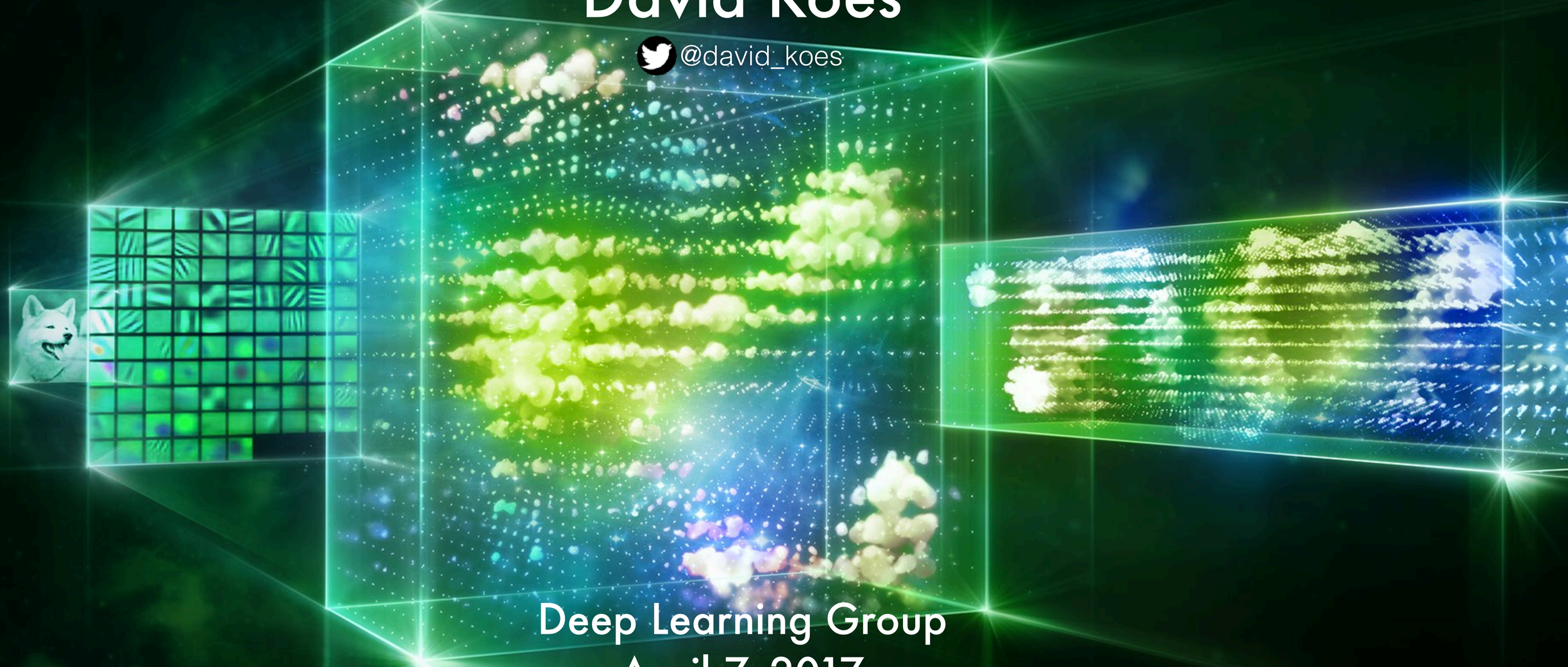
Visualizing Deep Neural Networks

David Koes

@david_koes



Deep Learning Group

April 7, 2017

Evaluating the visualization of what a Deep Neural Network has learned

Wojciech Samek† *Member, IEEE,* Alexander Binder†, Grégoire Montavon, Sebastian Bach, and Klaus-Robert Müller, *Member, IEEE,*

https://arxiv.org/pdf/1509.06321.pdf

Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler ZEILER@CS.NYU.EDU
Dept. of Computer Science, Courant Institute, New York University
Rob Fergus FERGUS@CS.NYU.EDU
Dept. of Computer Science, Courant Institute, New York University

https://arxiv.org/abs/1311.2901

On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, Wojciech Samek

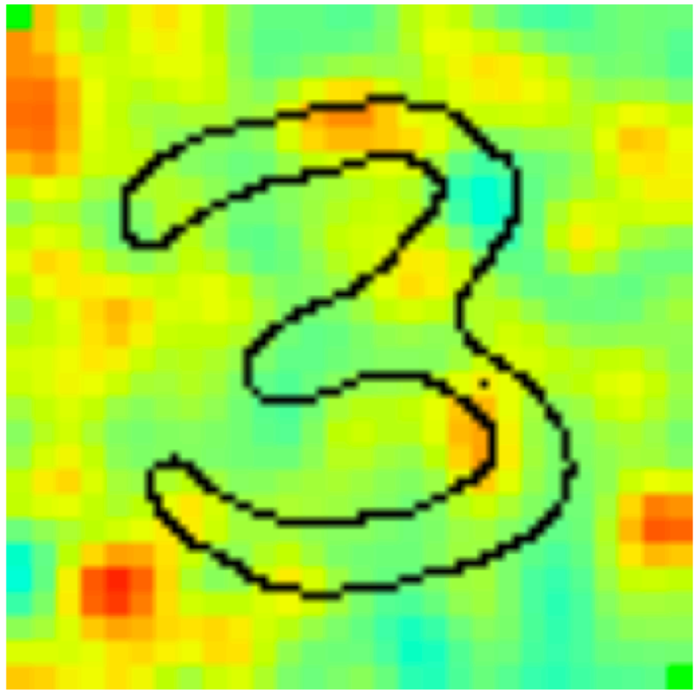
Published: July 10, 2015 • http://dx.doi.org/10.1371/journal.pone.0130140

http://dx.doi.org/10.1371/journal.pone.0130140

Heatmaps

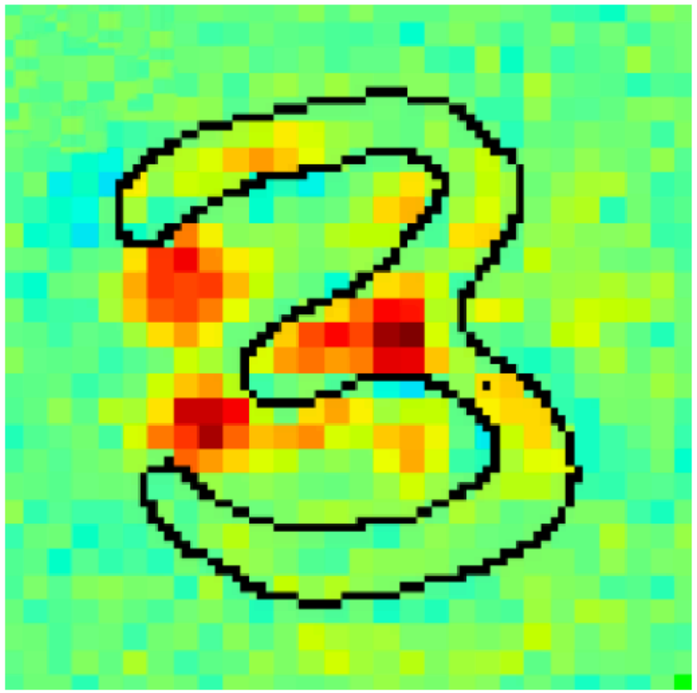
Map the neural network decision onto the input

Random

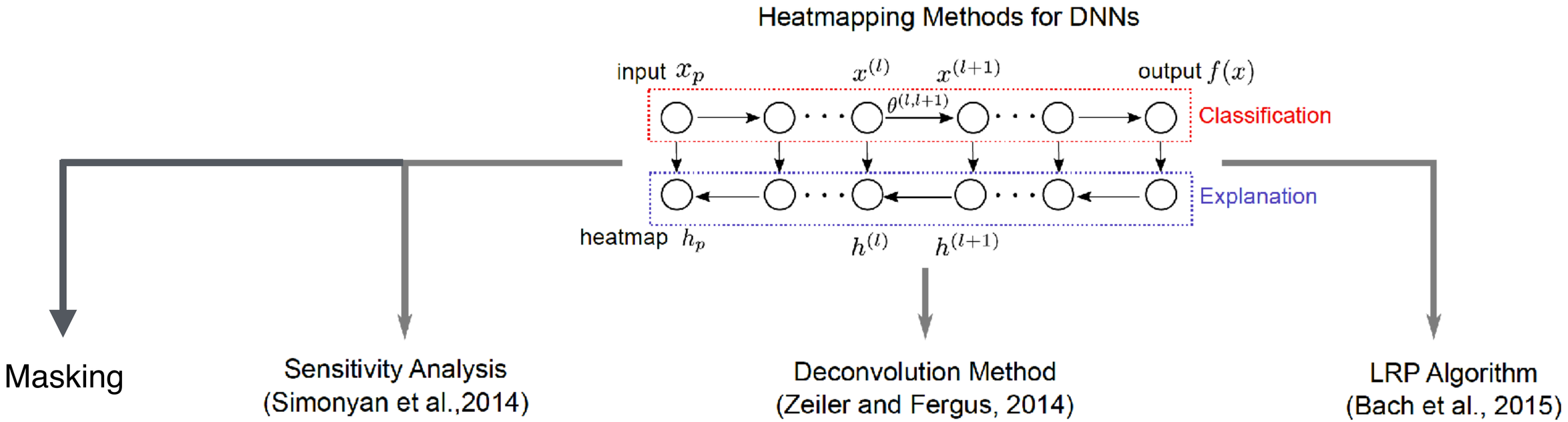


Segmentation

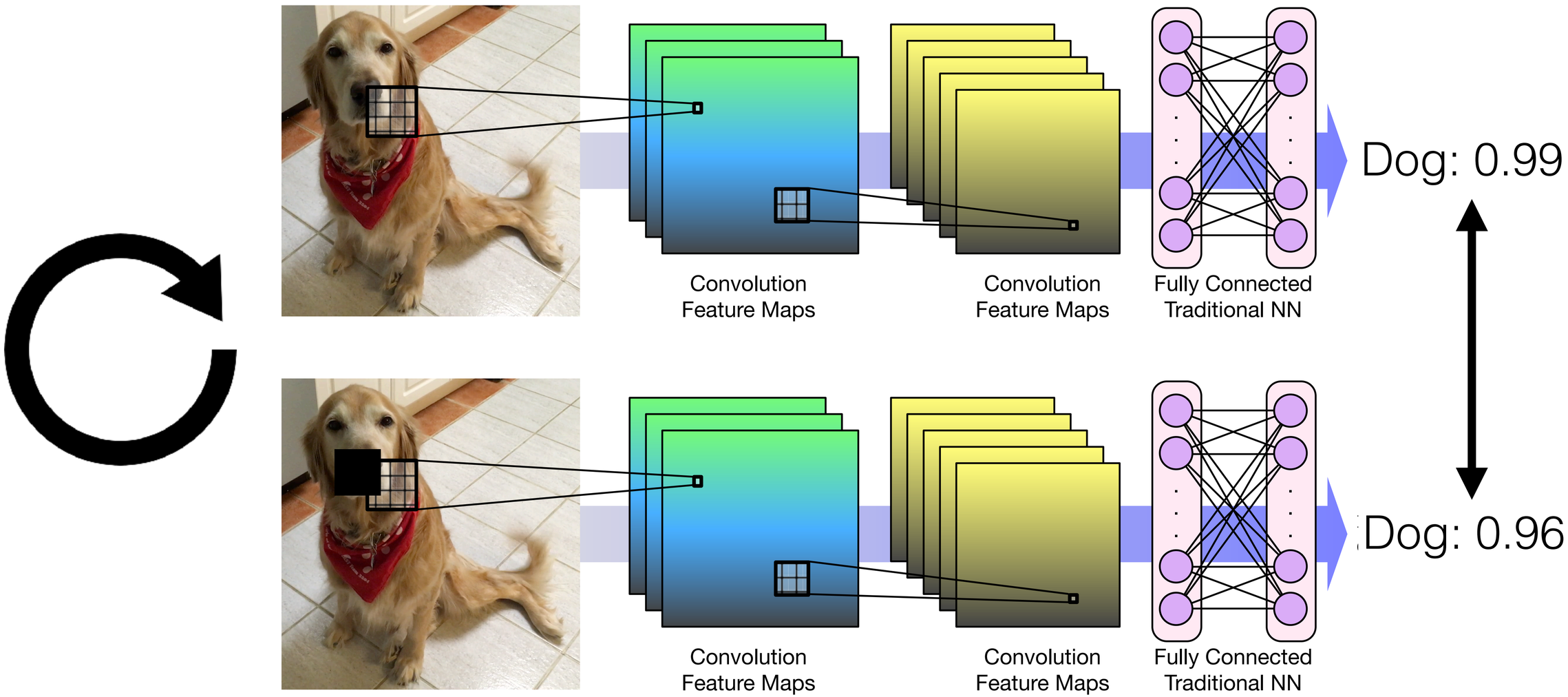
Relevance



Heatmapping DNNs



Masking


Convolution
Feature Maps
Convolution
Feature Maps
Fully Connected
Traditional NN
Dog: 0.99
Convolution
Feature Maps
Convolution
Feature Maps
Fully Connected
Traditional NN
Dog: 0.96

Sensitivity Analysis

Propagation

Chain rule for computing derivatives:

$$\nabla^{(l)} = \frac{\partial x^{(l+1)}}{\partial x^{(l)}} \nabla^{(l+1)}$$

Heatmap

local sensitivity (what makes a bird more/less a bird).

Relation to $f(x)$

$$h_p = \left|\left| \frac{\partial}{\partial x_p} f(x) \right|\right|_\infty$$

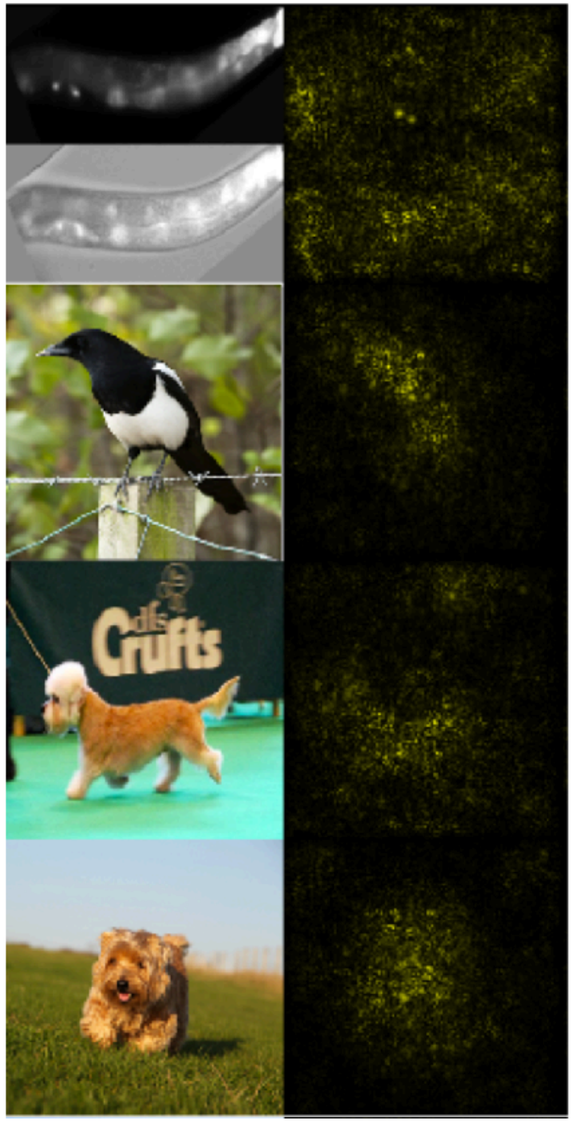
Applicable

any network with continuous locally differentiable neurons.

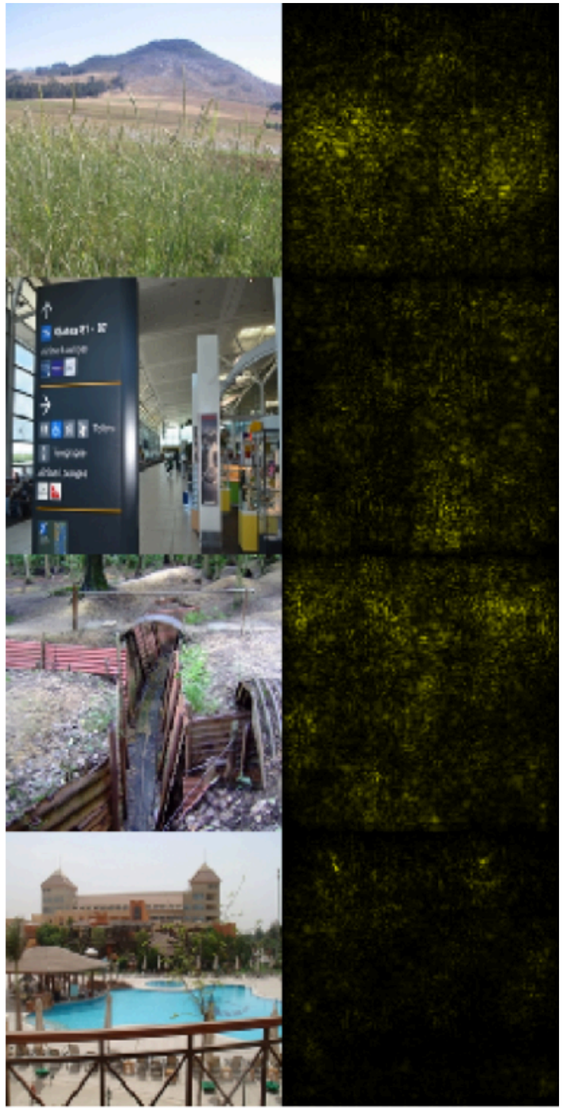
Drawbacks

(i) heatmap does not *fully* explain the image classification.

ILSVRC2012

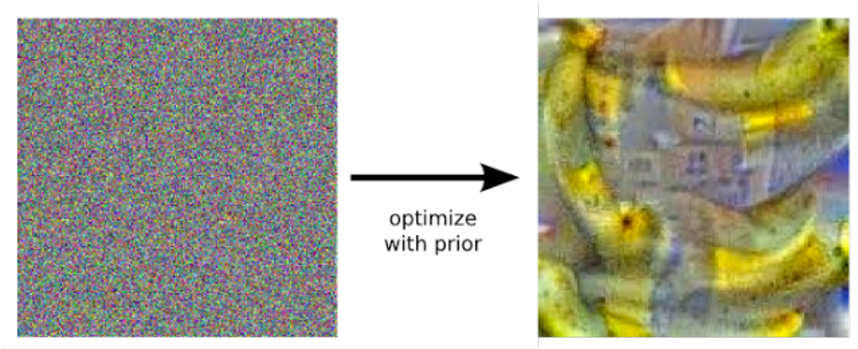


MIT Places



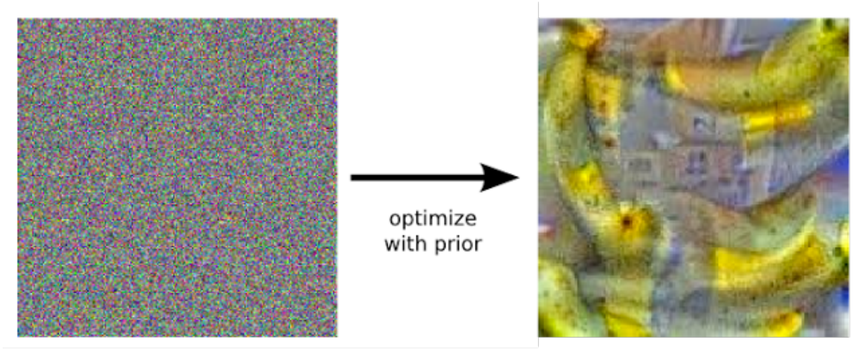
Sensitivity Analysis - Deep Dreams

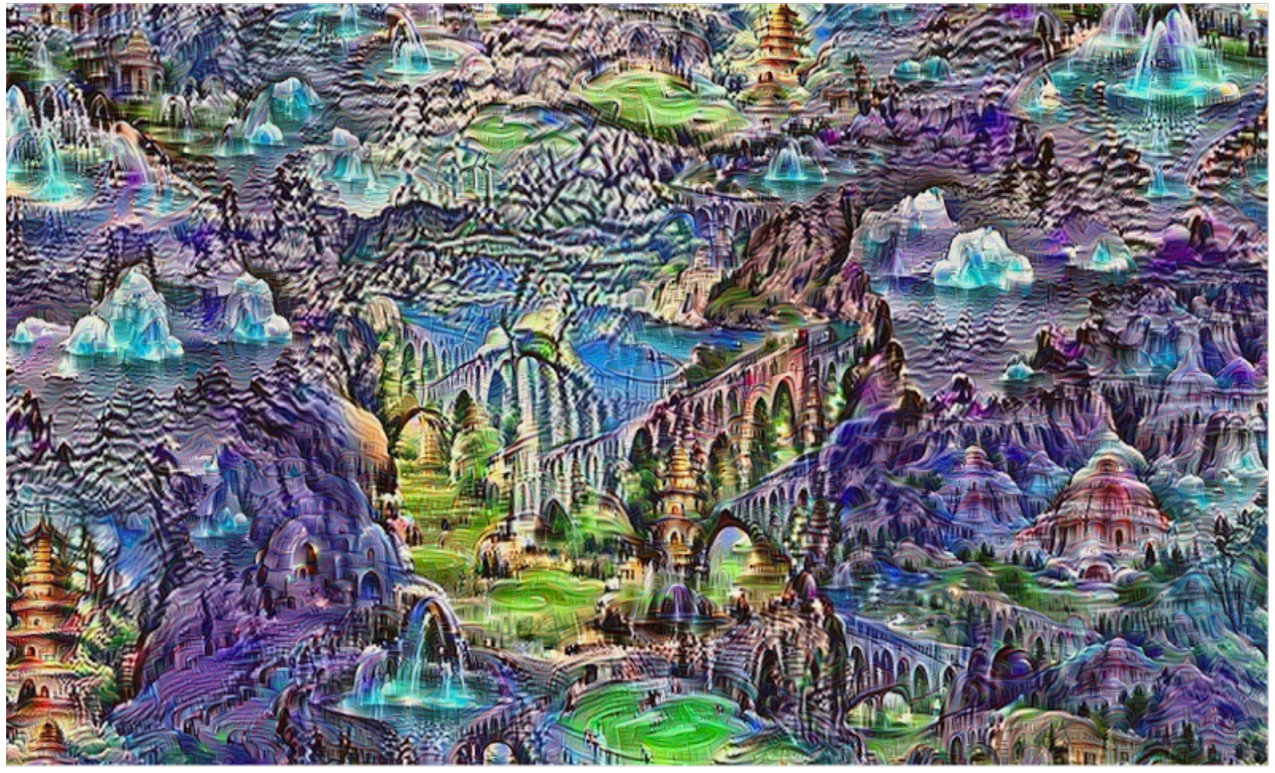
https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

