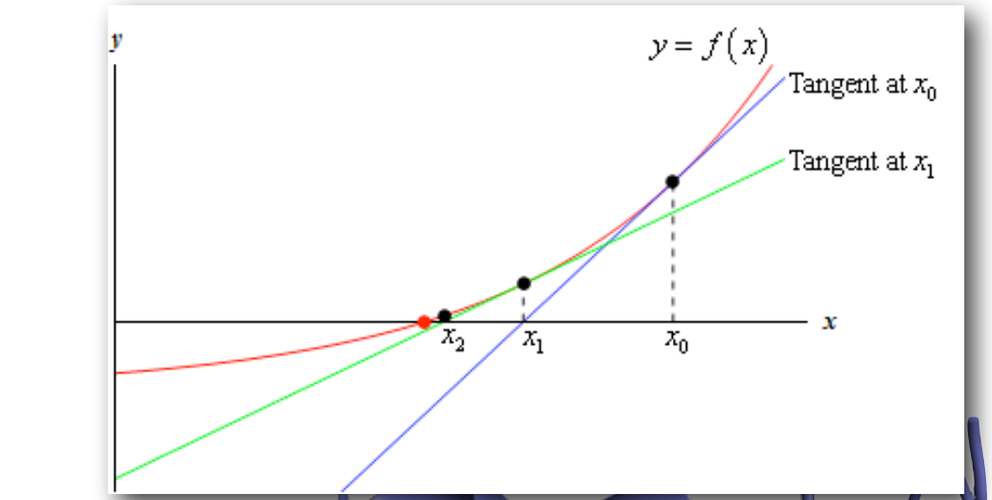


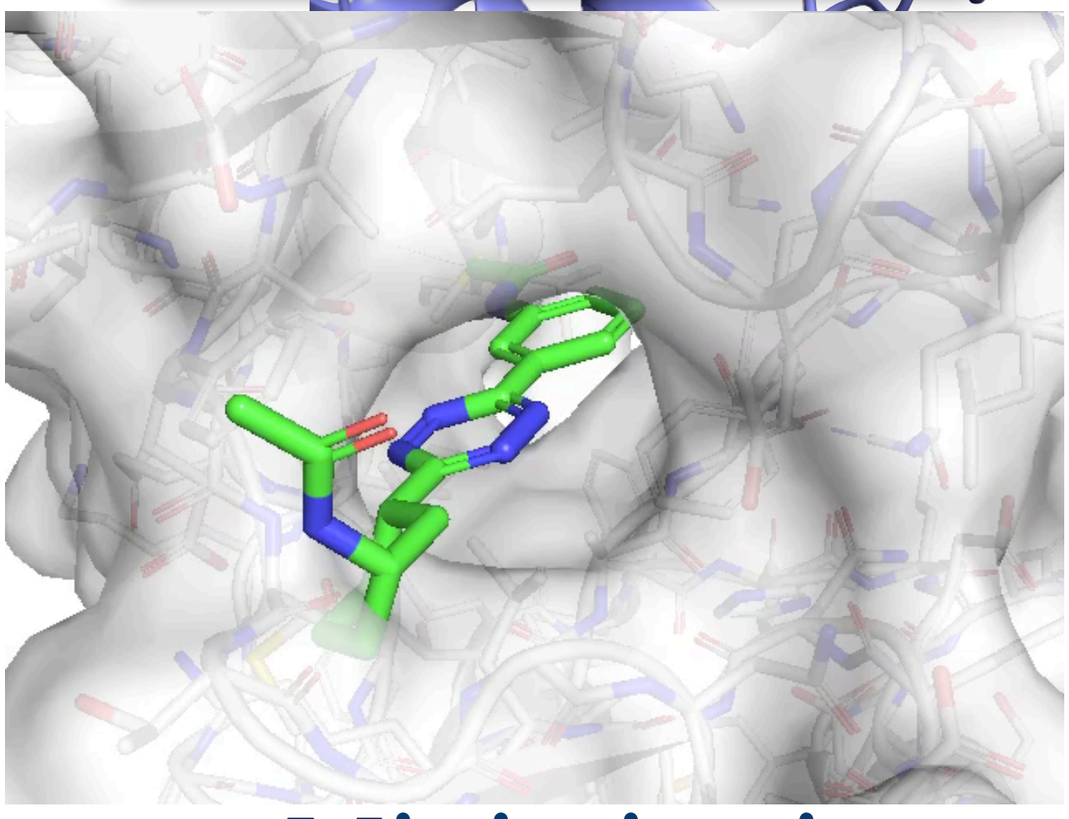
Two Phase Docking

1. Global Pose Estimation



Stochastic





Minimization

Search Algorithms

Systematic

search (BFS, DFS, A*)

Stochastic

Monte Carlo

Genetic/Evolutionary algorithms

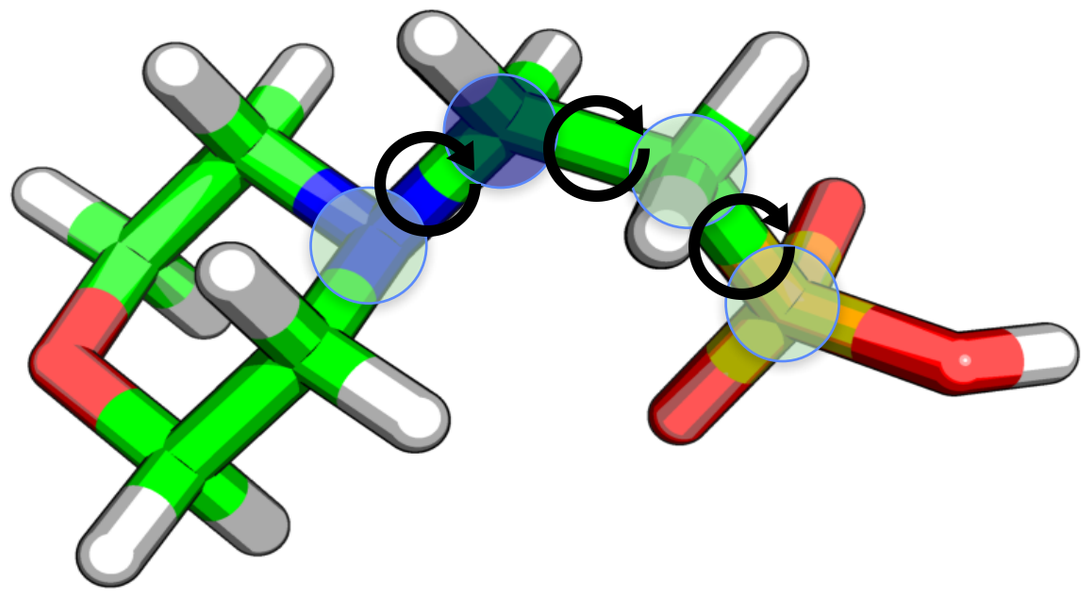
maintain a population of candidates
mutation and crossover

Monte Carlo

1. pick a random neighbor of q_i, which we will call q_j, with probability $\frac{1}{d}$, or $q_j = q_i$ with probability $1 - \frac{d_i}{d}$ if the degree of node i (d_i) is less than d
2. if $E_j \leq E_i$ then move to q_j
3. if $E_j > E_i$ then with probability $e^{-(E_j - E_i)/kT}$ move to q_j, otherwise stay in q_i

What is a neighbor?

What is E_i?



Since our goal is optimization, what else do we have to do?

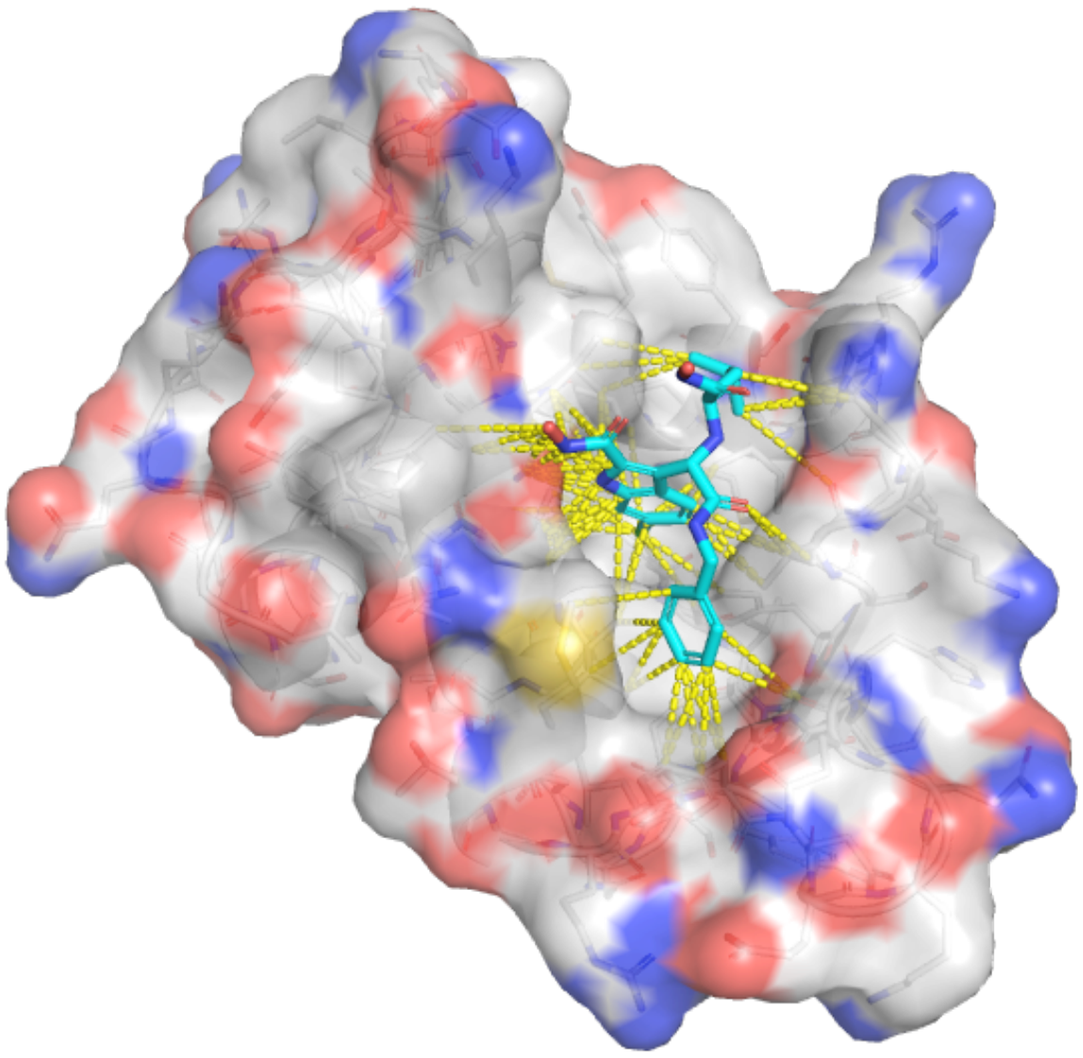
Local Optimization

Broyden–Fletcher–Goldfarb–Shanno (BFGS)

1. Obtain a direction $\mathbf{p}_k$ by solving $B_k \mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable stepsize α_k in the direction found in the first step, so
   $\alpha_k = \arg\min f(\mathbf{x}_k + \alpha \mathbf{p}_k)$.
3. Set $\mathbf{s}_k = \alpha_k \mathbf{p}_k$ and update $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$.
4. $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$.
5. $B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k} - \frac{B_k \mathbf{s}_k \mathbf{s}_k^{\mathrm{T}} B_k}{\mathbf{s}_k^{\mathrm{T}} B_k \mathbf{s}_k}$.

Approximation of Hessian

Receptor Based: Scoring



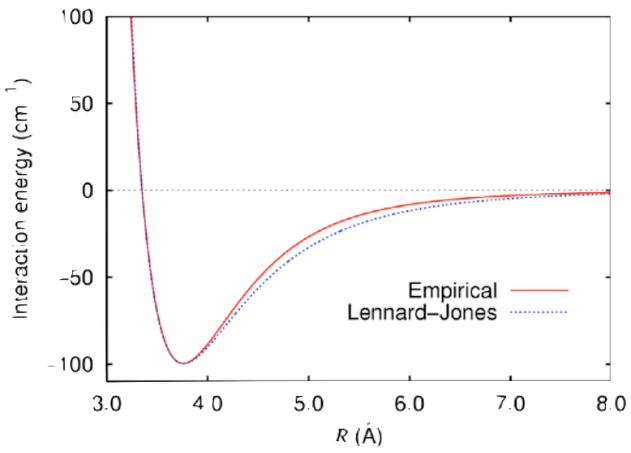
Dock 4.0

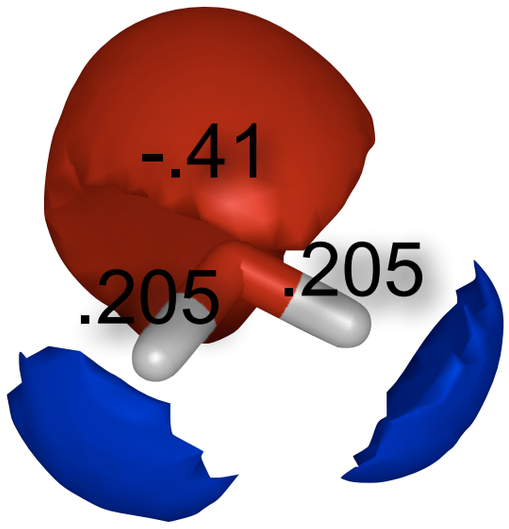
$$E = \sum_{i=1}^{lig} \sum_{j=1}^{rec} \left(\frac{A_{ij}}{r_{ij}^{a}} - \frac{B_{ij}}{r_{ij}^{b}} + 332 \frac{q_i q_j}{D r_{ij}} \right)$$

van der Waals

a = 12, b = 6

Lennard-Jones potential



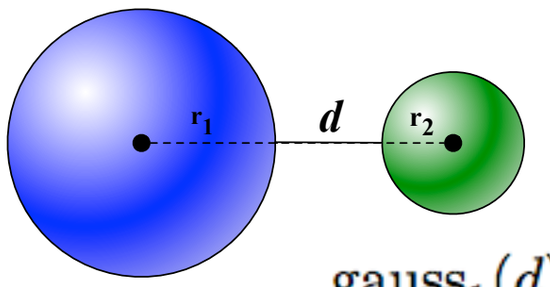


Coulomb's Law

q: partial charges

D: dielectrict constant

AutoDock Vina



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

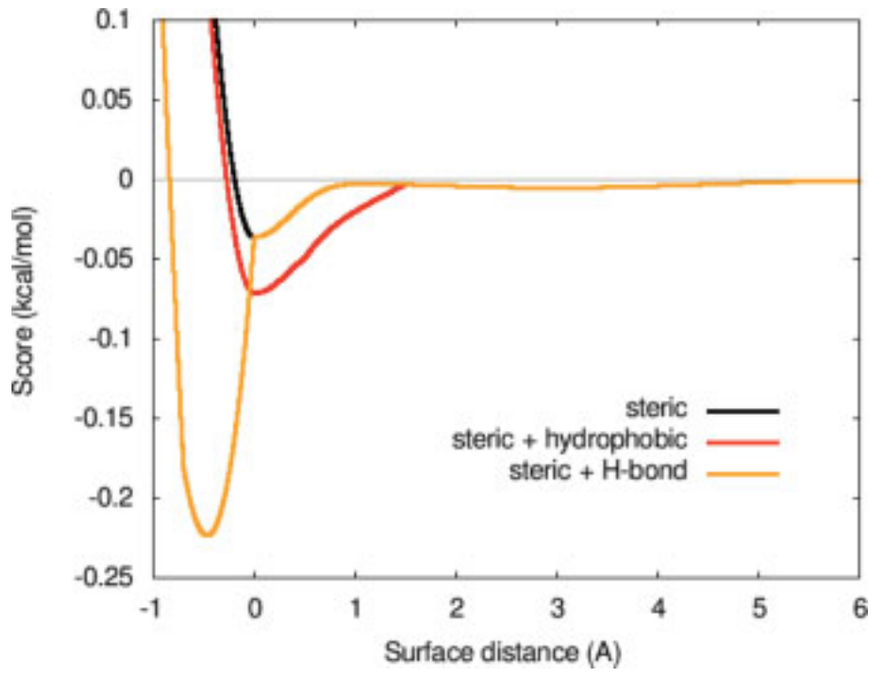
$$\text{gauss}_2(d) = w_{\text{guass}_2} e^{-((d-3)/2)^2}$$

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

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

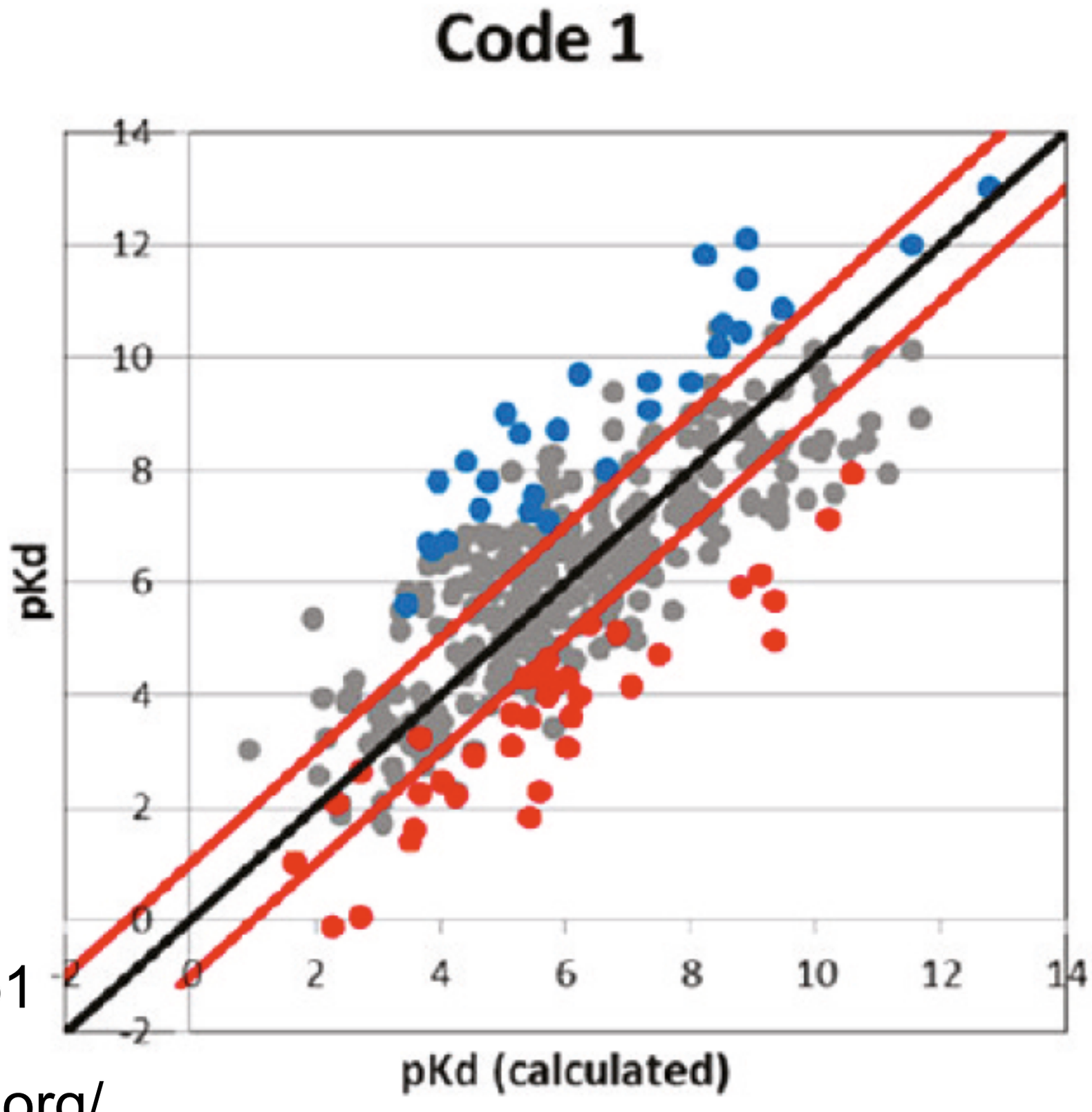
$$\text{hbond}(d) = \begin{cases} w_{\text{hbond}} & d < -0.7 \\ 0 & d > 0 \\ w_{\text{hbond}}(-\frac{10}{7} d) & otherwise \end{cases}$$

