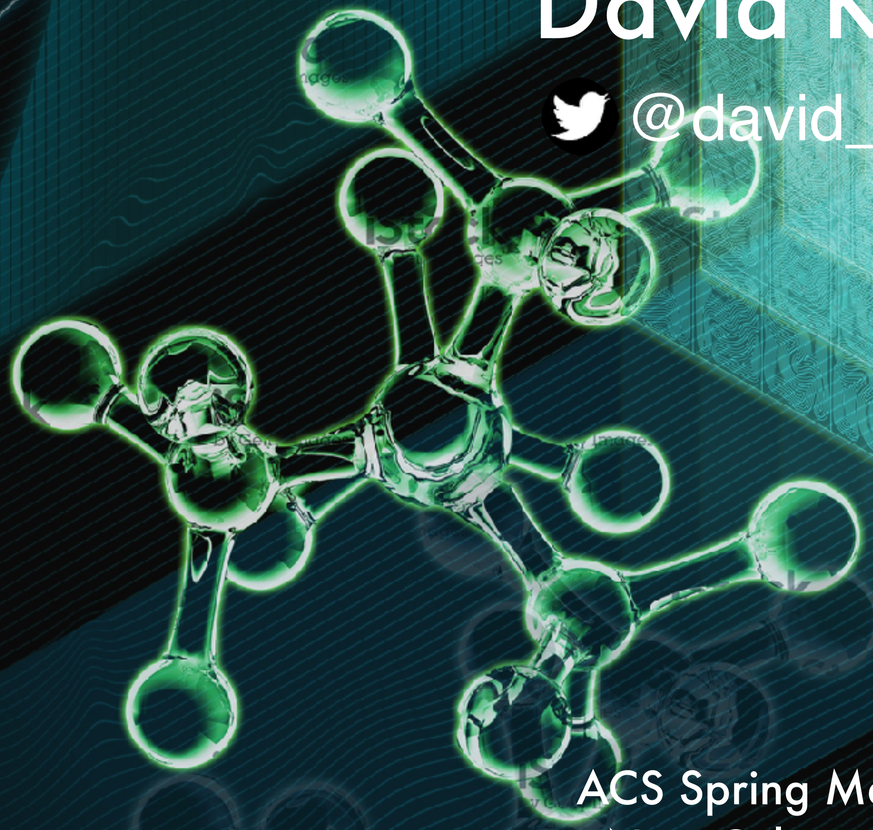
Protein-Ligand Scoring with Convolutional Neural Networks

David Koes

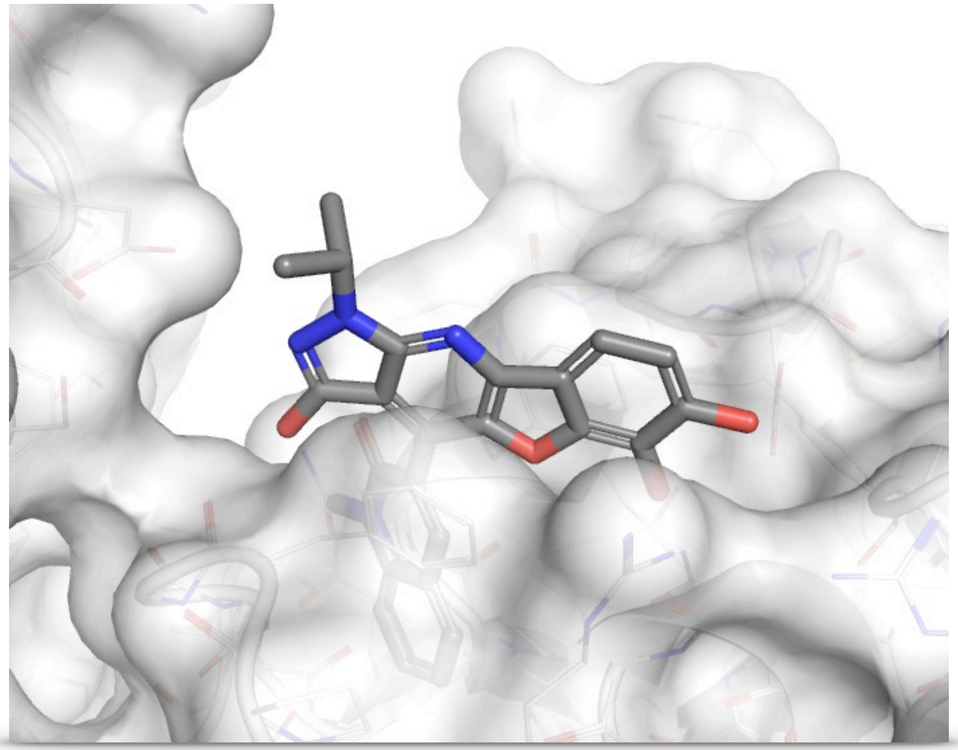
@david_koes



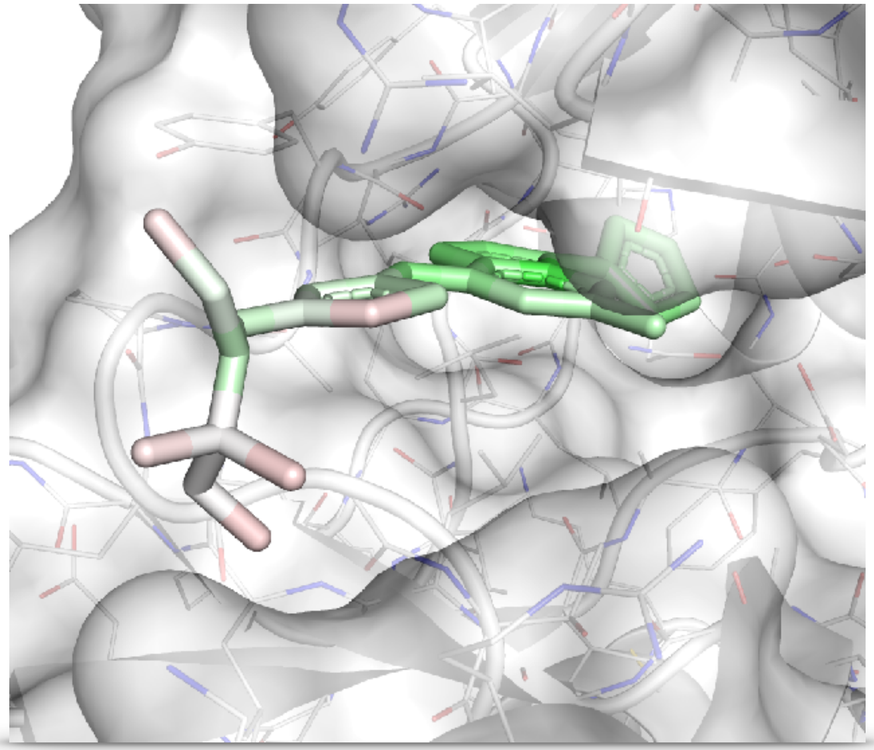
ACS Spring Meeting
New Orleans, LA
March 19, 2018

Structure Based Drug Design

Virtual Screening

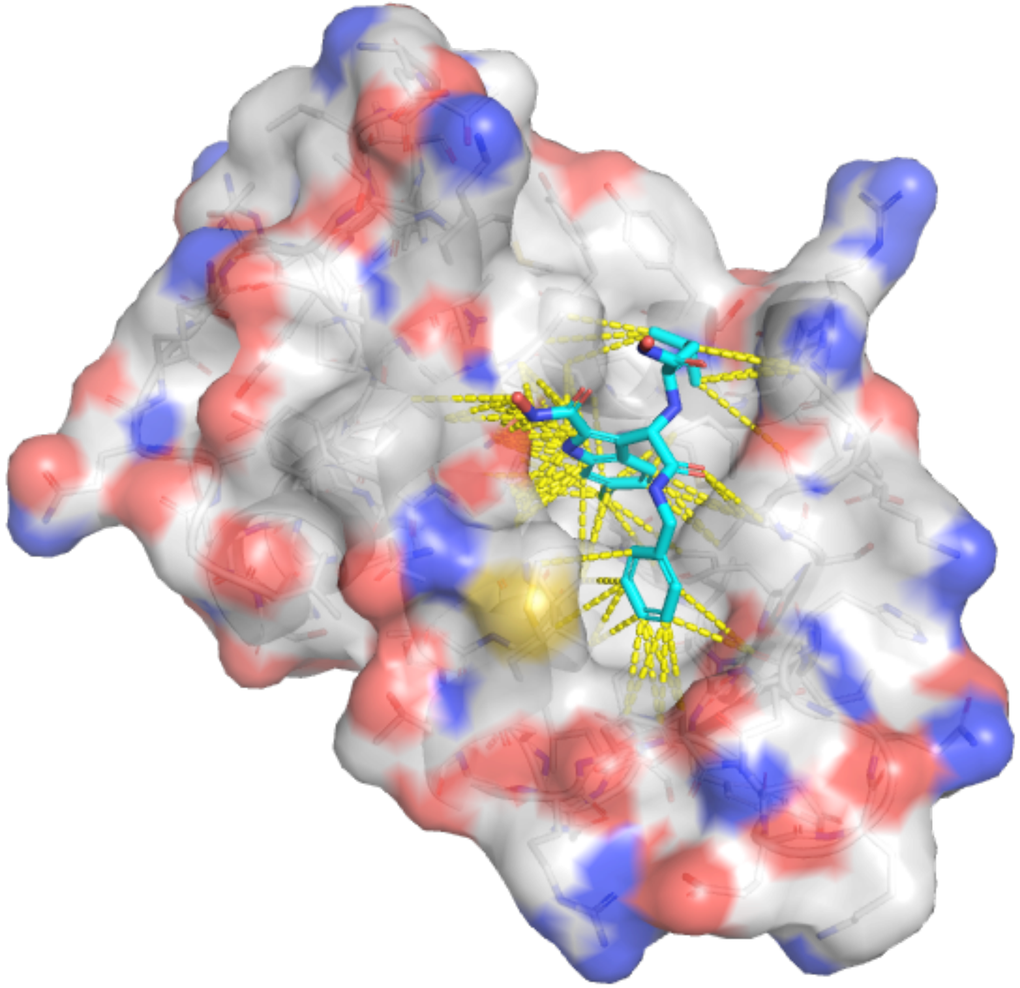


Lead Optimization



Pose Prediction → Binding Discrimination → Affinity Prediction

Protein-Ligand Scoring



AutoDock Vina

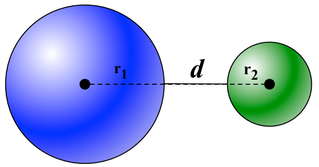
$$\text{gauss}_1(d) = w_{\text{guass}_1} e^{-(d/0.5)^2}$$

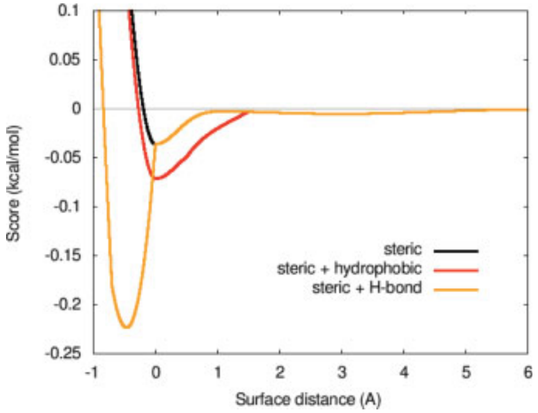
$$\text{gauss}_2(d) = w_{\text{guass}_2} e^{-((d-3)/2)^2}$$

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

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

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





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

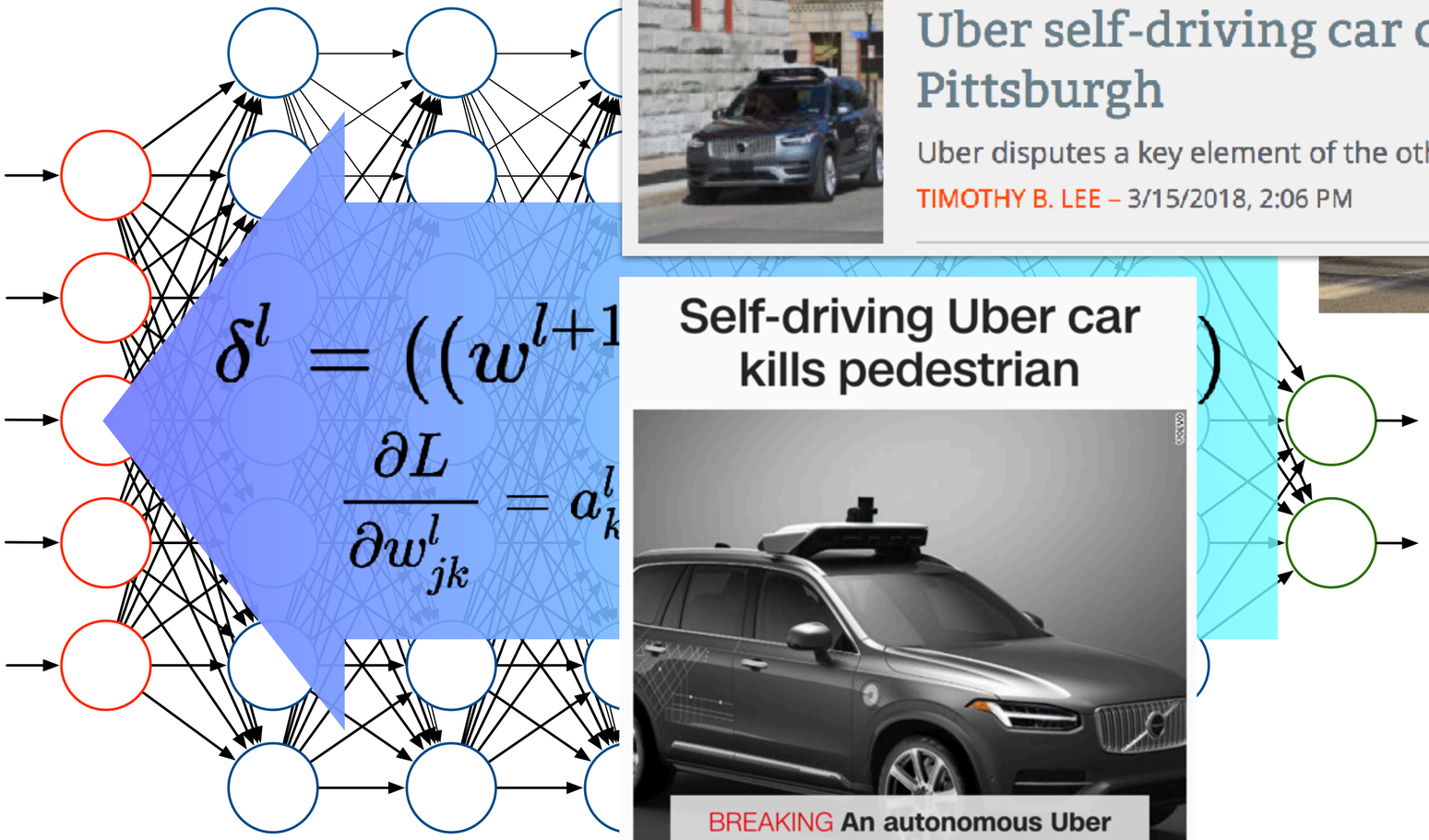
Can we do better?

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



$$\delta^l = ((w^{l+1}$$

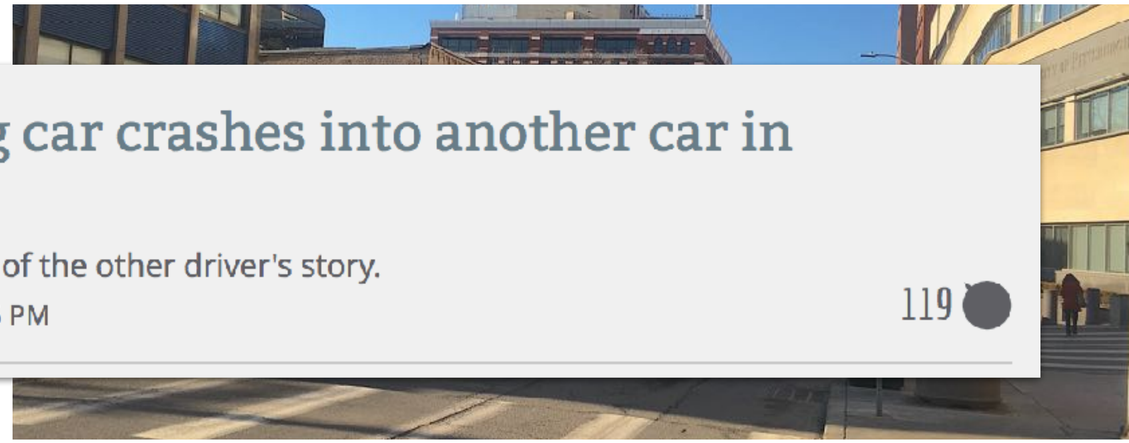
$$\frac{\partial L}{\partial w^l_{jk}} = a^l_k$$

Uber self-driving car crashes into another car in Pittsburgh

Uber disputes a key element of the other driver's story.