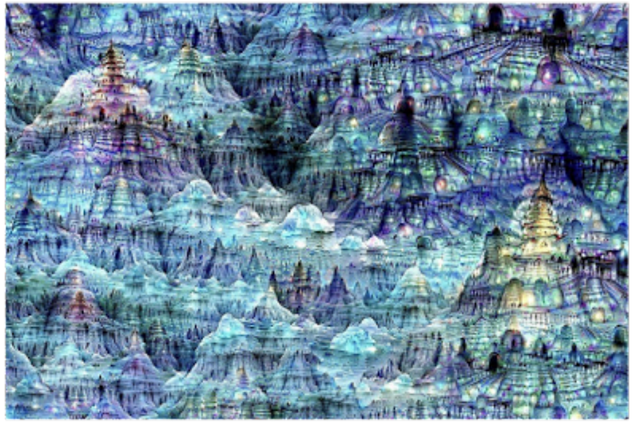
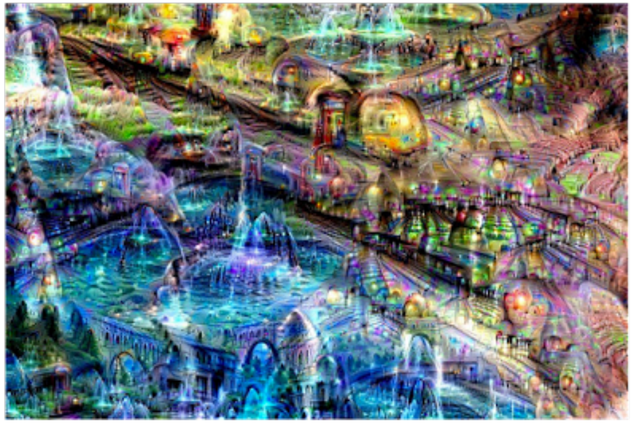


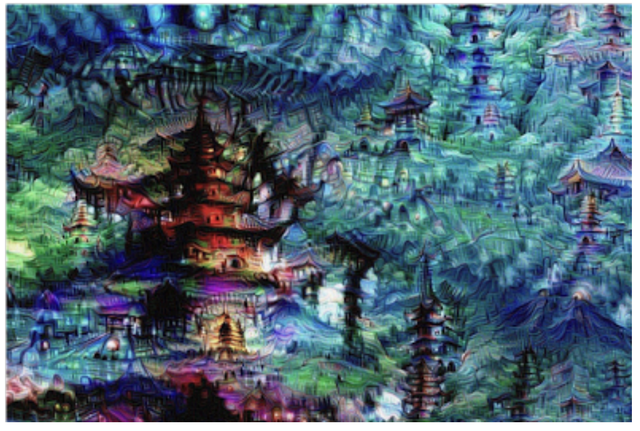
Sensitivity Analysis - Deep Dreams

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





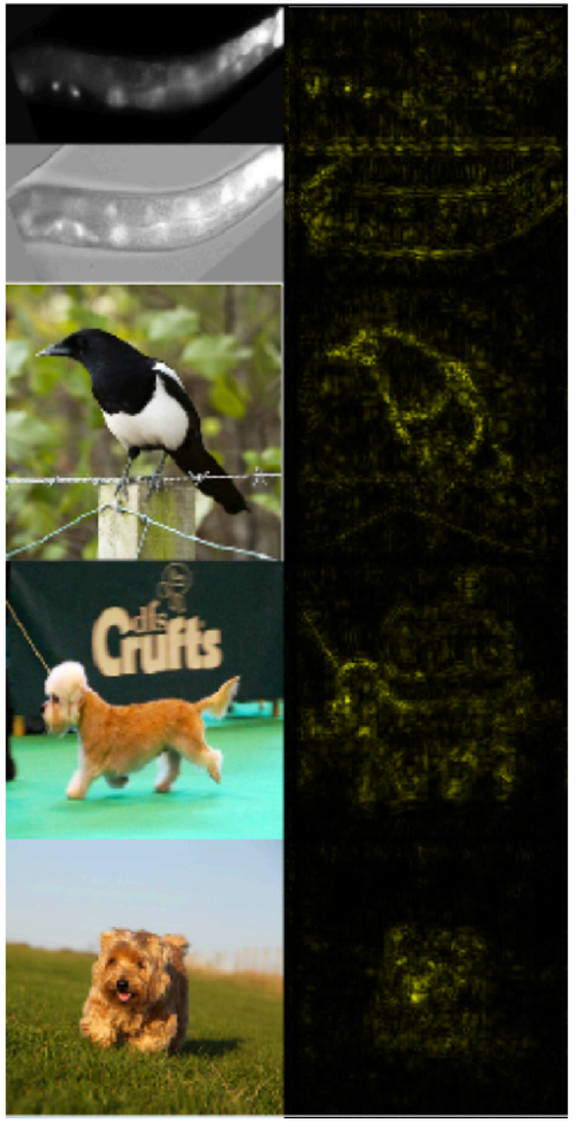





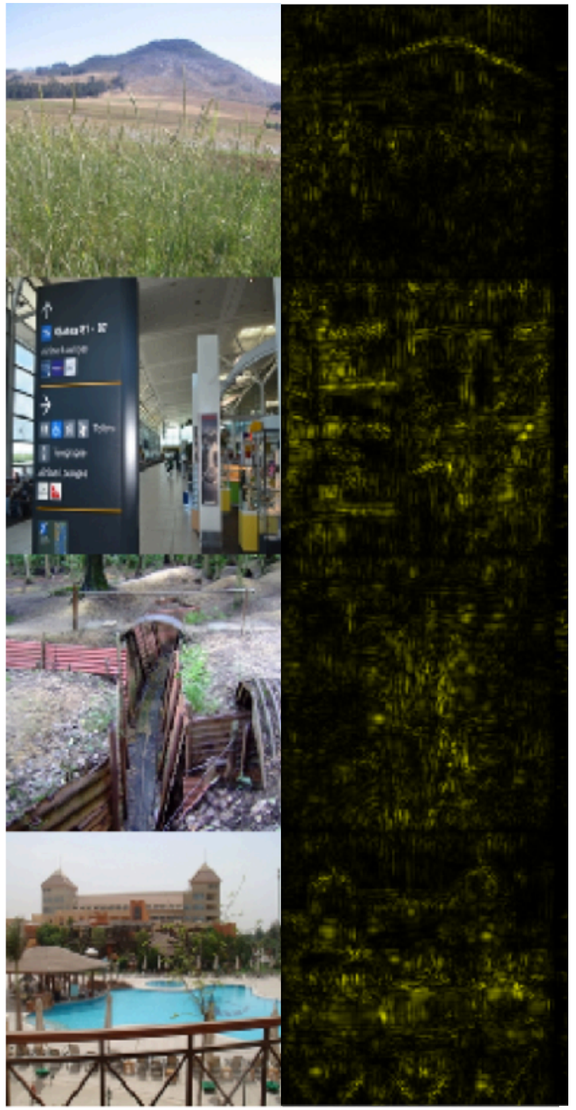
Deconvolution

Propagation	Backward mapping function: $h^{(l)} = m_{\text{dec}}(h^{(l+1)}; \theta^{(l,l+1)})$
Heatmap	matching input pattern for the classifed object in the image.
Relation to $f(x)$	not specified
Applicable	convolutional network with max-pooling and rectified linear units.
Drawbacks	(i) no direct correspondence between heatmap scores and contribution of pixels to the classification. (ii) image-specific information only from max-pooling layers.

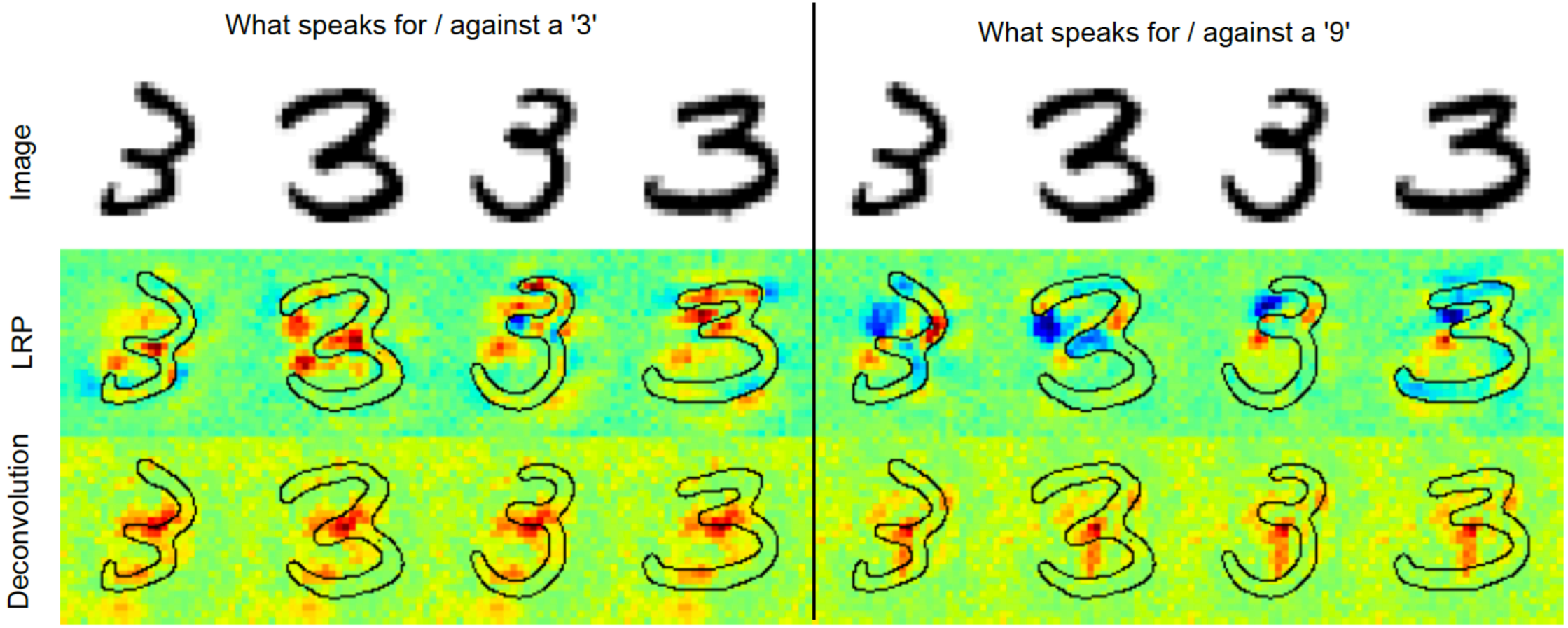
ILSVRC2012



MIT Places



Deconvolution



Layer-wise Relevance Propagation

Propagation

Backward mapping function
+ conservation principles:

$$h^{(l)} = m_{\text{lrp}}\left(h^{(l+1)}; x^{(l)}, \theta^{(l,l+1)}\right)$$

$$\sum_i h_i^{(l)} = \sum_j h_j^{(l+1)}$$

Heatmap

explanatory input pattern that indicates evidence for and against a bird.

Relation to $f(x)$

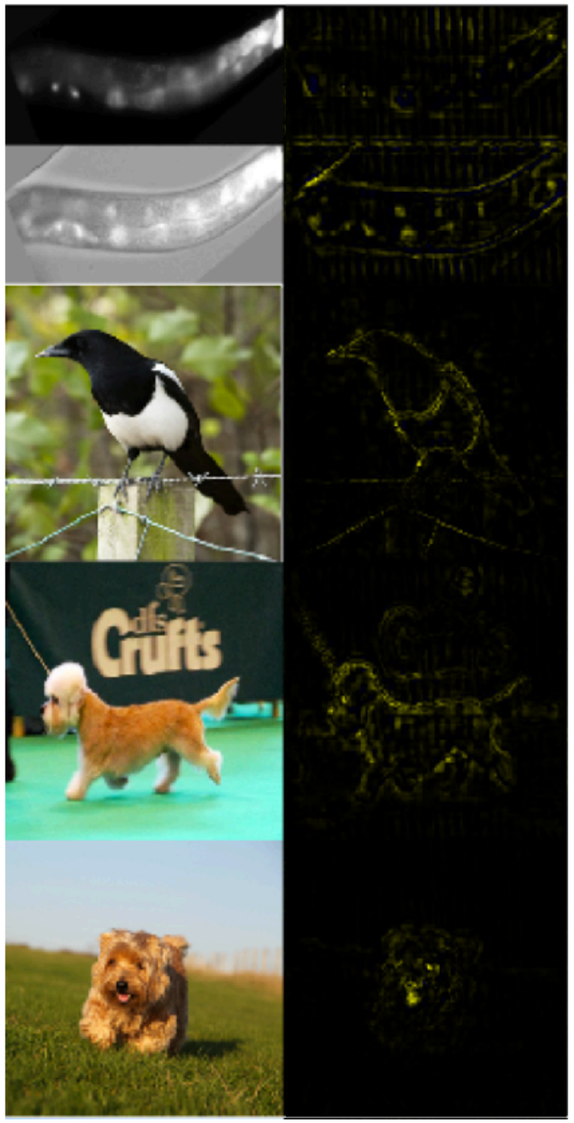
$$f(x) = \sum_p h_p$$

Applicable

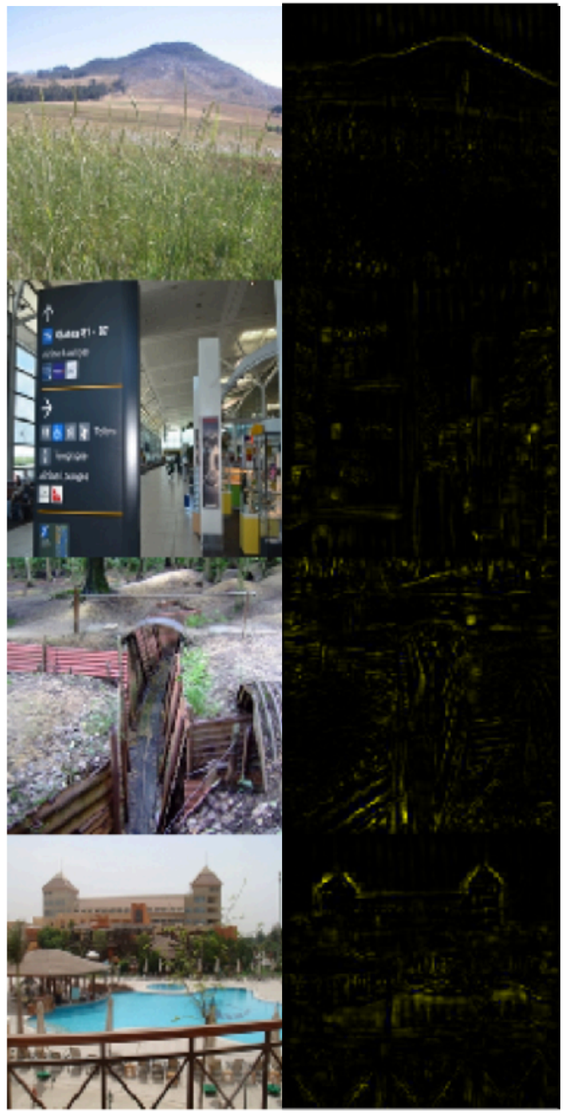
any network with monotonous activations
(even non-continuous units)

Drawbacks

ILSVRC2012



MIT Places



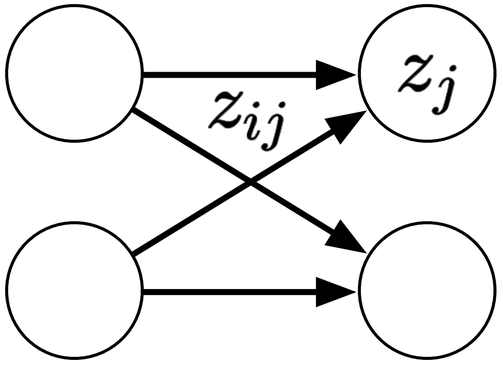
LRP

$$f(x) = \ldots = \sum_{d \in l+1} R_d^{(l+1)} = \sum_{d \in l} R_d^{(l)} = \ldots = \sum_{d} R_d^{(1)}$$

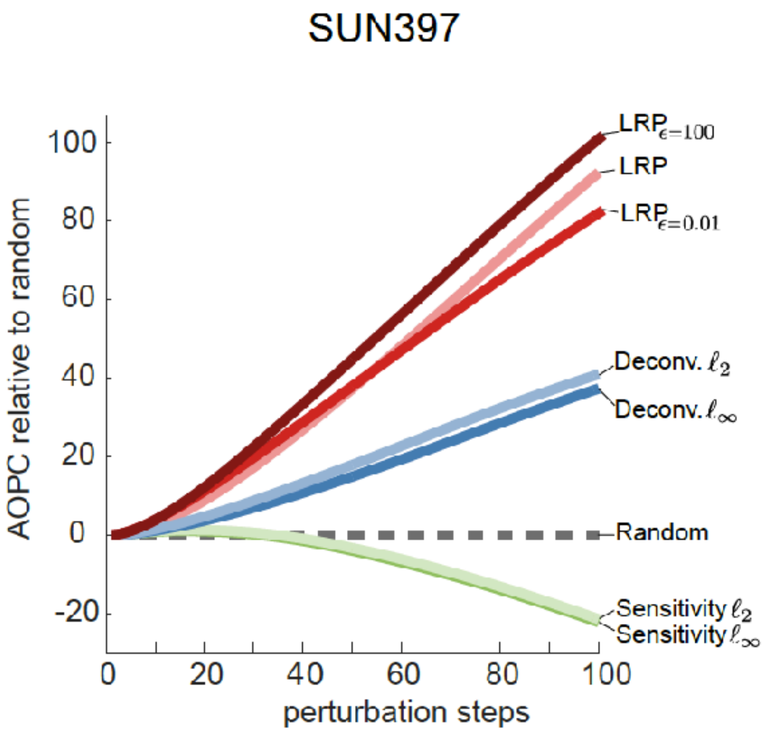
$$R_i^{(l)} = \sum_j \frac{z_{ij}}{\sum_{i'} z_{i'j} + \epsilon \, \mathrm{sign}(\sum_{i'} z_{i'j})} R_j^{(l+1)} \quad \textit{or} \quad R_i^{(l)} = \sum_j \left(\alpha \cdot \frac{z_{ij}^+}{\sum_{i'} z_{i'j}^+} + \beta \cdot \frac{z_{ij}^-}{\sum_{i'} z_{i'j}^-} \right) R_j^{(l+1)}$$

