

Deep Learning for Drug Discovery

David Koes

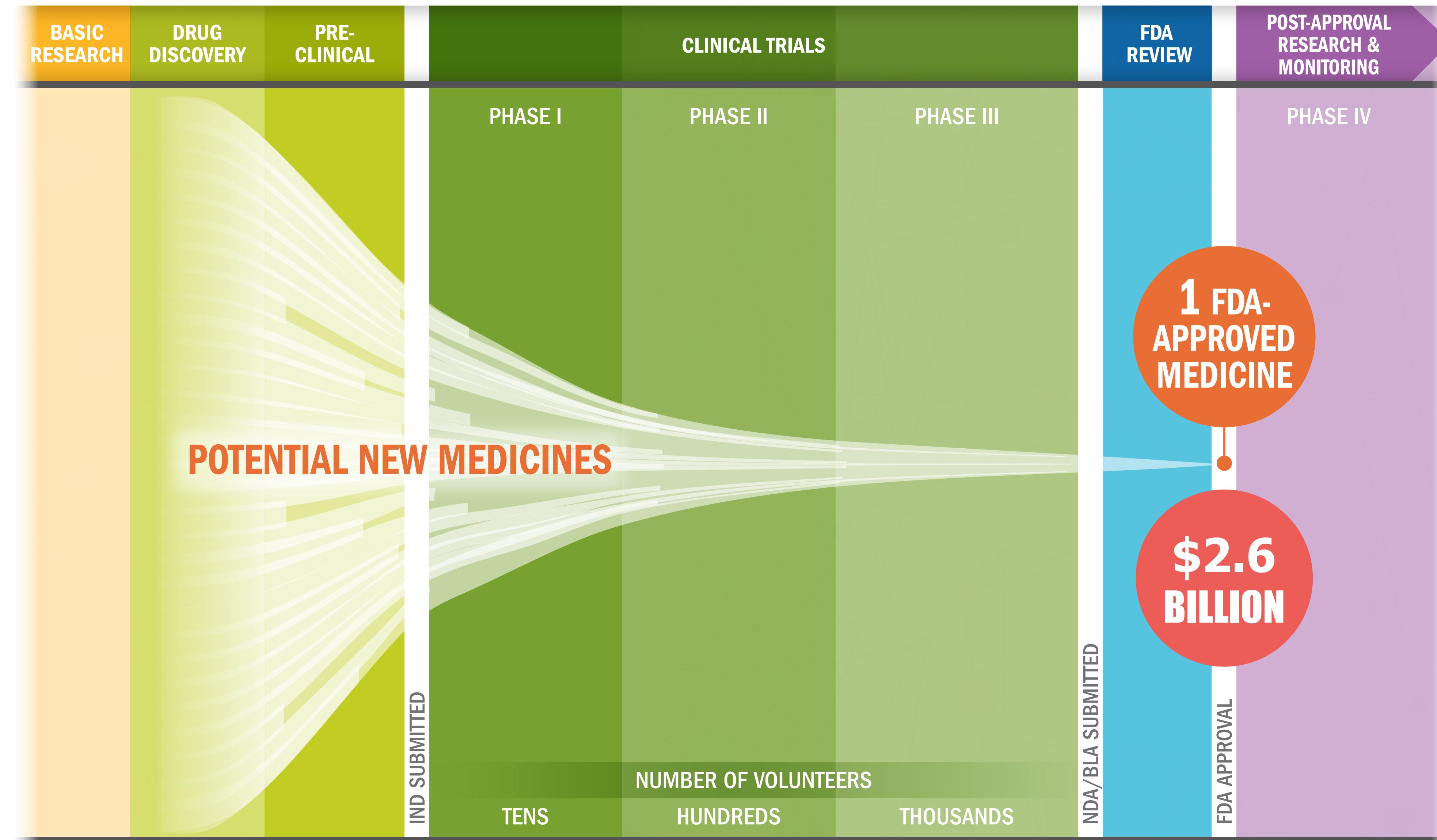


@david_koes

Drug Discovery Innovation Programme
Boston, MA
May 19, 2019

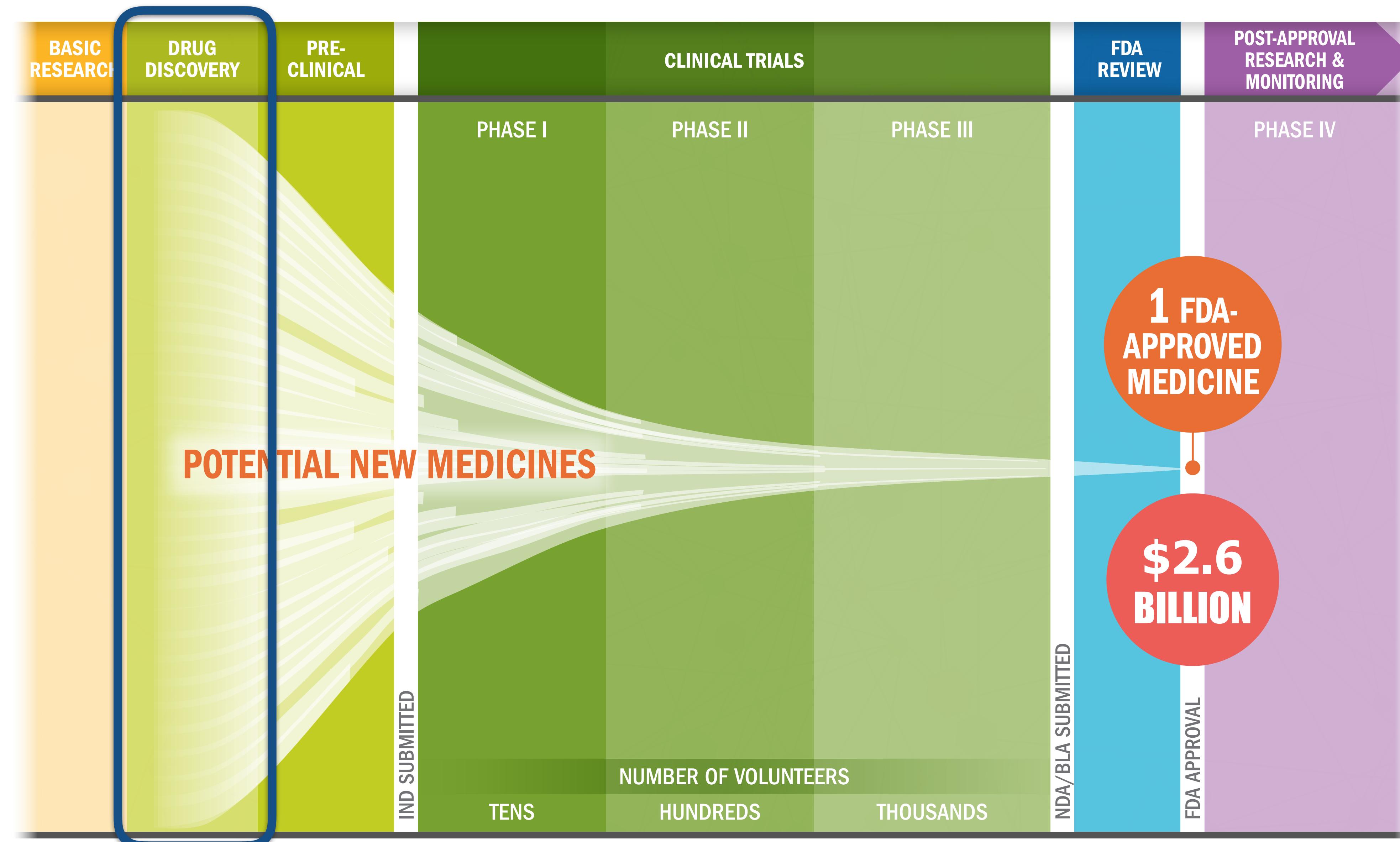


THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS



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

THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS



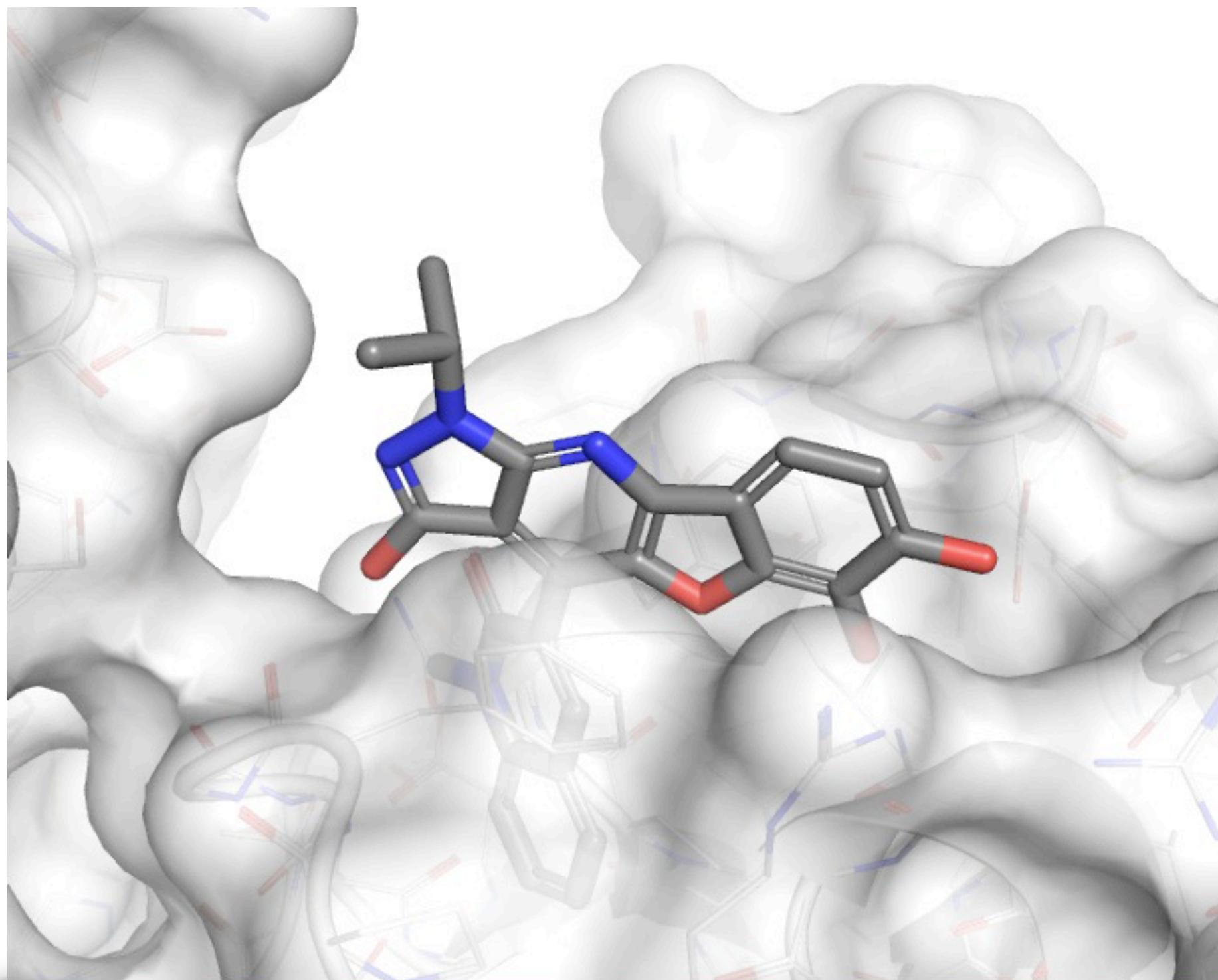
Source: Pharmaceutical Research and Manufacturers of America (<http://phrma.org>)

Structure Based Drug Design

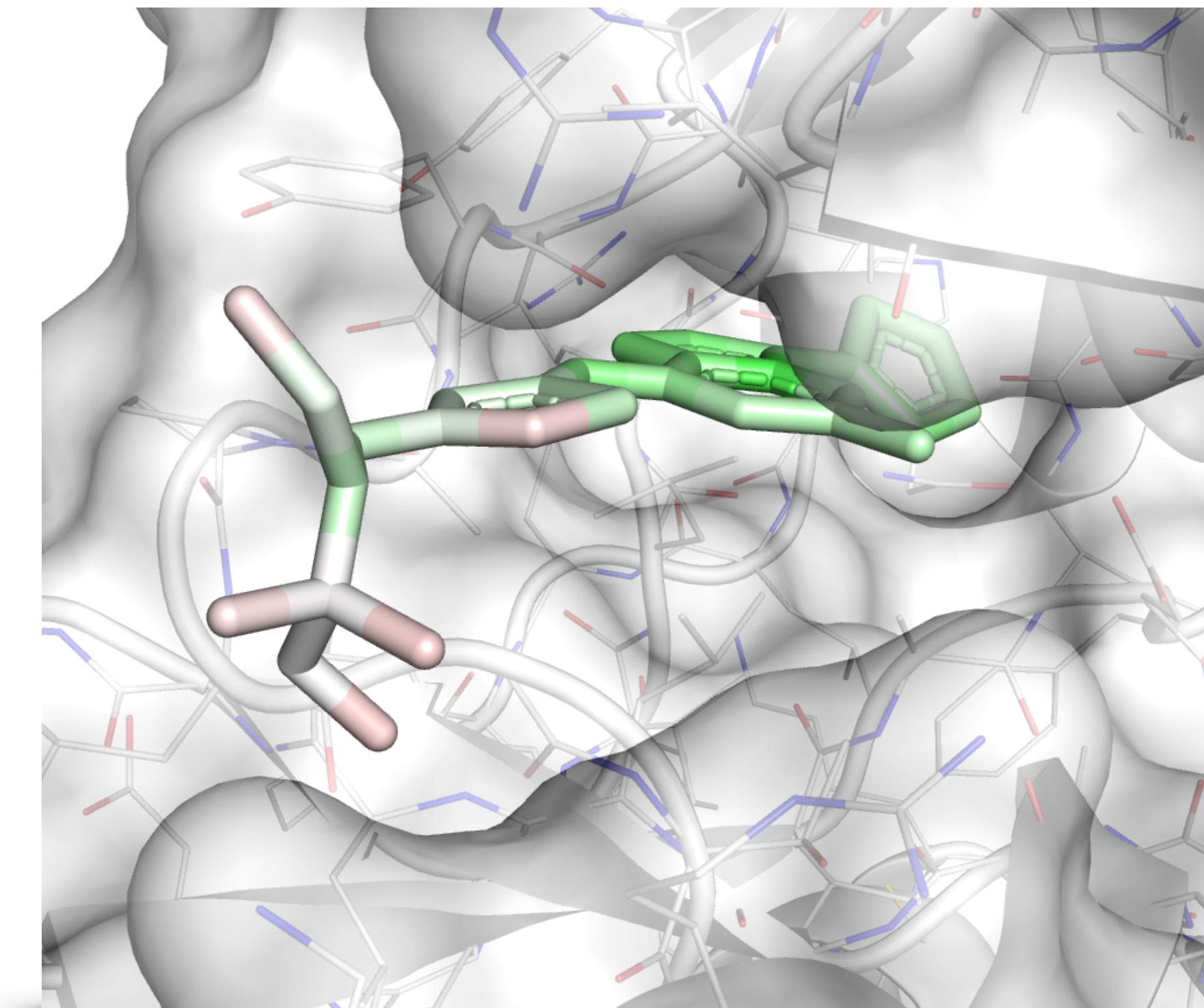
Pose Prediction

Binding Discrimination

Affinity Prediction



Virtual Screening



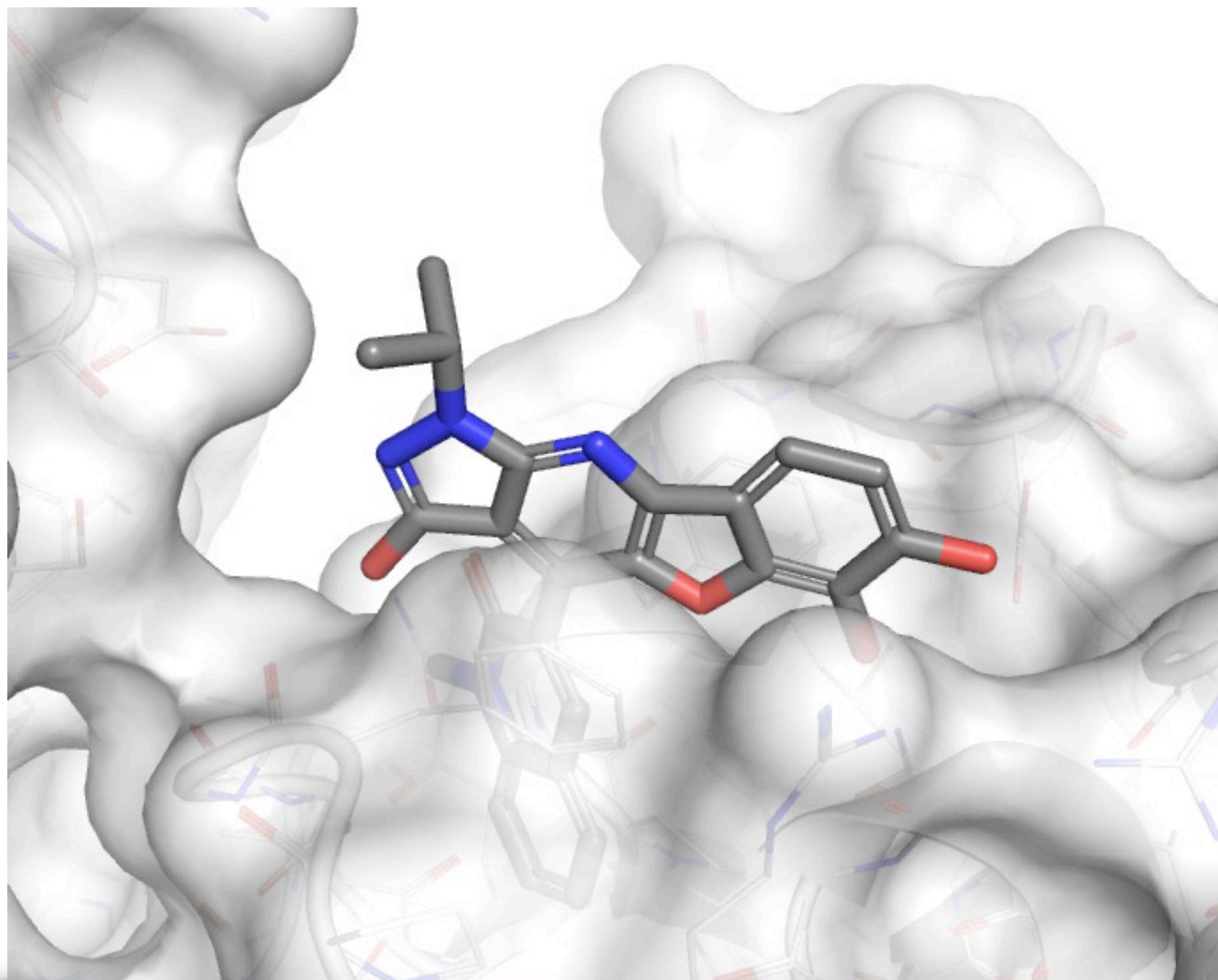
Lead Optimization

Structure Based Drug Design

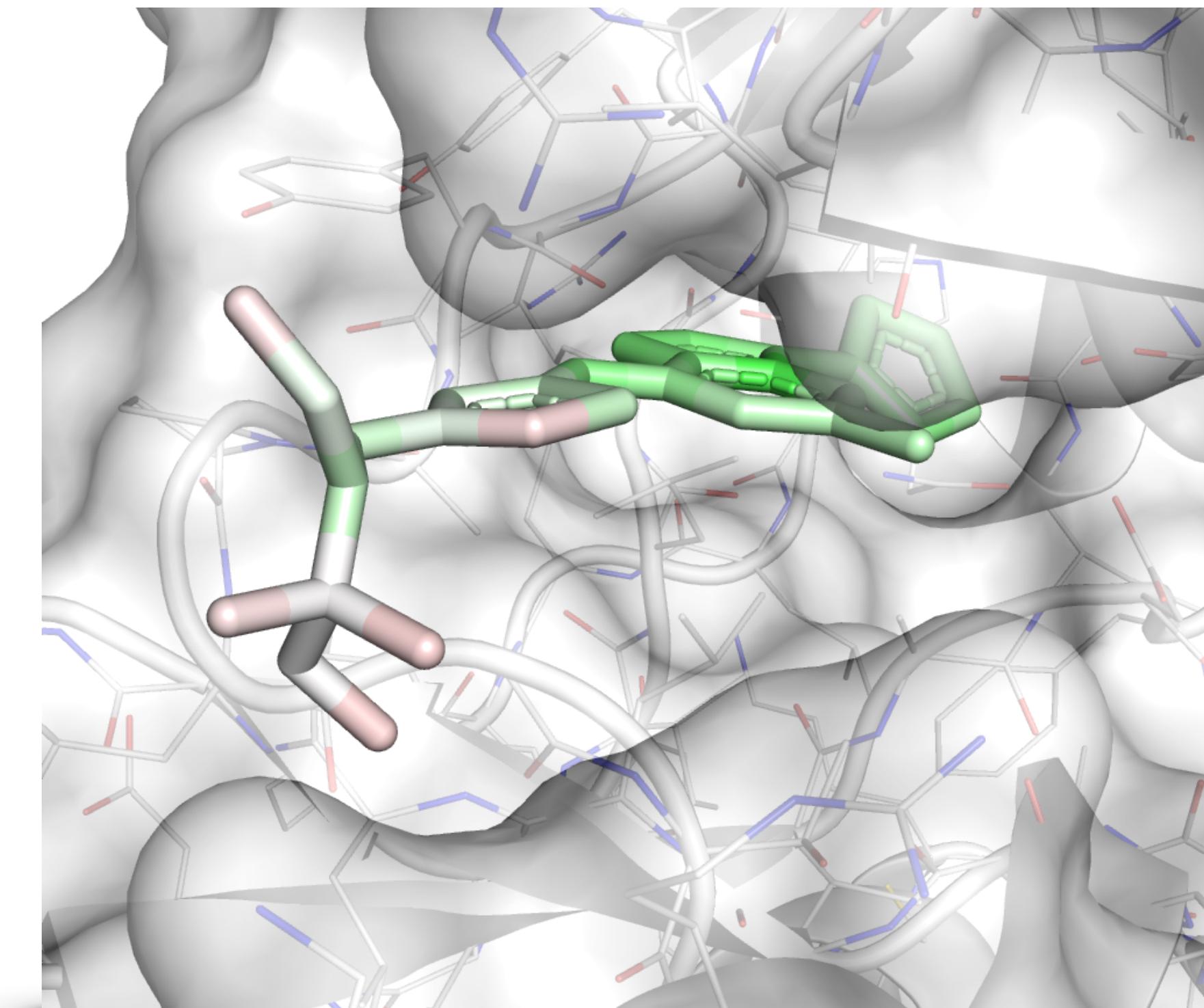
Pose Prediction

Binding Discrimination

Affinity Prediction



Virtual Screening

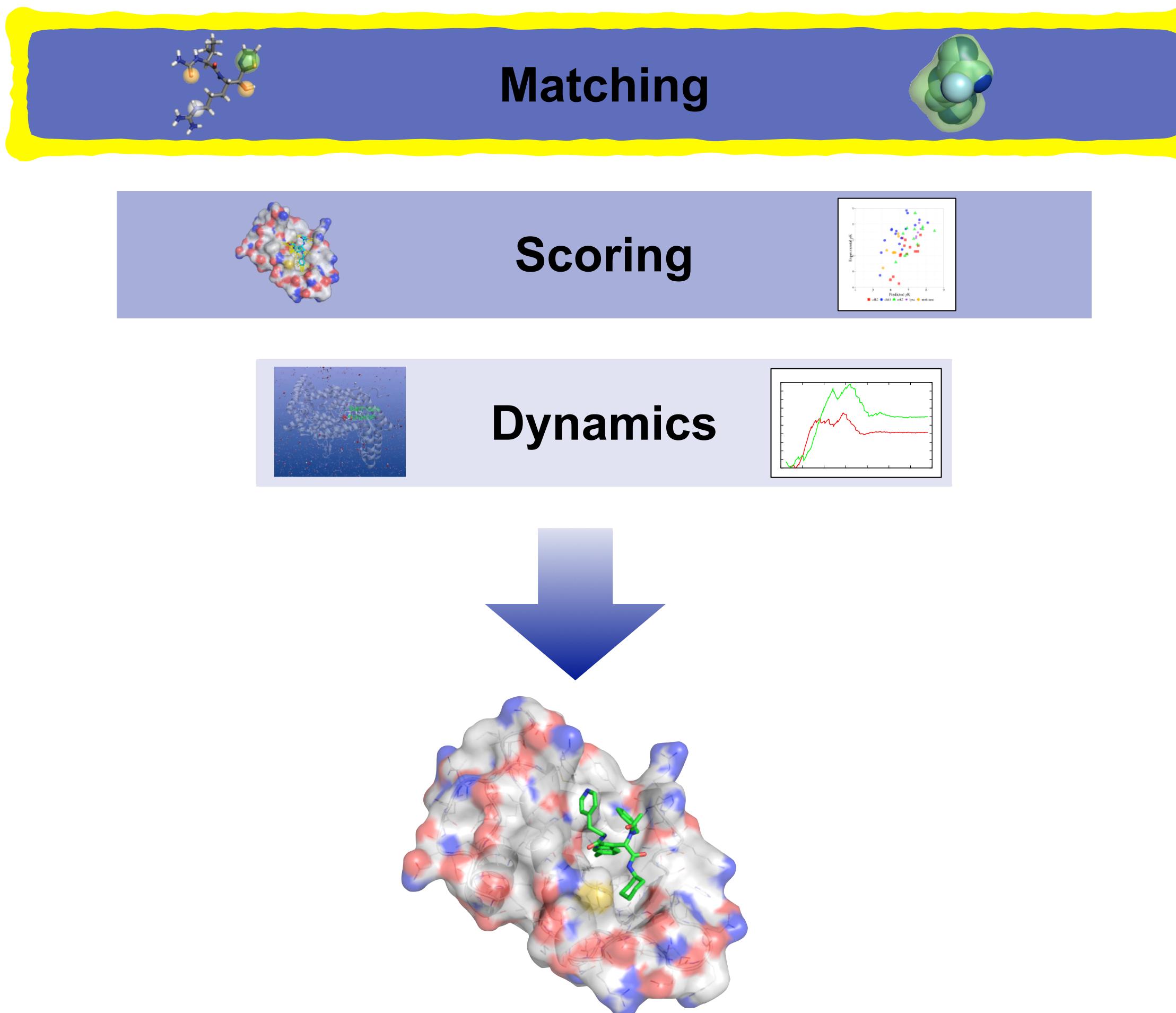


Lead Optimization

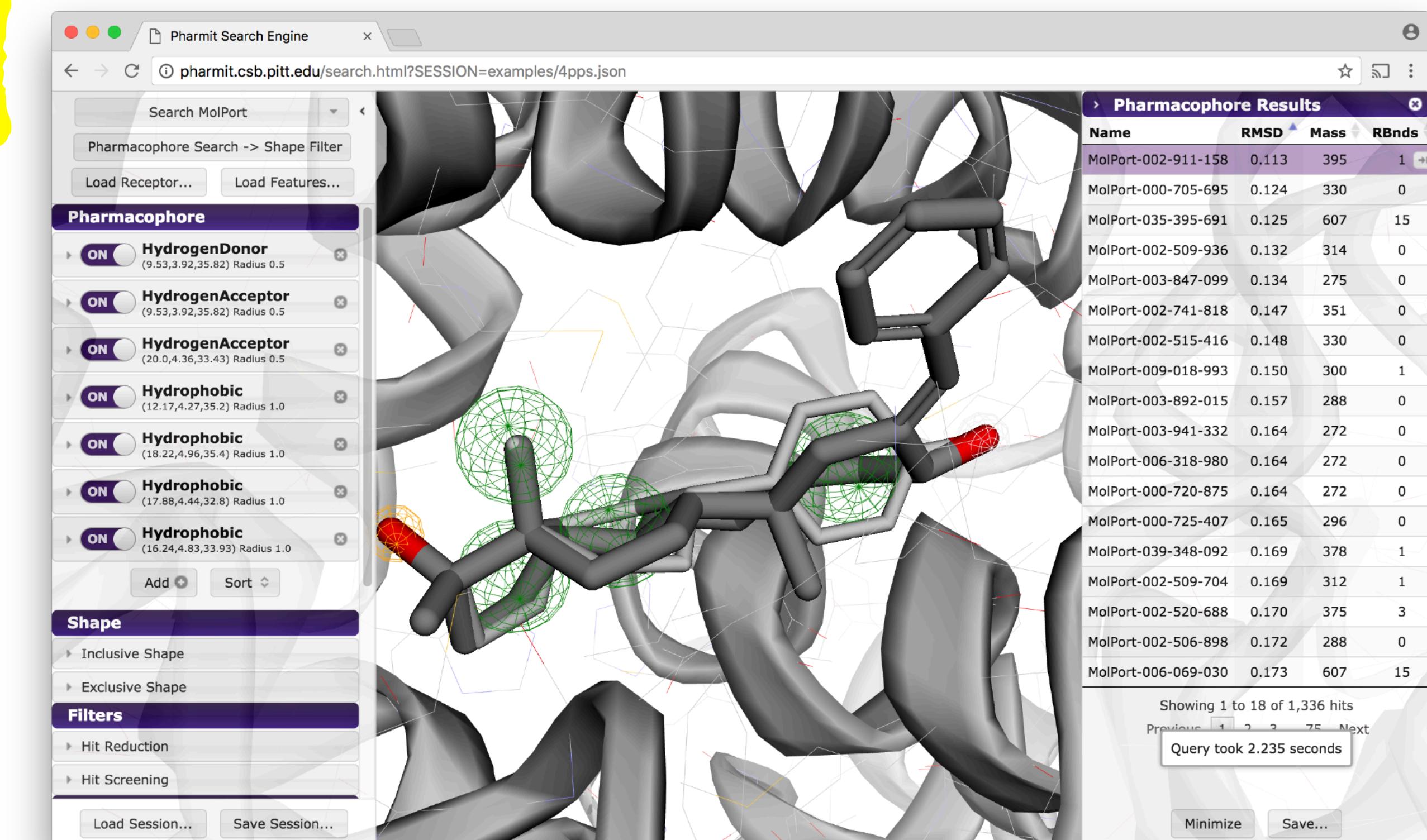
Purchasable



Accessible



Drug Discovery Funnel



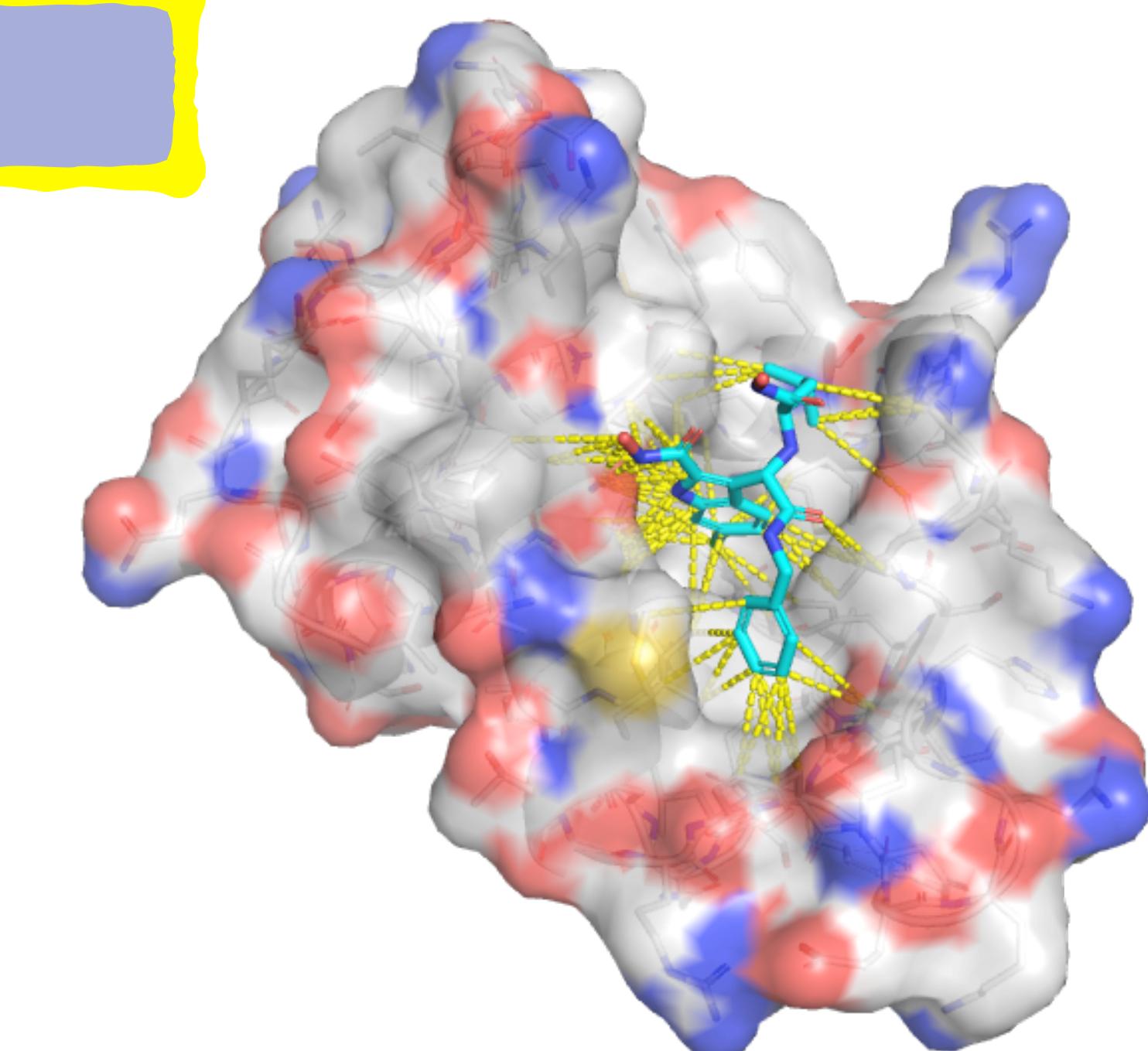
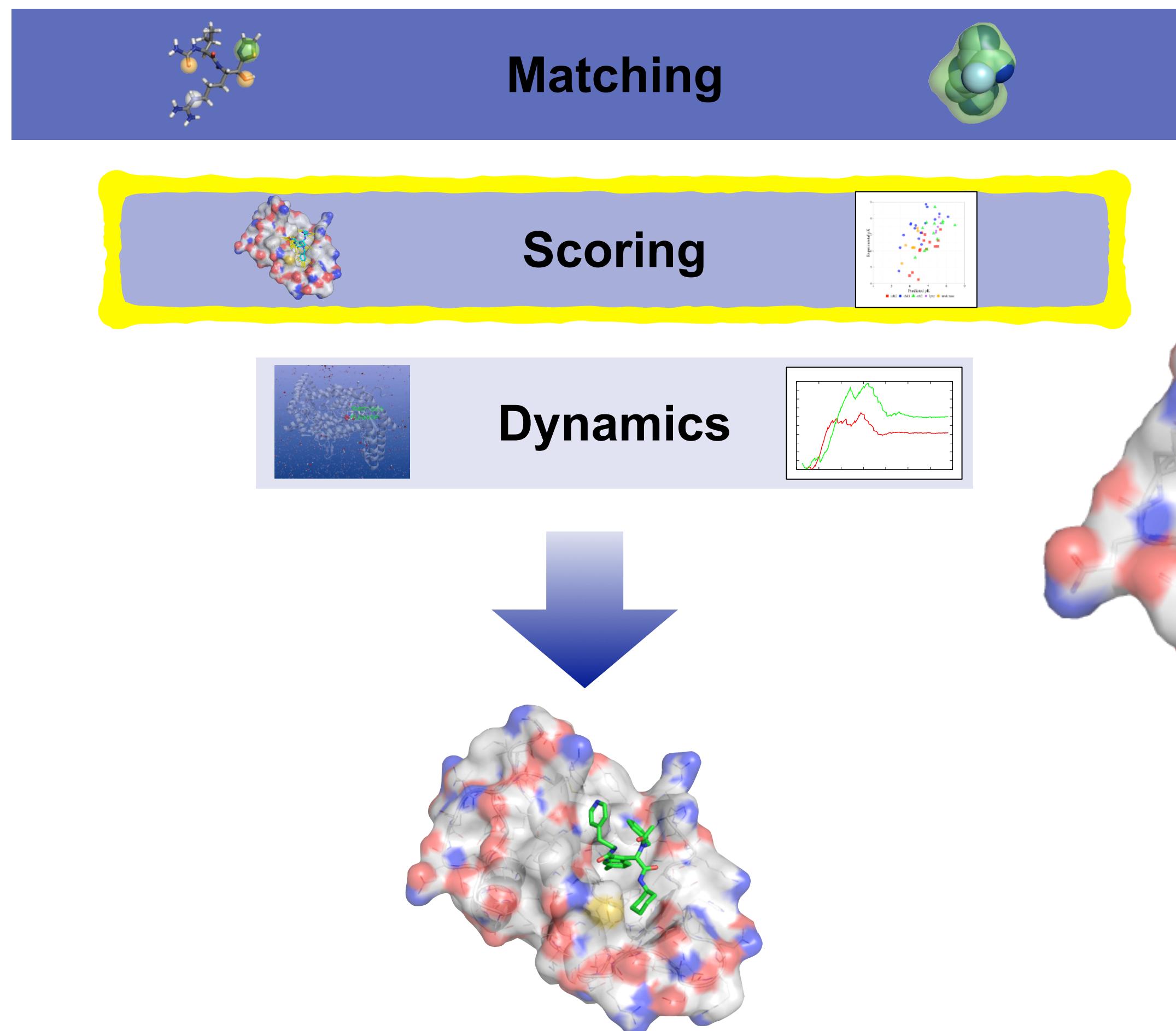
<http://pharmit.csb.pitt.edu>

Purchasable



Accessible

Drug Discovery Funnel



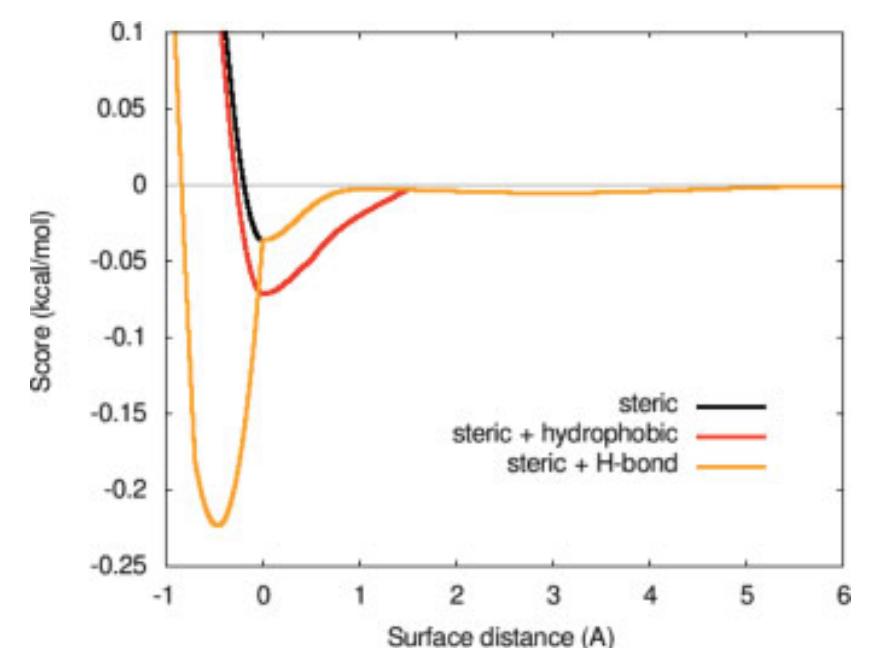
$$\text{gauss}_1(d) = w_{\text{gauss}_1} e^{-(d/0.5)^2}$$

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

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

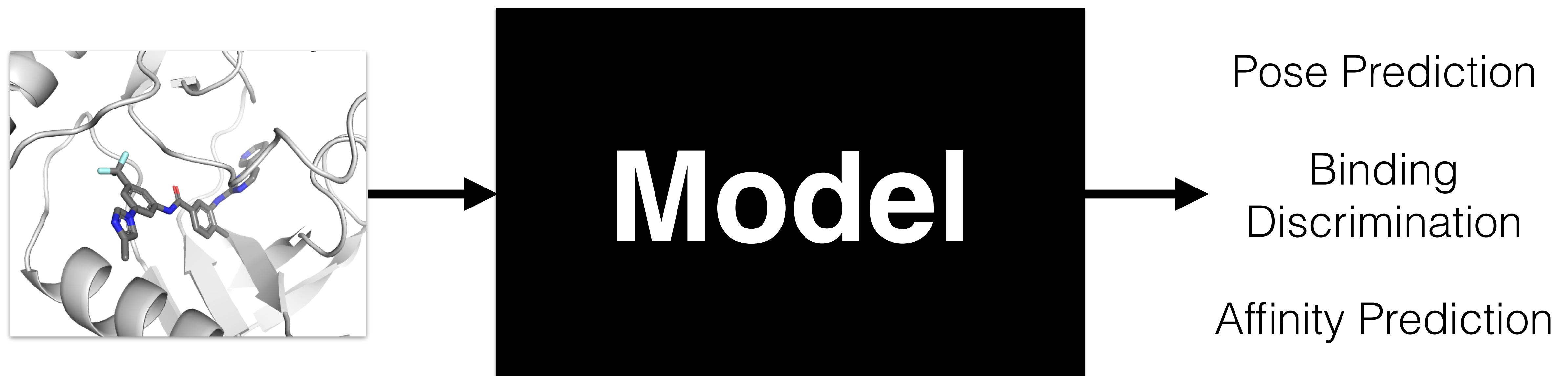
$$\text{hydrophobic}(d) = \begin{cases} w_{\text{hydrophobic}} & d < 0.5 \\ 0 & d > 1.5 \\ w_{\text{hydrophobic}}(1.5 - d) & \text{otherwise} \end{cases}$$

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

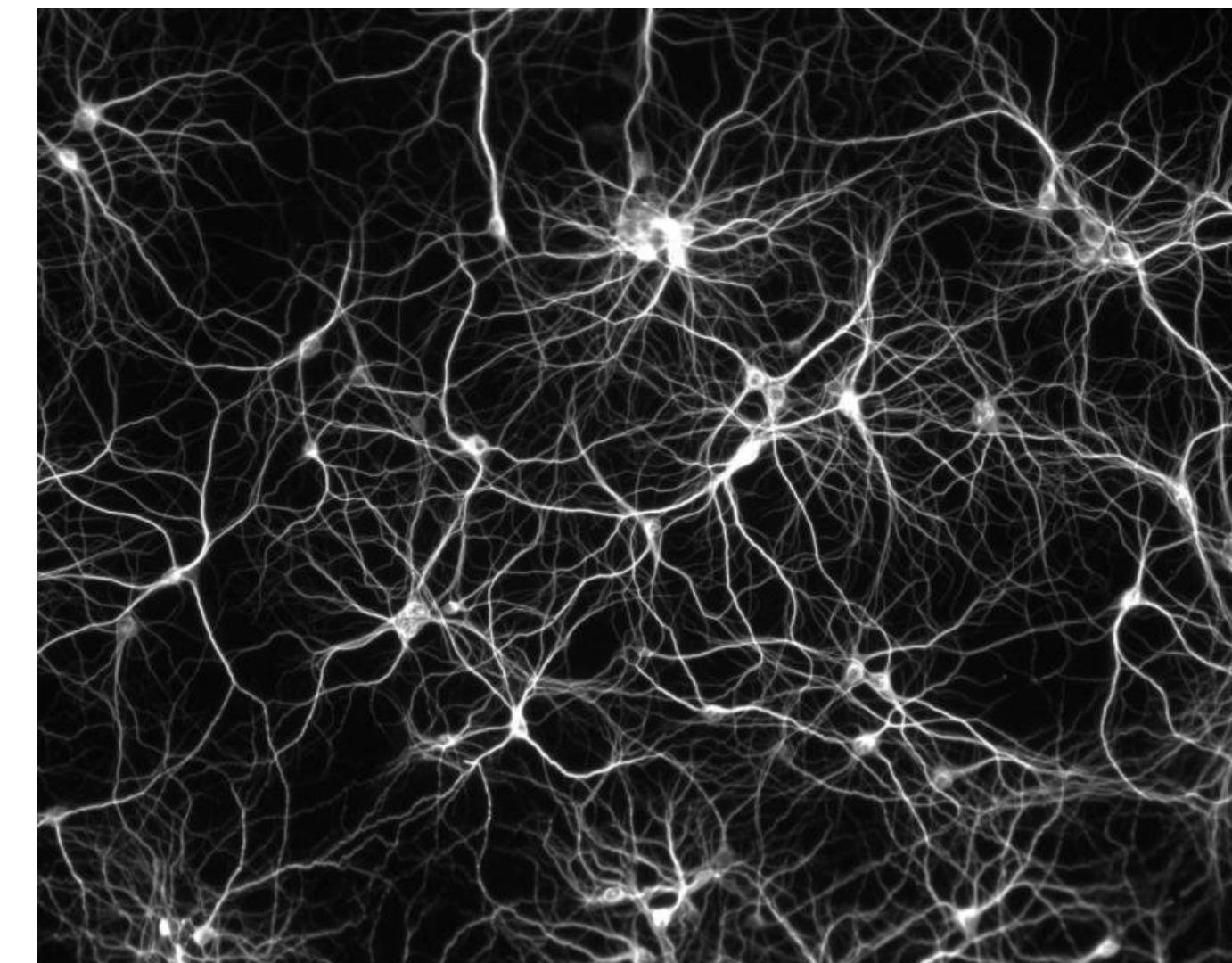
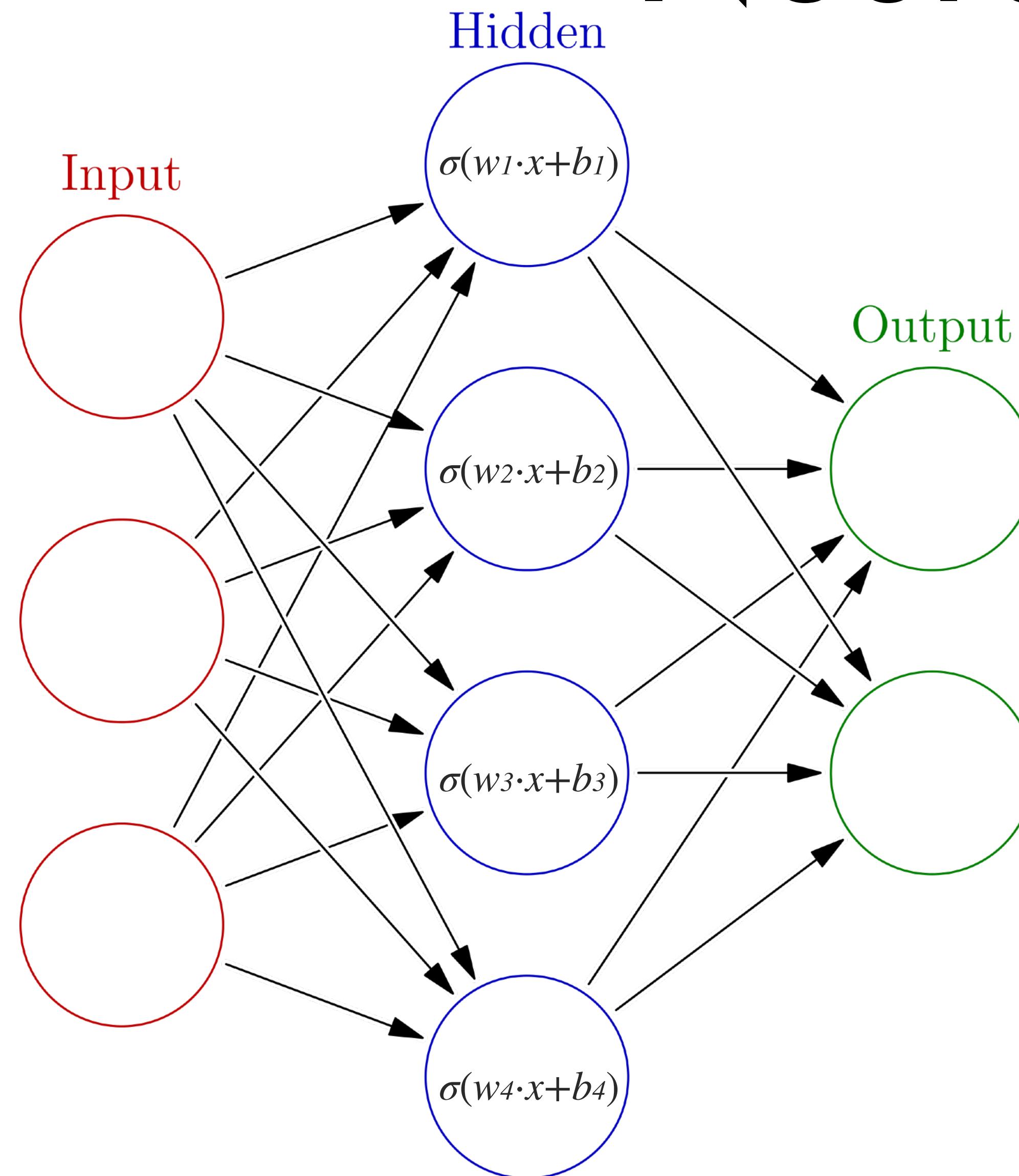


O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry* 31 (2010) 455-461

Protein-Ligand Scoring

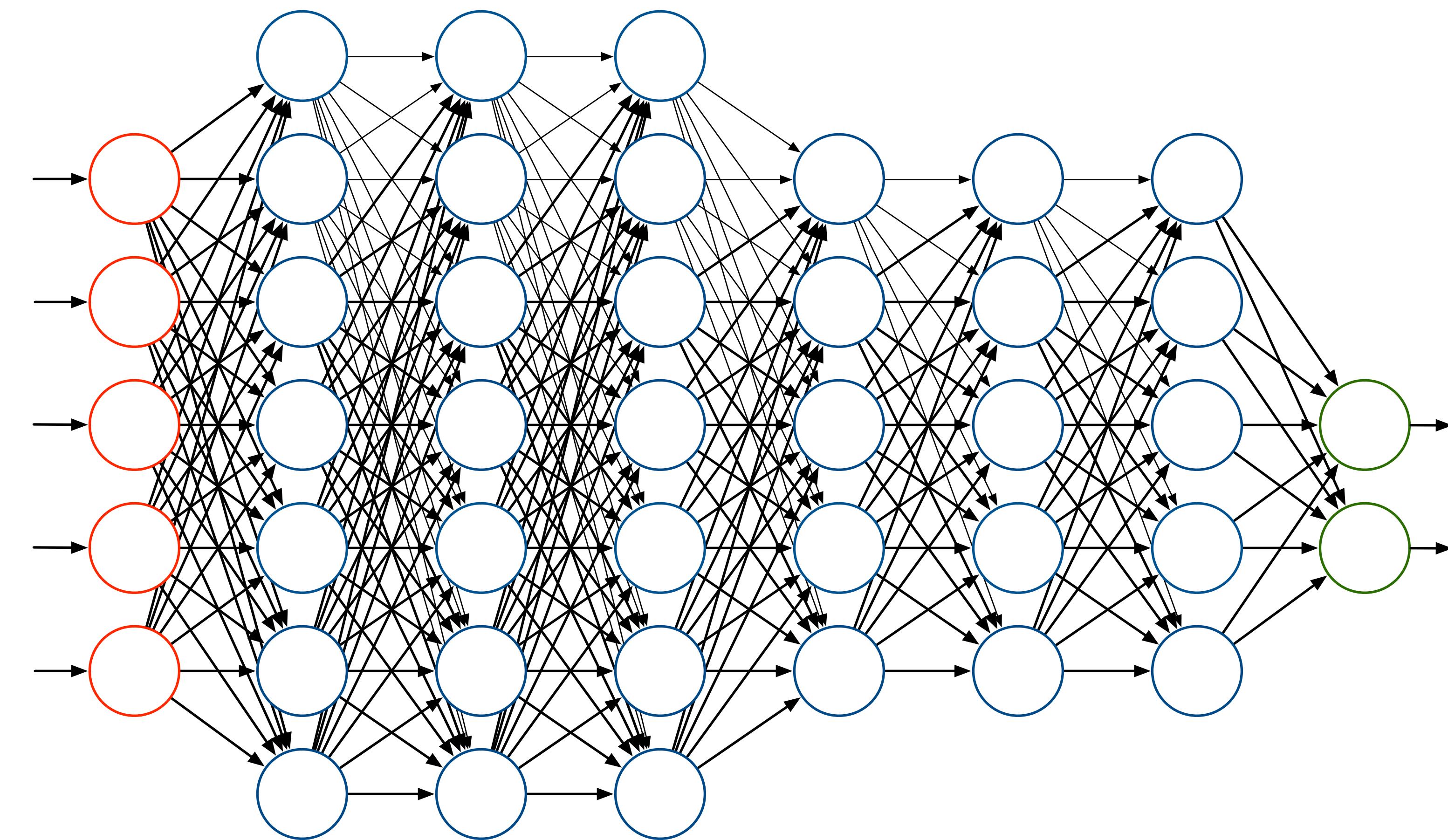


Neural Networks

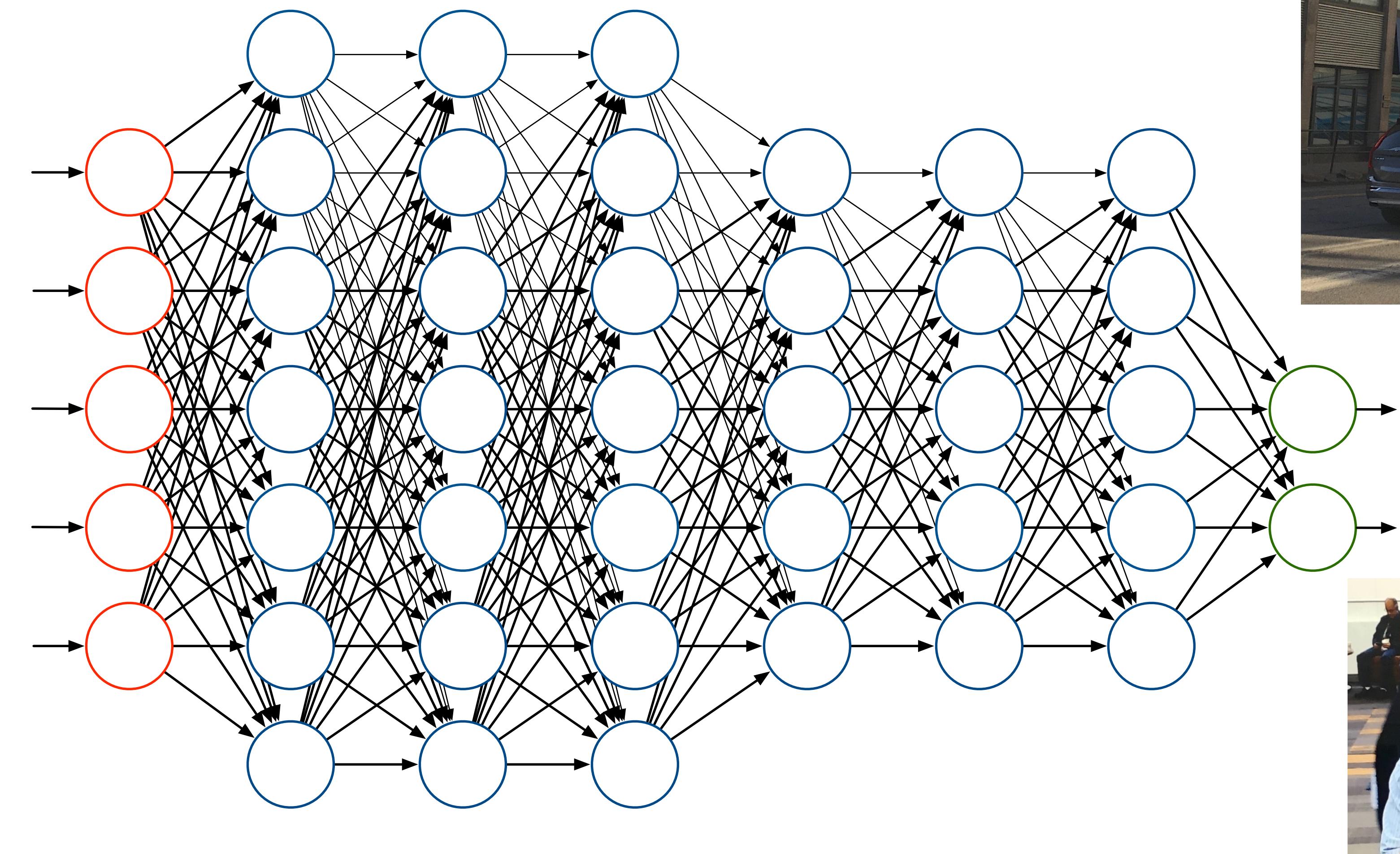


The **universal approximation theorem** states that, under reasonable assumptions, a feedforward **neural network** with a finite number of nodes **can approximate any continuous** function to within a given error over a bounded input domain.