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



Scoring

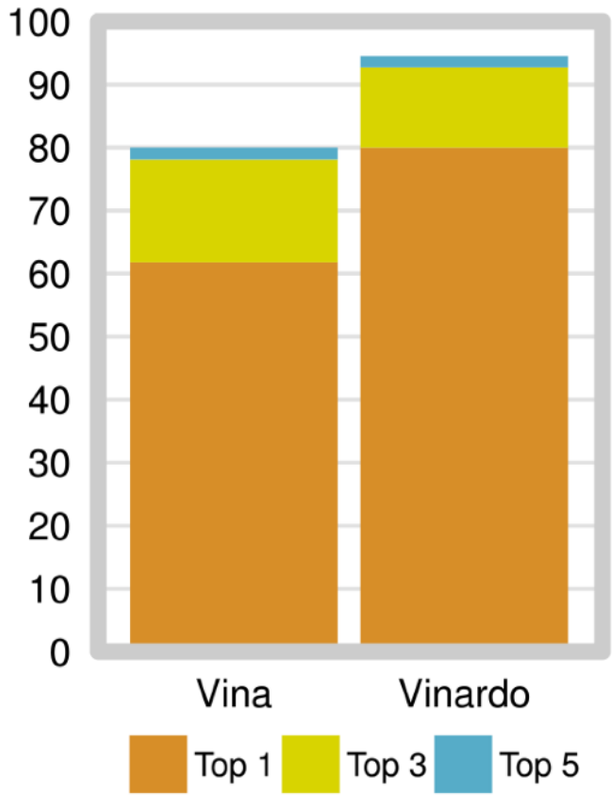
Ideally, score would equal affinity – but this is an unsolved problem.

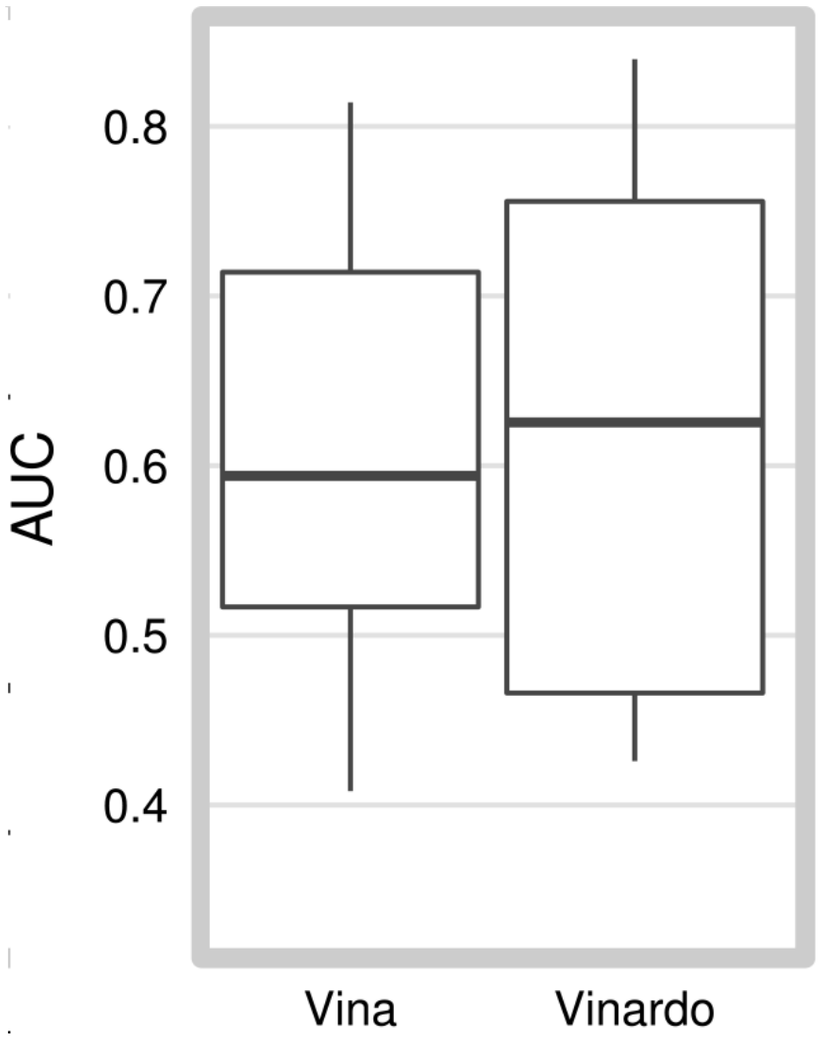


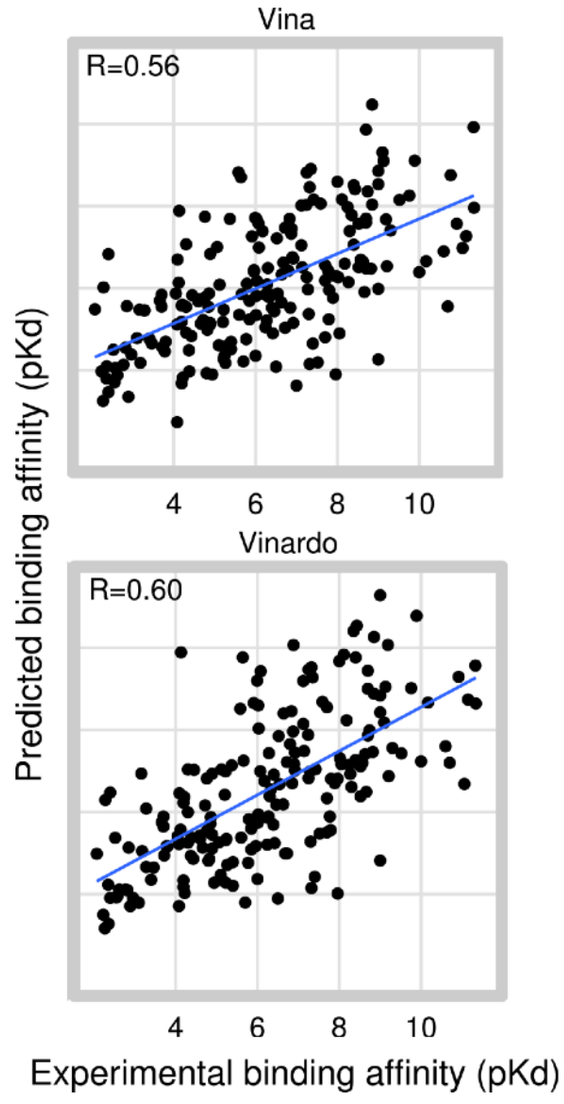
$R^2 = 0.58$
RMSE = 1.51

http://www.csardock.org/

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?

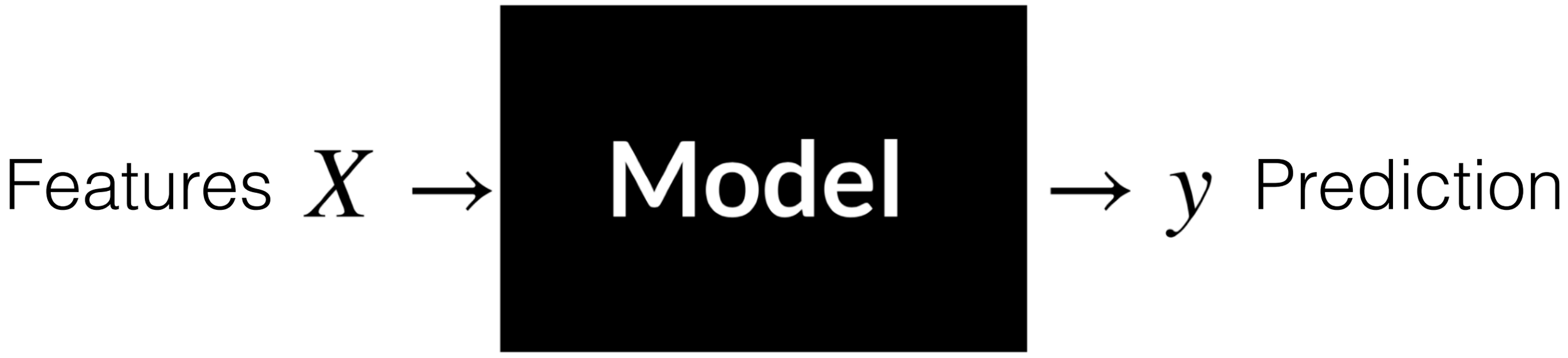
Key Idea: Leverage "big data"