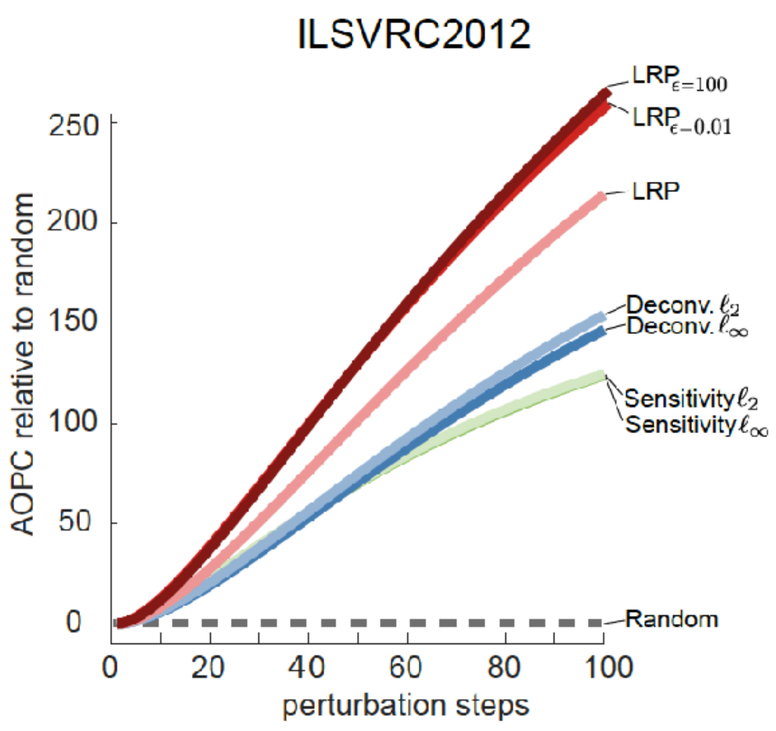
$$a_j^{(l+1)} = \sigma\left(\sum_i z_{ij} + b_j^{(l+1)} \right)$$

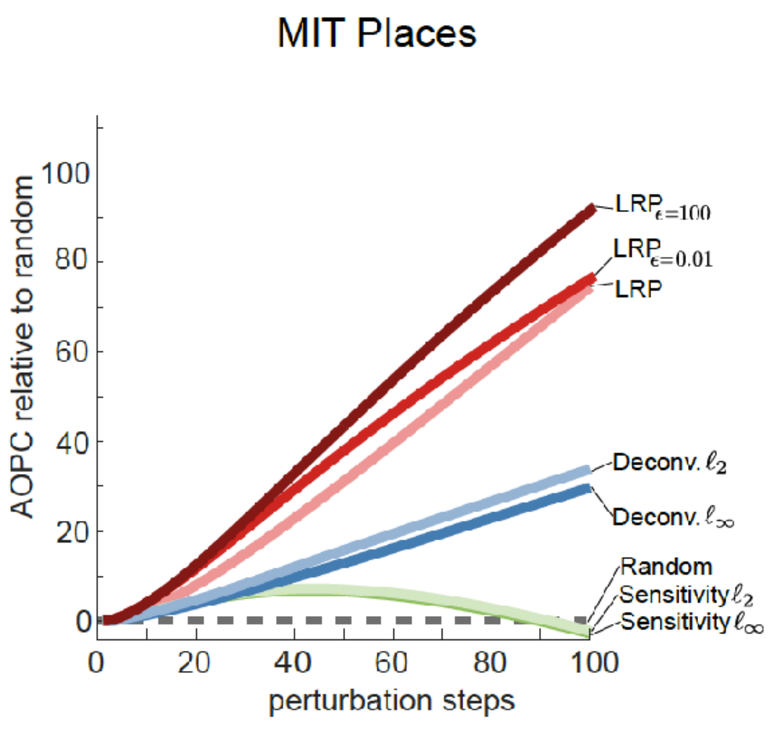
$$\text{with} \quad z_{ij} = a_i^{(l)} w_{ij}^{(l,l+1)}$$



Evaluating

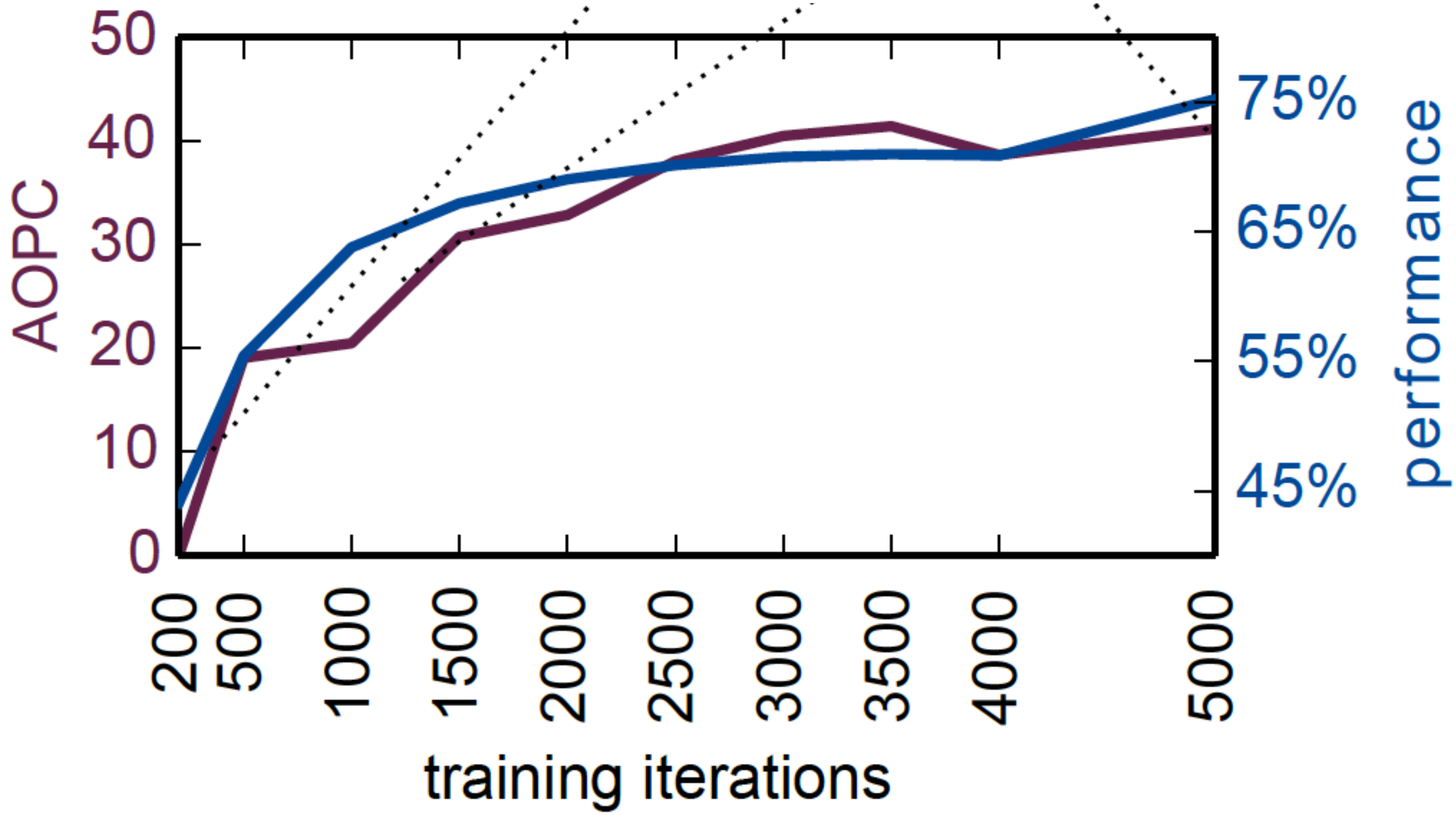




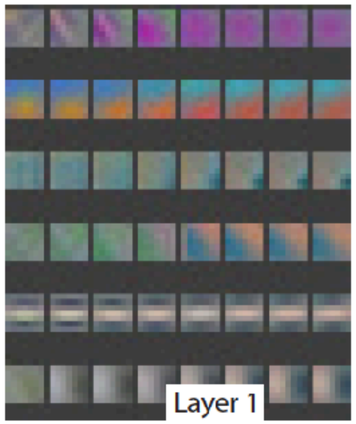


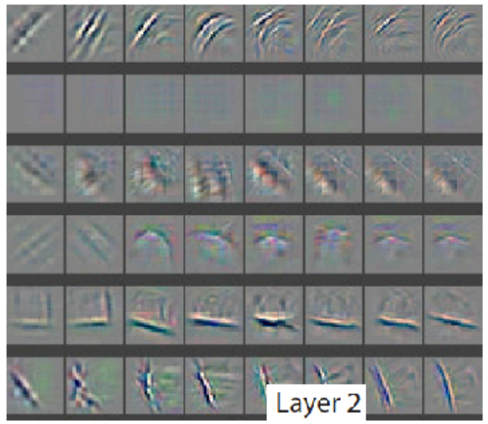
$$\text{AOPC} = \frac{1}{L+1}\left\langle \sum_{k=0}^{L} f(\boldsymbol{x}^{(0)}_{\text{MoRF}}) - f(\boldsymbol{x}^{(k)}_{\text{MoRF}}) \right\rangle_{p(\boldsymbol{x})}$$

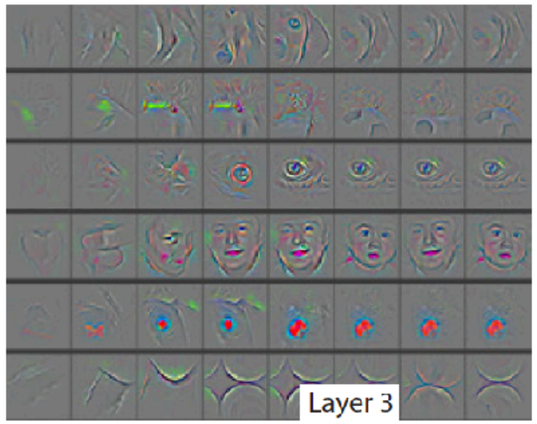
An interesting result?

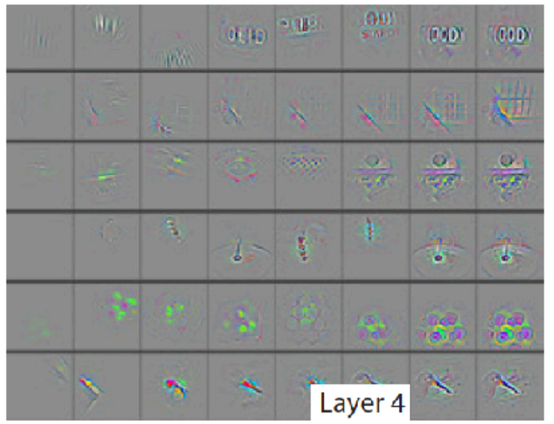


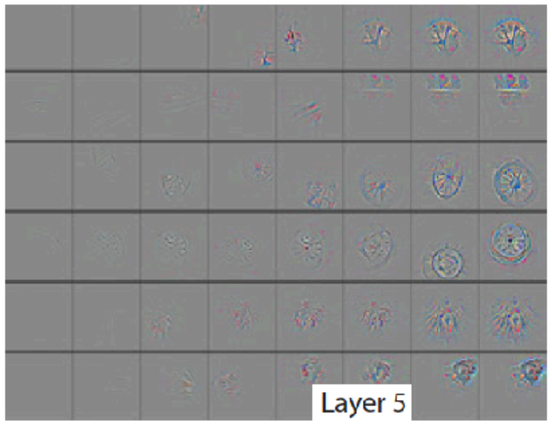
Beyond the output layer

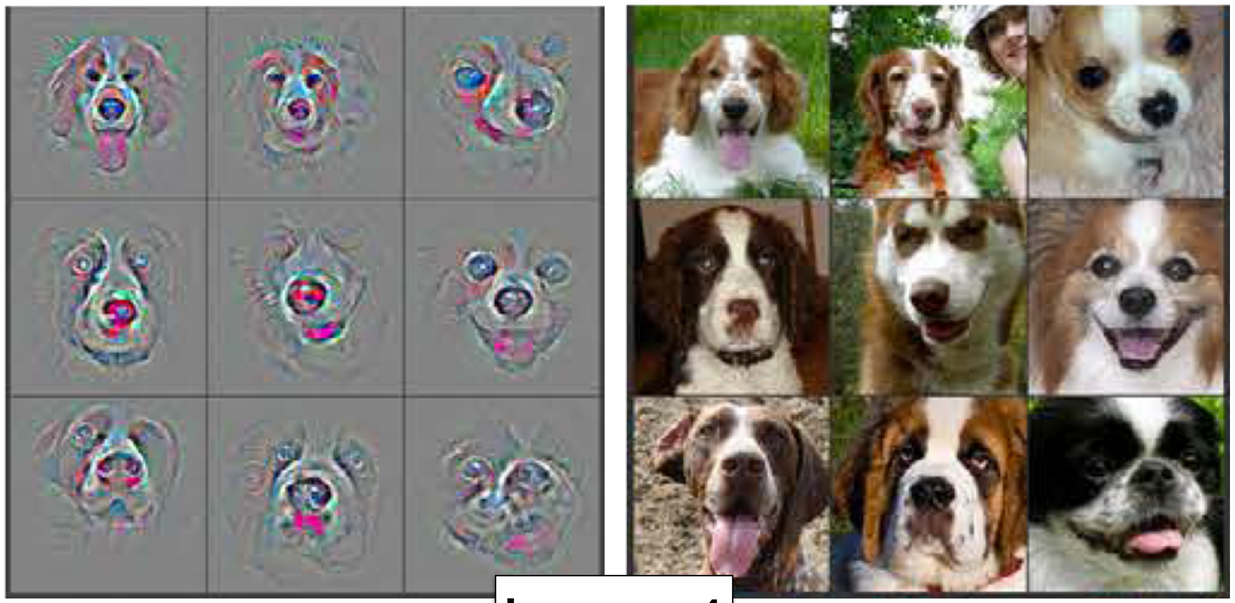










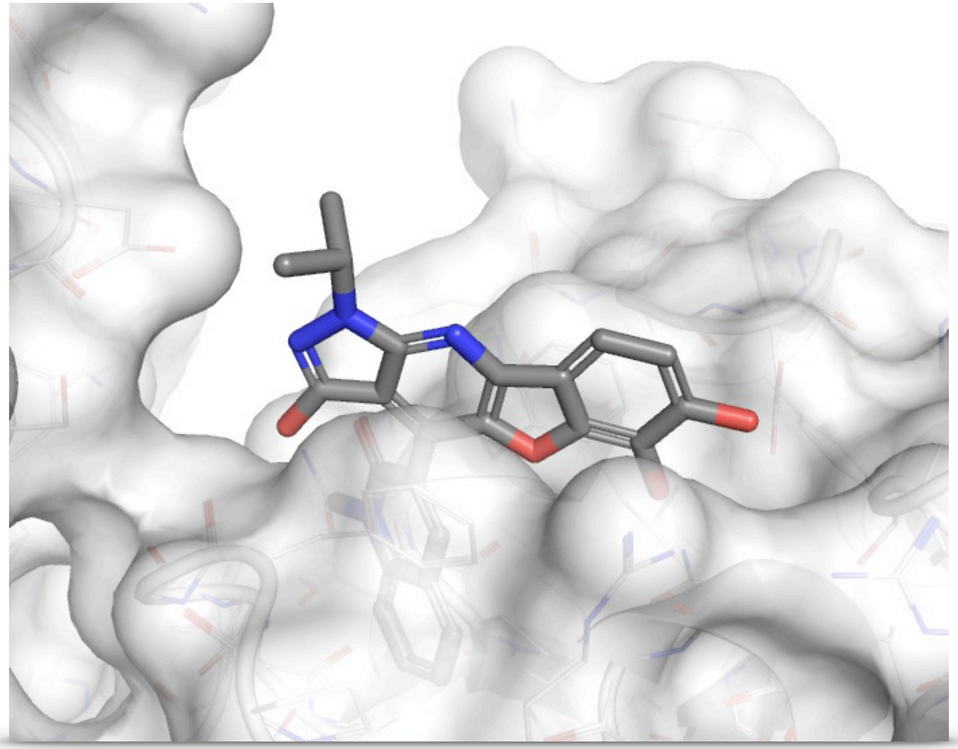


Layer 4

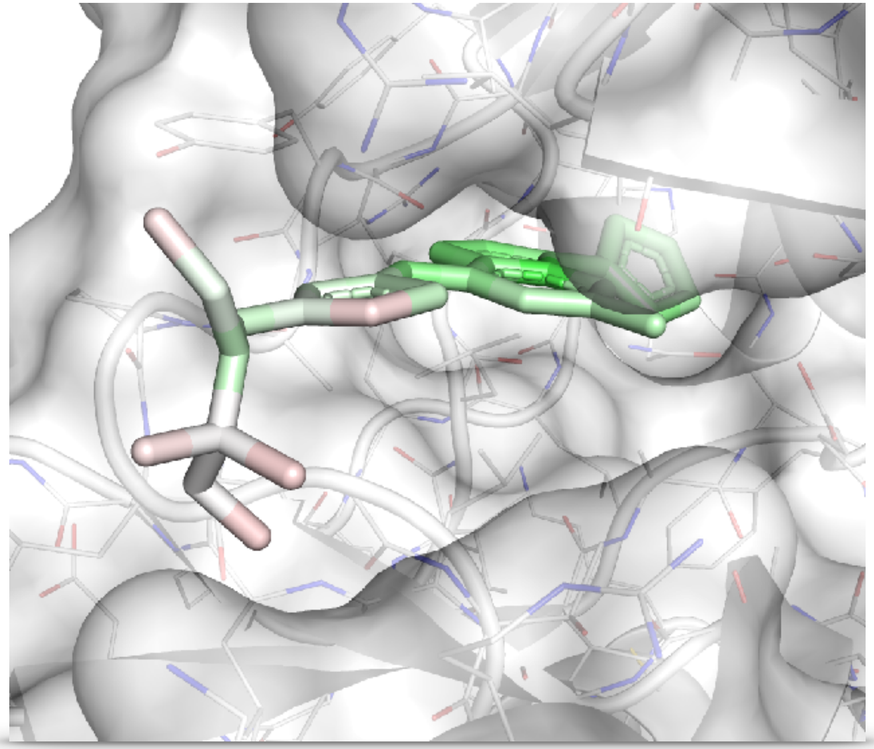
From the deconvolution paper

Structure Based Drug Design

Virtual Screening



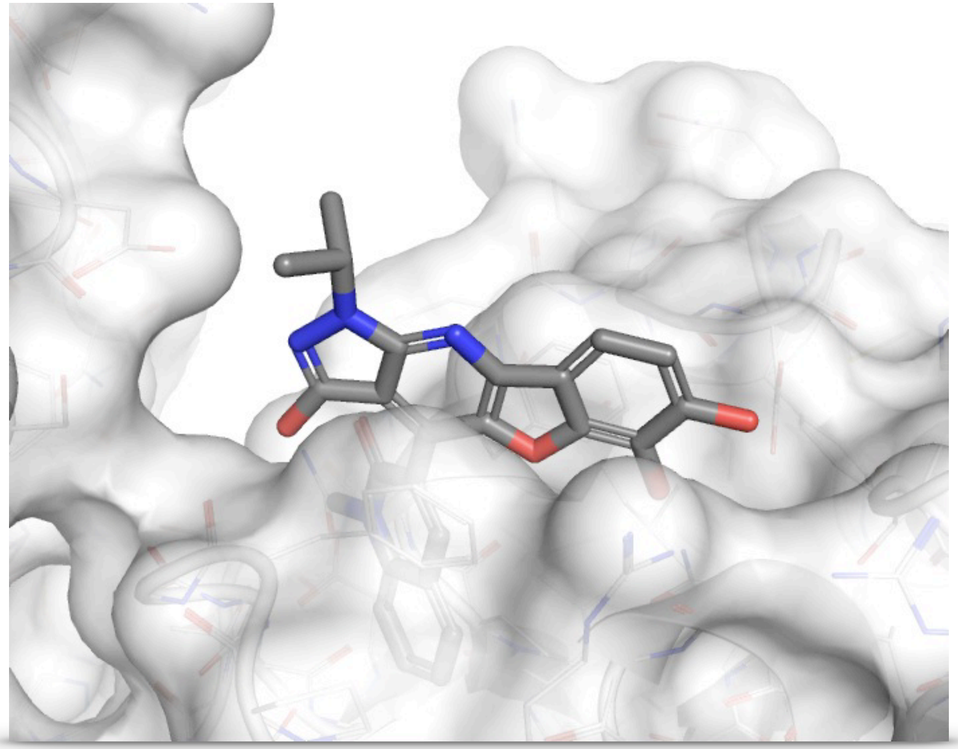
Lead Optimization



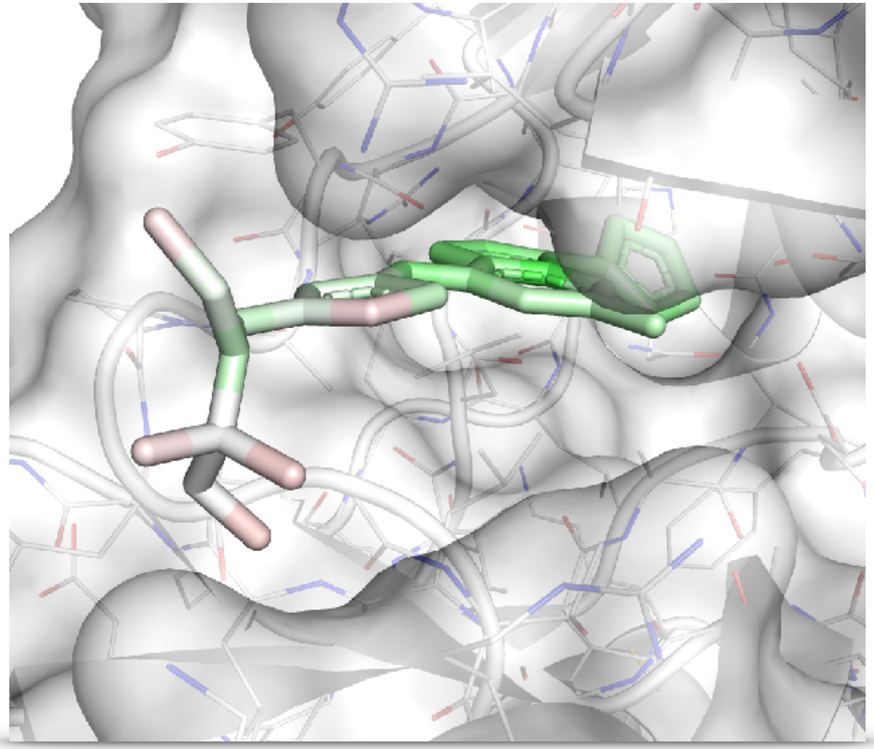
Pose Prediction → Binding Discrimination → Affinity Prediction