TIMOTHY B. LEE – 3/15/2018, 2:06 PM

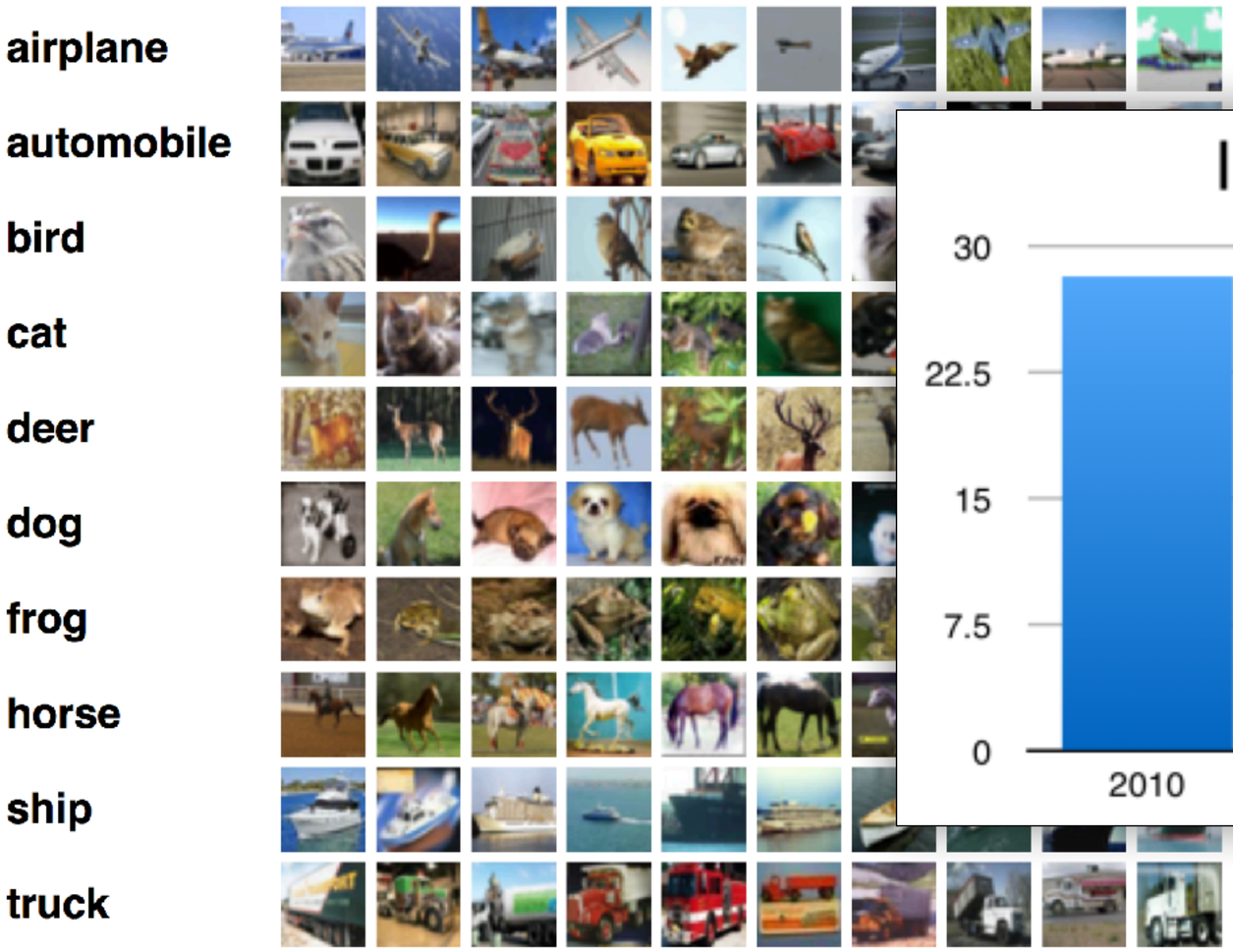
119

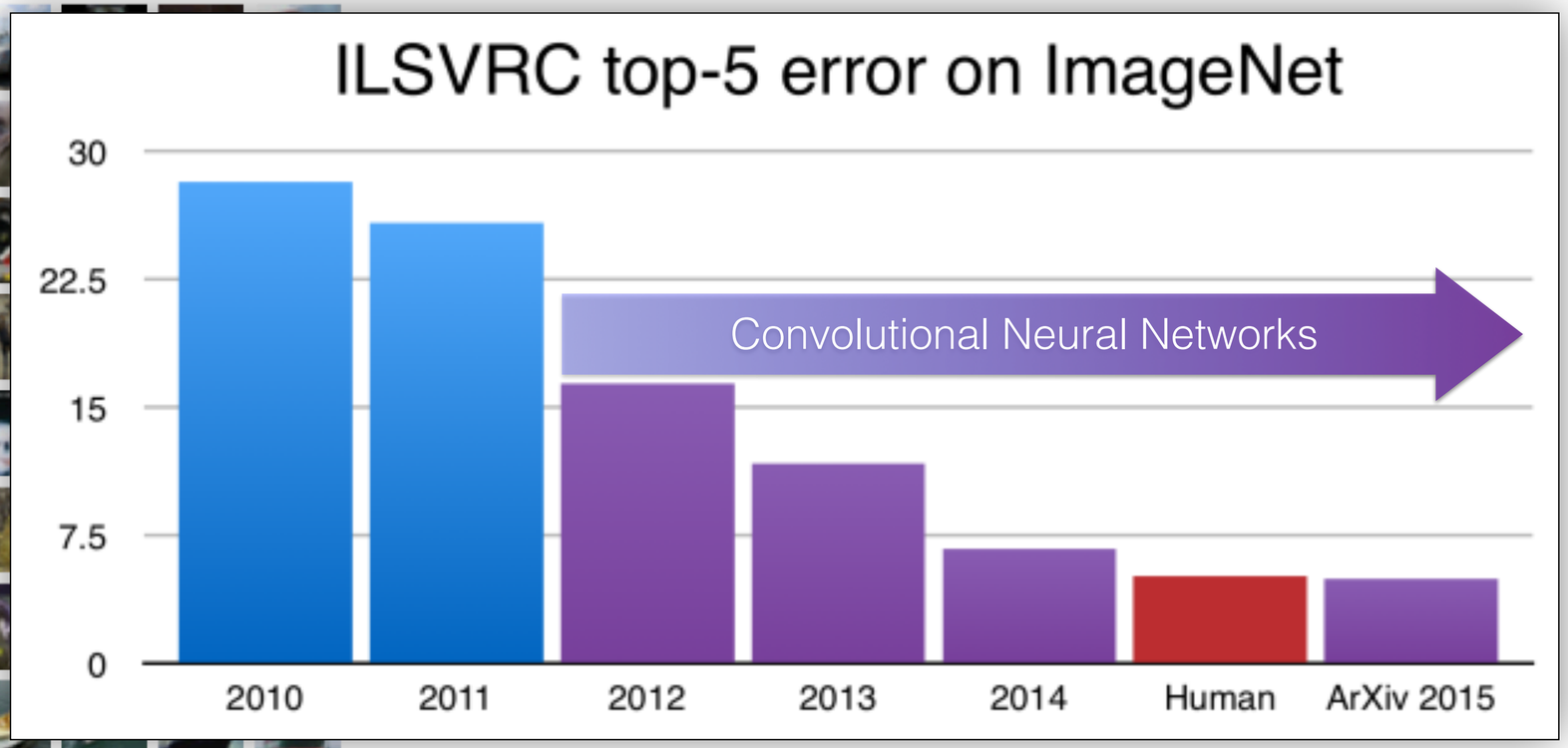


Self-driving Uber car kills pedestrian

BREAKING An autonomous Uber SUV struck and killed a pedestrian in Tempe, Arizona, Sunday night, Tempe police say

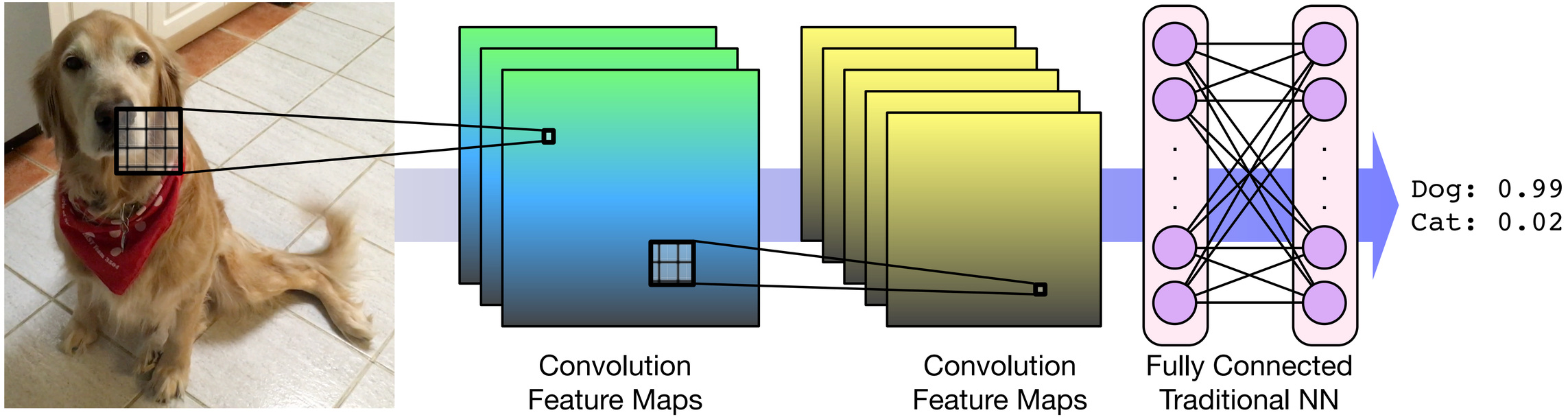
Image Recognition

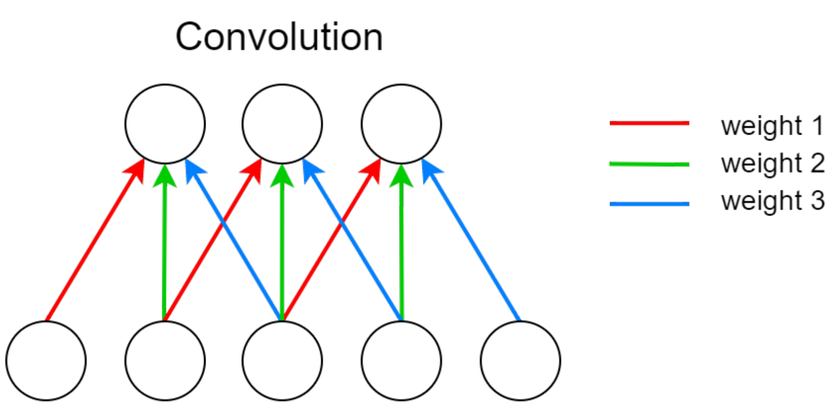


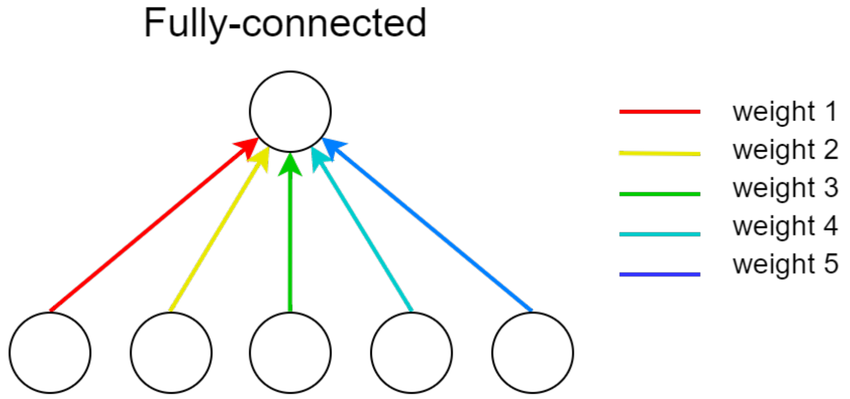


https://devblogs.nvidia.com

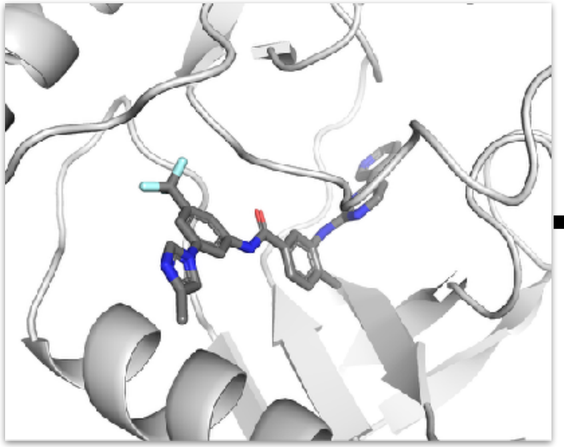
Convolutional Neural Networks







CNNs for Protein-Ligand Scoring



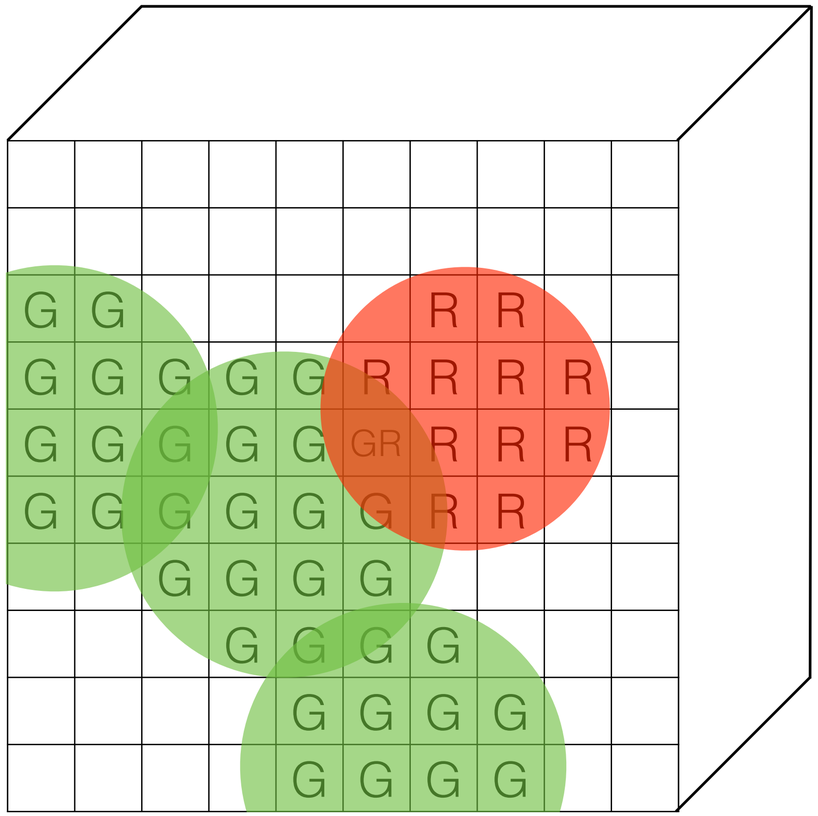
CNN

Pose Prediction

Binding Discrimination

Affinity Prediction

Protein-Ligand Representation



(R,G,B) pixel →

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

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

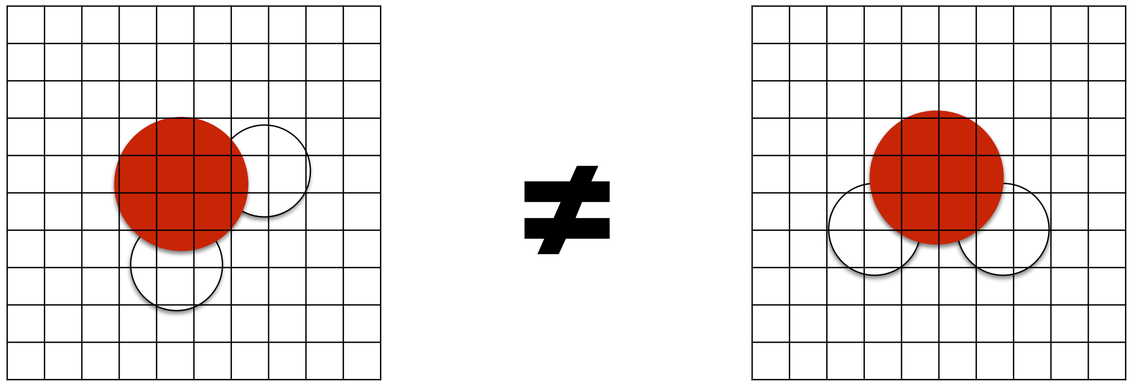
Why Grids?