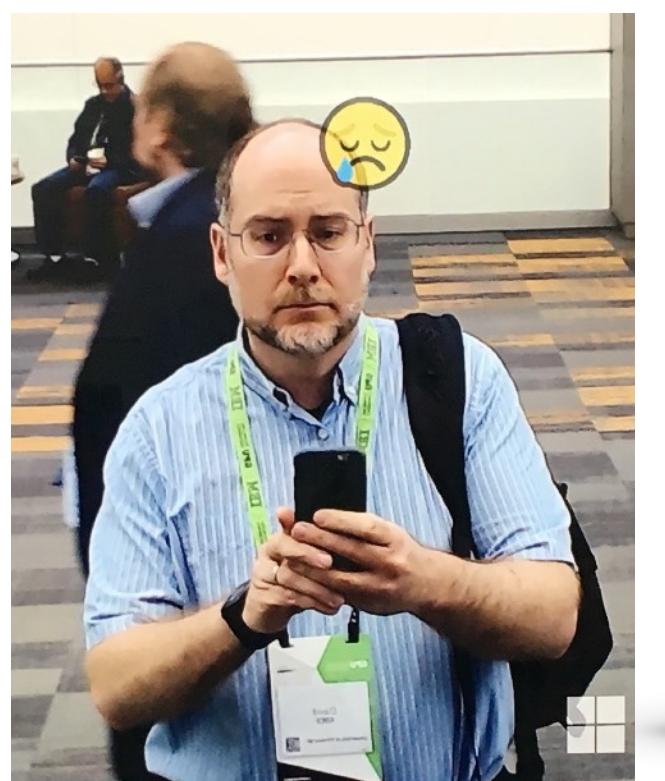
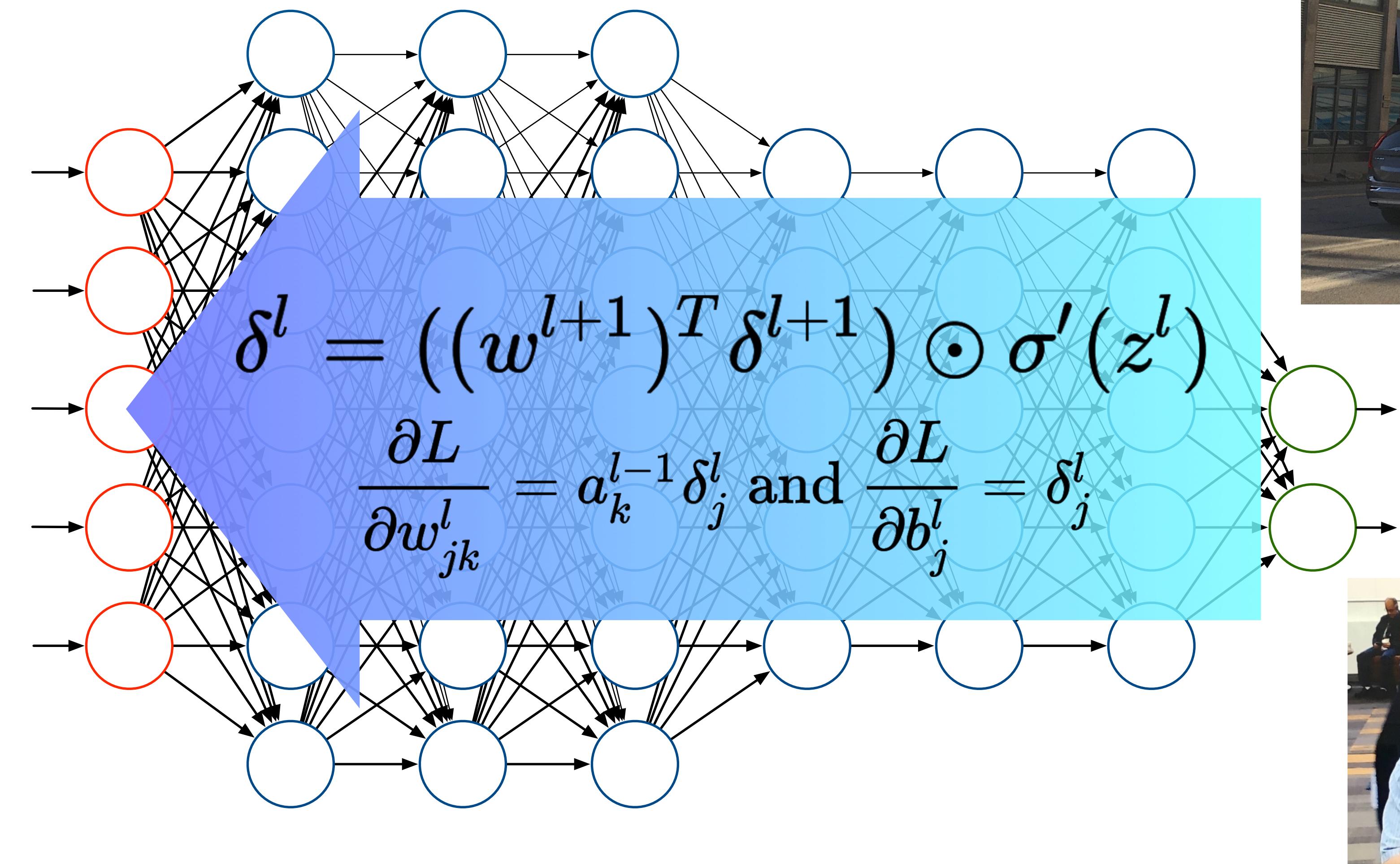
Deep Learning



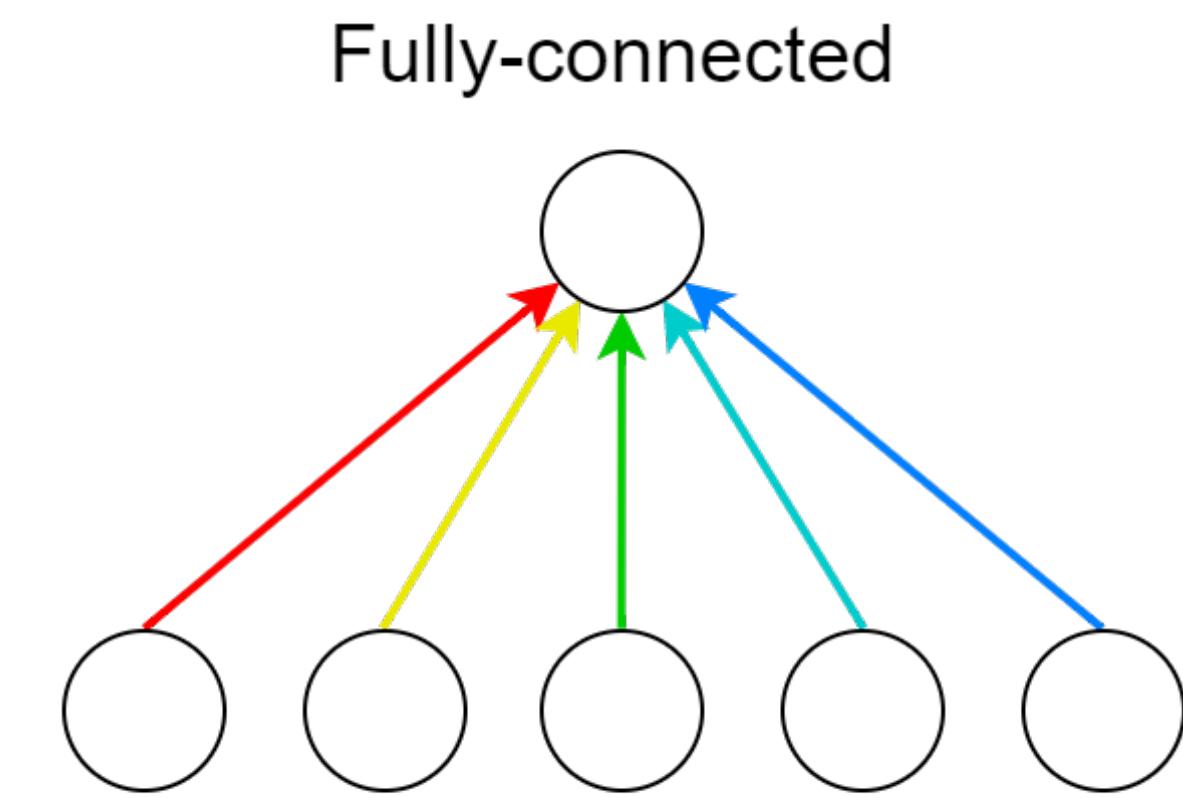
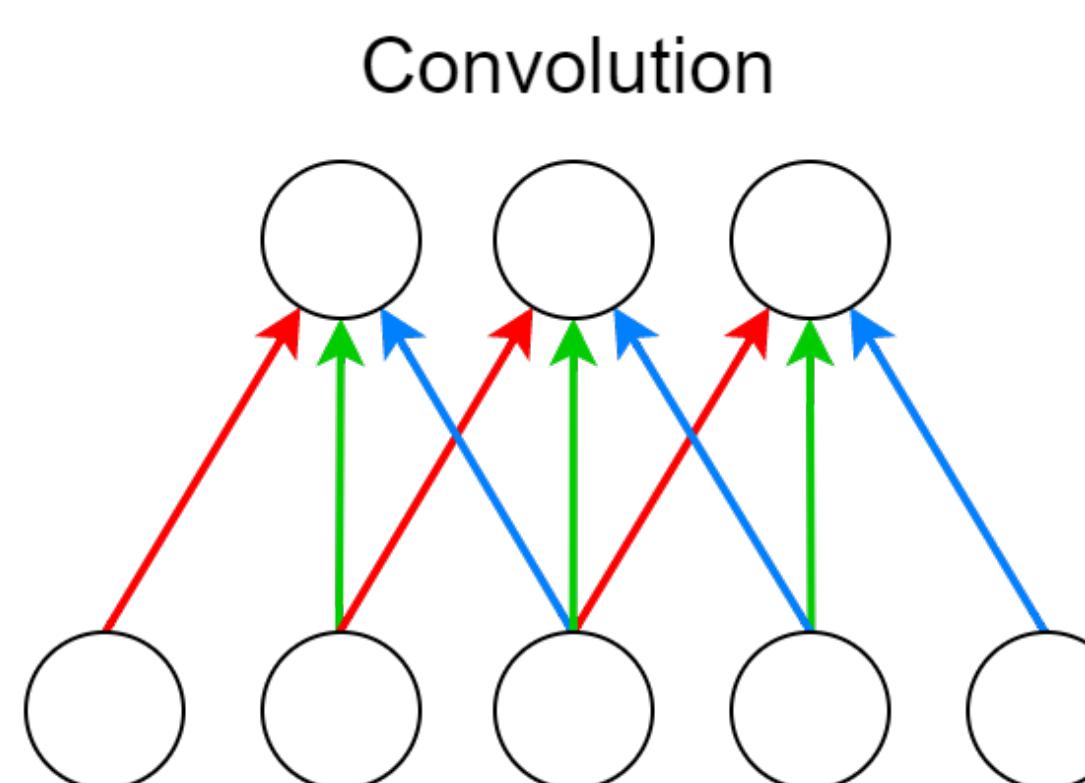
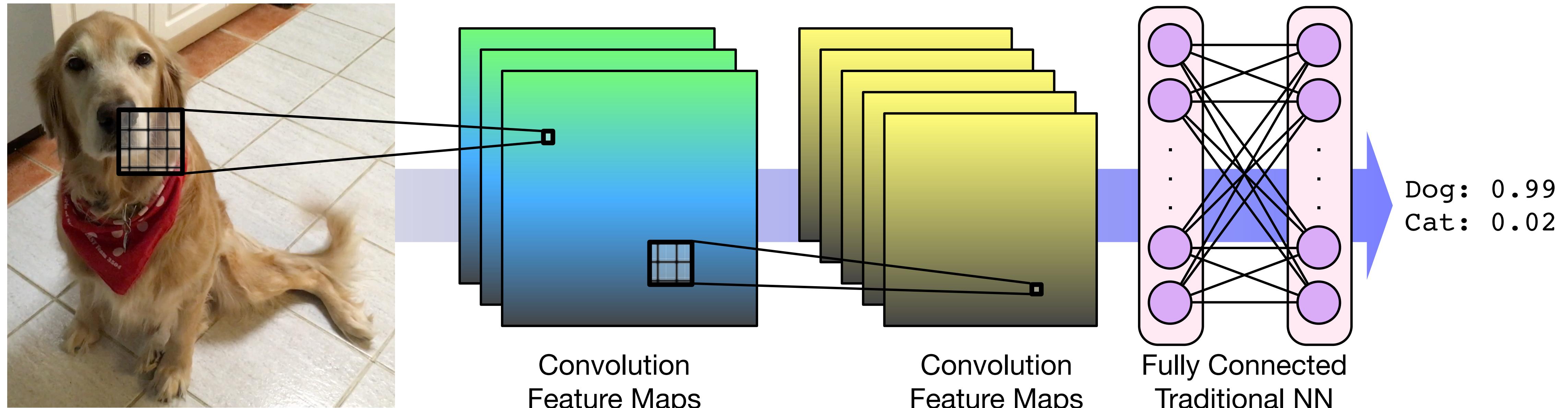
Deep Learning



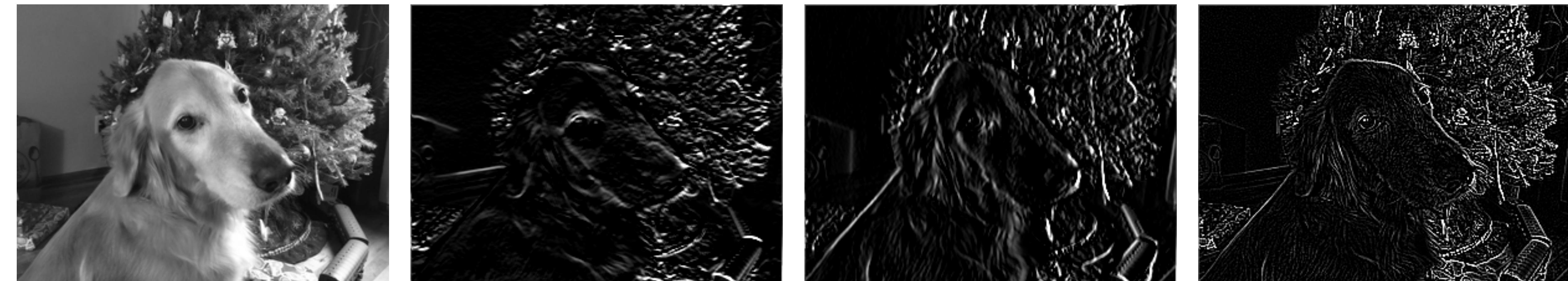
Deep Learning



Convolutional Neural Networks



Convolutional Filters

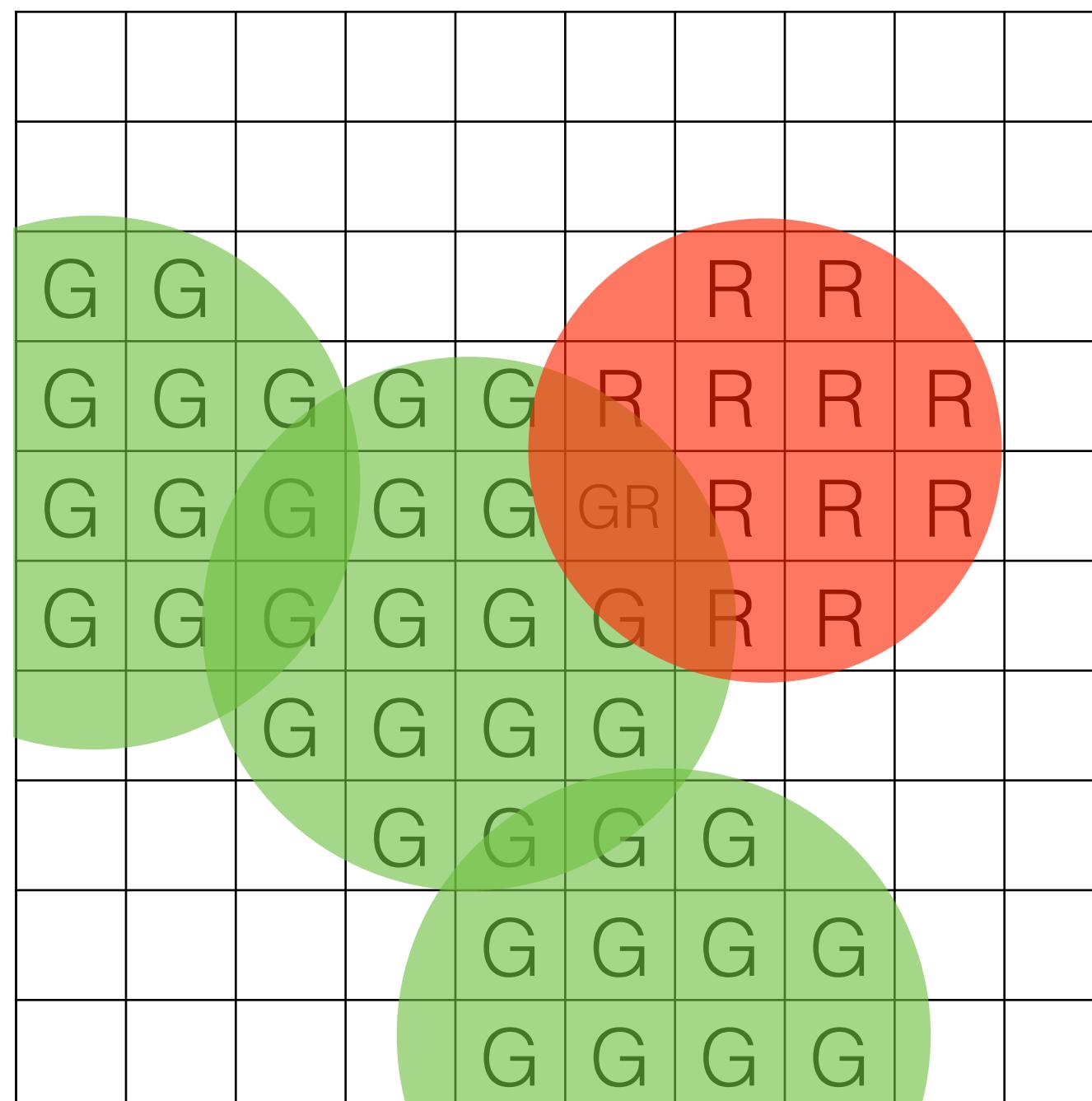


$$\begin{array}{ccc} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{array}$$

$$\begin{array}{ccc} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{array}$$

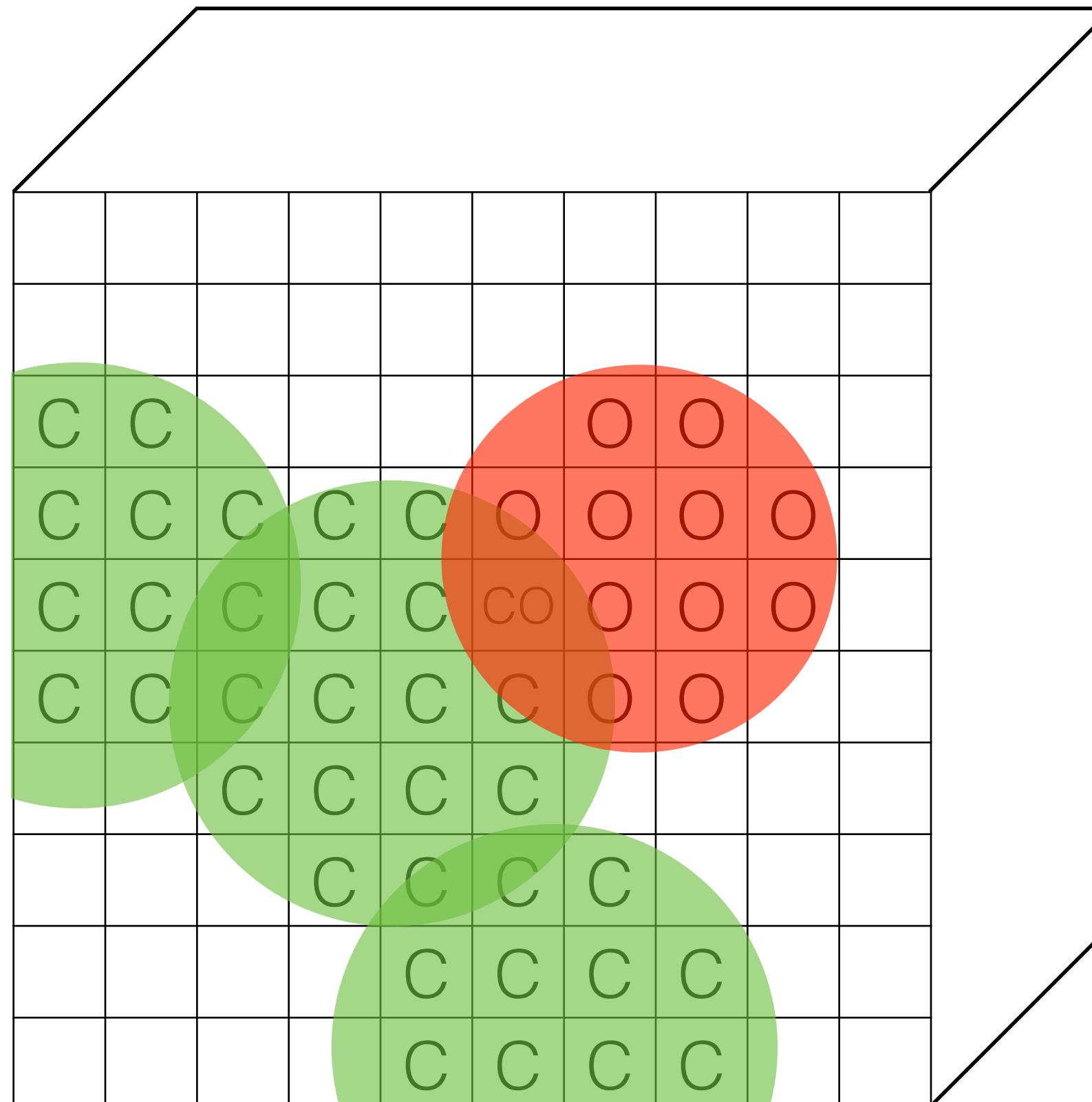
$$\begin{array}{ccc} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{array}$$

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**.

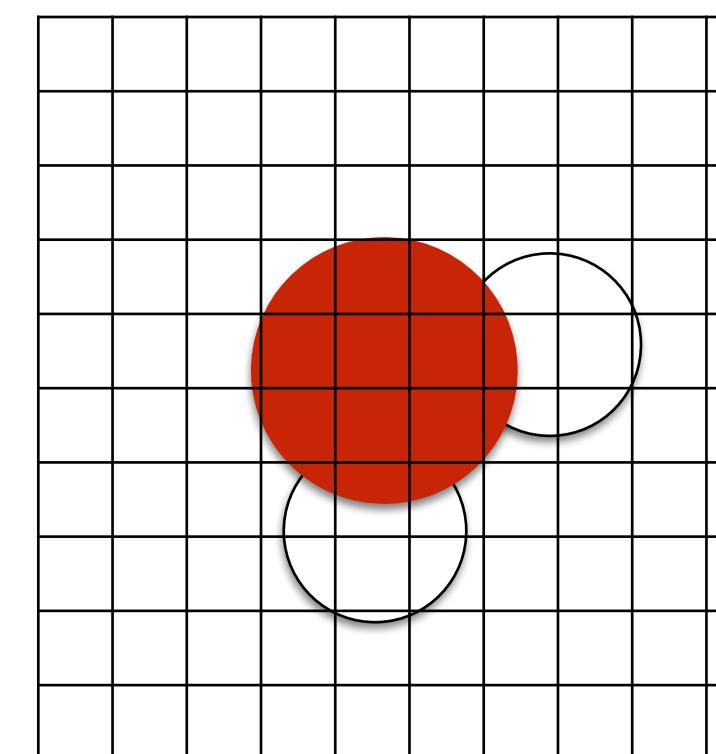
Why Grids?

Cons

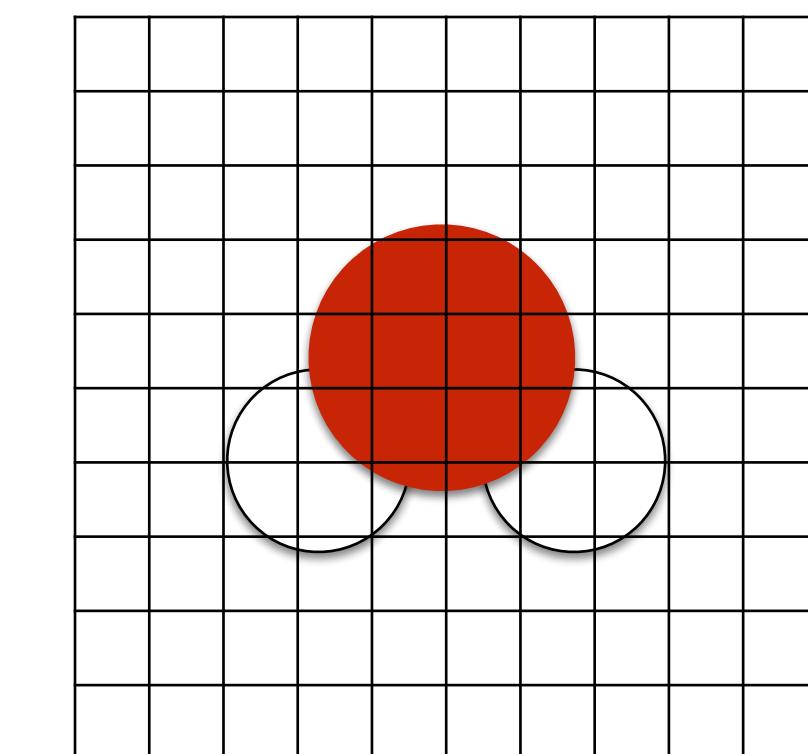
- coordinate frame dependent
- pairwise interactions not explicit

Pros

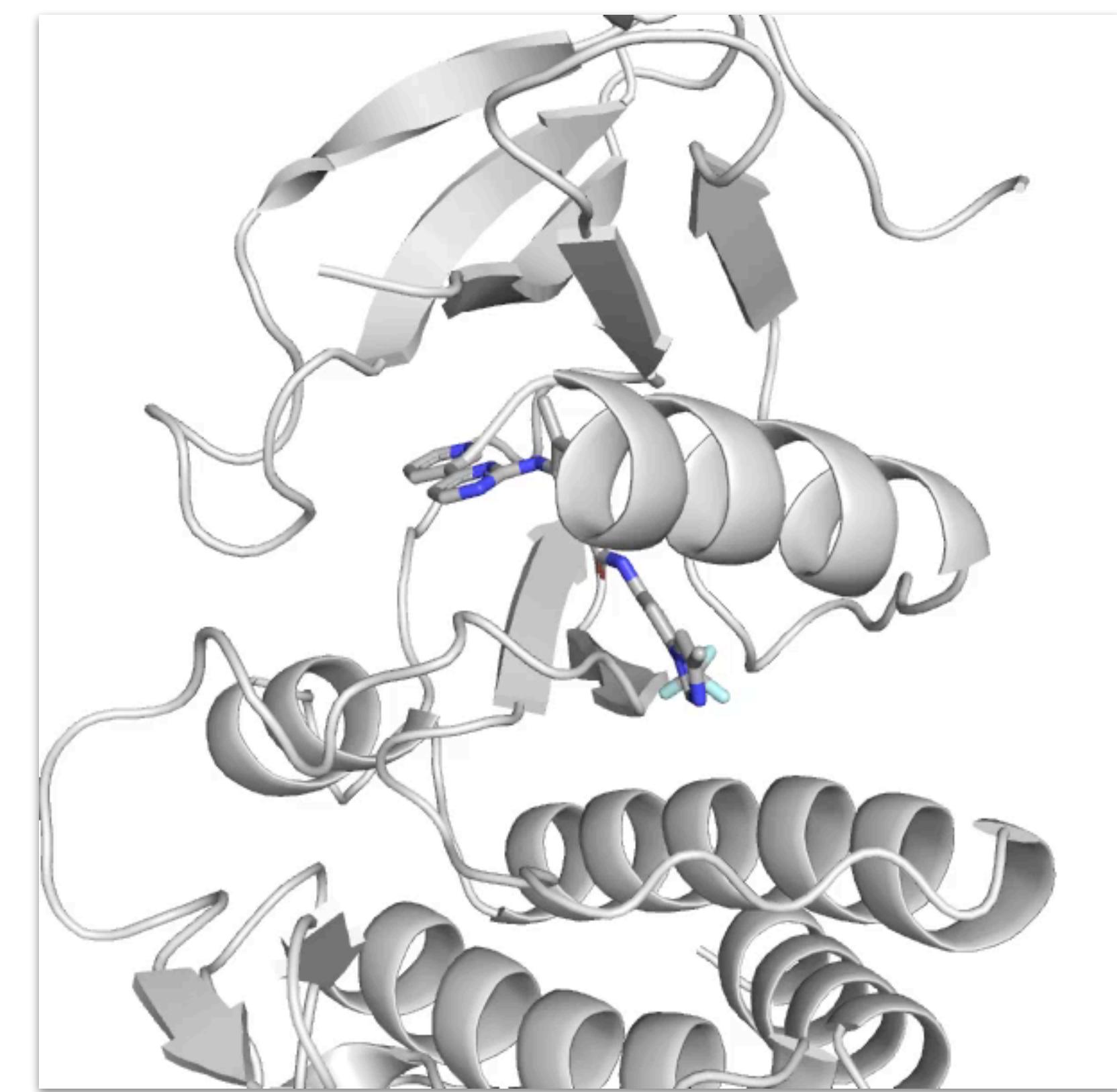
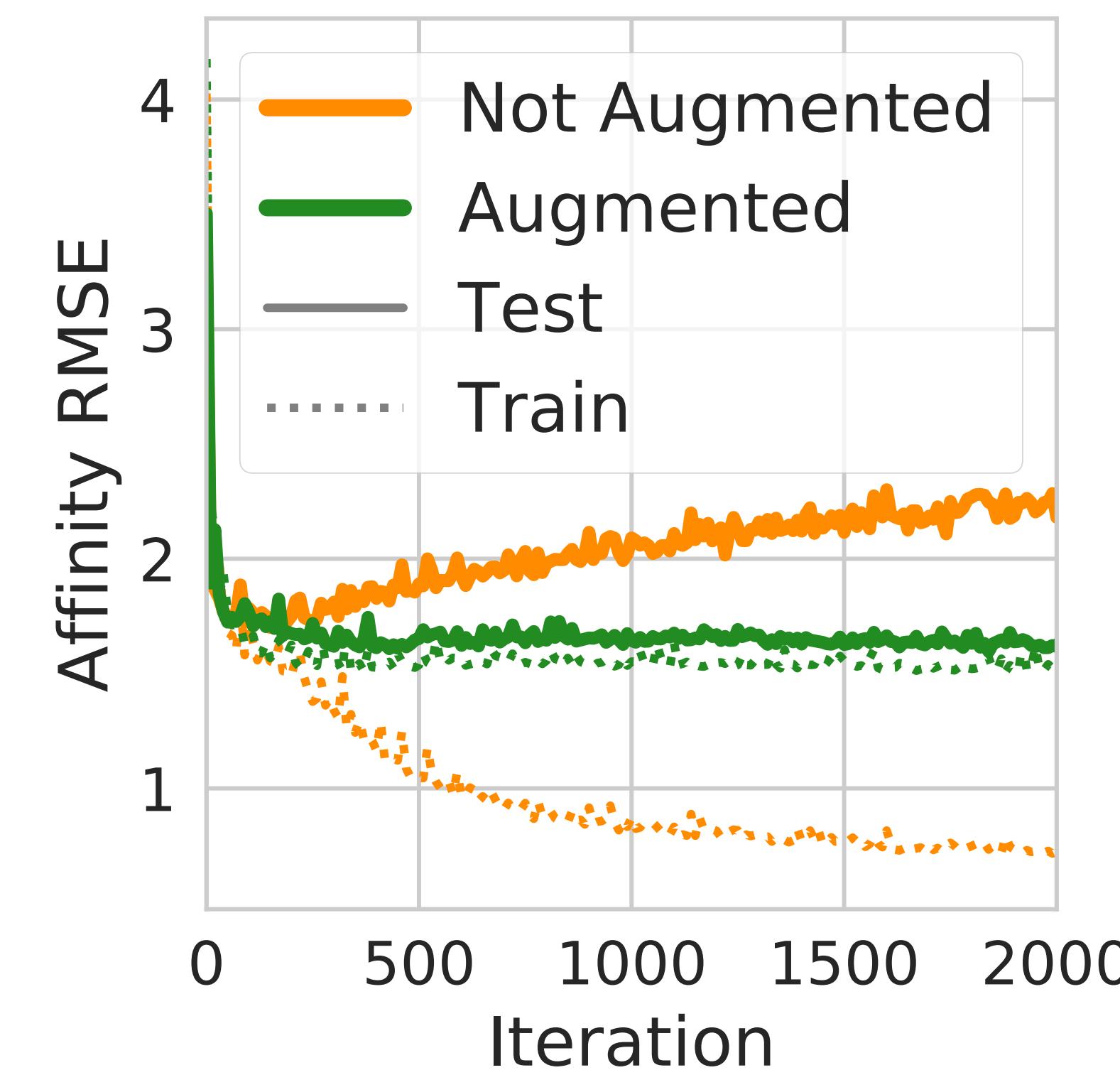
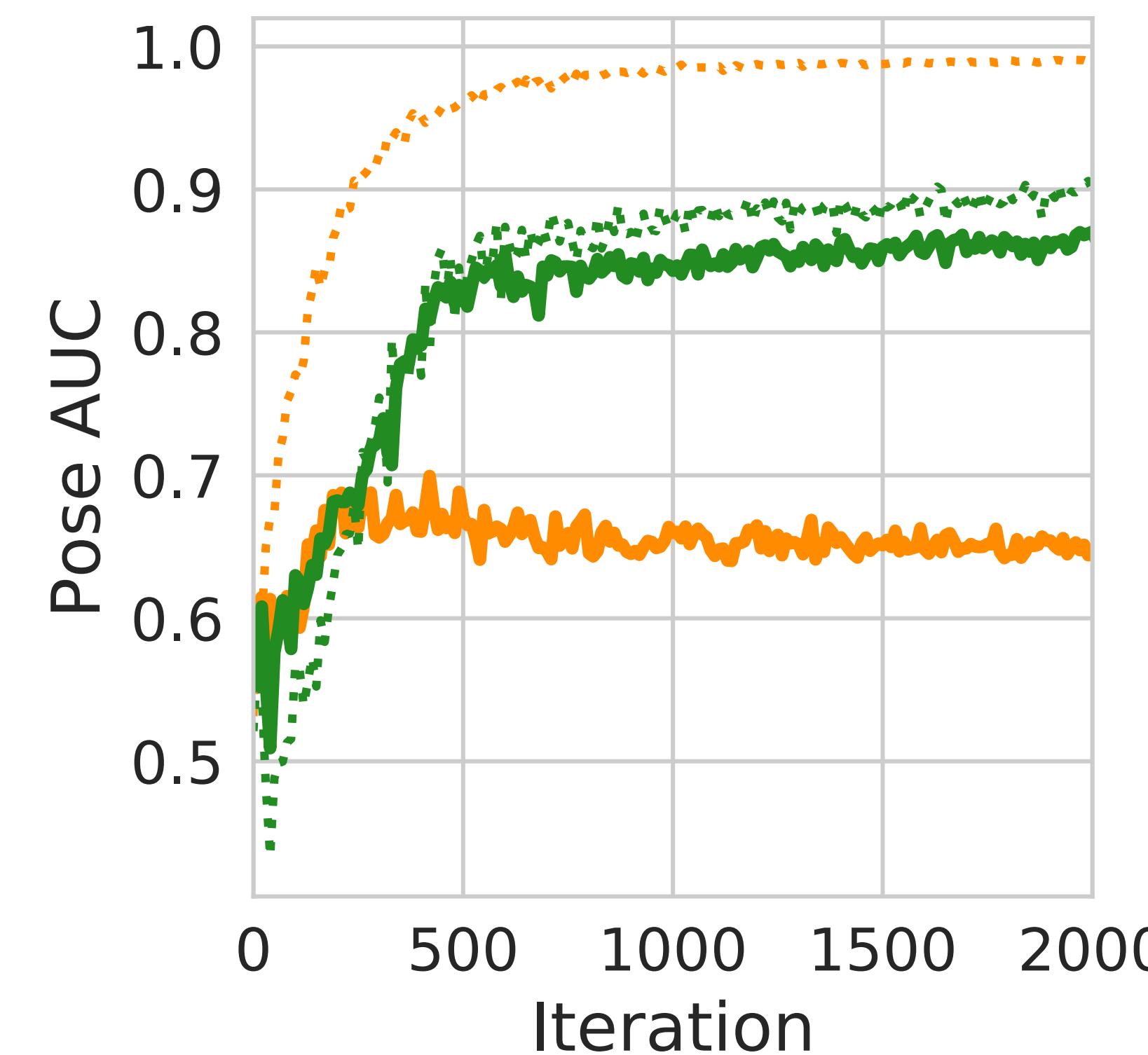
- clear spatial relationships
- amazingly parallel
- easy to interpret



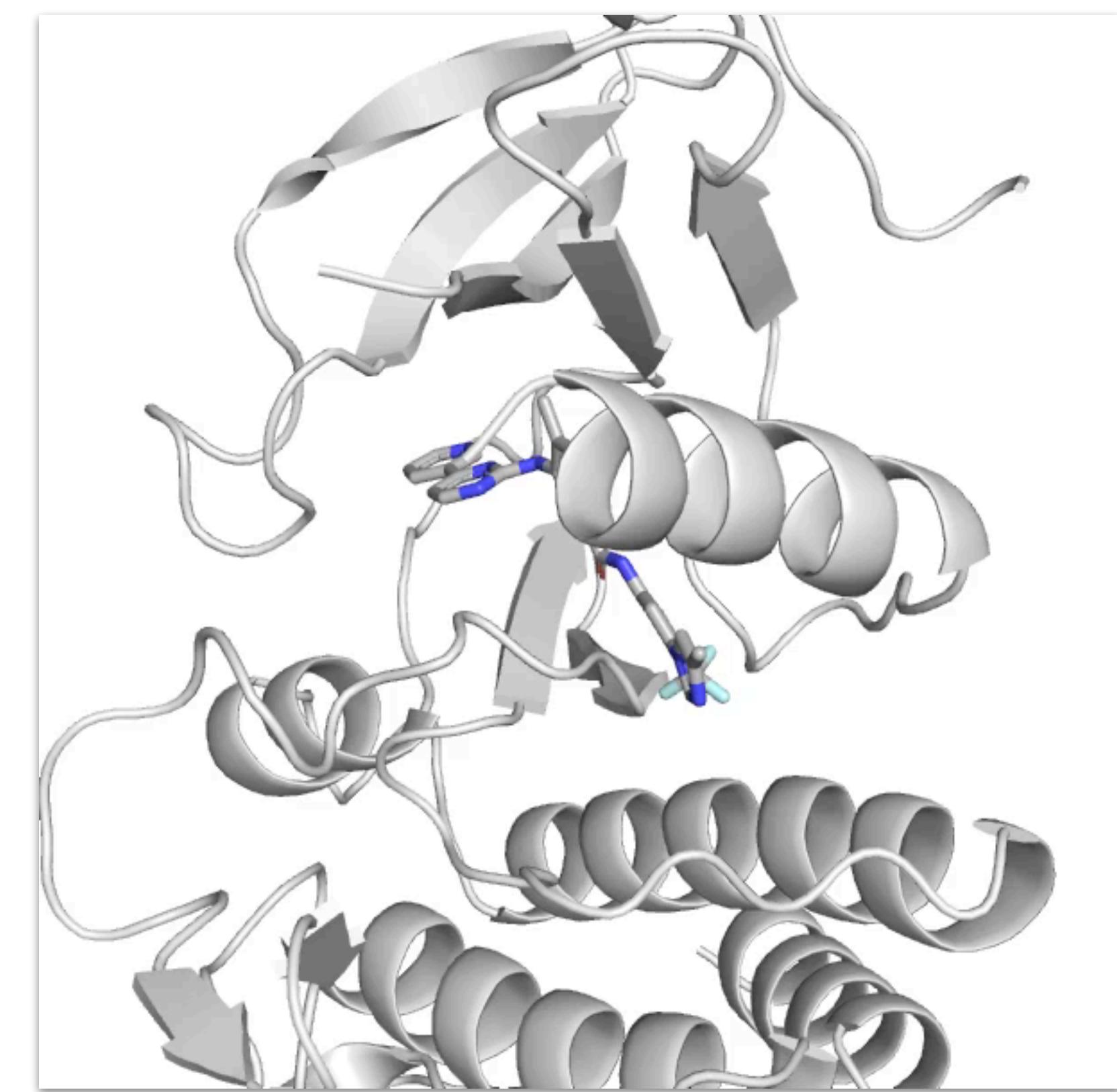
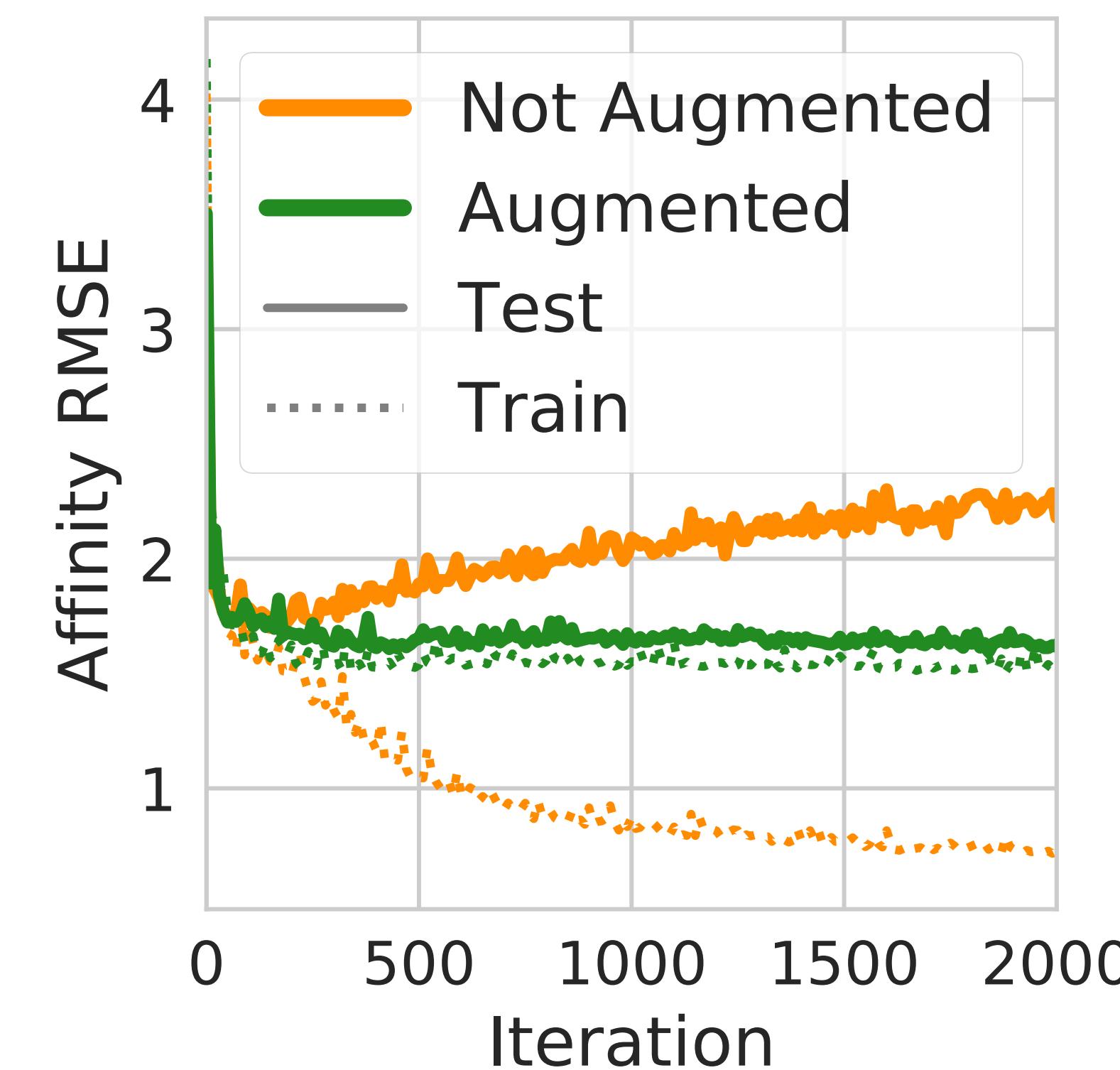
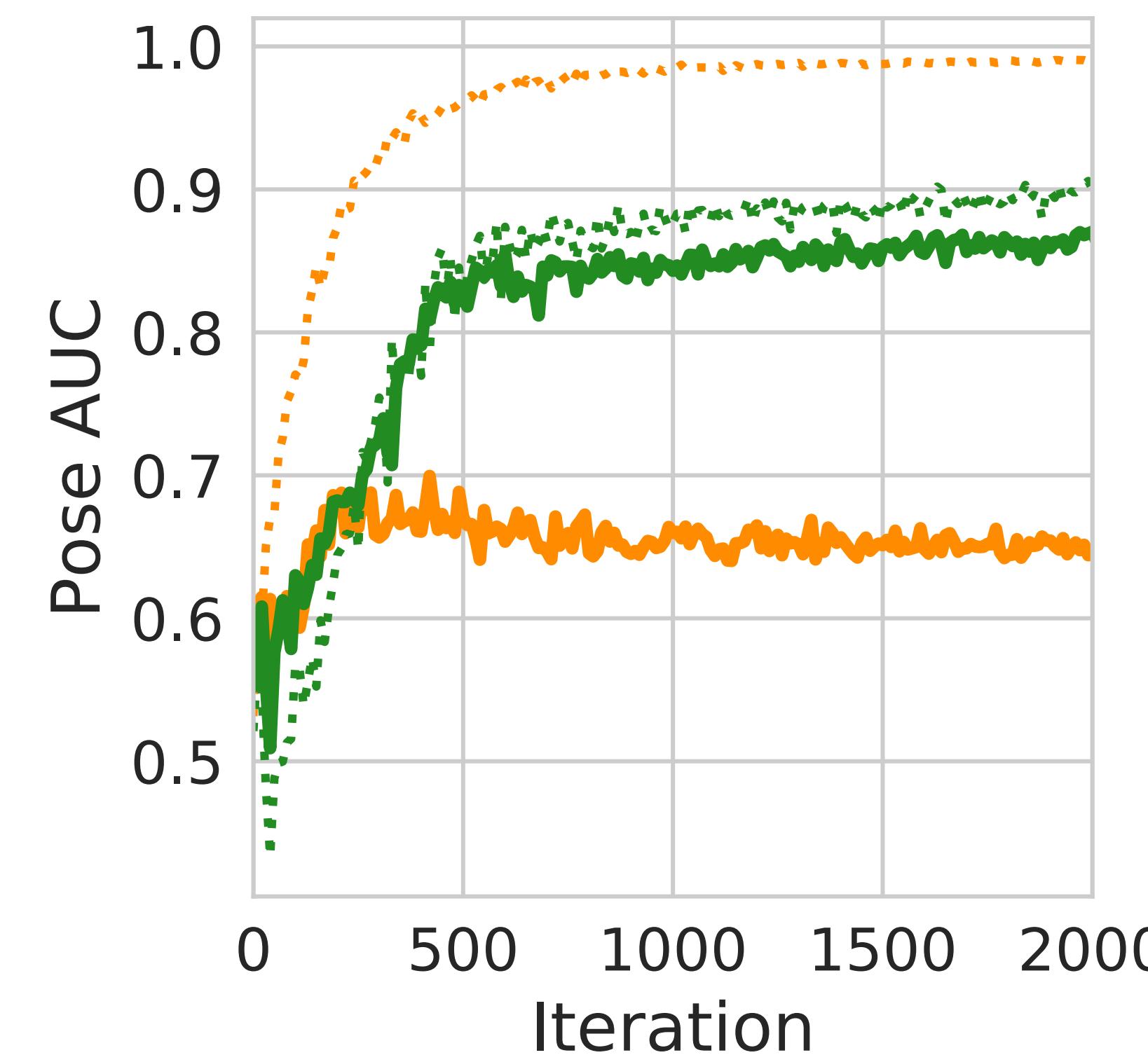
≠