231,655,275 bioactivities in PubCher

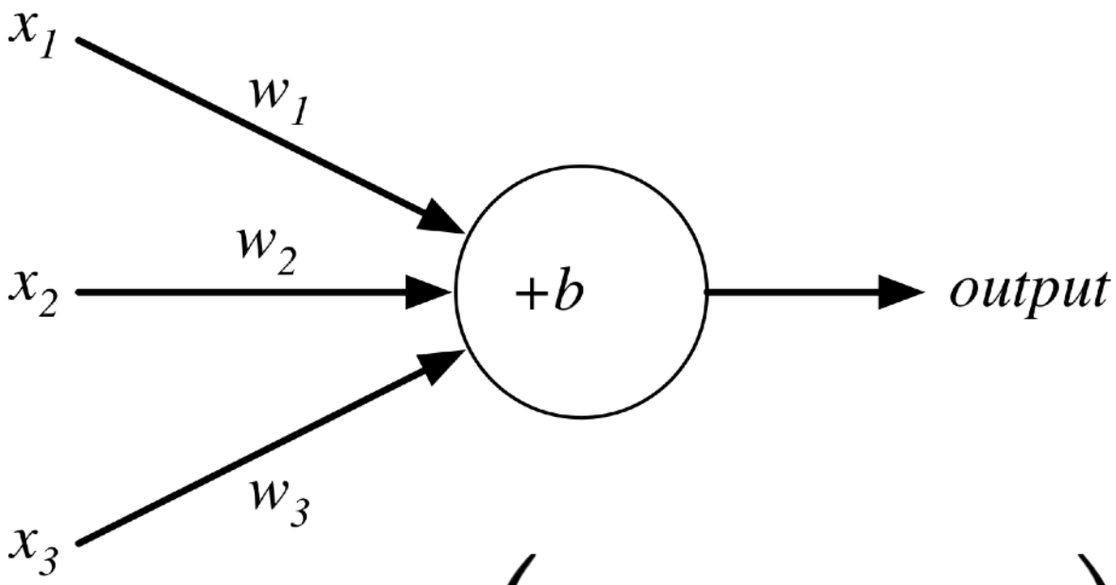
125,526 structures in the PDB

16,179 annotated complexes in PDBbind

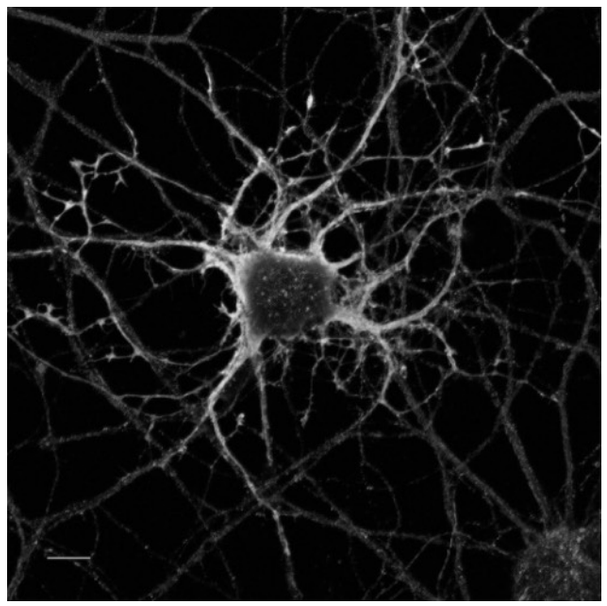
Machine Learning

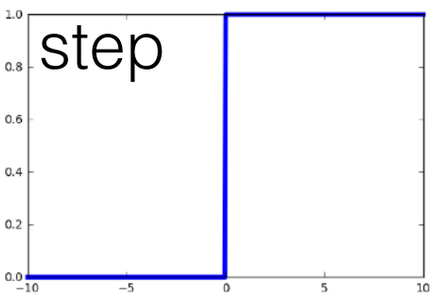


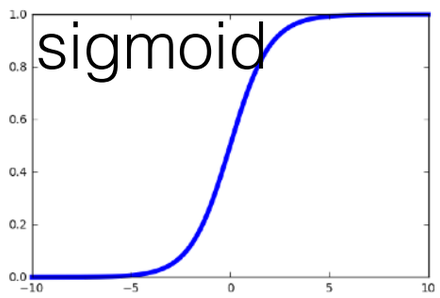
Neural Networks

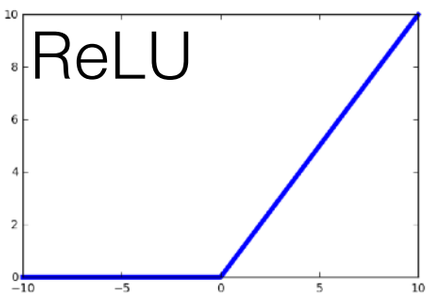


$$output = \sigma\left(\sum_i w_i x_i + b\right)$$

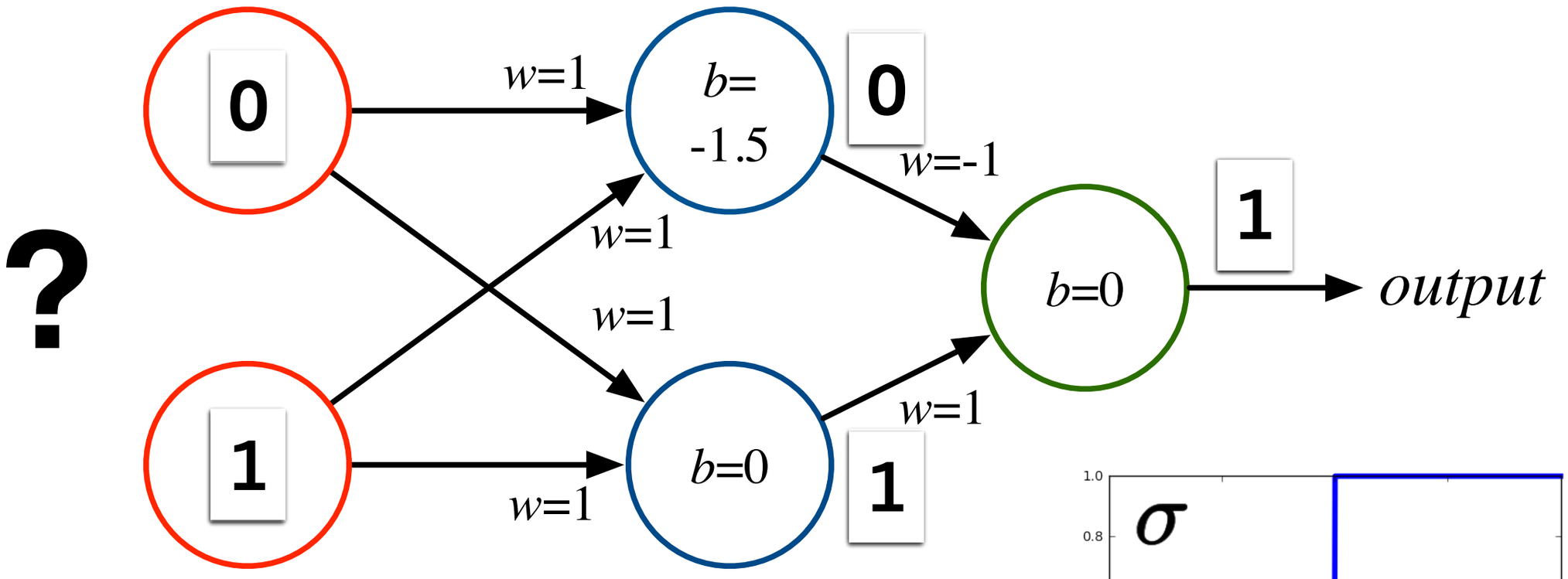


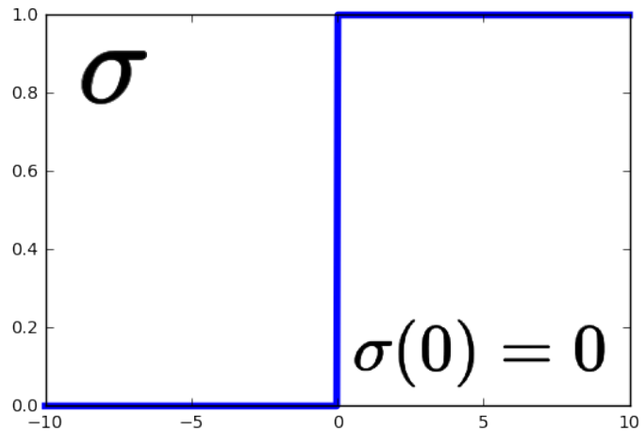




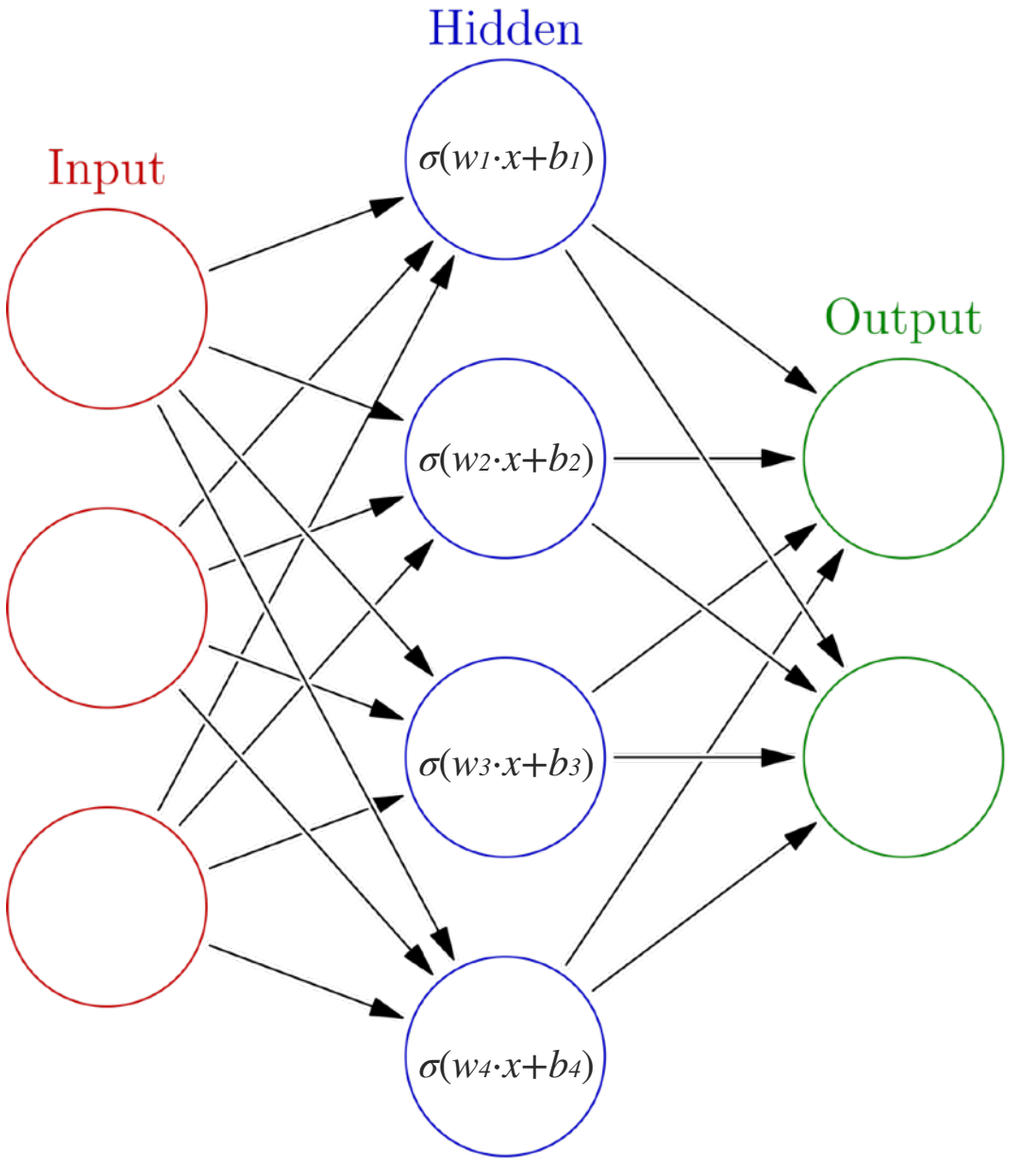


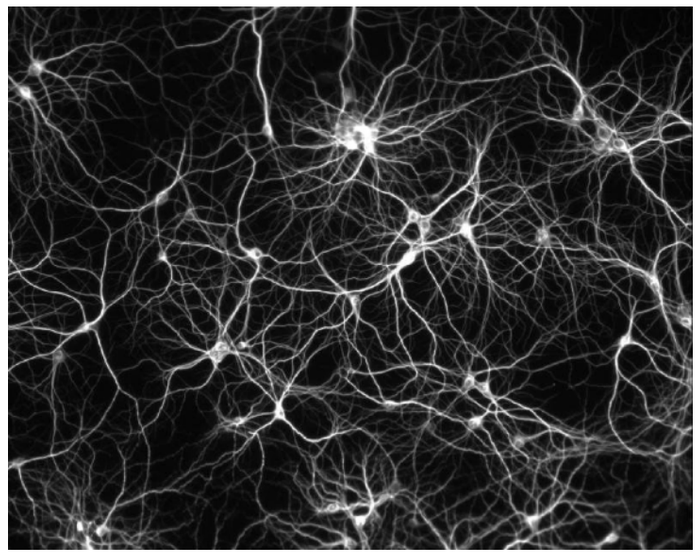
Neural Network Example





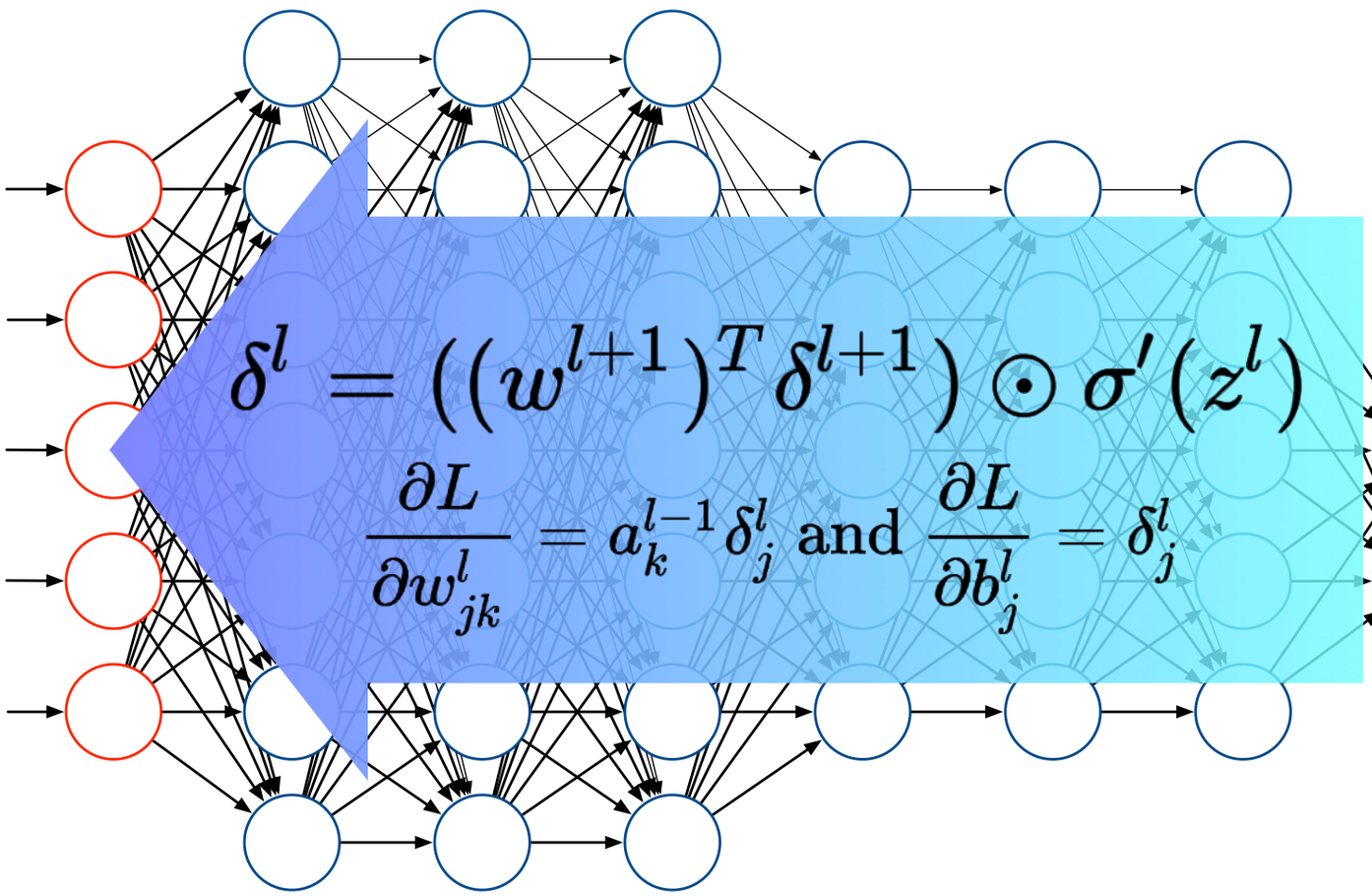
Neural Networks





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



Backpropagation

Backpropagation is an efficient algorithm for computing the partial derivatives needed by the gradient descent update rule. For a training example x and loss function L in a network with N layers:

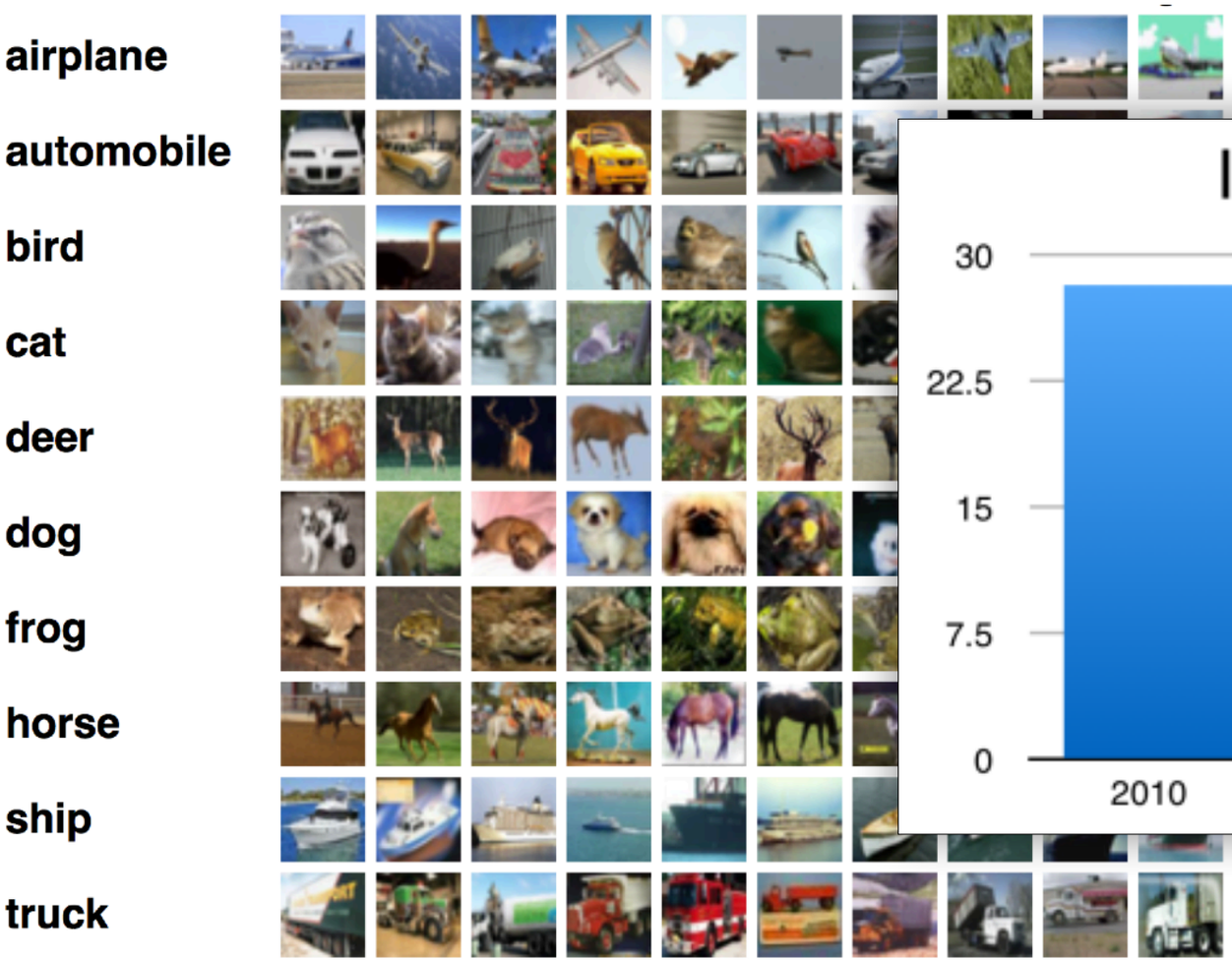
1. **Feedforward**. For each layer l compute
$$a^l = \sigma(z^l)$$
where z is the weighted input and a is the activation induced by x (these are vectors representing all the nodes of layer l).

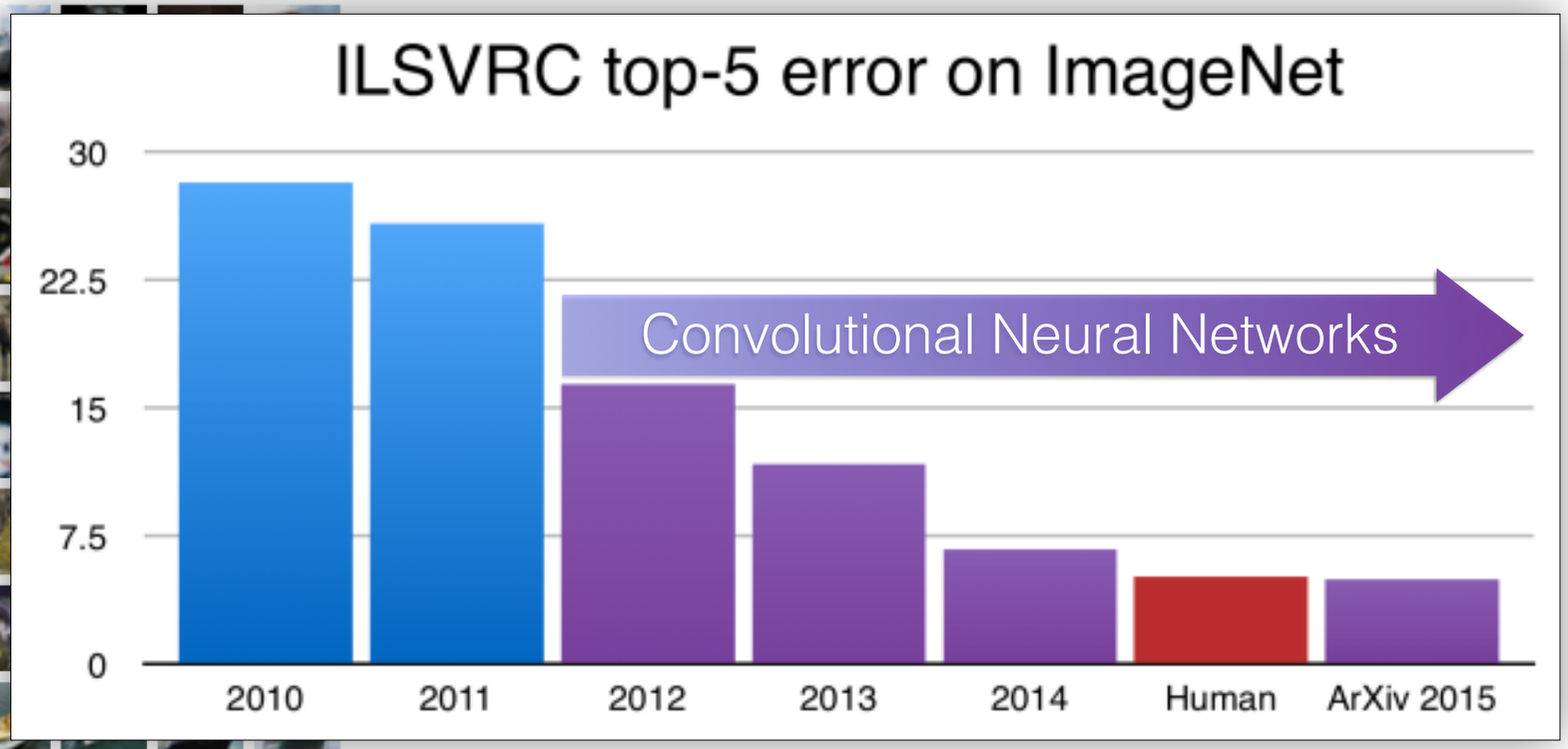
2. **Compute output error**
$$\delta^N = \nabla_a L \odot \sigma'(z^N)$$
where $\nabla_a L_j = \partial L / \partial a_j^N$, the gradient of the loss with respect to the output activations. $\odot$ is the elementwise product.

3. **Backpropagate the error**
$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

4. **Calculate gradients**
$$\frac{\partial L}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \text{ and } \frac{\partial L}{\partial b_j^l} = \delta_j^l$$

