# AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery

Izhar Wallach, Michael Dzamba, Abraham Heifets

Nithin Kannan

CS 371 Presentation

February 1, 2018

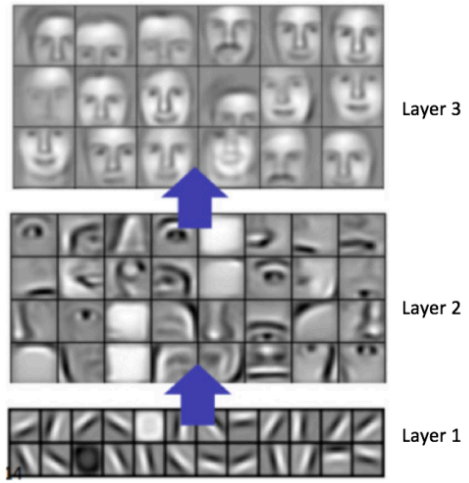
# Current Approaches to Drug Discovery

- Structure based algorithms look at the structure of target proteins to design ligands that bind
  - ➢Structure based algorithms have too many false positive examples
  - ➢These are expensive to experimentally verify

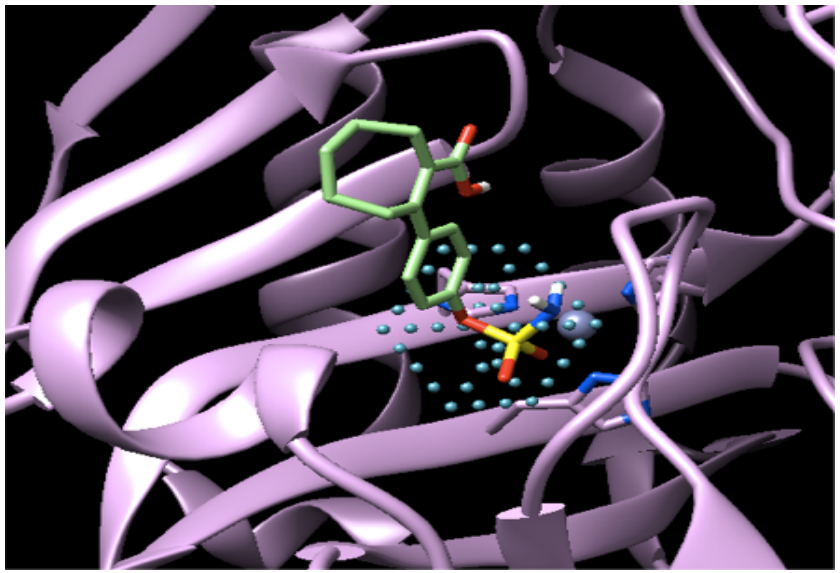
- ML algorithms attempt to learn motifs and structural similarities between ligands to predict ligand receptor interactions
  - ➢Training set not sufficient for accurate ligand based ML algorithms

# AtomNet

- Convolutional Neural Networks recognize the local structures and patterns that make successful ligands
- The neural network is able to pick up on aspects such as hydrogen bonding and aromaticity




Lee et al. 2009 A visualization of increasingly complex layers of a CNN



Wallach et al. 2015 Sulfonyl detection

# AtomNet Architecture

- The input to AtomNet is a 3D input convolved over a stack of hundreds of filters
  - The input can be thought of as an image of the 3D structure with each grid cell associated with a structure vector
- The output is a probability describing whether the ligand will bind strongly to the target protein
- The model has four convolutional layers and two fully connected hidden layers with 1024 neurons each
  - $128*5\nearrow 3$ , $256*3\nearrow 3$ , $256*3\nearrow 3$ , $256*3\nearrow 3$ (number of filters, filter dimension)

# The Directory of Useful Decoys Enhanced (DUDE) Dataset

- The dataset contains a diverse set of active molecules (ligands) for certain target sets of proteins
  - For every active molecule, there is a set of property matched decoys, that are inactive

- Similar active molecules are removed by clustering using scaffold similarity in order to reduce analogue bias when training/testing AtomNet
  - Scaffold Similarity: structural similarity (number of ring and chain atoms)