Data Augmentation



Data Augmentation



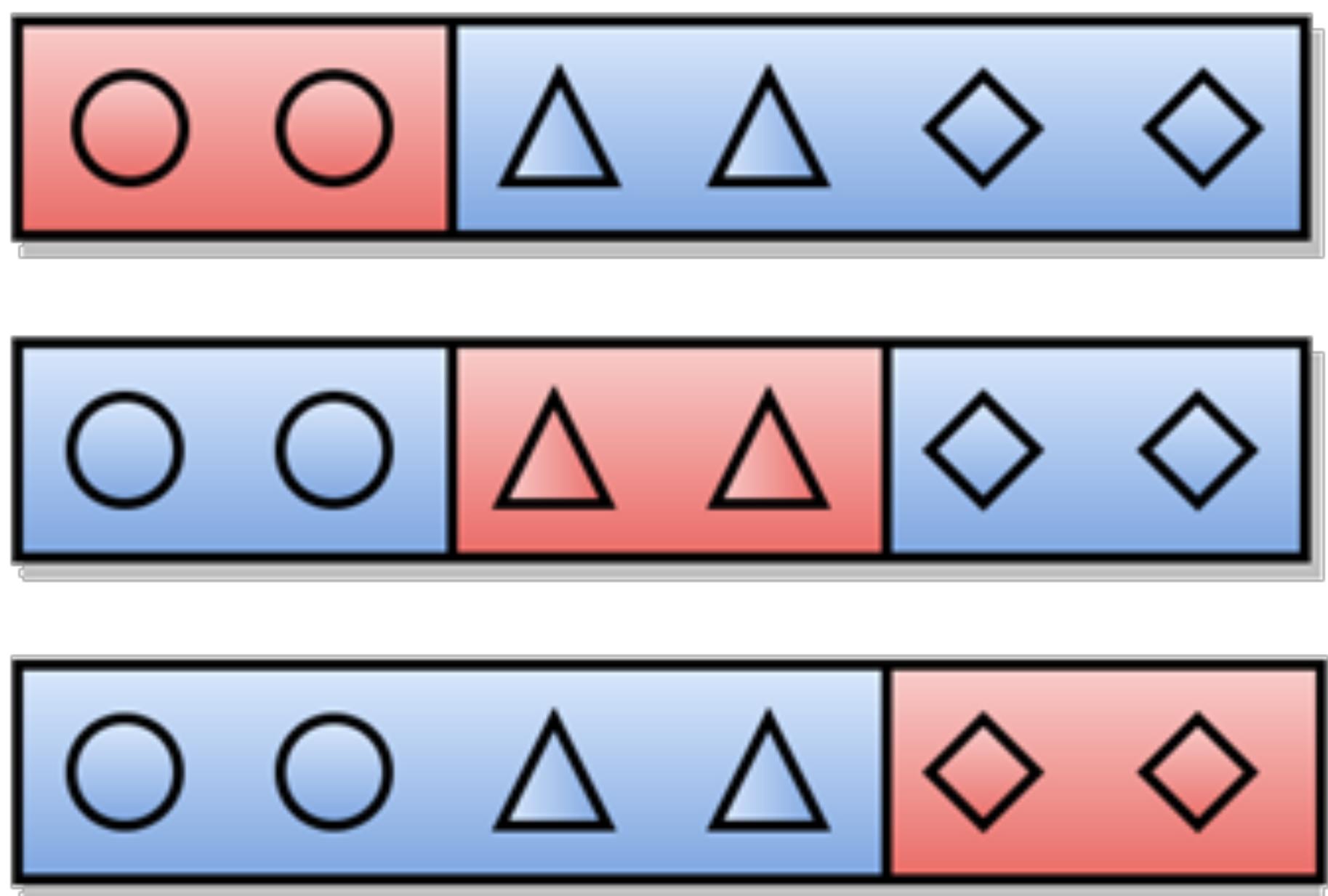
Training



PDBbind 2016 refined set

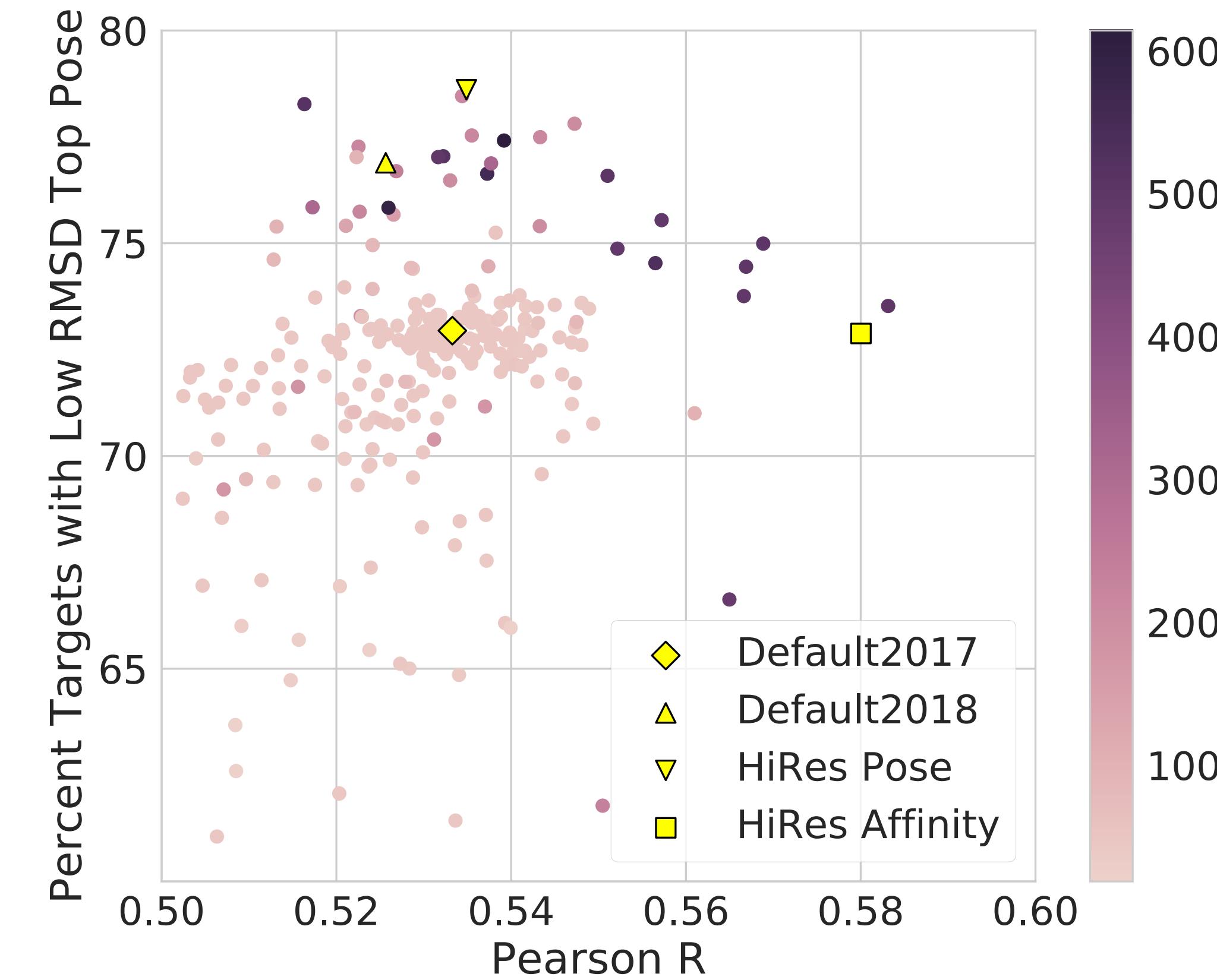
- **4056** protein-ligand complexes
- diverse targets
- wide range of affinities
- generate poses with AutoDock Vina
- include minimized crystal pose

Clustered Cross-validation

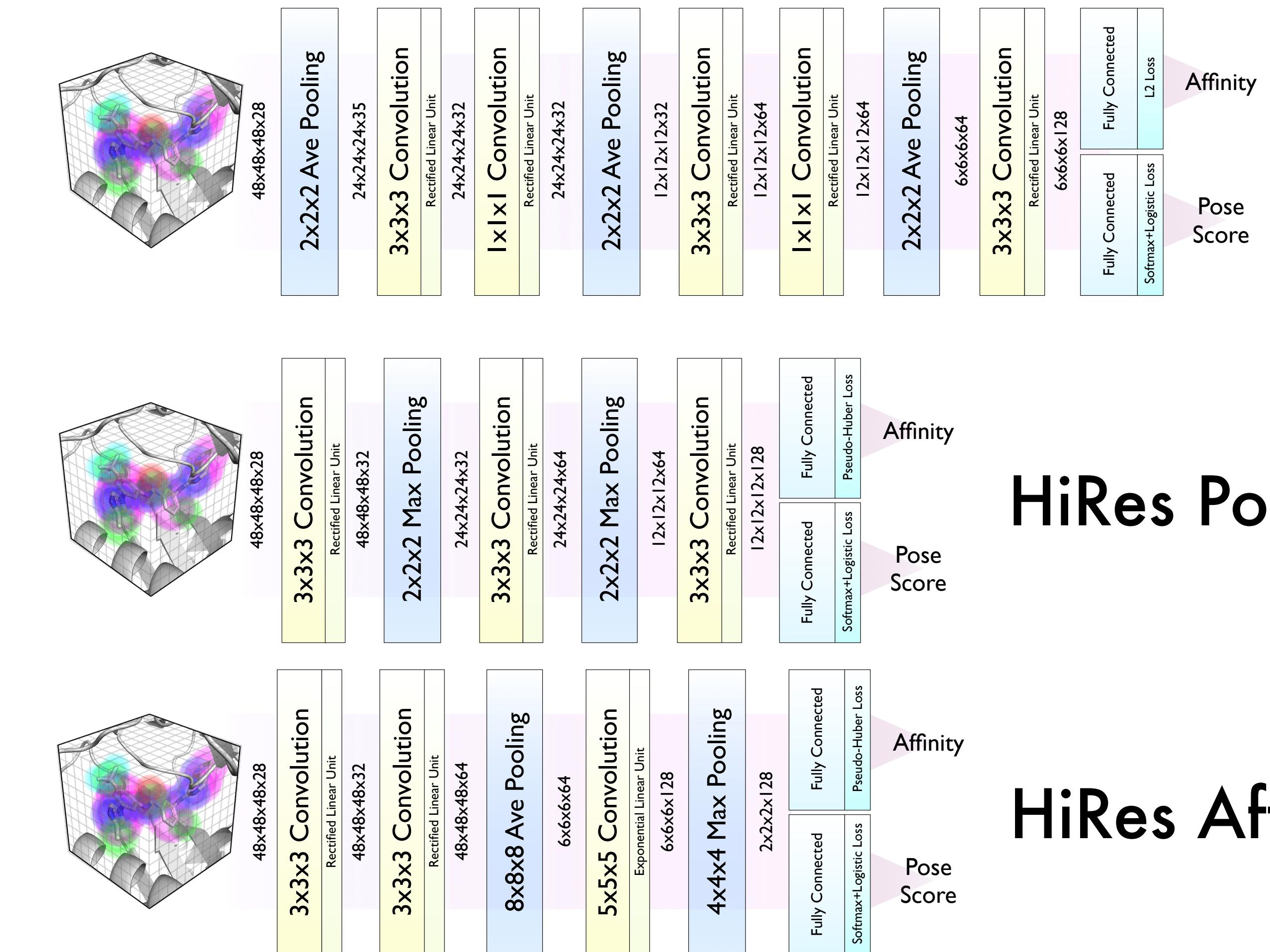


Target sequence similarity < 0.5
AND
Ligand similarity < 0.9

Optimized Models

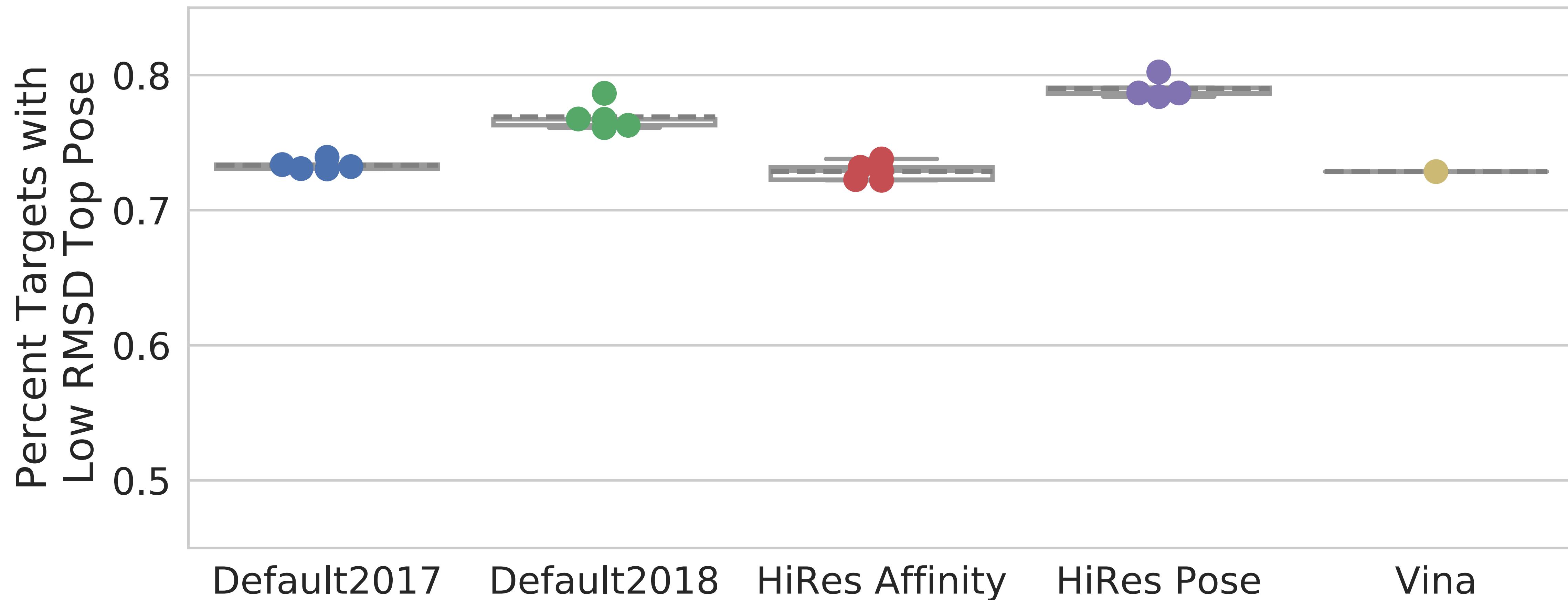


Default2018



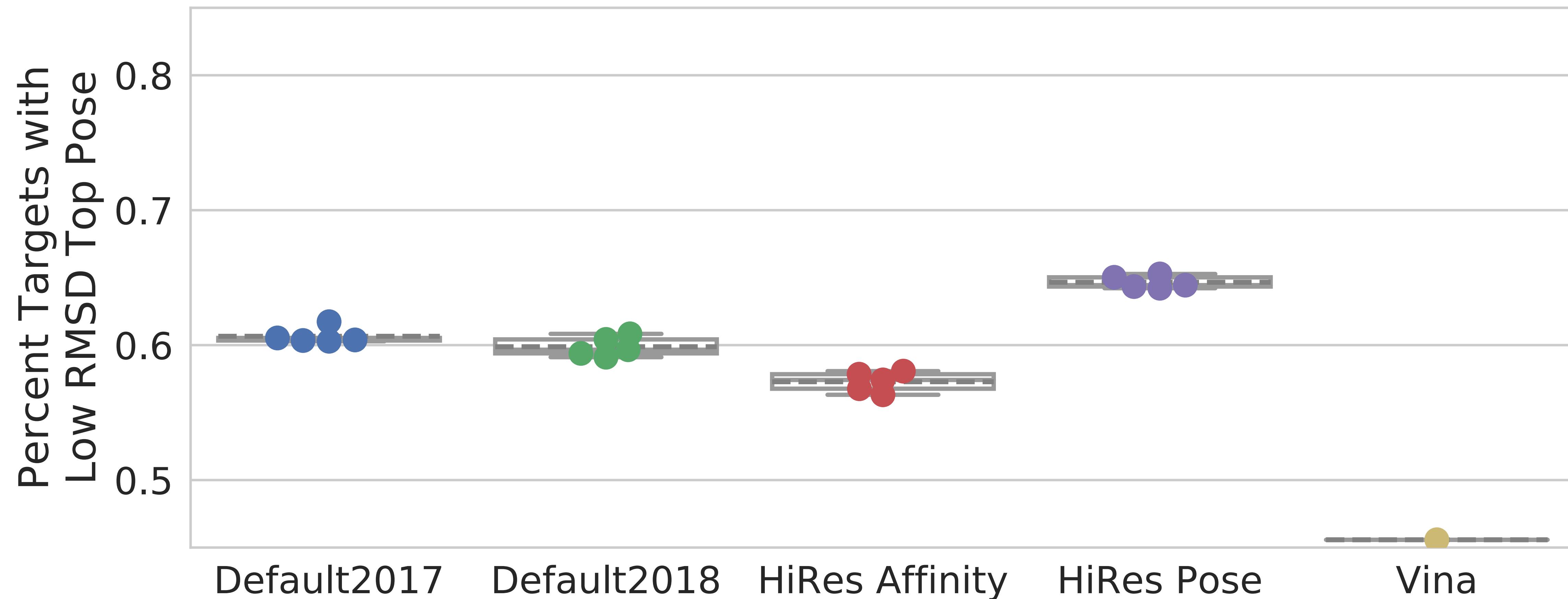
Pose Results

Redocked Pose

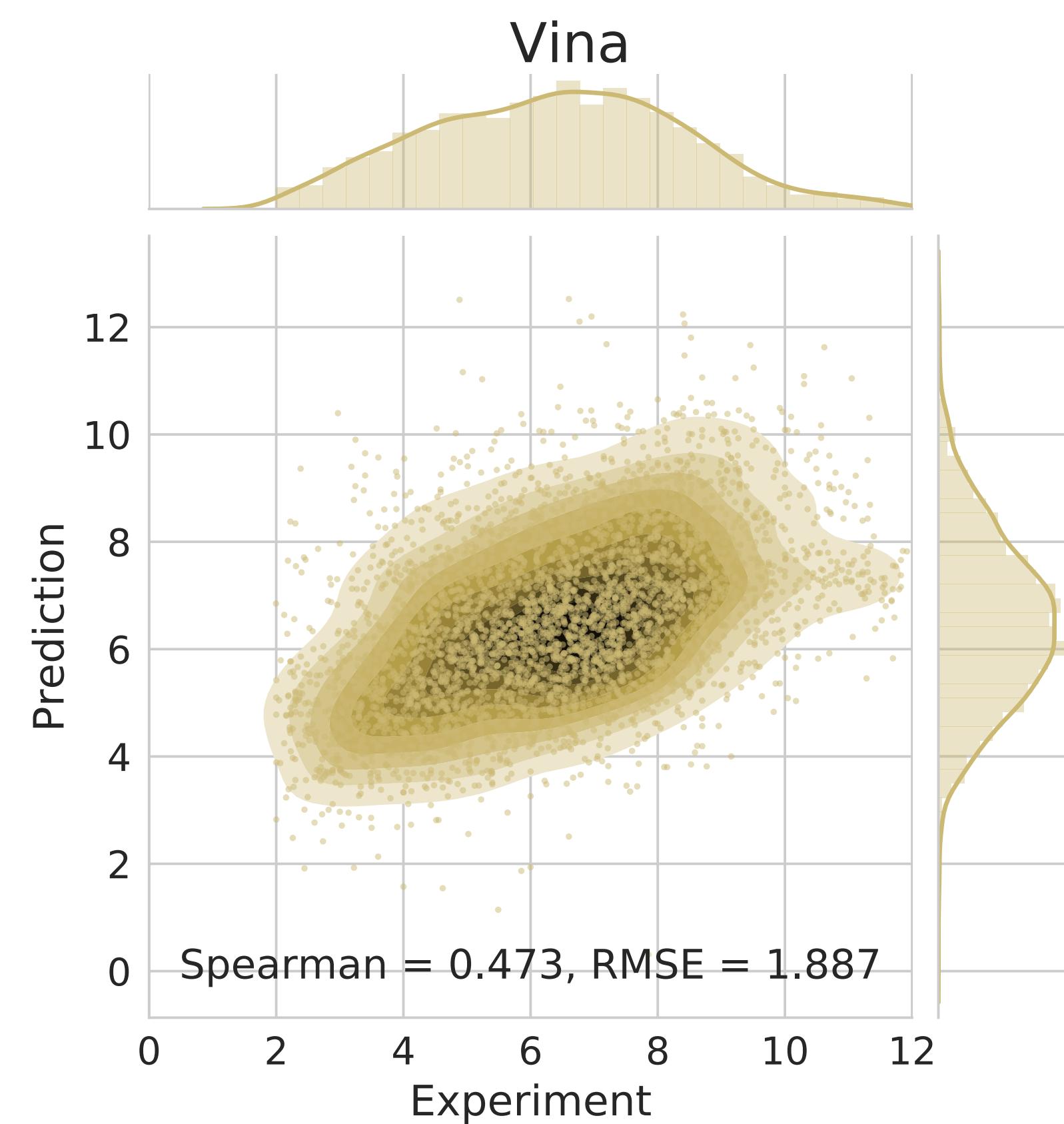
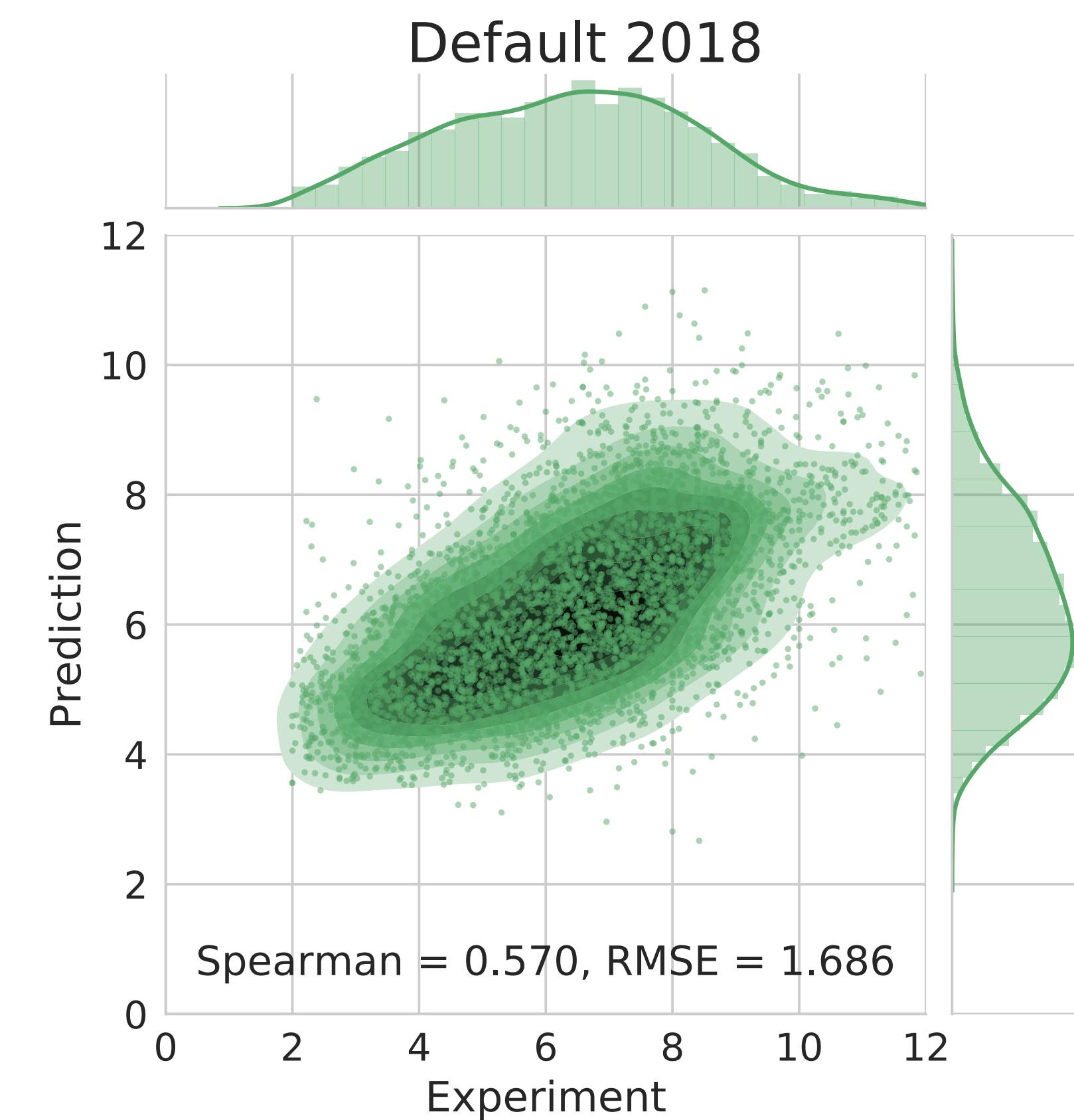
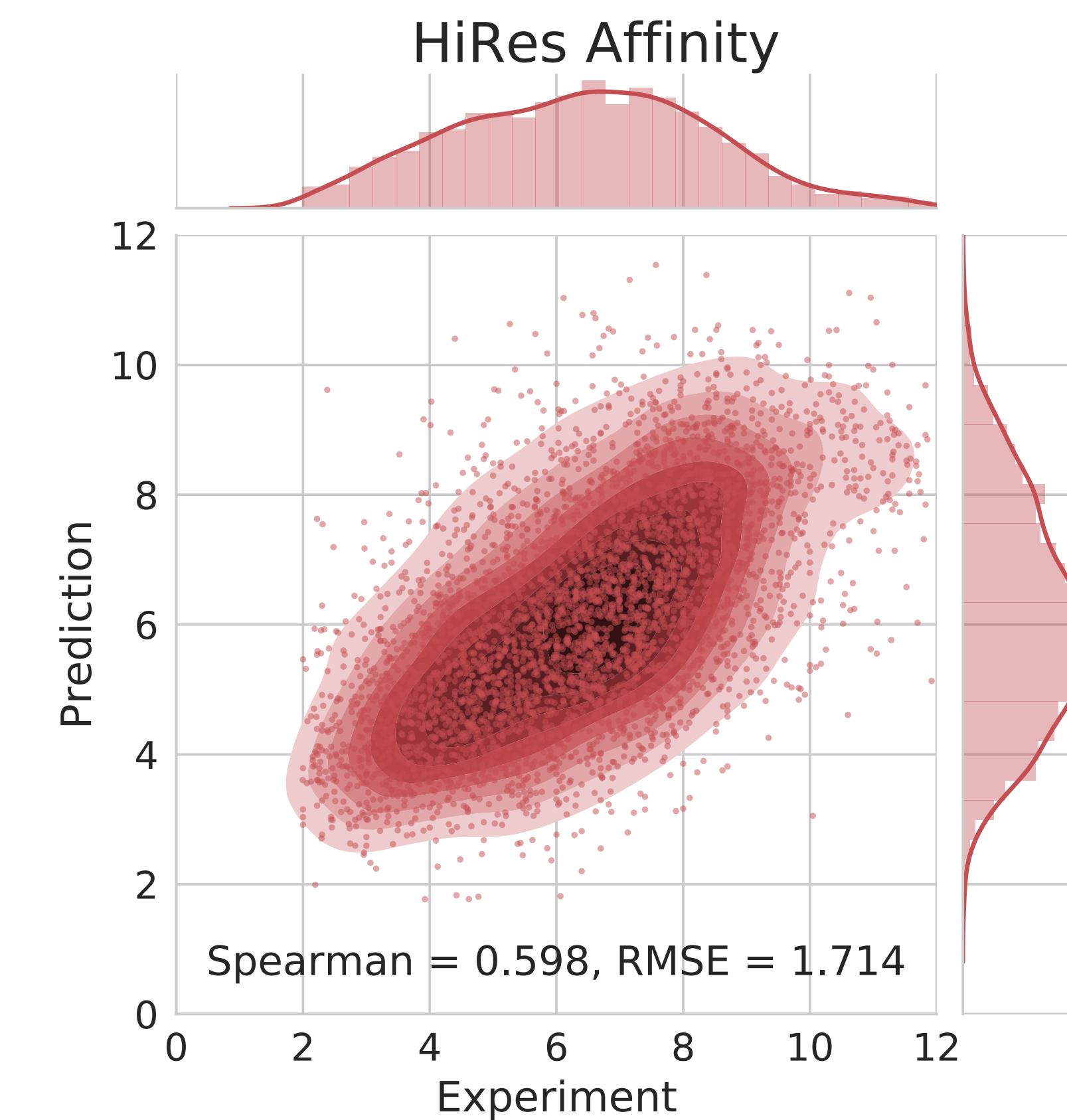


Pose Results

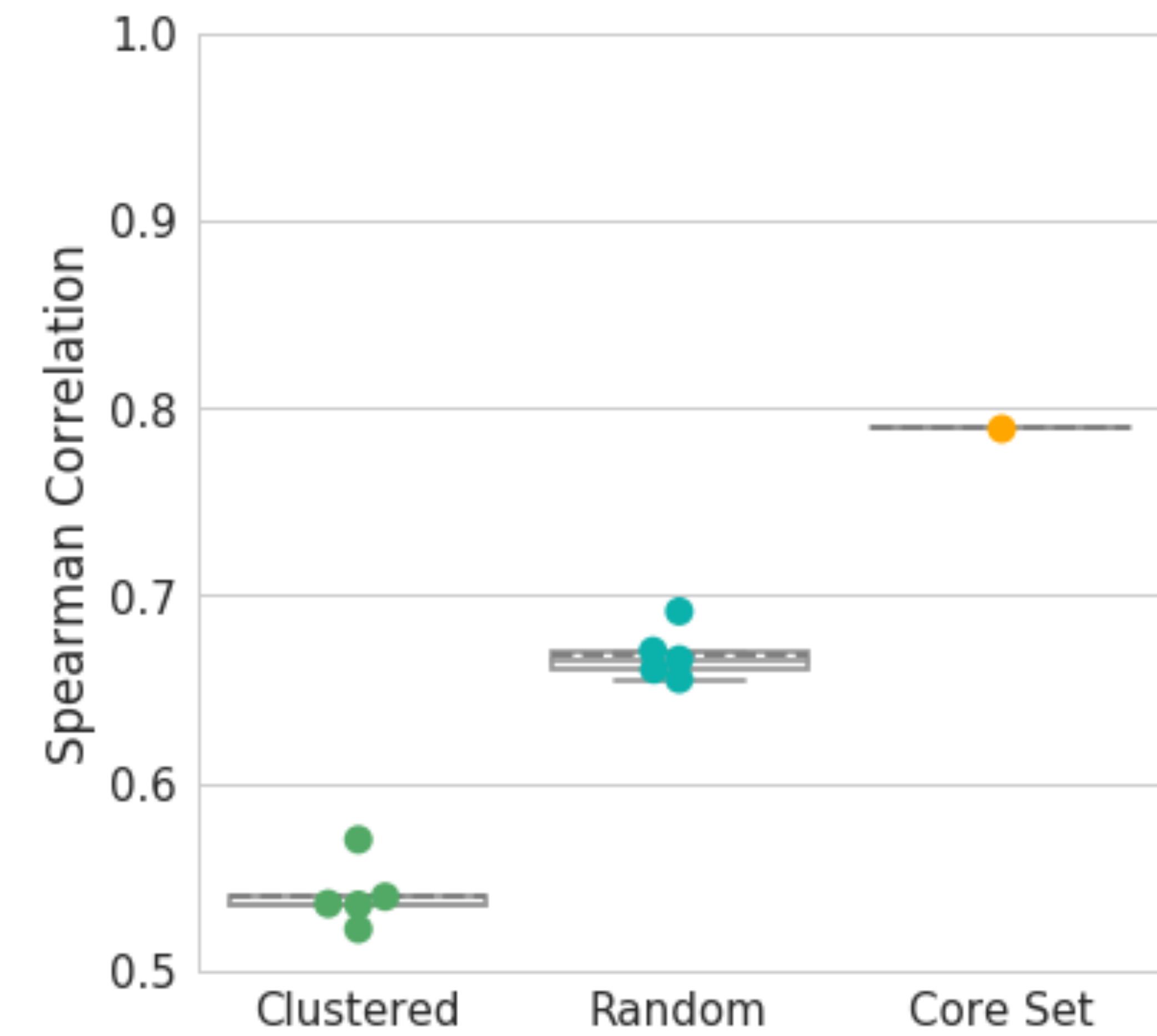
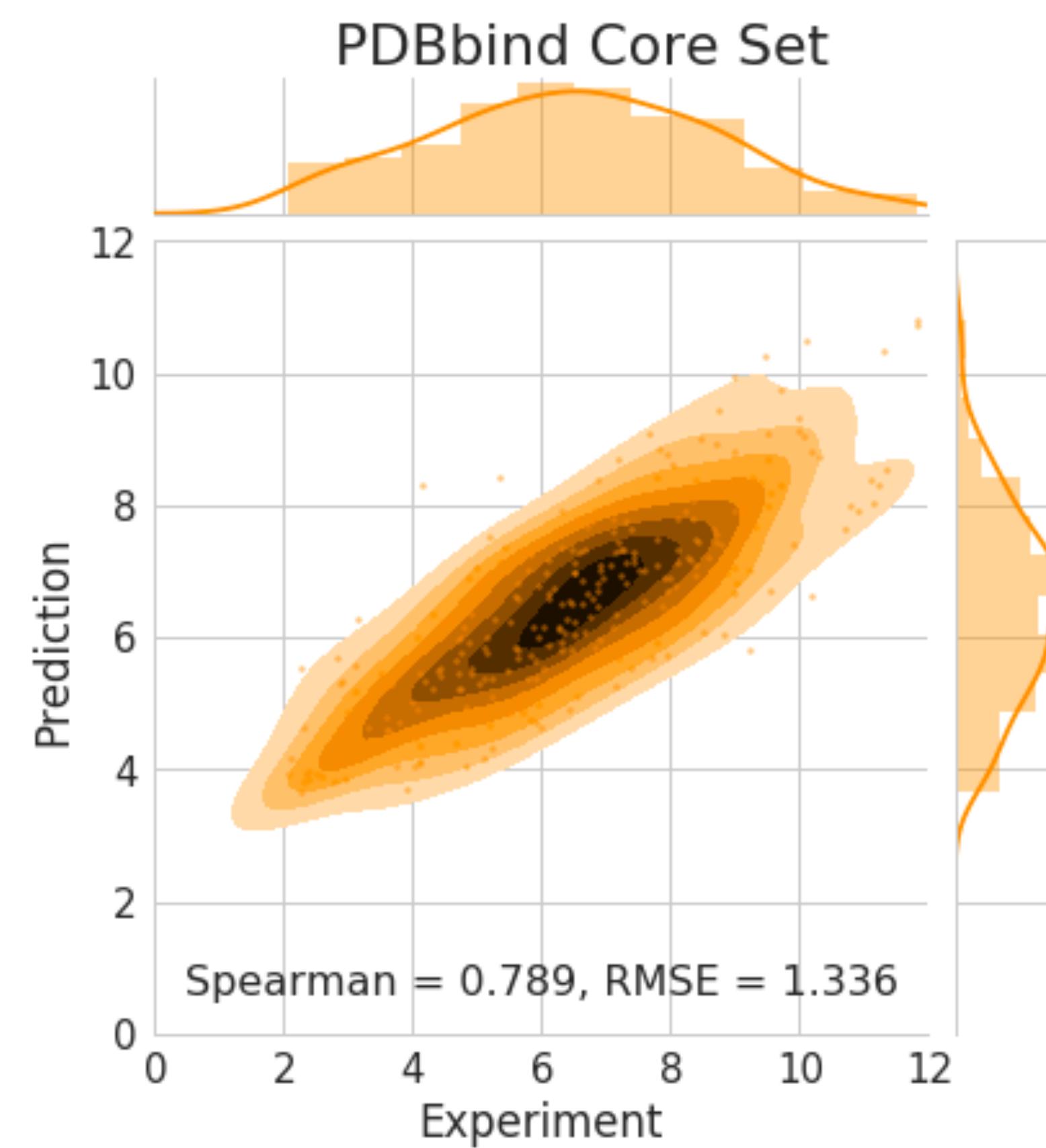
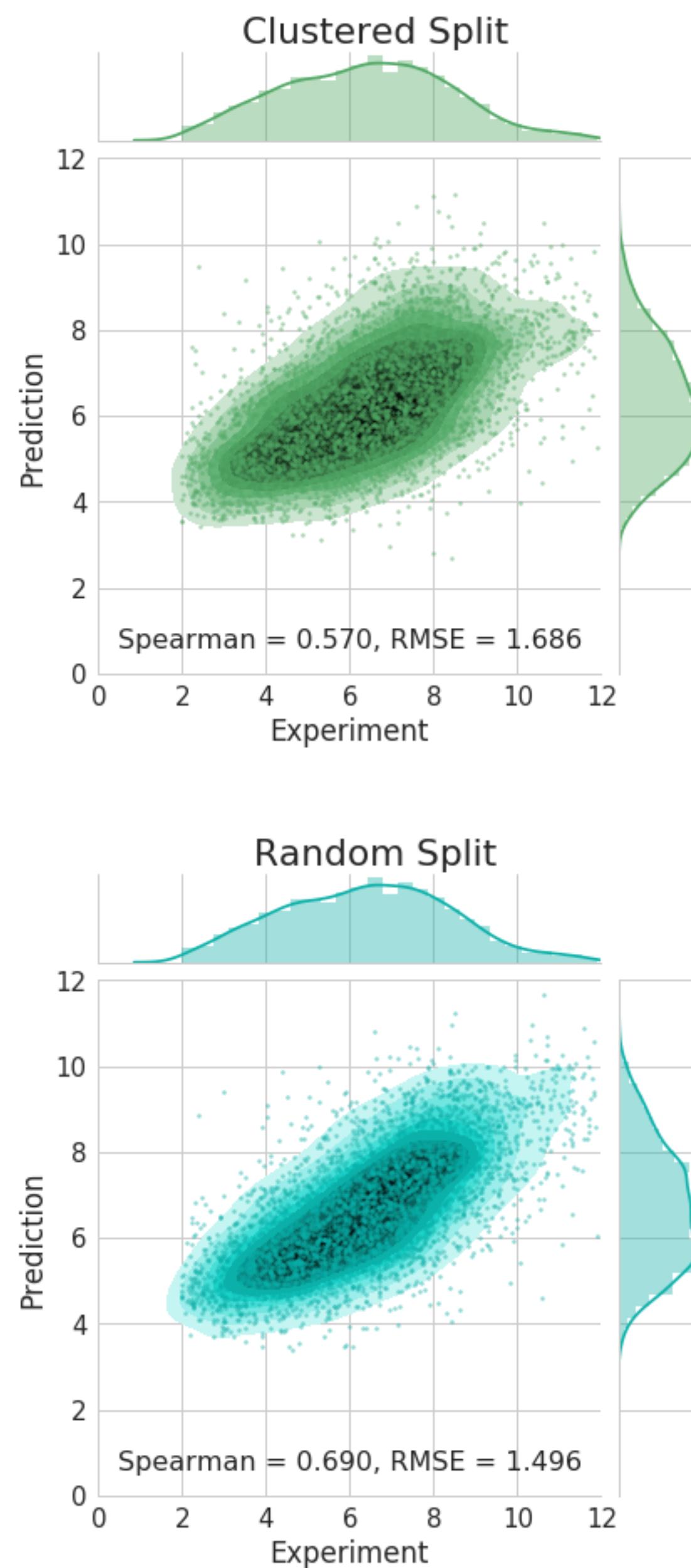
Crossdocked Pose



Affinity Results



Affinity Results



Virtual Screening

Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data

Fergus Imrie[†] , Anthony R. Bradley^{#¶§}, Mihaela van der Schaar^{¶⊥}, and Charlotte M. Deane^{*†} 

[†] Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3LB, U.K.

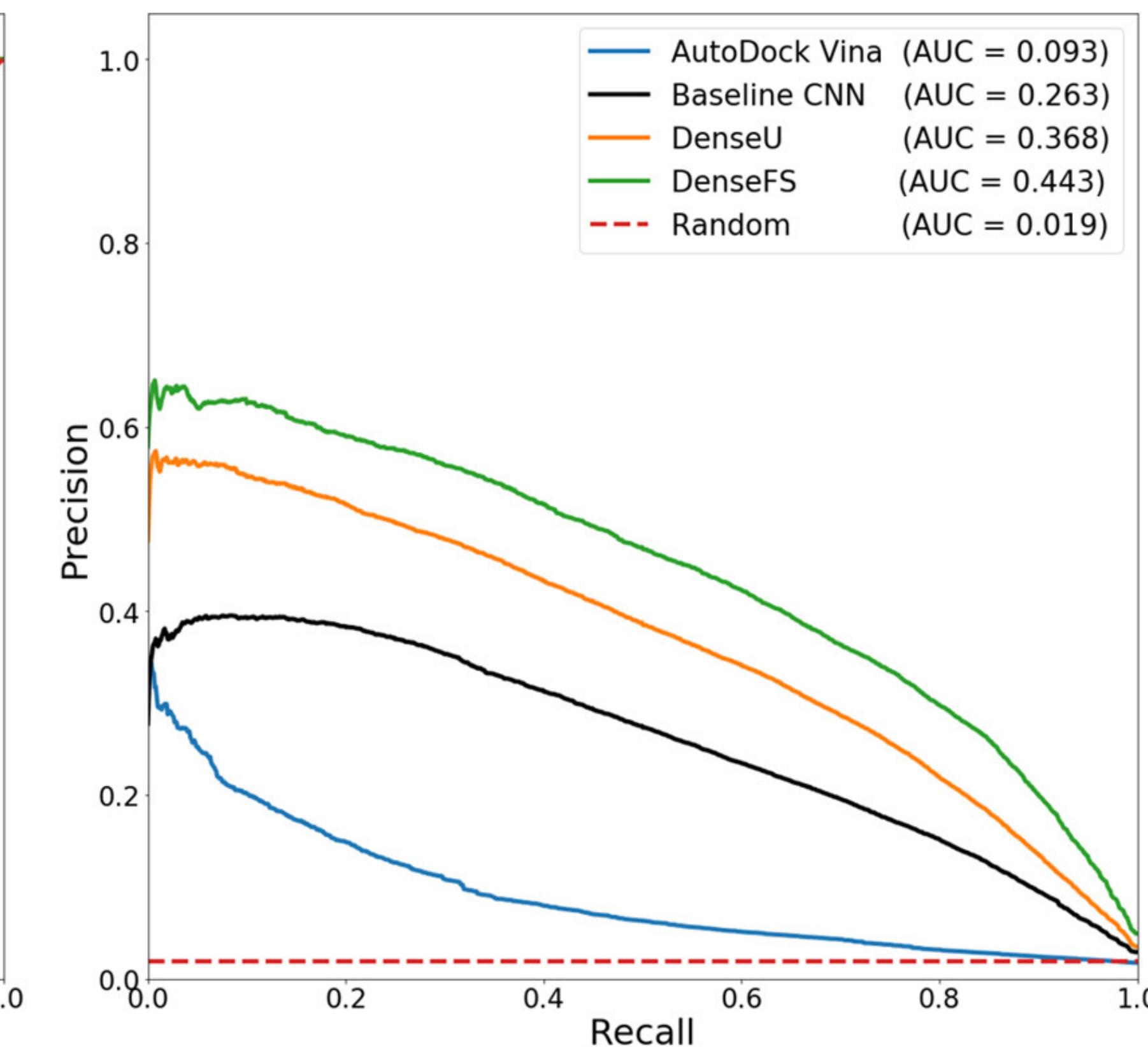
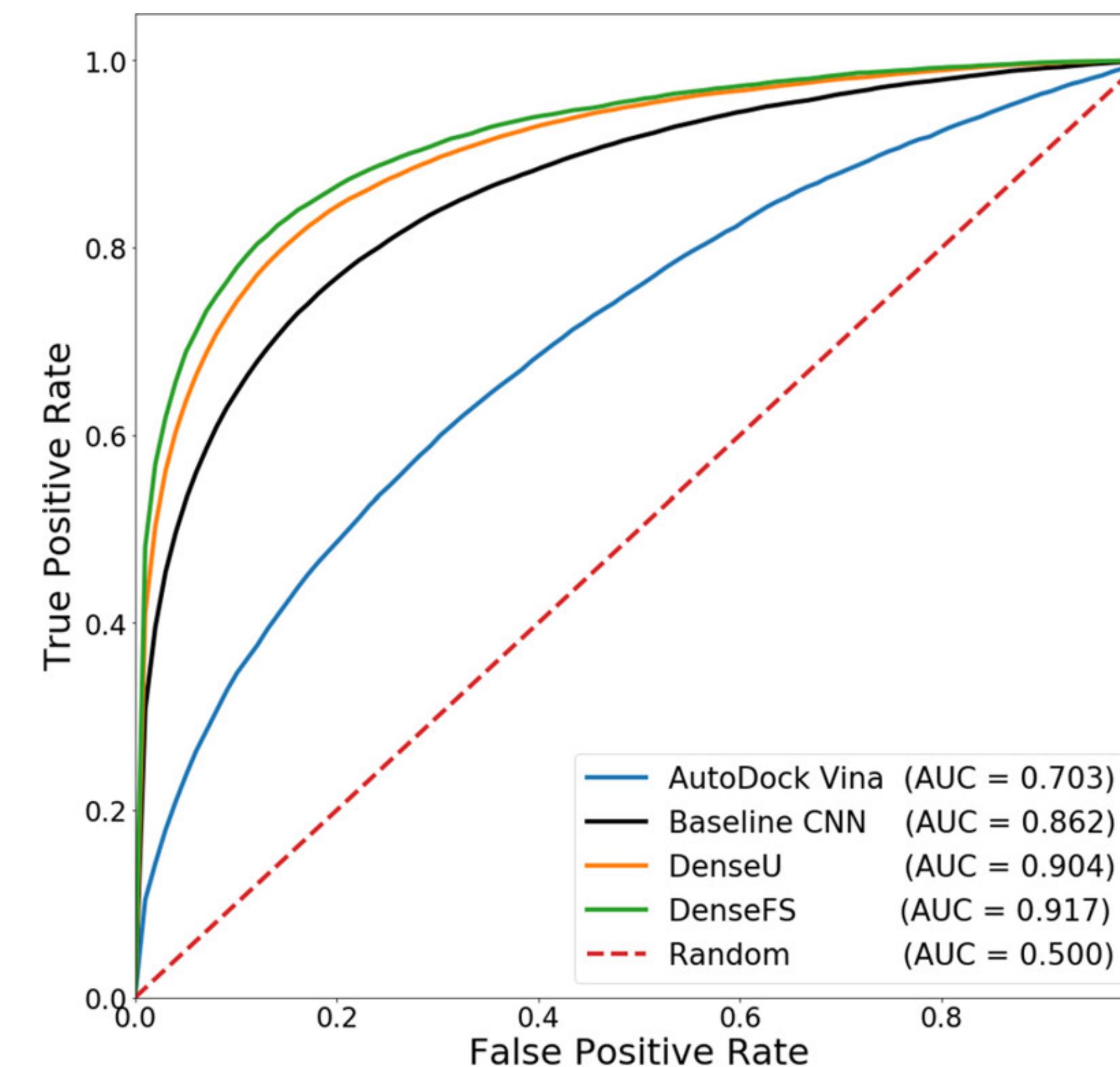
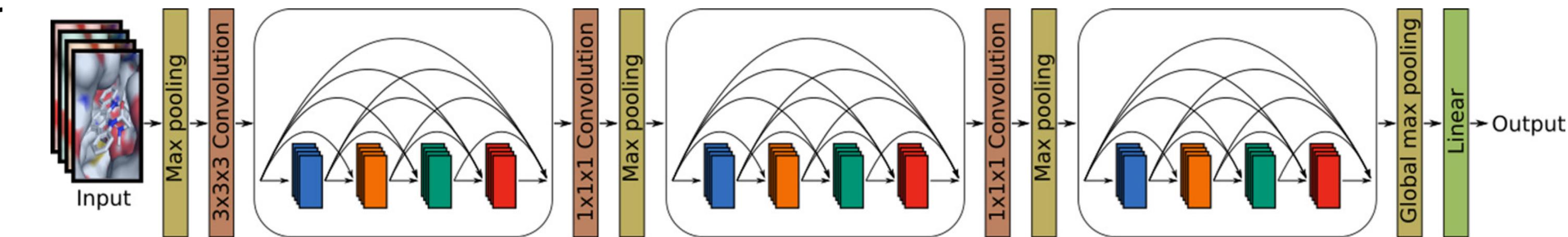
[‡] Structural Genomics Consortium, University of Oxford, Oxford OX3 7DQ, U.K.