Pros

- clear spatial relationships
- amazingly parallel
- easy to interpret

Cons

- *coordinate frame dependent*
- pairwise interactions not explicit



Training Data

Pose Prediction

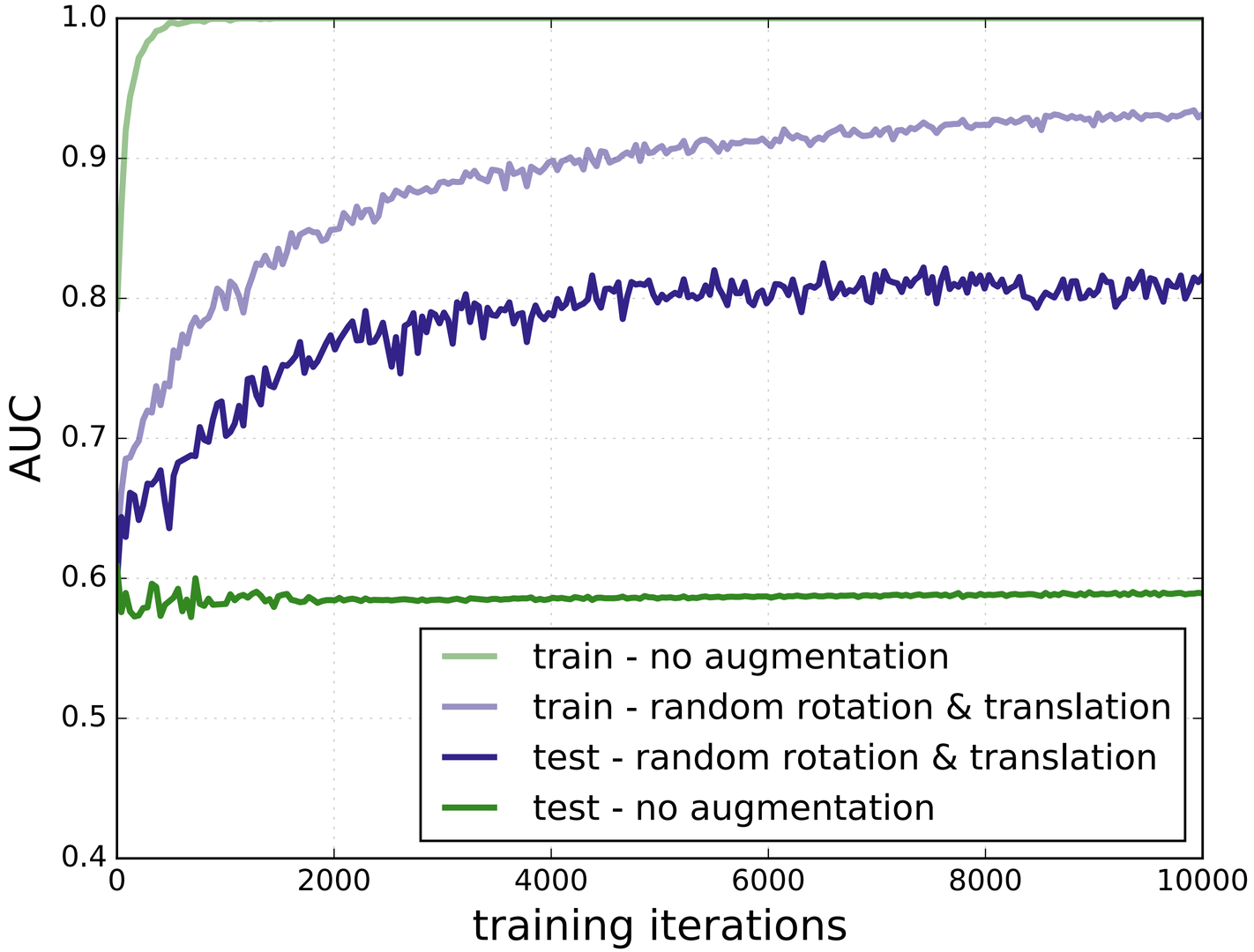
4056 protein-ligand complexes

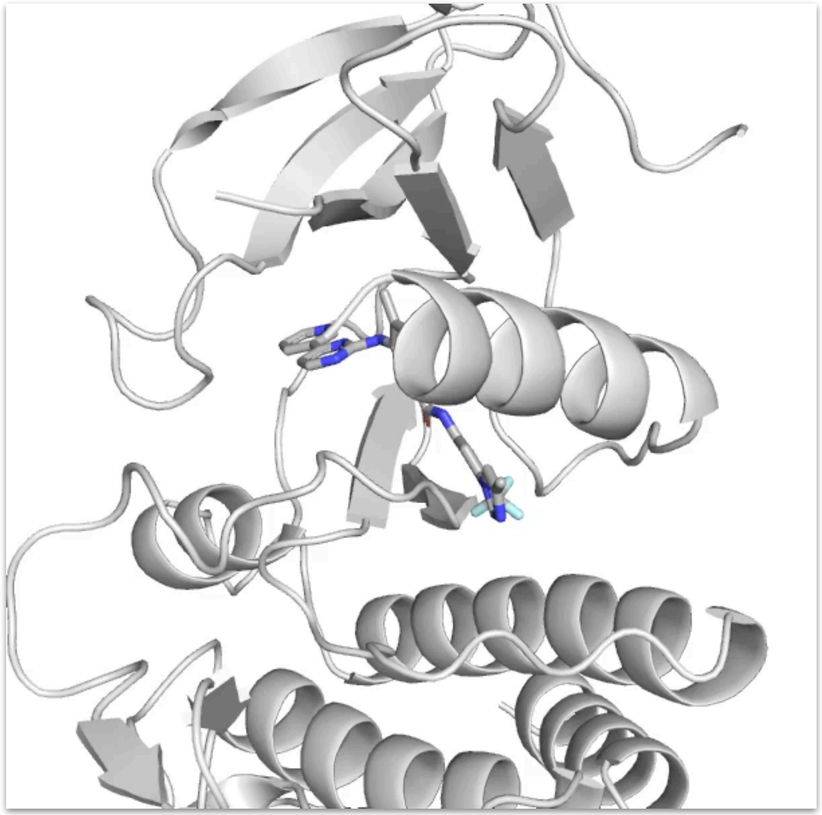
- diverse targets
- wide range of affinities
- generate poses with AutoDock Vina
- include minimized crystal pose

Affinity Prediction

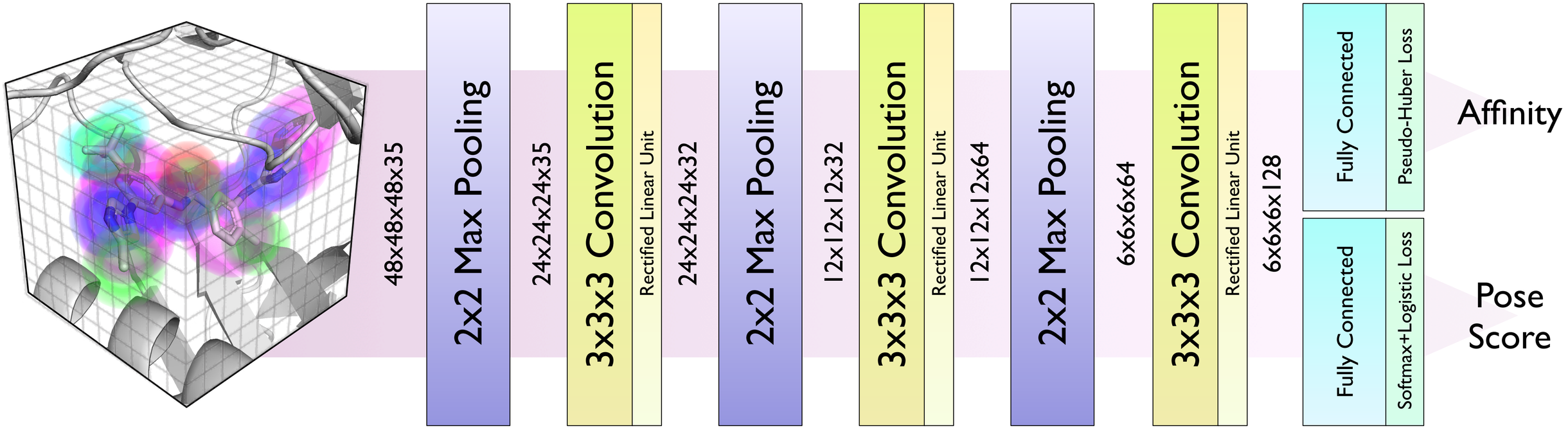
- 8,688 low RMSD poses
- assign known affinity
- **regression problem**