- Used for training and testing of AtomNet

# The ChEMBL-20 PMD Dataset

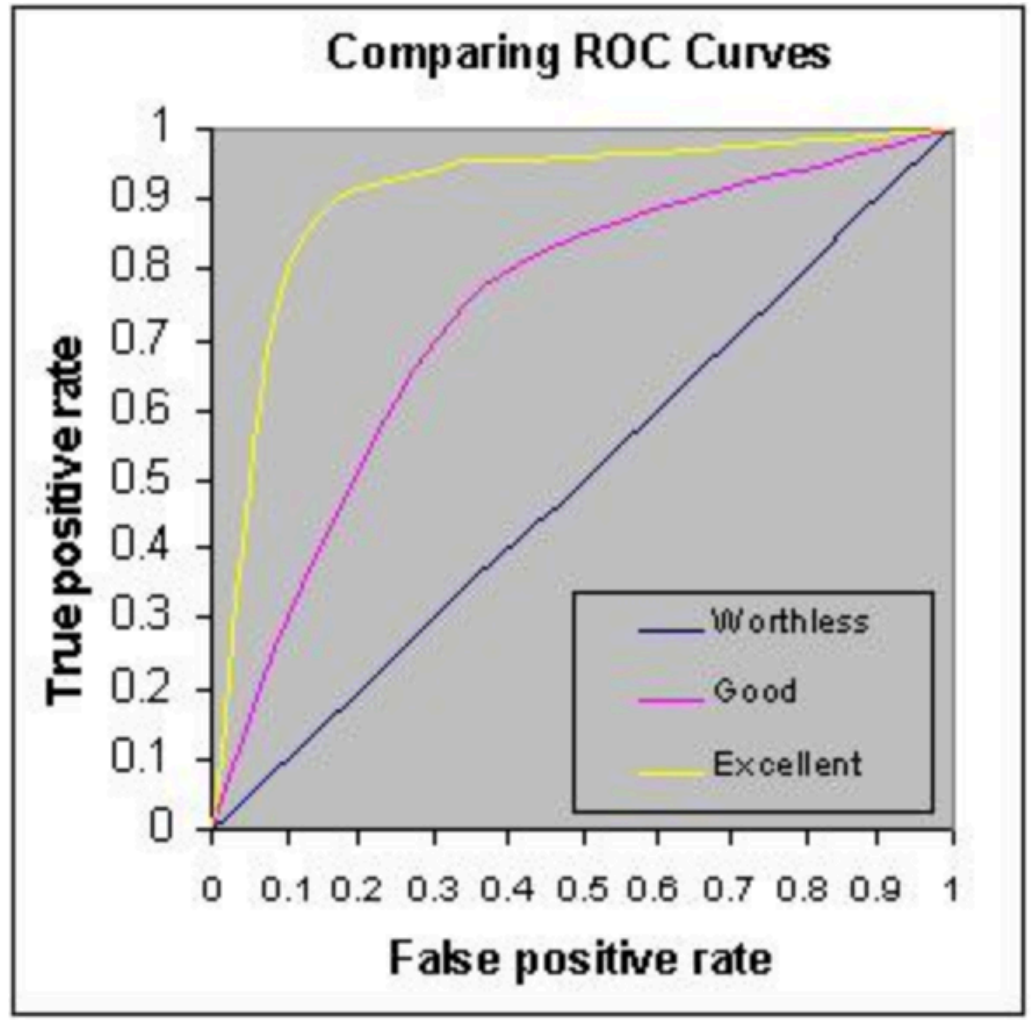
- The dataset is a set of active ligands compiled by the European Molecular Biology Laboratory
- Constructed in a manner analogous to the DUDE dataset
- 30 Property Matched Decoys (PMD) associated with each active molecule
- Bemis-Murcko scaffolds were used to cluster the molecules to avoid analogue bias

# Experimentally Verified Inactives

- The problem with PMD datasets is that they require decoys to have a sufficiently different 2D fingerprint
- This allows the construction of many decoys without expensive experimental validation
- However, this in itself creates a bias in what our model is learning
- Therefore, these experimentally verified inactives served to provide a more representative dataset and force our model to properly classify activity on adversarial examples

# AUC metric

- This curve looks at the true positive rate and false positive rate at different binding thresholds
- Ideally, we would like to have our model only predict true positives, giving us an AUC of 1
- A similarly important metric is the log AUC curve, which places more emphasis on areas of the AUC curve with lower false positive rates




UNMC 2010: Varying AUC curves

# Results of the AtomNet Model

- The receiver operating characteristic is a graph of the true positive rate vs the false positive rate.

- Discovering true positives with as few false positives as possible streamlines the drug discovery process

| | | AUC | | Adjusted logAUC | |
|---|---|---|---|---|---|
| | | Mean | Median | Mean | Median |
| ChEMBL-20 PMD | AtomNet | 0.781 | 0.792 | 0.317 | 0.328 |
| | Smina | 0.552 | 0.544 | 0.04 | 0.021 |
| DUDE-30 | AtomNet | 0.855 | 0.875 | 0.321 | 0.355 |
| | Smina | 0.7 | 0.694 | 0.153 | 0.139 |
| DUDE-102 | AtomNet | 0.895 | 0.915 | 0.385 | 0.38 |
| | Smina | 0.696 | 0.707 | 0.138 | 0.132 |
| ChEMBL-20 inactives | AtomNet | 0.745 | 0.737 | 0.145 | 0.133 |
| | Smina | 0.607 | 0.607 | 0.054 | 0.044 |

Wallach et al. 2015 AtomNet's performance on various datasets

# Strengths

- The convolutional layers pick up on local chemical structures, using interactions between these to notice increasingly complex relations in our model
- The model is robust, working on a variety of different proteins
- Does well compared to other structure based models on the DUDE dataset

# Limitations

- No attempt made to alter hyperparameters of the model
- The code is proprietary, meaning others can’t improve the model
- Did not try different input representations
- The model doesn’t use physics

# Learning Deep Architectures for Interaction Prediction in Structure-based Virtual Screening

Critique by Daniel Hsu

# Virtual Screening

- Virtual screening - computational drug discovery
- Identifies candidates for ligands to bind proteins
- Structure-based: Use binding capacity and structure

# Difficulties

- Complexity of chemical space: $10^{60}$ [1]
- Commercially available compounds: $10^{7}$ [2]
- High false positive rate of identified ligands [3]
- Limited datasets for structure-based virtual screening

[1] RS Bohacek, C McMartin, and WC Guida. The art and practice of structure-based drug design: A molecular modeling perspective. Medicinal research reviews, 16(1):3–50, 1996.
[2] JJ Irwin, T Sterling, MM Mysinger, ES Bolstad, and RG Coleman. Zinc: a free tool to discover chemistry for biology. Journal of chemical information and modeling, 52(7):1757–1768, 2012.
[3] Deng, N.; Forli, S.; He, P.; Perryman, A.; Wickstrom, L.; Vijayan, R. S.; Tiefenbrunn, T.; Stout, D.; Gallicchio, E.; Olson, A. J.; Levy, R. M. Distinguishing Binders from false positives by free energy calculations: fragment screening against the flap site of HIV protease. J. Phys. Chem. B 2015, 119, 976–988.