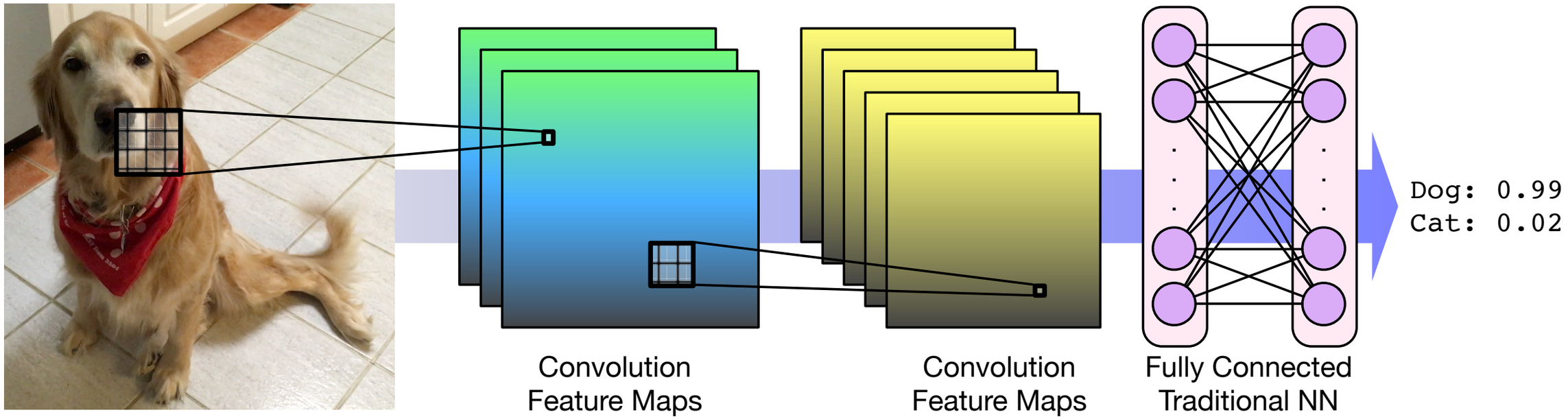
Image Recognition

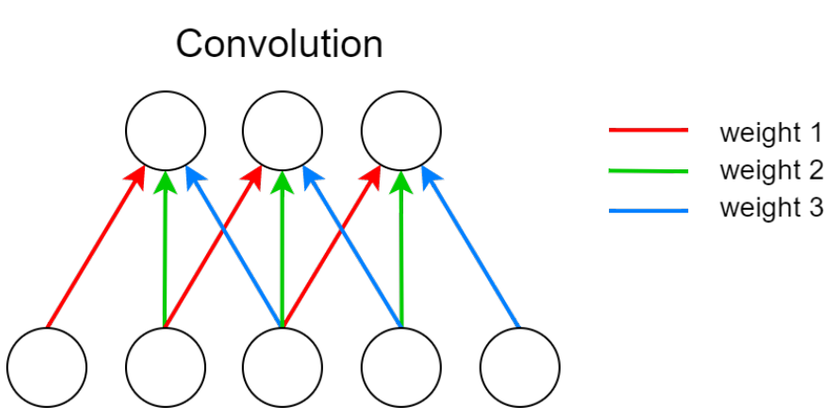


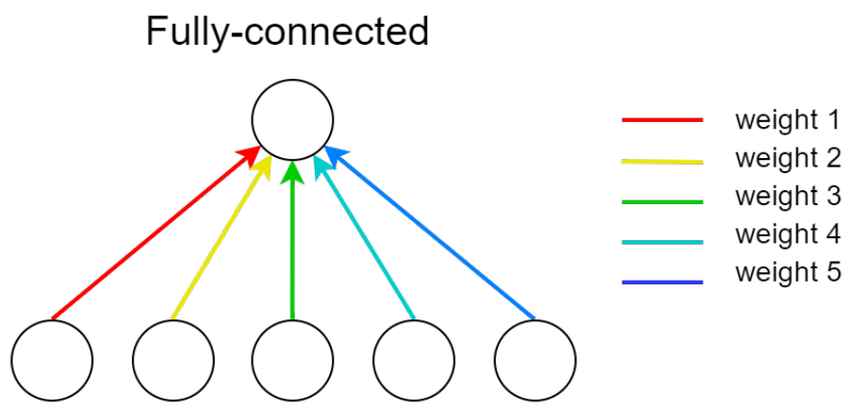


https://devblogs.nvidia.com

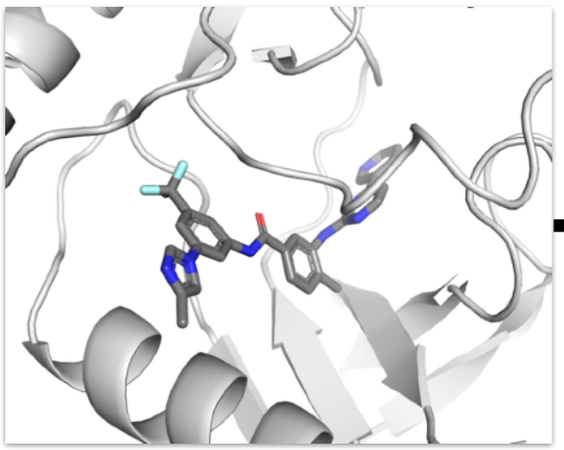
Convolutional Neural Networks







CNNs for Protein-Ligand Scoring



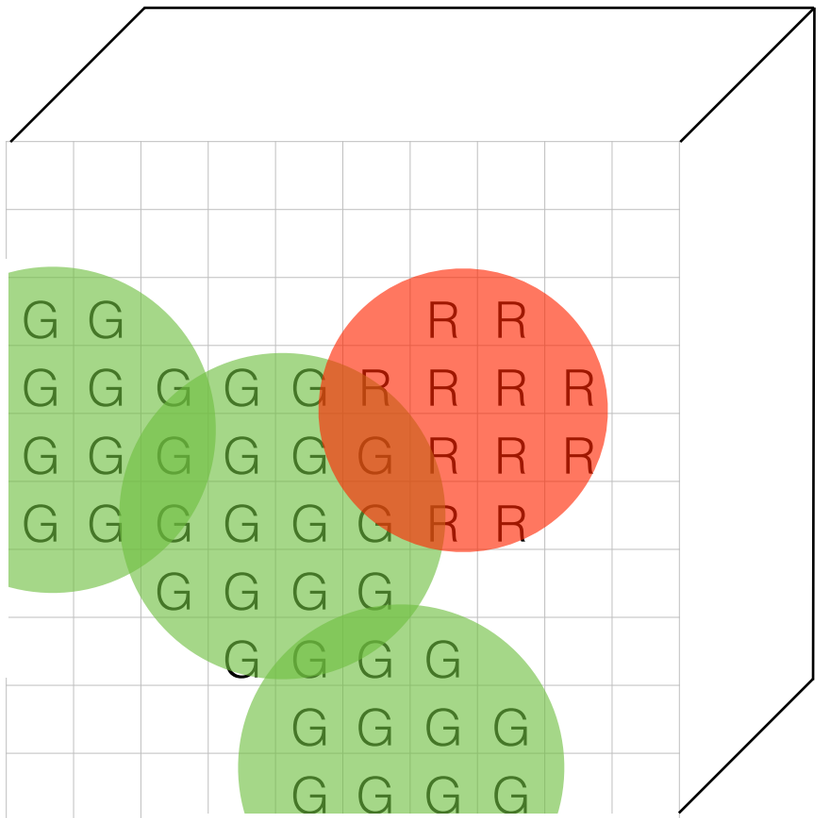
CNN

Pose Prediction

Binding Discrimination

Affinity Prediction

Protein-Ligand Representation

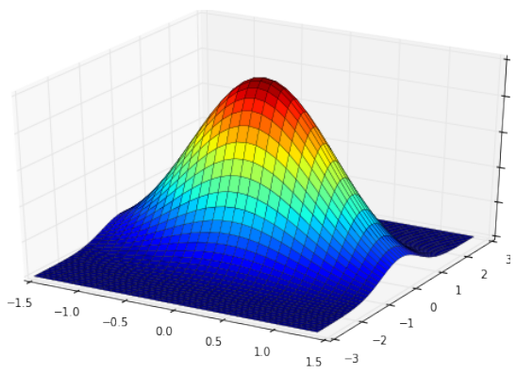


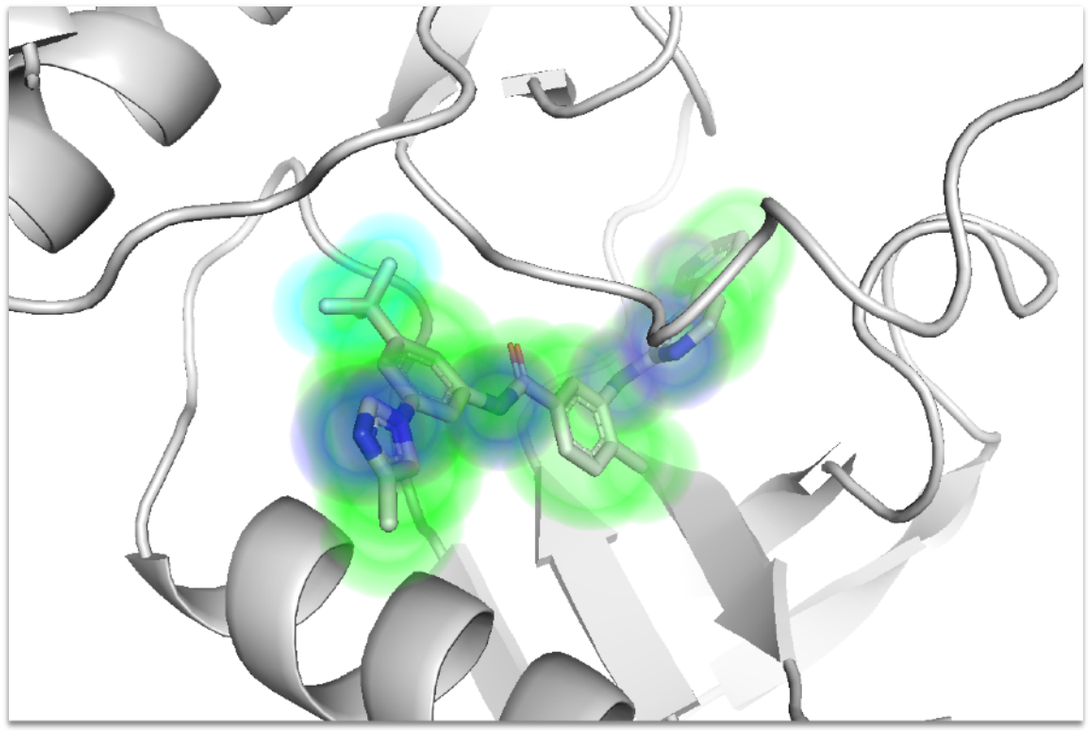
(R,G,B) pixel → (Carbon, Nitrogen, Oxygen,…) **voxe**

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

Atom Density

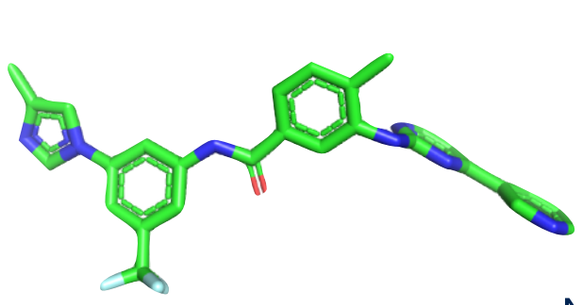
$$A(d,r) = \begin{cases} e^{-\frac{2d^2}{r^2}} & 0 \le d < r \\ \frac{4}{e^2 r^2} d^2 - \frac{12}{e^2 r} d + \frac{9}{e^2} & r \le d < 1.5r \\ 0 & d \ge 1.5r \end{cases}$$





Gaussian

Atom Types

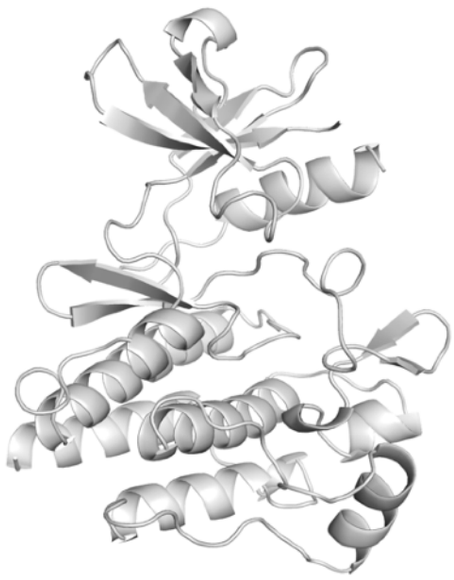


Ligand

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

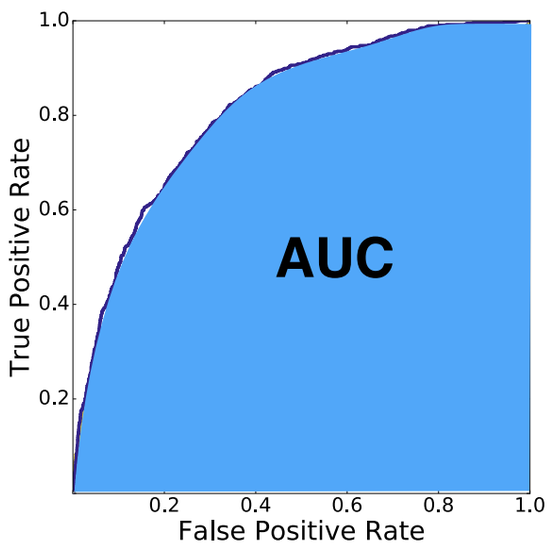
PDBbind

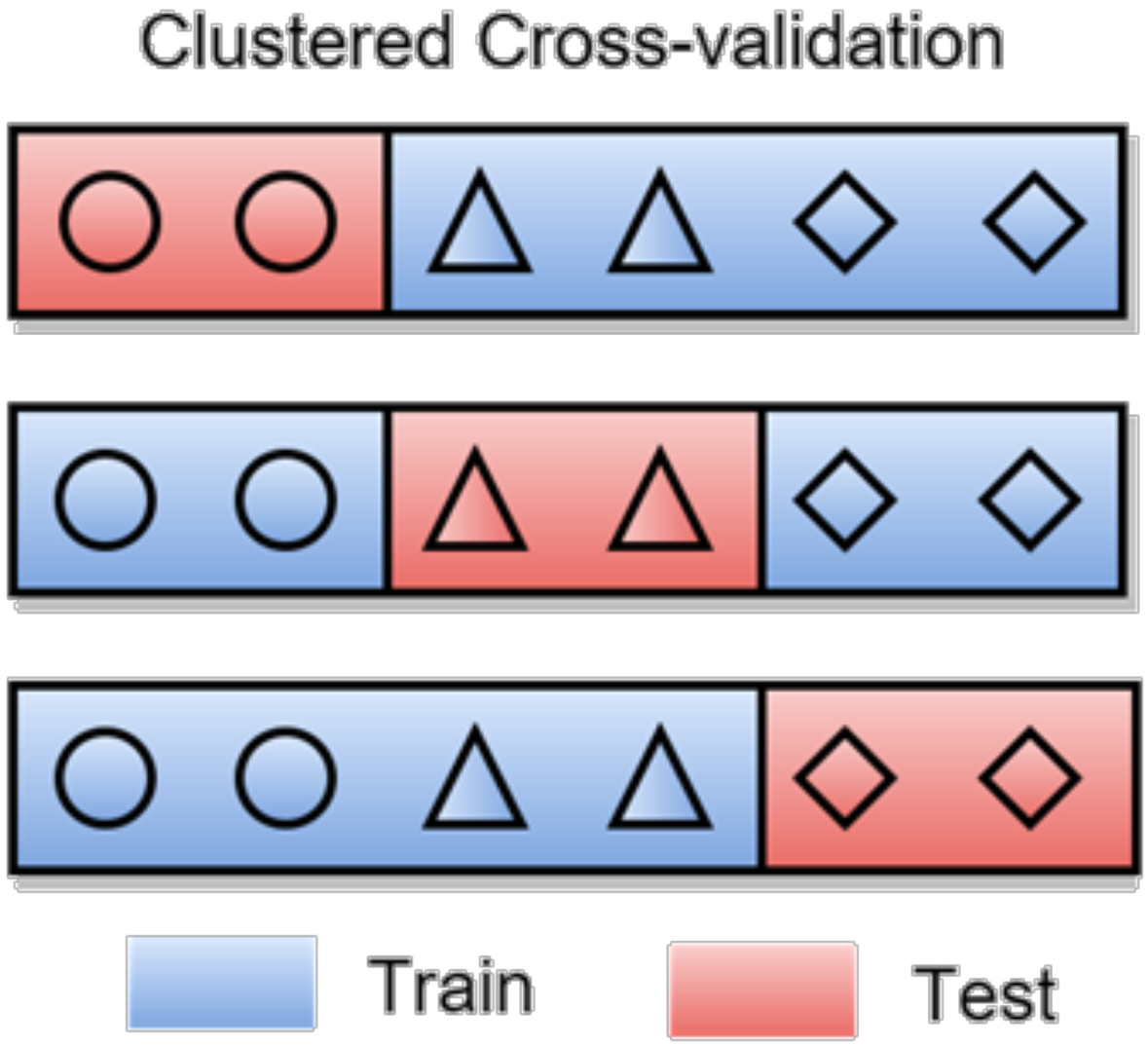
- 8,688 low RMSD poses
- assign known affinity
- **regression problem**

Model Evaluation

CSAR: >90% similar targets kept in same fold

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

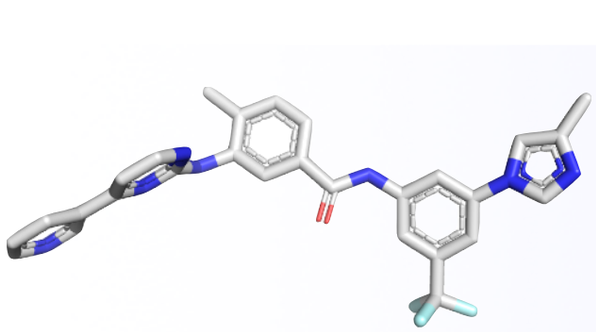


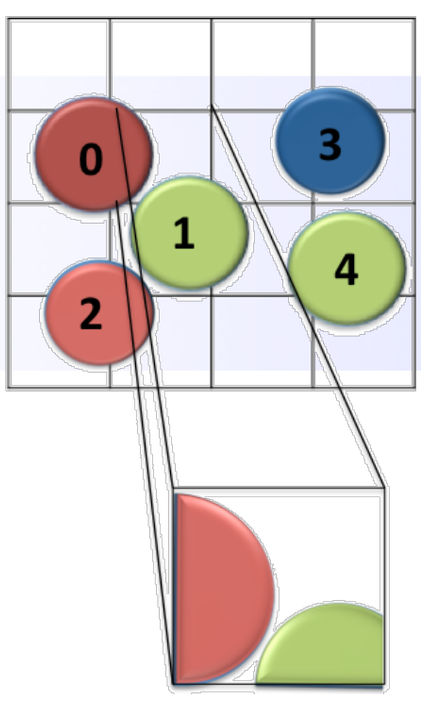


Model Training

Custom **`MolGridDataLayer`**

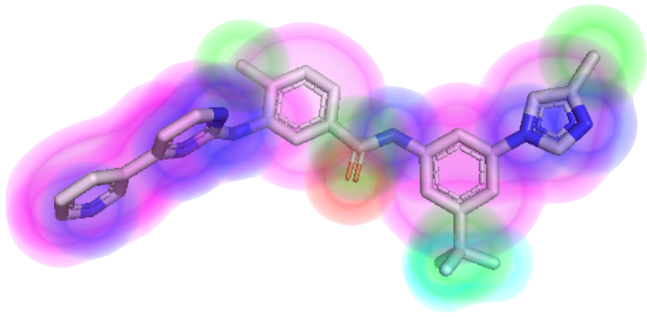
Parallelize over *atoms* to obtain a mask of atoms that overlap each grid region
Use exclusive scan to obtain a list of atom indices from the mask
Parallelize over *grid points*, using reduced atom list to avoid $O(N_{atoms})$ check