[#] Department of Chemistry, University of Oxford, Oxford OX1 3TA, U.K.

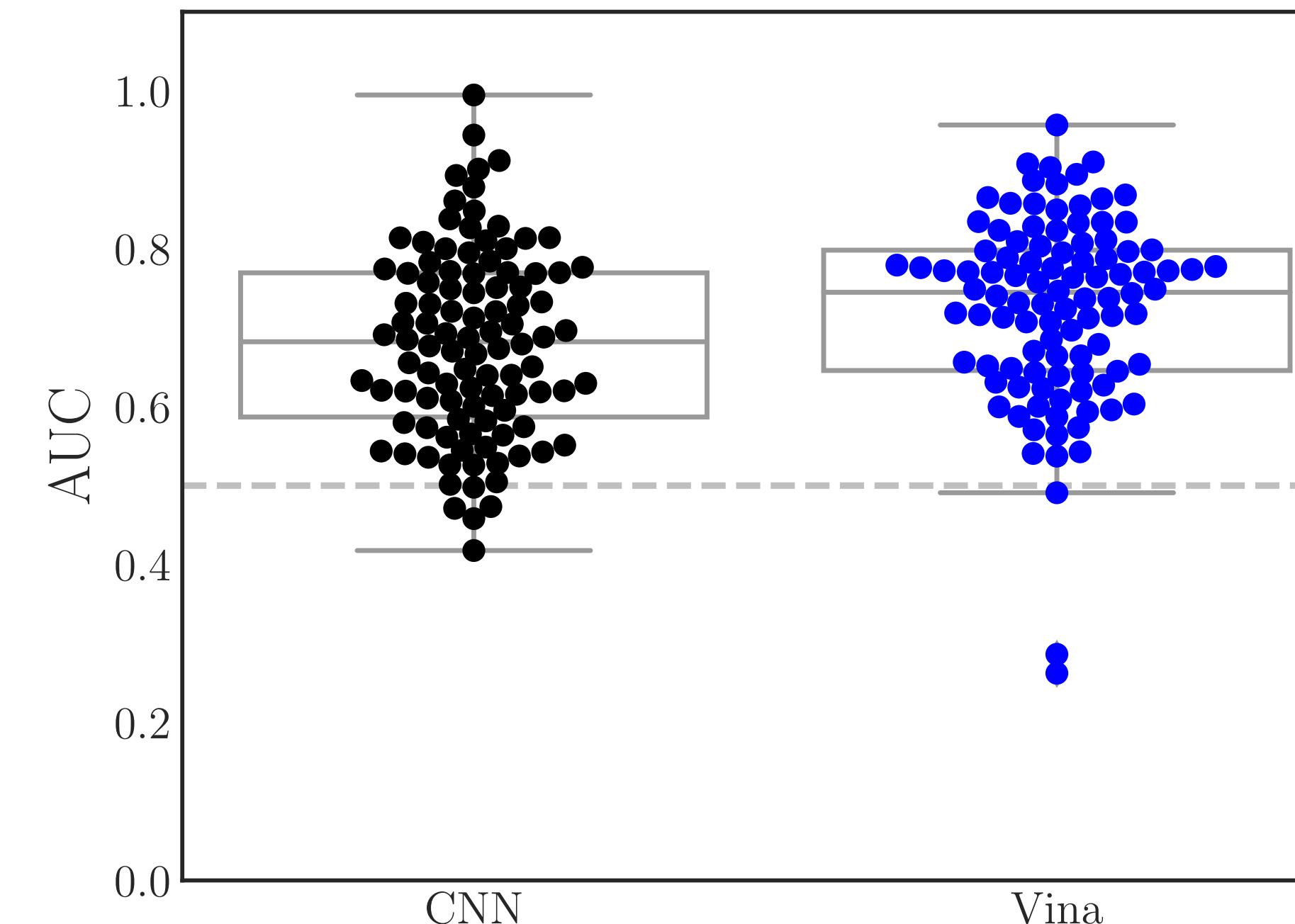
[§] Diamond Light Source Ltd., Didcot OX11 0DE, U.K.

^{*} Department of Engineering, University of Oxford, Oxford OX1 3PJ, U.K.

[⊥] Alan Turing Institute, London NW1 2DB, U.K.



Virtual Screening



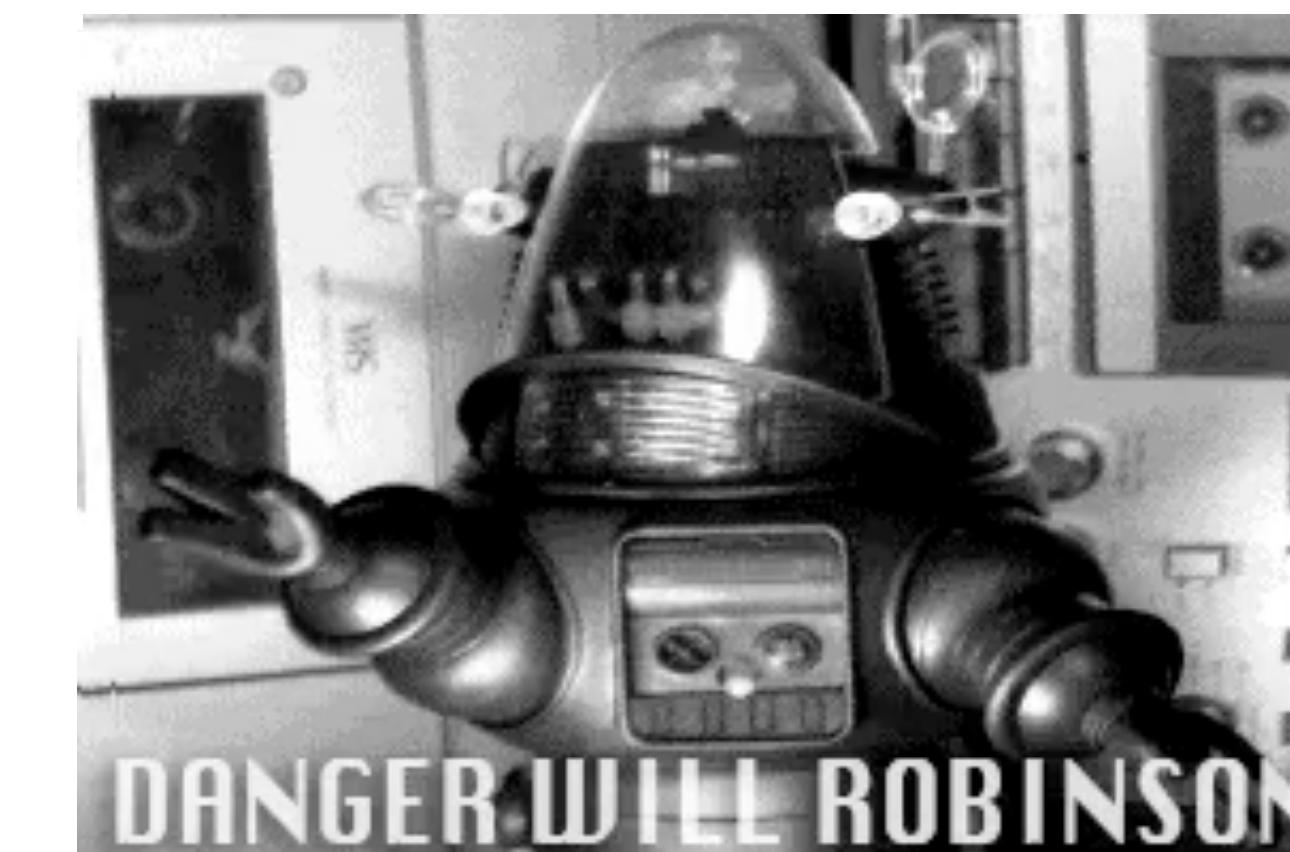
In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening

Jochen Sieg [Florian Flachsenberg](#) [and Matthias Rarey*](#)

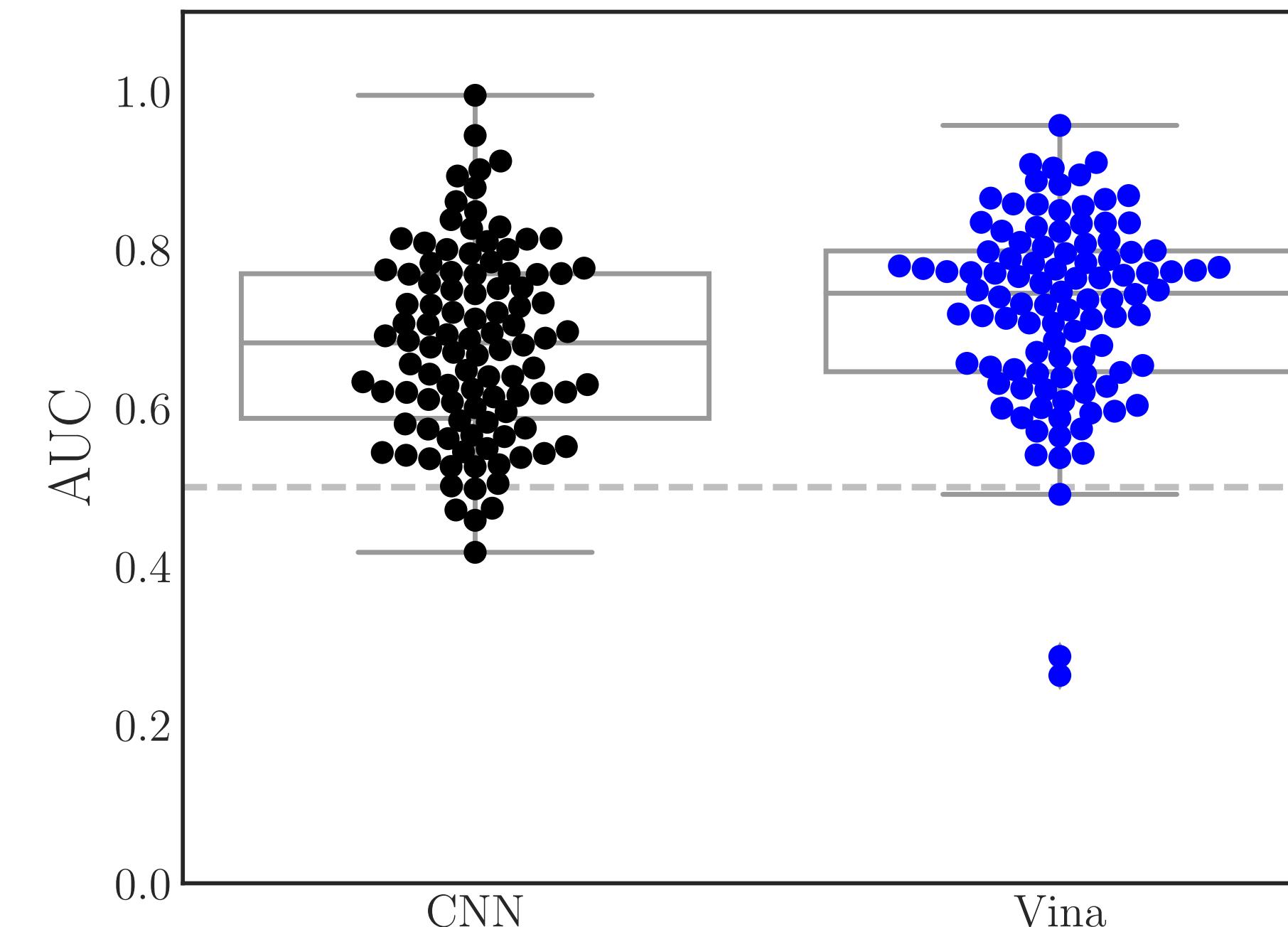
Universität Hamburg, ZBH - Center for Bioinformatics, Research Group for Computational Molecular Design,
Bundesstraße 43, 20146 Hamburg, Germany

Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening

Preprint submitted on 24.03.2019, 15:39 and posted on 25.03.2019, 12:58 by Lieyang Chen, Anthony Cruz,
Steven Ramsey, Callum J. Dickson, José S. Duca, Viktor Hornak, David R. Koes, [Tom Kurtzman](#)



Virtual Screening



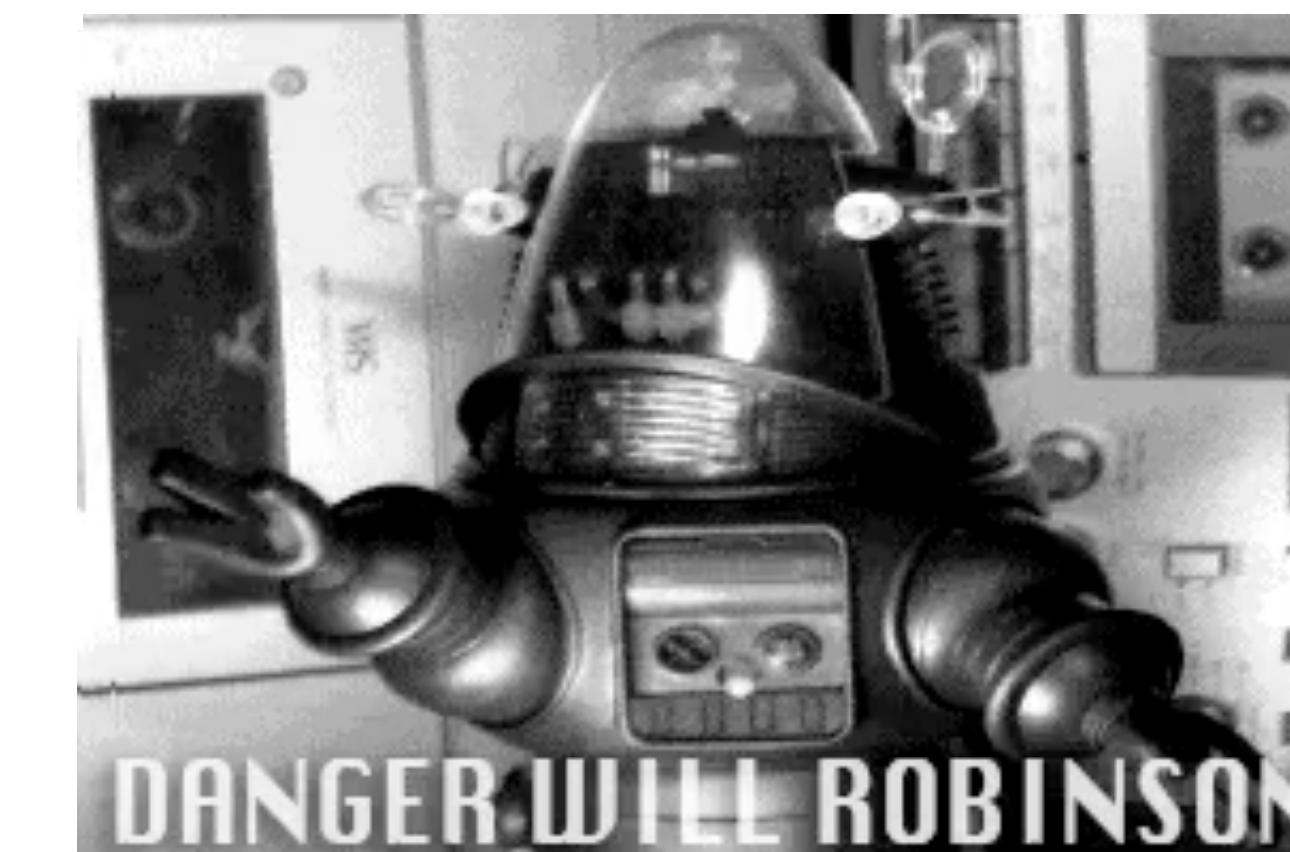
In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening

Jochen Sieg , Florian Flachsenberg , and Matthias Rarey* 

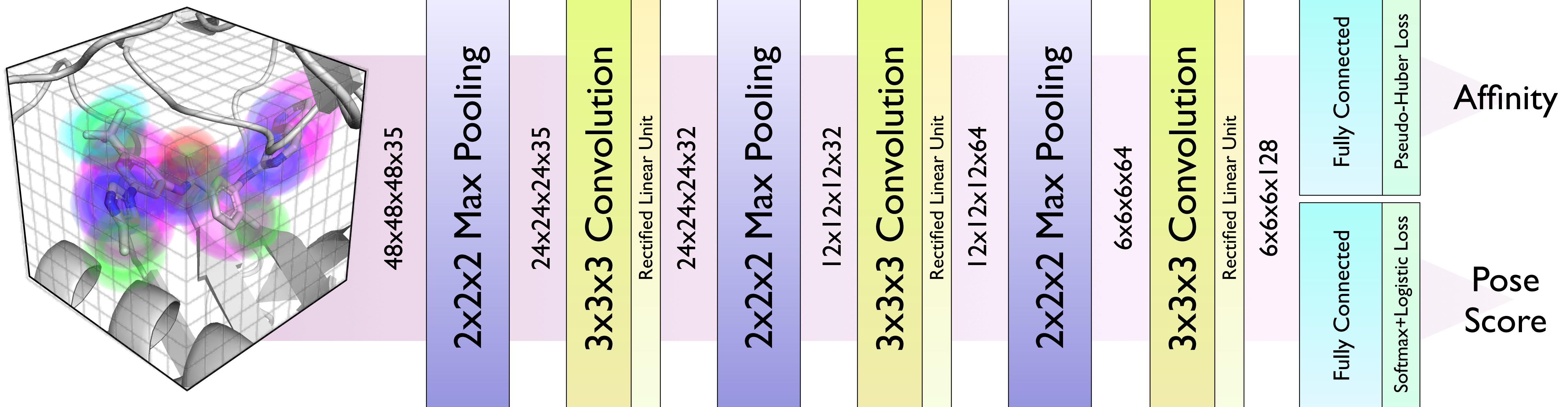
Universität Hamburg, ZBH - Center for Bioinformatics, Research Group for Computational Molecular Design,
Bundesstraße 43, 20146 Hamburg, Germany

Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening

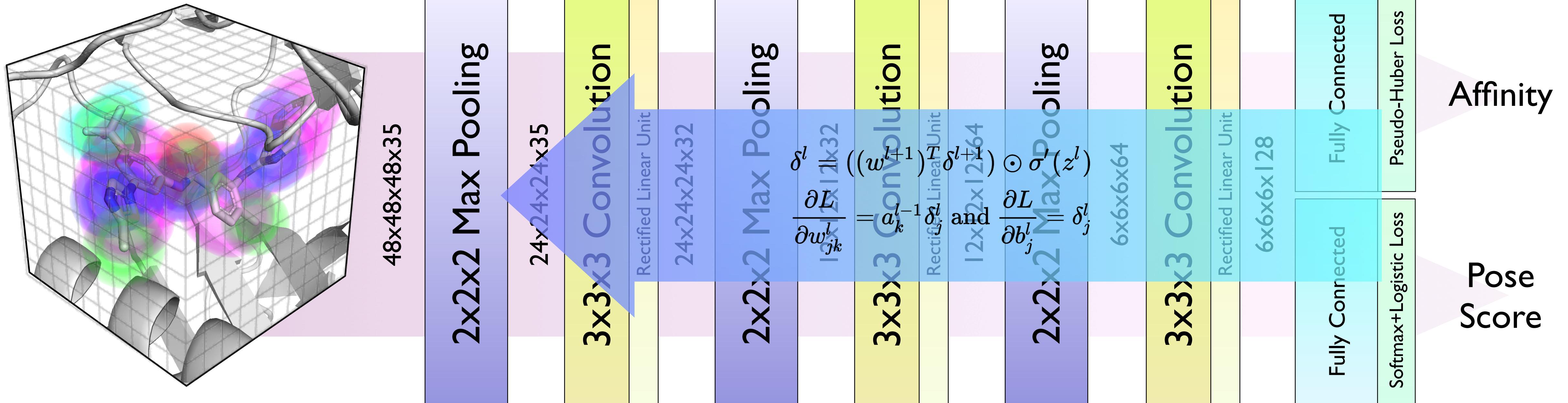
Preprint submitted on 24.03.2019, 15:39 and posted on 25.03.2019, 12:58 by Lieyang Chen, Anthony Cruz,
Steven Ramsey, Callum J. Dickson, José S. Duca, Viktor Hornak, David R. Koes, Tom Kurtzman



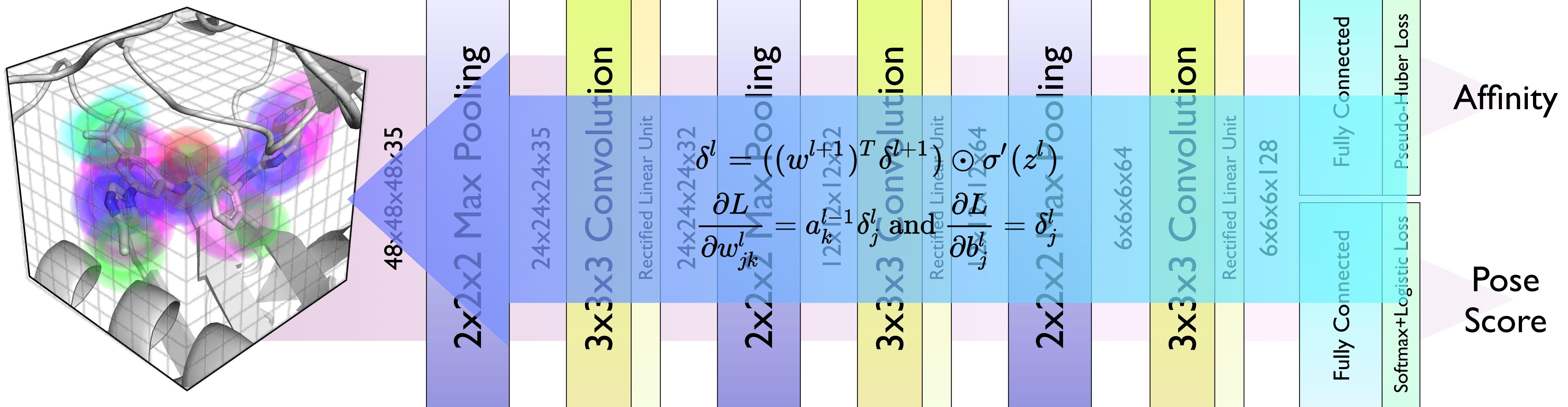
Beyond Scoring



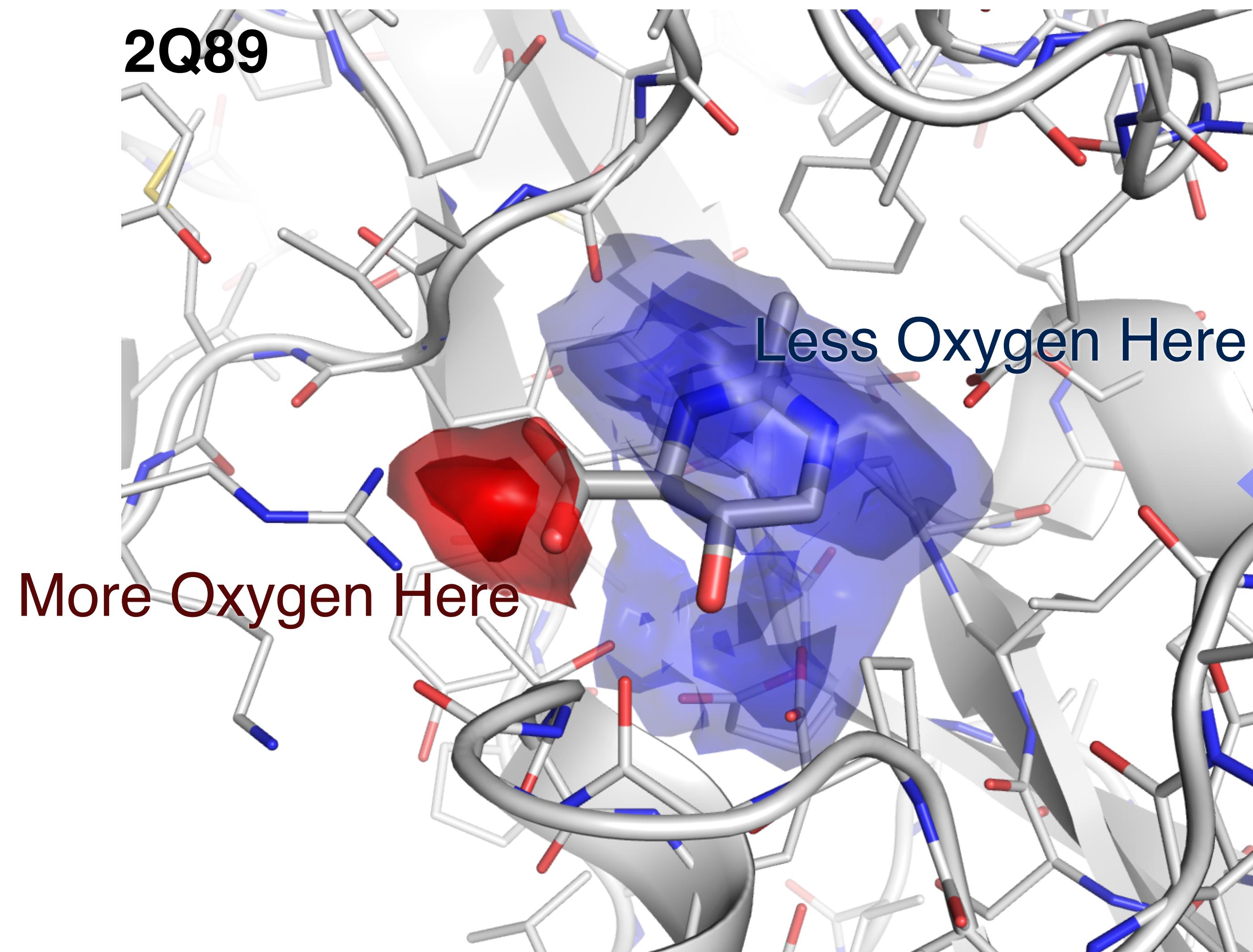
Beyond Scoring



Beyond Scoring

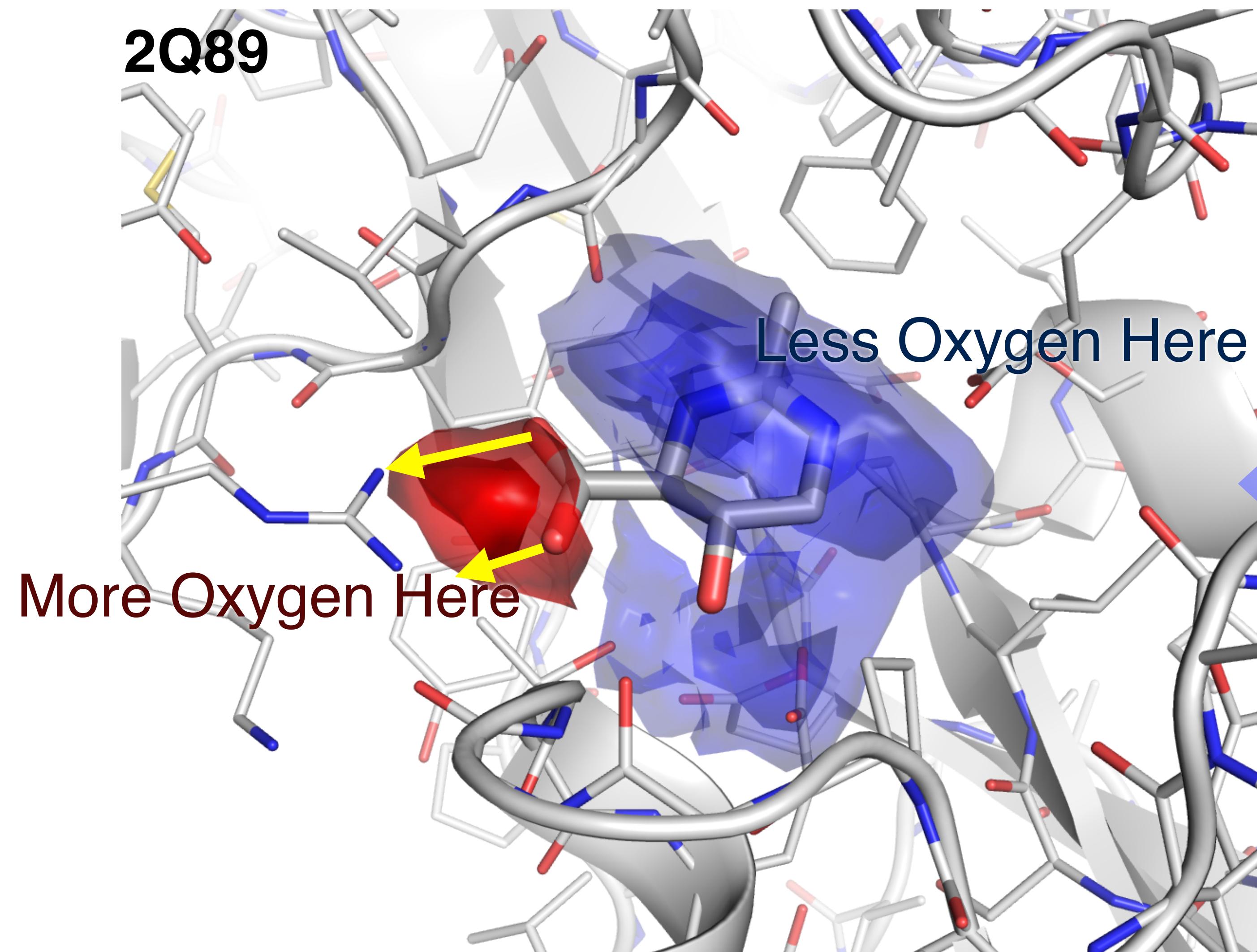


Beyond Scoring

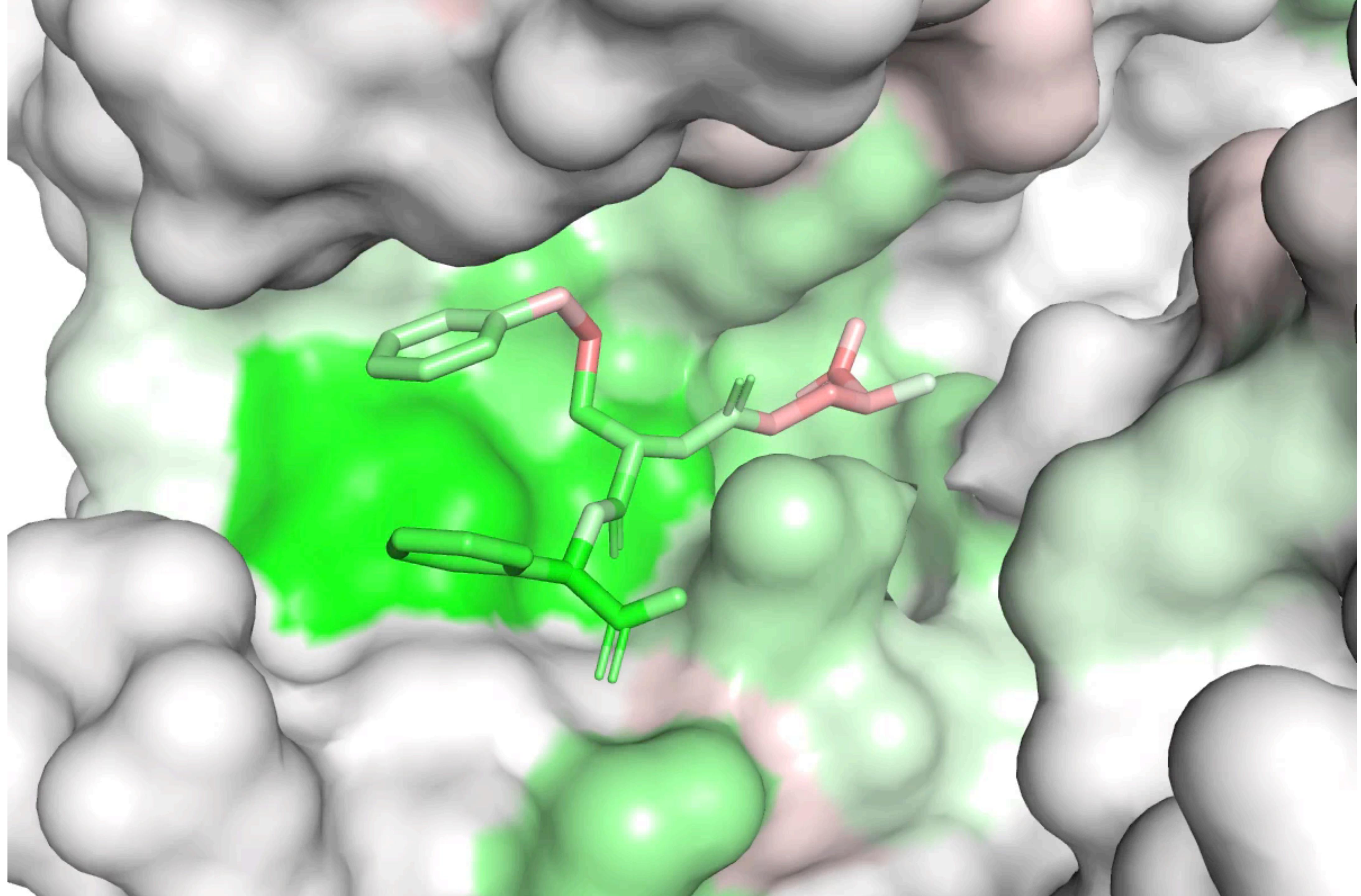


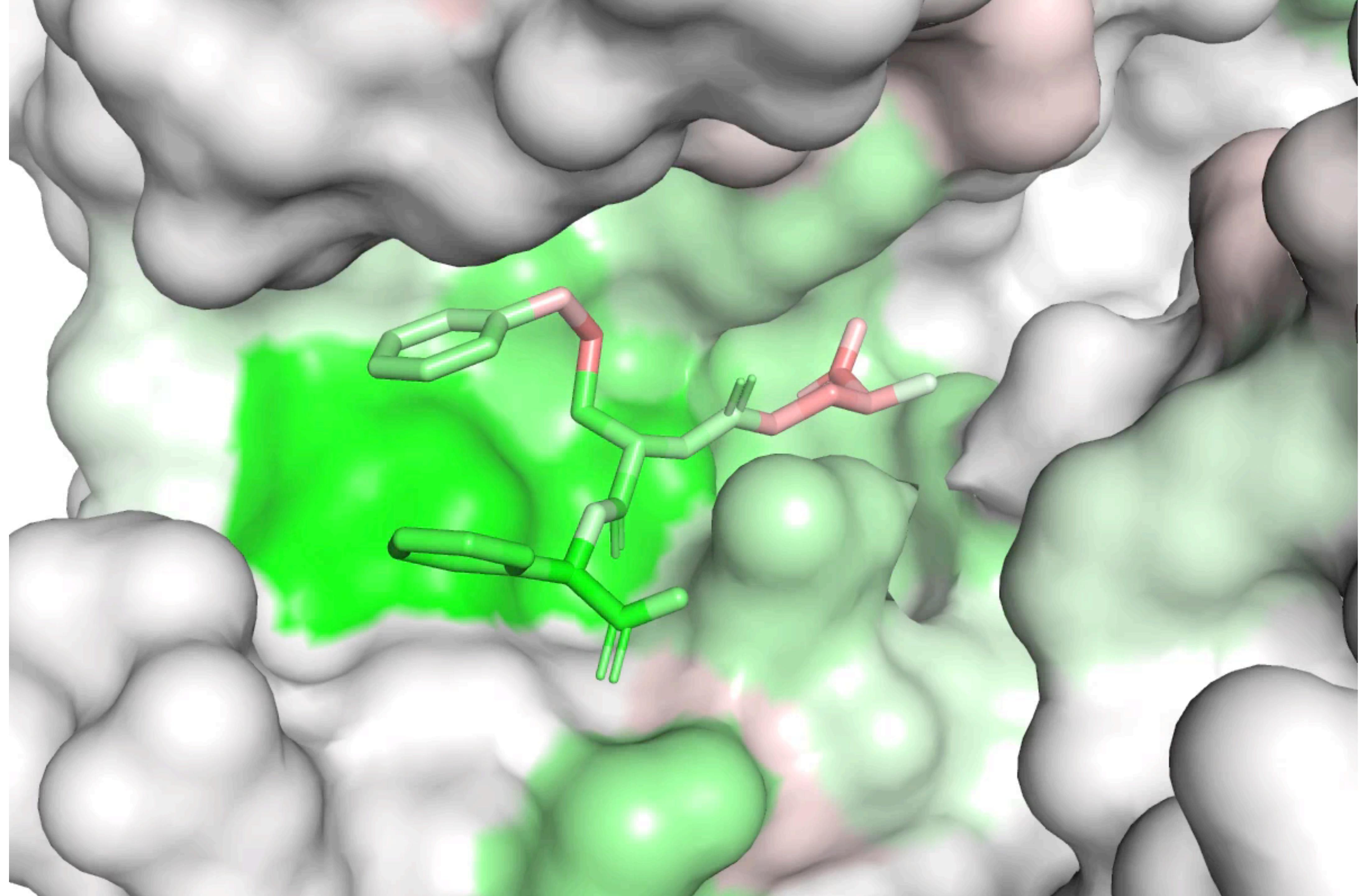
$$\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