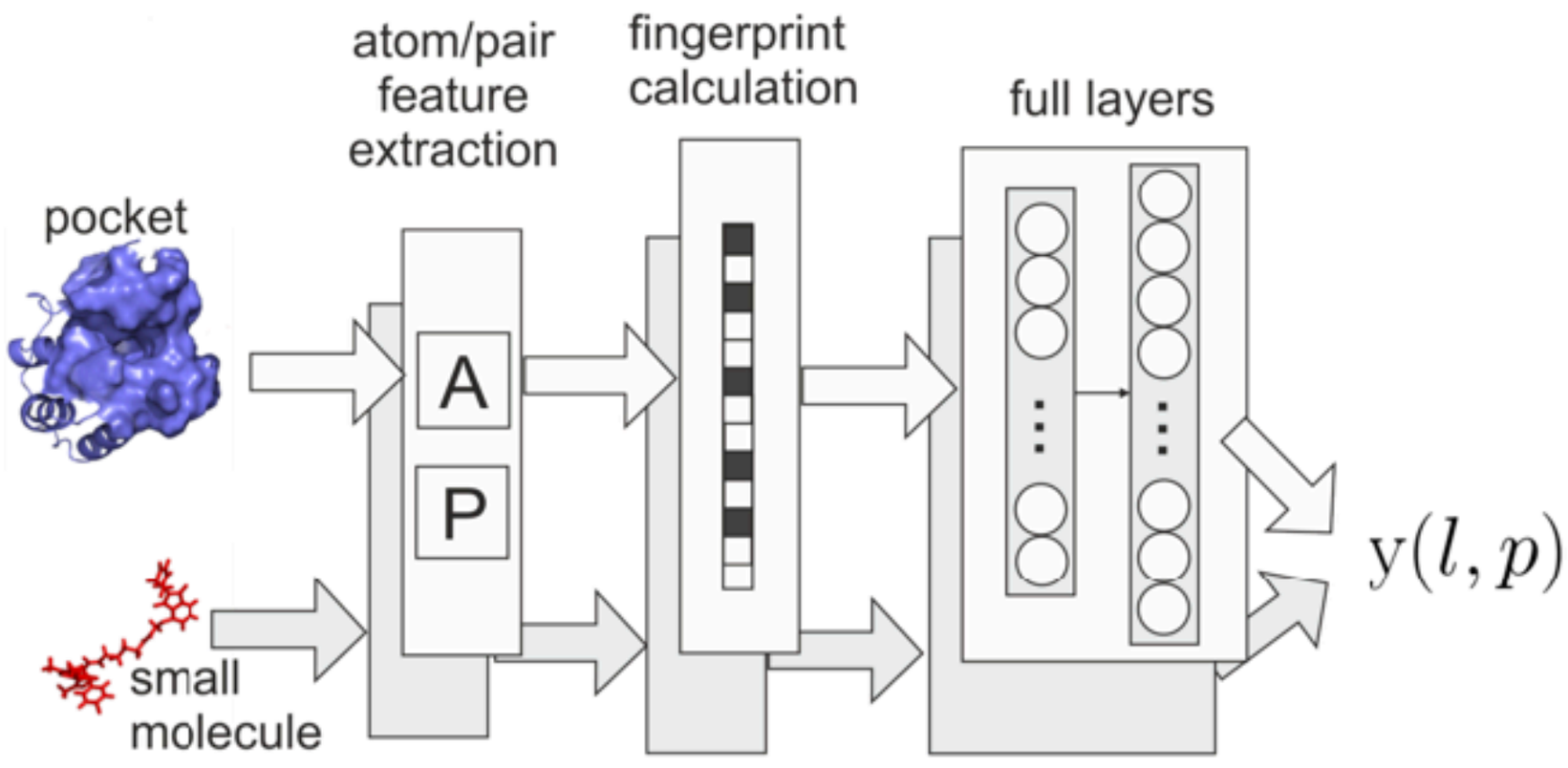
# Approach

- Use deep learning for structure-based virtual screening
- Predict binding capacity of protein-molecule pair
- Propose new benchmark dataset more suitable for structure-based virtual screening

# Model Pipeline

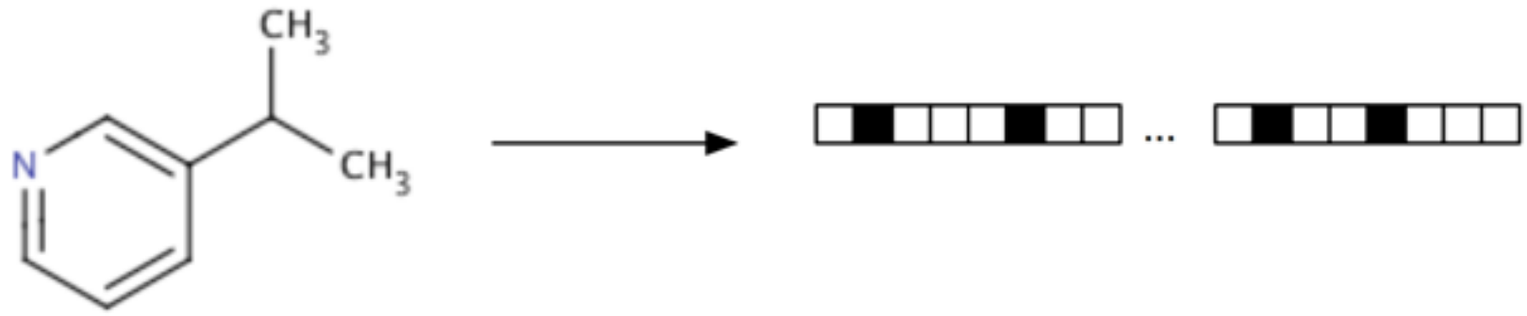
- Process protein and small molecule separately into two fixed-size descriptions (fingerprint vectors)