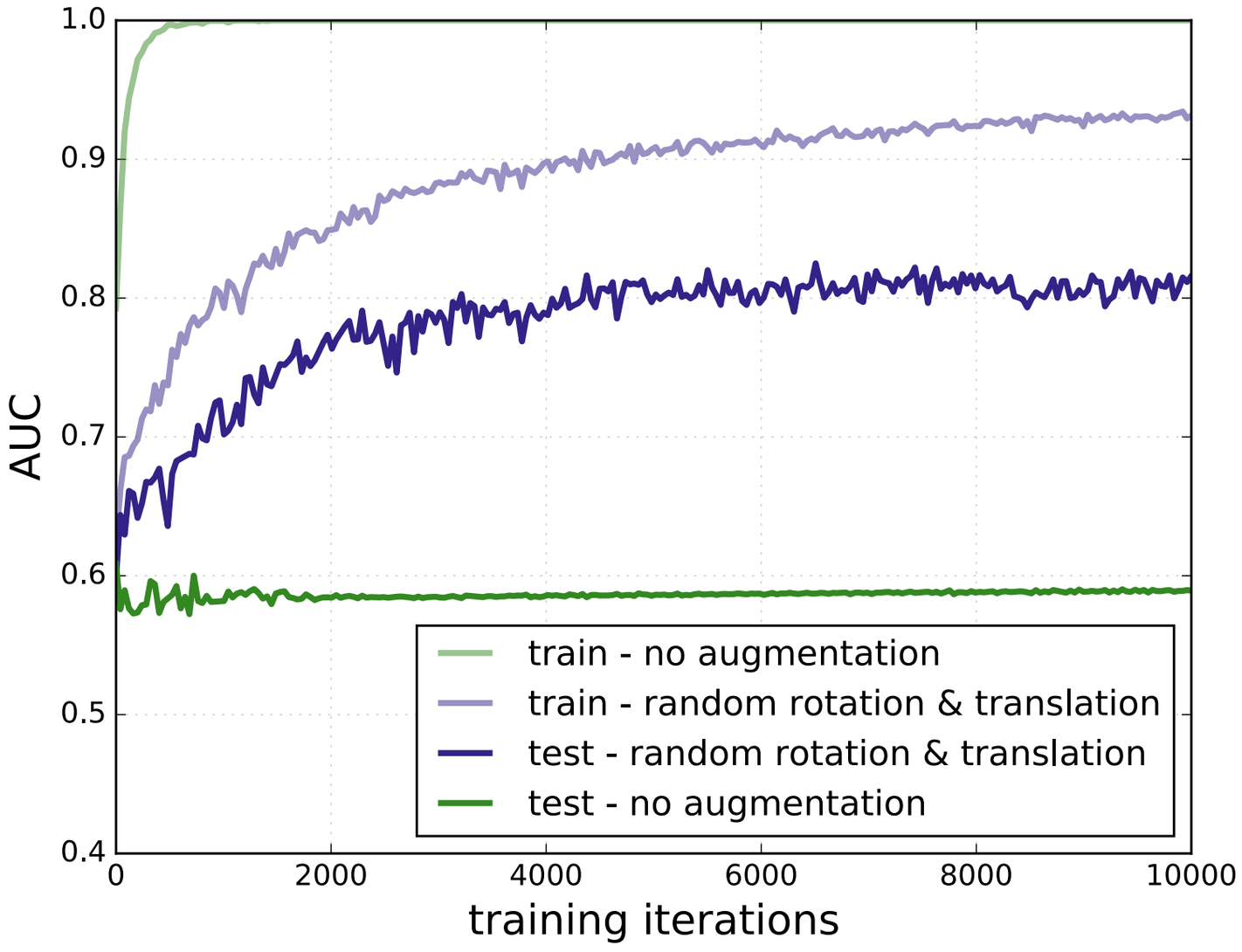
For example, consider subgrid 5:

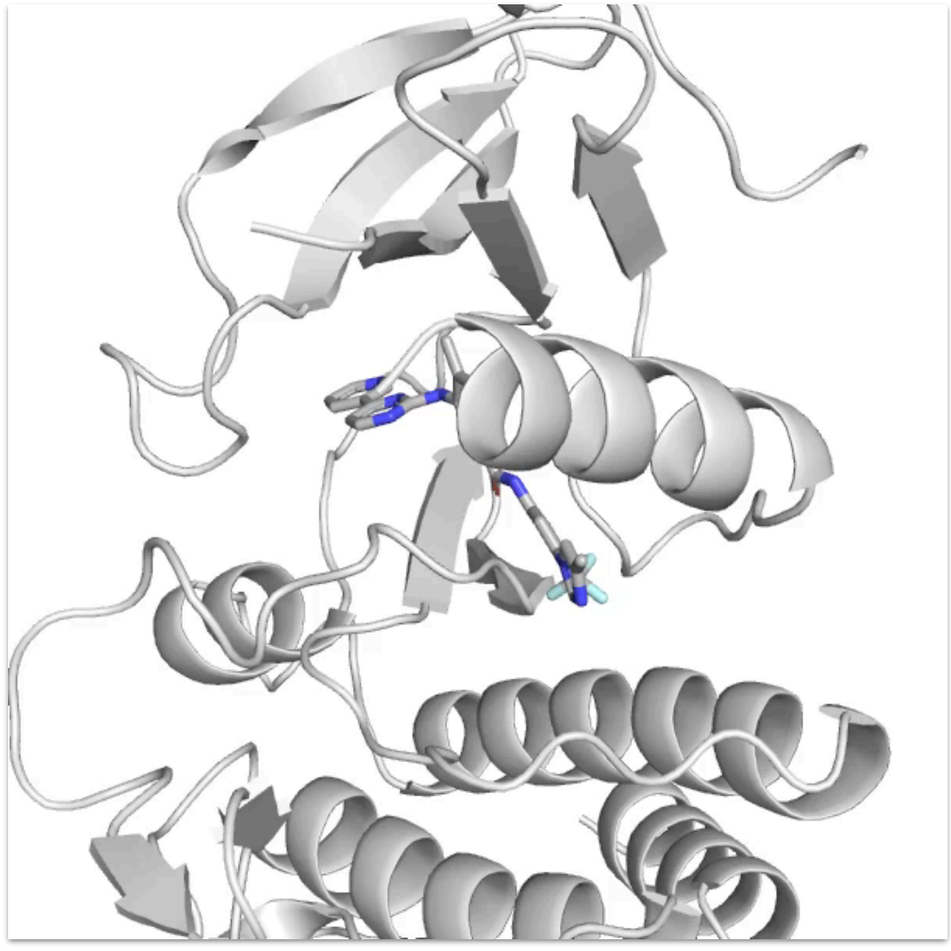
Atom mask:	1	1	0	0	0
Exclusive scan:	0	1	2	2	2
Final indices:	0	1			