Structure Based Drug Design

Virtual Screening

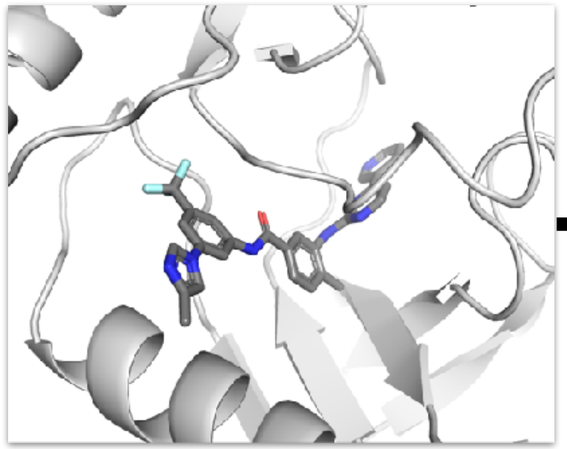


Lead Optimization



Pose Prediction → Binding Discrimination → Affinity Prediction

CNNs for Protein-Ligand Scoring



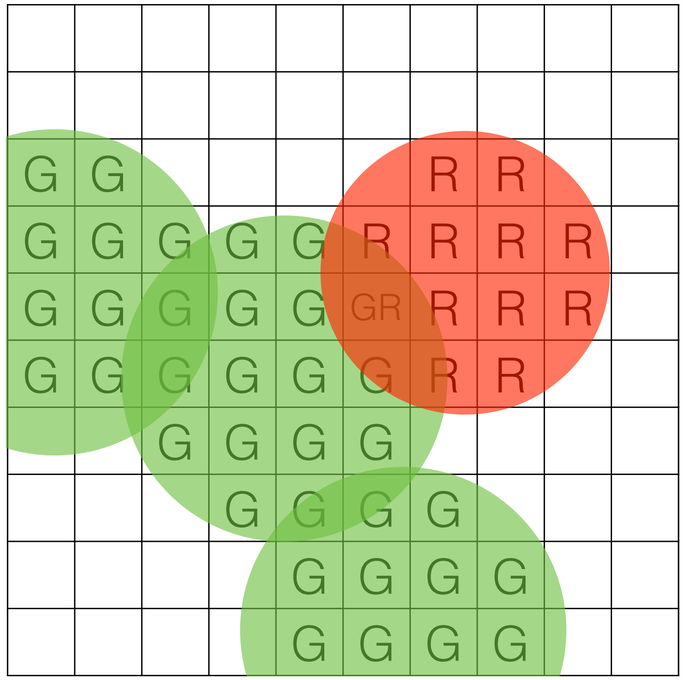
CNN

Pose Prediction

Binding Discrimination

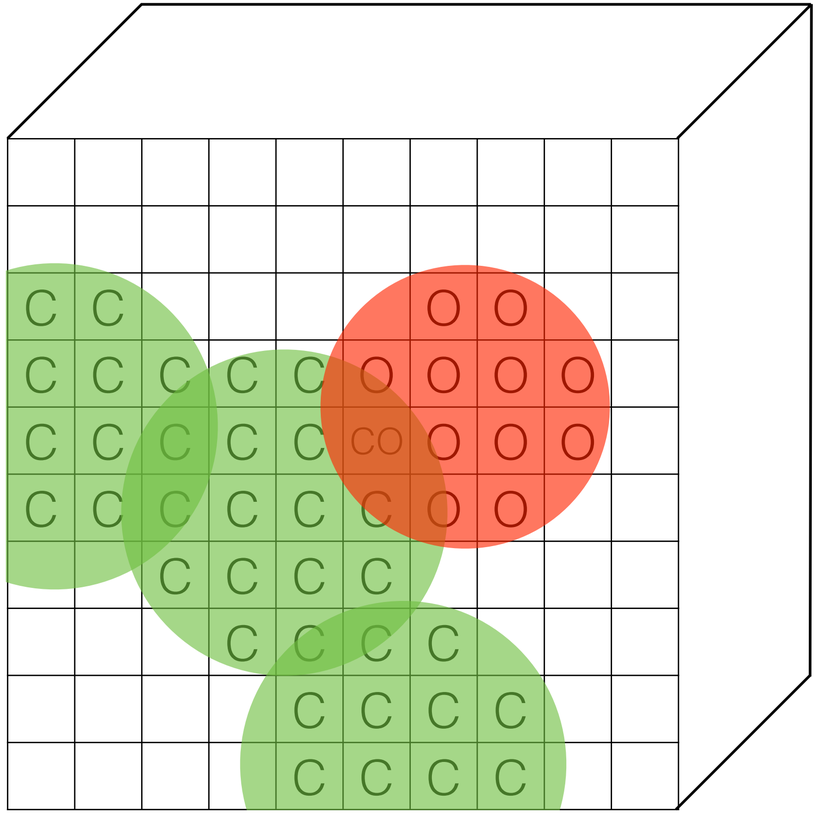
Affinity Prediction

Protein-Ligand Representation



(R,G,B) pixel

Protein-Ligand Representation



(R,G,B) pixel →

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

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

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)

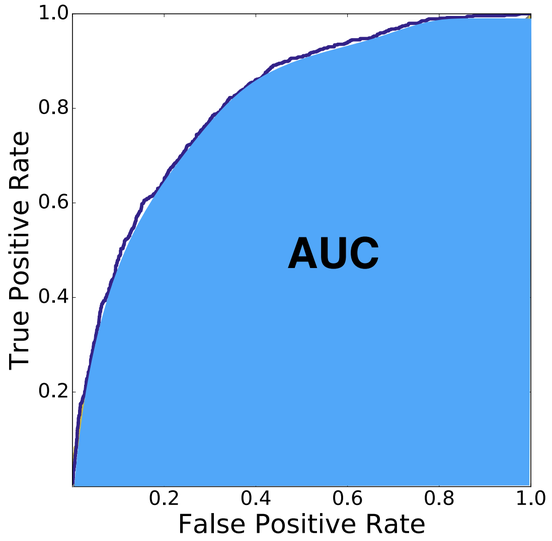
4056 protein-ligand complexes

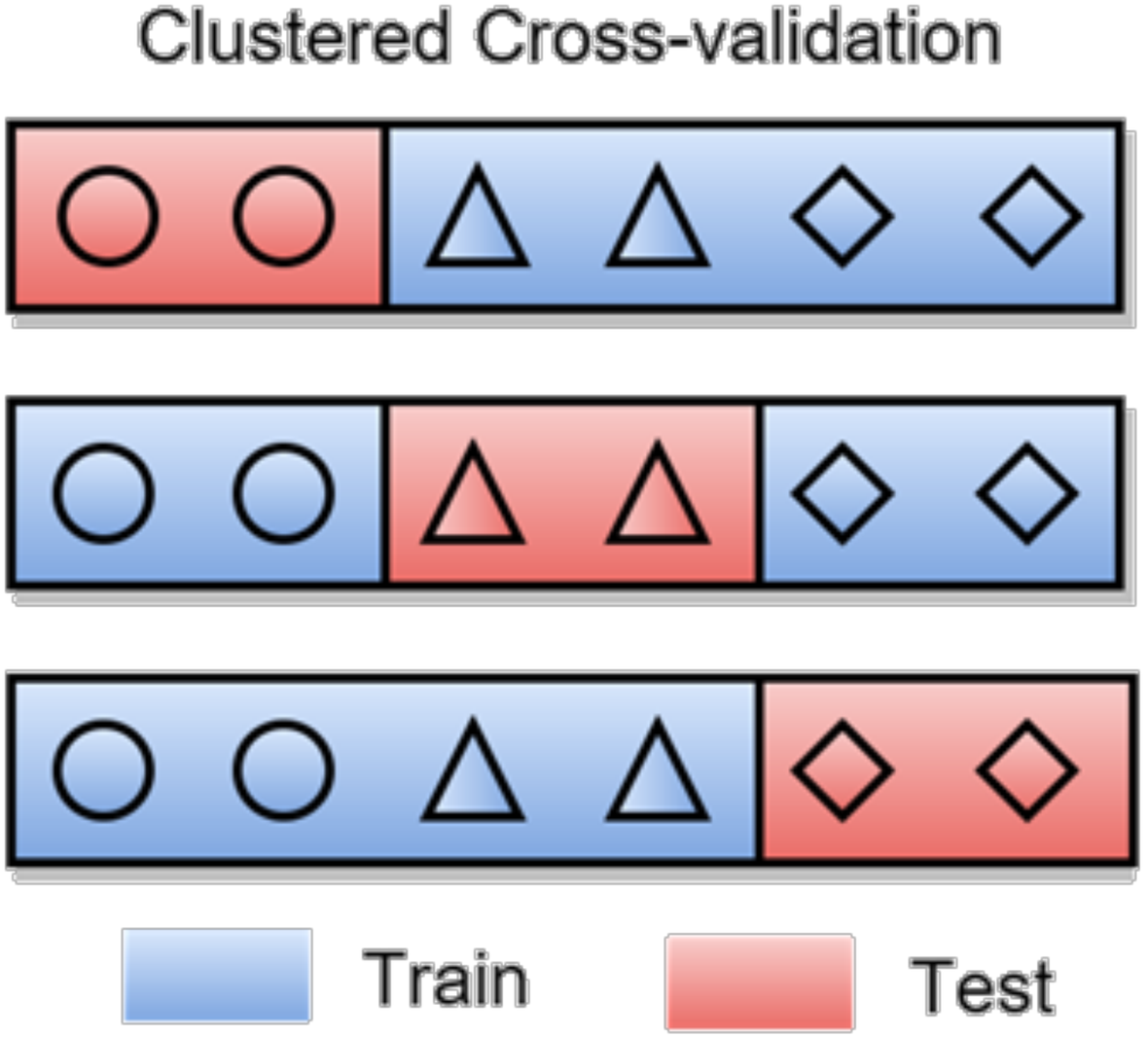
- diverse targets
- wide range of affinities
- **generate poses** with AutoDock Vina
- include minimized crystal pose
  - 8,688 <2Å RMSD (actives)
  - 76,743 >4Å RMSD (decoys)

Model Evaluation

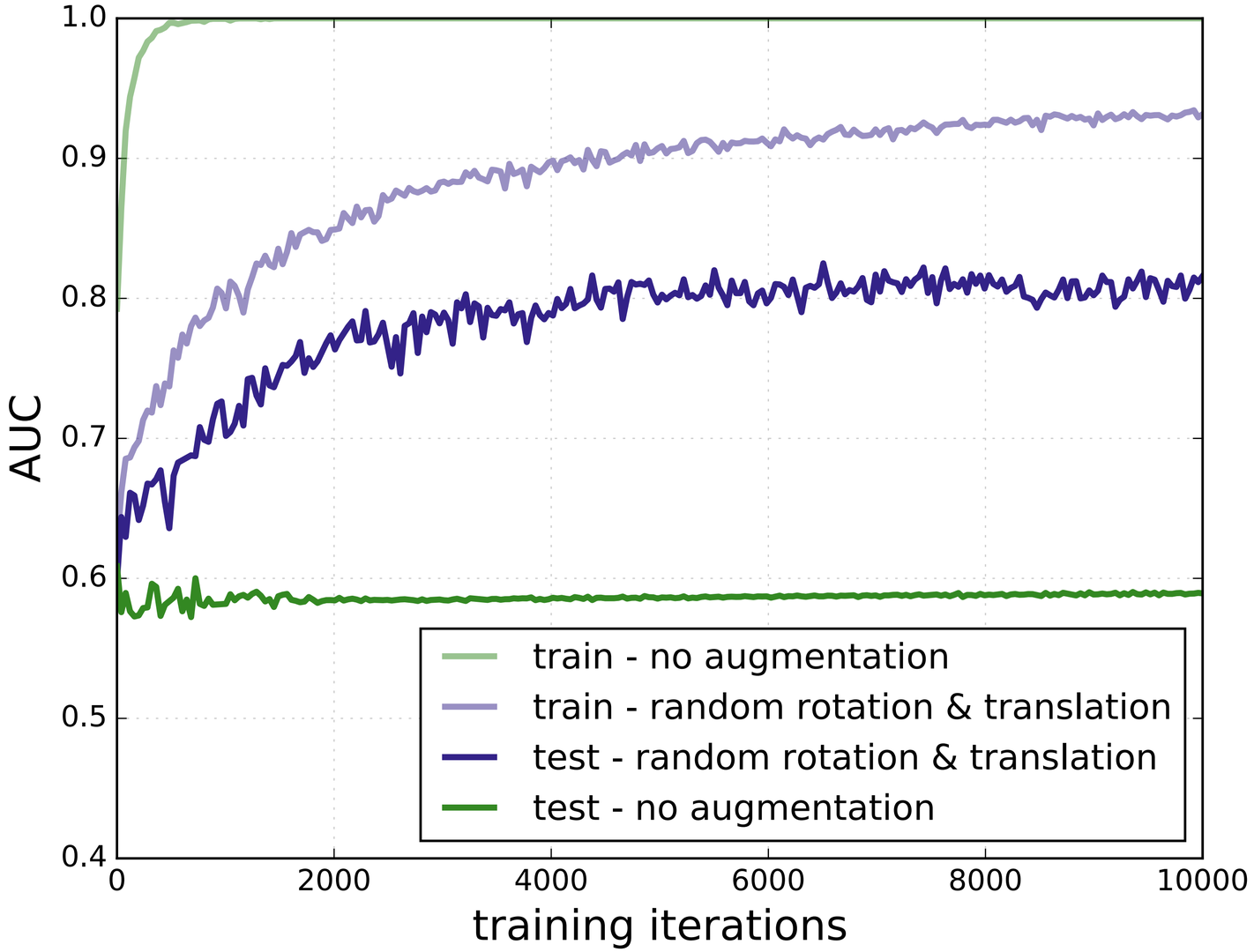
CSAR: >90% similar targets kept in same fold

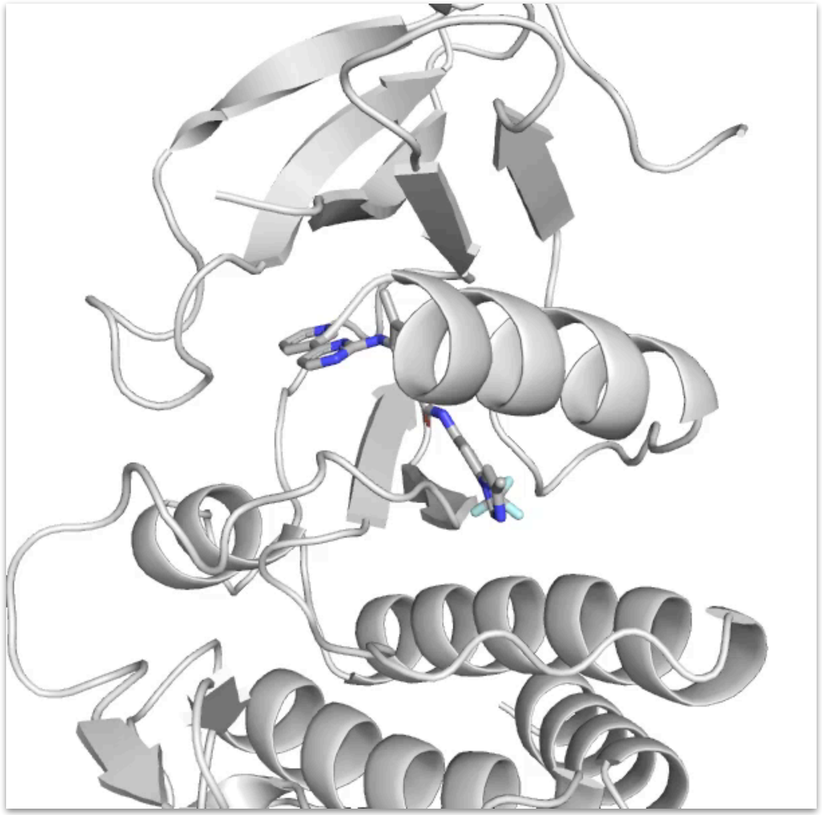
DUD-E & PDBbind: >80% similar targets kept in same fold