- Use neural nets to further transform fingerprints



# Ligand Representation: Fingerprints

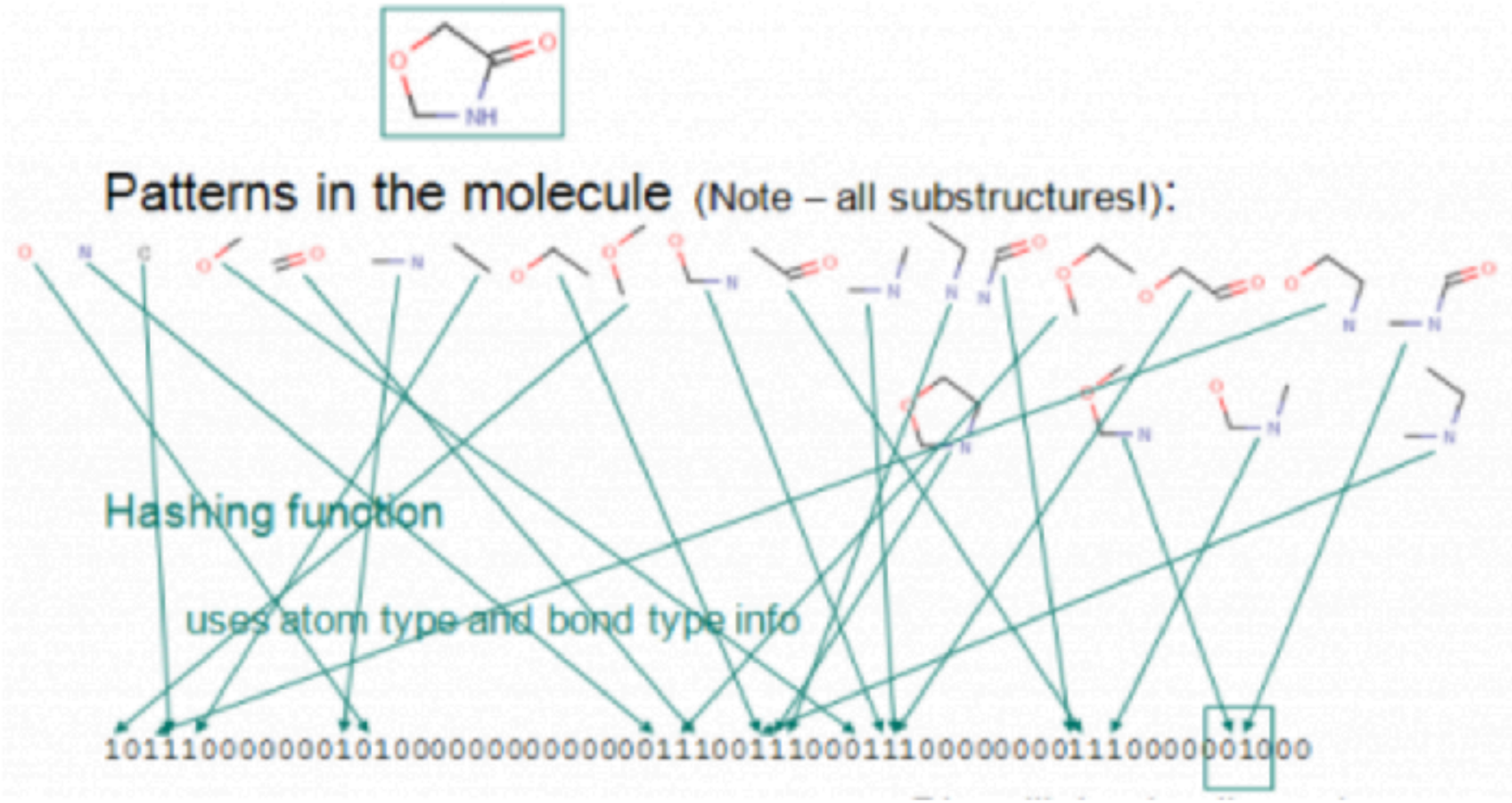
- Convert molecule into binary vector
- 1D representation of a molecule



# ECFP Fingerprints

- Hashing - hash concatenated features of neighborhood of atom
- Indexing - Set 1 into index of feature vector
- Sensitive to small perturbations in molecular structure