Data Augmentation

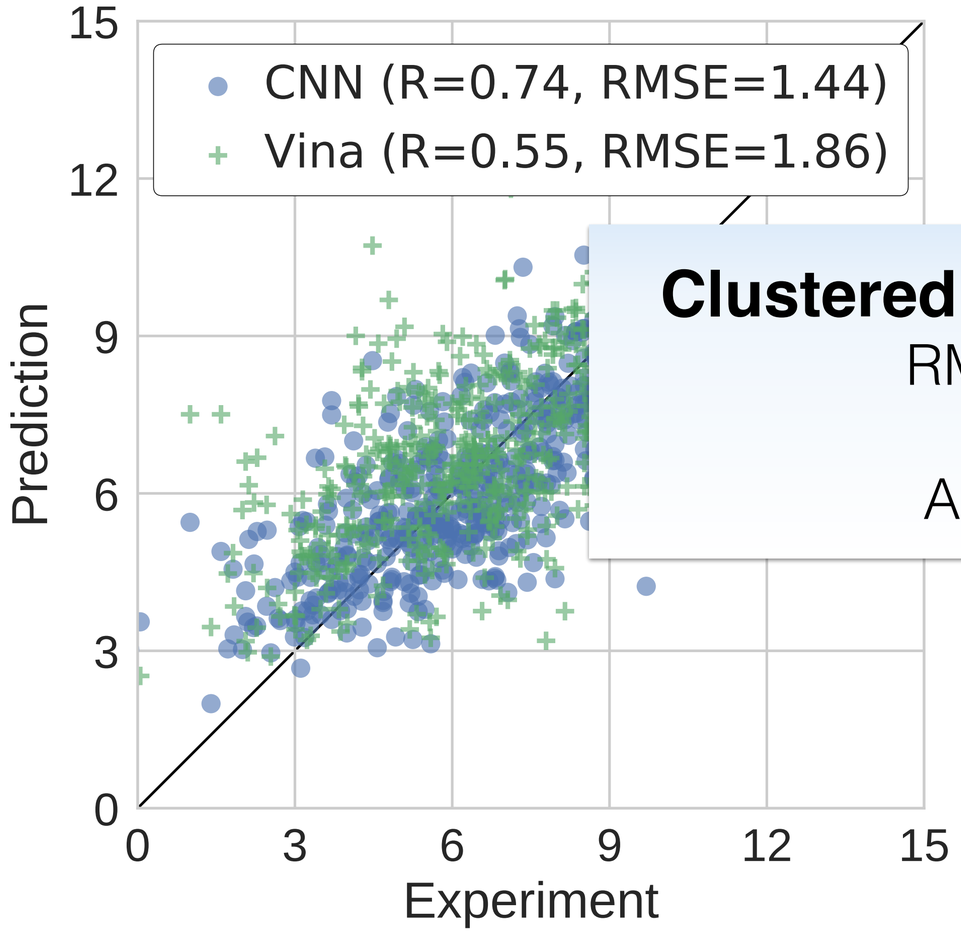




Model


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

Results

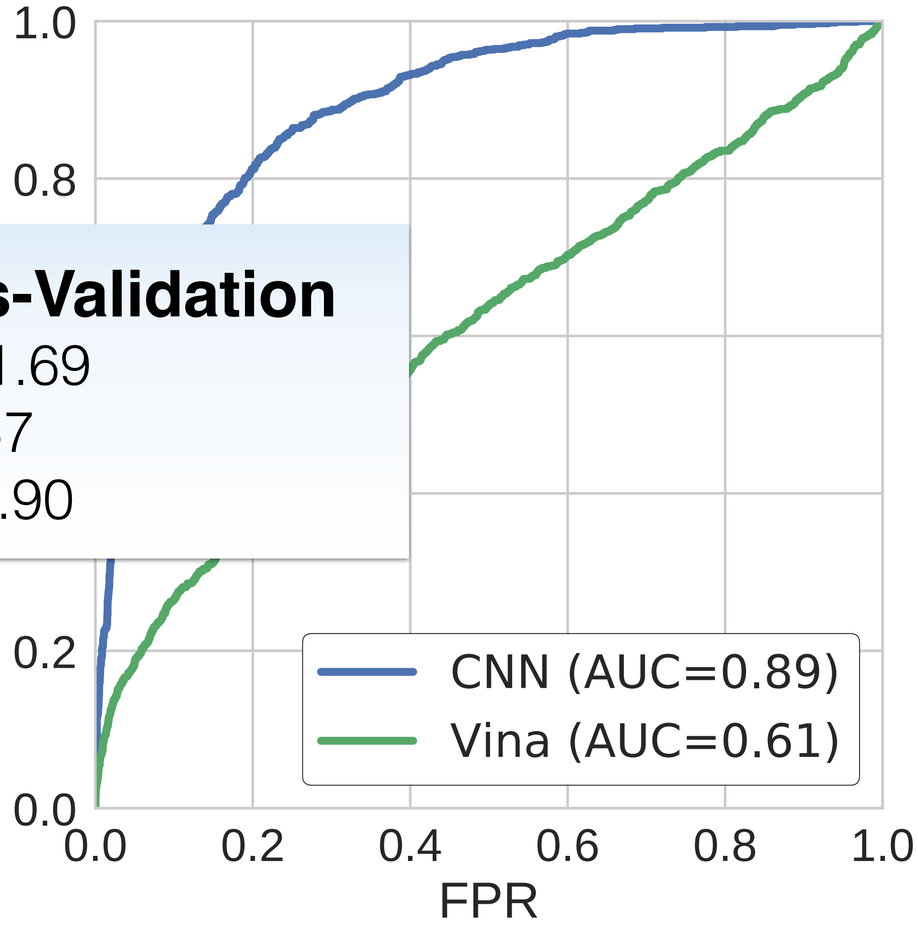


Clustered Cross-Validation

RMSE = 1.69

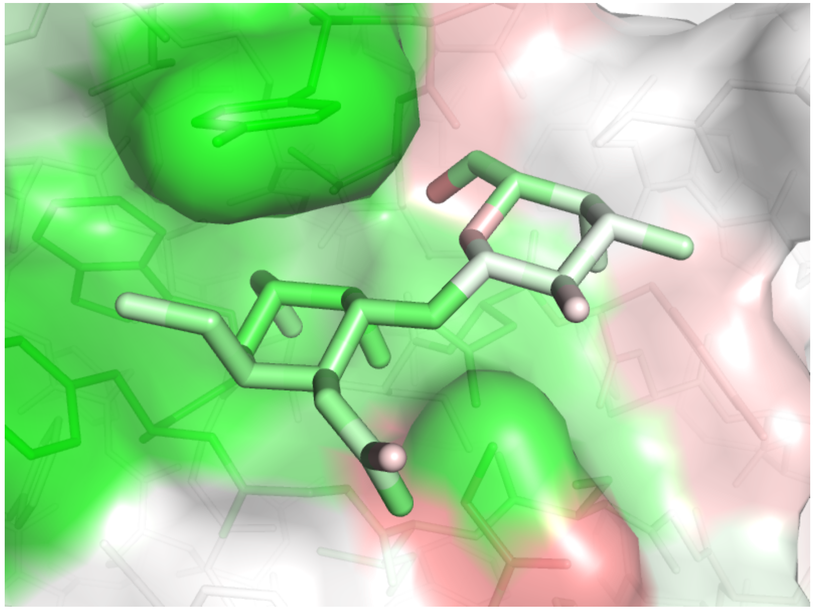
R = 0.57

AUC = 0.90

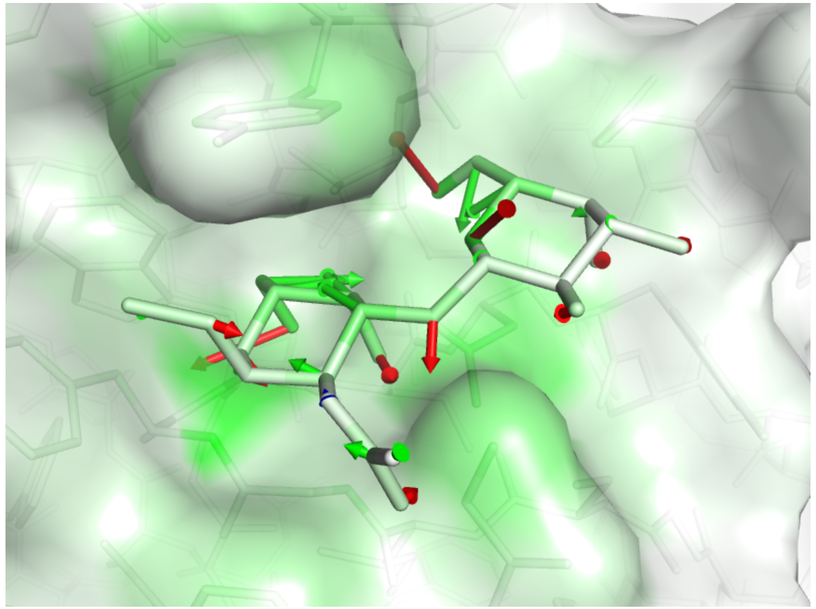


Trained on PDBbind refined; tested on CSAR

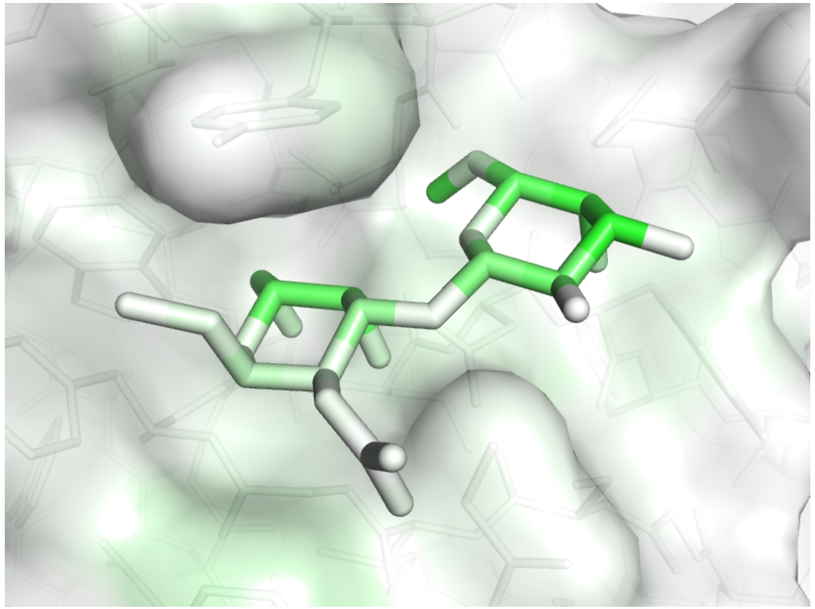
Visualization



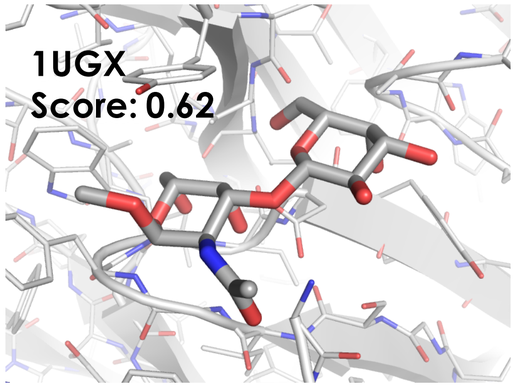
masking



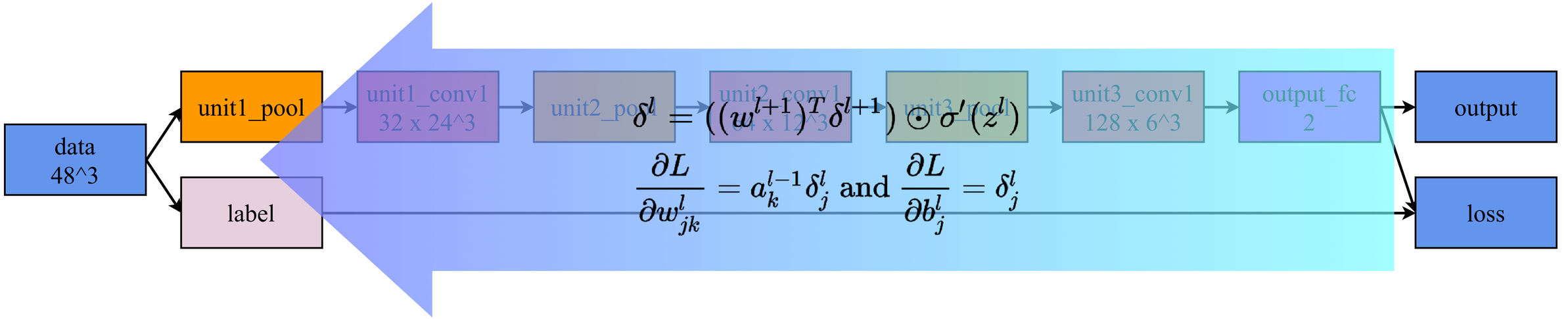
gradients



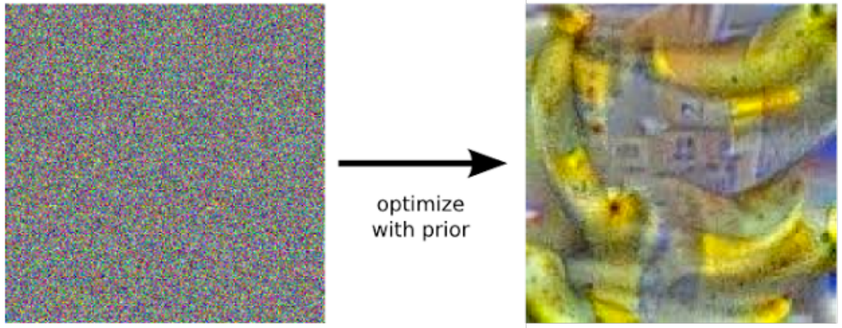
layer-wise relevance



Beyond Scoring

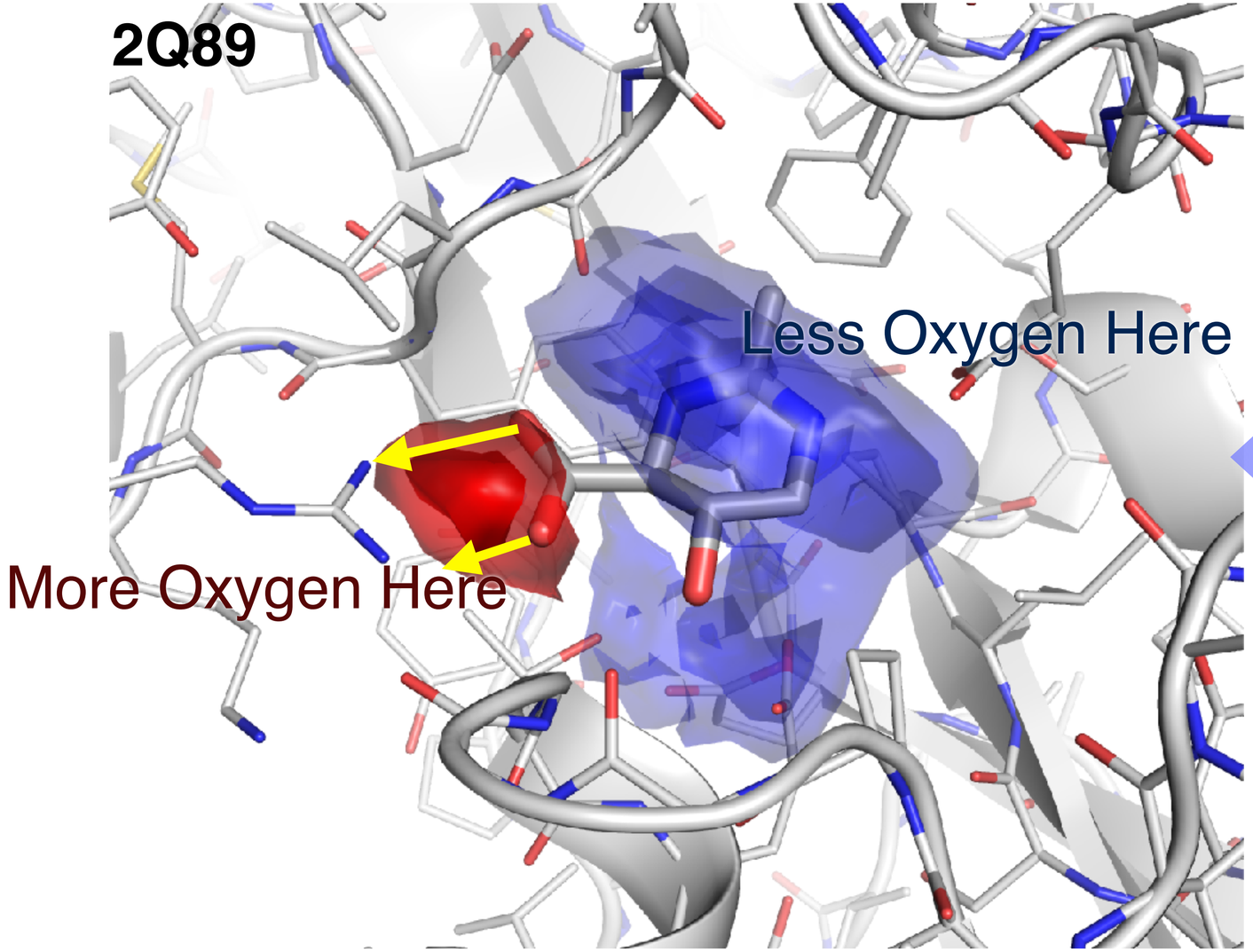


Deep Dreams