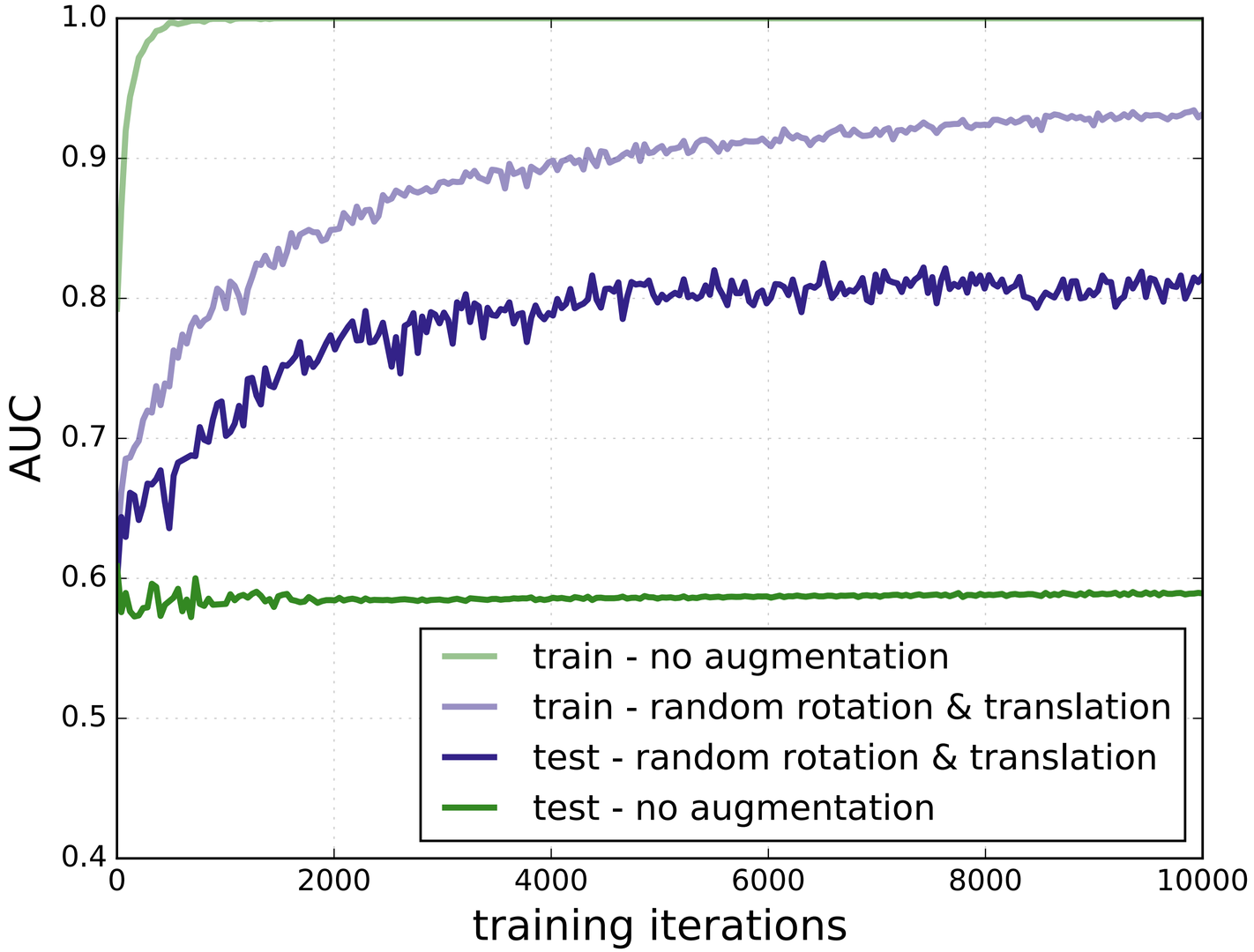


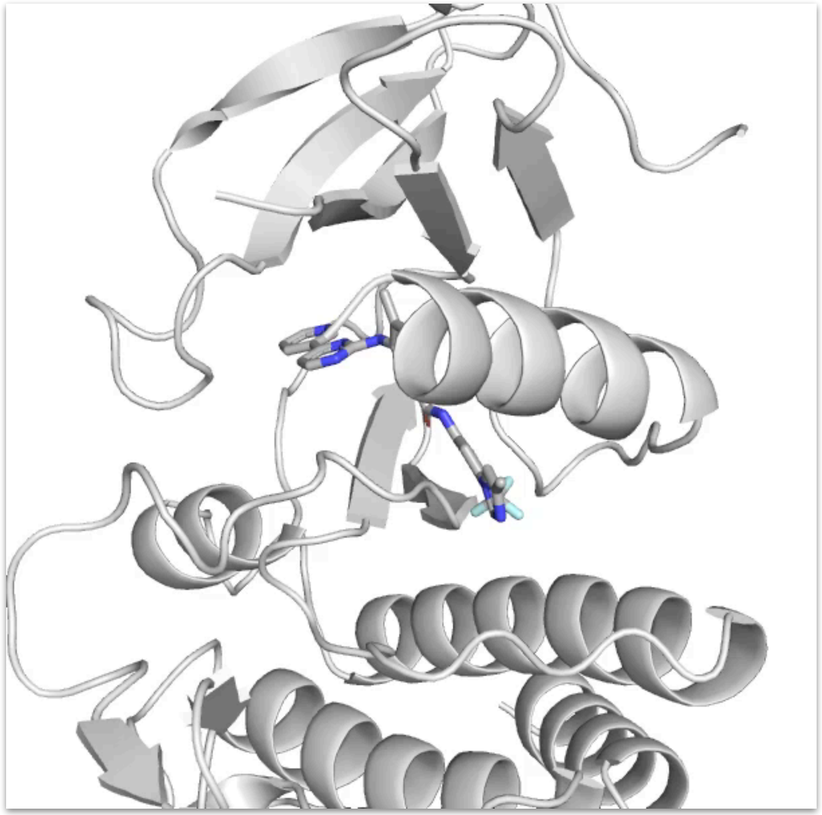
Data Augmentation



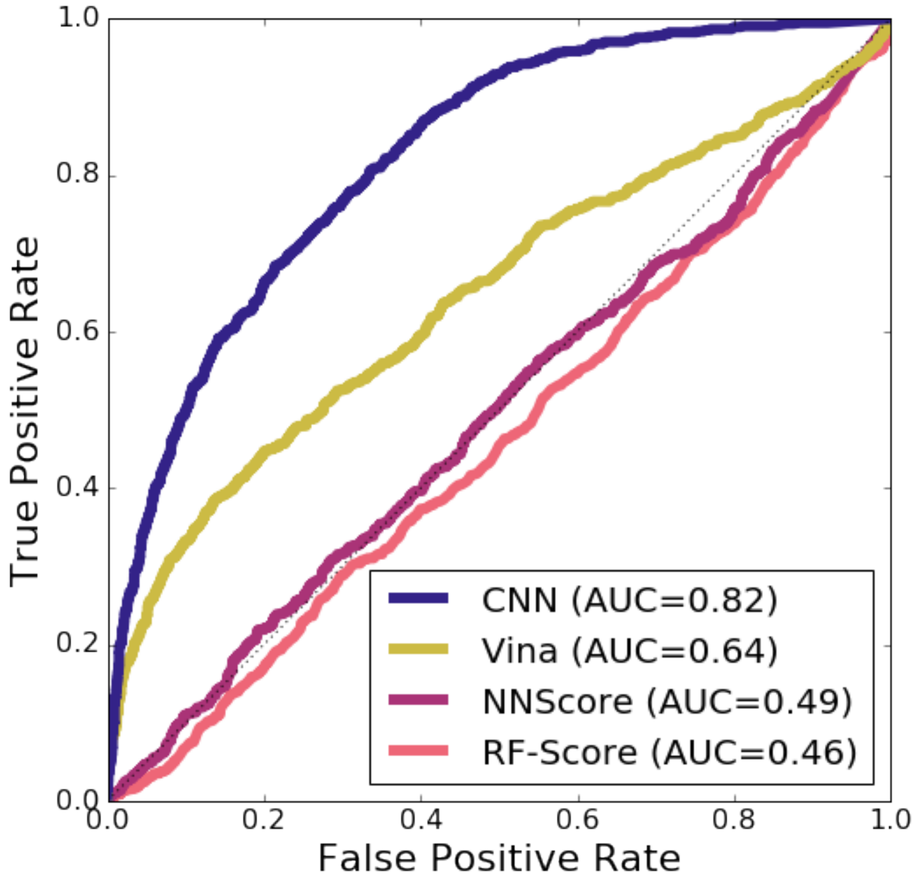


Data Augmentation

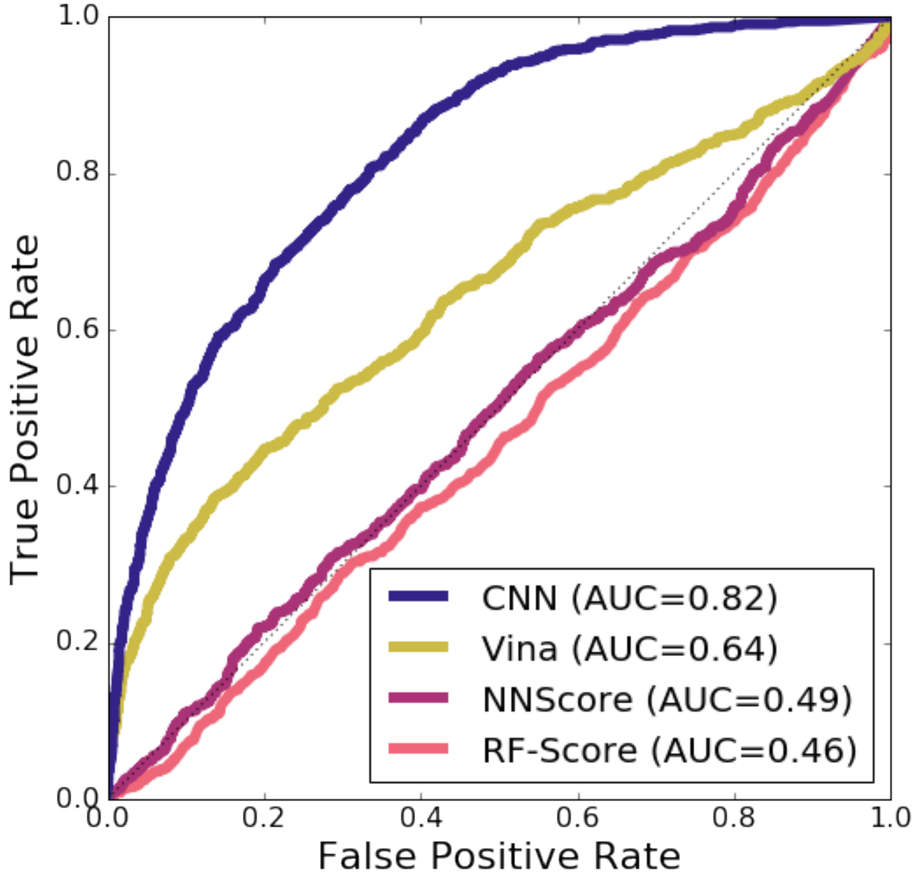




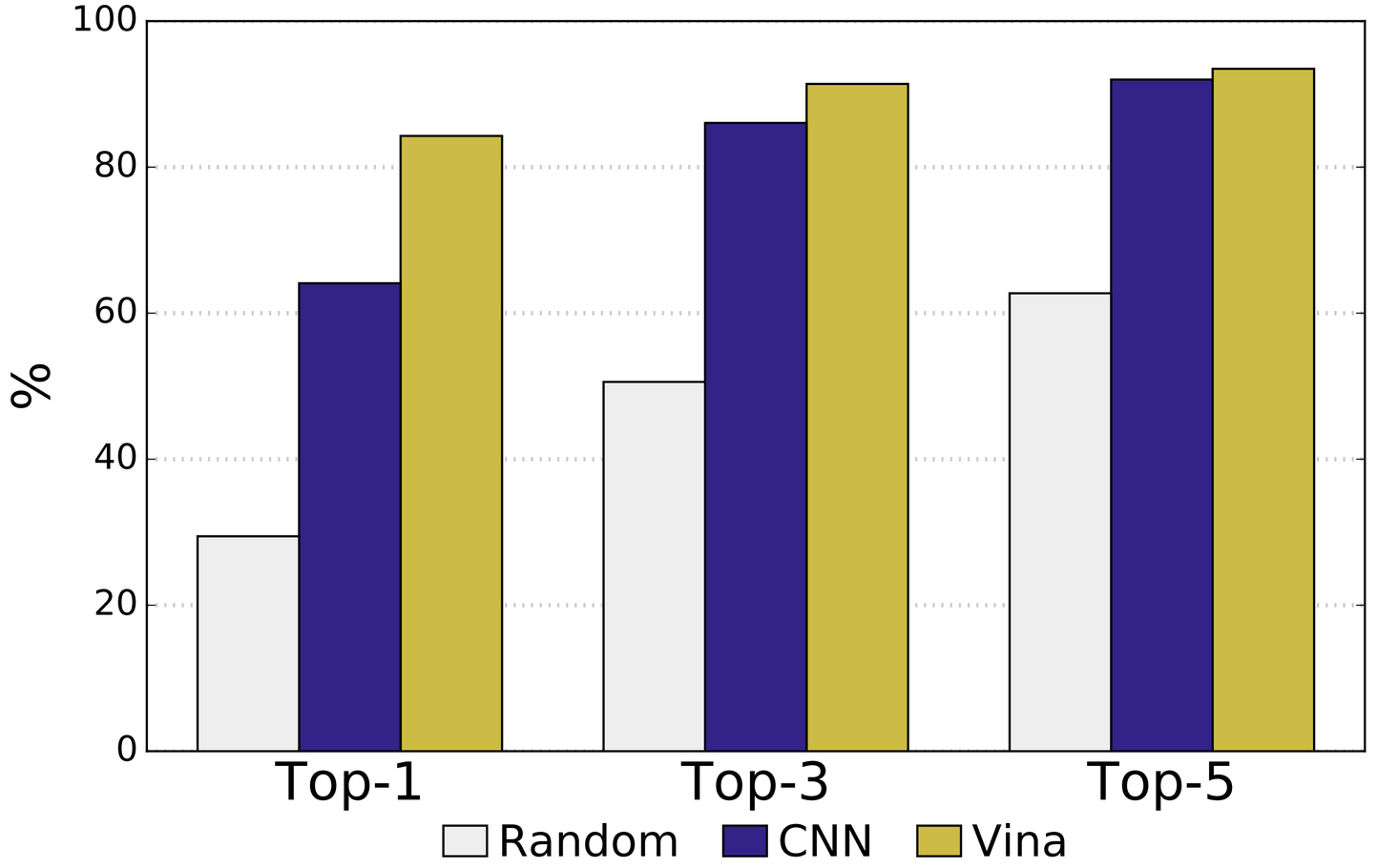
Pose Prediction (CSAR)



Pose Prediction (CSAR)

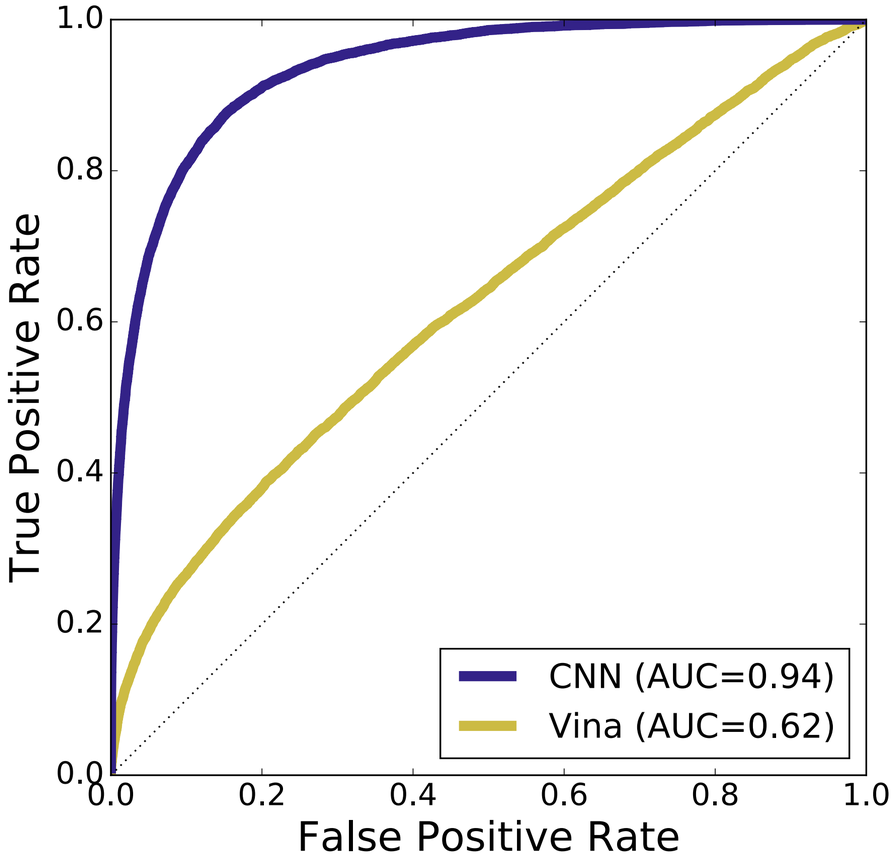


inter-target ranking

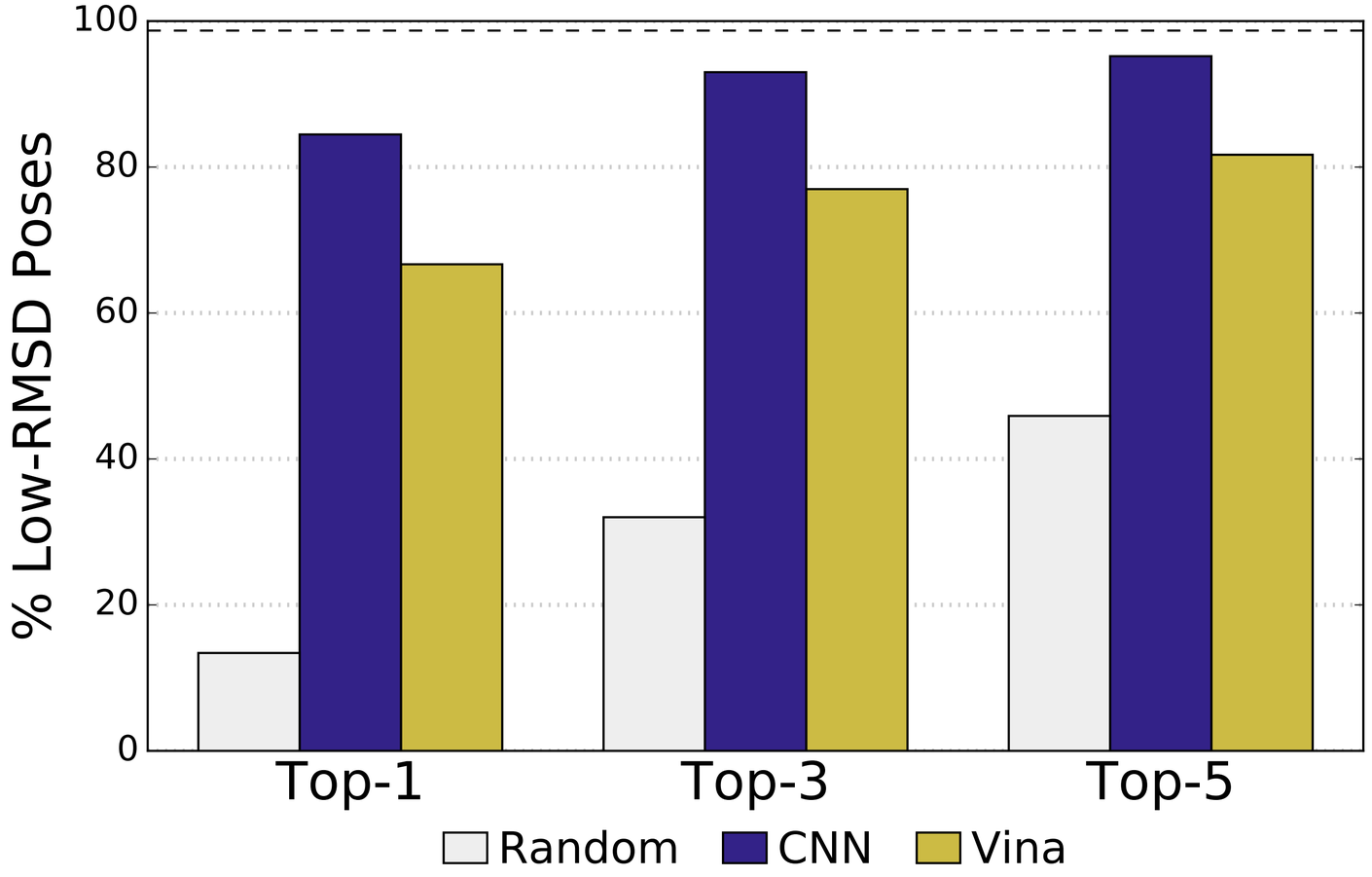


intra-target ranking

Pose Prediction (PDBbind)