# ECFP Fingerprints



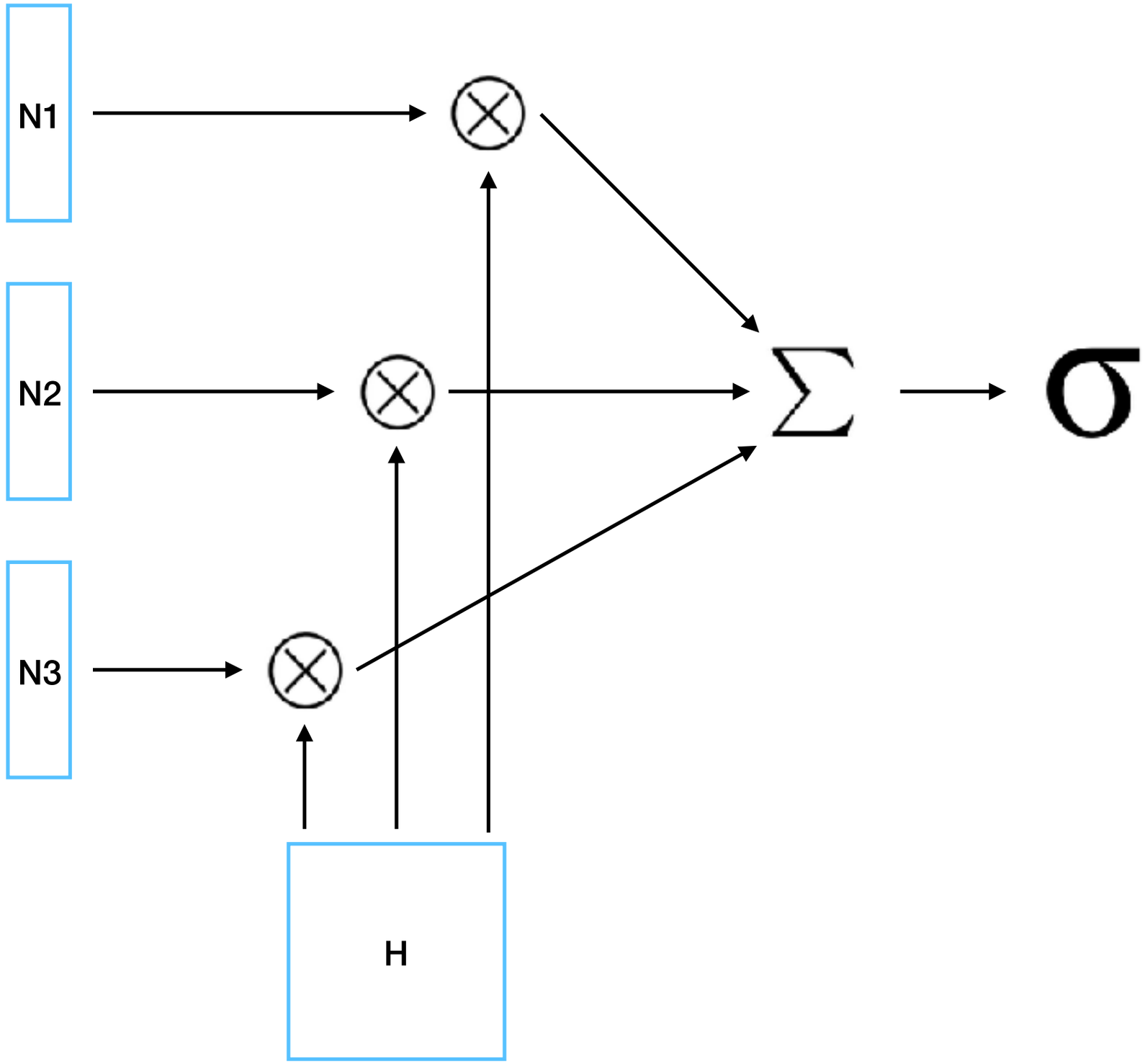

https://chemaxon.com/products/screen-suite

# Atom Convolution Fingerprints

- Initialize vector representation of each with its element, connectivity, number of hydrogen bonds, etc

- Update vector with convolution of weight matrix on neighbor atoms

- Apply non-linearity to updated vector

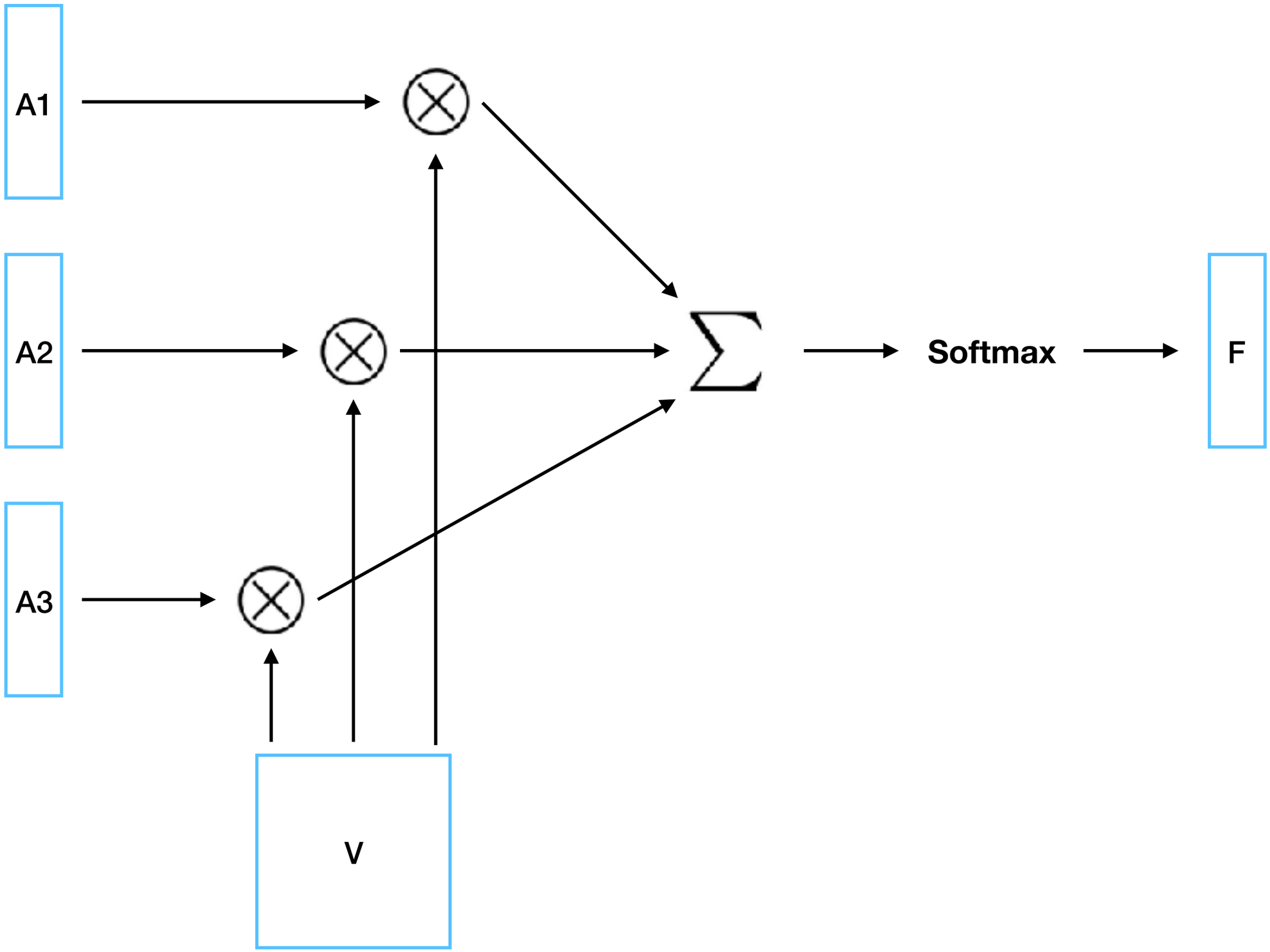

N1
N2
N3
H
∑
σ

# Atom Convolution Fingerprints

- Obtain fingerprint of ligand by convolving another matrix with the final atom vectors to obtain combined sum