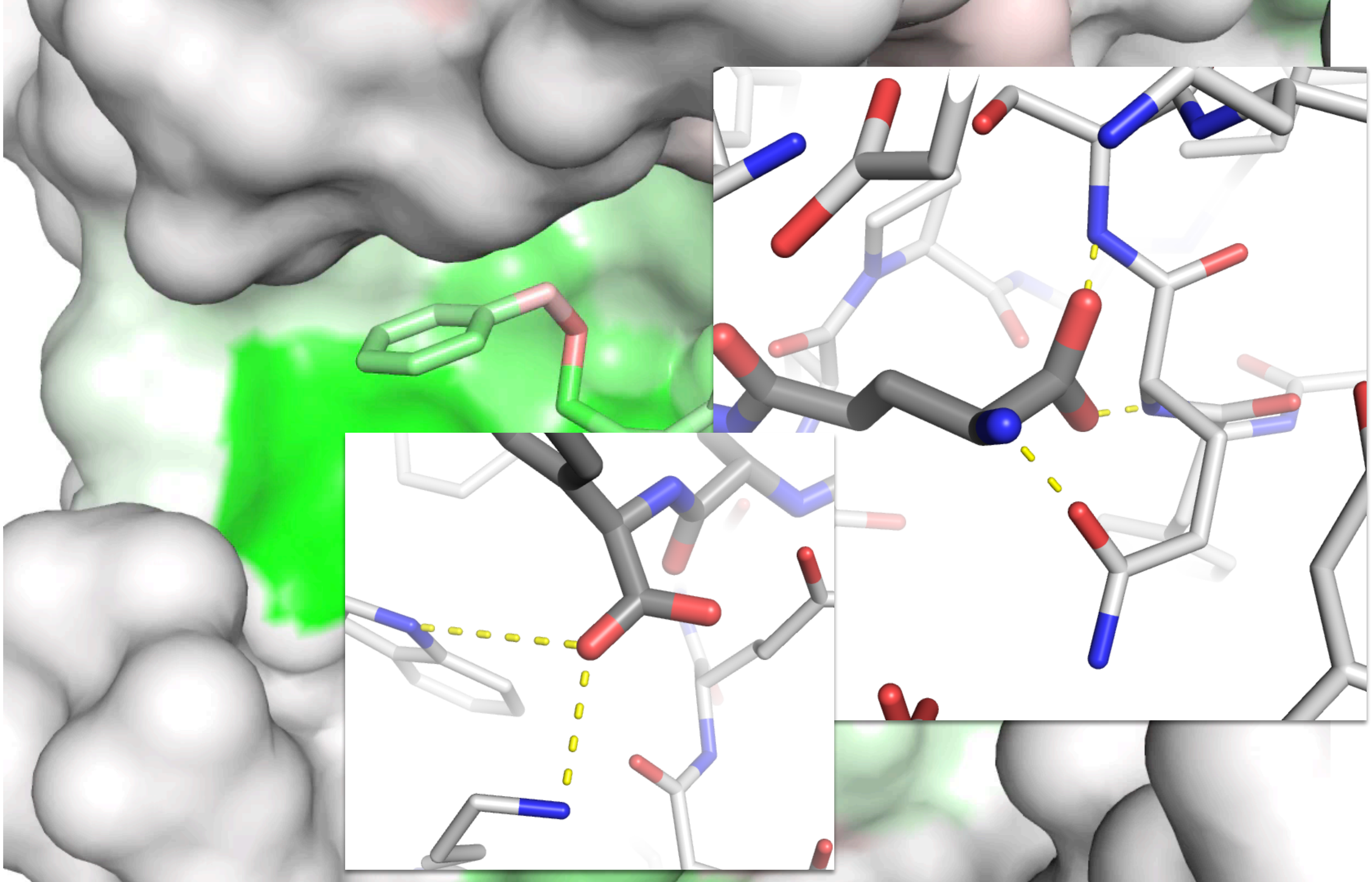


https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

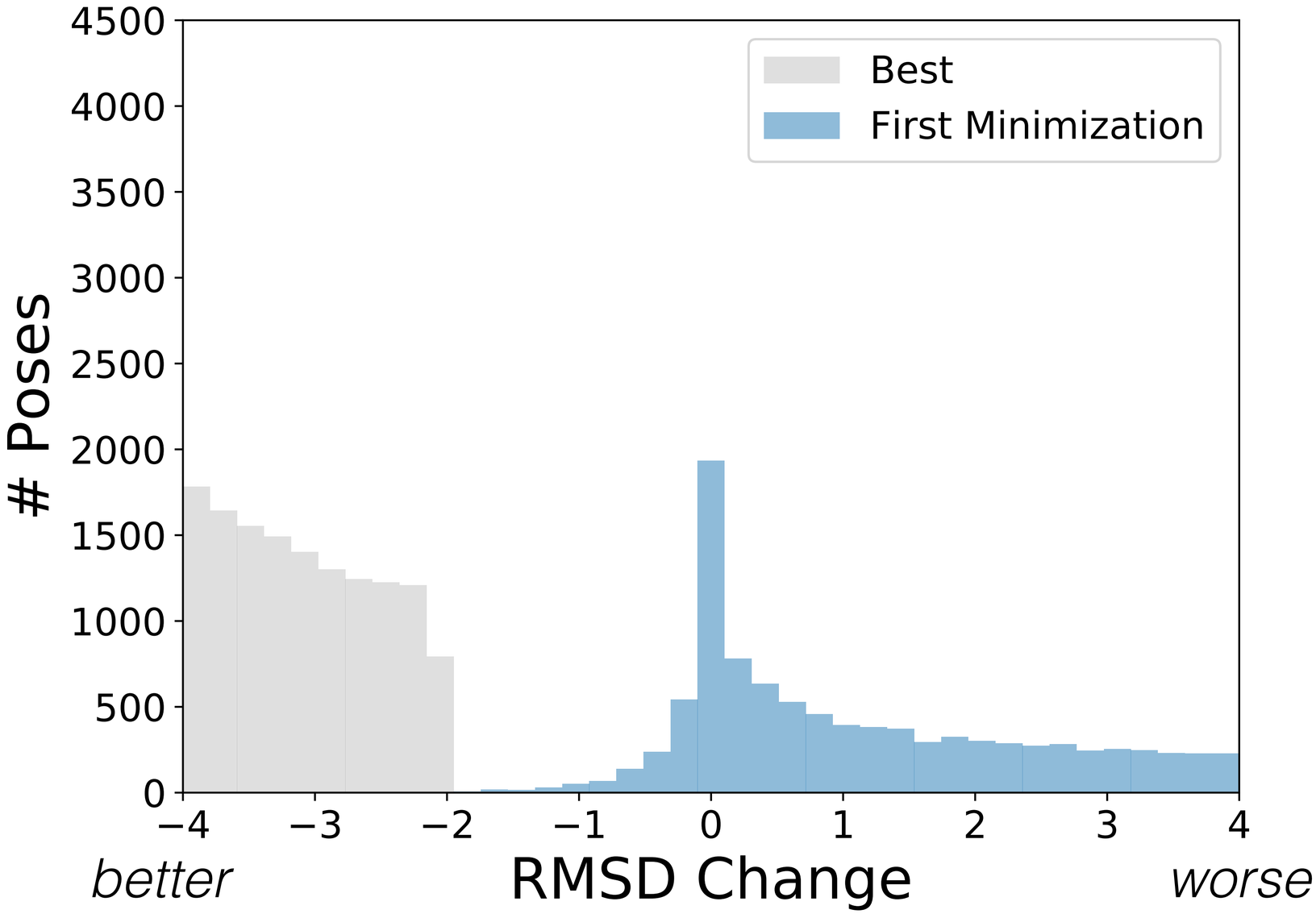
Beyond Scoring

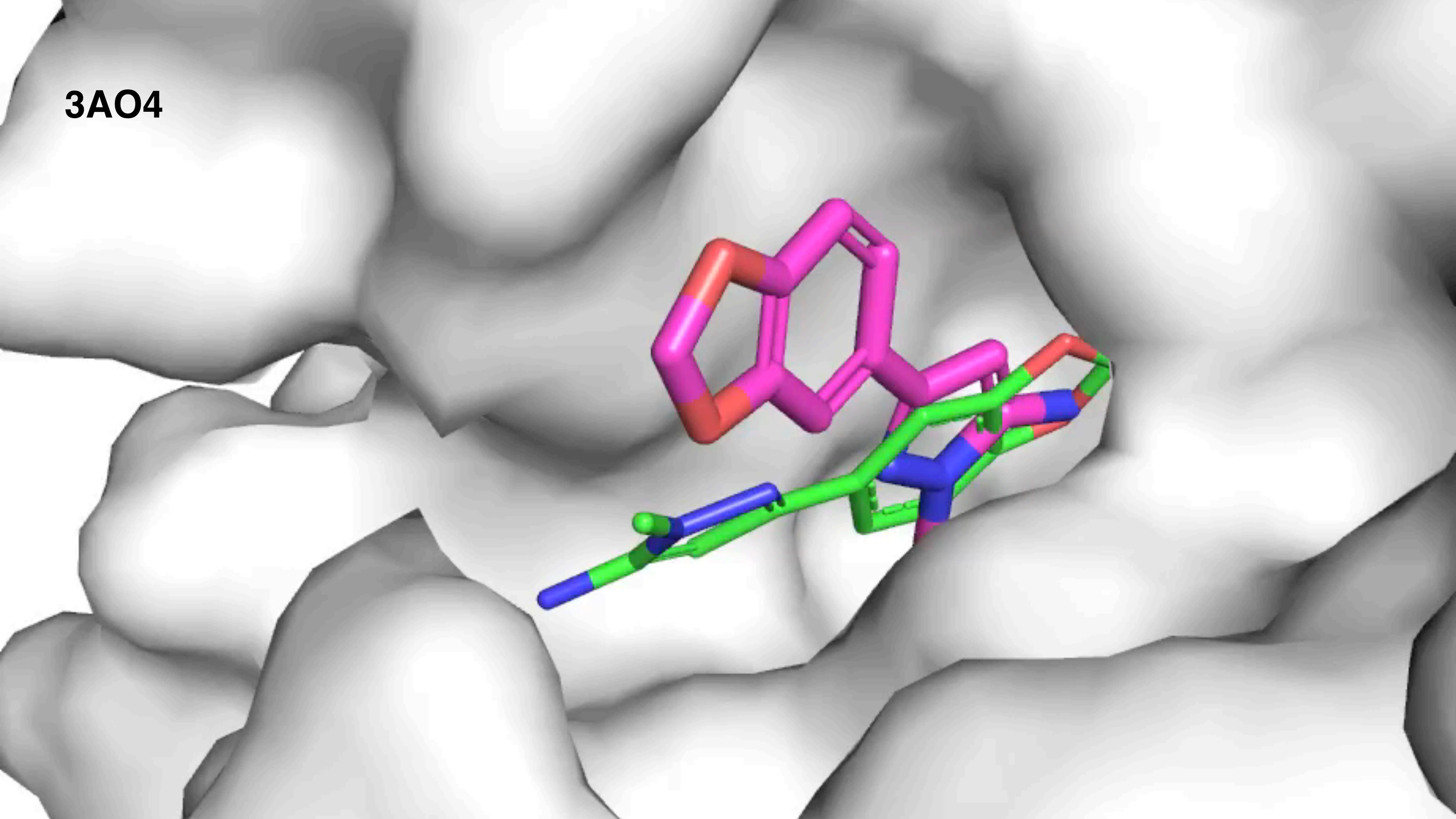


$$\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

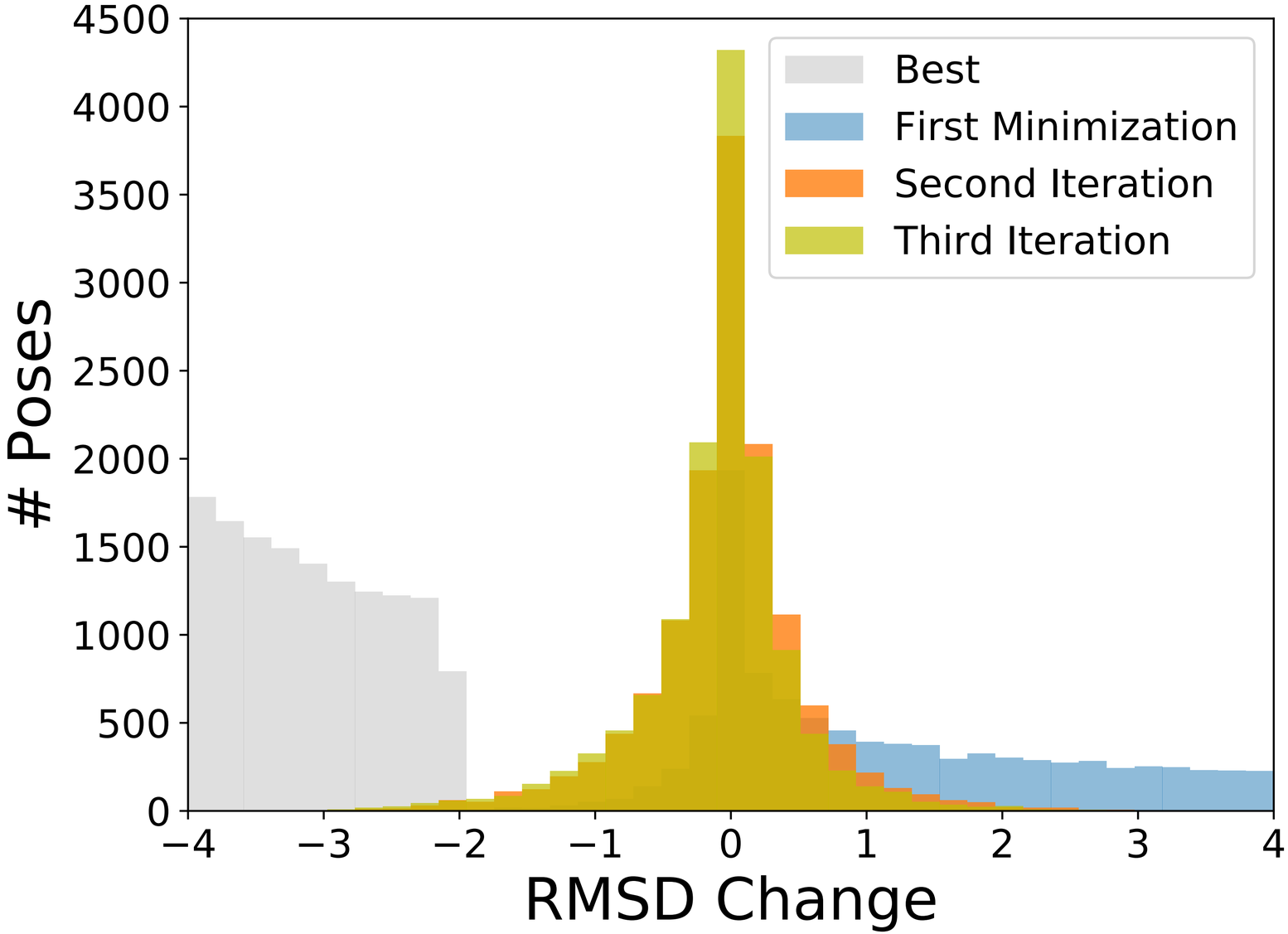


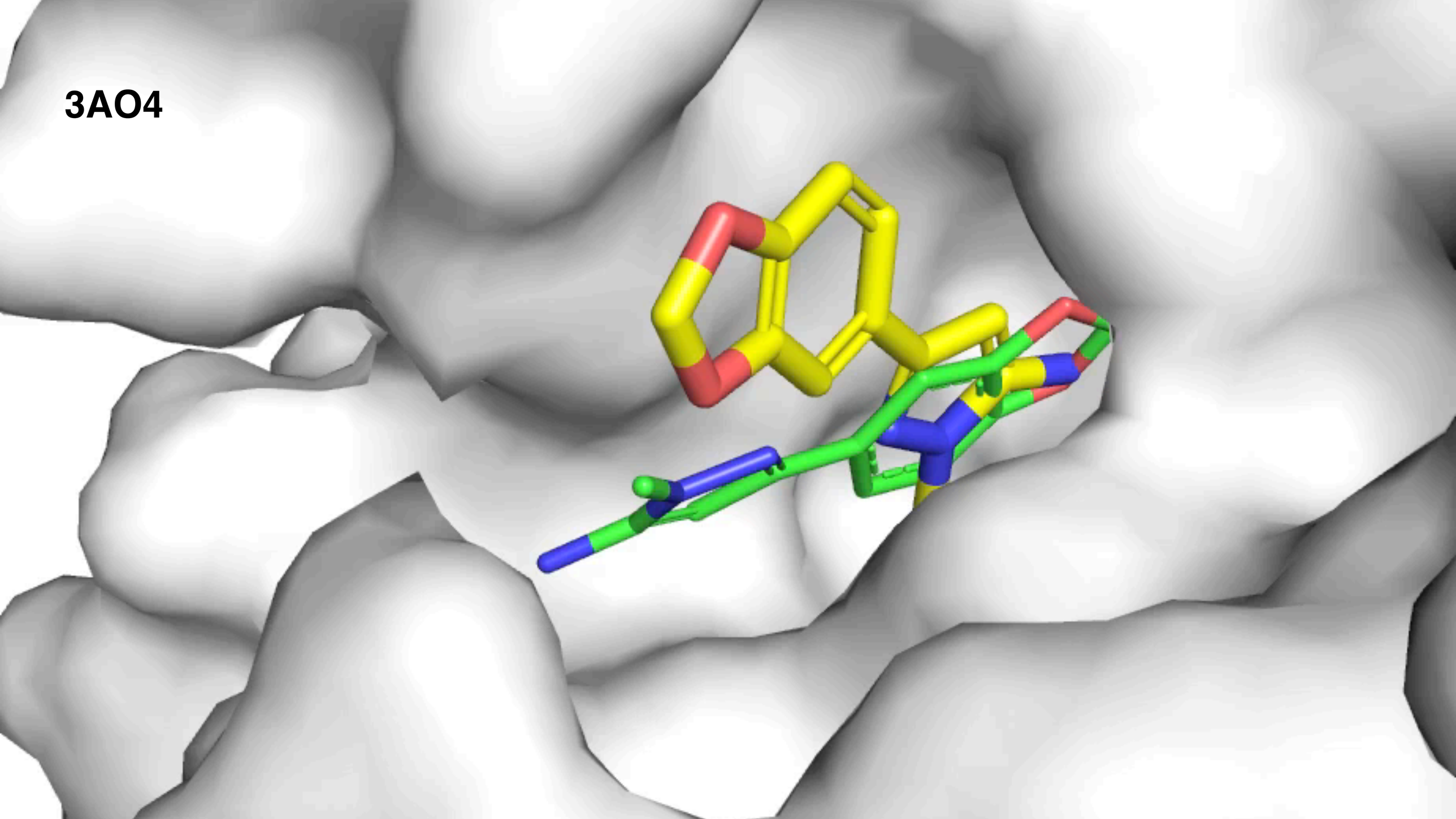
Minimizing Low RMSD Poses




3AO4

Iterative Refinement




3AO4

Docking

vina/smina/gnina

Sampling

MCMC

MCMC

MCMC

MCMC

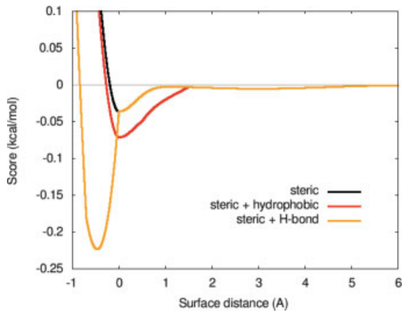
MCMC

⋮

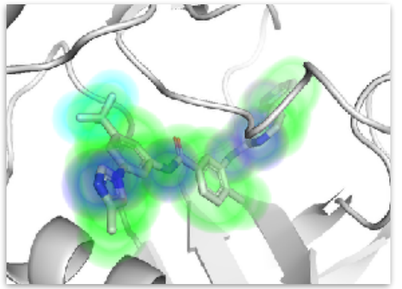
N (50) independent Monte Carlo chains
Scored with grid-accelerated Vina
Best identified pose retained