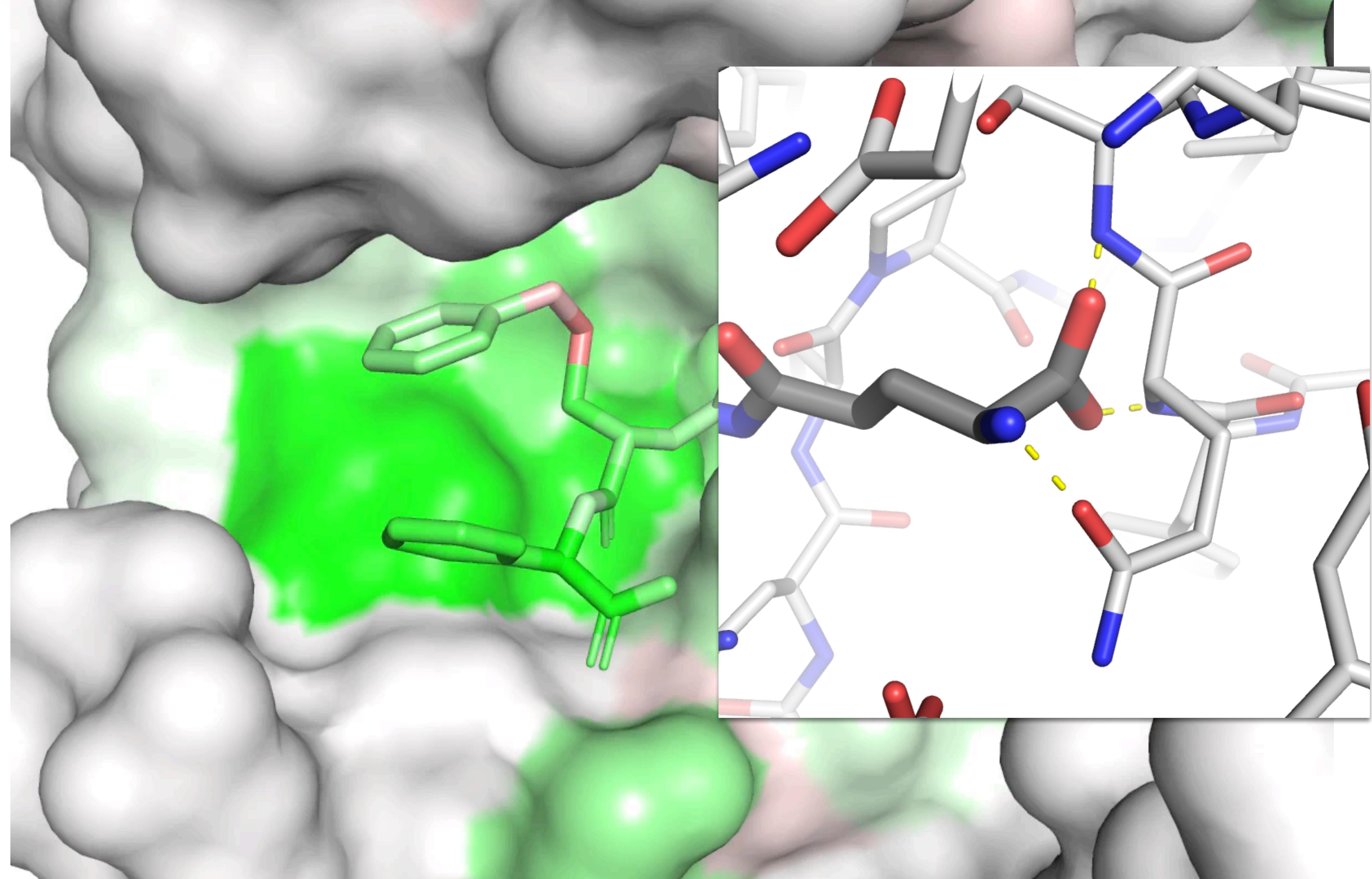
Beyond Scoring

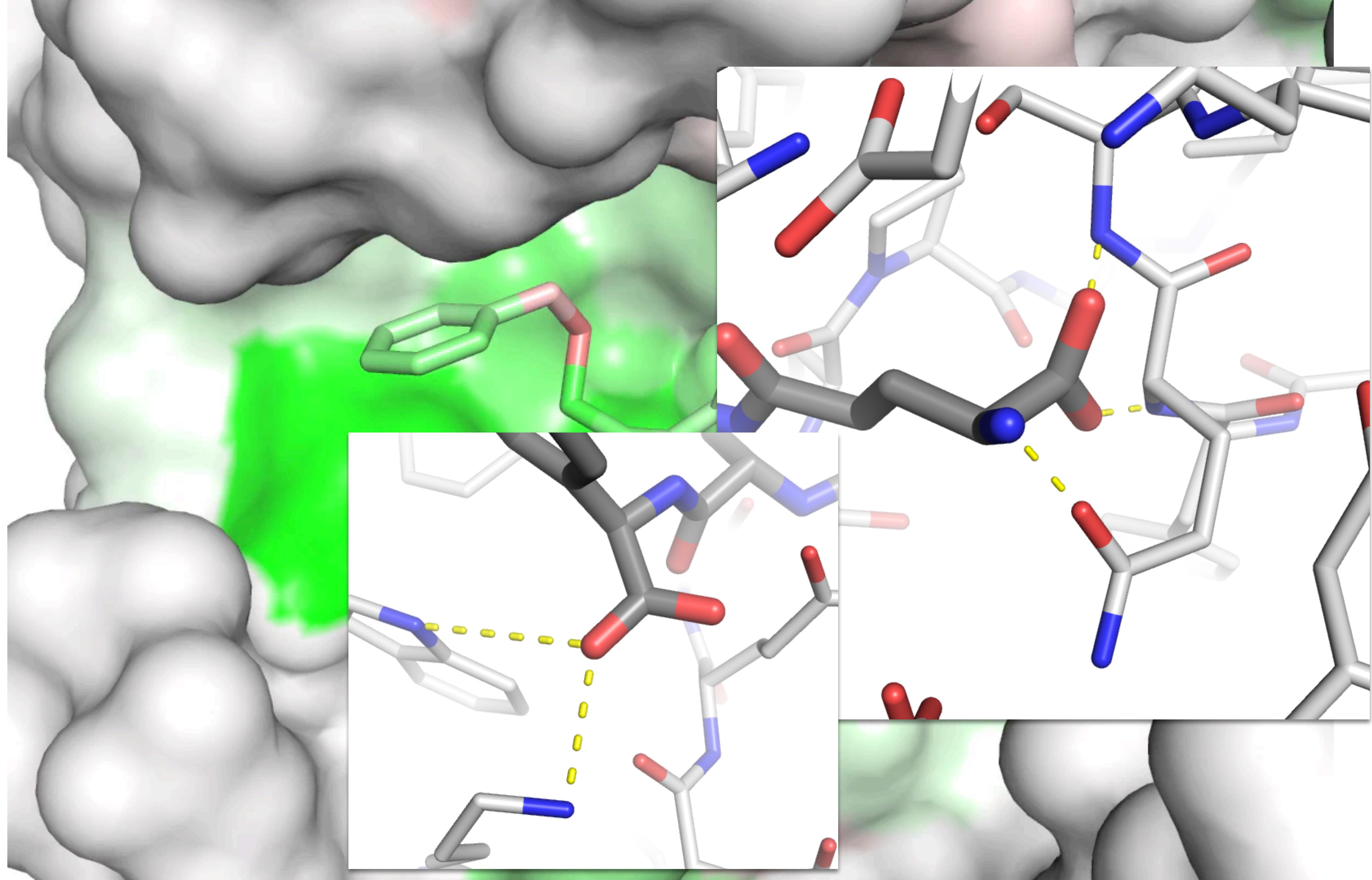


$$\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

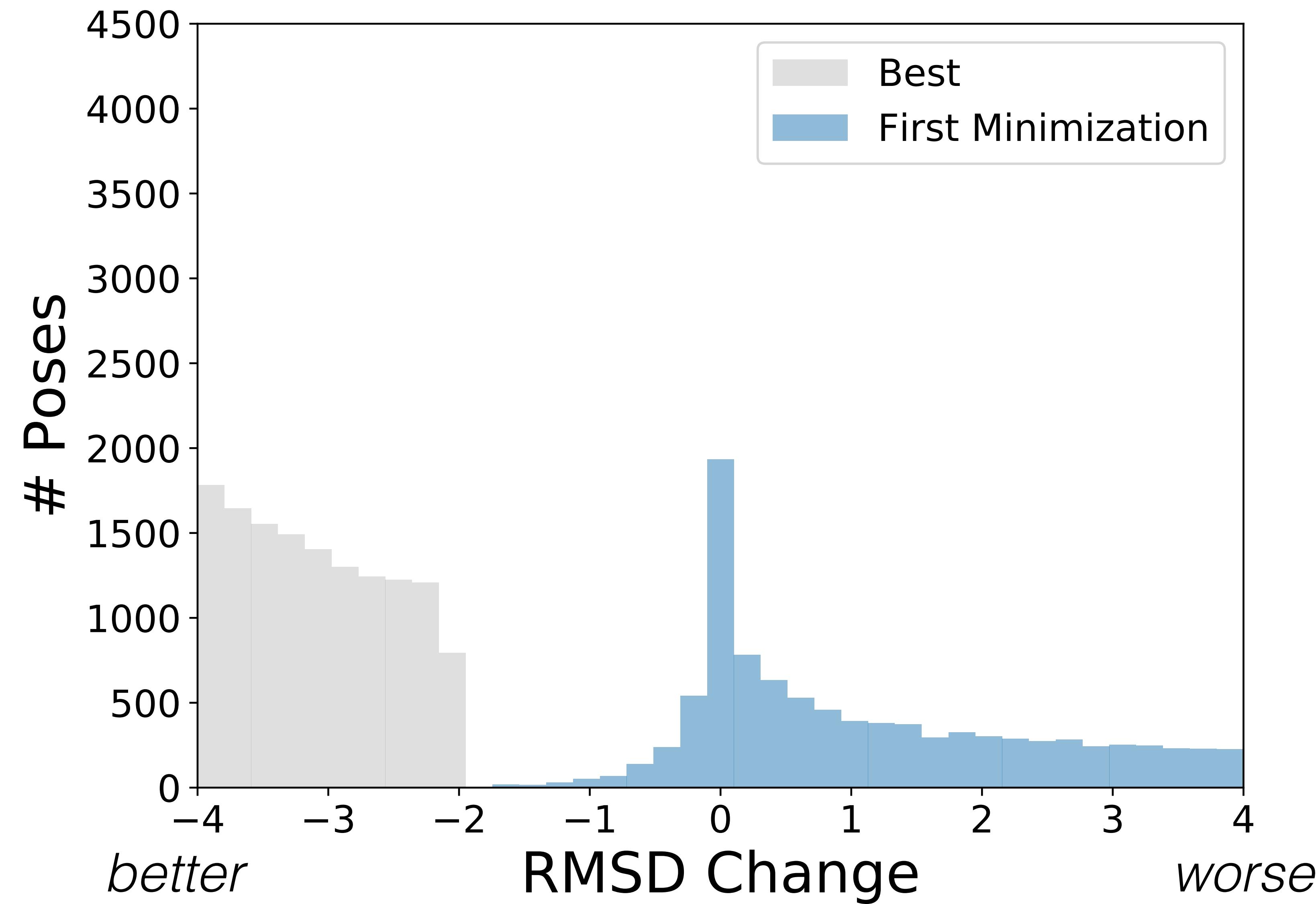




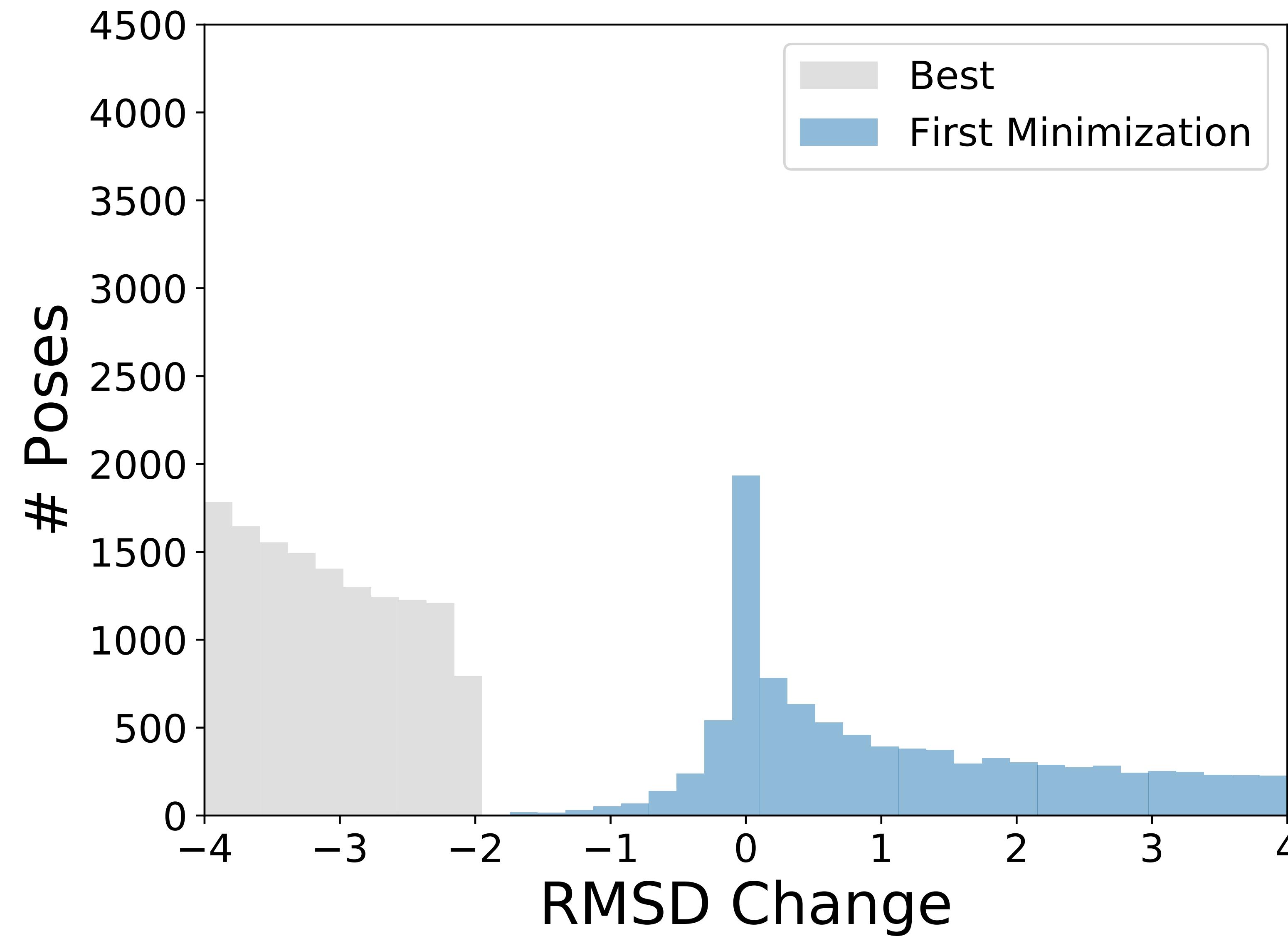




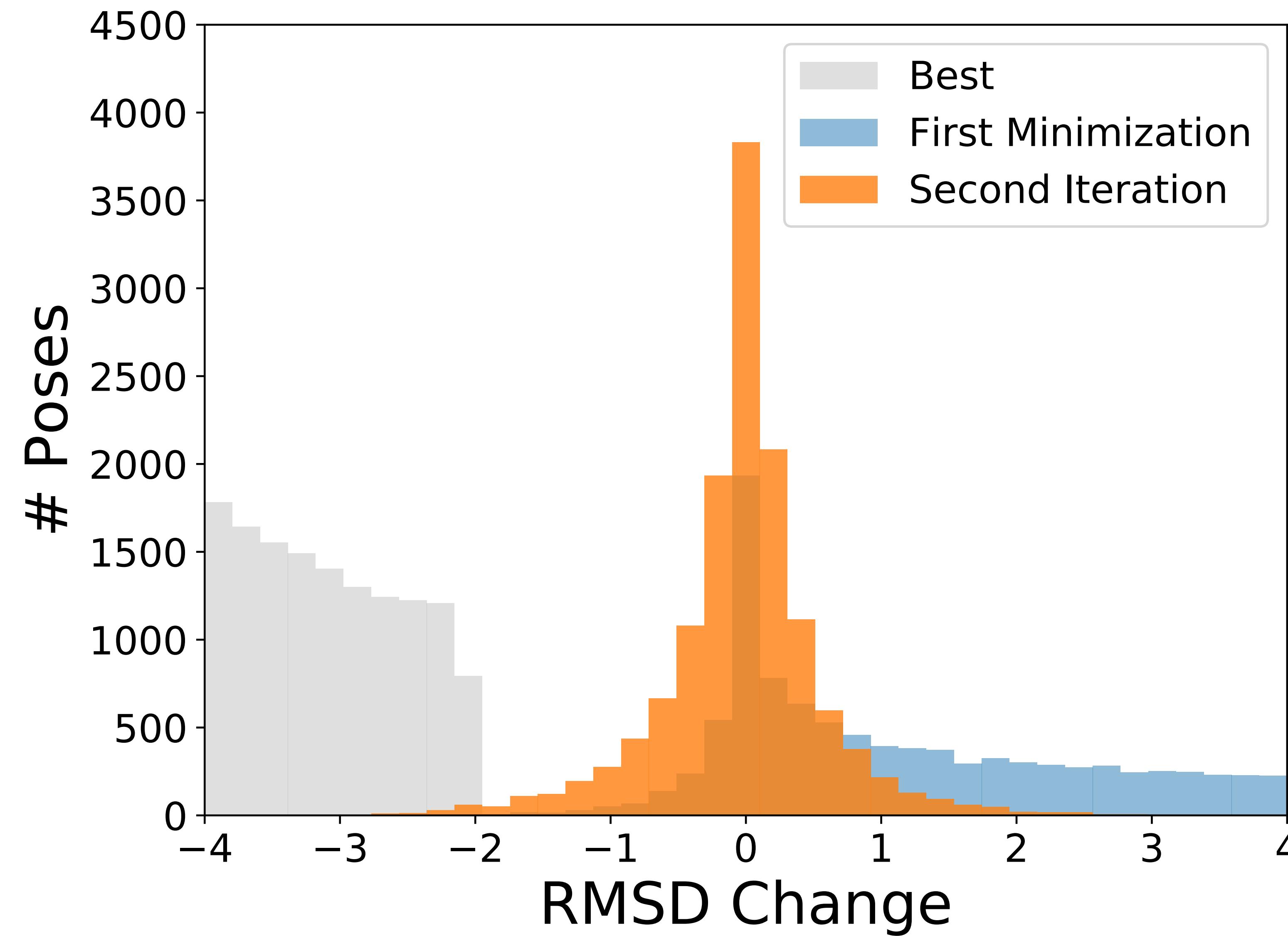
Minimizing Low RMSD Poses



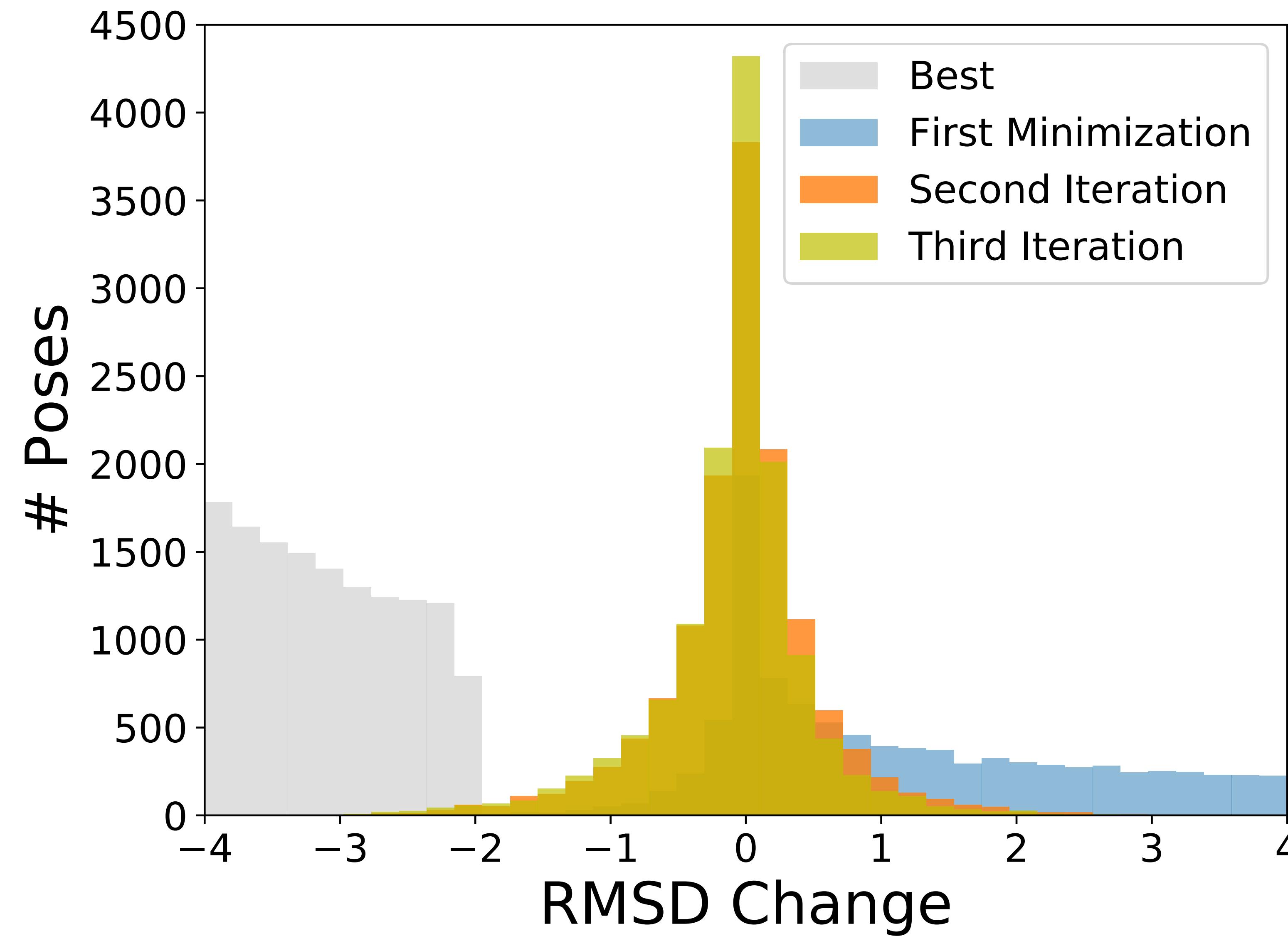
Iterative Refinement

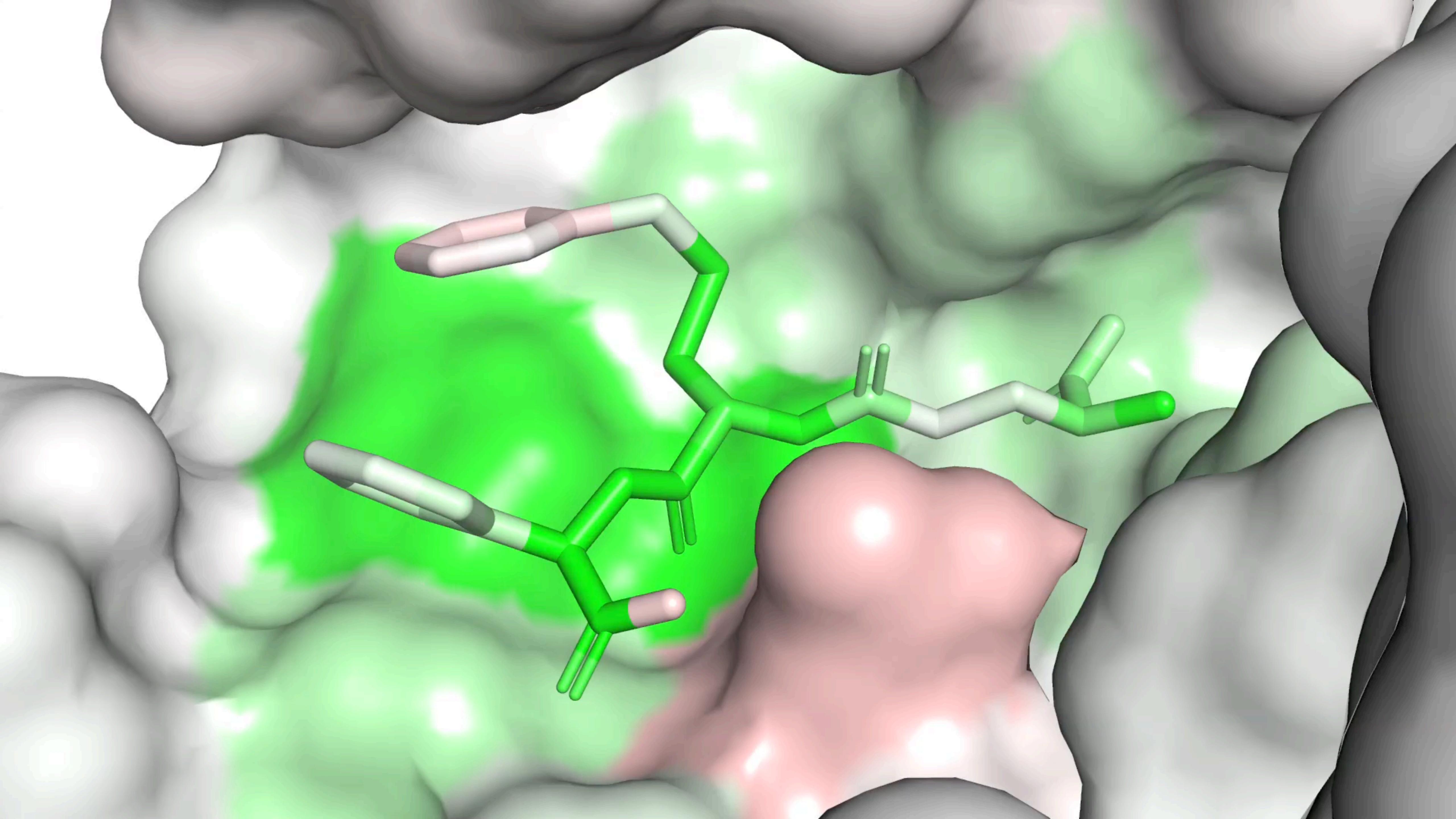


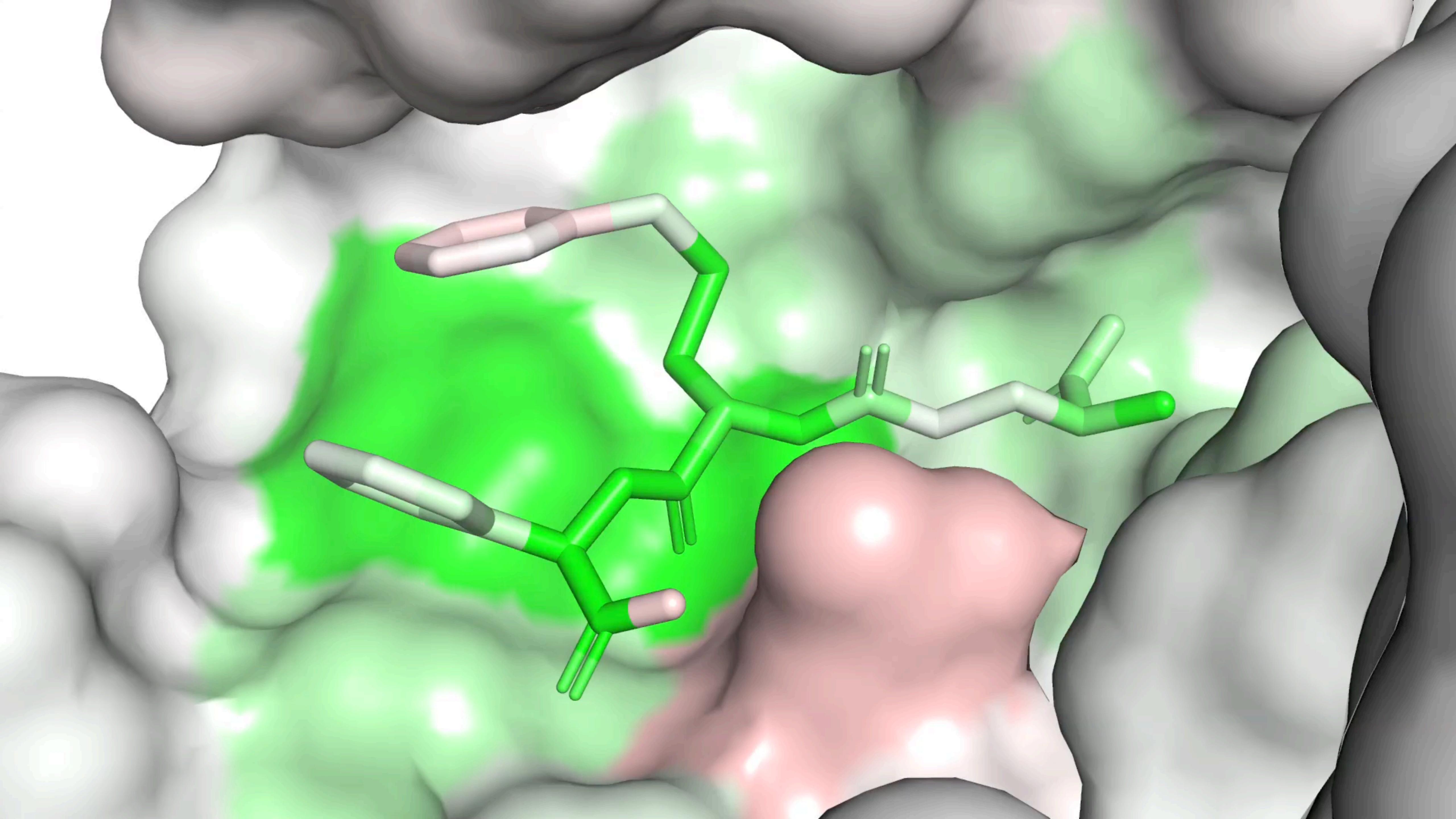
Iterative Refinement



Iterative Refinement

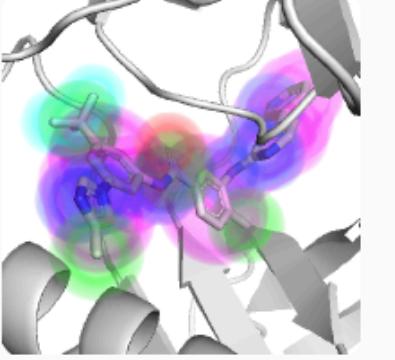






Search or jump to...

Pull requests Issues Marketplace Explore



gnina

Repositories 4 **People** 9 **Teams** 1 **Projects** 0 **Settings**

Find a repository... Type: All ▾ Language: All ▾

gnina
gnina is not smina/vina

cheminformatics computational-chemistry drug-discovery

C++ ★ 35 26 Updated 11 days ago

scripts

Jupyter Notebook ★ 1 13 BSD-3-Clause Updated on Sep 17, 2018

models

Trained caffe models

★ 6 9 Updated on Aug 23, 2018

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

Copyright © 2017 American Chemical Society

 Cite this: *J. Chem. Inf. Model.* 2017, 57, 4, 942-957

 RIS Citation 

arXiv.org > stat > arXiv:1710.07400

Statistics > Machine Learning

Ligand Pose Optimization with Atomic Grid-Based Convolutional Neural Networks

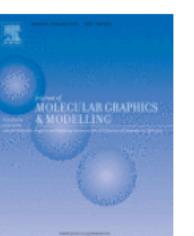
Matthew Ragoza, Lillian Turner, David Ryan Koes

(Submitted on 20 Oct 2017)



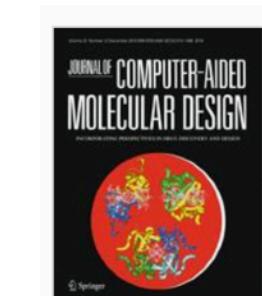
Journal of Molecular Graphics and Modelling

Volume 84, September 2018, Pages 96-108



Visualizing convolutional neural network protein-ligand scoring

Joshua Hochuli, Alec Helbling, Tamar Skaist, Matthew Ragoza, David Ryan Koes  



Journal of Computer-Aided Molecular Design

pp 1-16 | [Cite as](#)

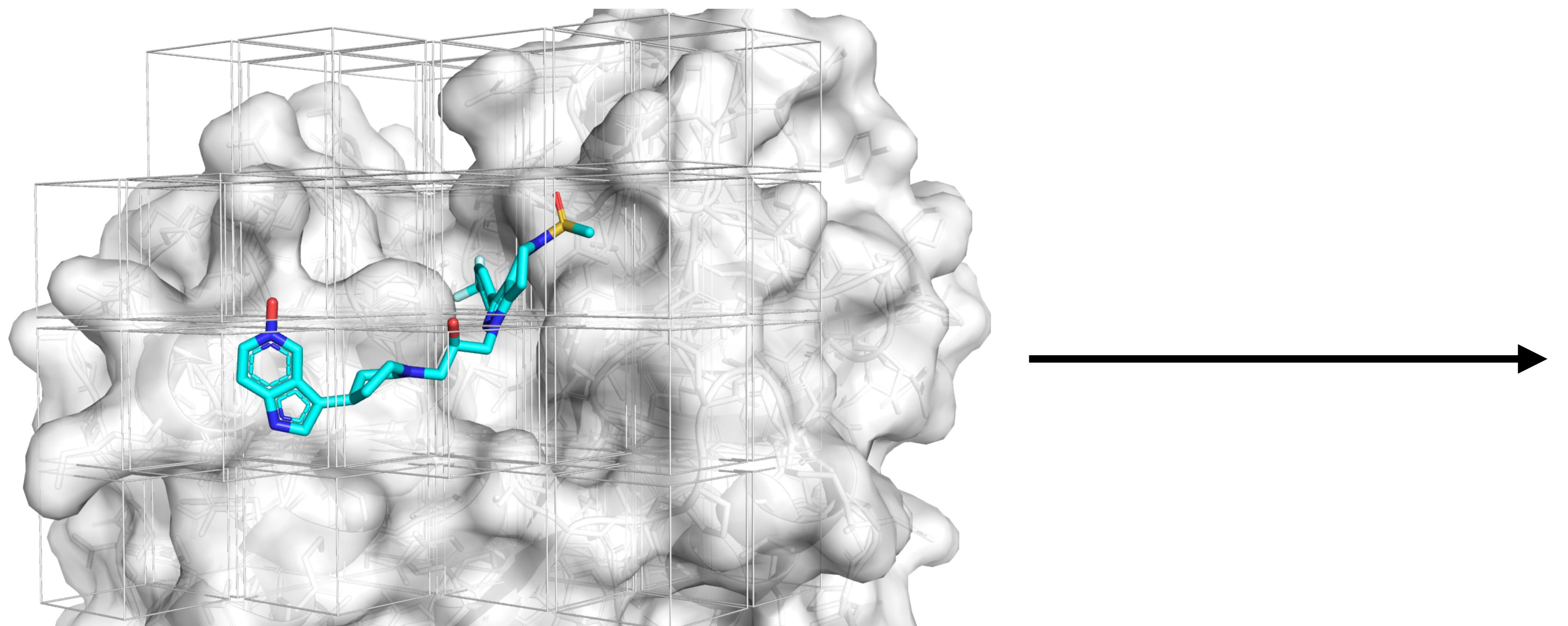
Convolutional neural network scoring and minimization in the D3R 2017 community challenge

Authors

Authors and affiliations

Jocelyn Sunseri, Jonathan E. King, Paul G. Francoeur, David Ryan Koes 

libmolgrid

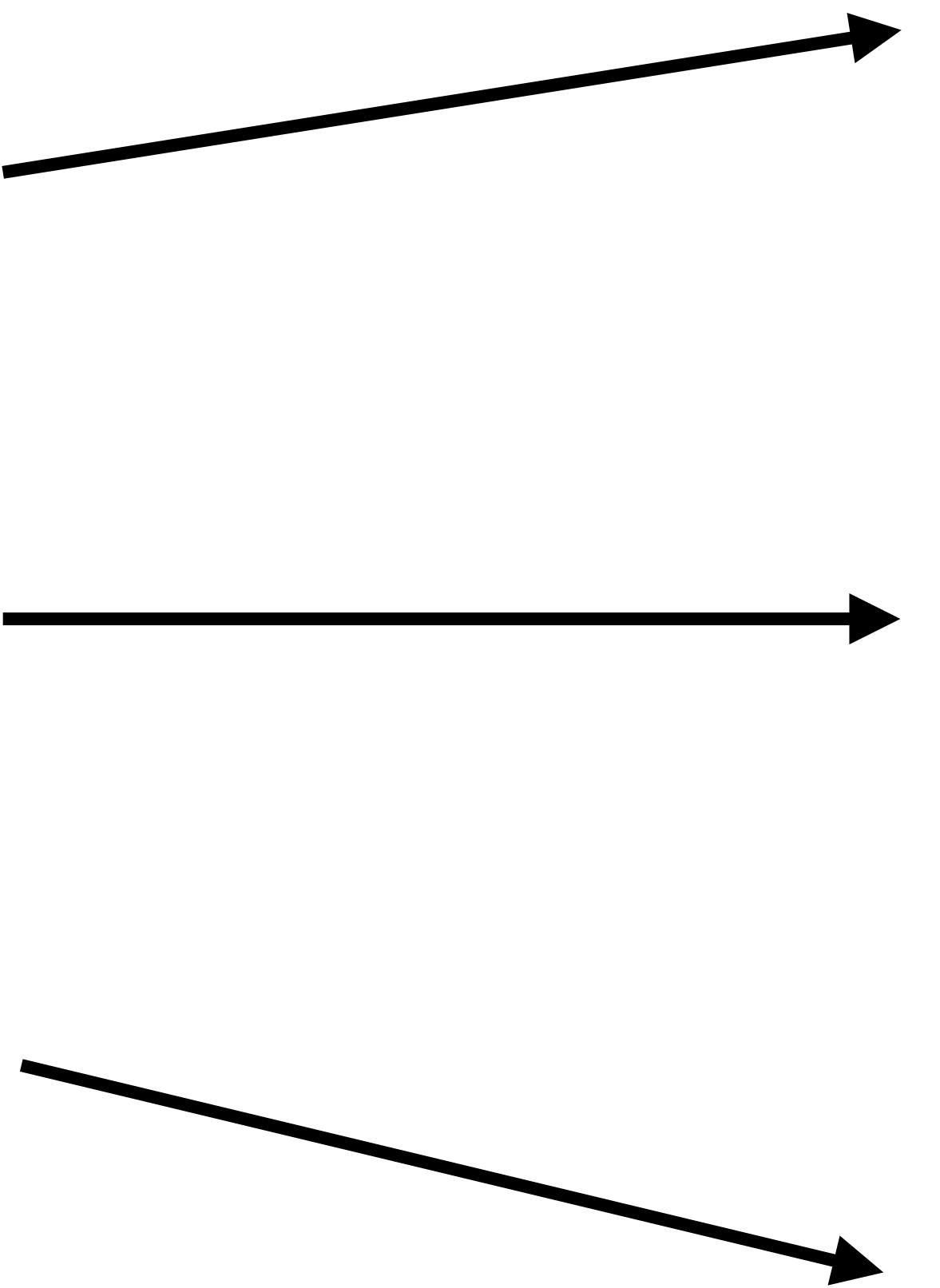
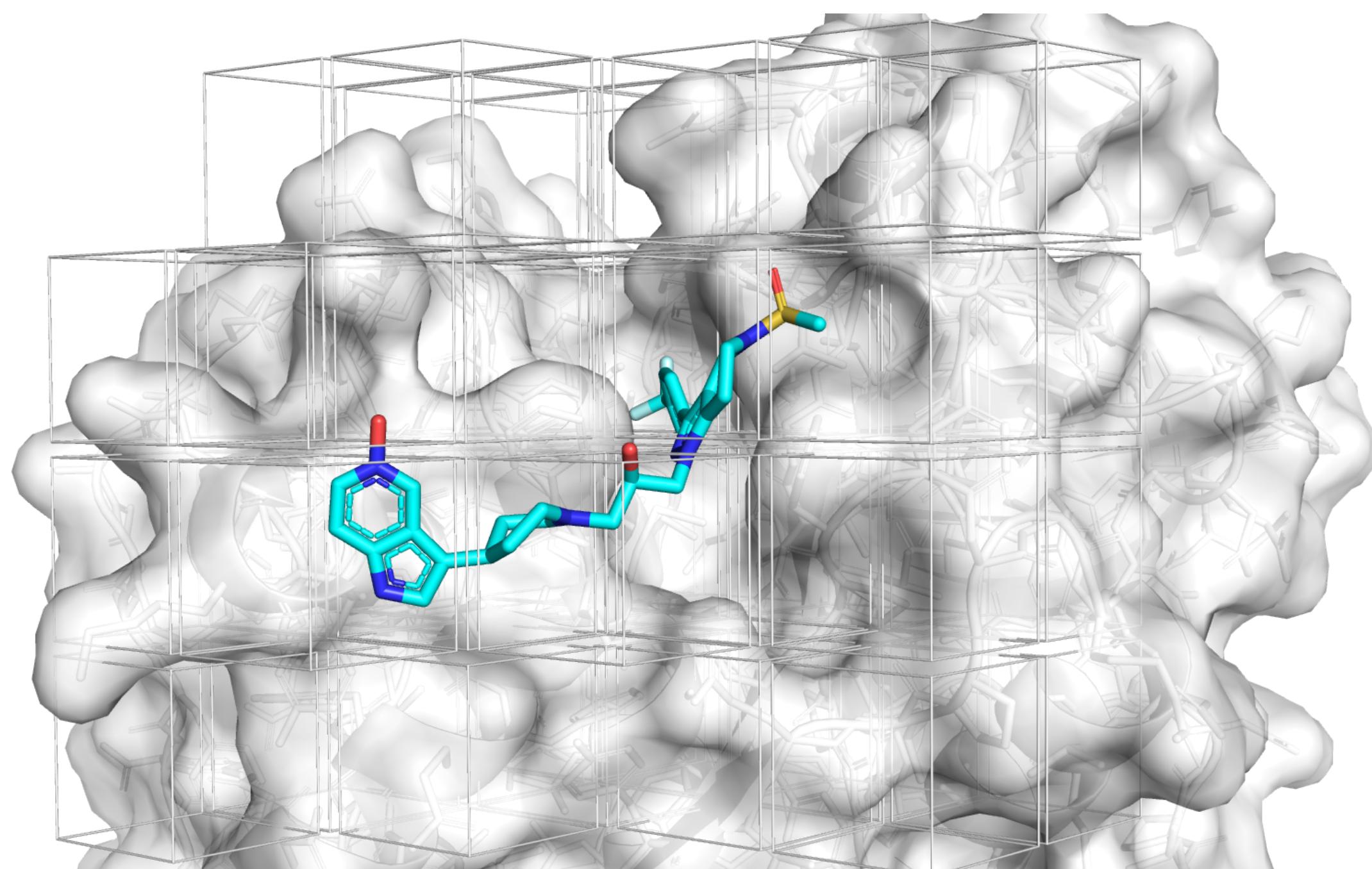


Caffe

providing support for:

- balanced, randomized, stratified batches
- temporal and spatial recurrences
- generation of tensors from molecular input data, and not just grids either!

libmolgrid



PYTORCH

Deep Learning with PyTorch

Caffe

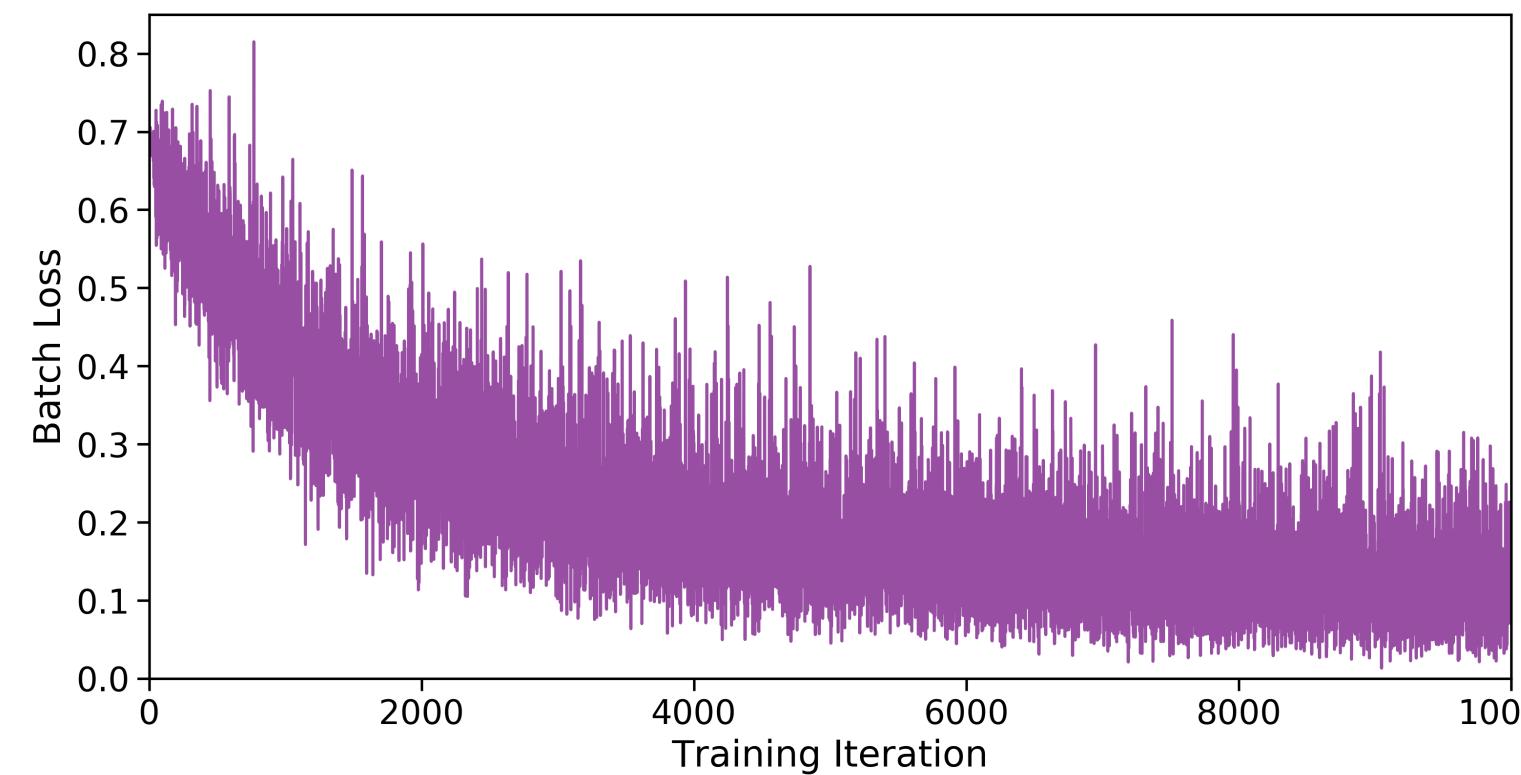
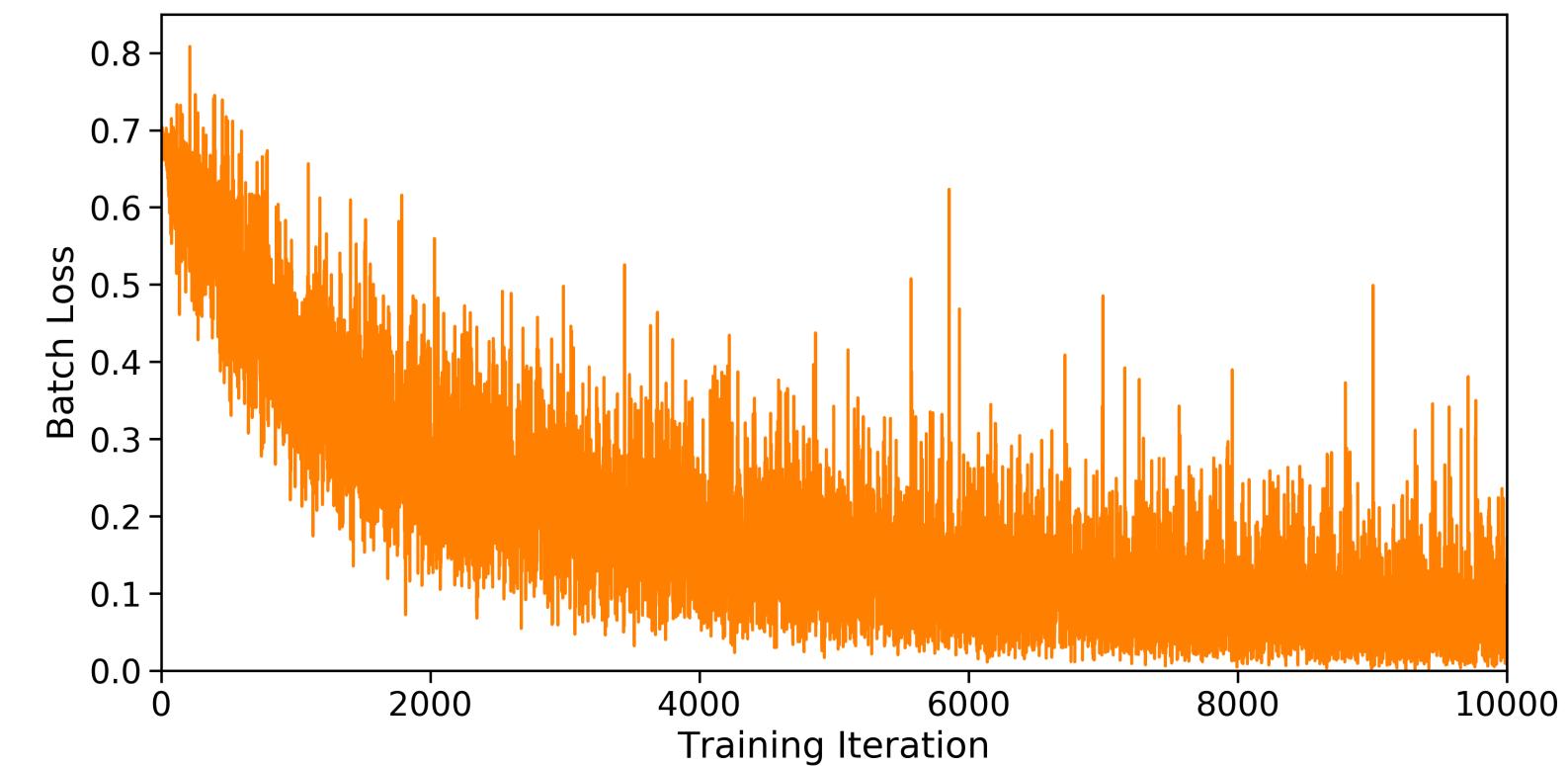
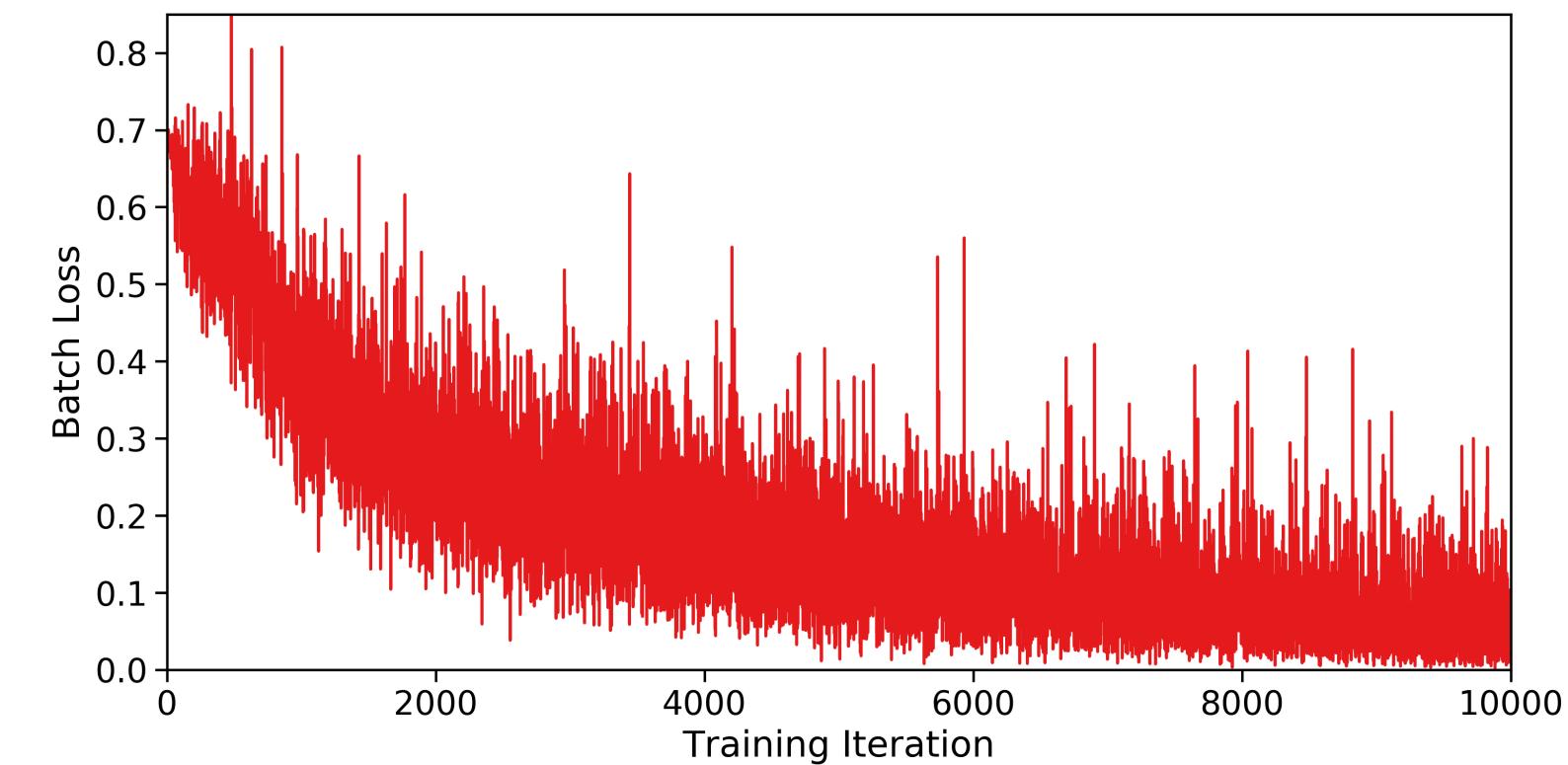
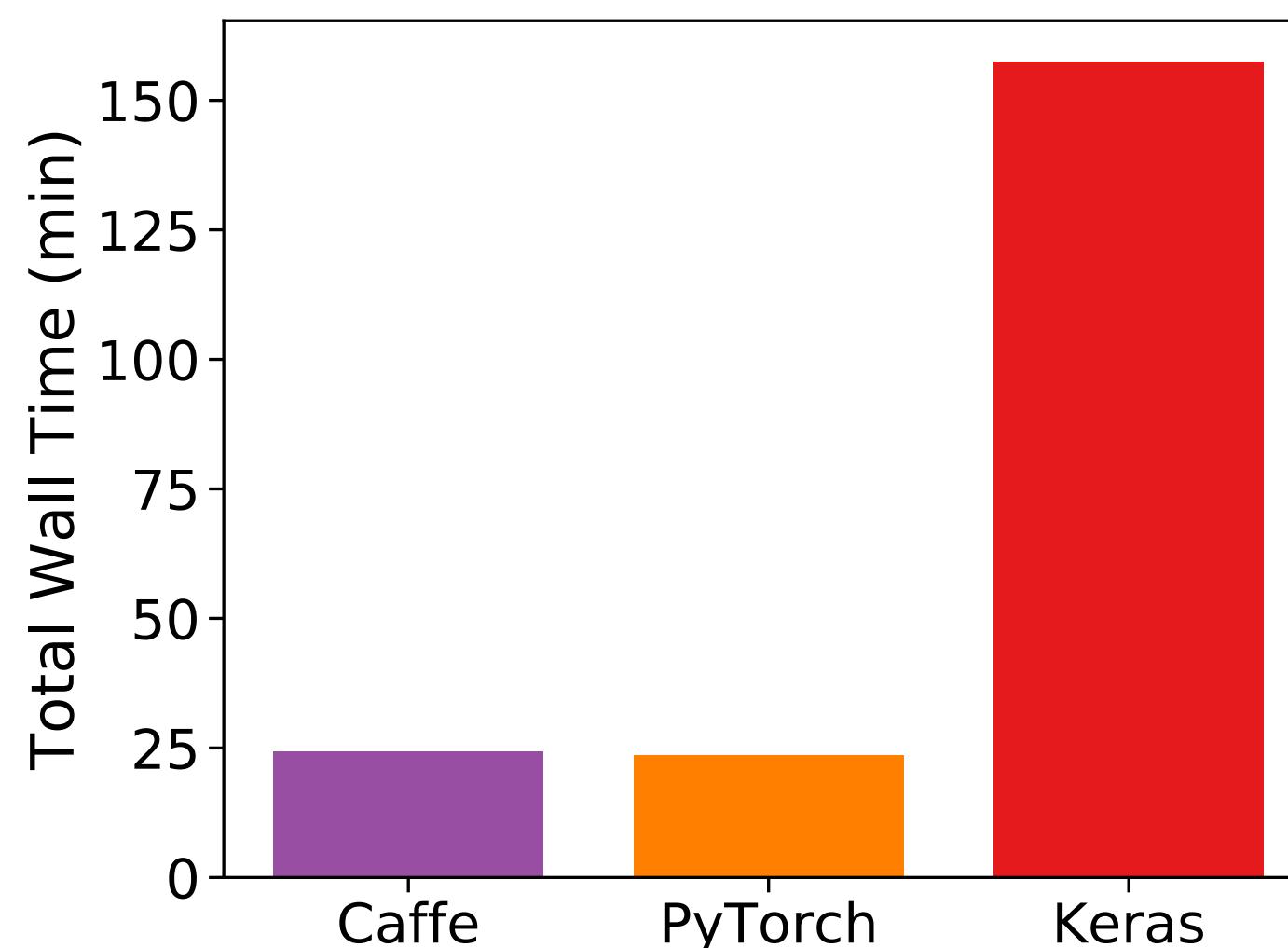
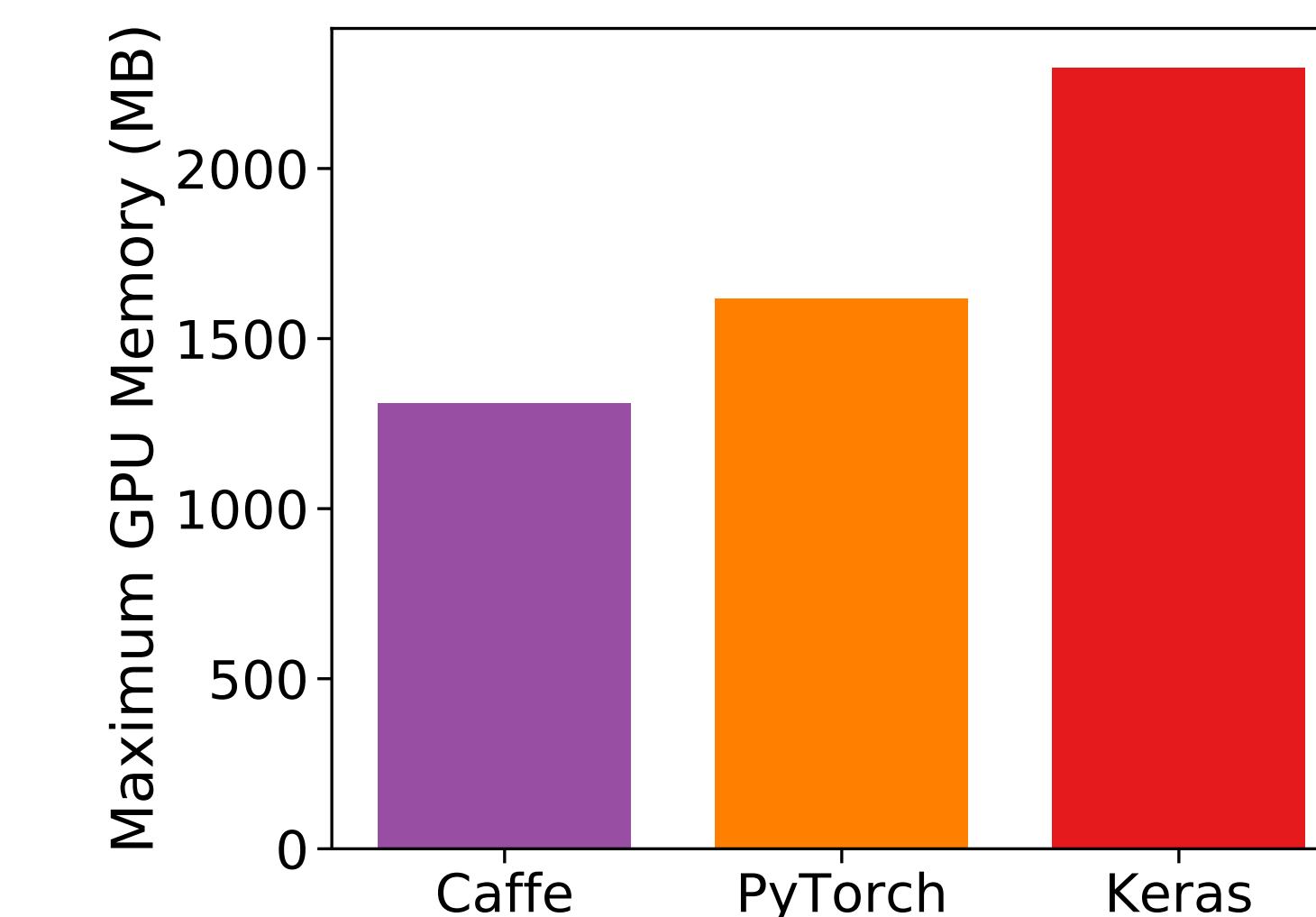