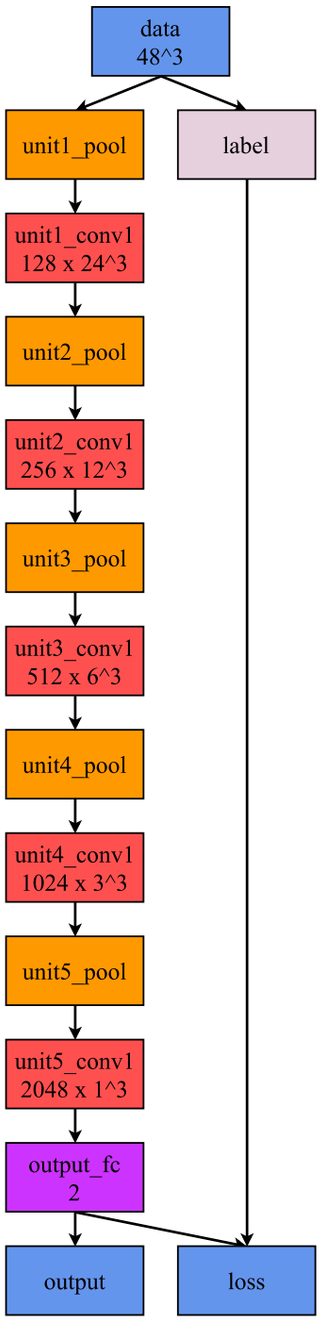
NVIDIA

Data Augmentation





Model Optimization



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

Atom Density Type
- Boolean
- Gaussian

Radius Multiple

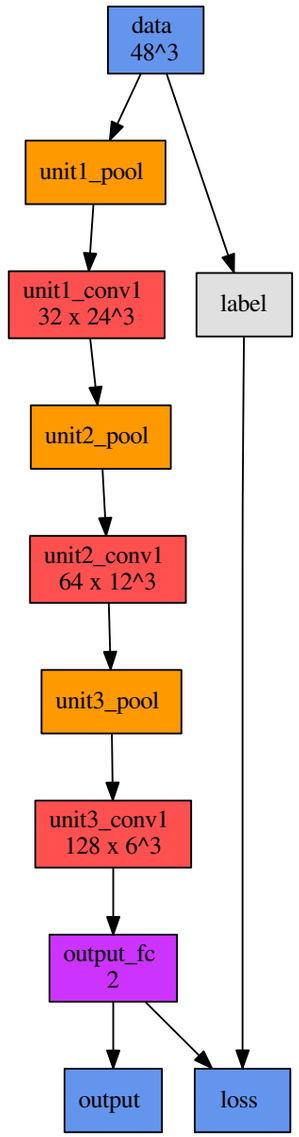
Resolution

Pooling

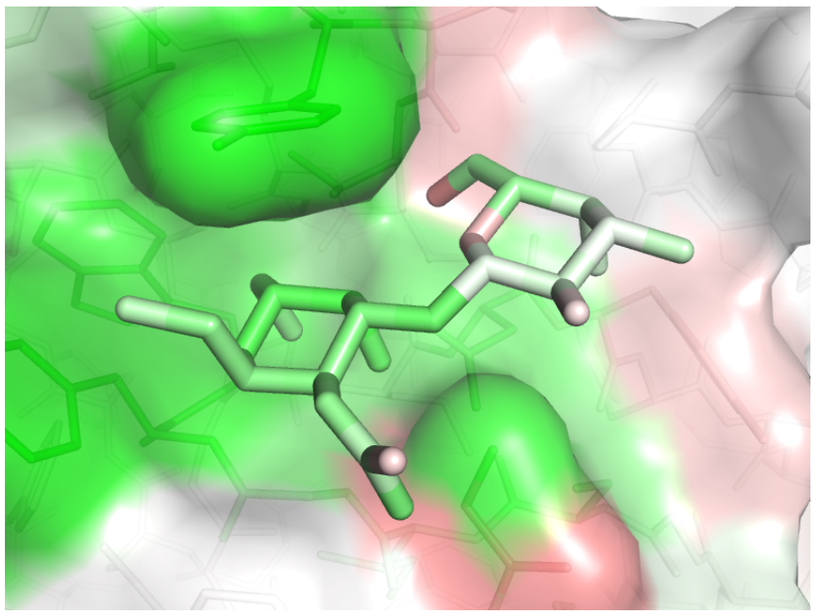
Depth

Width

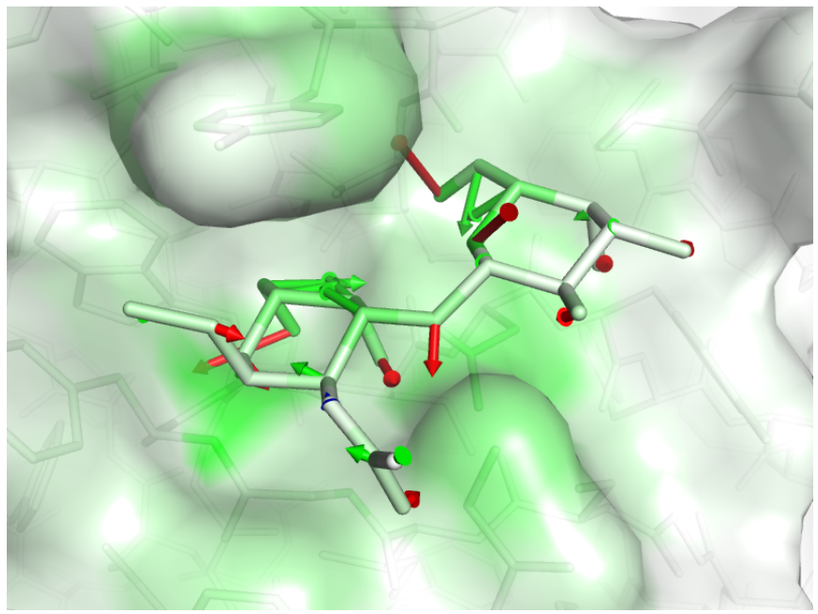
Fully Connected Layers



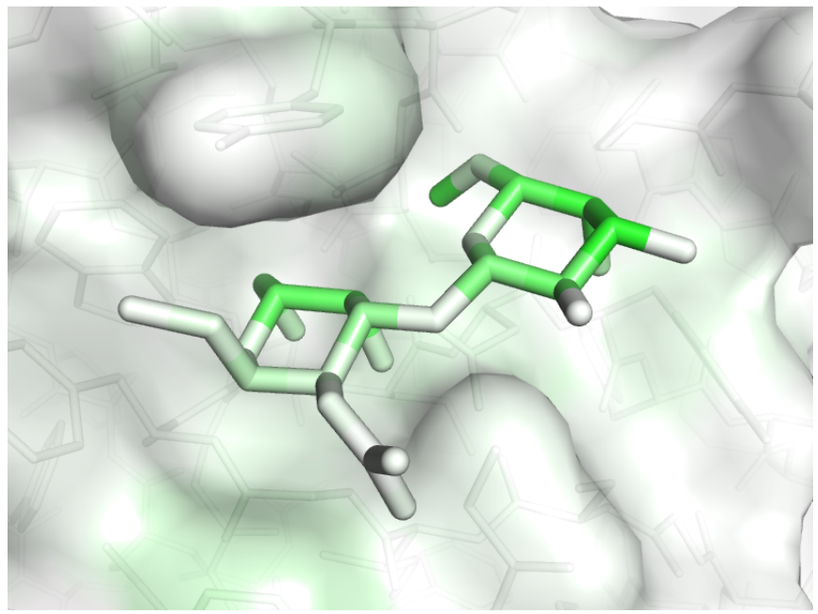
Visualization



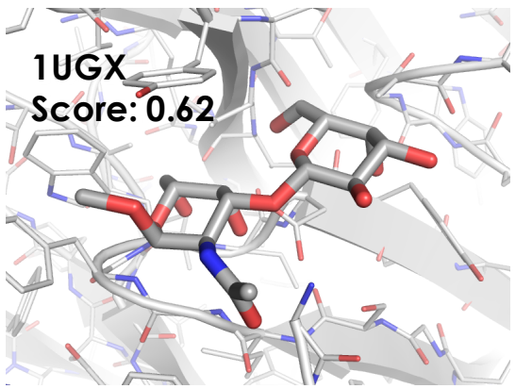
masking



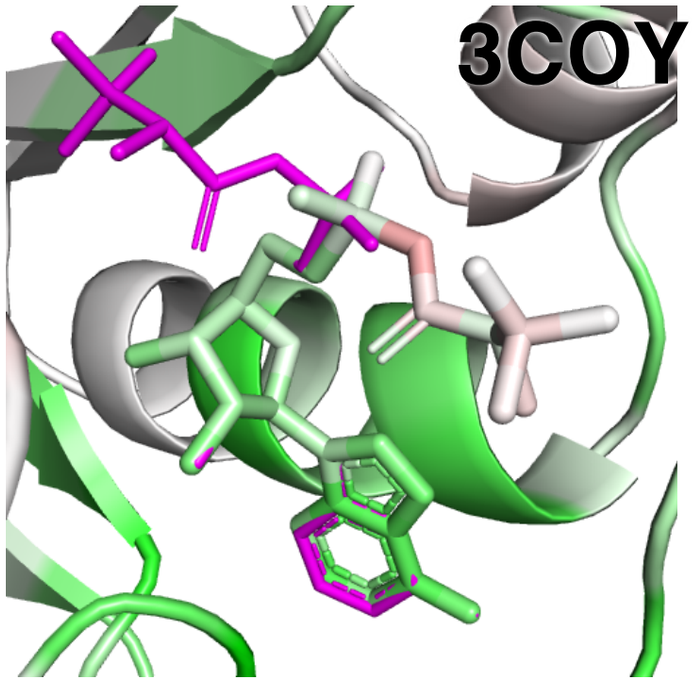
gradients

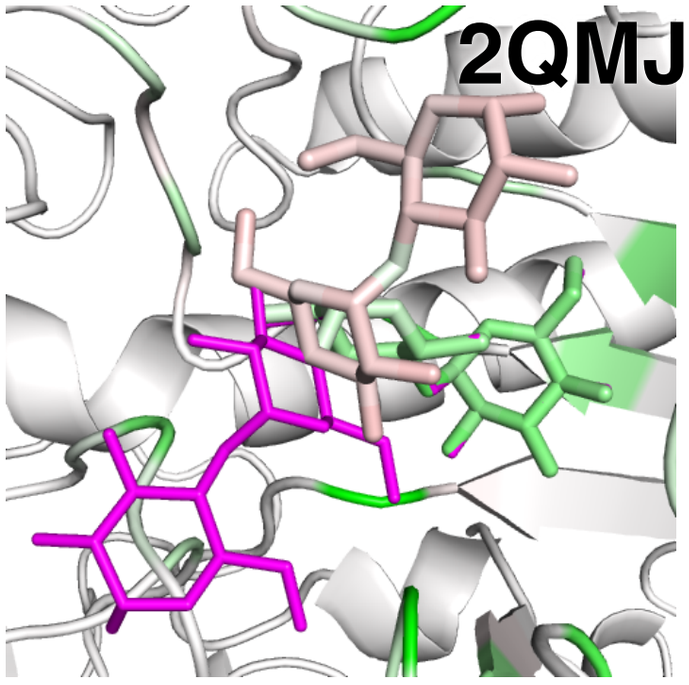


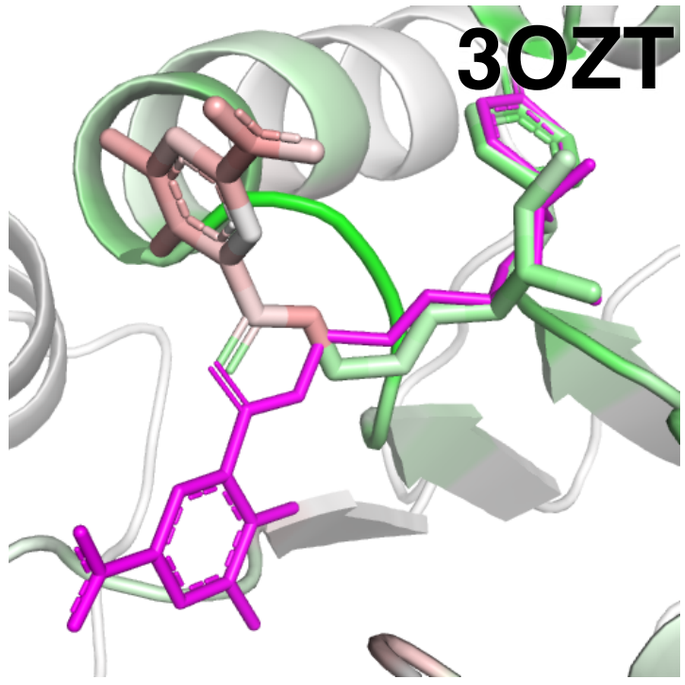
layer-wise relevance



Pose Sensitivity

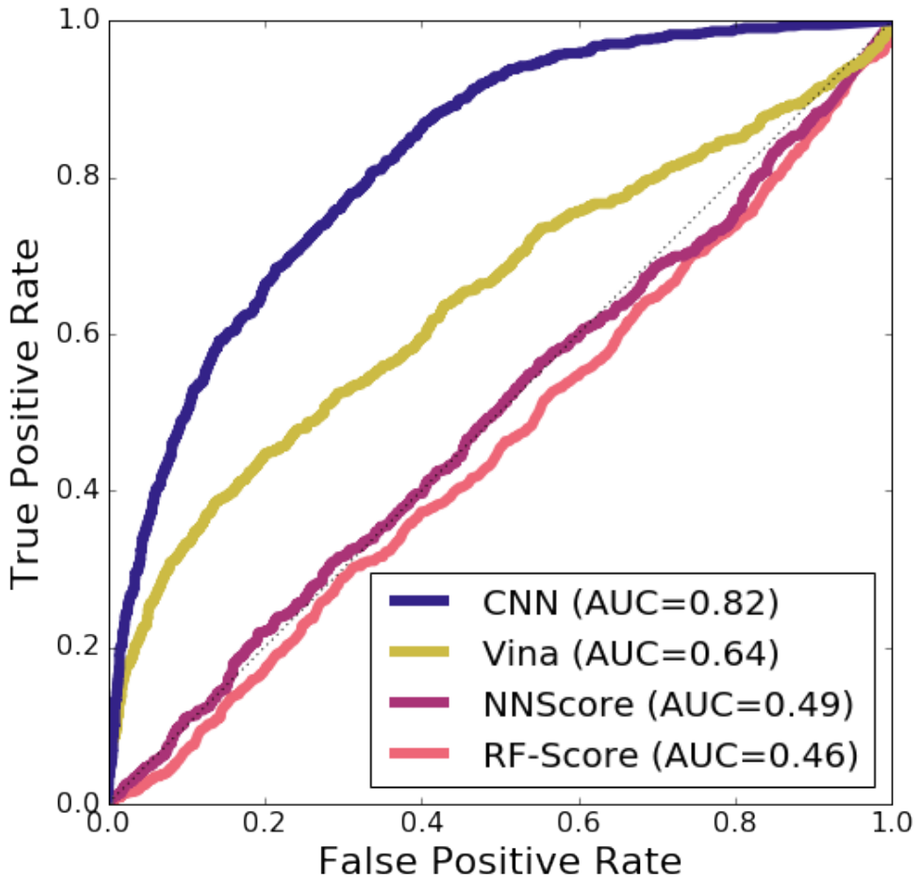




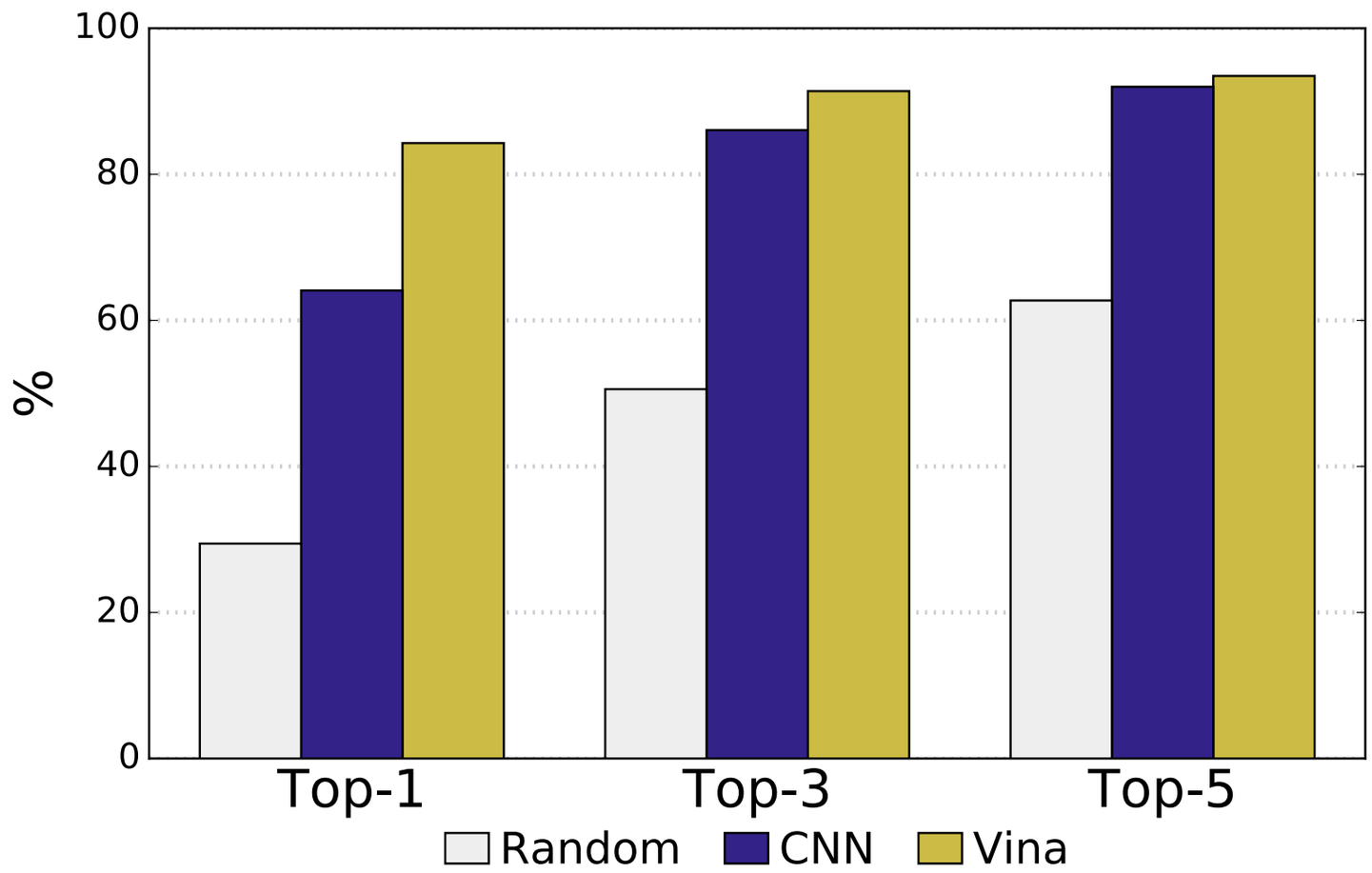


Partially Aligned Poses

Pose Prediction (CSAR)

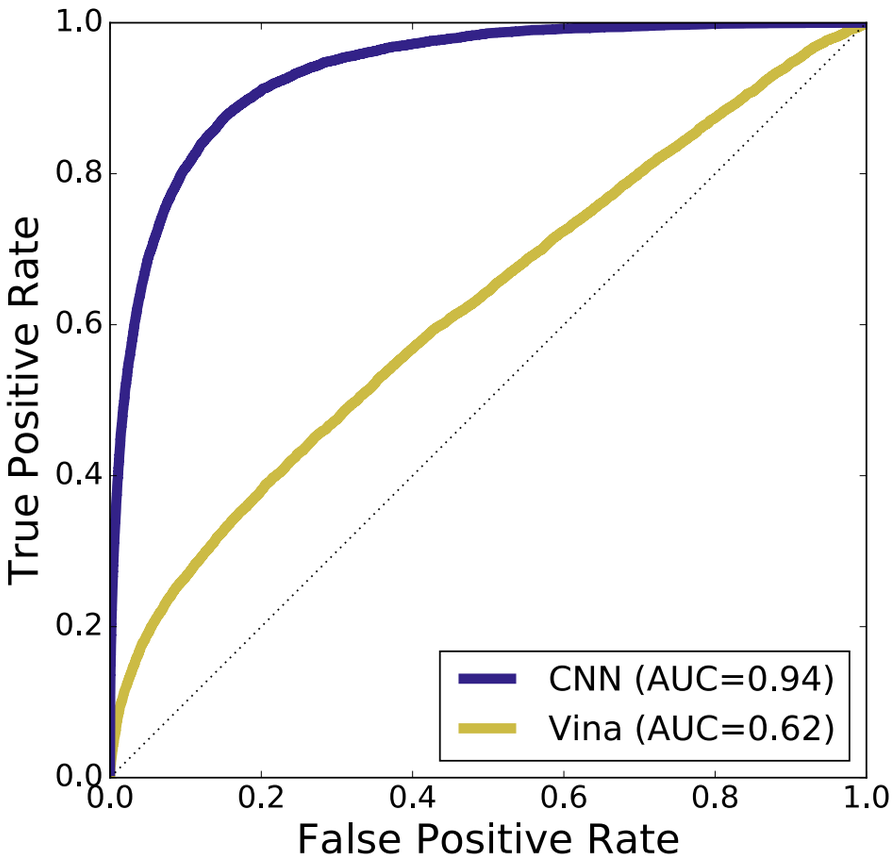


inter-target ranking

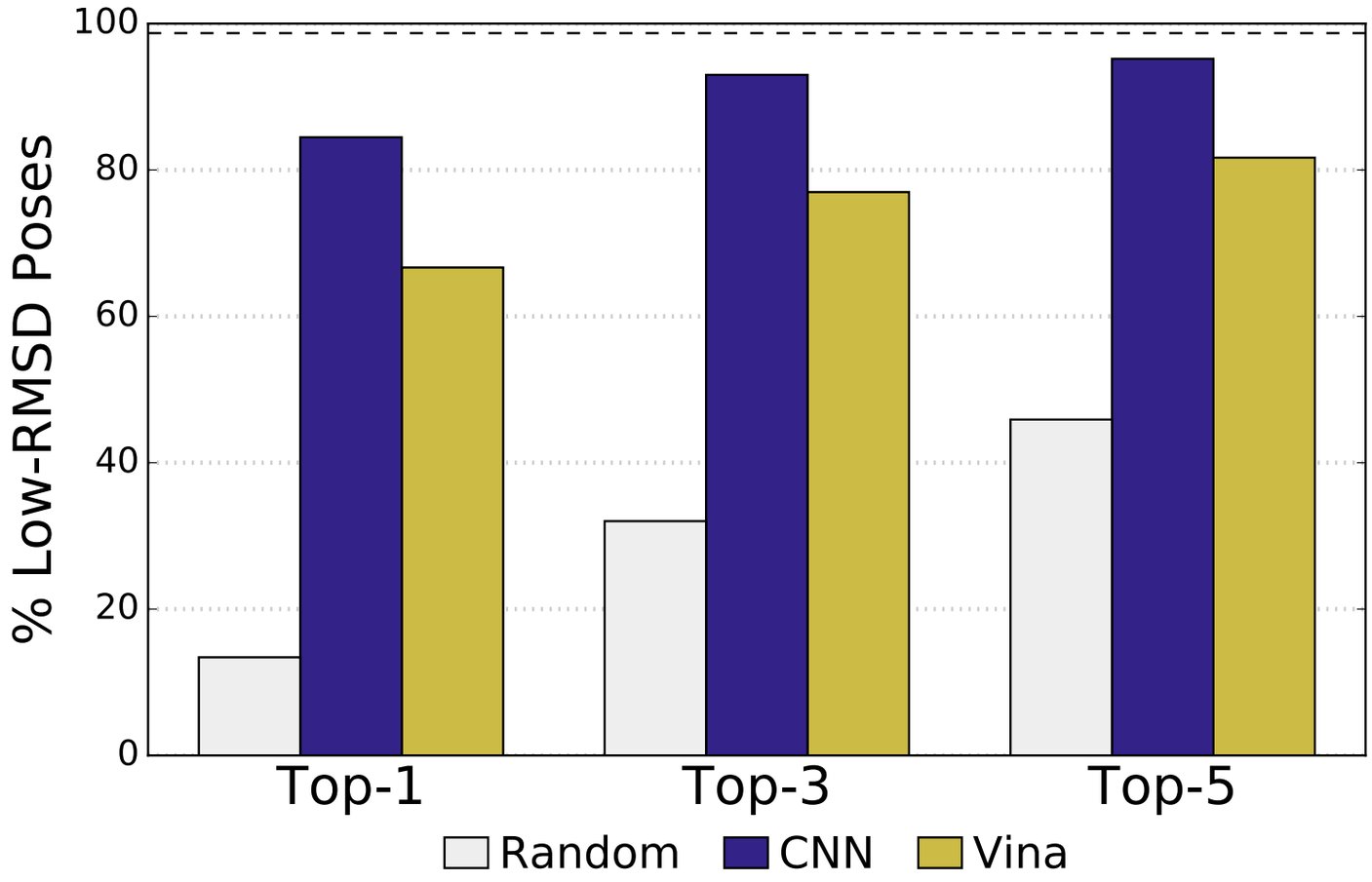


intra-target ranking

Pose Prediction (PDBbind)