best poses

Refinement



Vina



CNN

Rescoring

CNN
pose
affinity

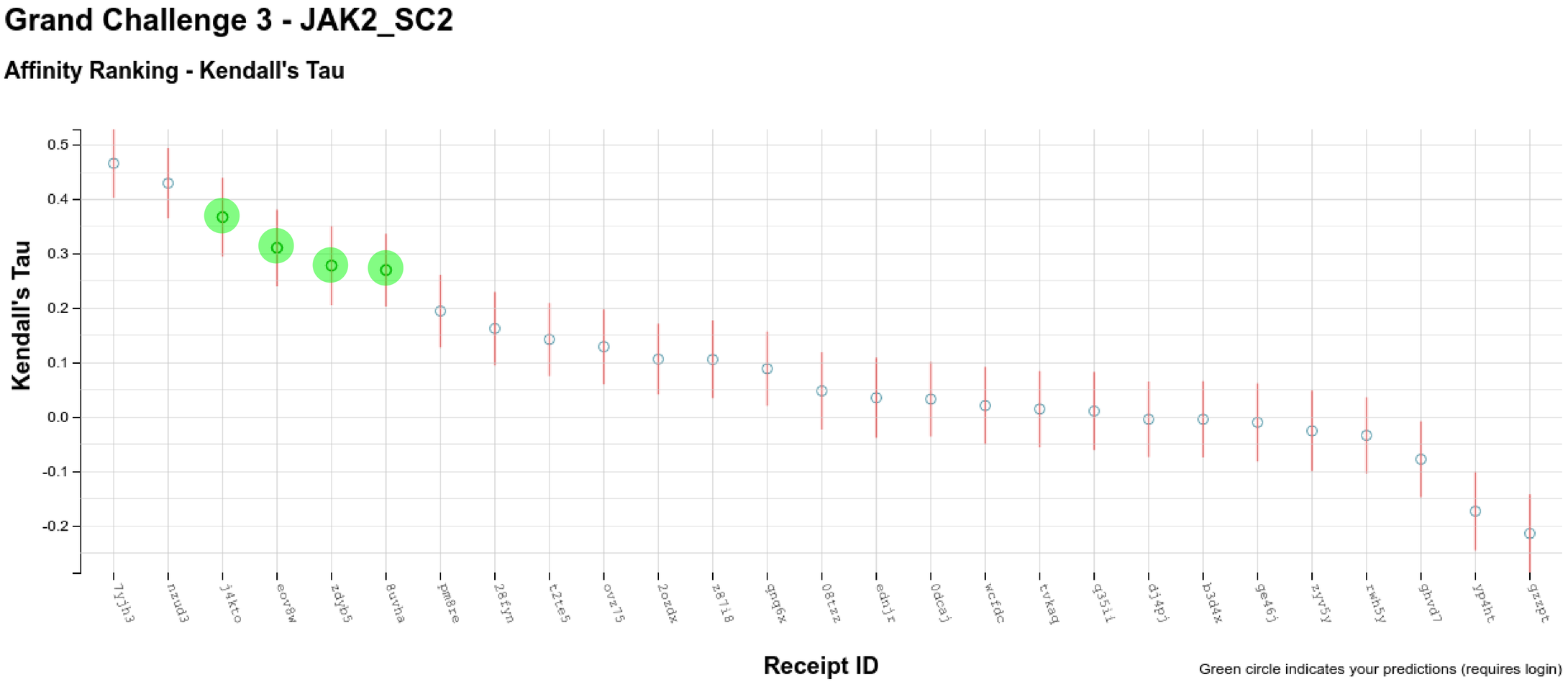
Prospective Evaluation: D3R

Grand Challenge 3

Spearman Correlation

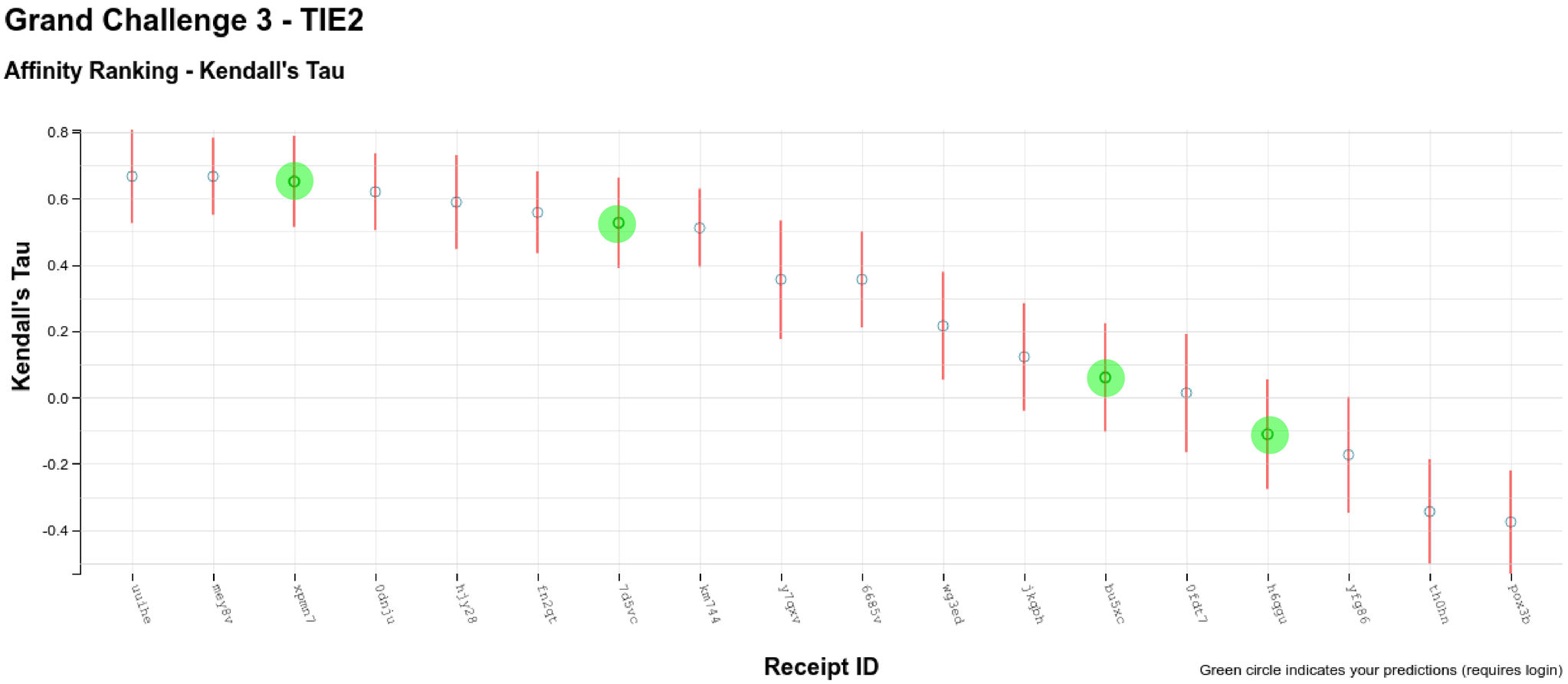
	cnn_docked_affinity	cnn_rescore_affinity	cnn_docked_scoring	cnn_rescore_scoring	vina
cat	0.0701	0.154	-0.0351	0.178	**0.179**
p38a	-0.0784	-0.116	-0.329	-0.305	**-0.0631**
vegfr2	0.366	**0.484**	0.434	0.448	0.414
jak2	**0.428**	0.338	0.39	0.27	0.106
jak2_sub3	**0.68**	0.369	-0.372	0.159	-0.633
tie2	0.648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

Grand Challenge 3



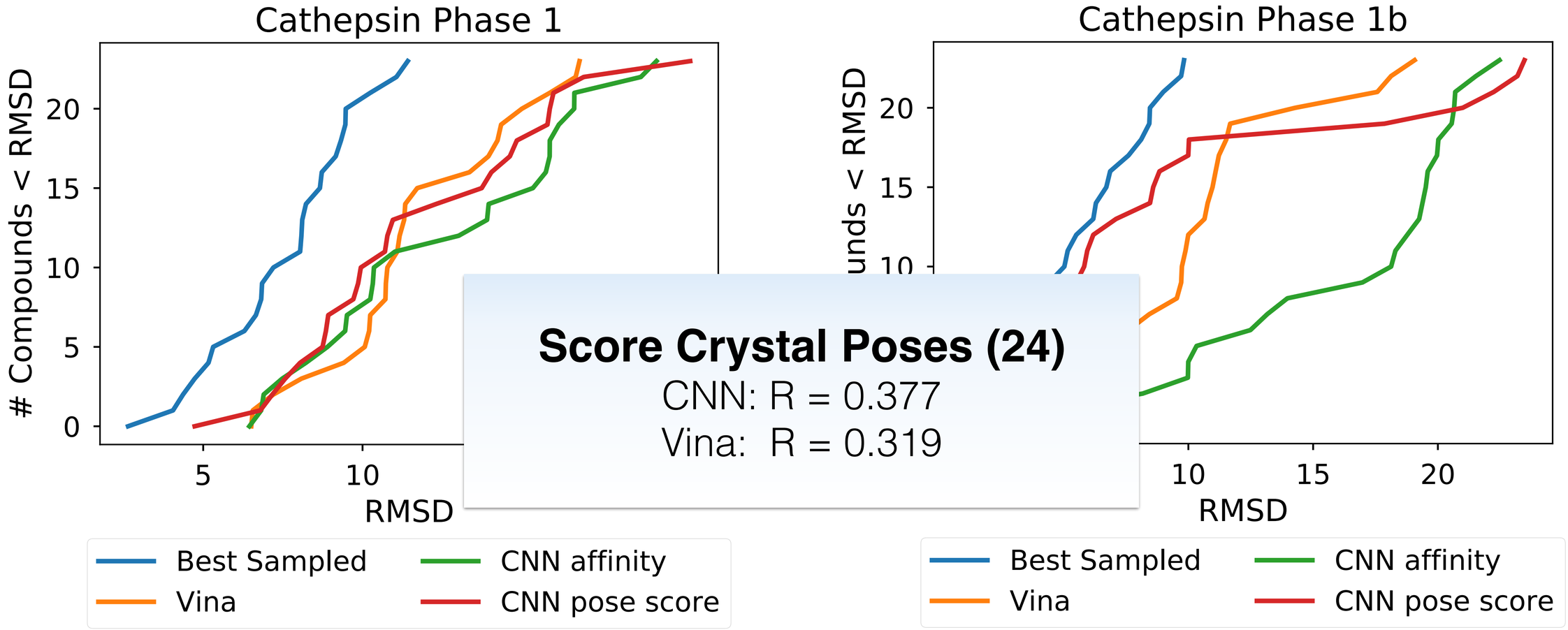
Grand Challenge 3

Grand Challenge 3



Grand Challenge 3

GC3: Pose Prediction



Score Crystal Poses (24)

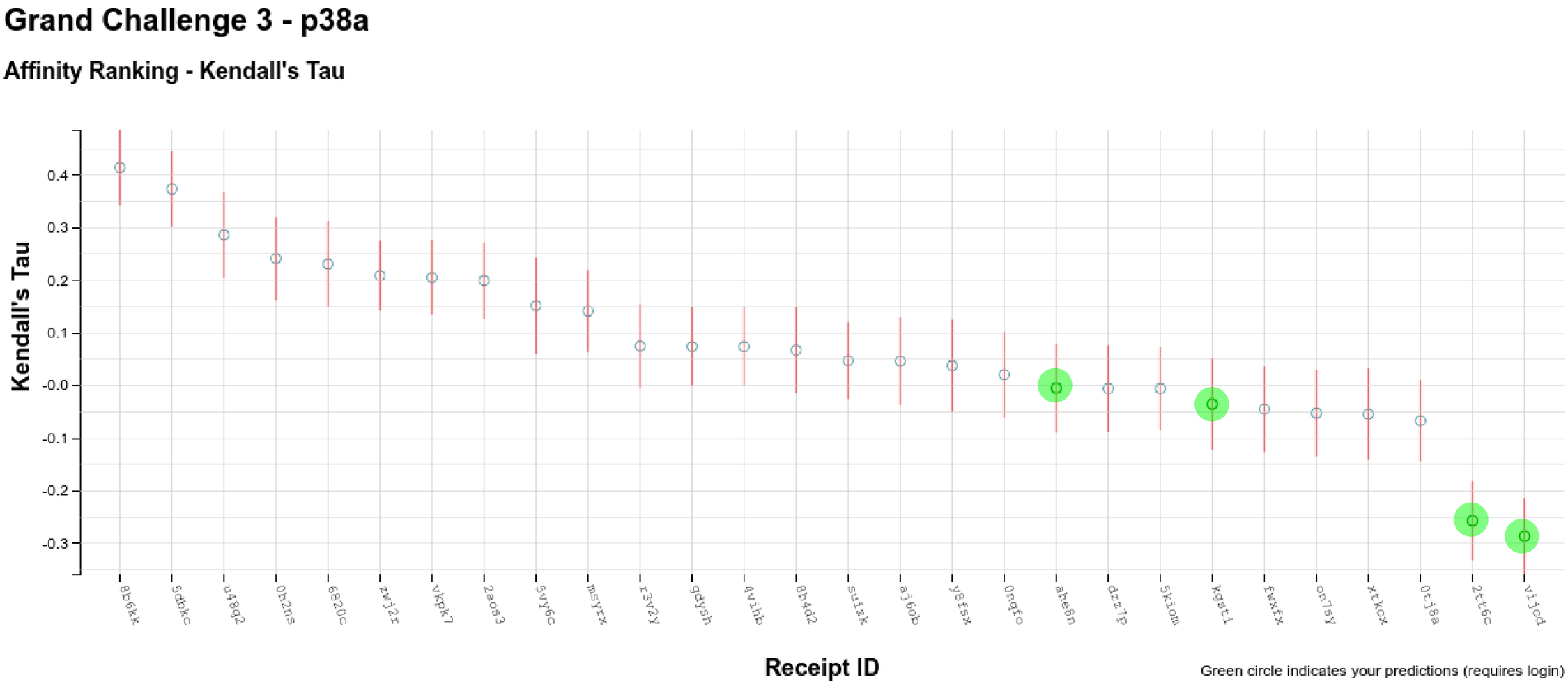
CNN: R = 0.377

Vina: R = 0.319

cross-docking

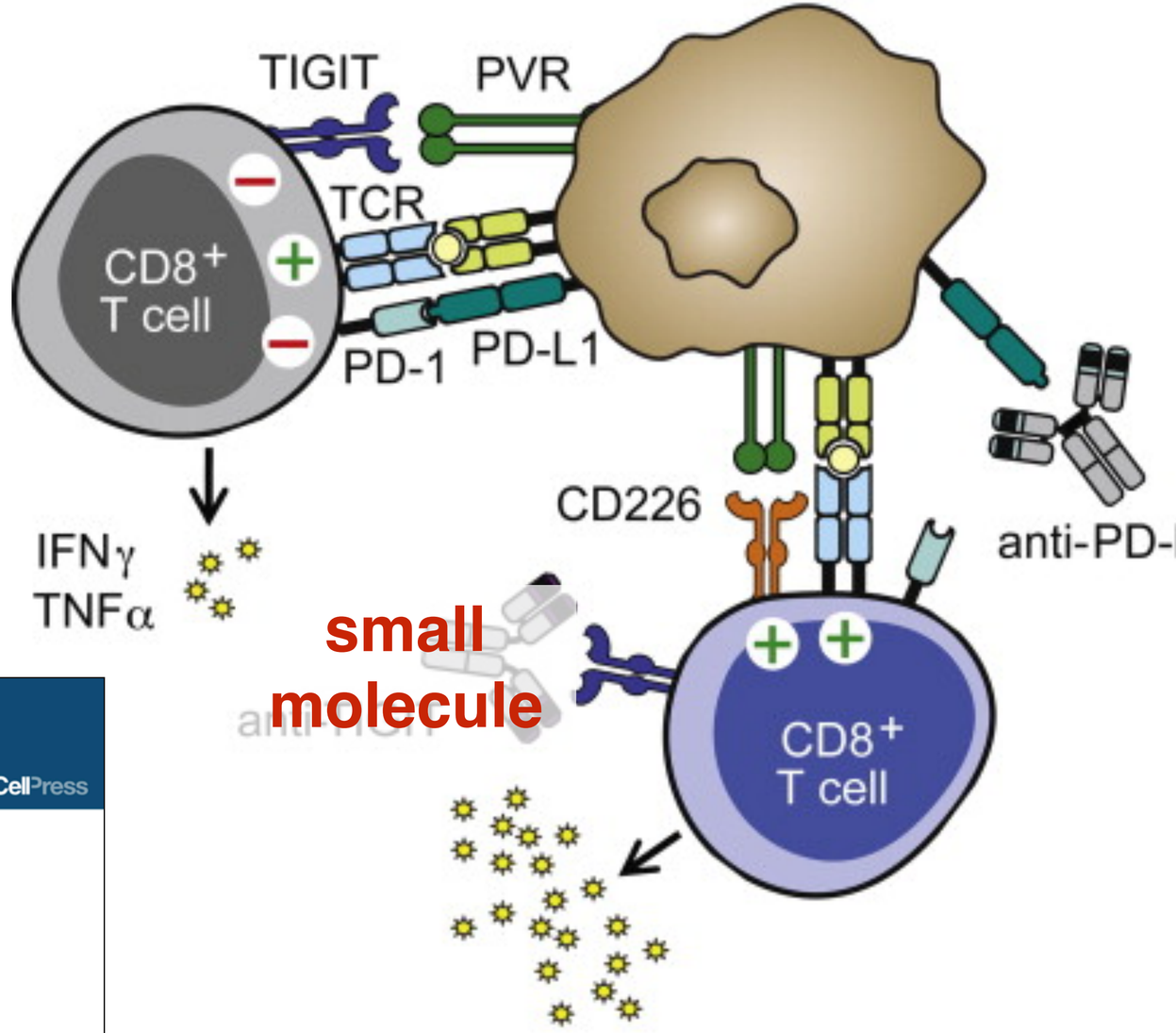
redocking

Grand Challenge 3



Prospective Evalution: TIGIT

Can we block TIGIT/ PVR interaction with a small molecule?