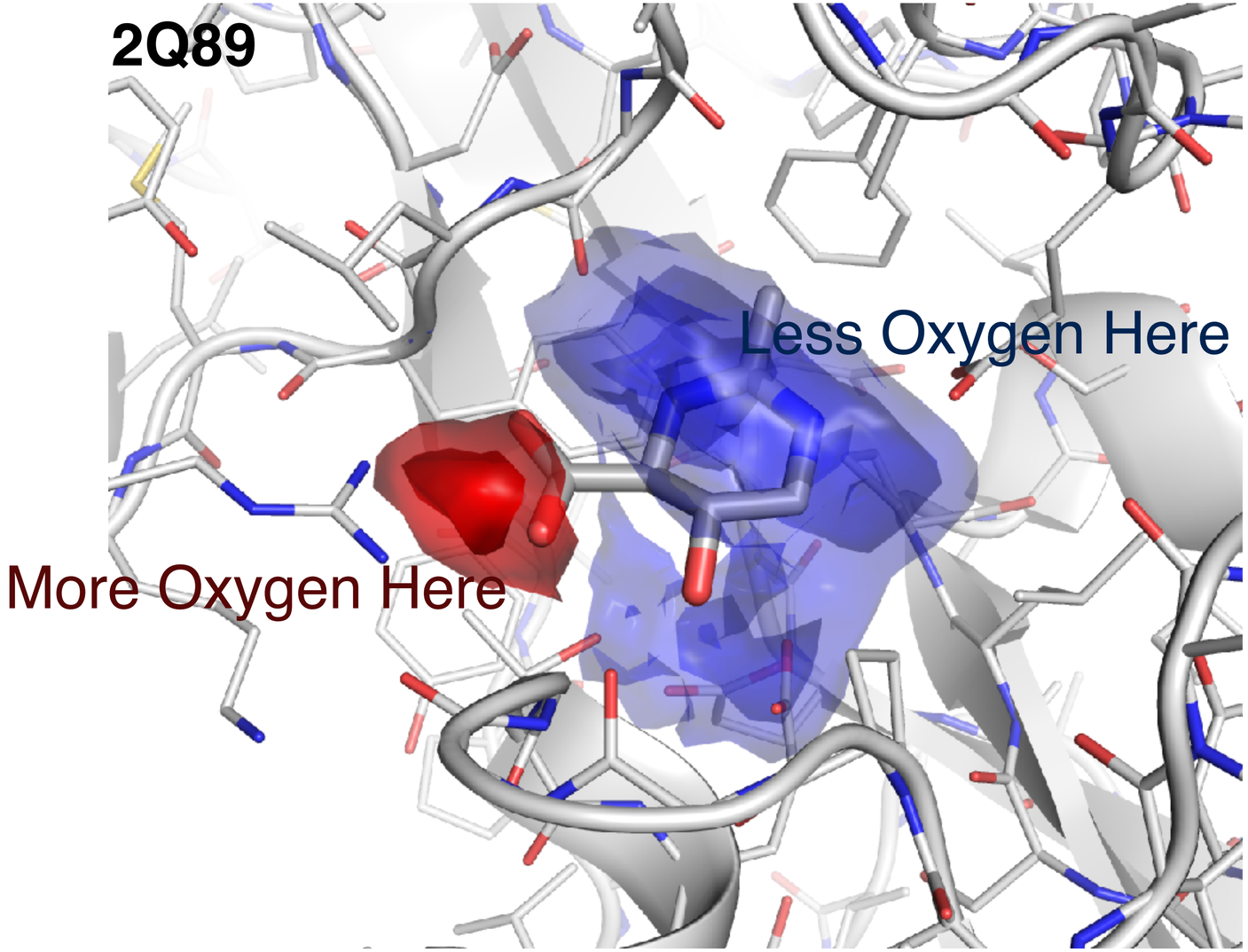


inter-target ranking

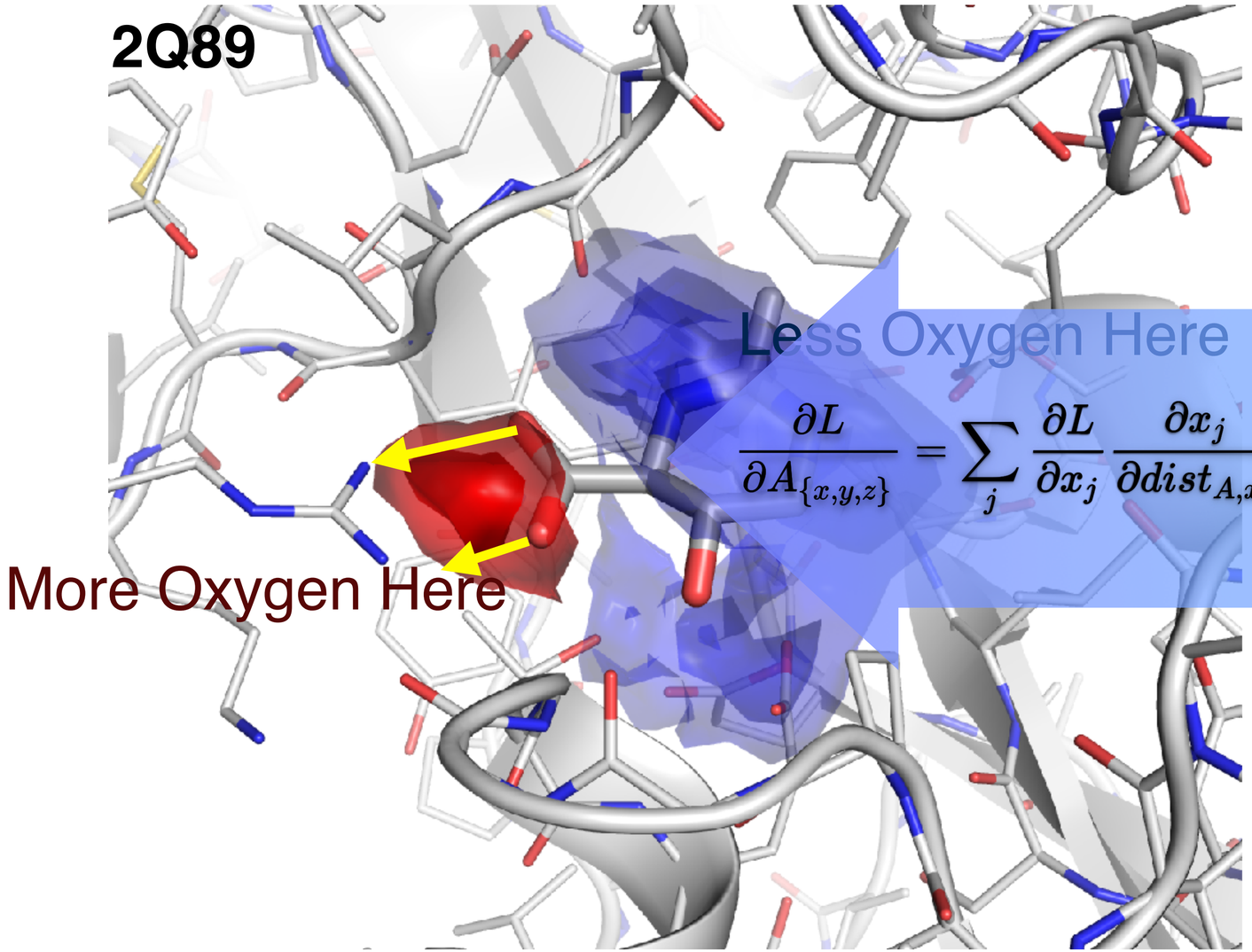


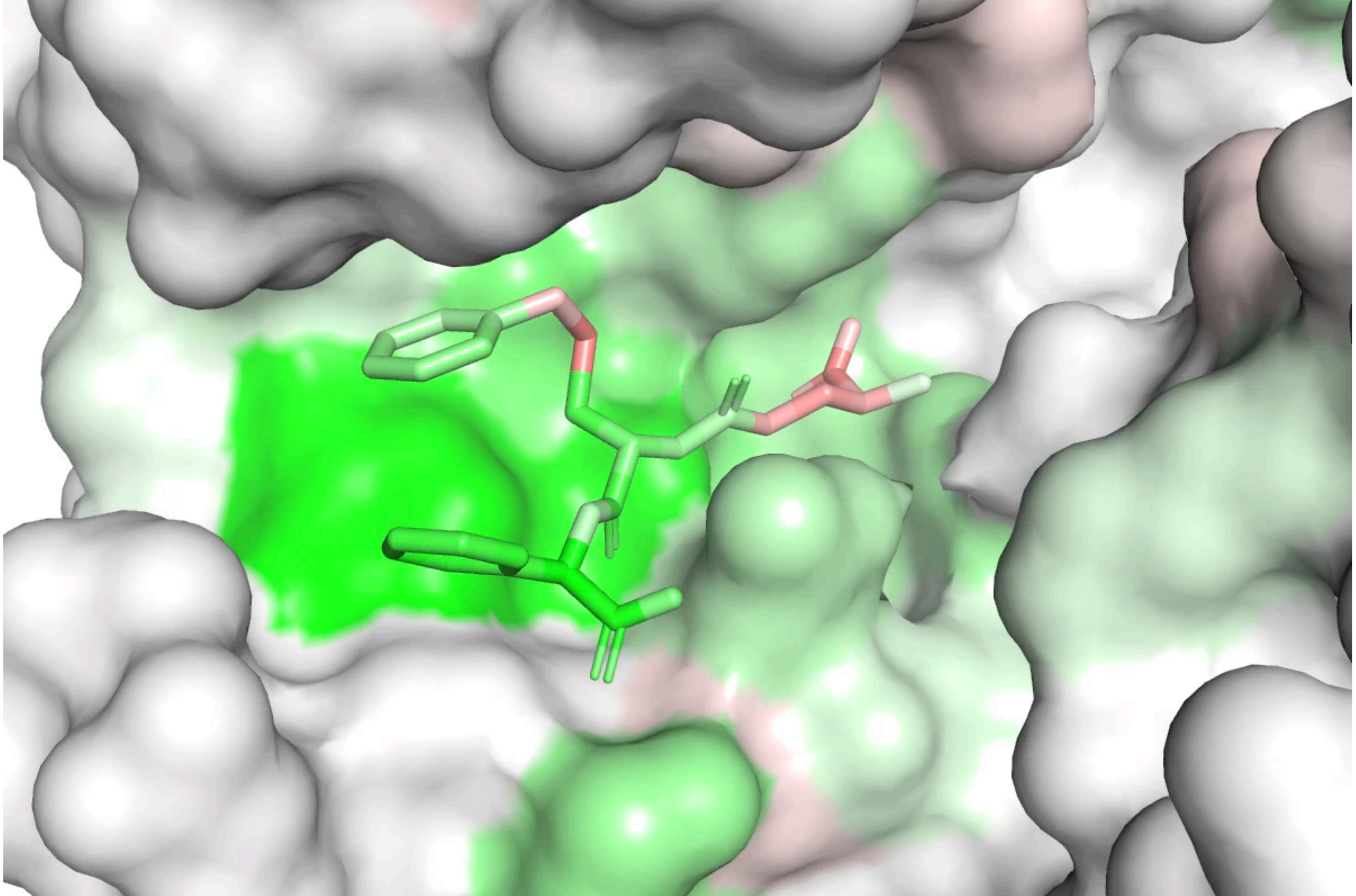
intra-target ranking

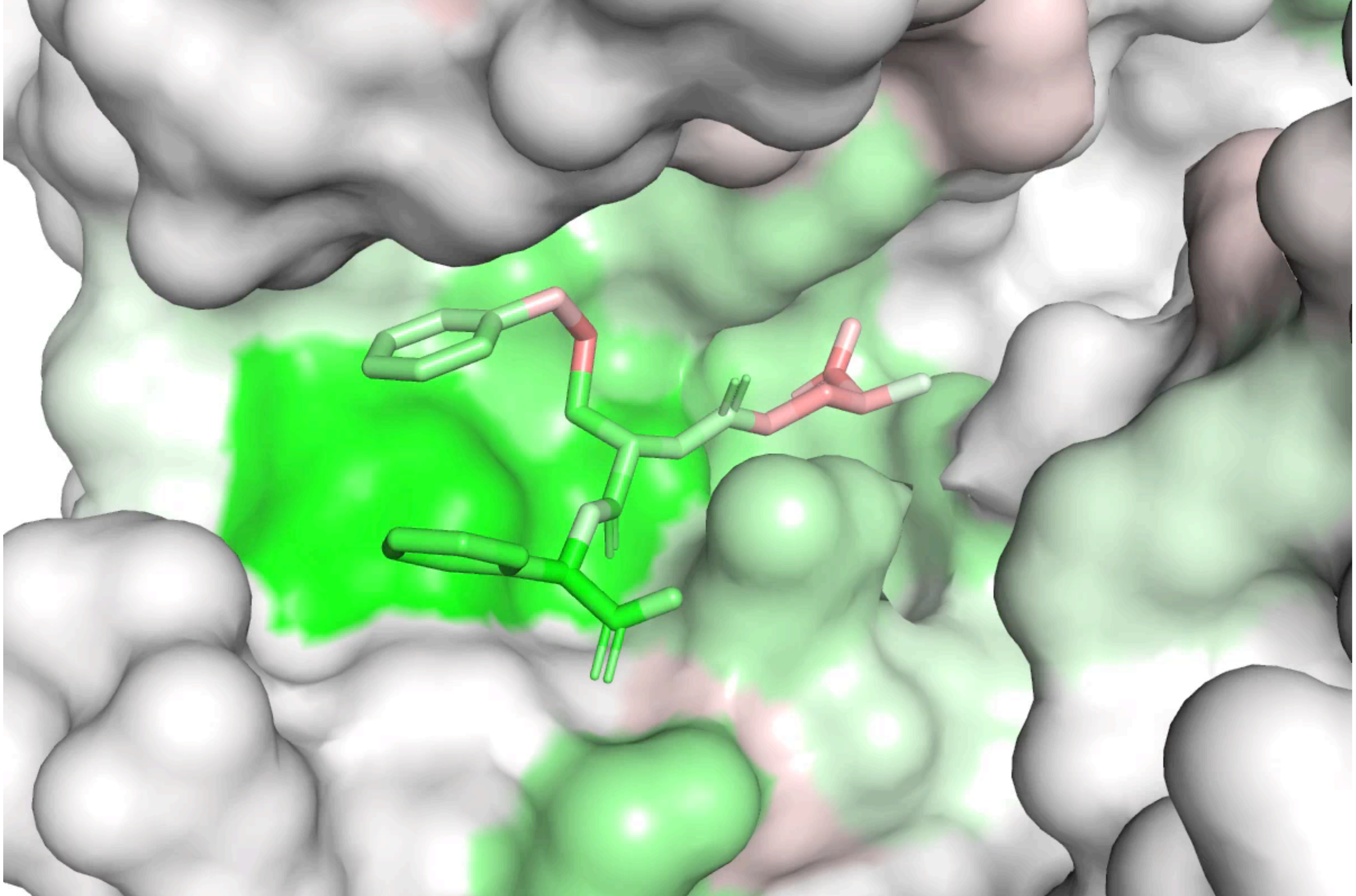
Beyond Scoring

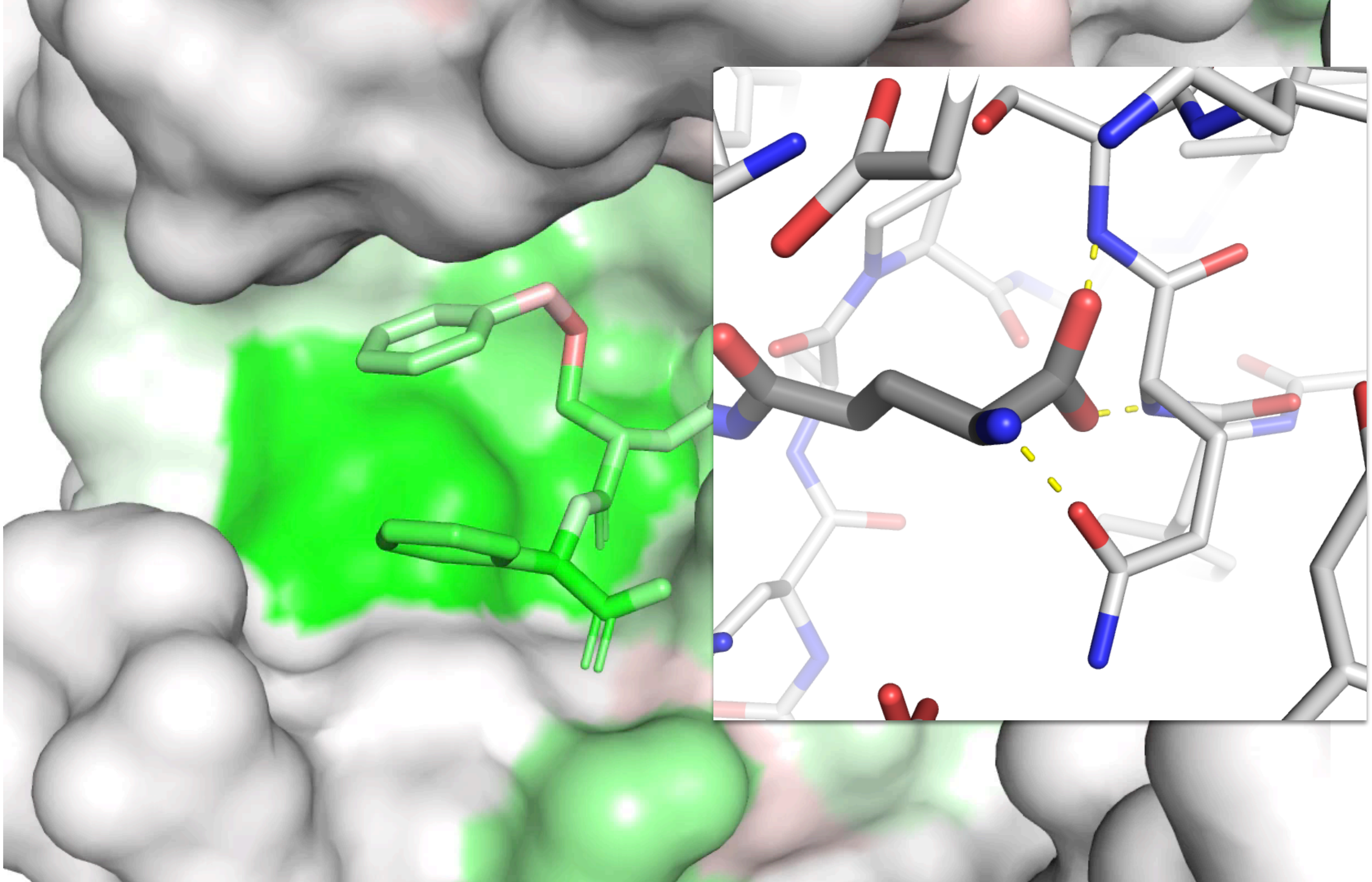


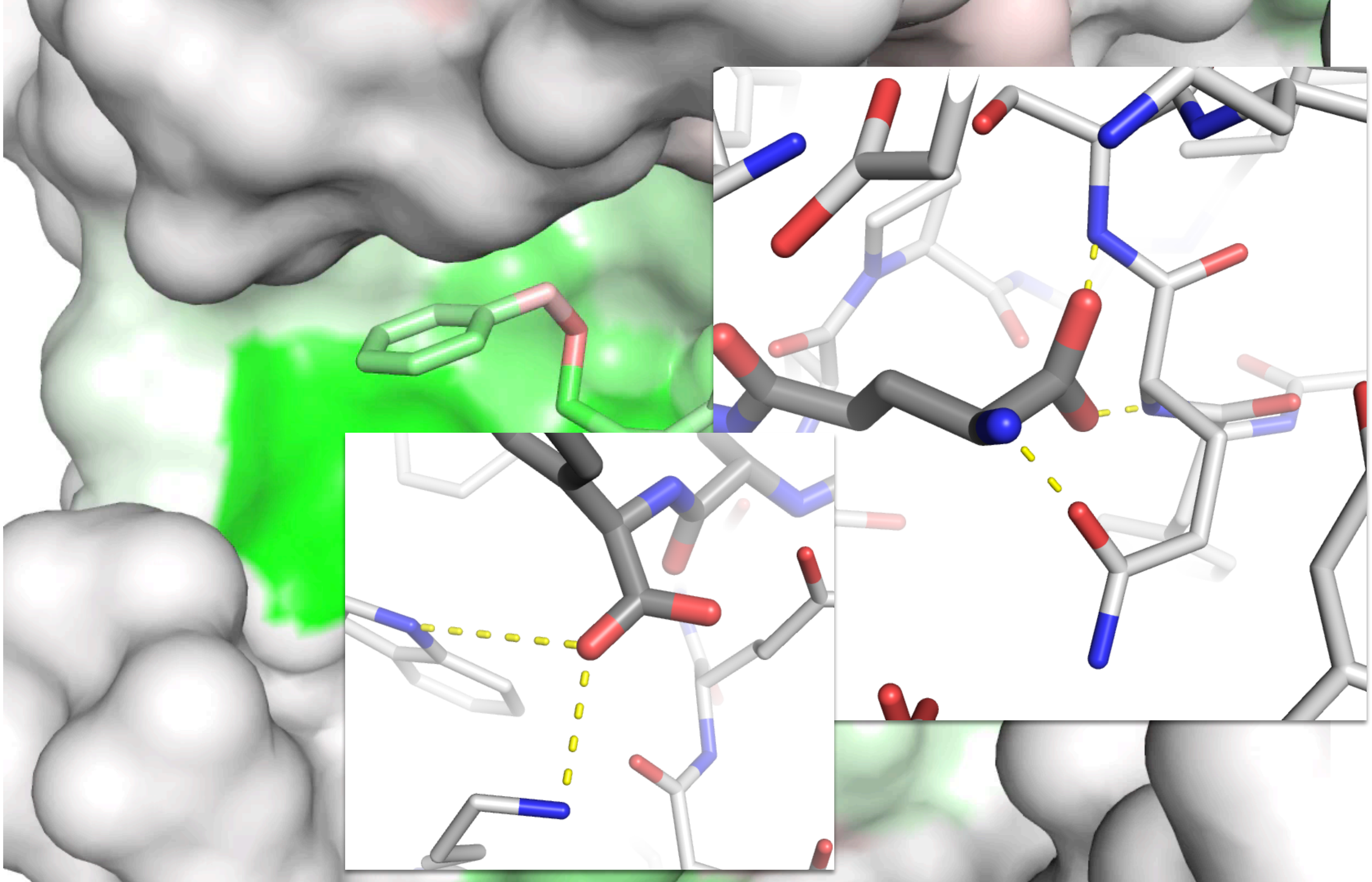
Beyond Scoring



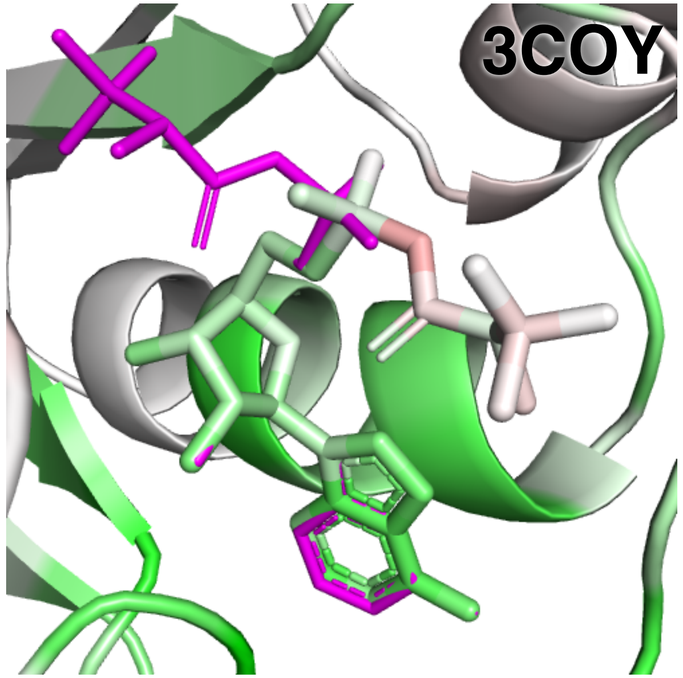


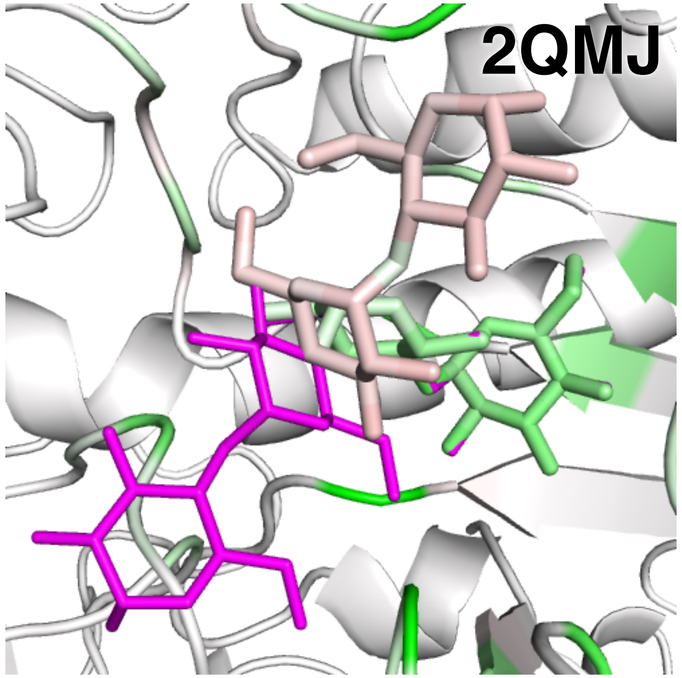


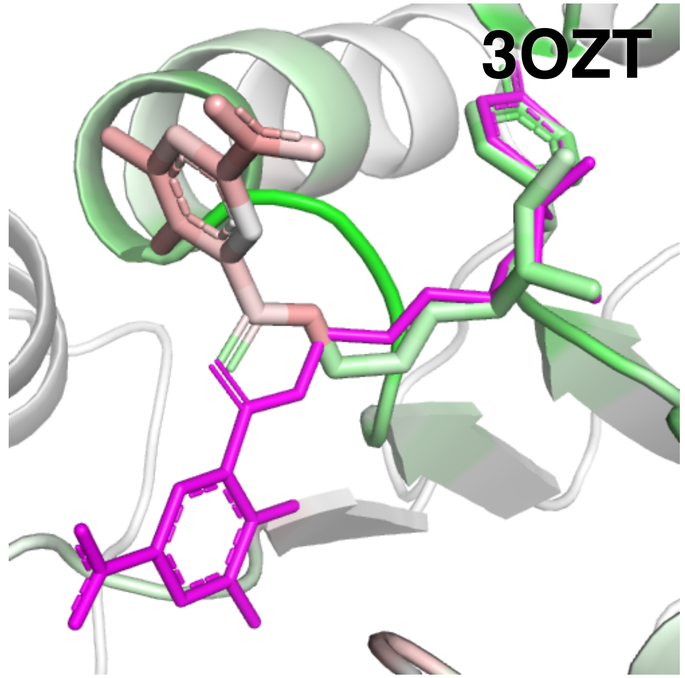




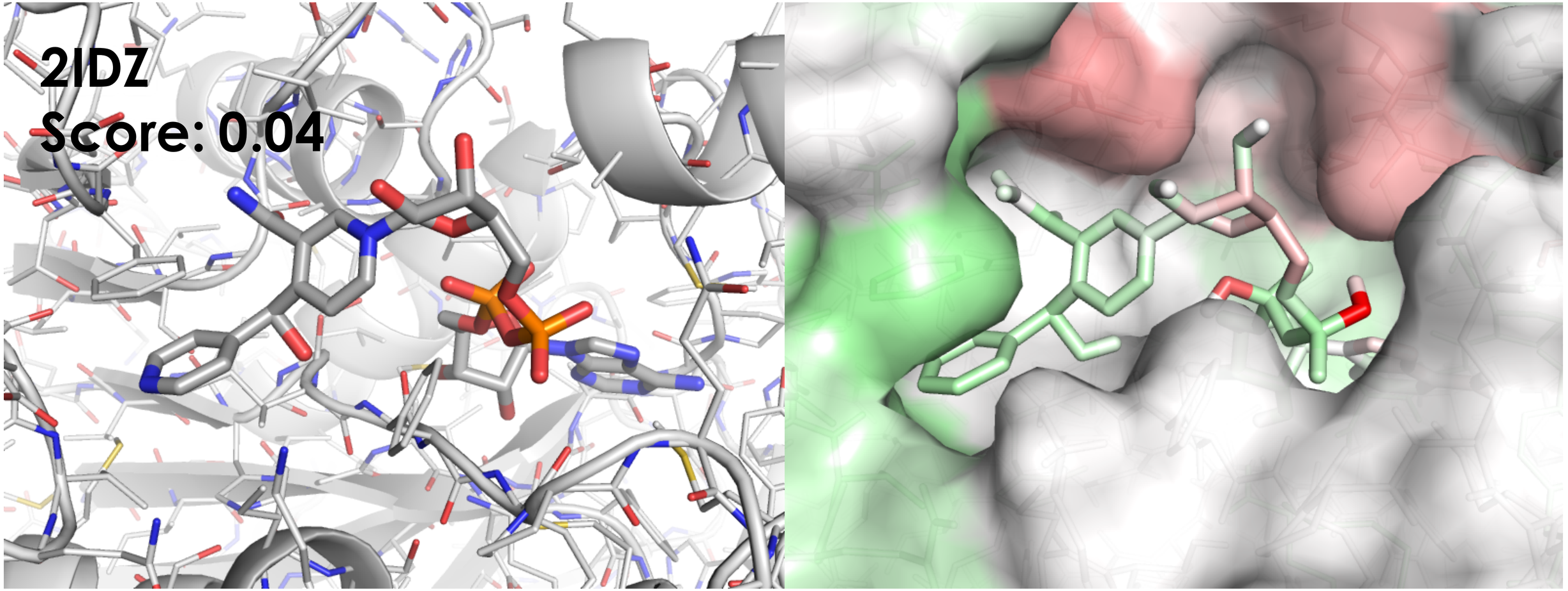