- Traditional ECFP fingerprint is binary vector, so approximate this with a softmax operation

- Softmax also makes this operation differentiable


A1
A2
A3
V
Softmax
F

# DUD-E experiment

- Used model to classify active ligands vs decoys using ECFP only: 0.904 AUC

- Perhaps dataset mostly contains information on differences between ligands and decoys, instead of interactions between ligand and proteins

- Suspicious because model makes conclusions on binding of ligands and proteins without using protein information

# PDBBind + DUD-E

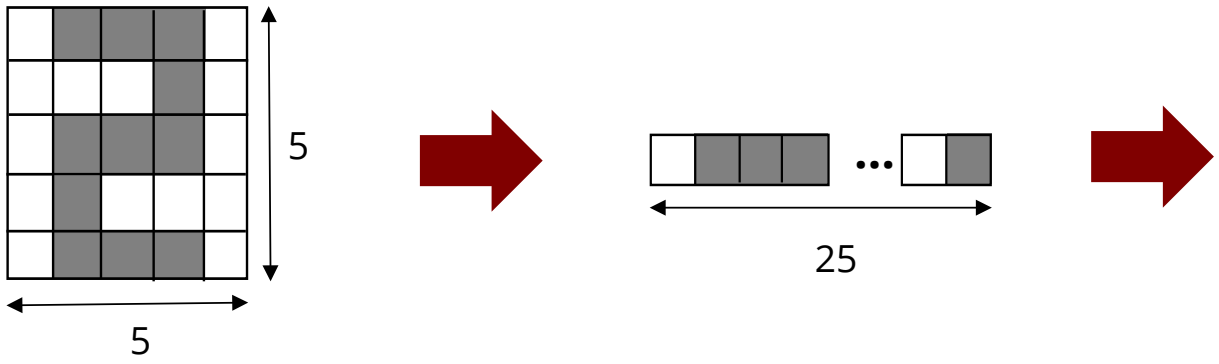
- Use PDBBind for training and DUD-E for testing
- No decoys, negative examples from random sampling
- Fingerprint with network achieved 0.714 AUC, better than fingerprint with ECFP and other algorithms.

# Conclusion

- Deep learning architecture for predicting binding potential
- Form fingerprints using atom convolution instead of ECFP
- Propose new benchmark with PDBBind and DUD-E
- Criticism: Deep learning may not be “cheating” if it can make predictions on binding potential using only ligand information