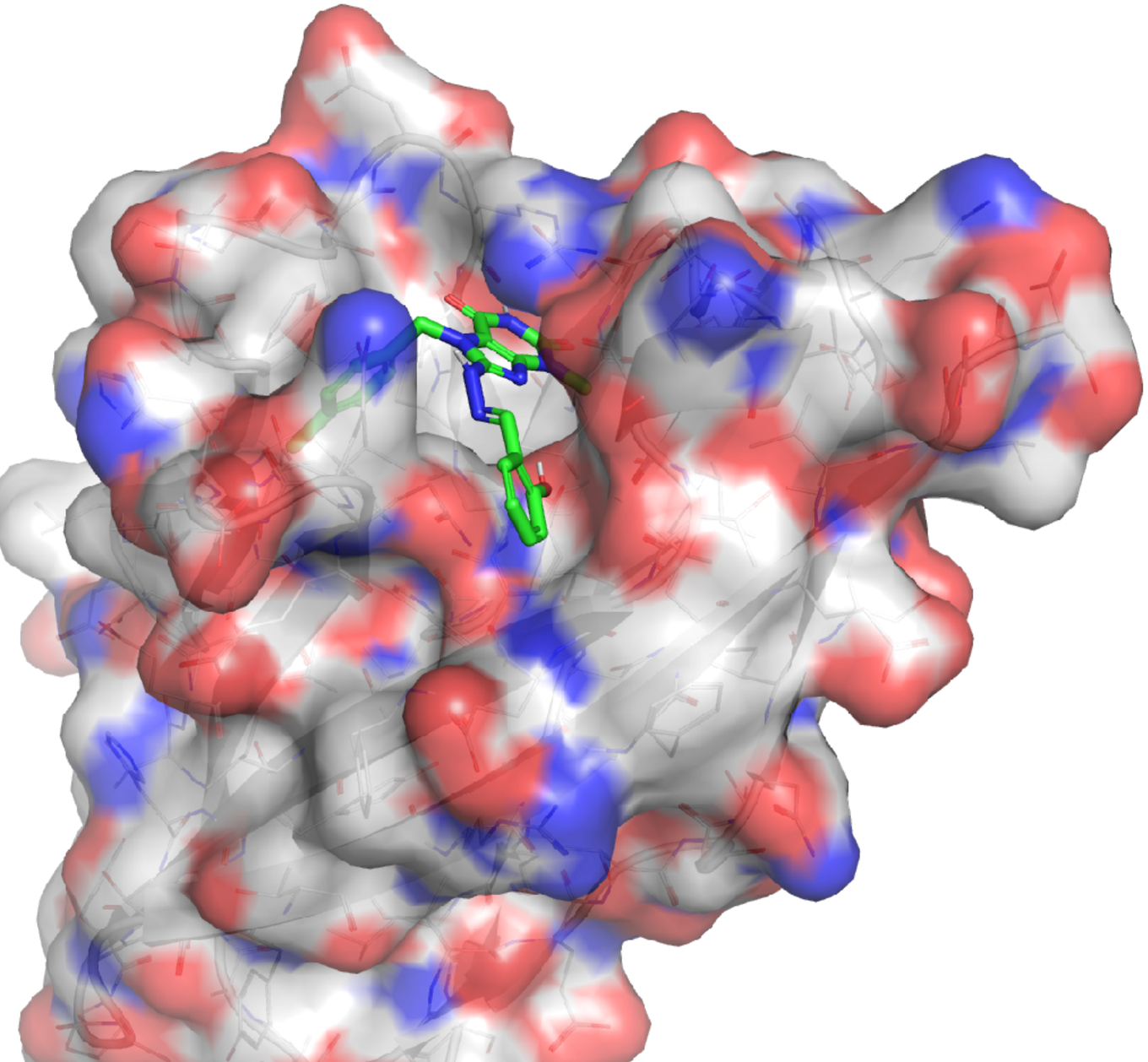
Cancer Cell
Article

CellPress

The Immunoreceptor TIGIT Regulates Antitumor and Antiviral CD8+ T Cell Effector Function

Robert J. Johnston,[1] Laetitia Comps-Agrar,[2] Jason Hackney,[3] Xin Yu,[1] Mahrukh Huseni,[4] Yagai Yang,[5] Summer Park,[6] Vincent Javinal,[5] Henry Chiu,[7] Bryan Irving,[1] Dan L. Eaton,[2] and Jane L. Grogan[1,*]
[1]Department of Cancer Immunology
[2]Department of Protein Chemistry
[3]Department of Bioinformatics and Computational Biology
[4]Department of Oncology Biomarker Development
[5]Department of Translational Oncology
[6]Department of Translational Immunology
[7]Department of Biochemical and Cellular Pharmacology
Genentech, 1 DNA Way, South San Francisco, CA 94080, USA
*Correspondence: grogan.jane@gene.com
http://dx.doi.org/10.1016/j.ccell.2014.10.018

Screening

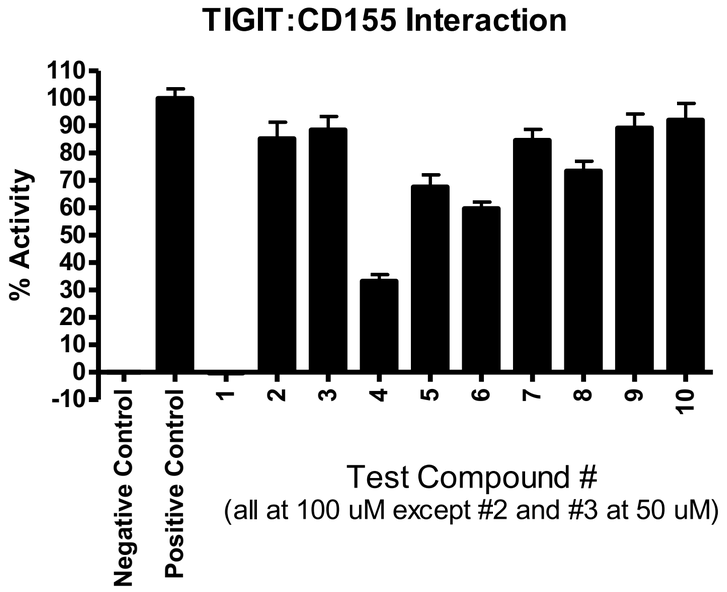


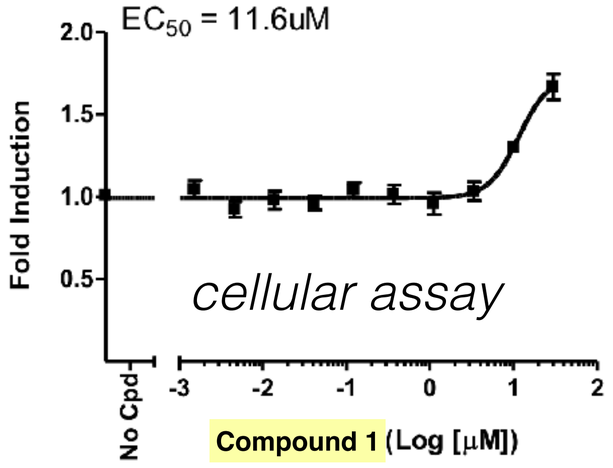
10 diverse compounds selected for screening

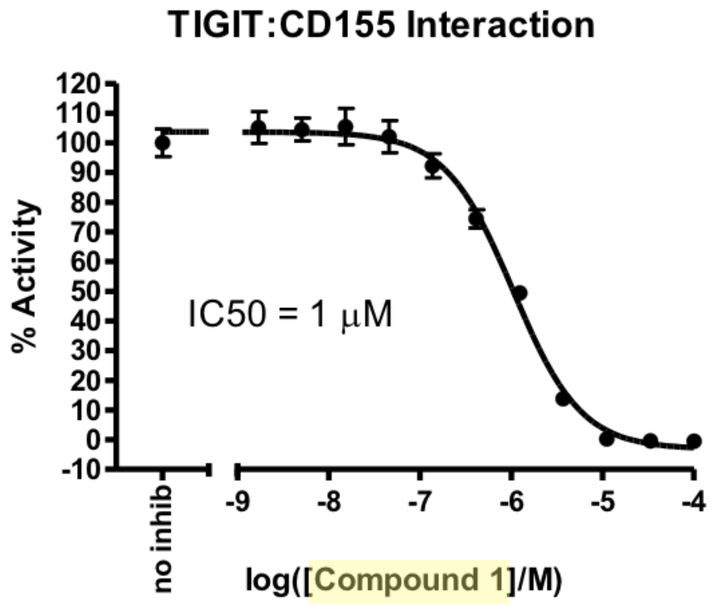
- **top ranked by Vina**
- **top ranked by CNN**

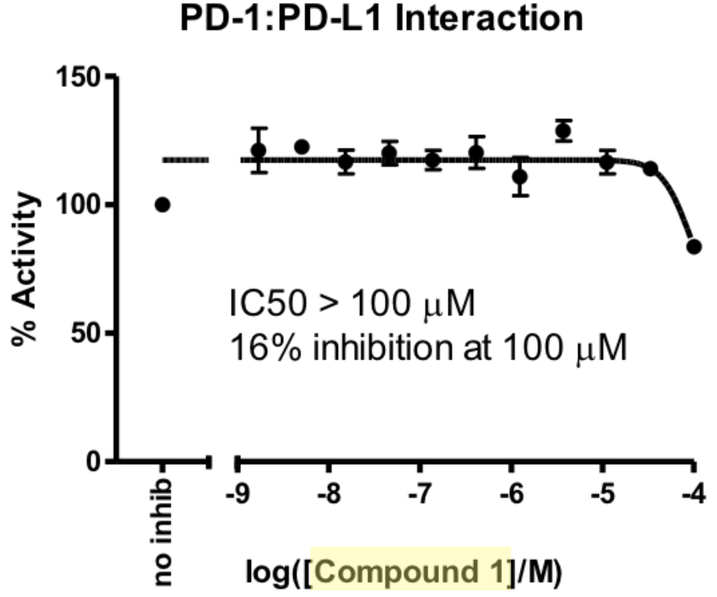
Name	CNN Affinity	CNN Score	Vina
Compound 1	7.69807	0.994763	85.95
Compound 2	5.57909	0.0180277	-8.12632
Compound 3	6.73692	0.0624742	-9.81935
Compound 4	6.87897	0.953488	-3.81378
Compound 5	6.32813	0.209807	-8.60293
Compound 6	5.689	0.0437	-8.991
Compound 7	4.368	0.022	-9.34722
Compound 8	4.81	0.072	-6.81787
Compound 9	5.22	0.032	-6.264
Compound 10	6.67	0.361	6.1053

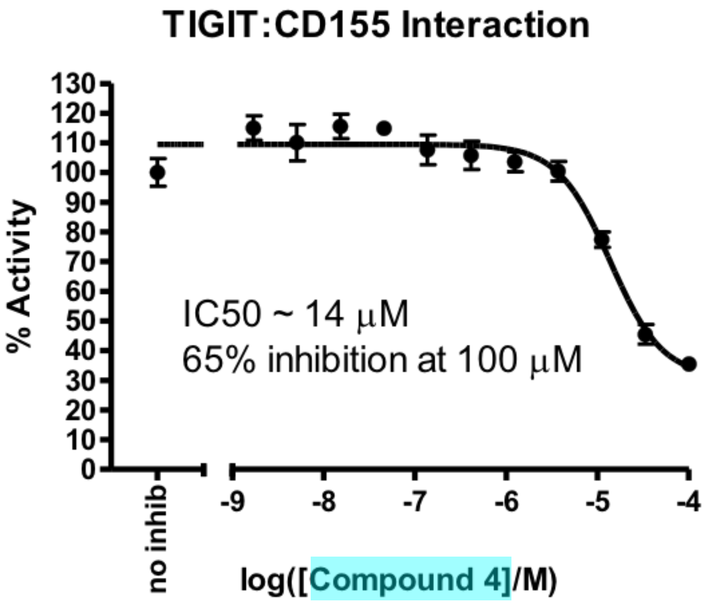
Results

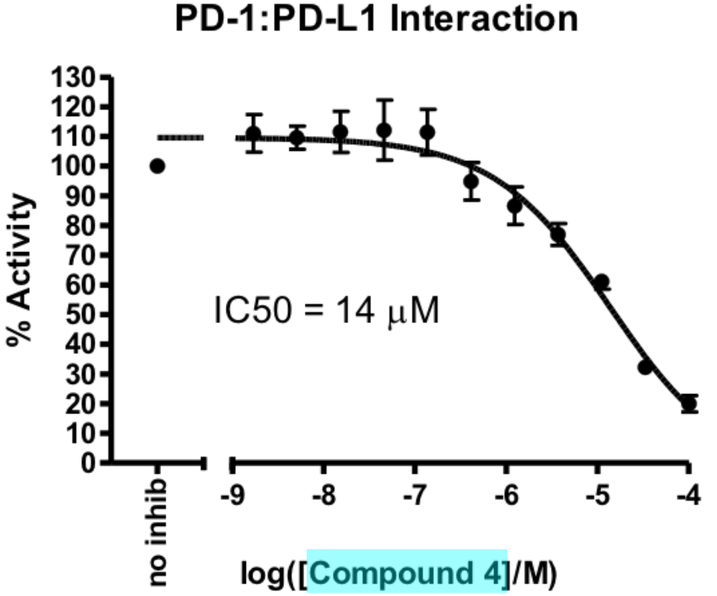

TIGIT:CD155 Interaction
% Activity
Negative Control
Positive Control
Test Compound #
(all at 100 uM except #2 and #3 at 50 uM)


EC50 = 11.6uM
Fold Induction
cellular assay
No Cpd
Compound 1 (Log [μM])

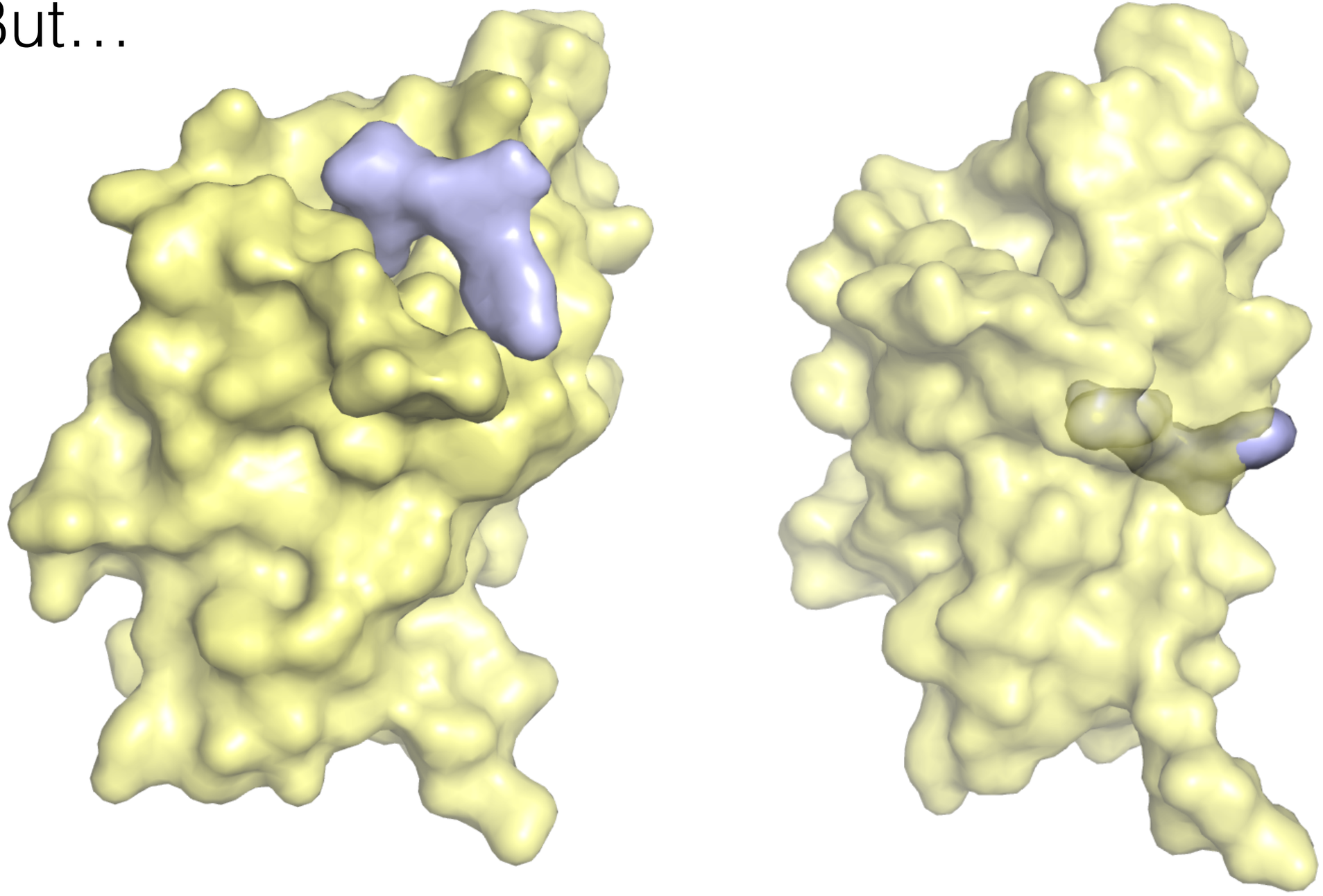

TIGIT:CD155 Interaction
% Activity
IC50 = 1 μM
no inhib
log([Compound 1]/M)


PD-1:PD-L1 Interaction
% Activity
IC50 > 100 μM
16% inhibition at 100 μM
no inhib
log([Compound 1]/M)


TIGIT:CD155 Interaction
% Activity
IC50 ~ 14 μM
65% inhibition at 100 μM
no inhib
log([Compound 4]/M)