Visual Examples



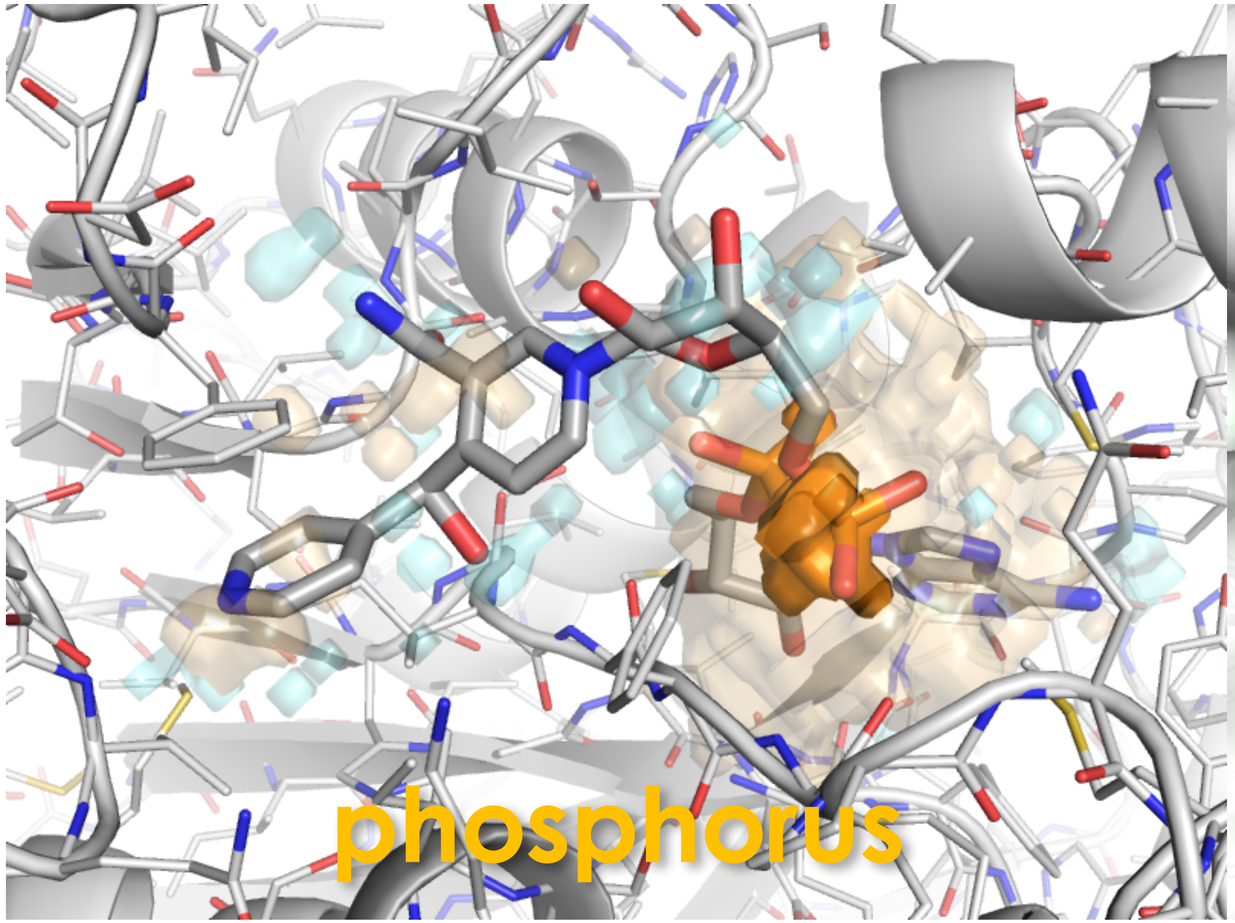


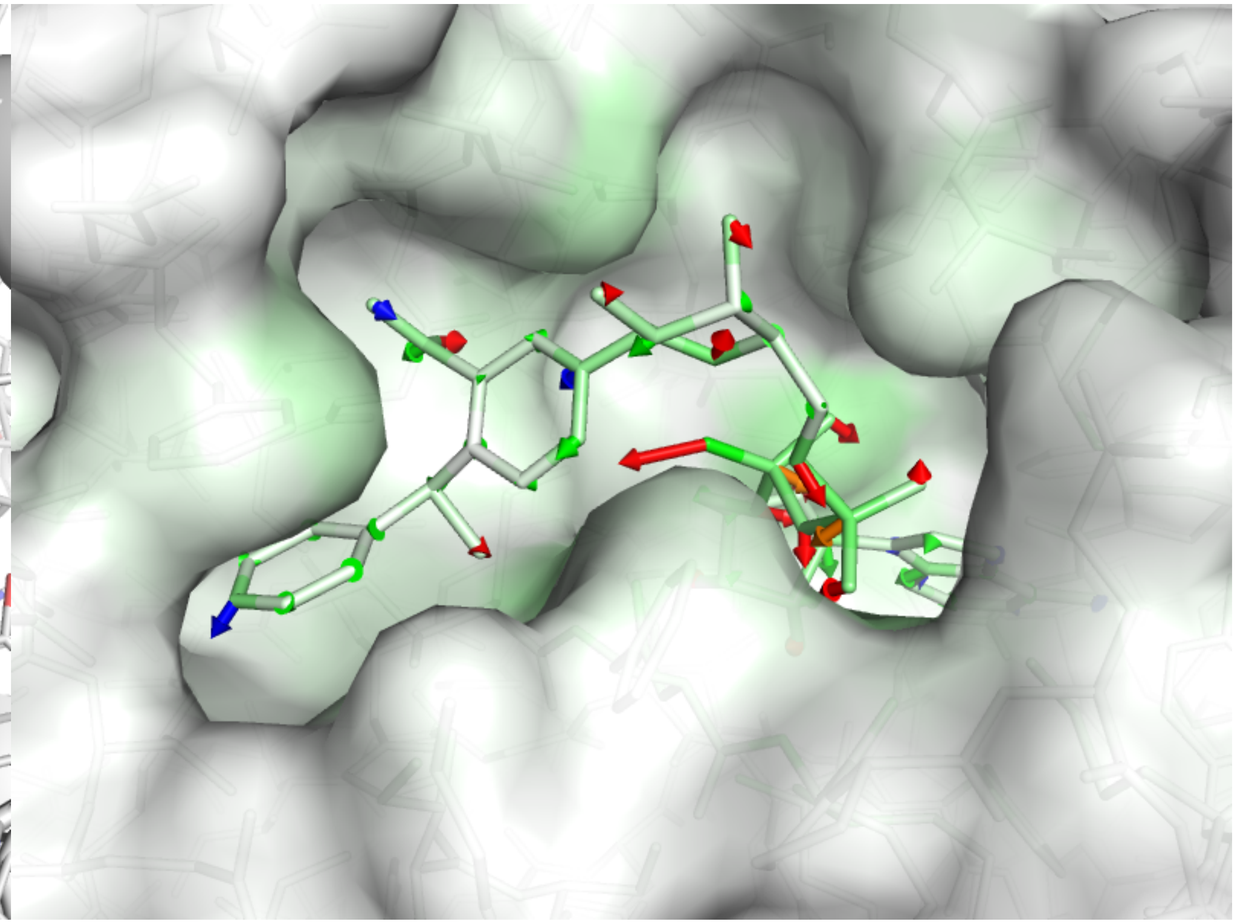


Partially Aligned Poses
Masking

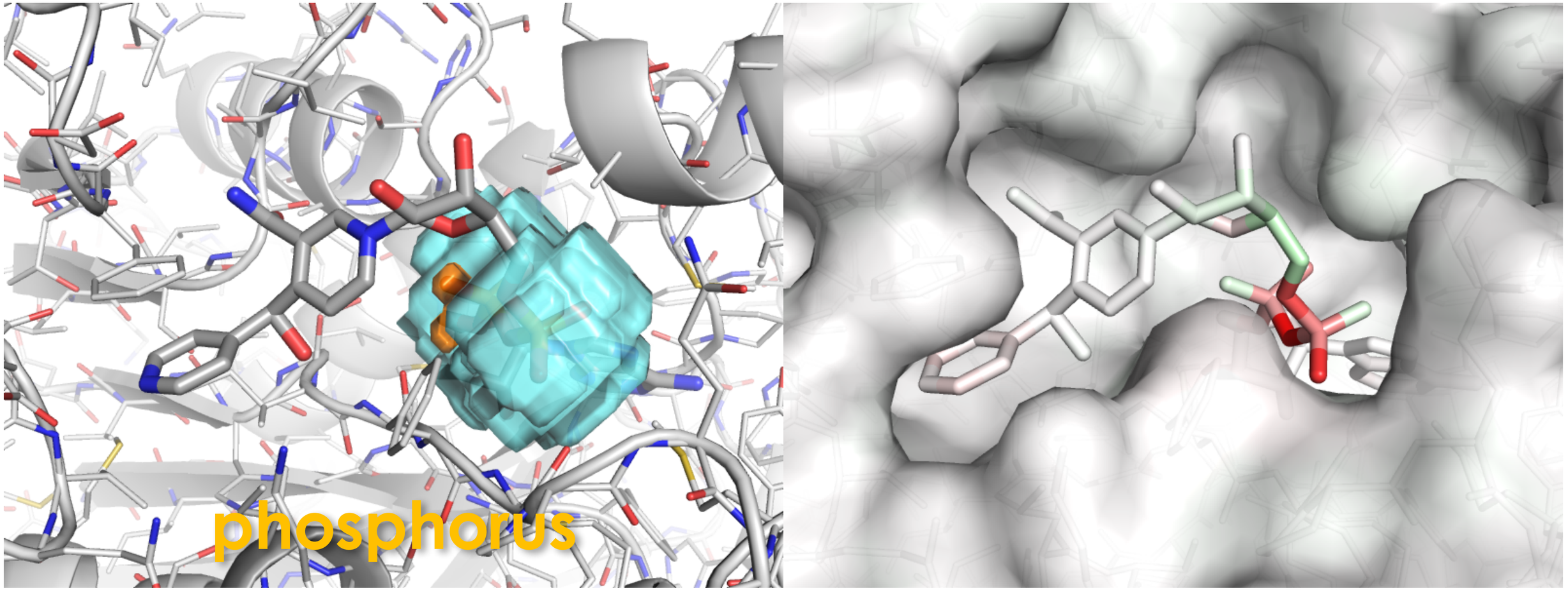


Masking

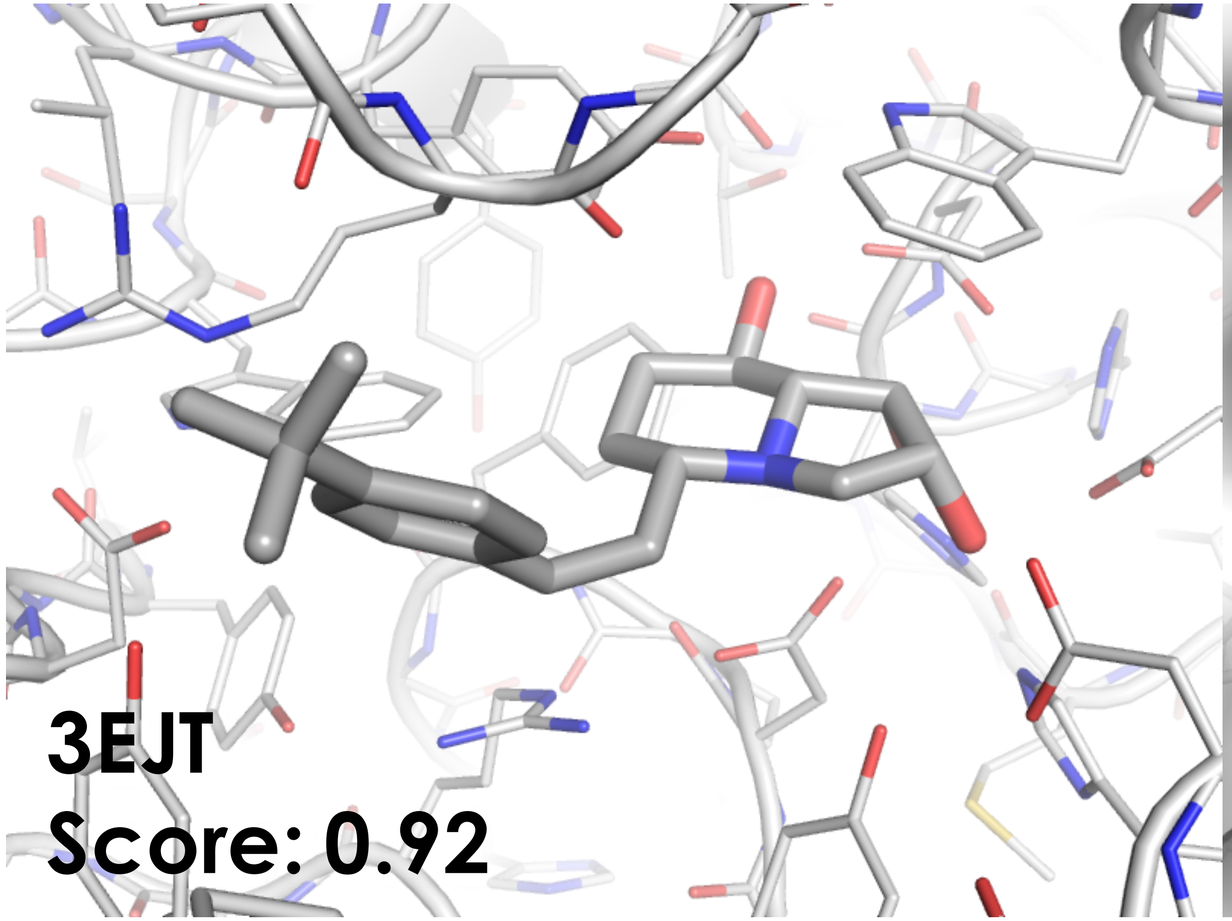


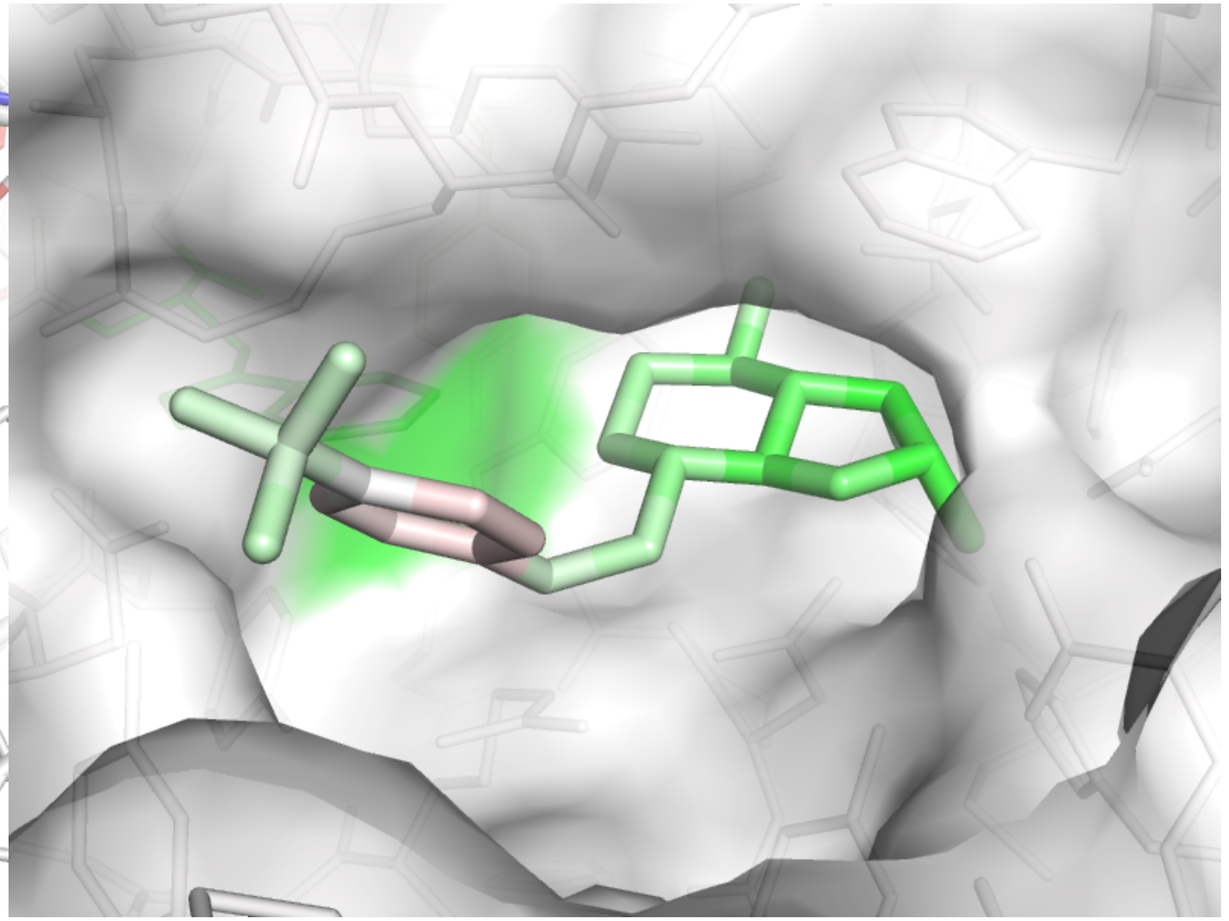


Gradients

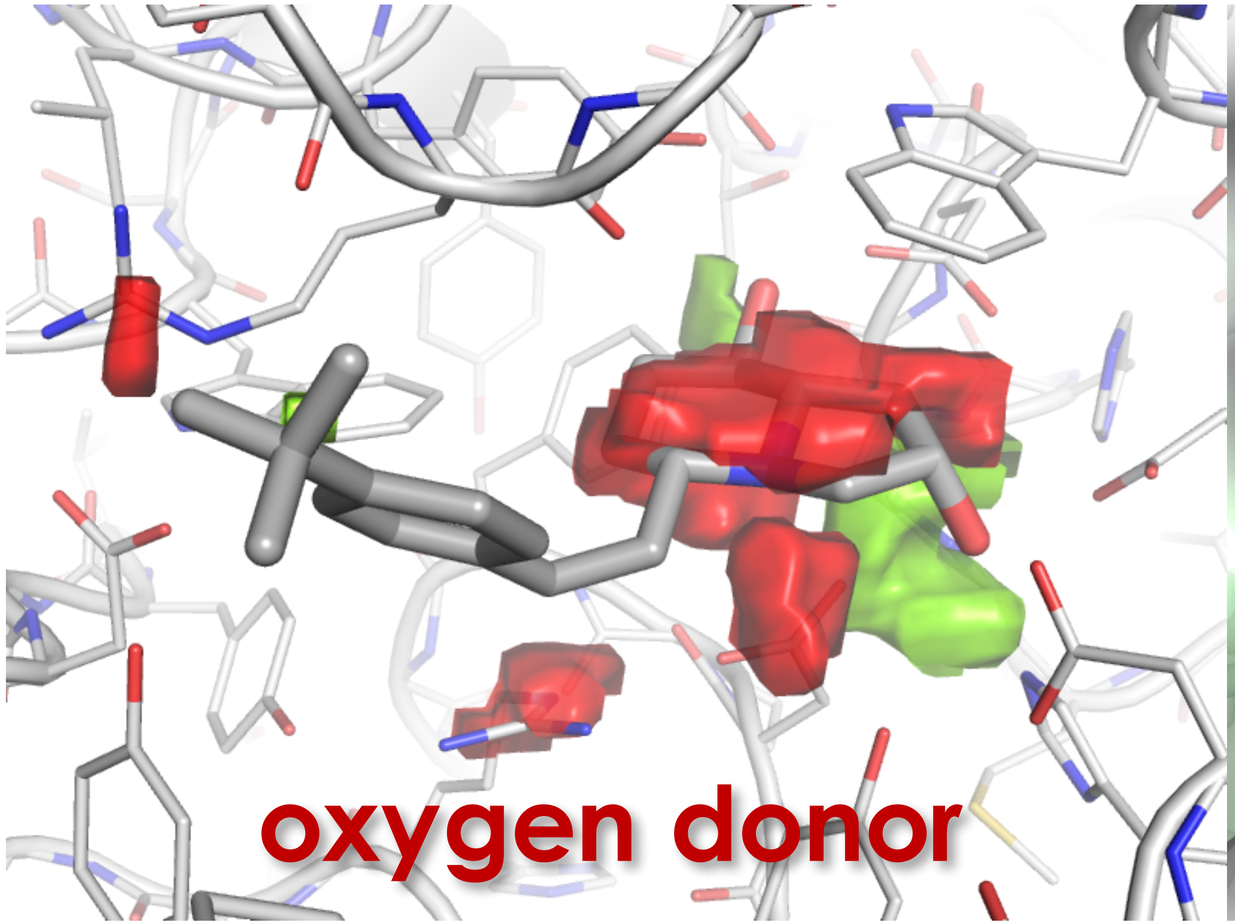


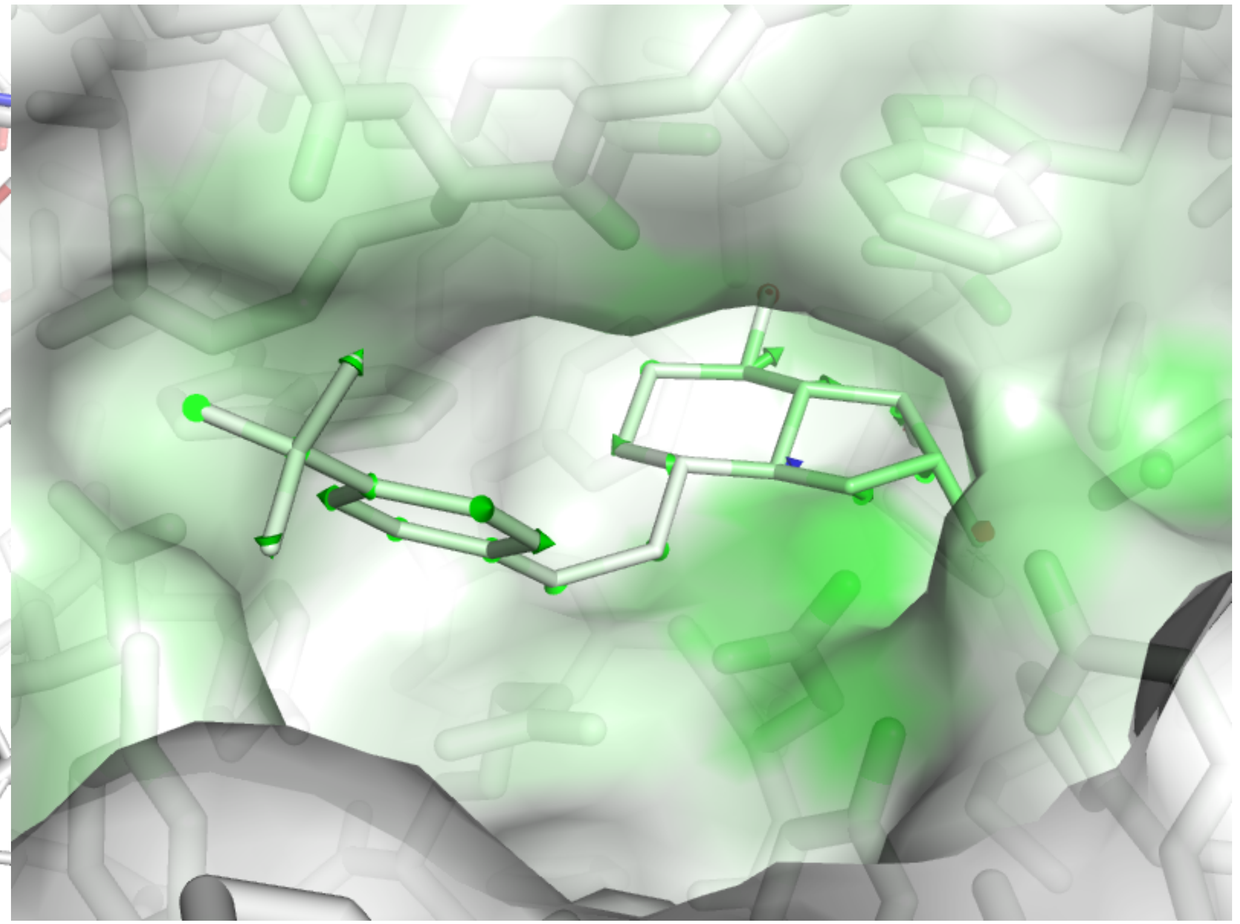
Relevance



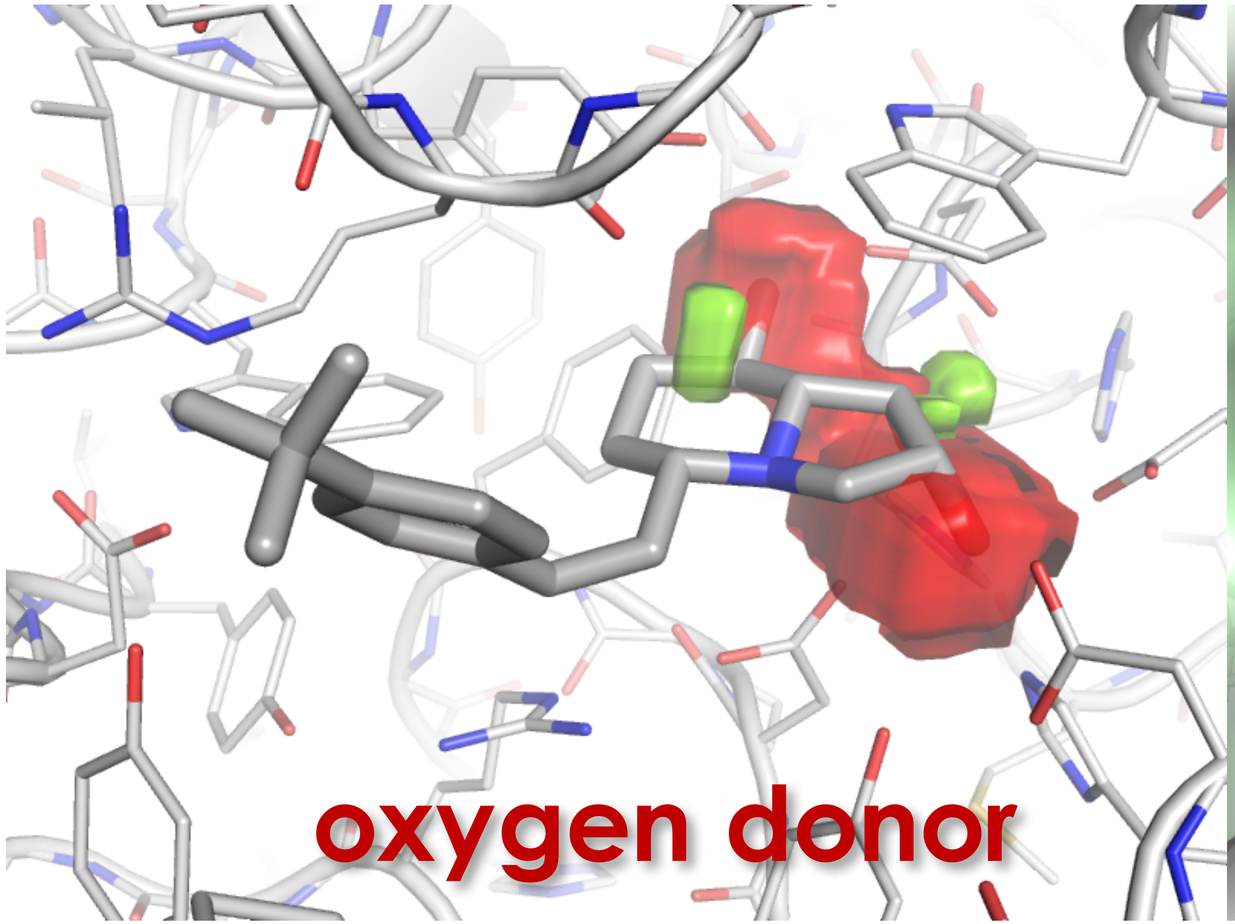


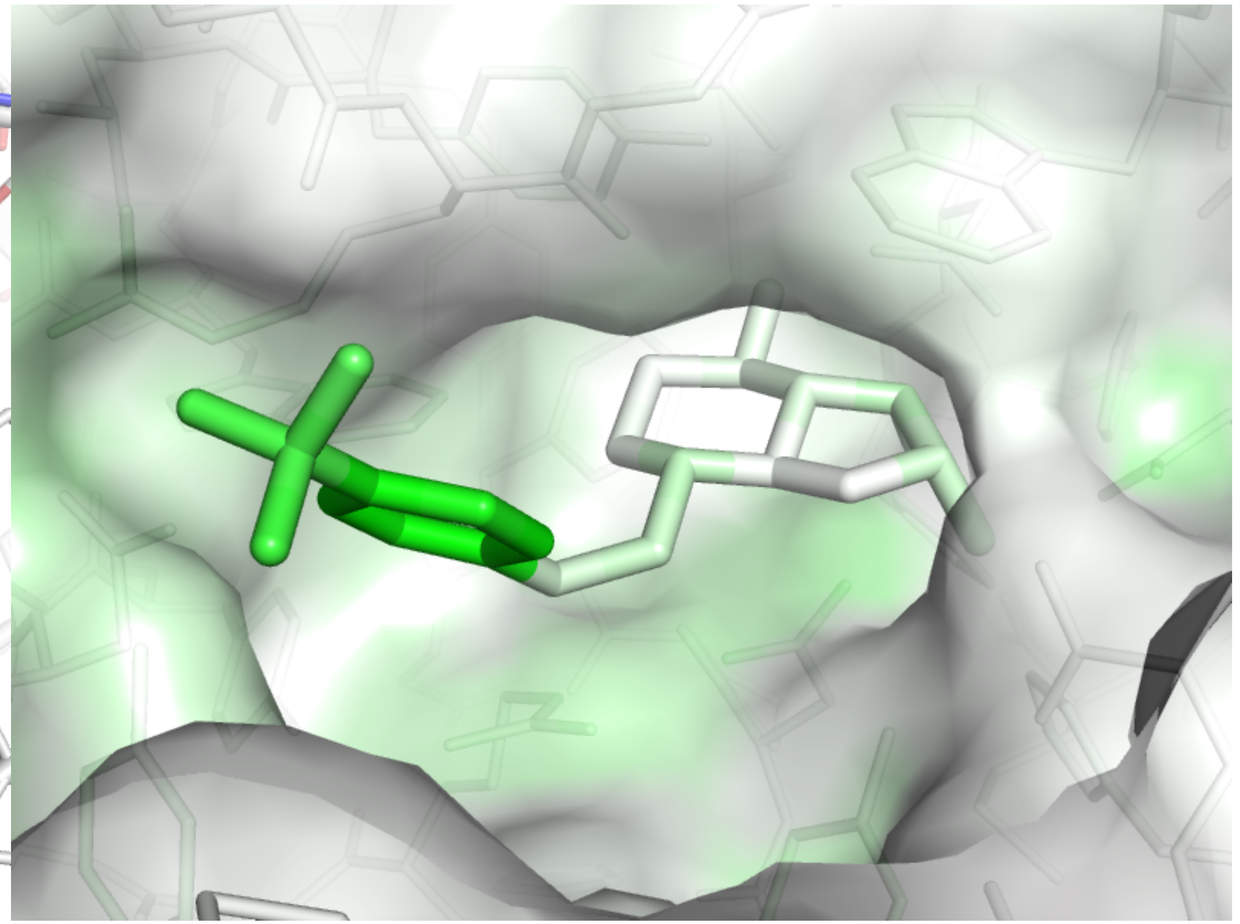