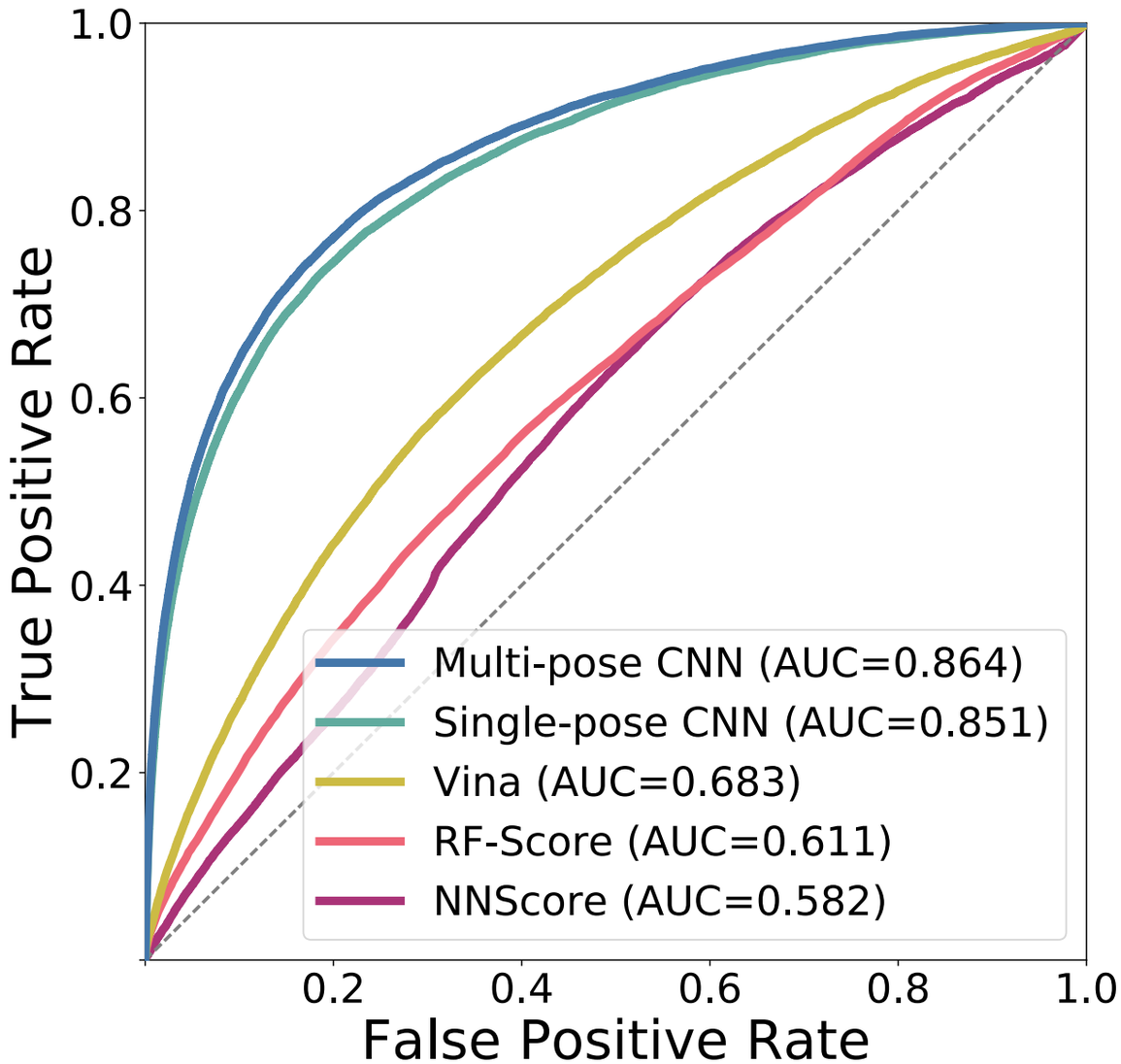


inter-target ranking



intra-target ranking

Binding Determination

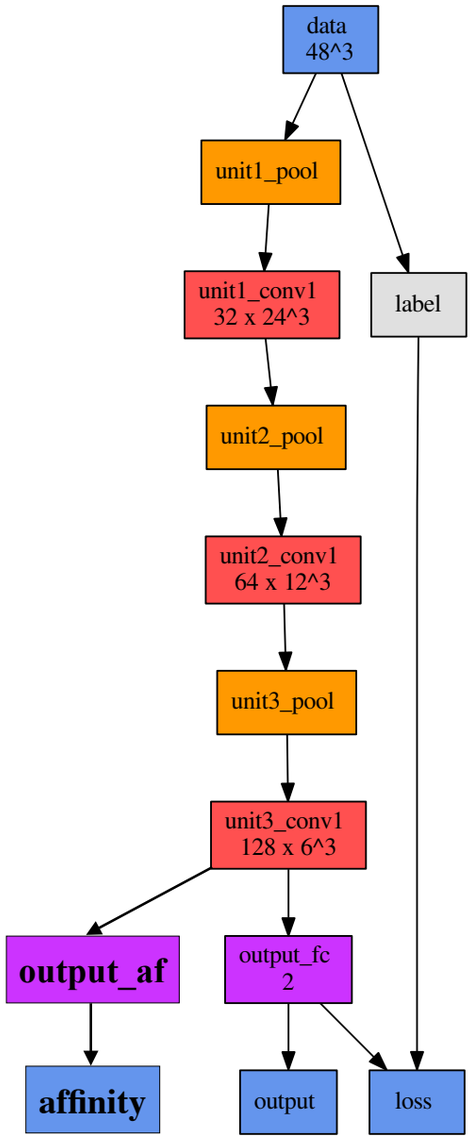


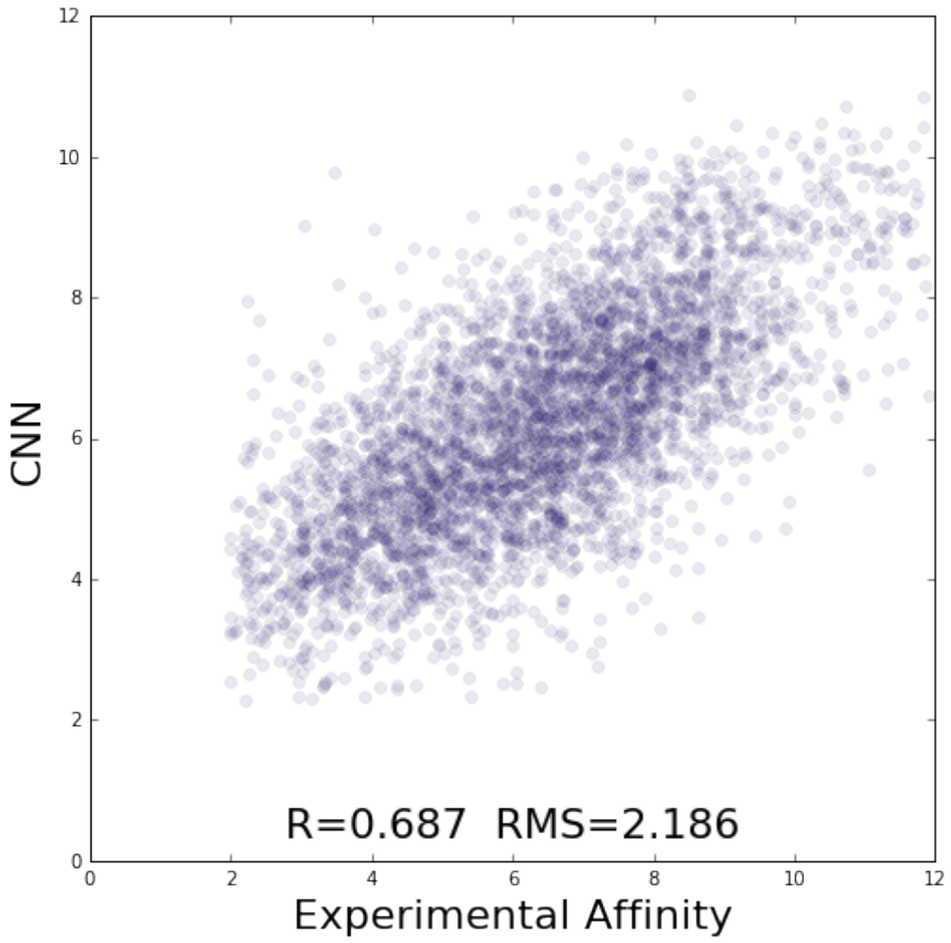
D U D • E

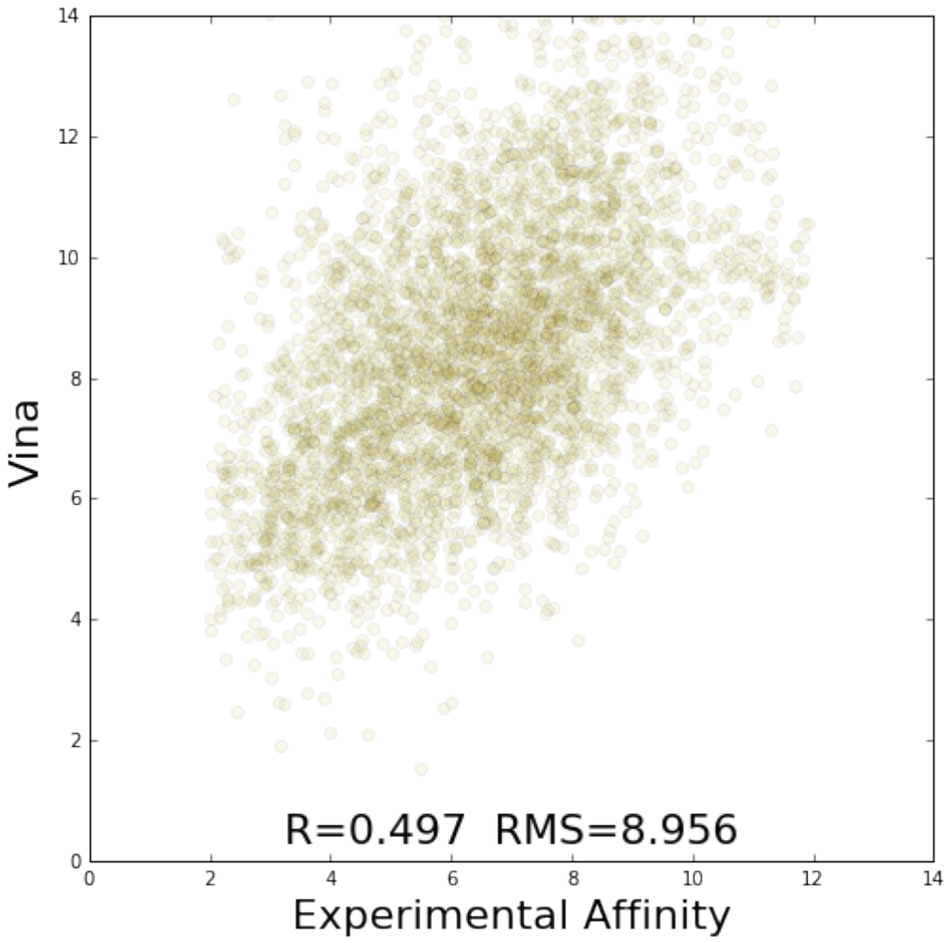
102 targets

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

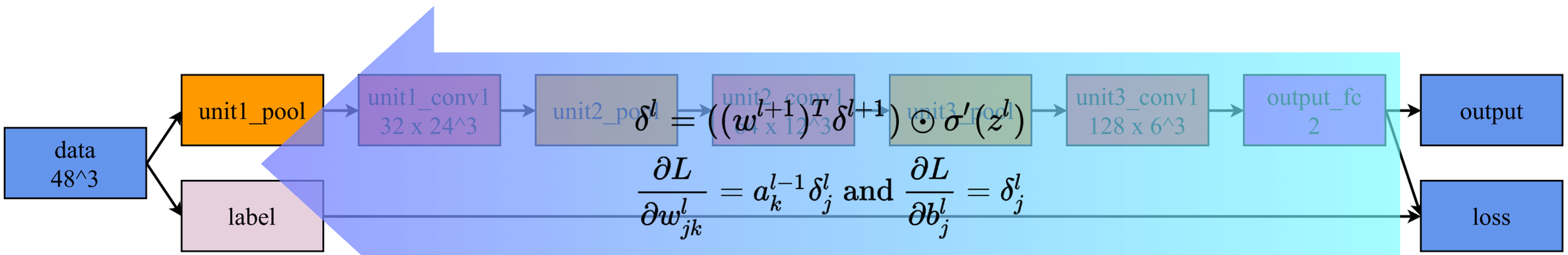
Affinity Prediction



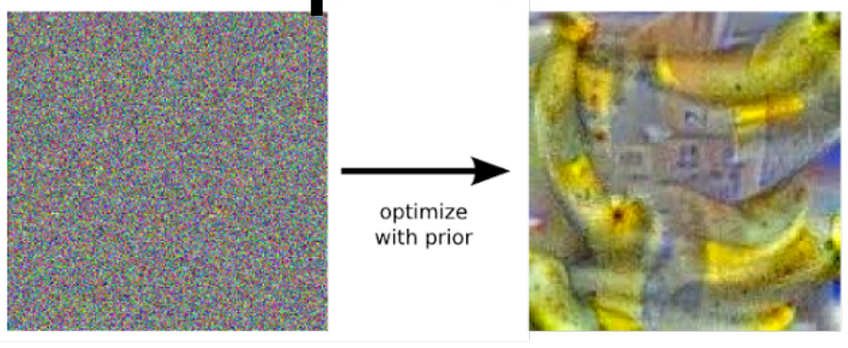




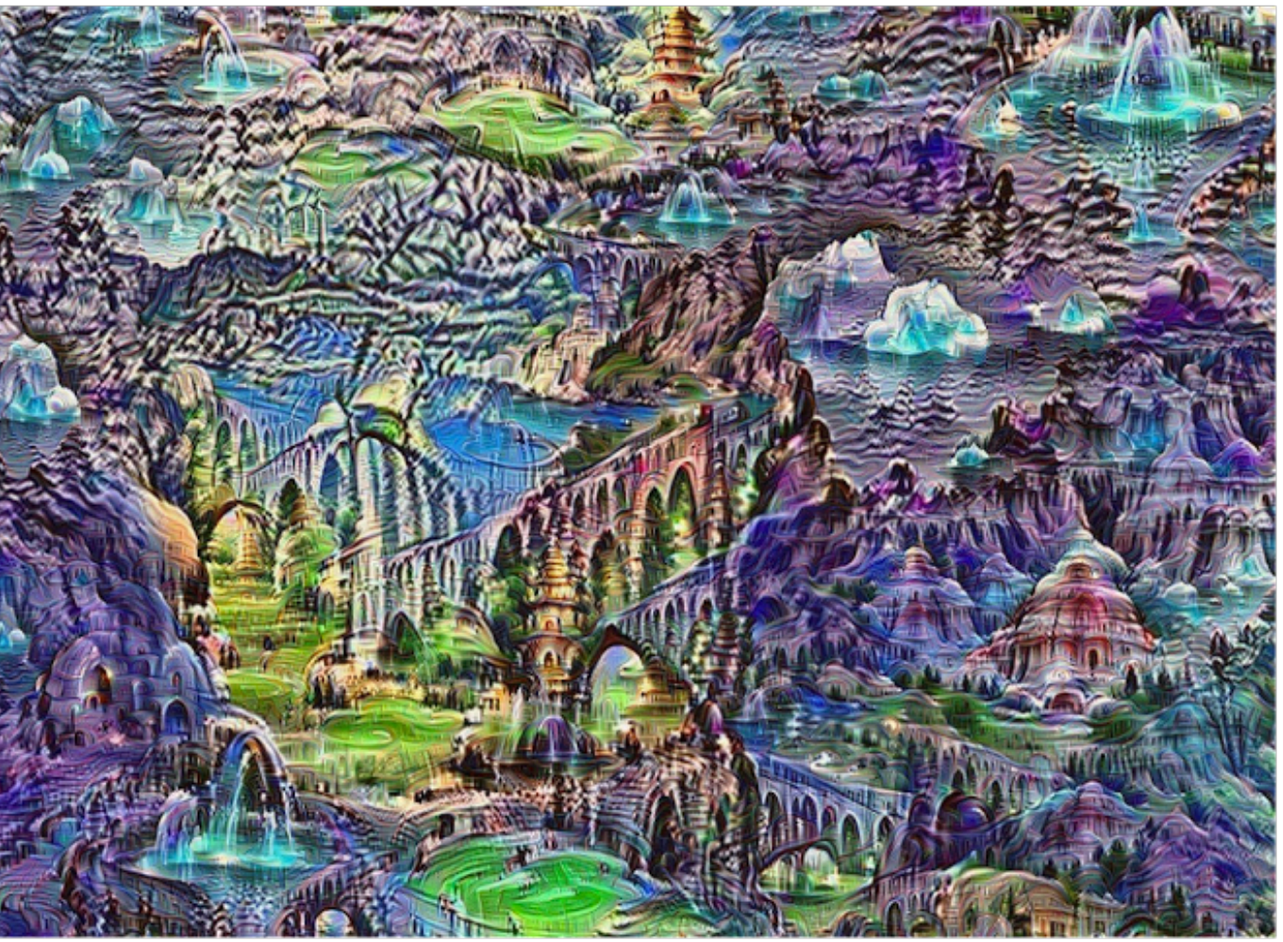
Beyond Scoring

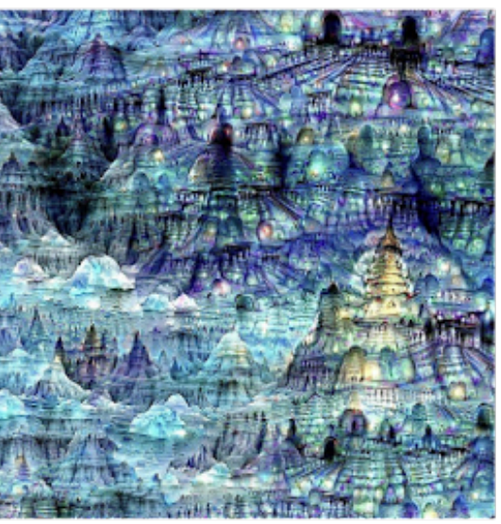


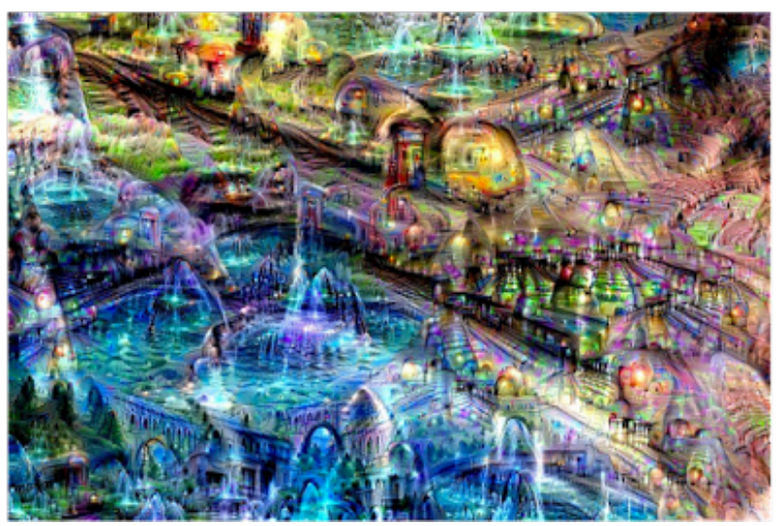
Deep Dreams

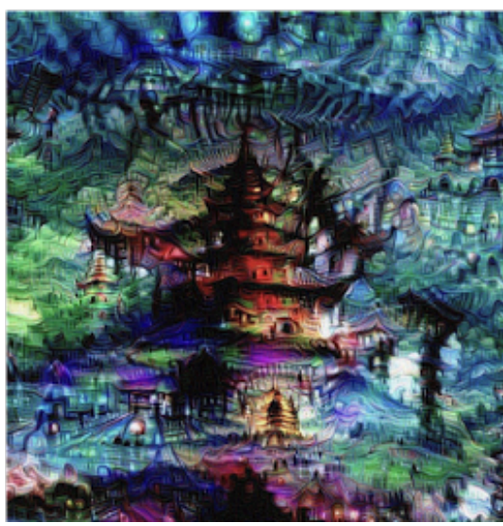


https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

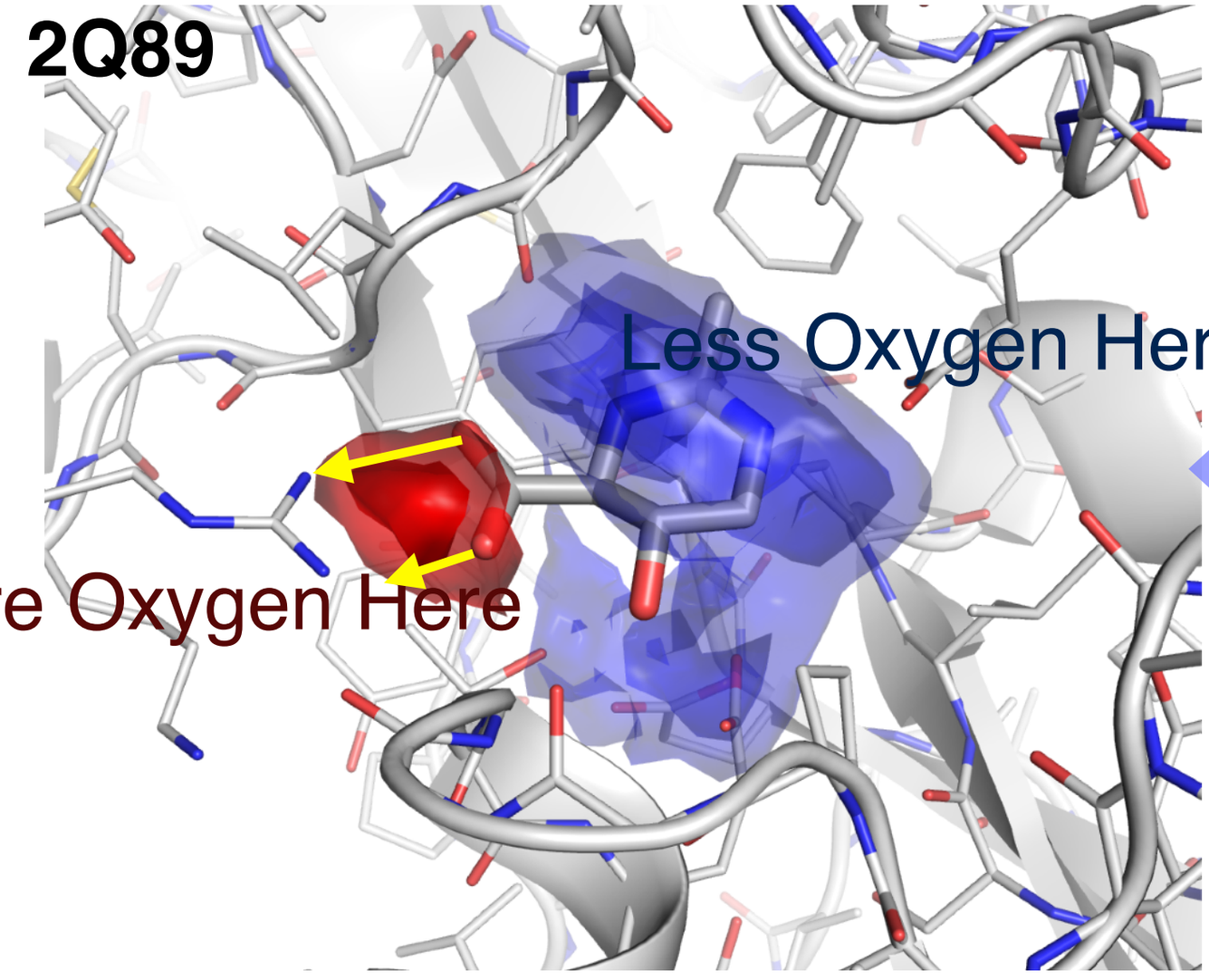








Beyond Scoring

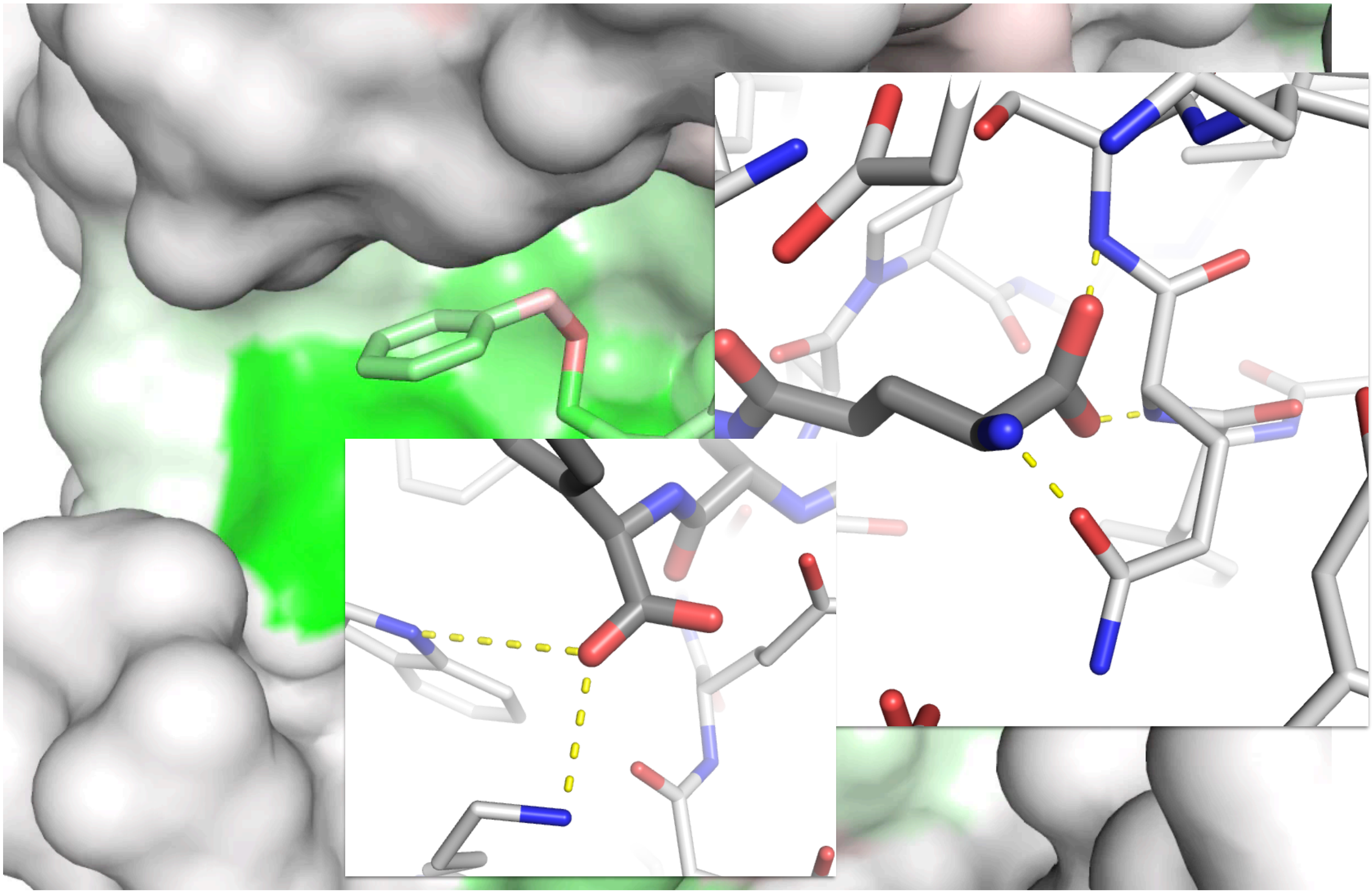


$$\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

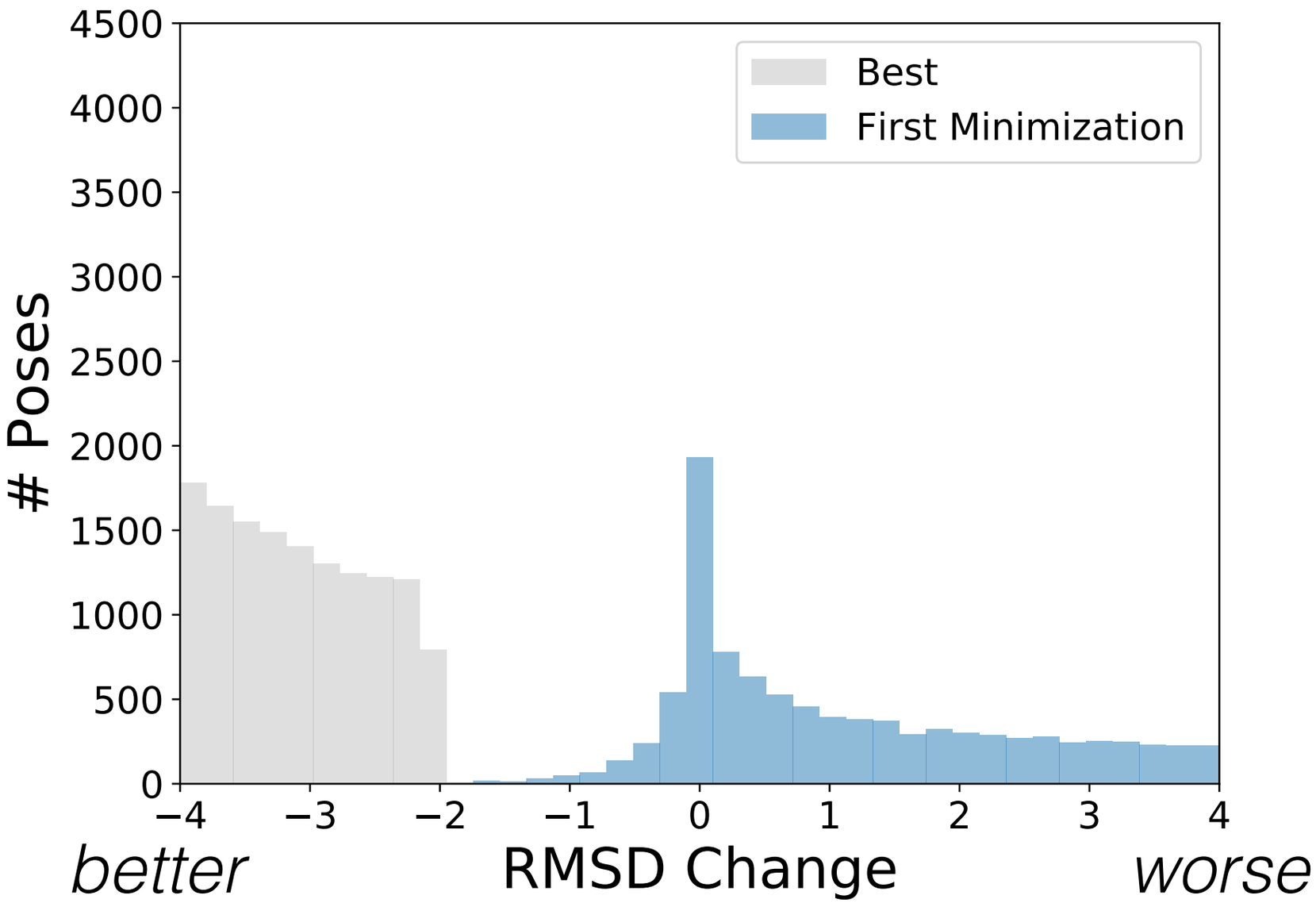
data 48^3

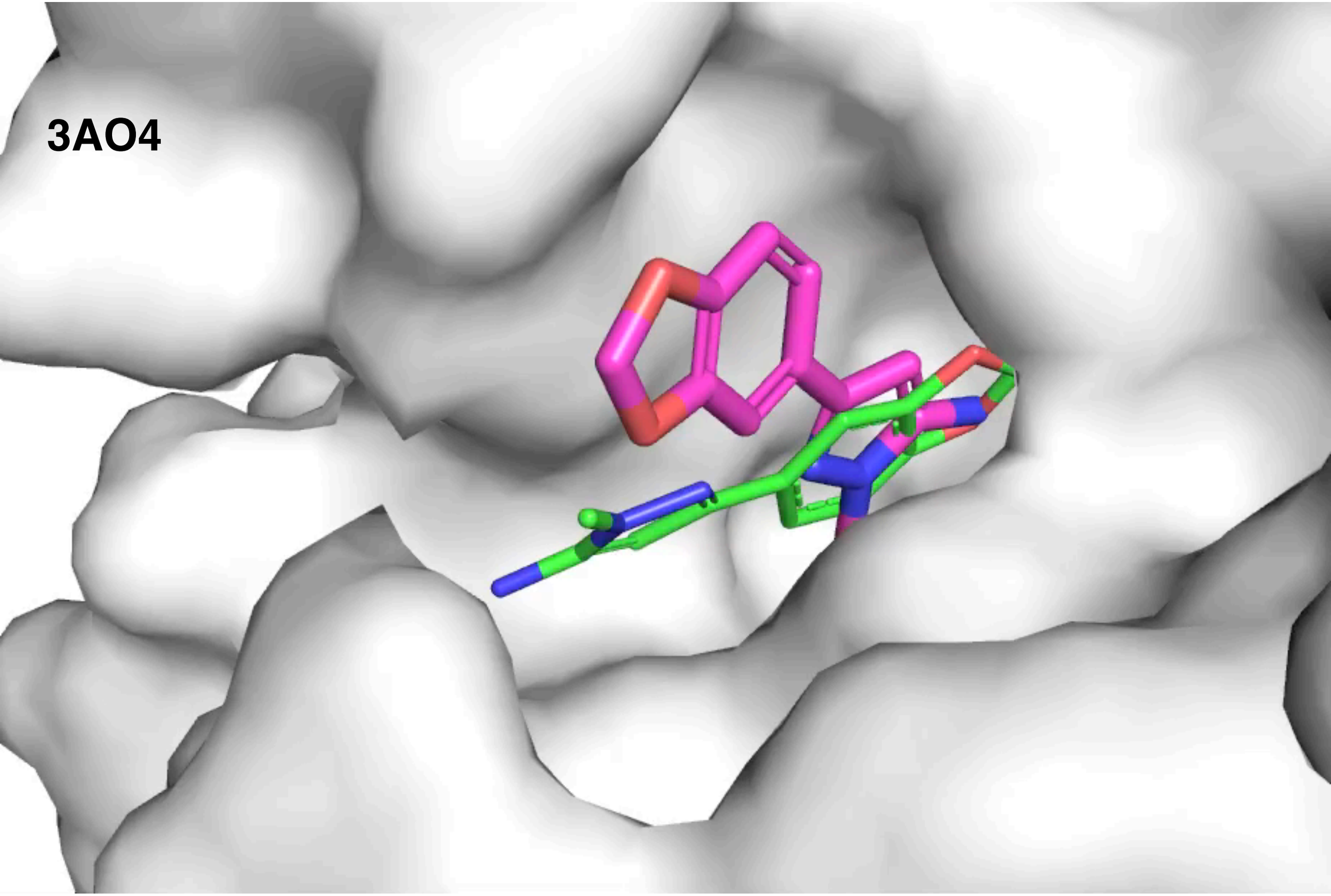
unit1_pool

label

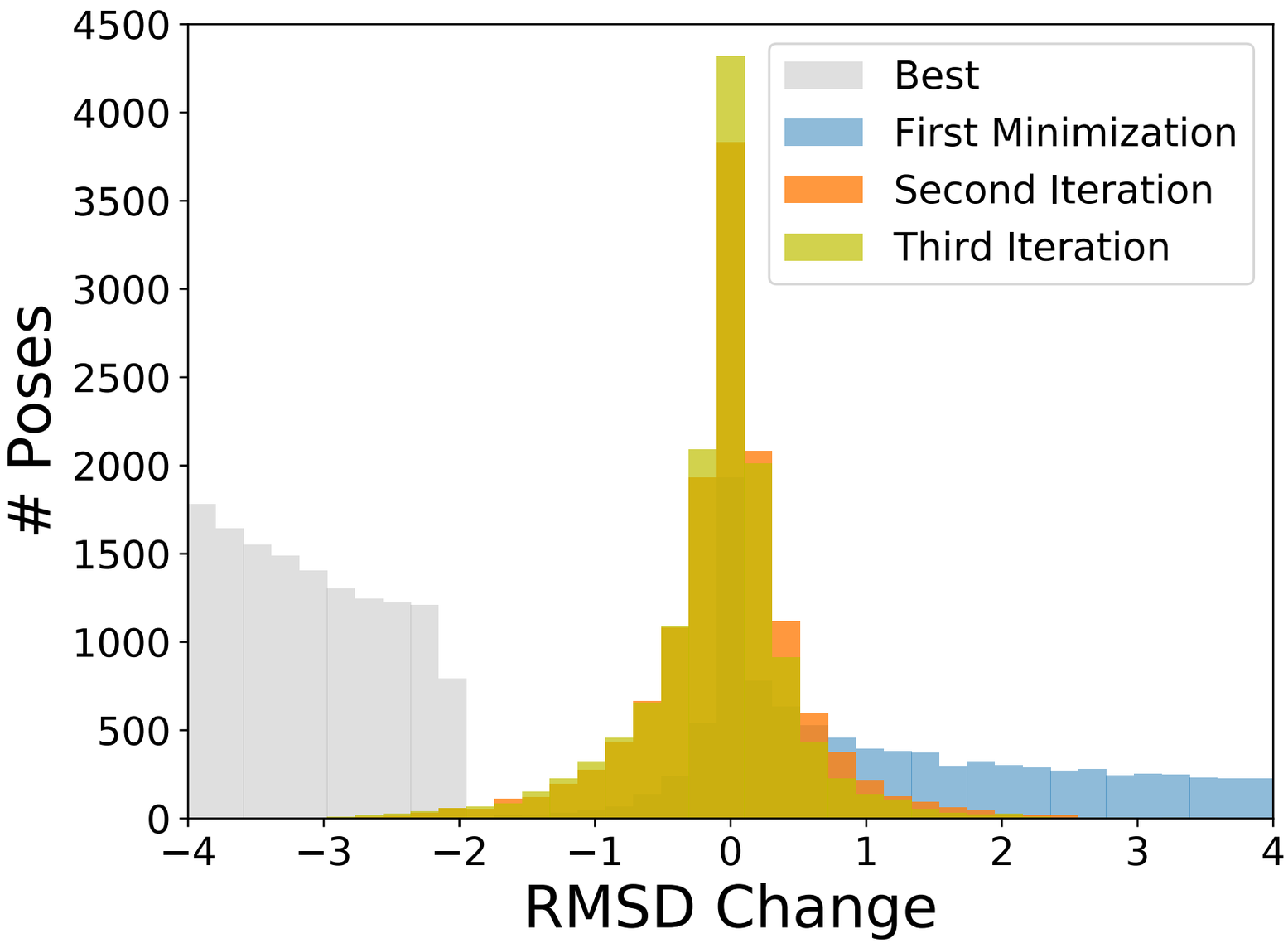


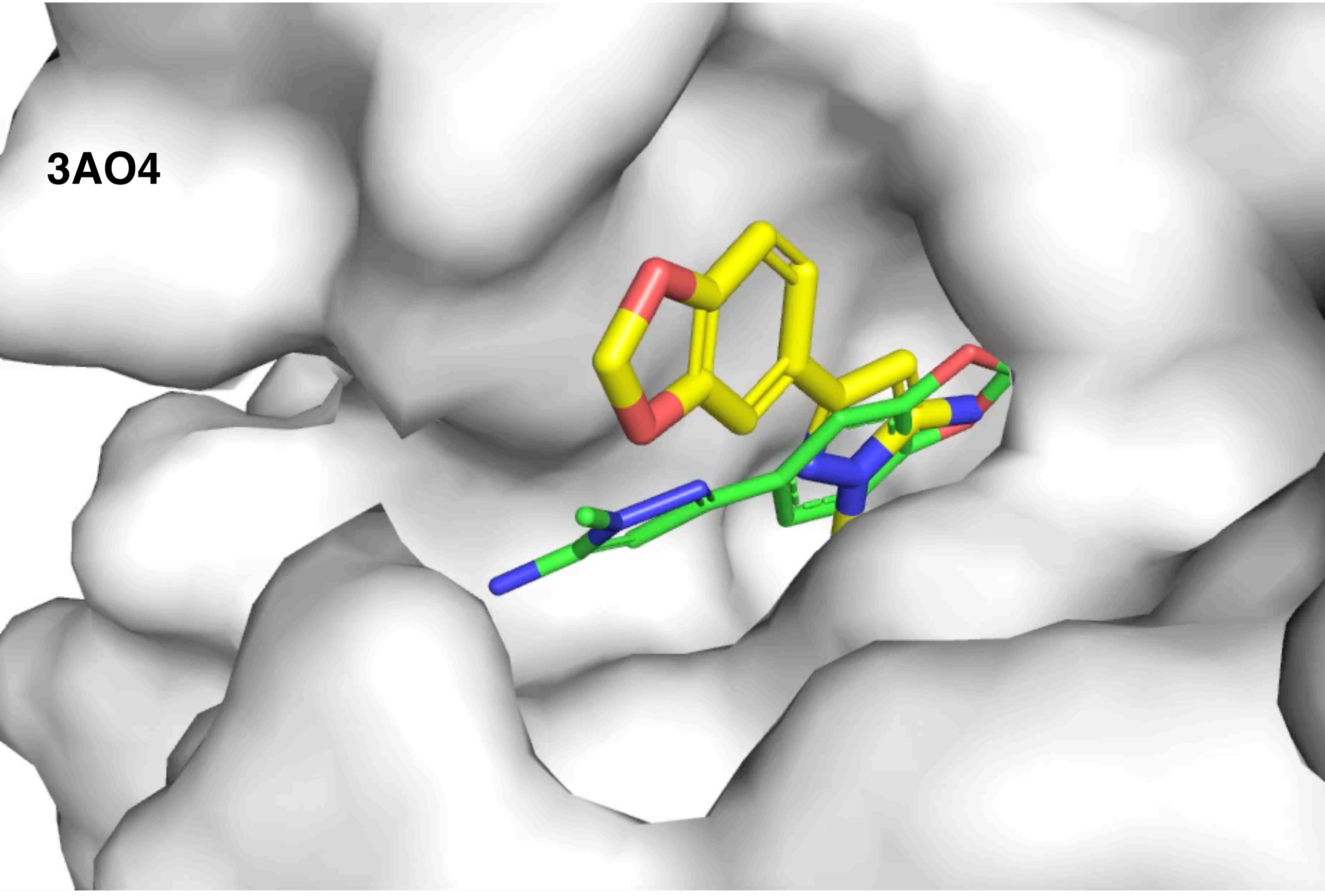
Minimizing Low RMSD Poses