TensorFlow

providing support for:

- balanced, randomized, stratified batches
- temporal and spatial recurrences
- generation of tensors from molecular input data, and not just grids either!

libmolgrid

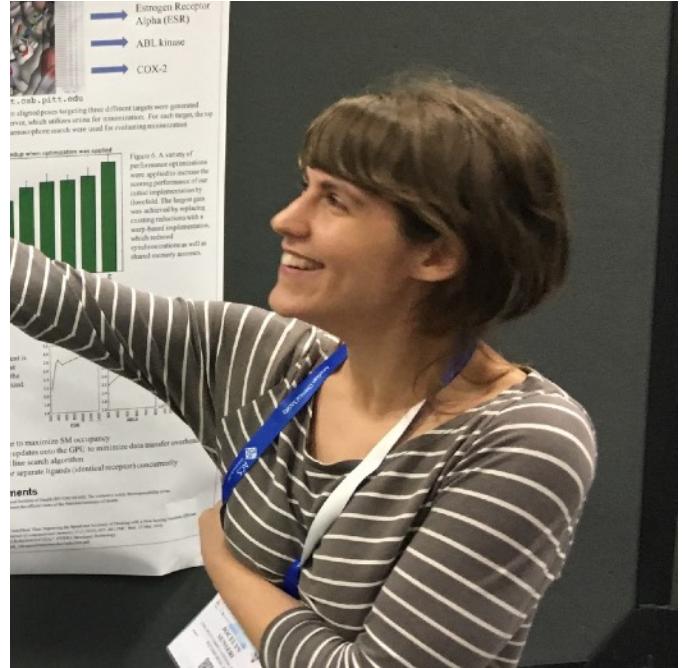
Caffe Training**PyTorch Training****Keras Training****GPU Performance****GPU Memory Utilization**

```
e = molgrid.ExampleProvider(balanced=True, shuffle=True)
e.populate('examples.txt')

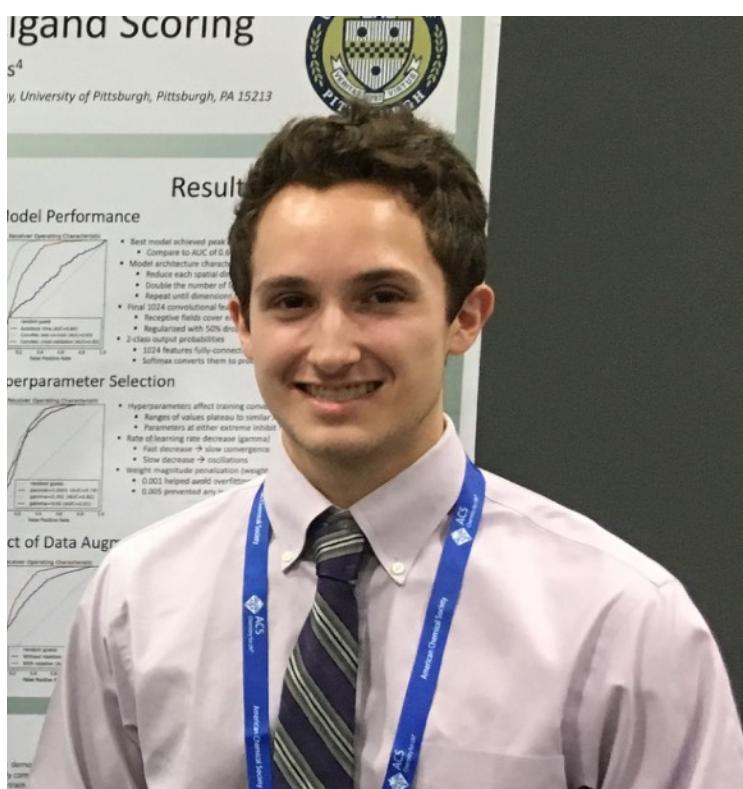
gmaker = molgrid.GridMaker()

batch = e.next_batch(batch_size)
gmaker.forward(batch, input_tensor,
               random_translation=0, random_rotation=True)
```

Acknowledgements



Jocelyn Sunseri



Matt Ragoza
Tomohide Masuda



Paul Francoeur



Department of
Computational and
Systems Biology



National Institute of
General Medical Sciences
R01GM108340

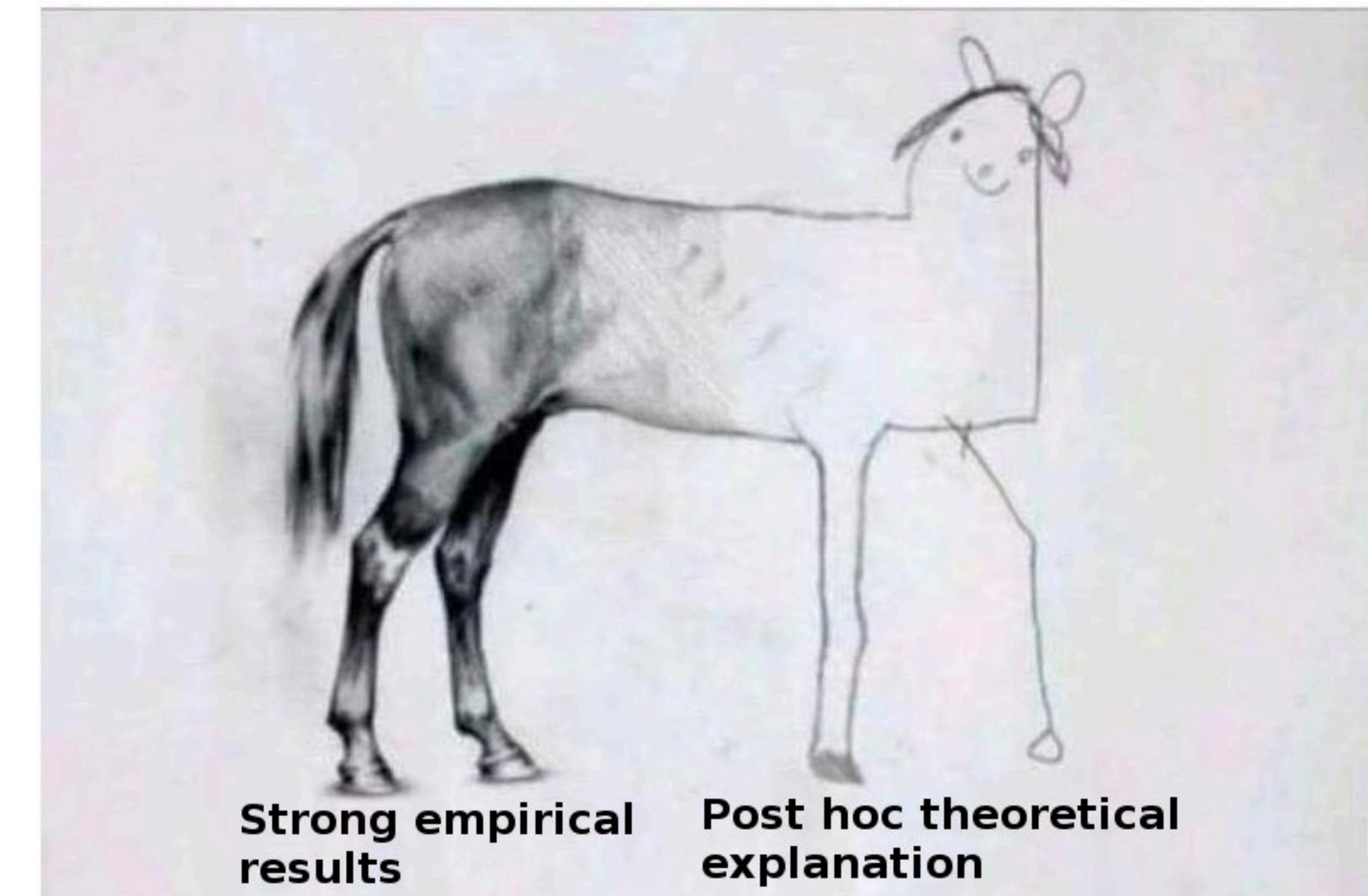


 github.com/gnina

 <http://bits.csb.pitt.edu>

 [@david_koes](https://twitter.com/david_koes)

Anatomy of a deep learning paper



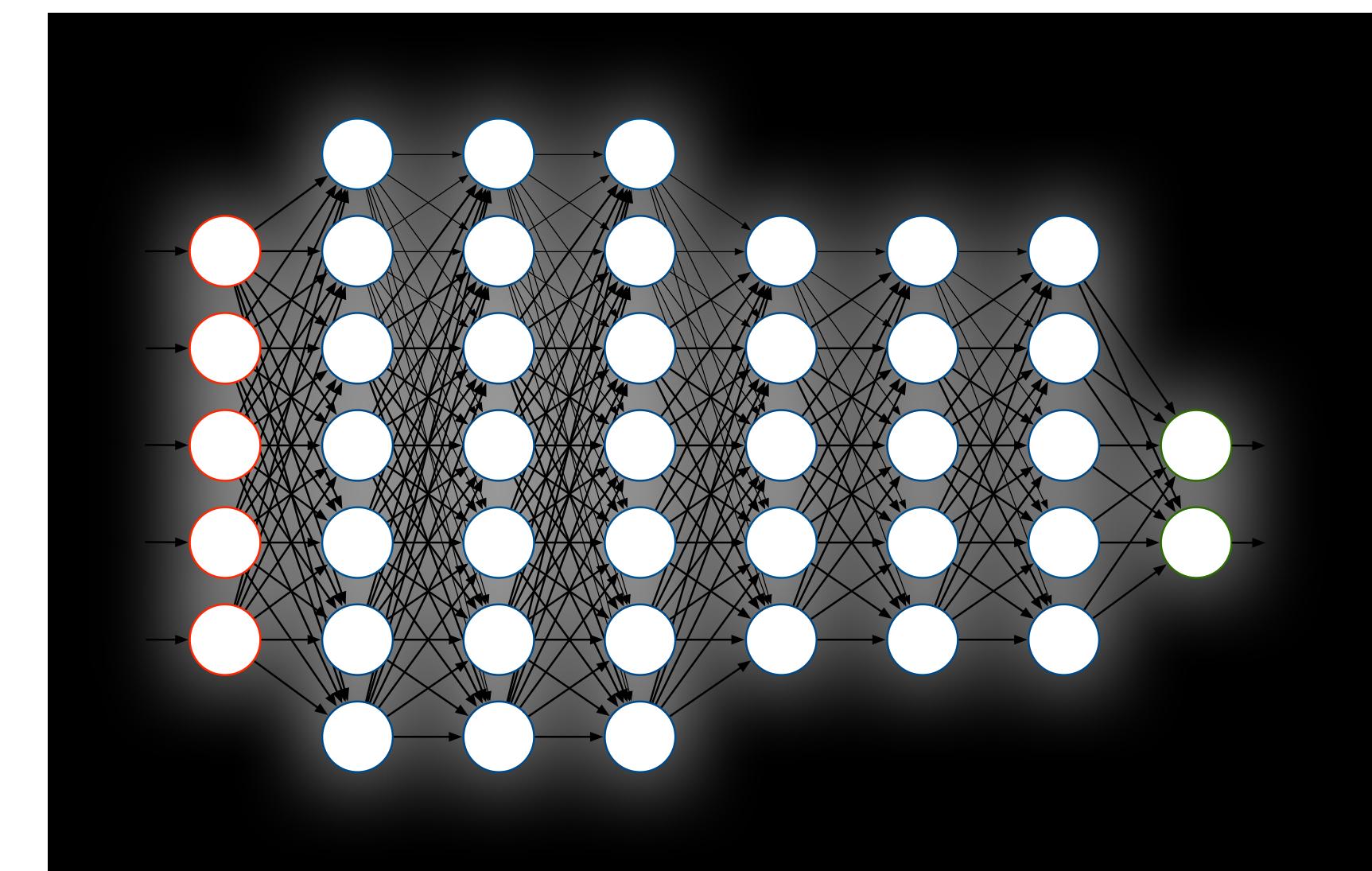
**Strong empirical
results**

**Post hoc theoretical
explanation**

Generative Modeling

Discriminative Model

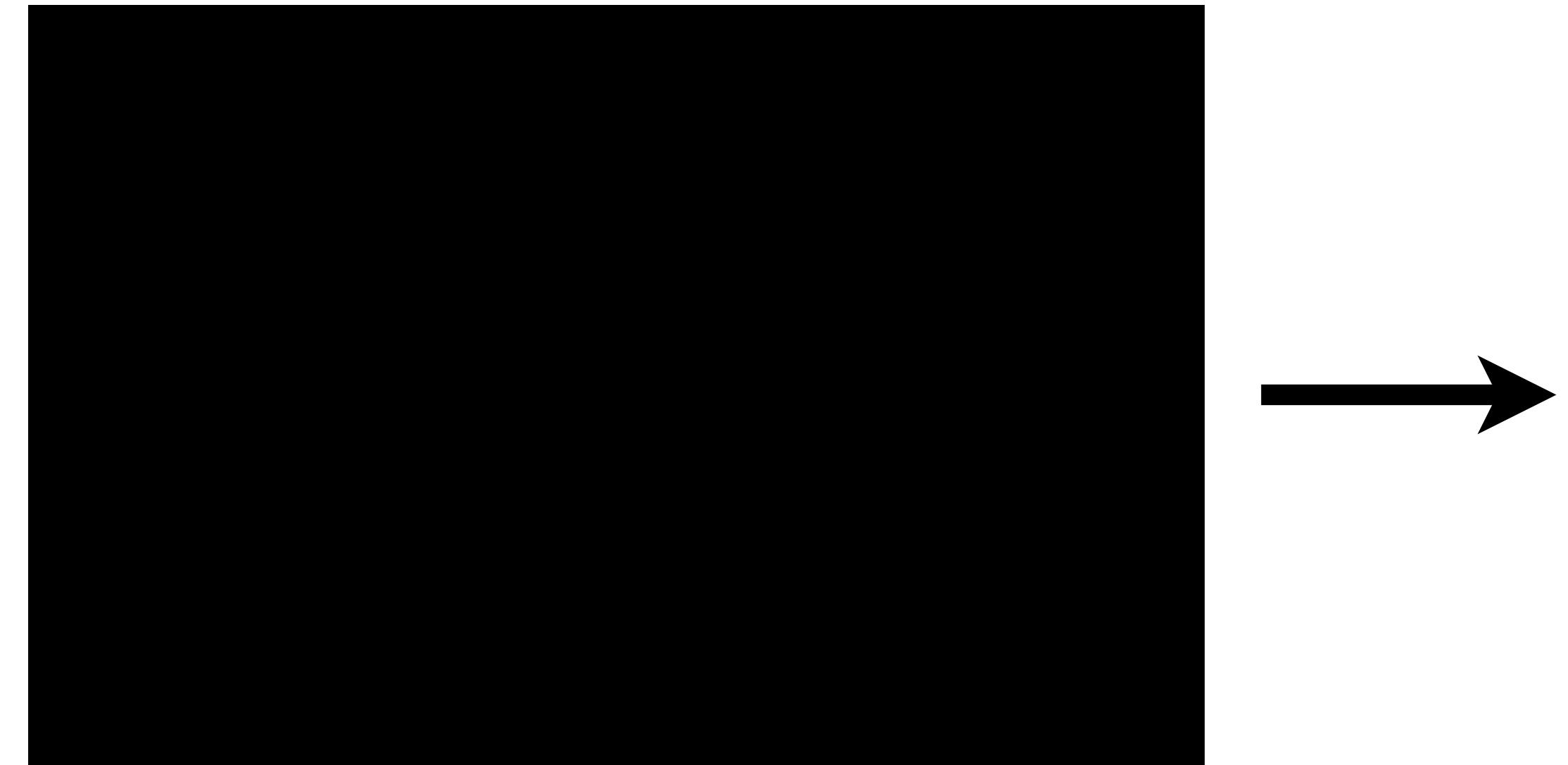
Features \mathbf{X} →



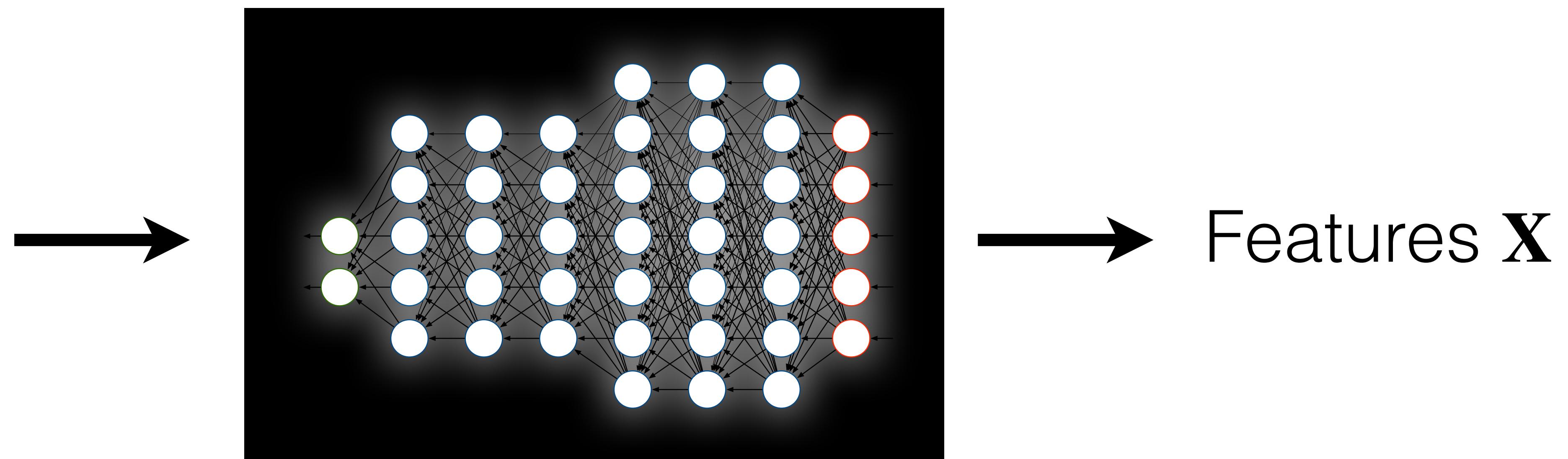
→ Prediction \mathbf{y}

Generative Model

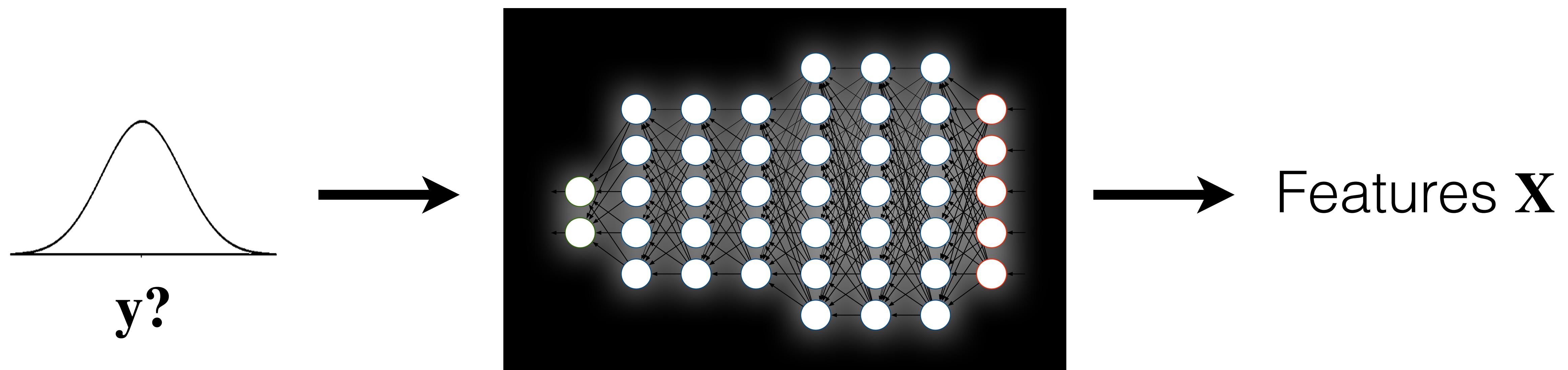
Features \mathbf{X} →



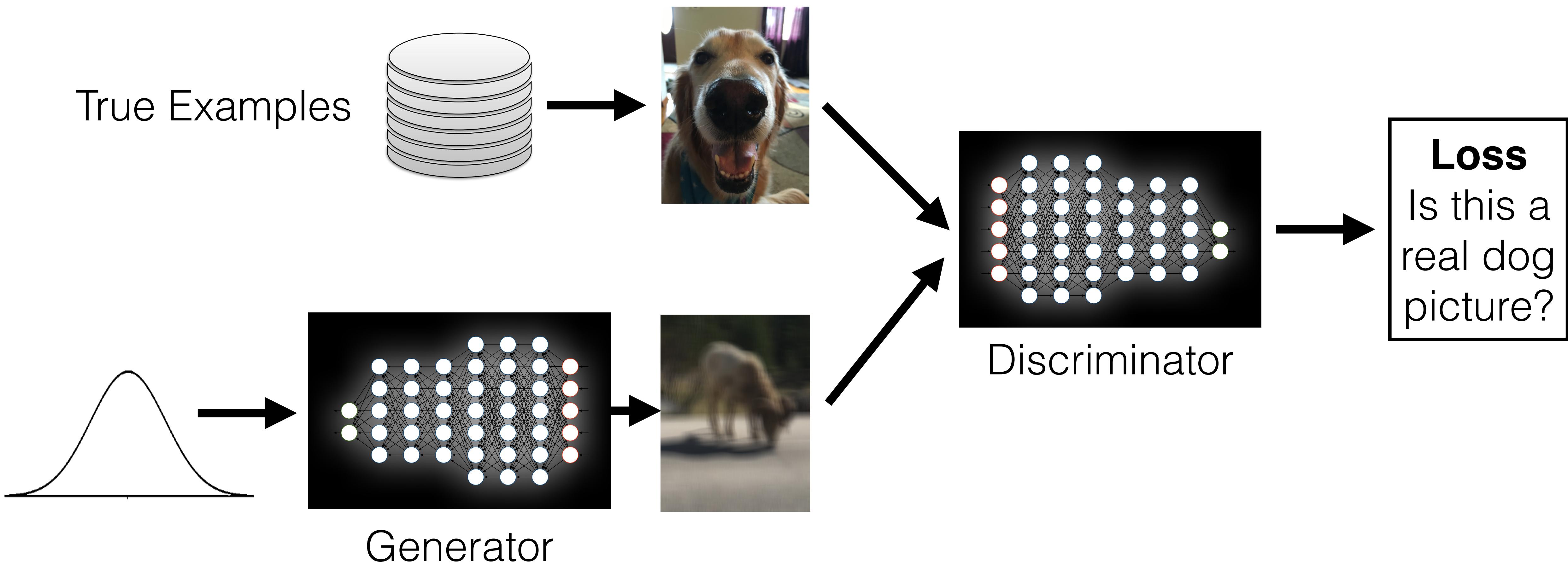
Generative Model



Generative Model



Generative Adversarial Networks



Generative Adversarial Networks



Ian Goodfellow @goodfellow_ian · 2h

4.5 years of GAN progress on face generation. arxiv.org/abs/1406.2661

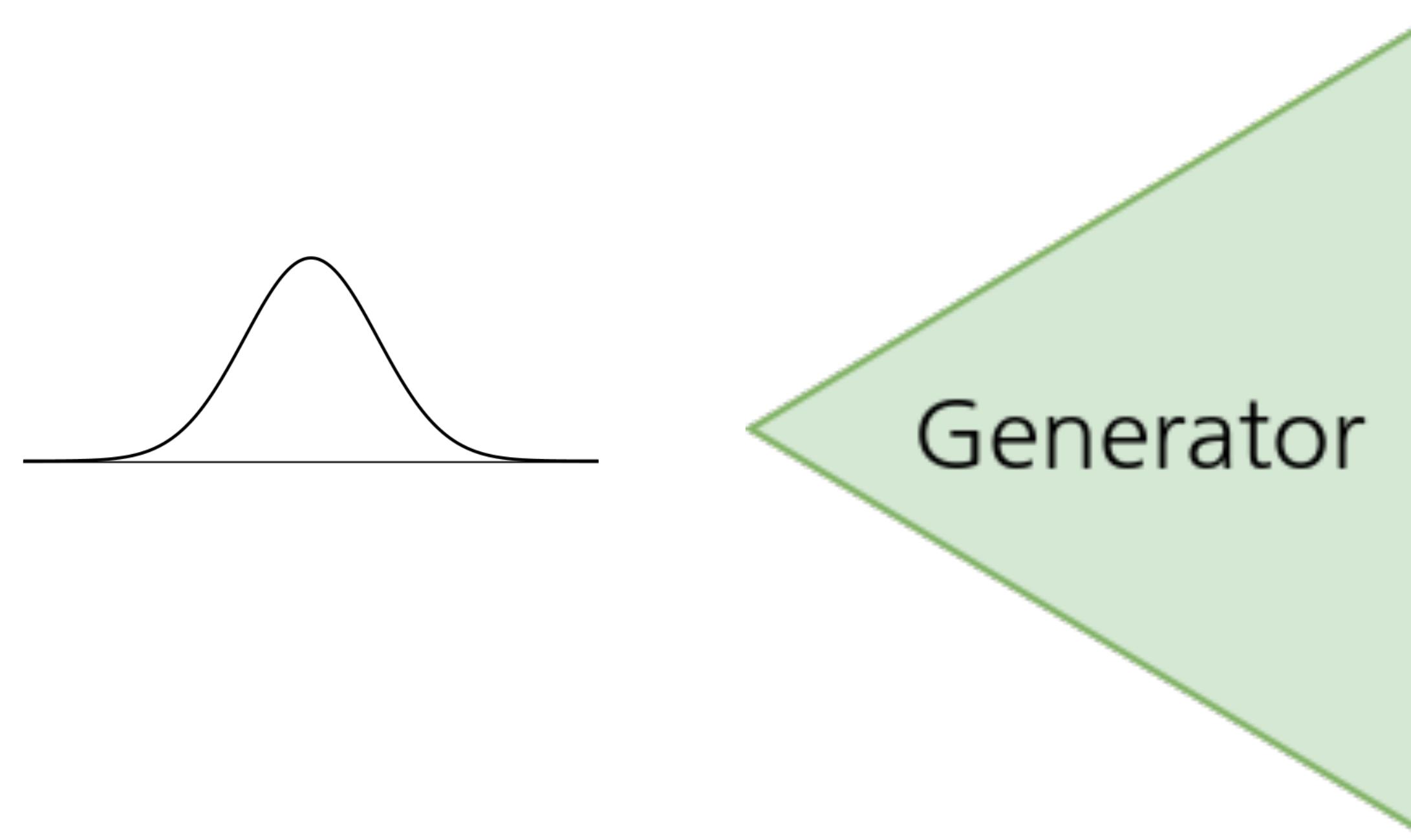
arxiv.org/abs/1511.06434 arxiv.org/abs/1606.07536 arxiv.org/abs/1710.10196

arxiv.org/abs/1812.04948

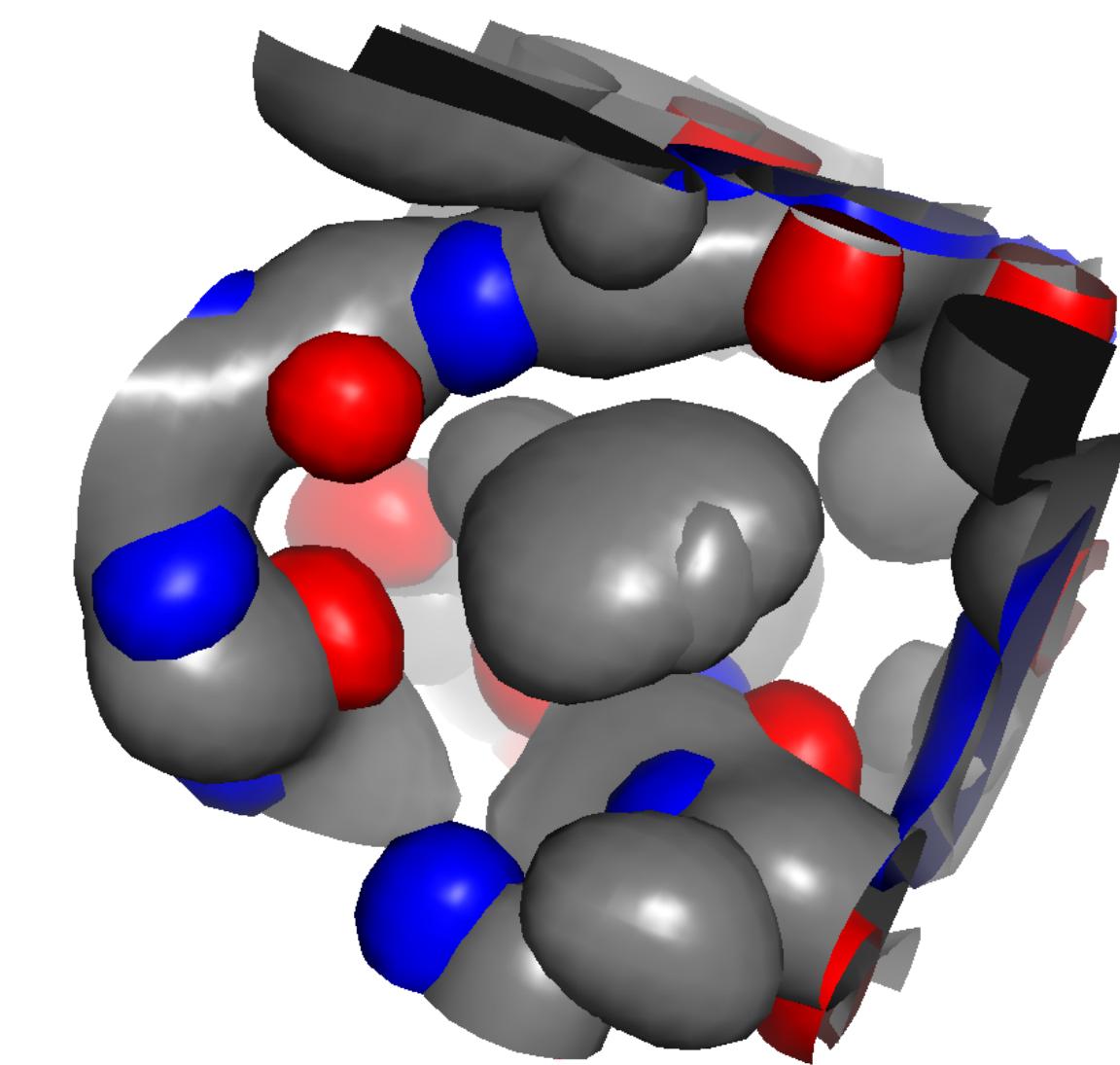
<https://thispersondoesnotexist.com>

Generative Models

Generative models approximate a data distribution directly. They can map samples from one distribution (noise or input data) to realistic samples from an output distribution of interest.

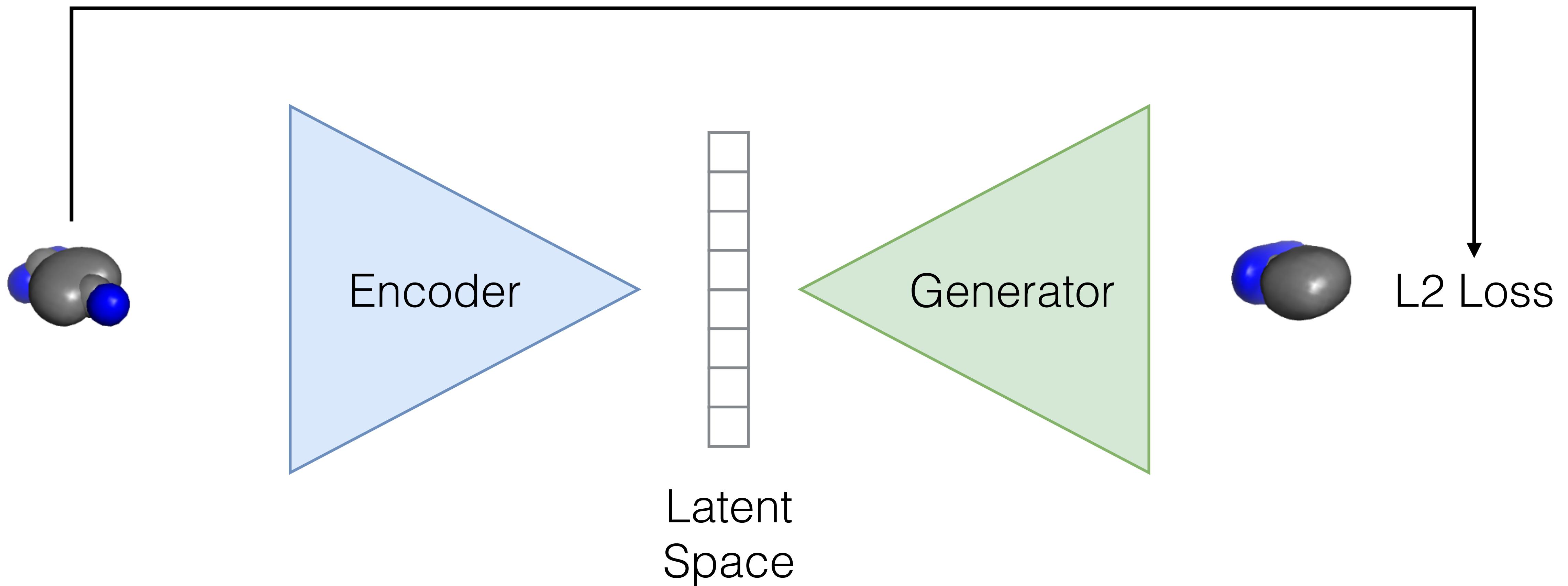


noise sample



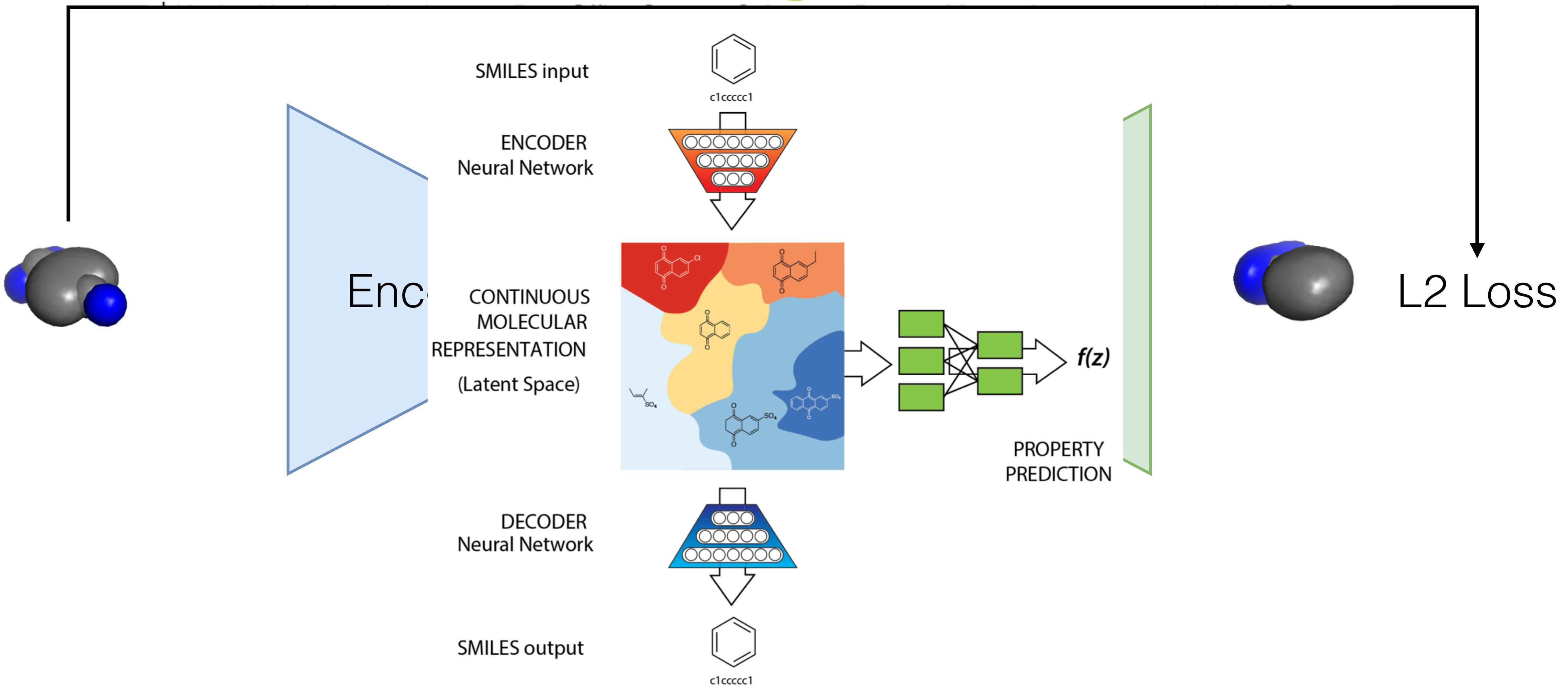
generated receptor & ligand grid

Autoencoding

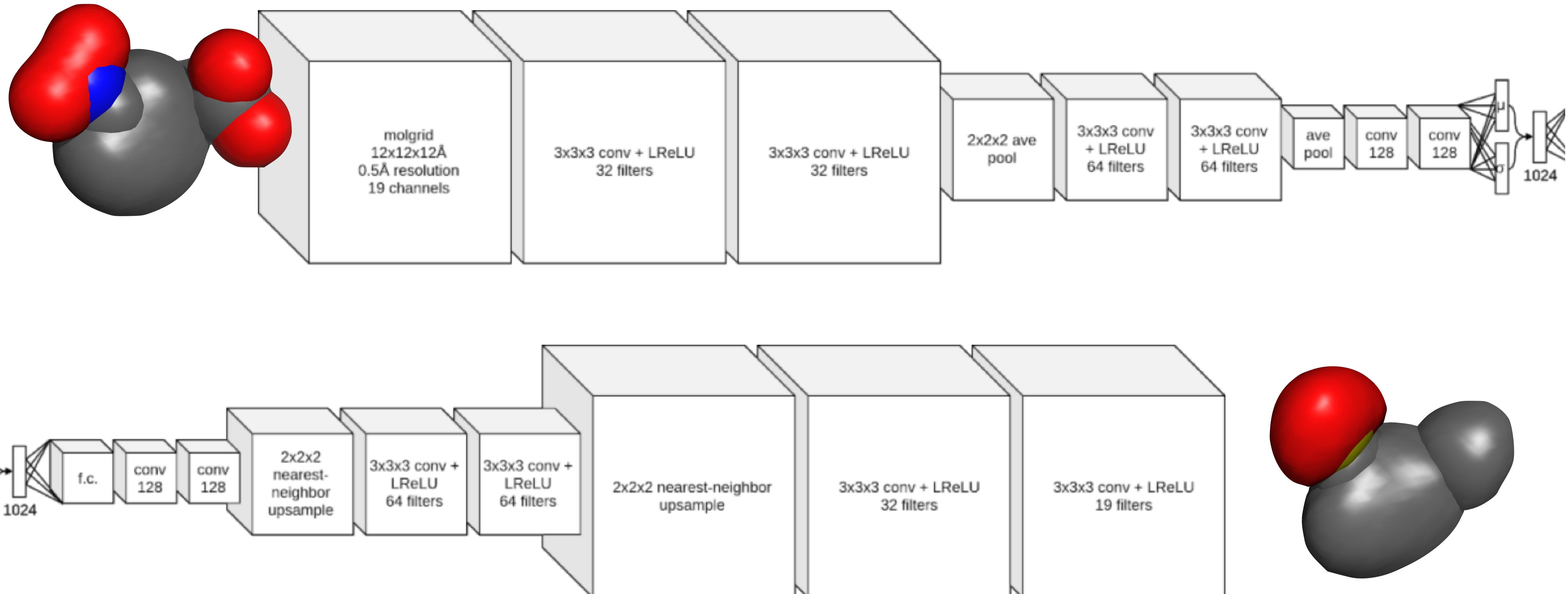


Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

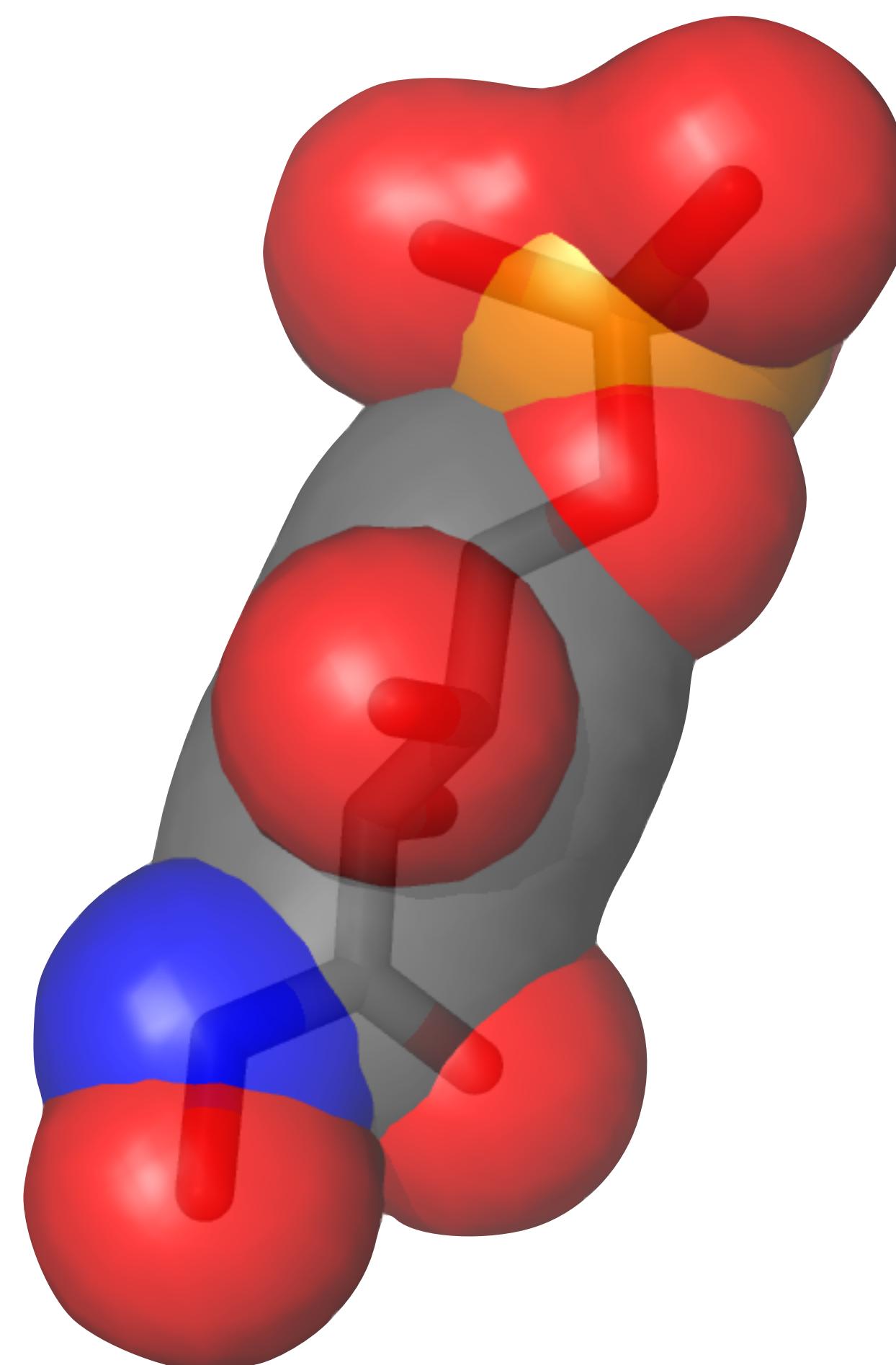
Rafael Gómez-Bombarelli^{†‡} , Jennifer N. Wei^{‡#} , David Duvenaud^{¶†} , José Miguel Hernández-Lobato^{§#} , Benjamín Sánchez-Lengeling[‡], Dennis Sheberla[‡] , Jorge Aguilera-Iparraguirre[†], Timothy D. Hirzel[†], Ryan P. Adams^{¶VI}, and Alán Aspuru-Guzik^{*‡⊥} 