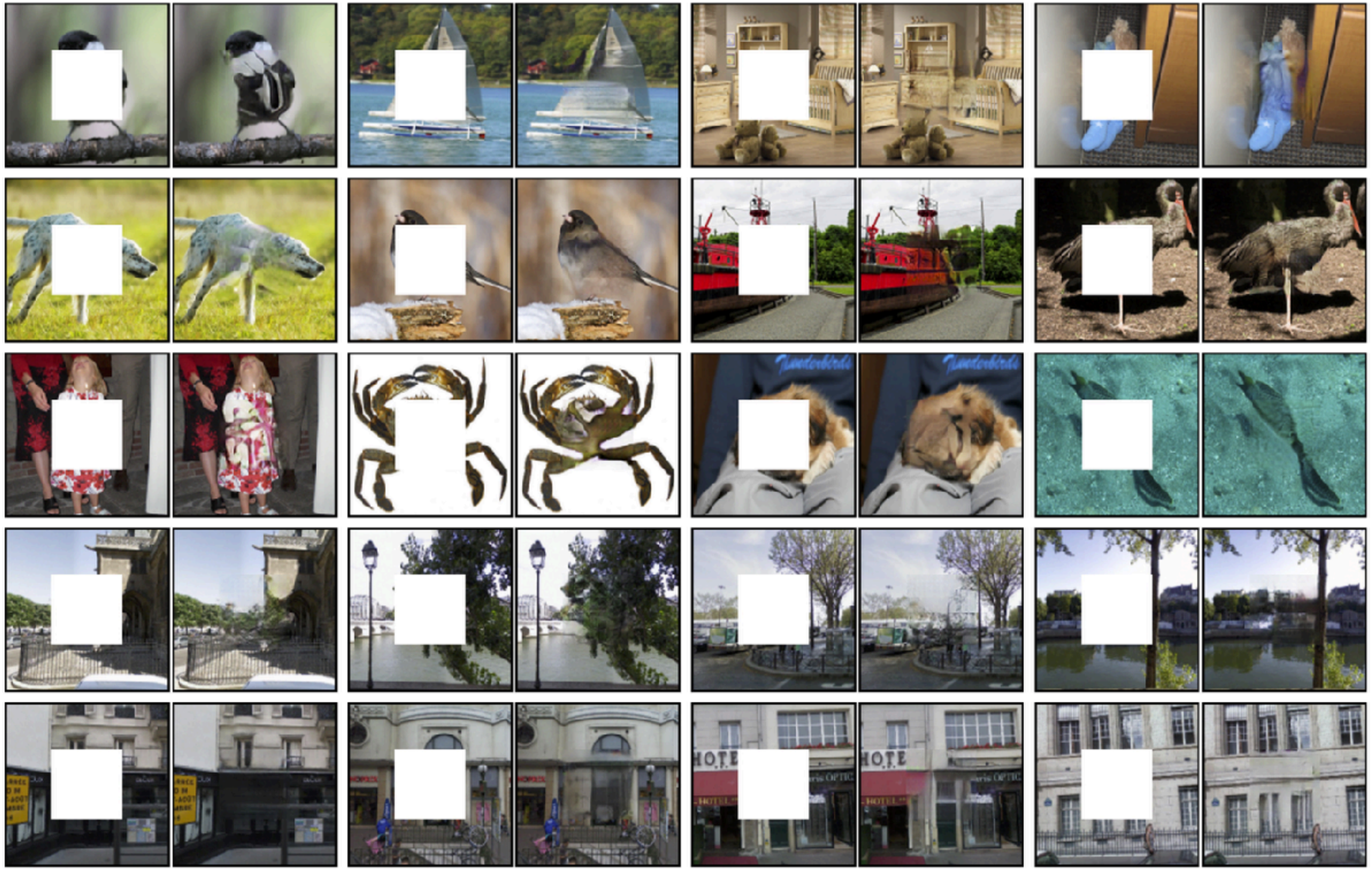

PD-1:PD-L1 Interaction
% Activity
IC50 = 14 μM
no inhib
log([Compound 4]/M)

But…



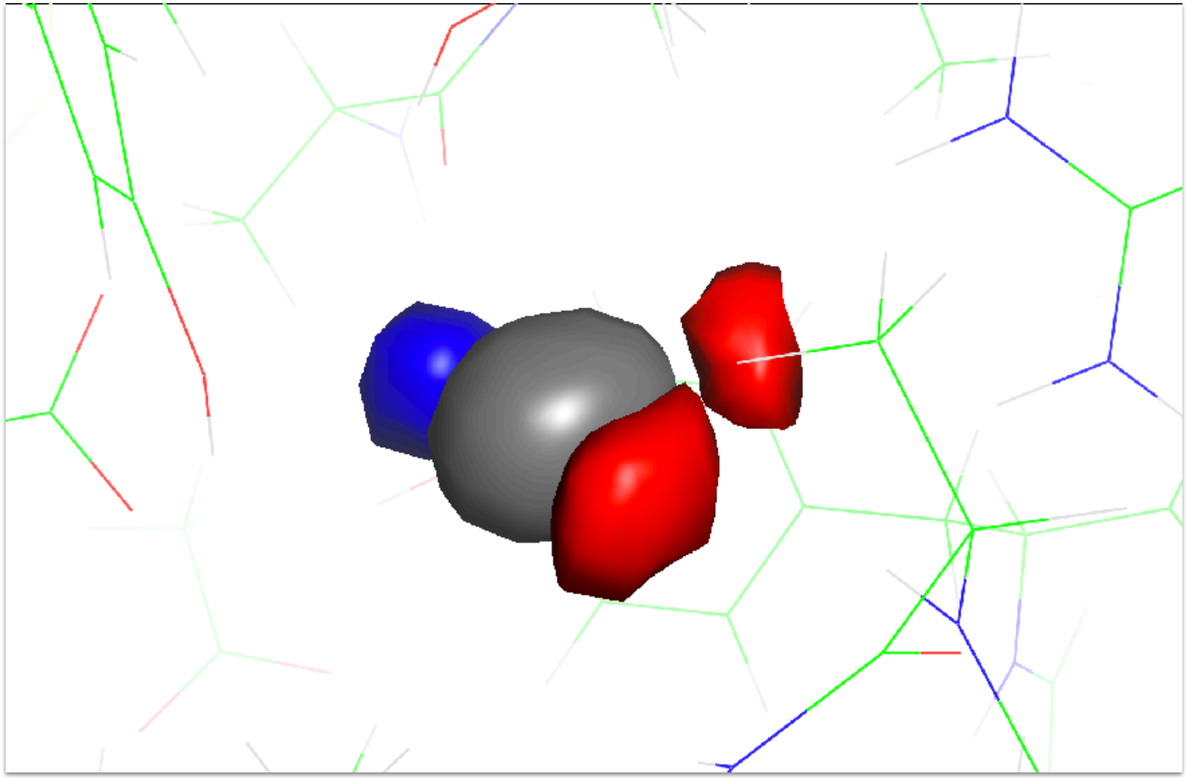
and now for something completely different…

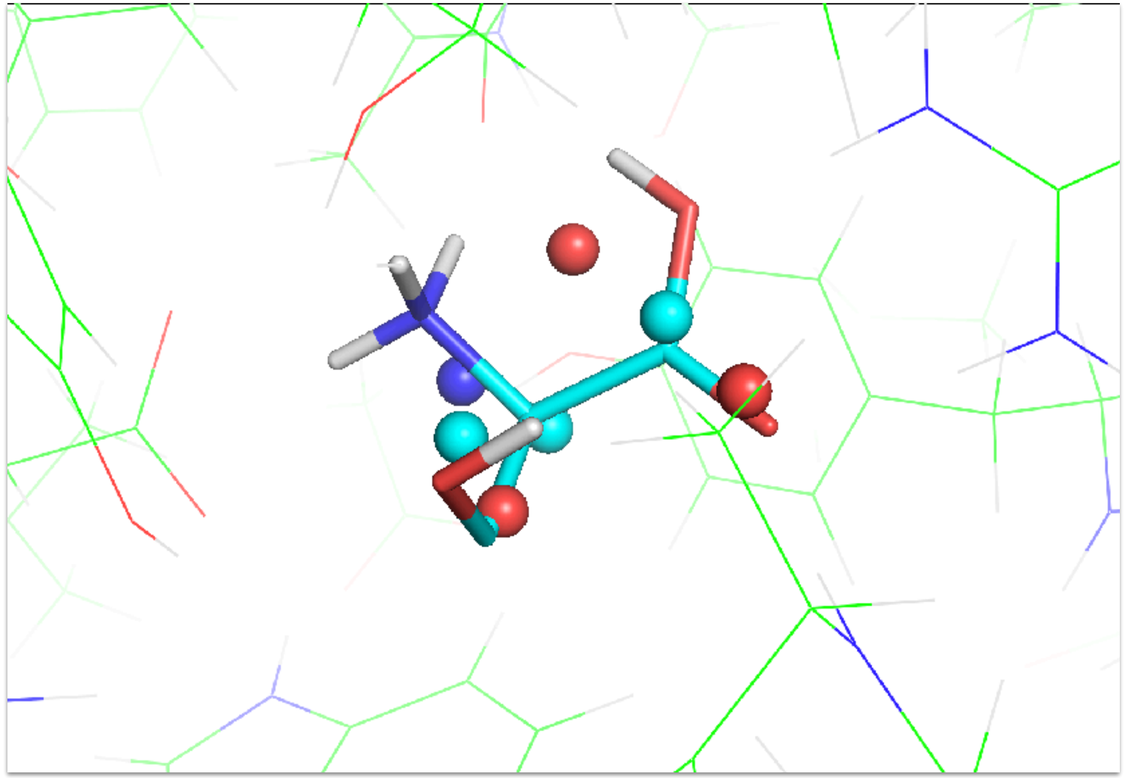
Context Encoding



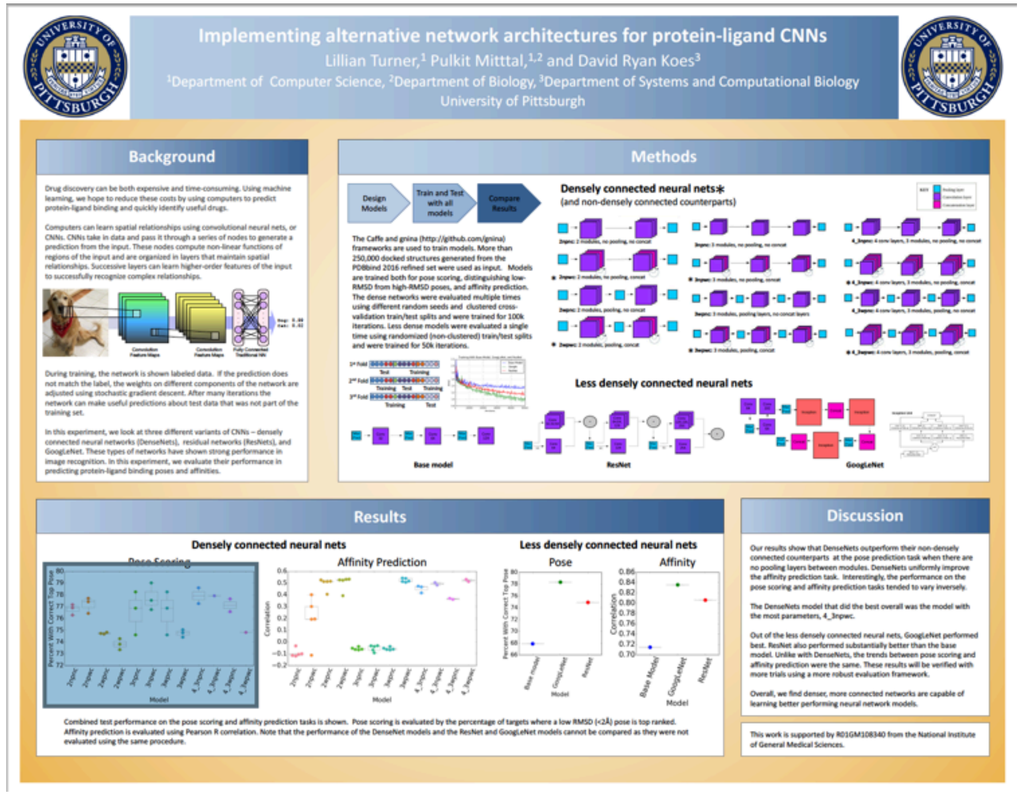
http://people.eecs.berkeley.edu/~pathak/context_encoder/

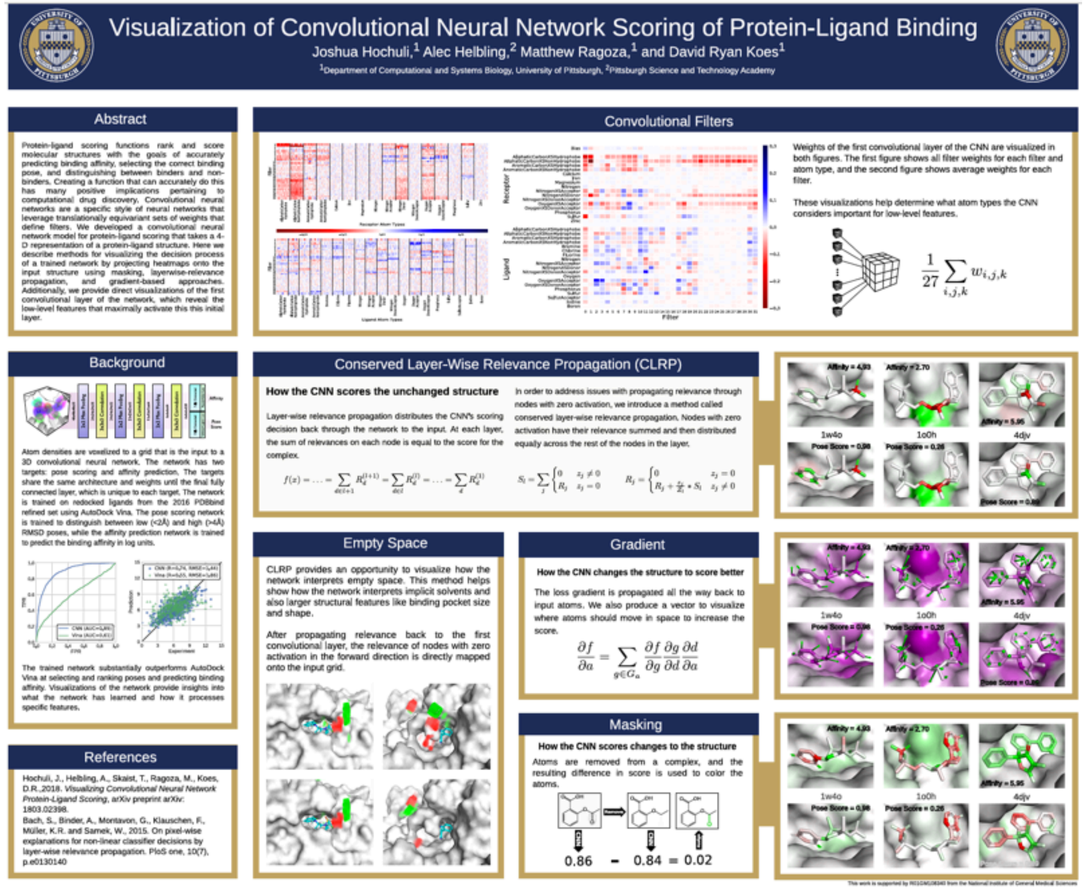
Molecular Context Encoding





Shameless Plug



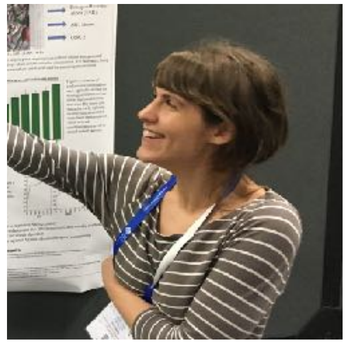


☆ COMP 191: Alternative network architectures for protein-ligand scoring

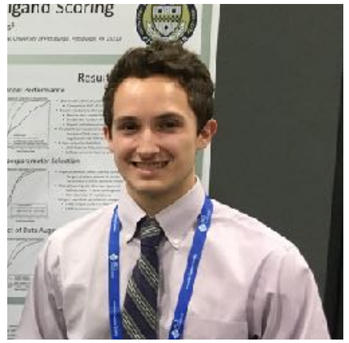
☆ COMP 320: GPU molecular docking with convolutional neural network scoring functions

☆ COMP 226: Visualizing convolutional neural network scoring of protein-ligand binding

Acknowledgements



Jocelyn Sunseri



Matt Ragoza

Josh Hochuli

Lily Turner

Group Members

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

Department of Computational and Systems Biology

TECBio REU

AI GRANT

National Institute of General Medical Sciences
R01GM108340

NVIDIA

github.com/gnina

http://bits.csb.pitt.edu

@david_koes

Computer Facts
@computerfact

concerned parent: if all your friends jumped off a bridge would you follow them?
machine learning algorithm: yes.

2:20 PM · Mar 15, 2018