# Machine learning for structure-based virtual screening

# What are Neural Networks?

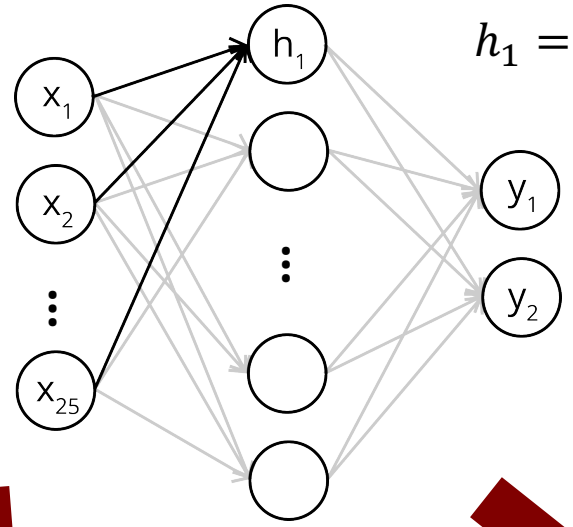


# What are Neural Networks?

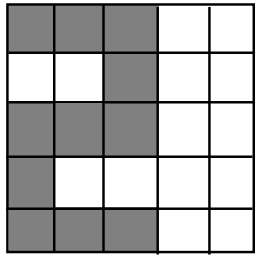






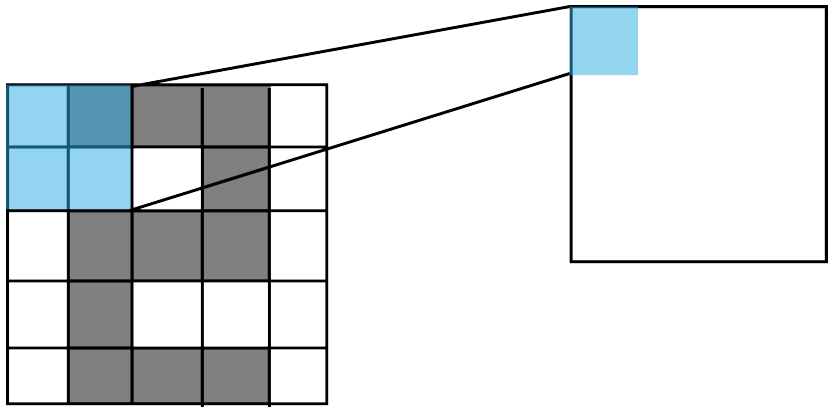


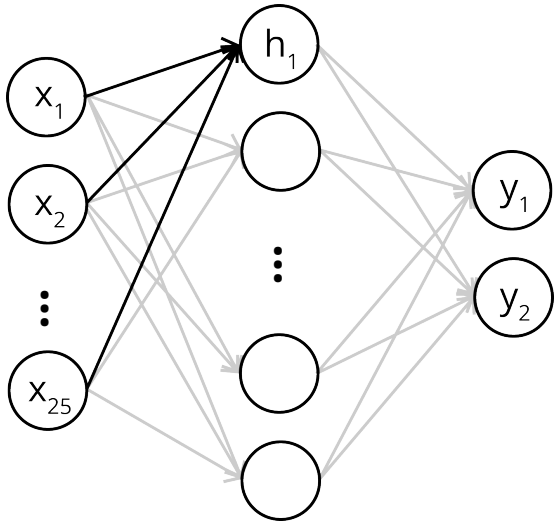

$$h_1 = \sigma(\sum_{i}^{25} w_{i1}x_i + b_1)$$



# Convolutional Neural Networks

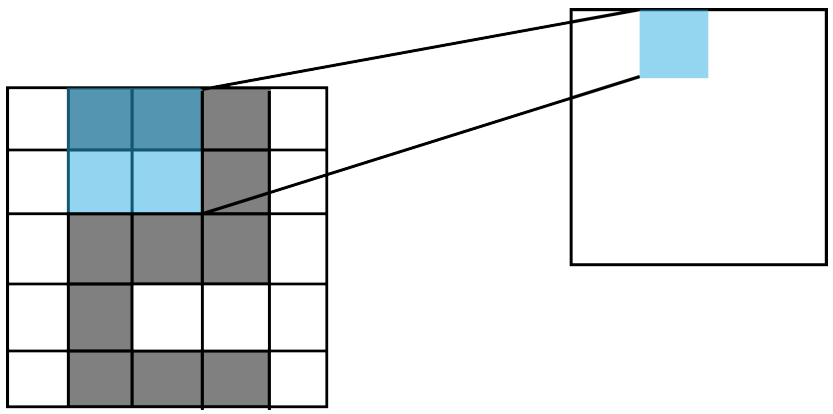
Local connectivity in 2D space. Similar to human vision.






Krizhevsky, A. et al. Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012)

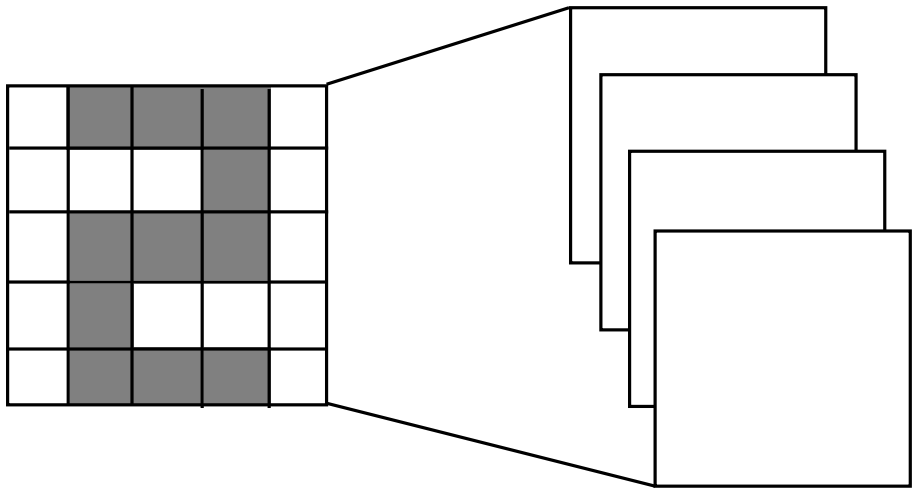
# Convolutional Neural Networks



Shares the same weights and biases. Identifies a certain pattern and is translationally independent.

Krizhevsky, A. et al. Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012)