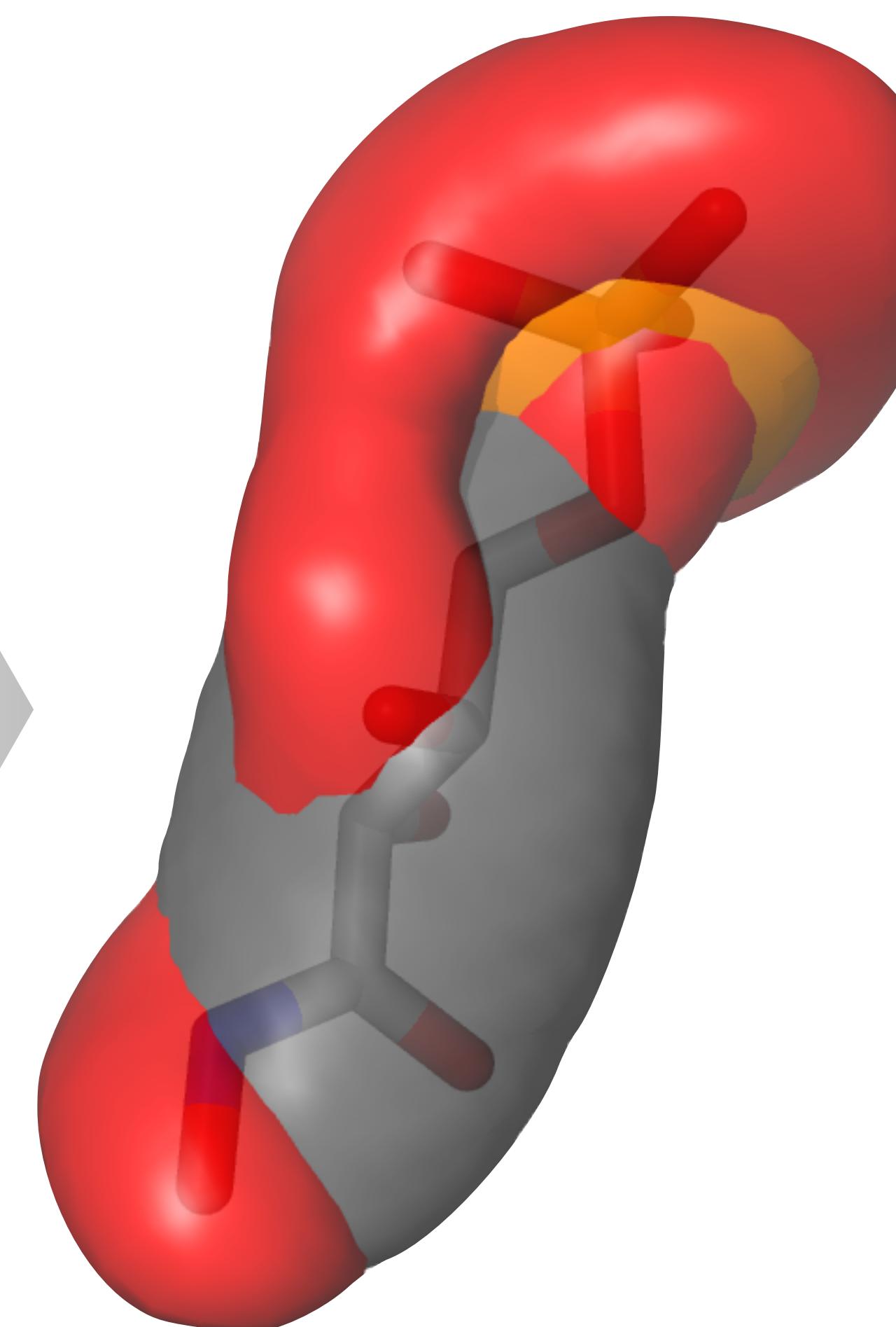
Model



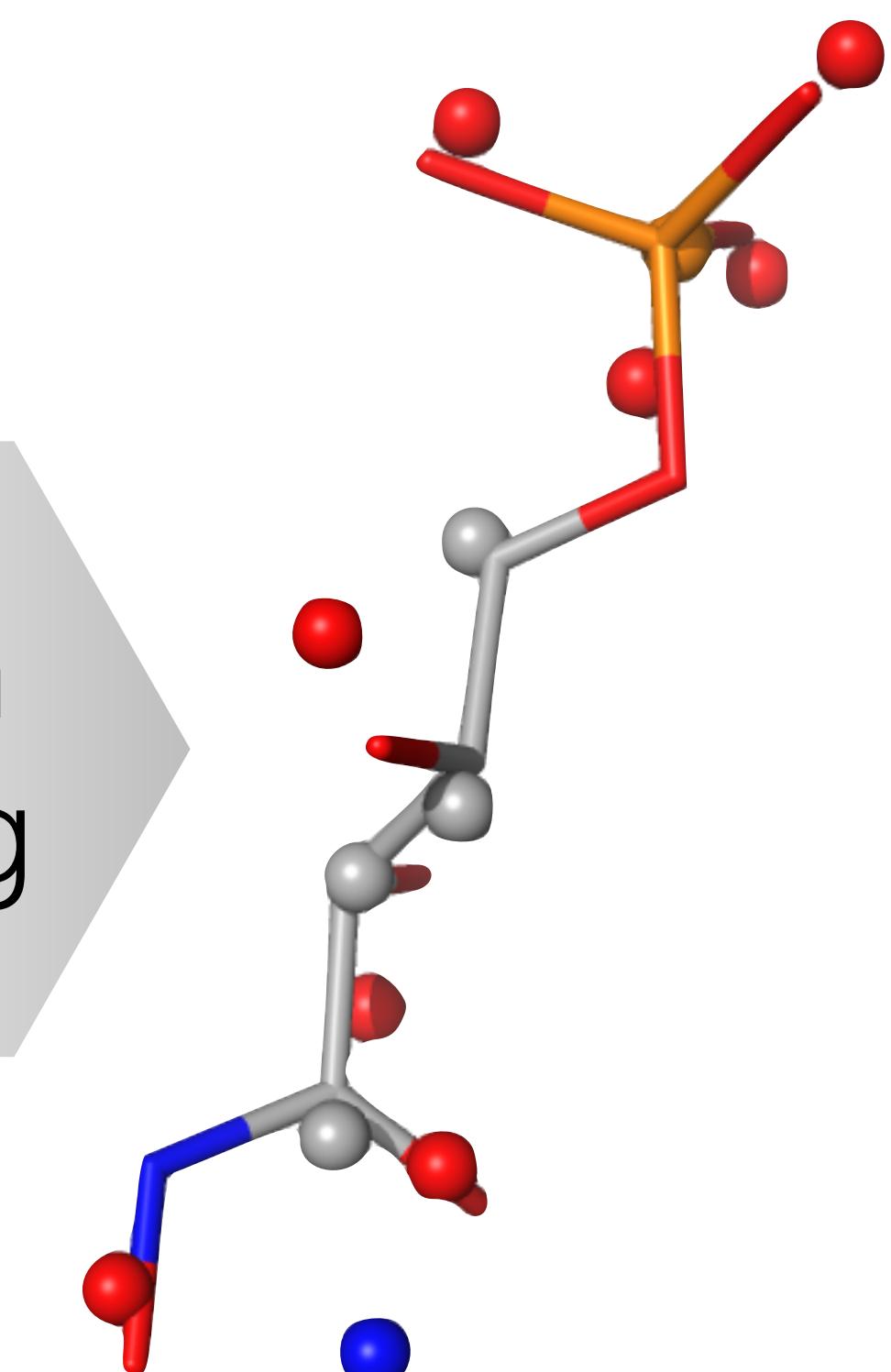
Variational Autoencoding Examples



VAE

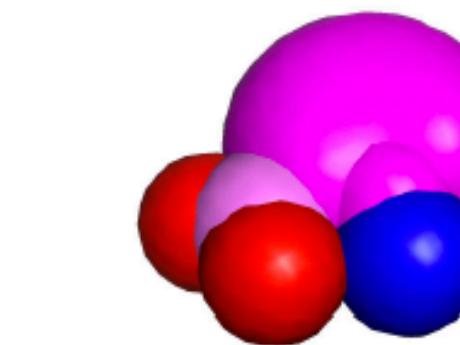
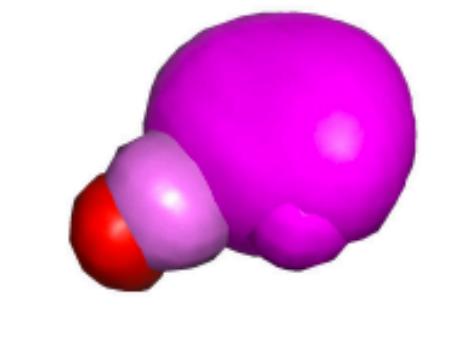
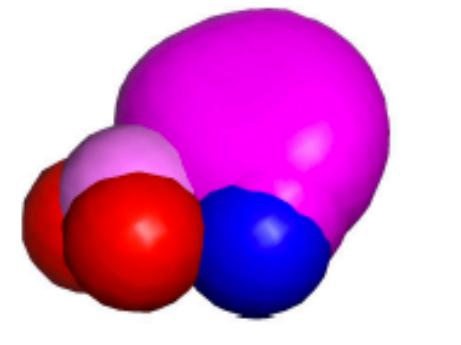
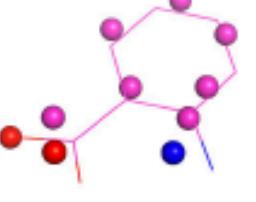
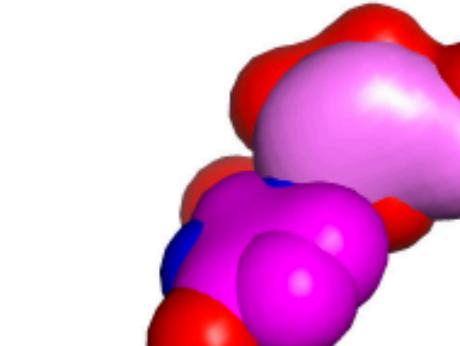
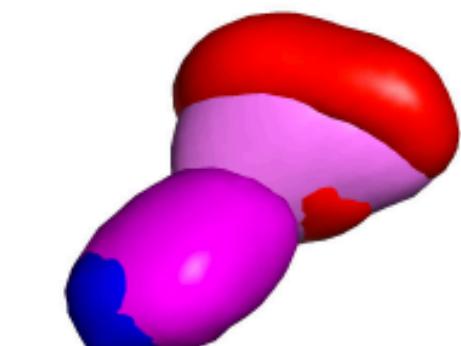
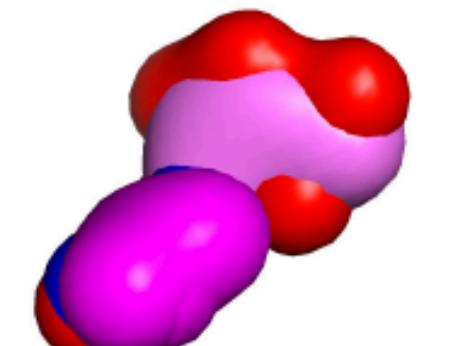
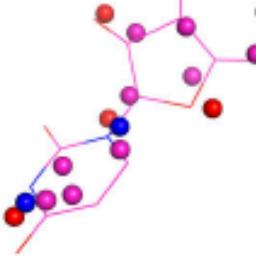
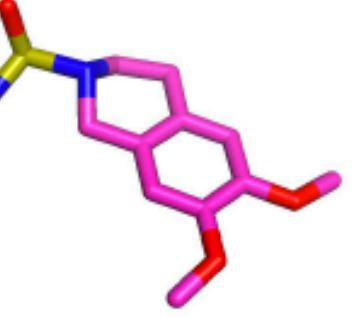
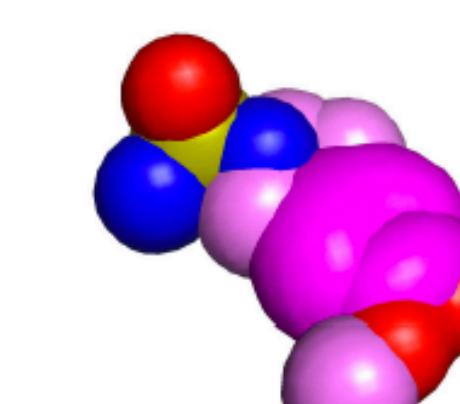
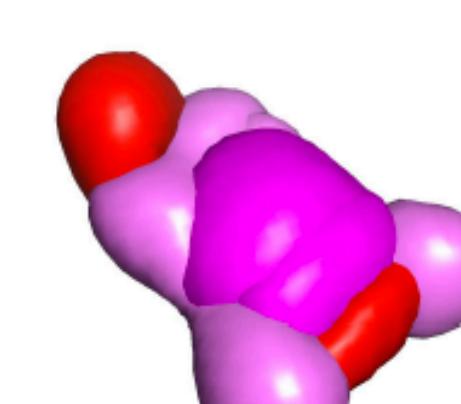
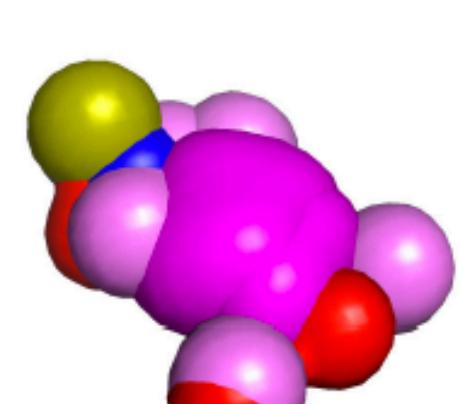
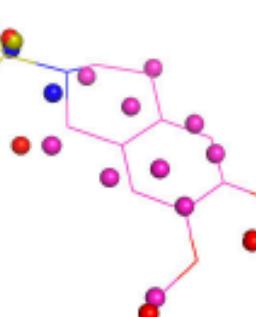
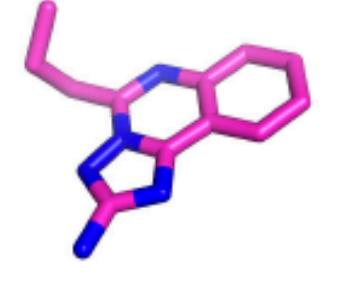
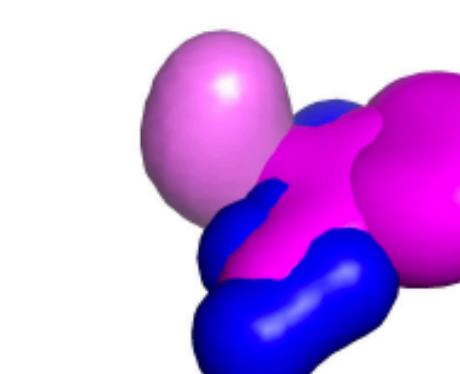
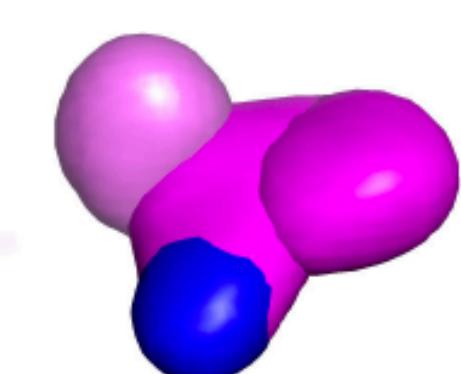
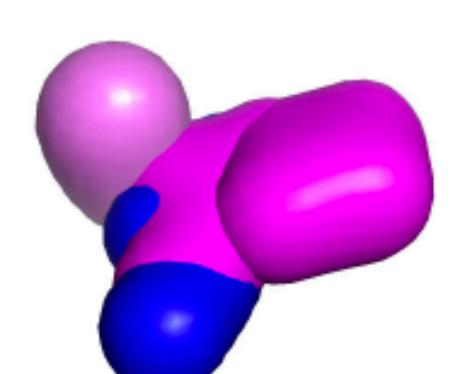
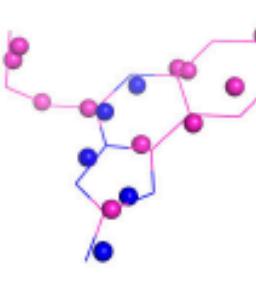


Atom
Fitting

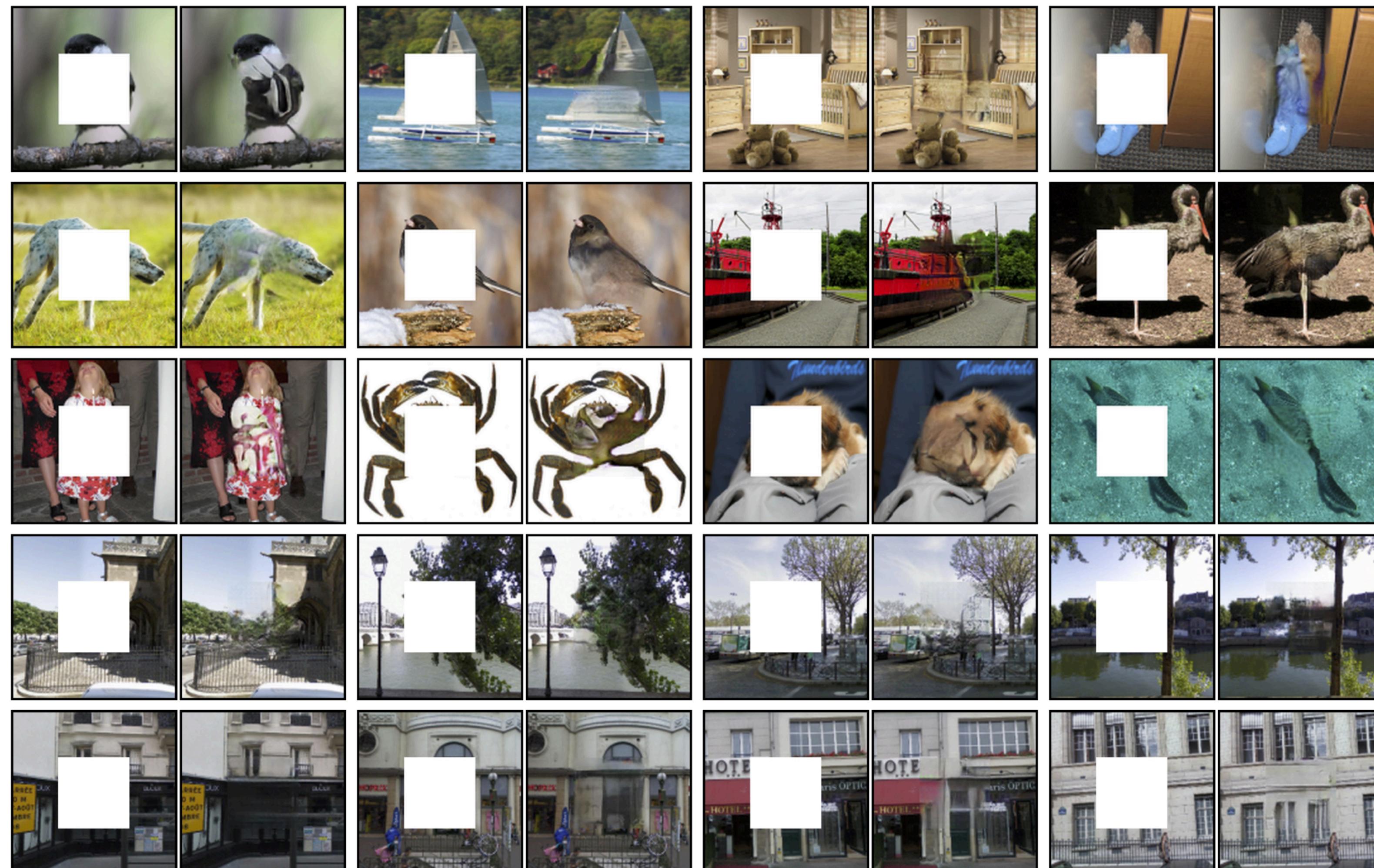


2BES

Variational Autoencoding Examples

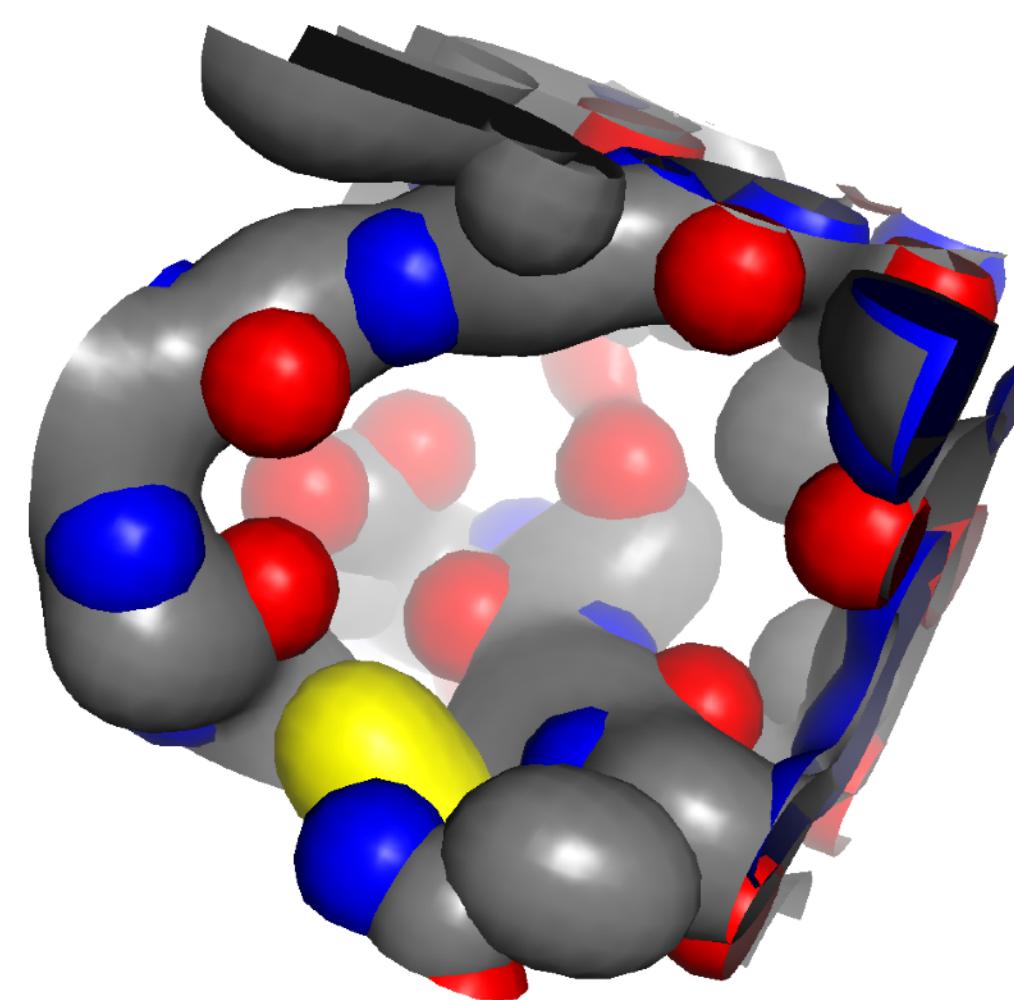
PDB	True structure	True density	Gen. density	Fit density	Fit structure	Gen. L2 distance	Fit L2 distance	Fit RMSD
3h78						9.4053	8.3141	0.6160
4jx9						13.8545	9.7198	0.8820
3igp						14.8525	12.5245	1.2066
4cwf						11.4730	9.0564	0.6725

Context Encoding

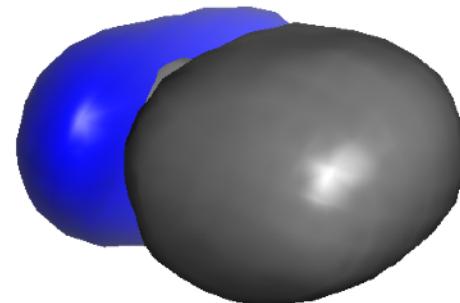
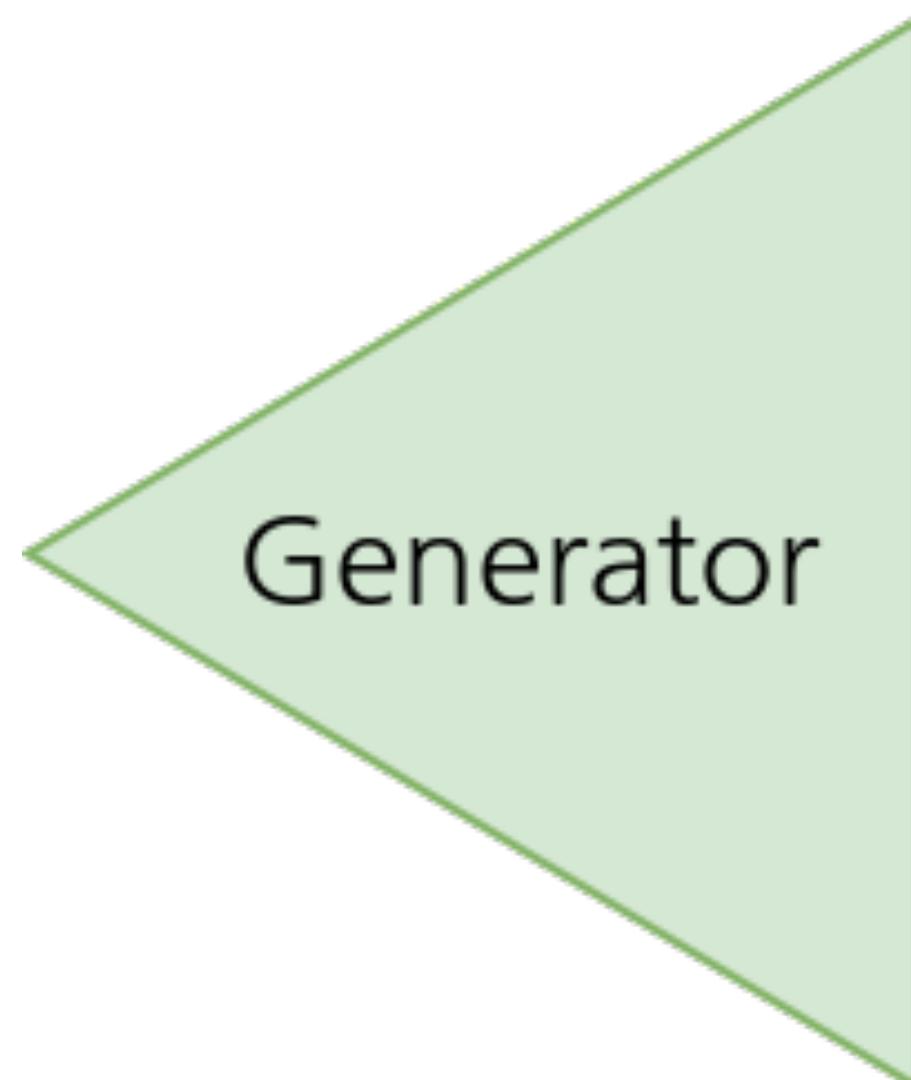