Masking



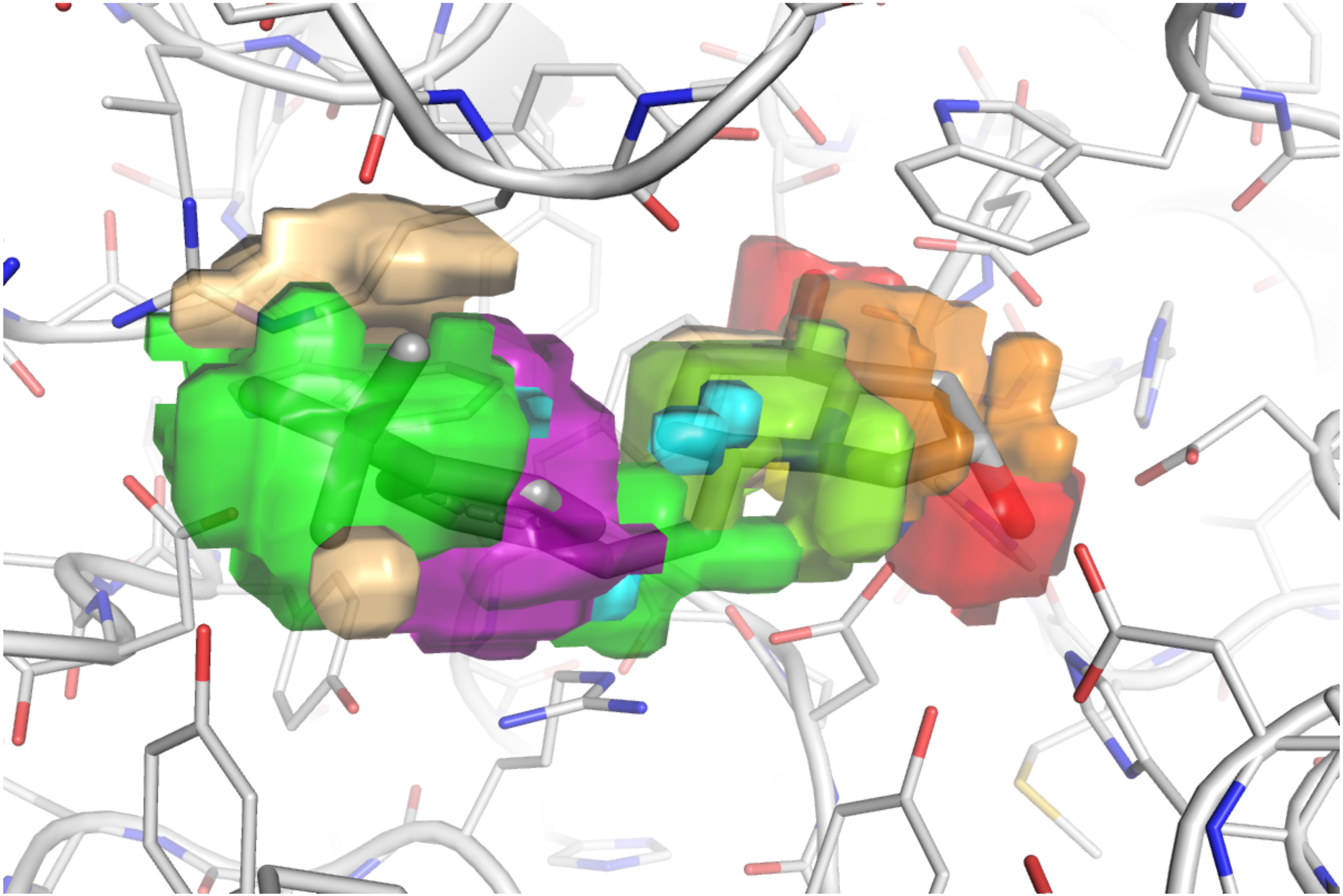


Gradients

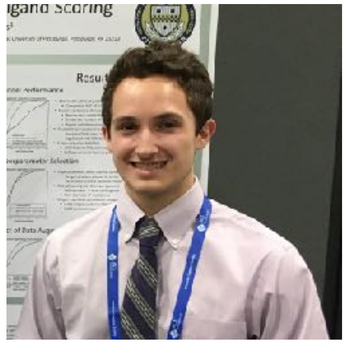




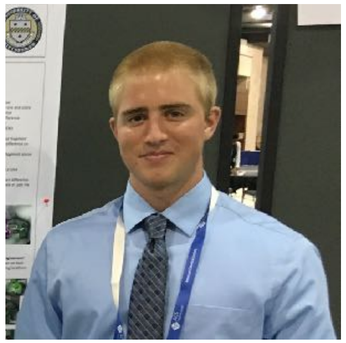
Relevance



Acknowledgements

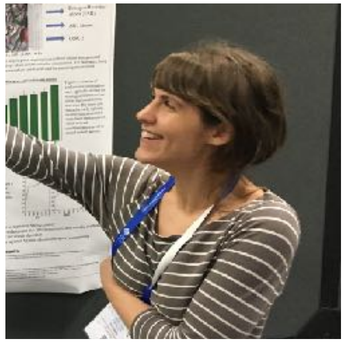


Matt Ragoza



Josh Hochuli

Elisa Idrobo



Jocelyn Sunseri

Group Members

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

Department of Computational and Systems Biology

TECBIO REU

National Institute of General Medical Sciences
R01GM108340