3AO4

Iterative Refinement




3AO4

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

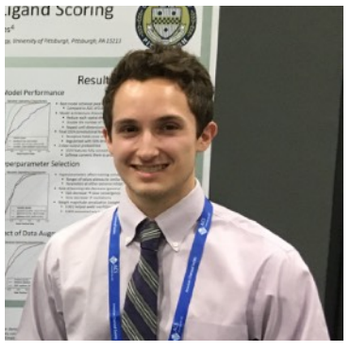
(Submitted on 6 Feb 2015)

Protein–Ligand Scoring with Convolutional Neural Networks

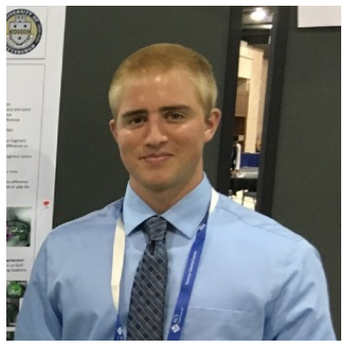
Matthew Ragoza†‡, Joshua Hochuli‡¶, Elisa Idrobo§, Jocelyn Sunseri∥, and David Ryan Koes*∥
†Department of Neuroscience, ‡Department of Computer Science, ¶Department of Biological Sciences, and ∥Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States
§ Department of Computer Science, The College of New Jersey, Ewing, New Jersey 08628, United States

J. Chem. Inf. Model., **2017**, *57* (4), pp 942–957
DOI: 10.1021/acs.jcim.6b00740
Publication Date (Web): April 3, 2017

Acknowledgements

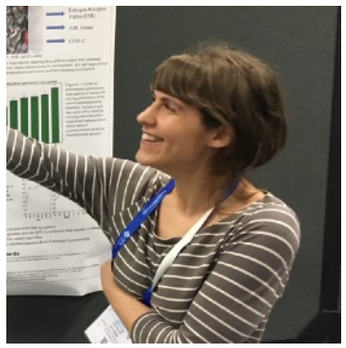


Matt Ragoza



Josh Hochuli

Elisa Idrobo



Jocelyn Sunseri

Group Members

Jocelyn Sunseri
Matt Ragoza
Josh Hochuli
Pulkit Mittal
Alec Helbling
Aaron Zheng
Sharanya Bandla
Faiha Khan
Lily Turner

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