http://people.eecs.berkeley.edu/~pathak/context_encoder/

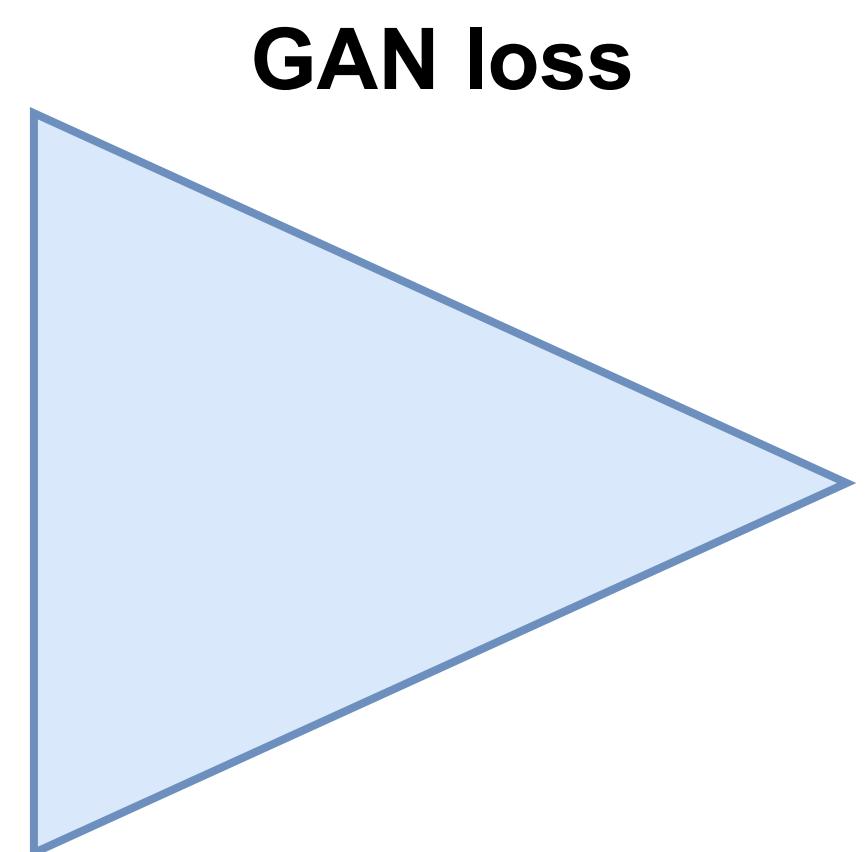
Context Encoding



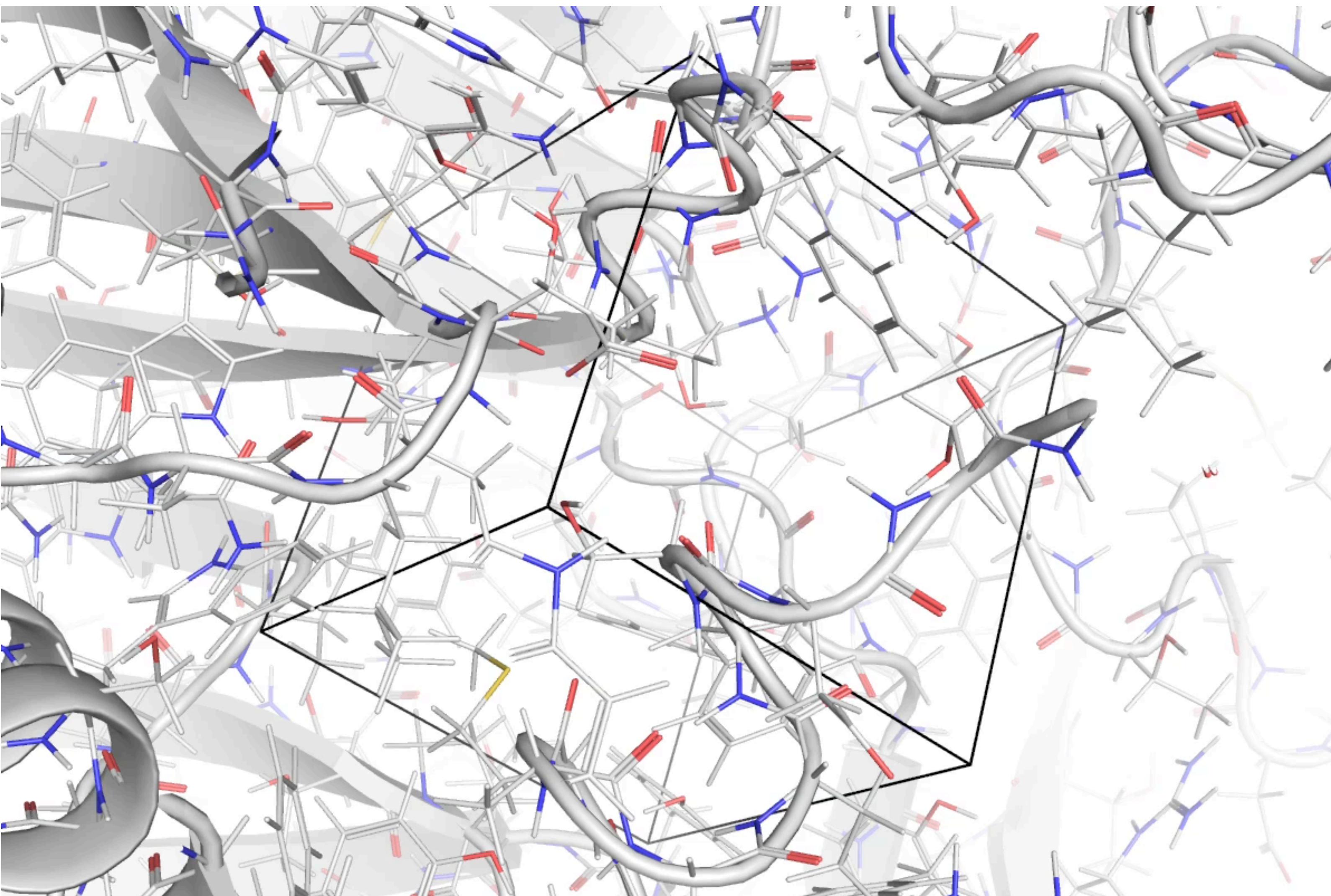
receptor grid



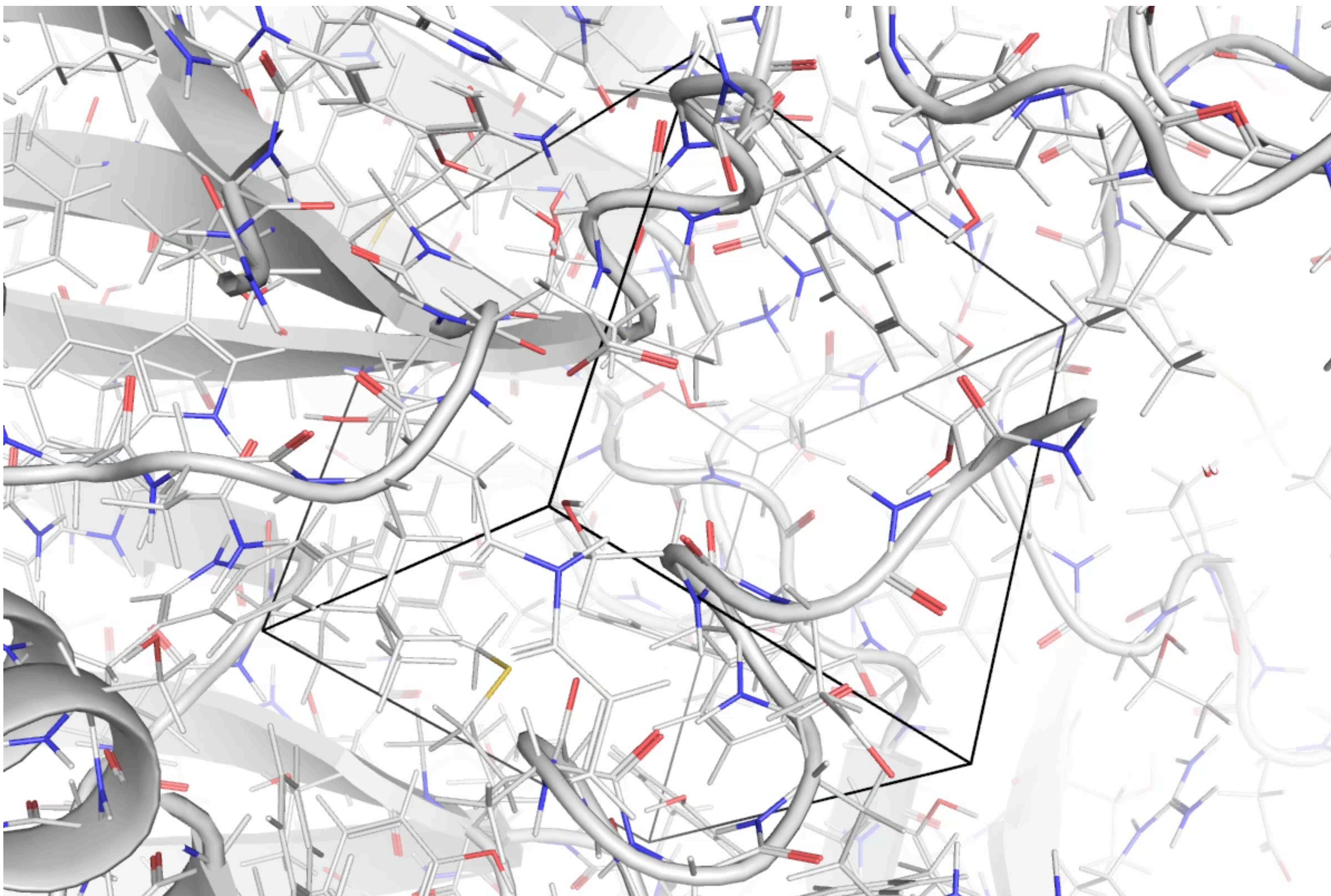
generated ligand grid



Conditioning on the Receptor

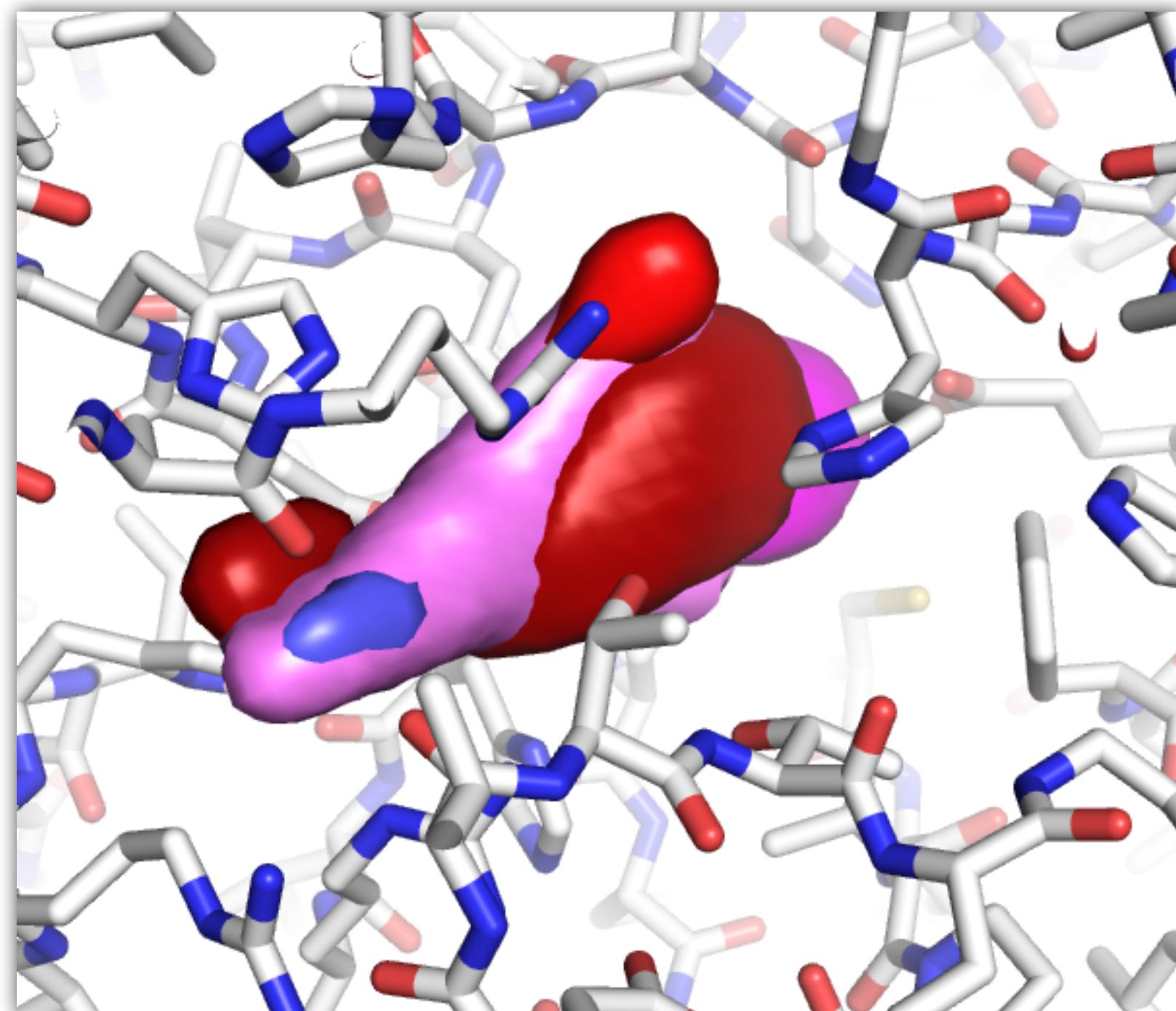


Conditioning on the Receptor

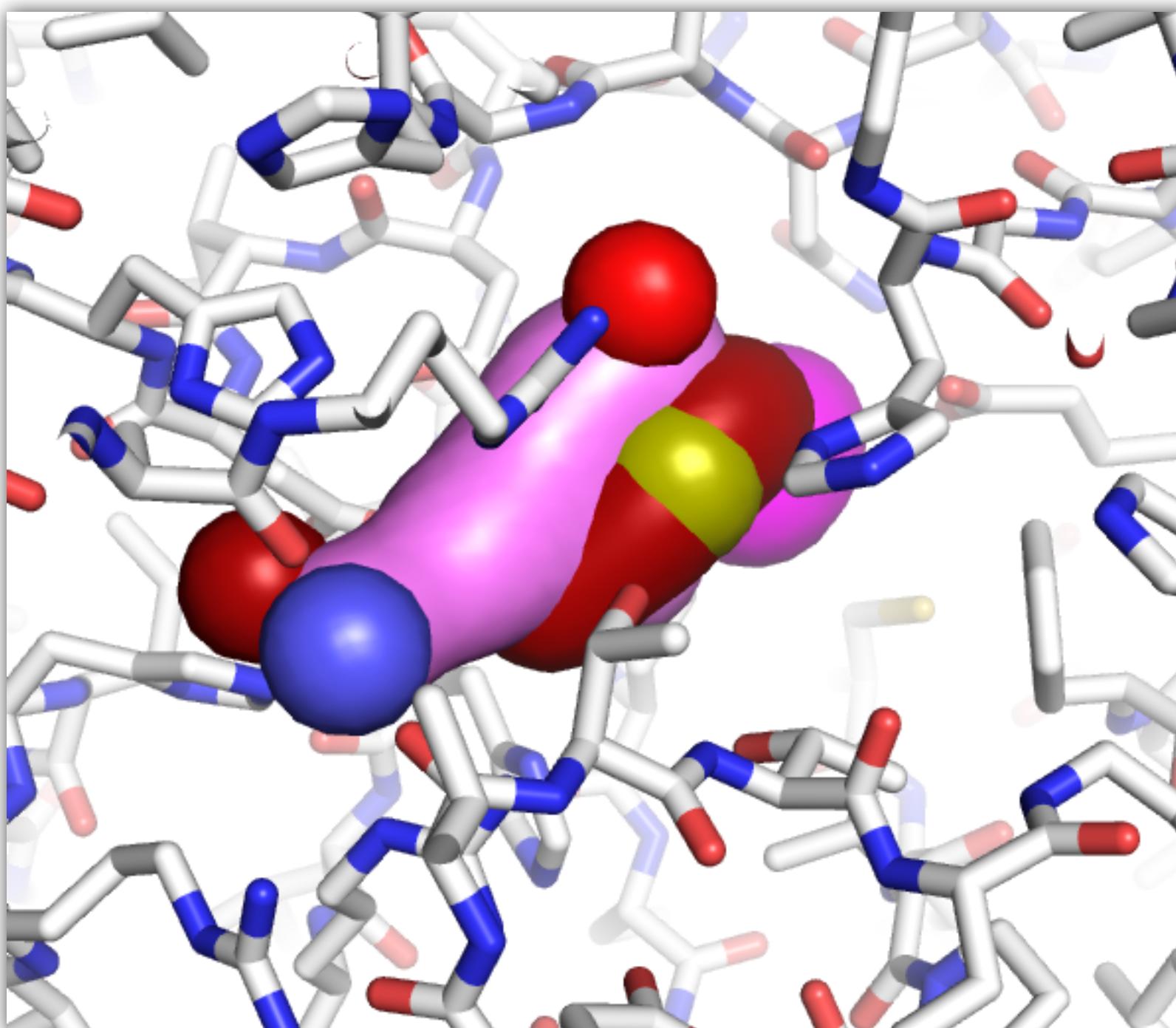


Context Encoding with Fully Convolutional Network

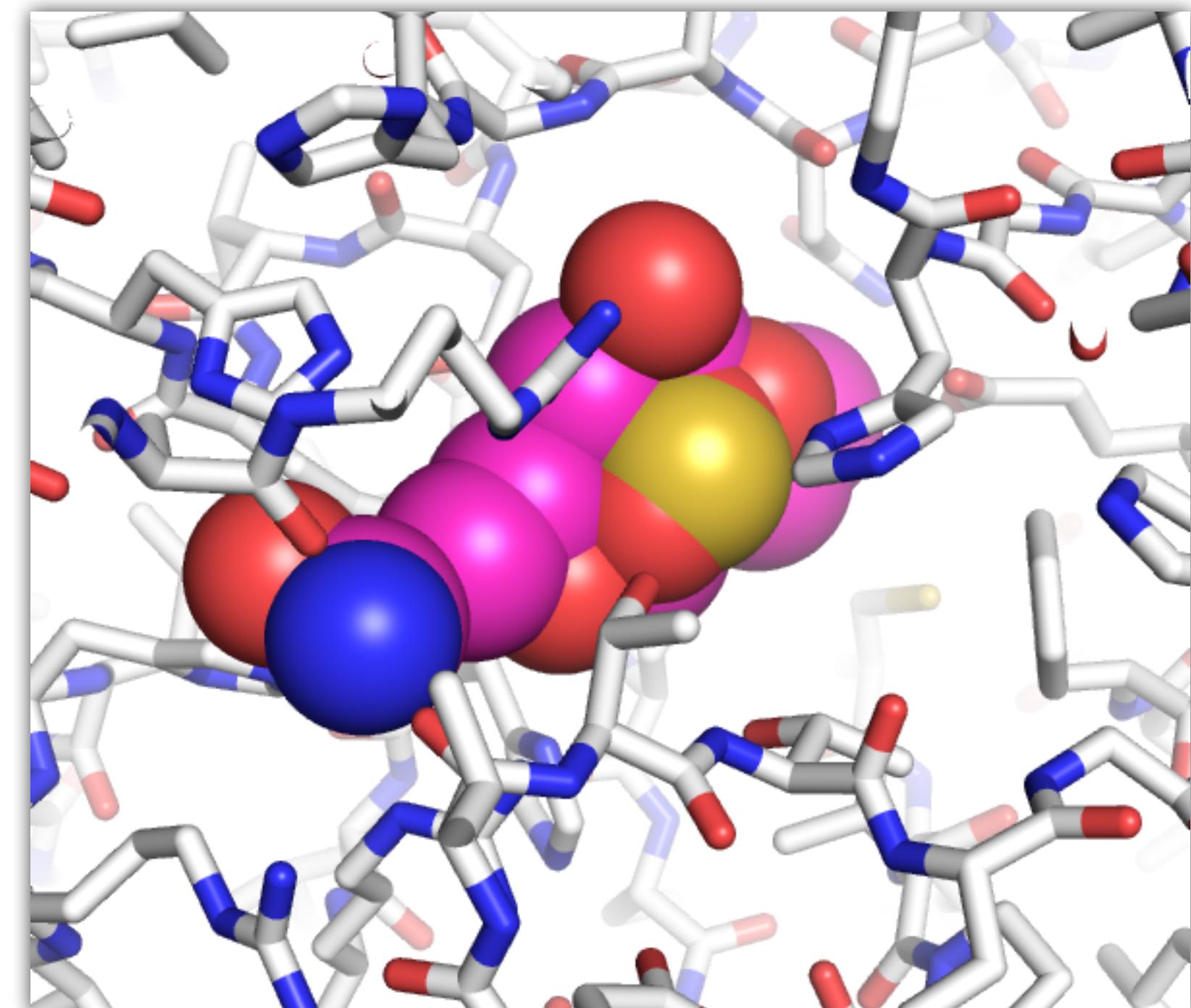
1m5w



Generated



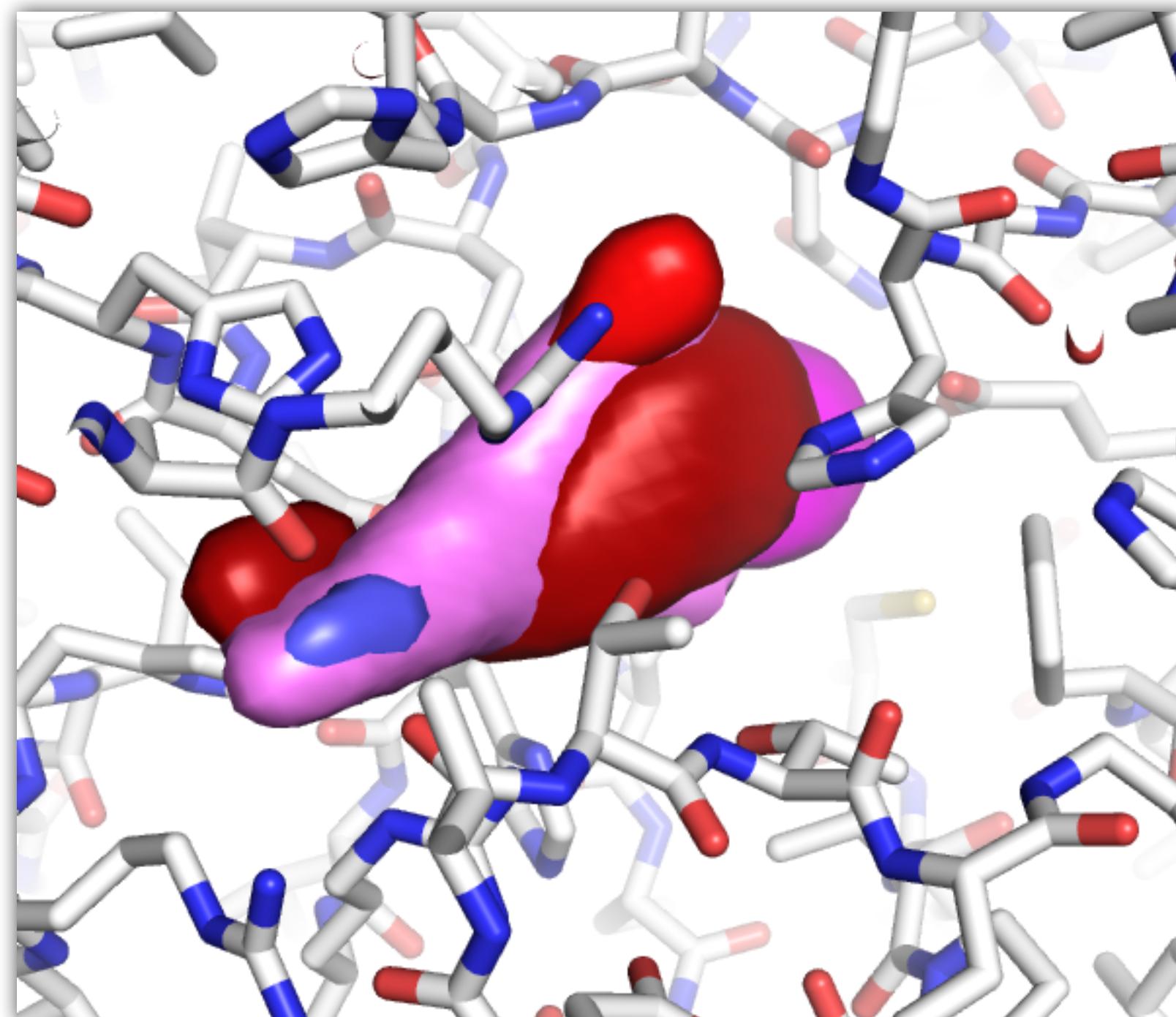
Fit Densities



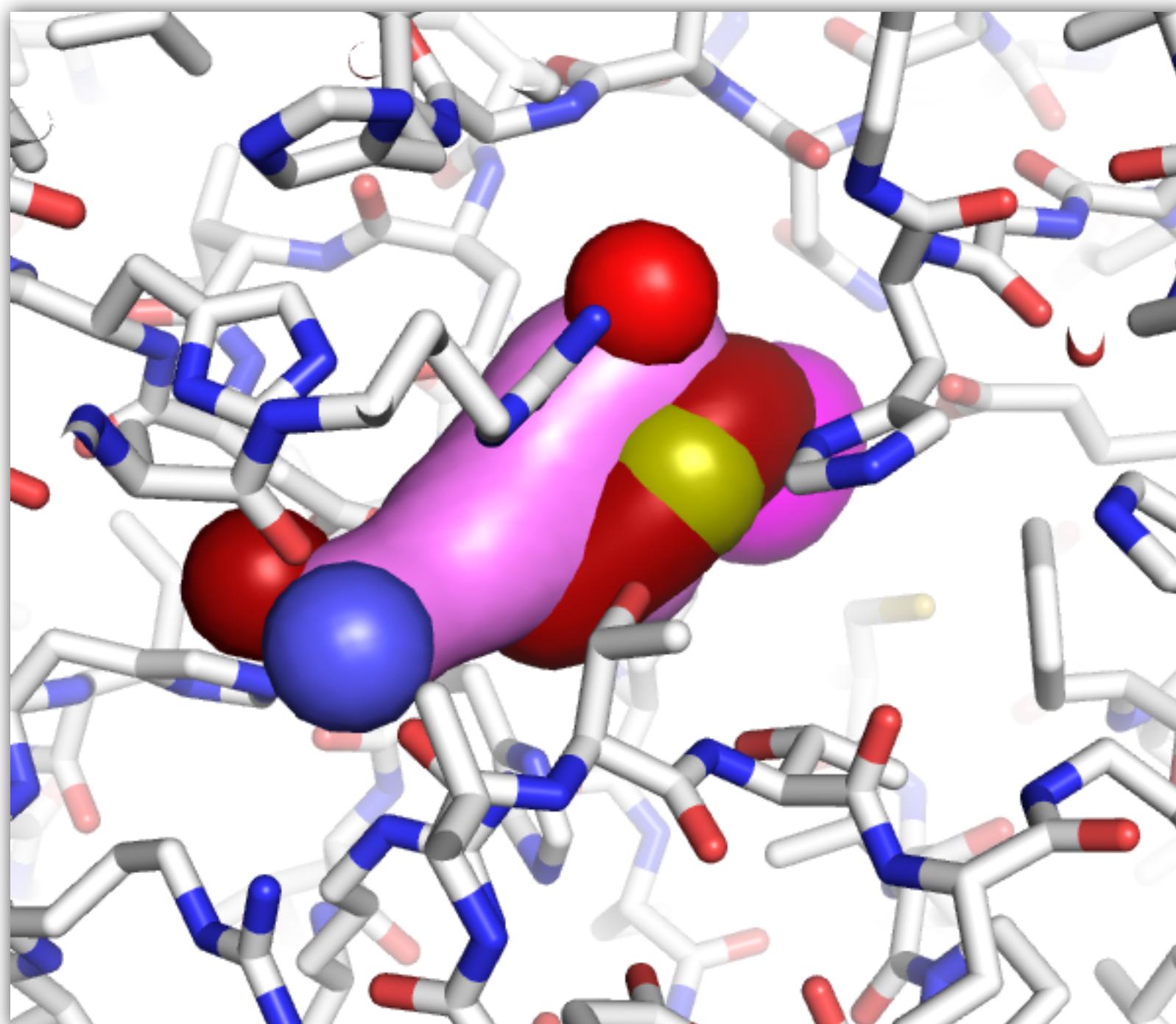
Fit Atoms

Context Encoding with Fully Convolutional Network

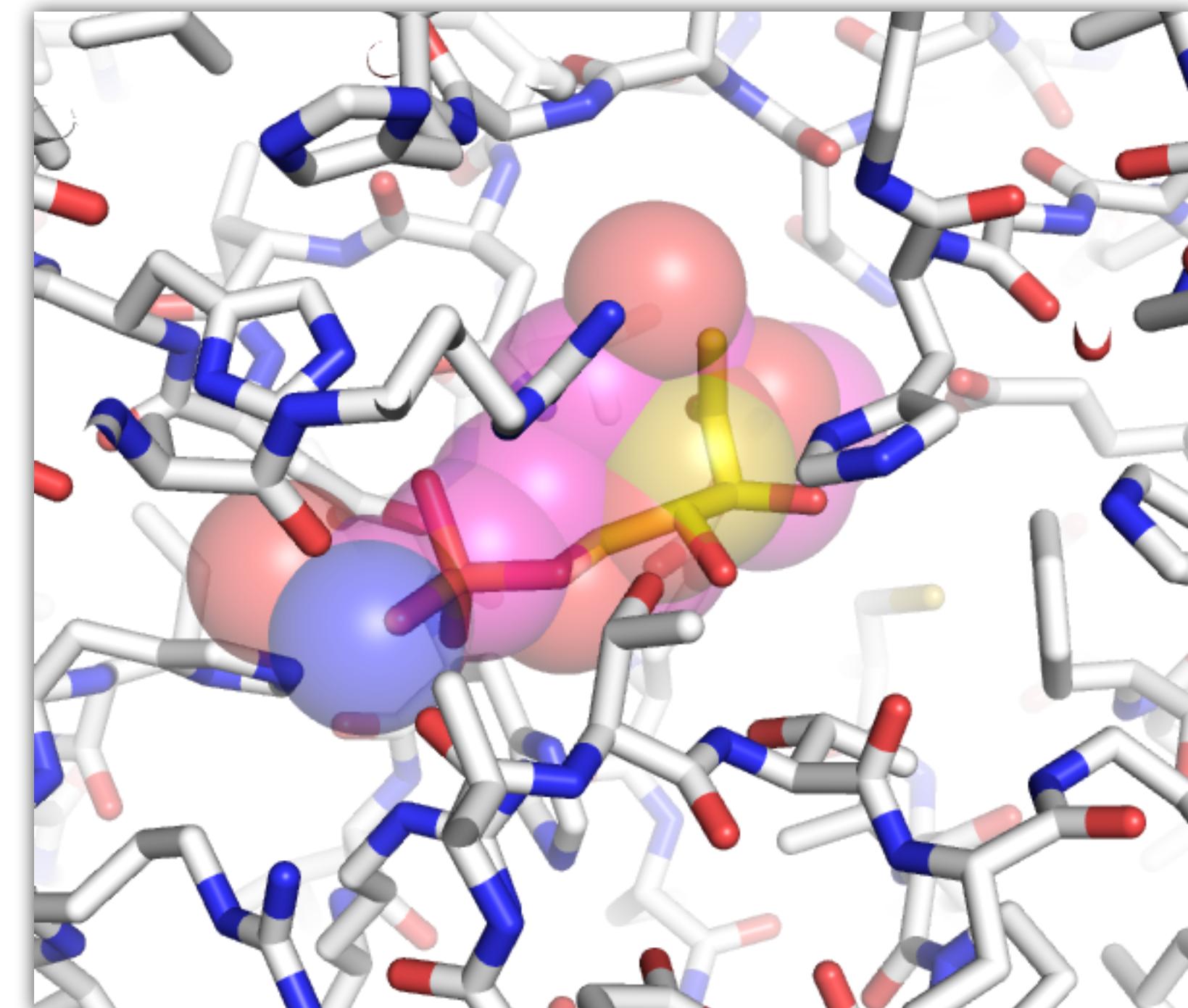
1m5w



Generated



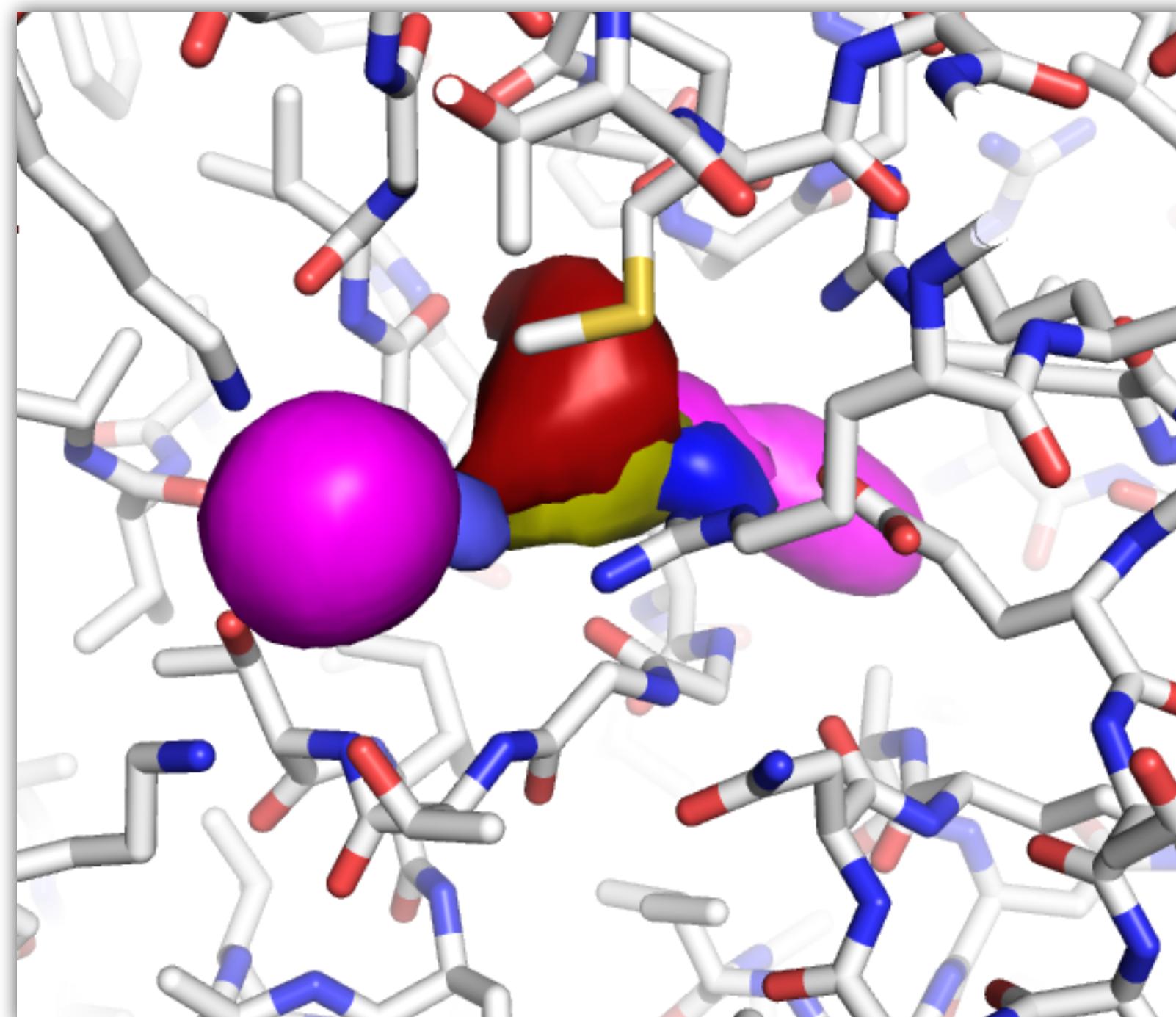
Fit Densities



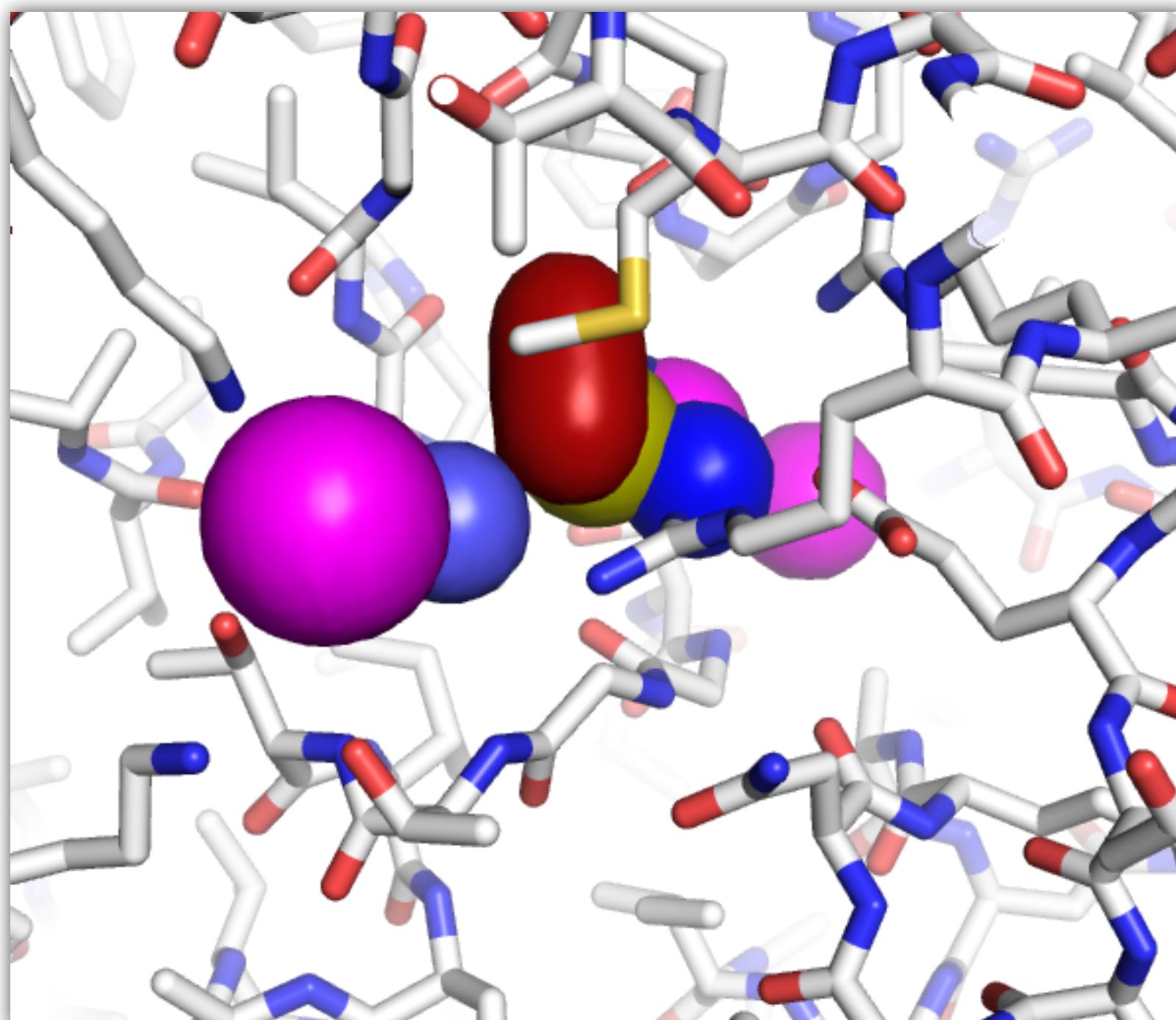
Fit Atoms

Context Encoding with Fully Convolutional Network

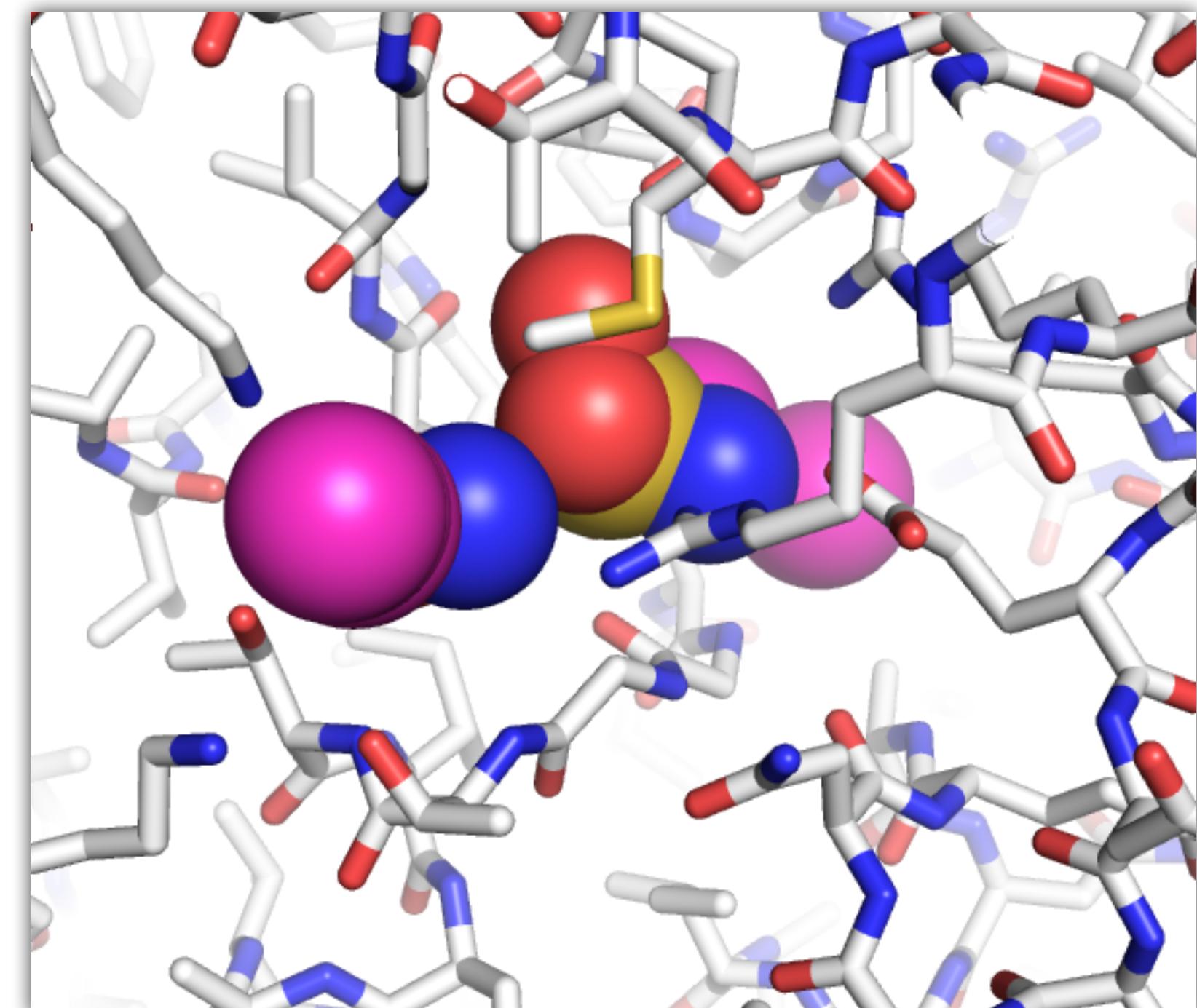
3bxg



Generated



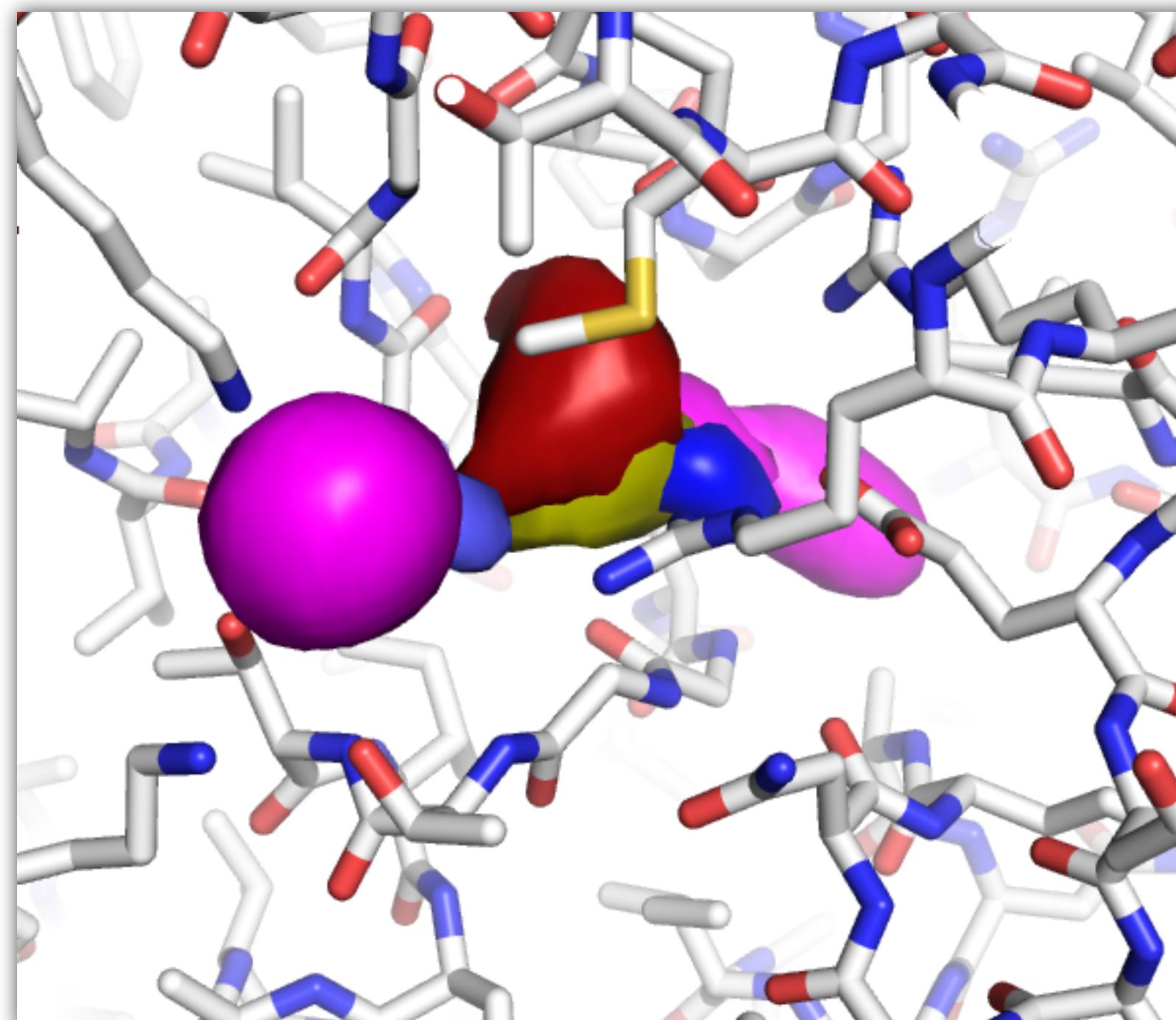
Fit Densities



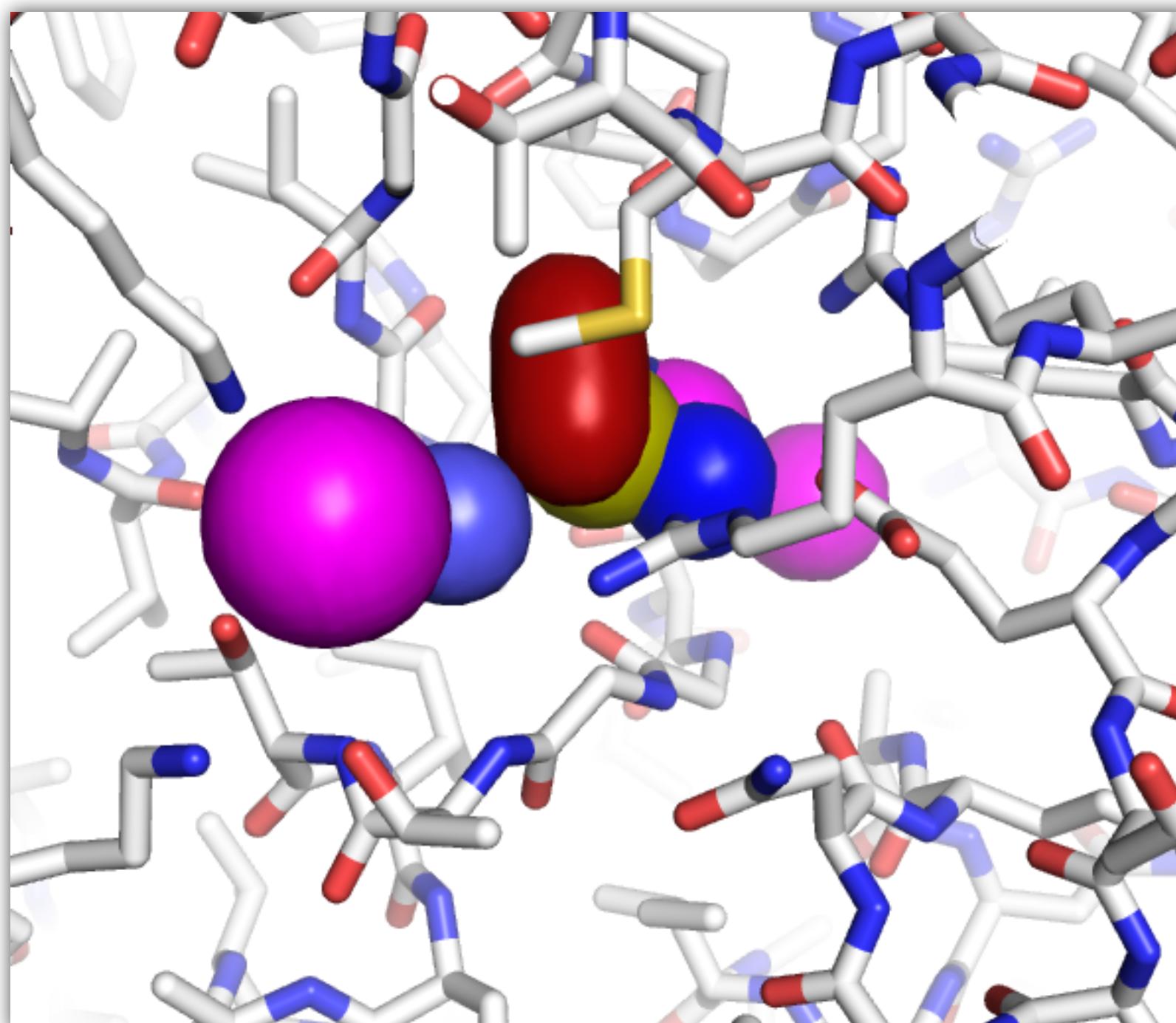
Fit Atoms

Context Encoding with Fully Convolutional Network

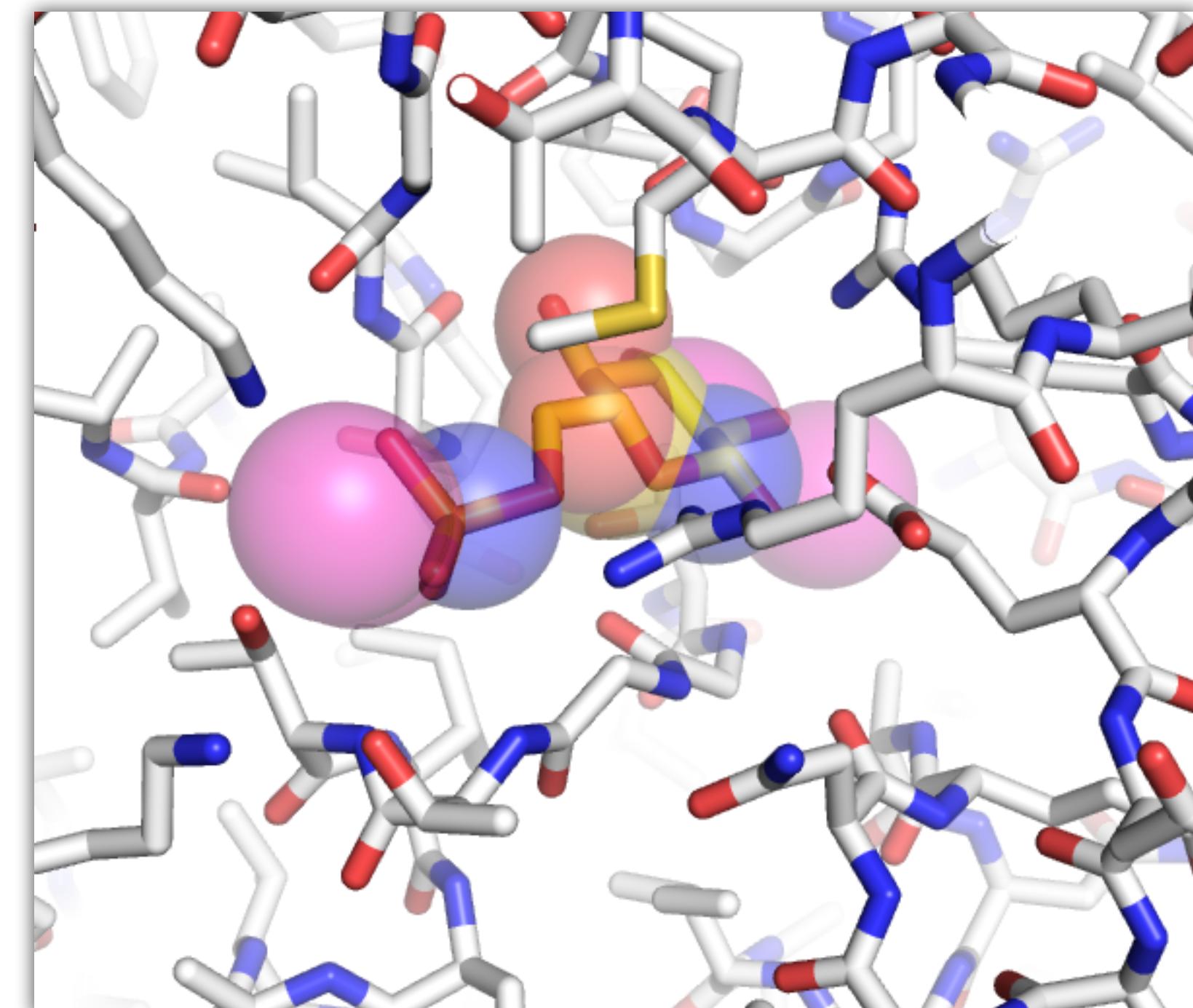
3bxg



Generated



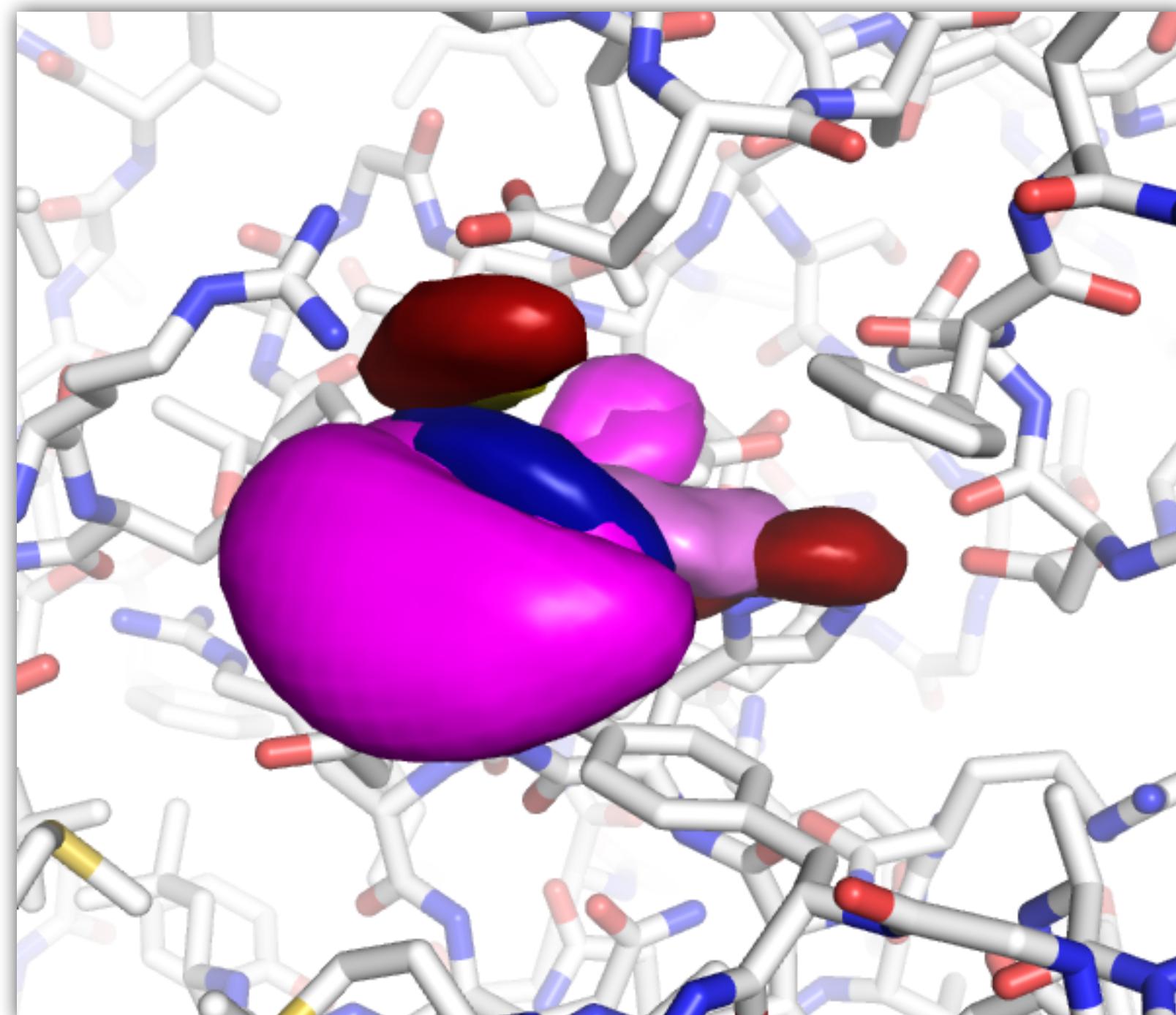
Fit Densities



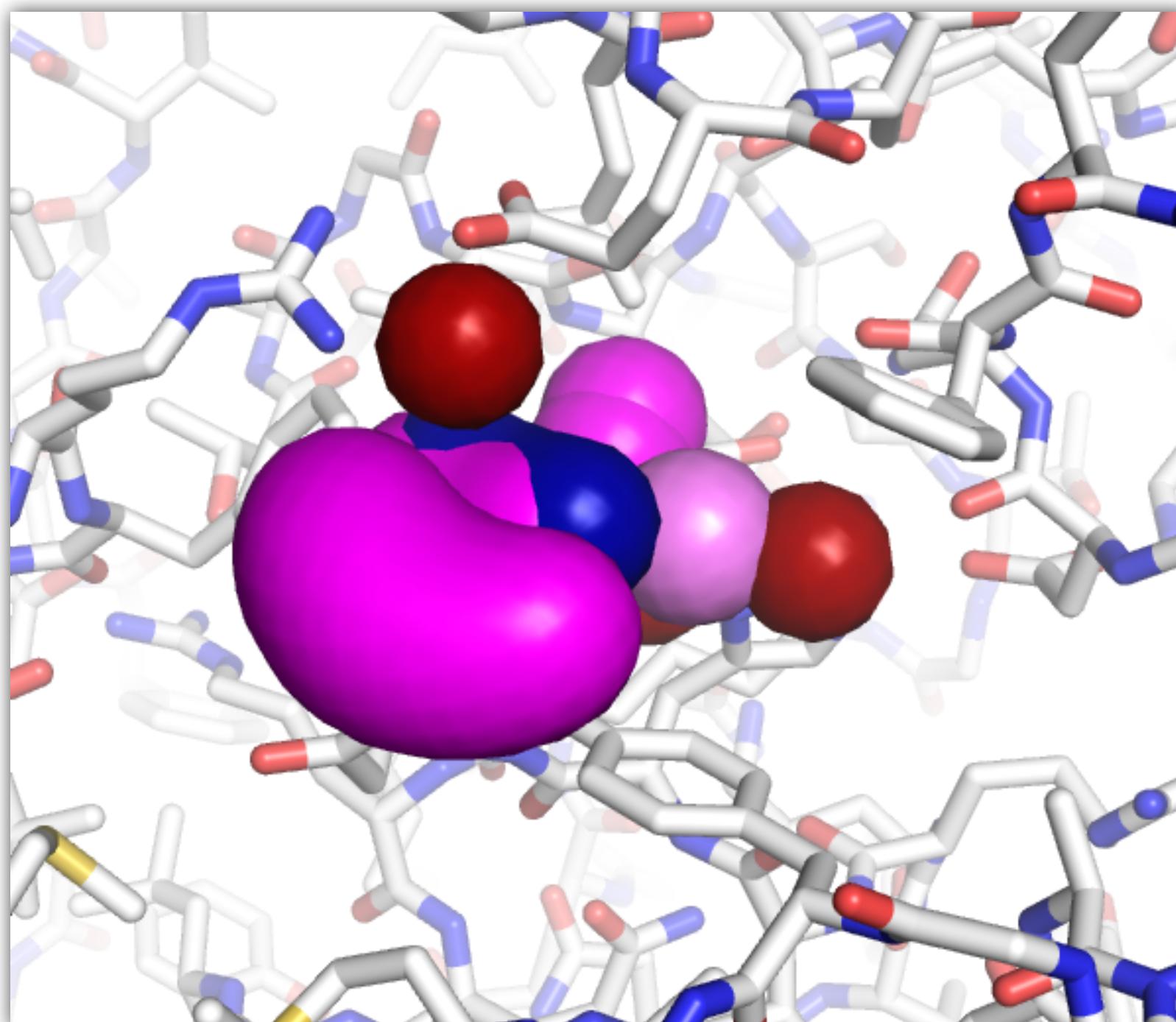
Fit Atoms

Context Encoding with Fully Convolutional Network

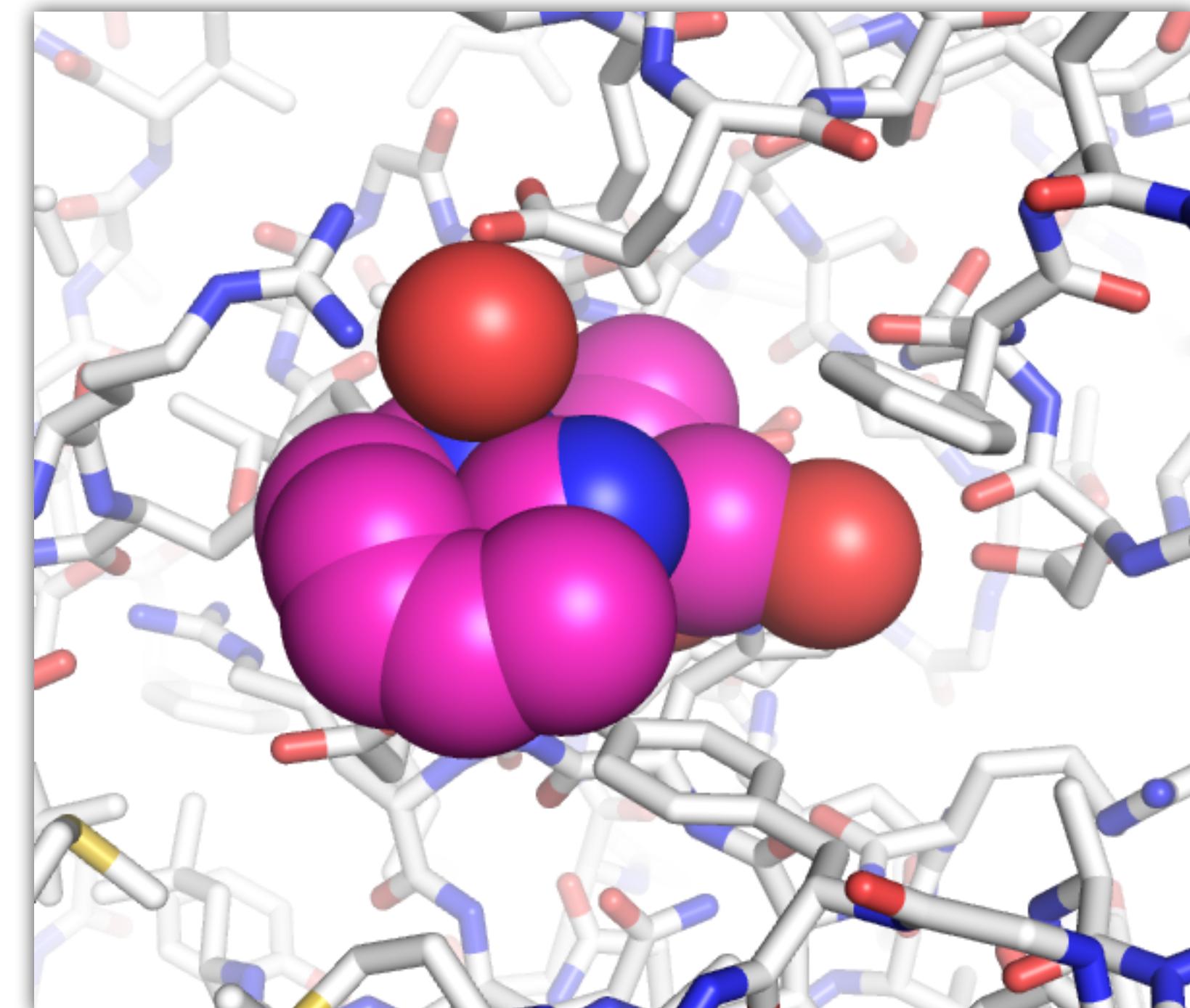
3ebp



Generated



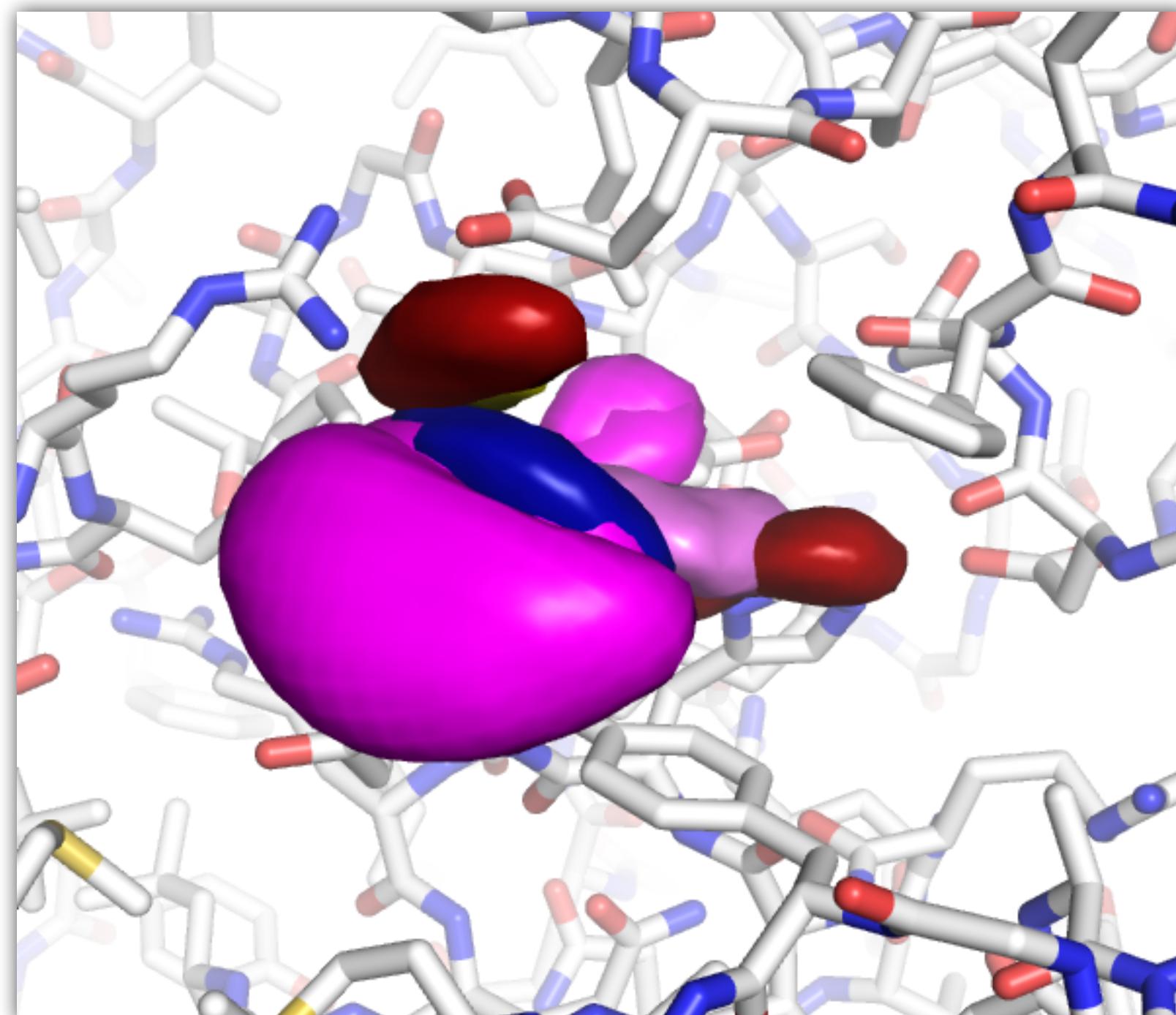
Fit Densities



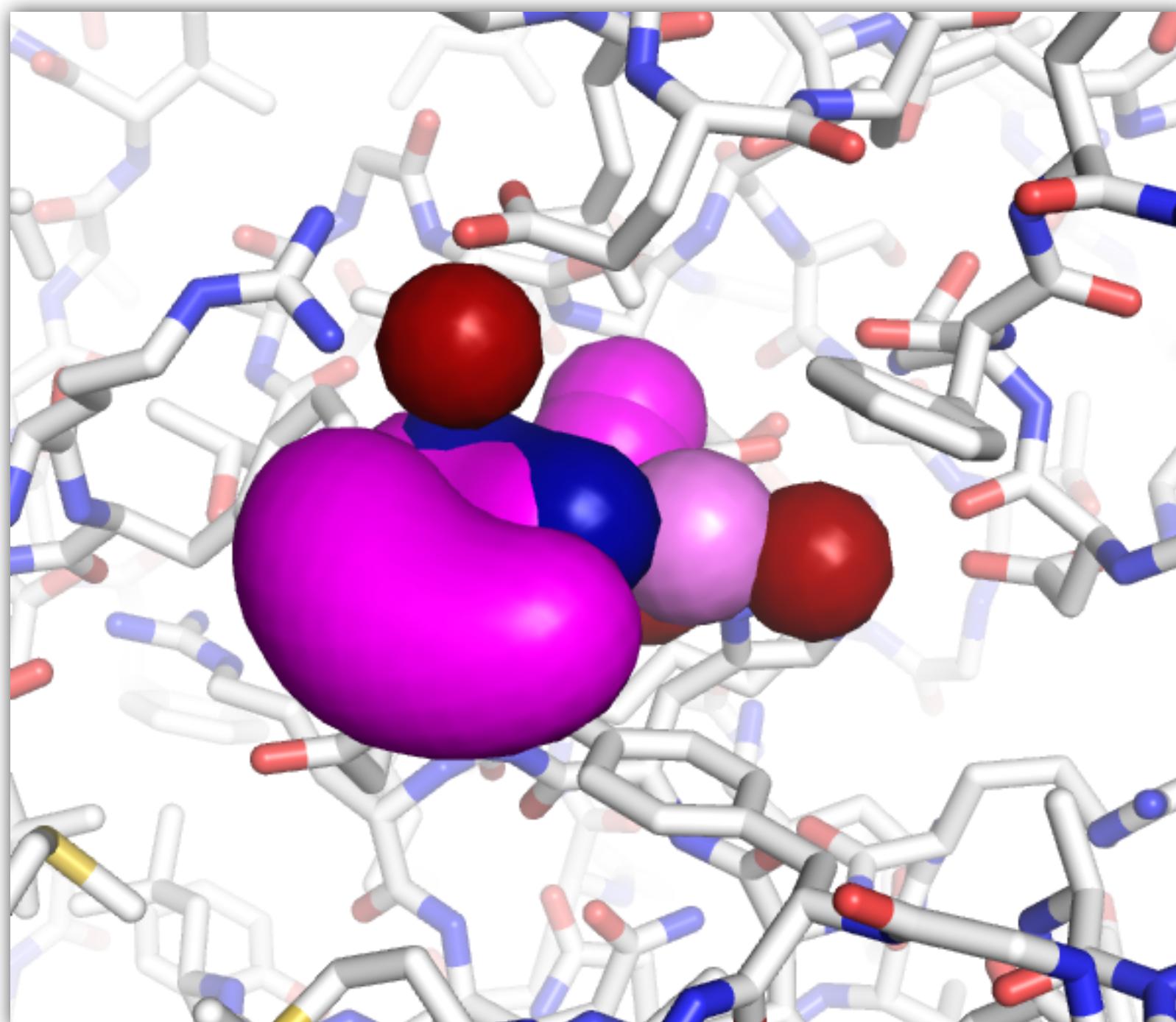
Fit Atoms

Context Encoding with Fully Convolutional Network

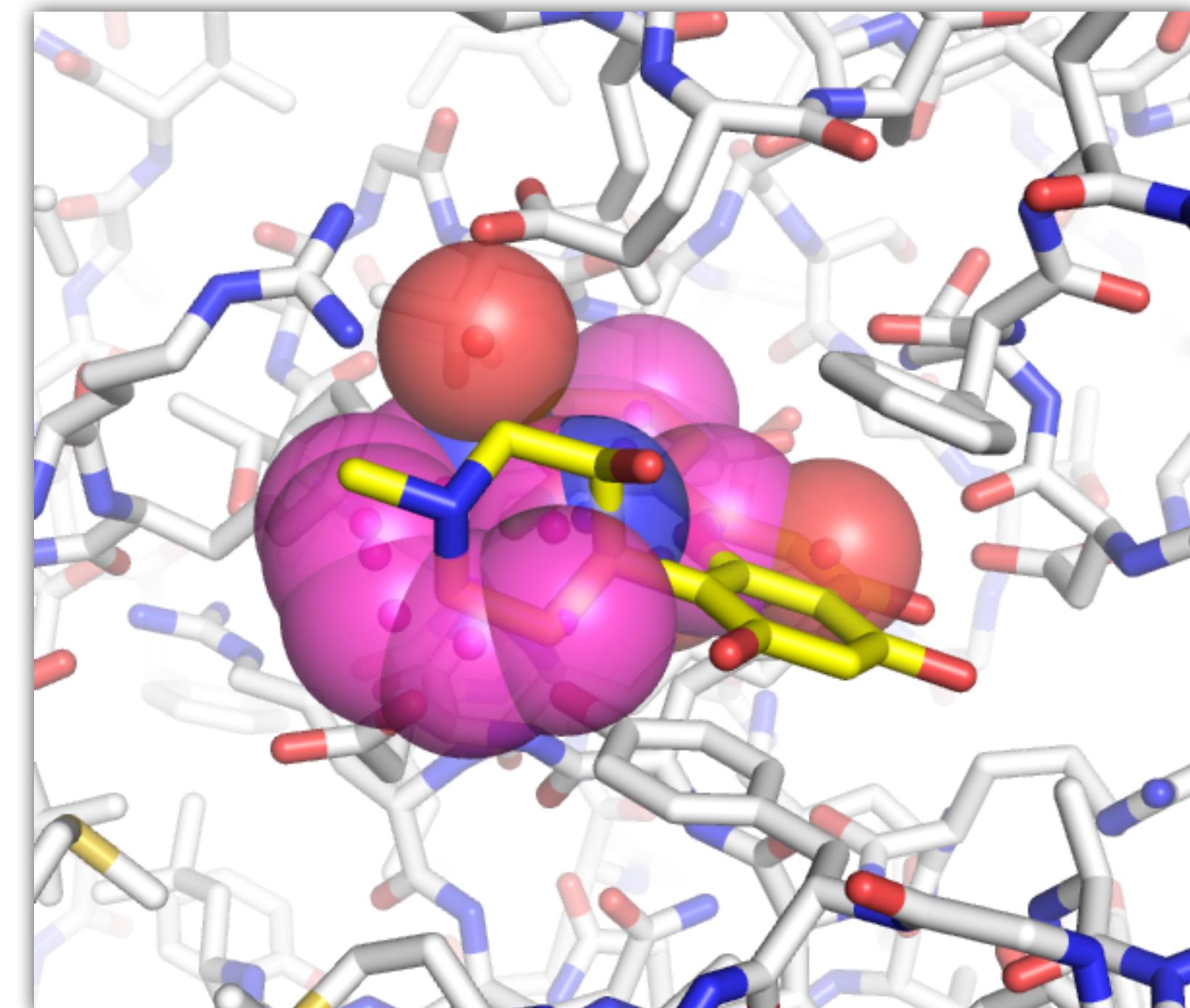
3ebp



Generated



Fit Densities



Fit Atoms