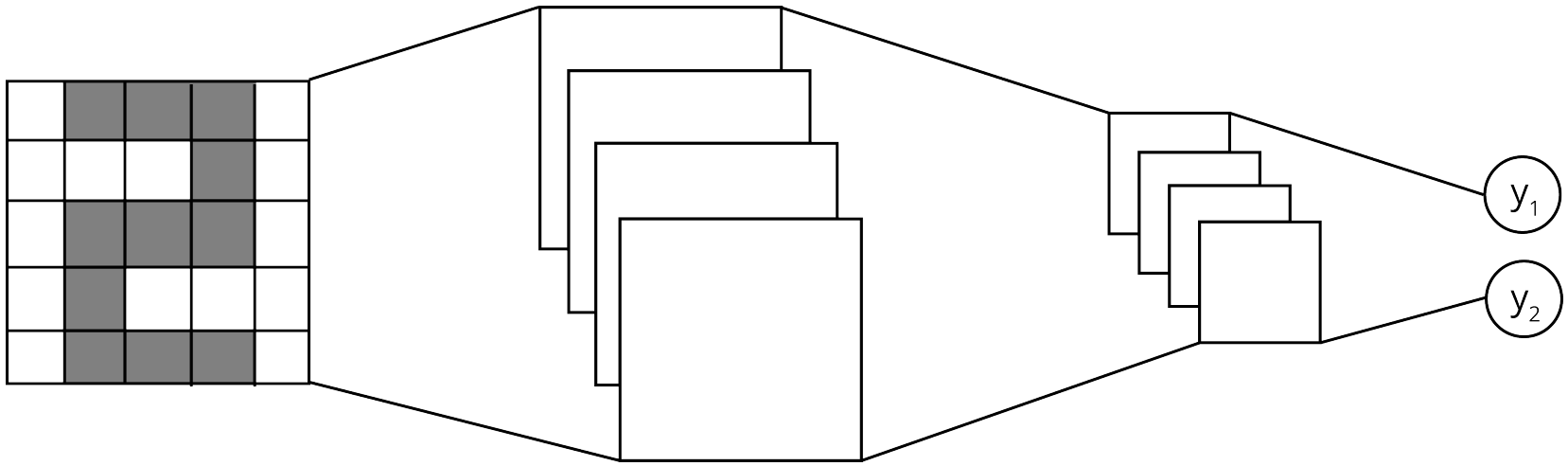
# Convolutional Neural Networks



Each feature map learns to identify a certain pattern in the image

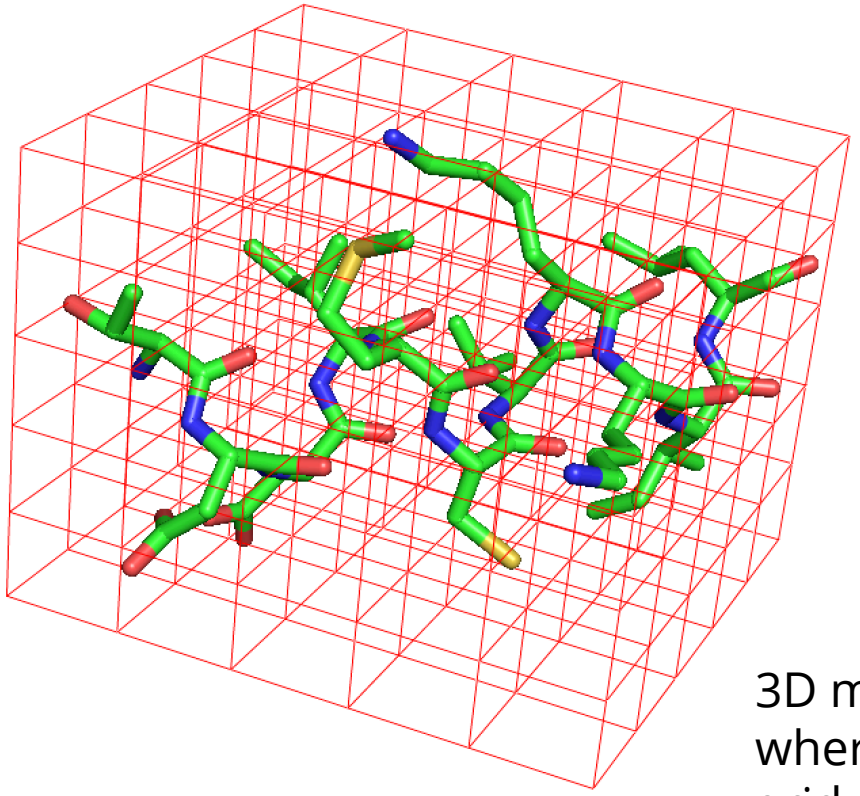
Krizhevsky, A. et al. Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012)

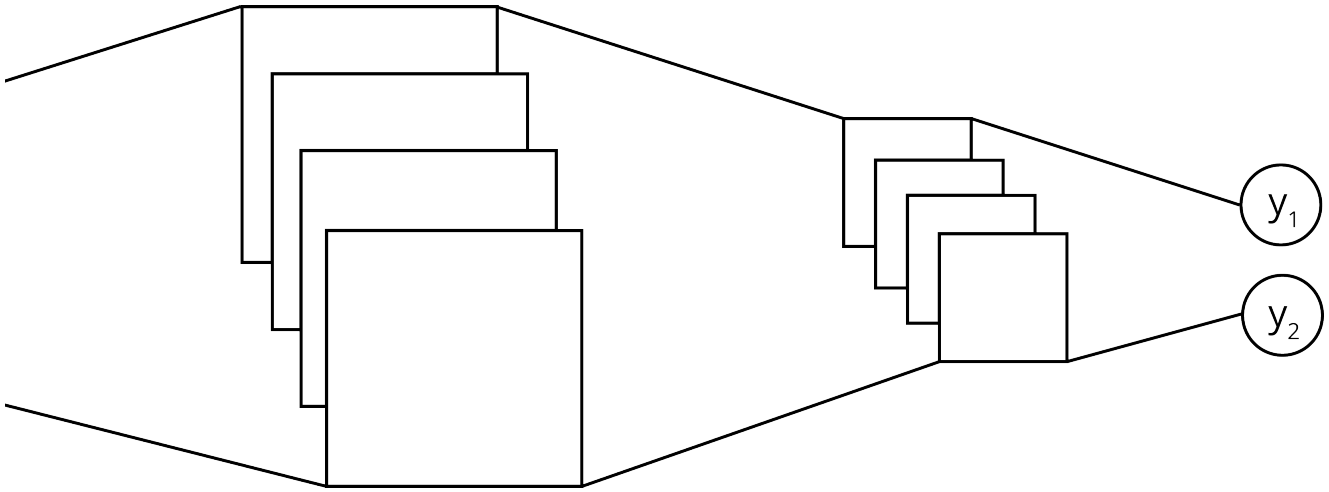
# Convolutional Neural Networks




Krizhevsky, A. et al. Proc. Advances in Neural Information Processing Systems 25 1090–1098 (2012)

# Convolutional Neural Networks





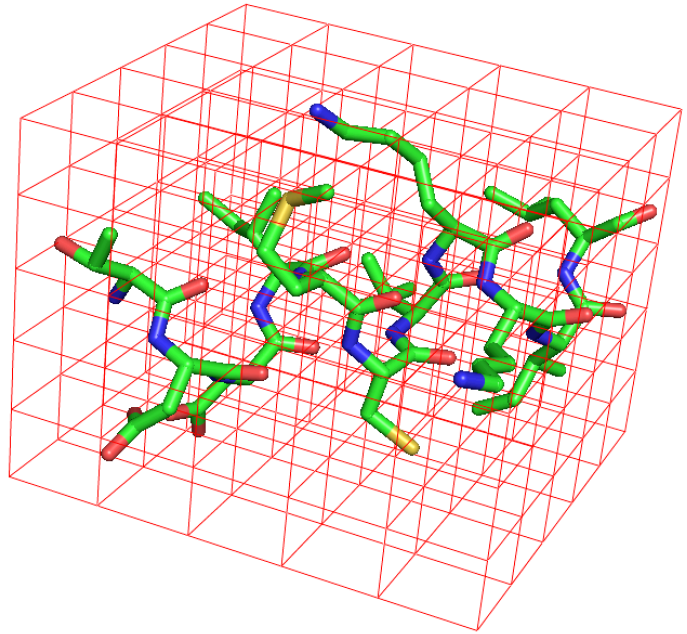

3D mesh of a molecule where each point in the grid contains information about the atom types.

# Protein-Ligand Scoring with Convolutional Neural Networks

Ragoza M. et al. J. Chem. Inf. Model. 57 942−957 (2017)

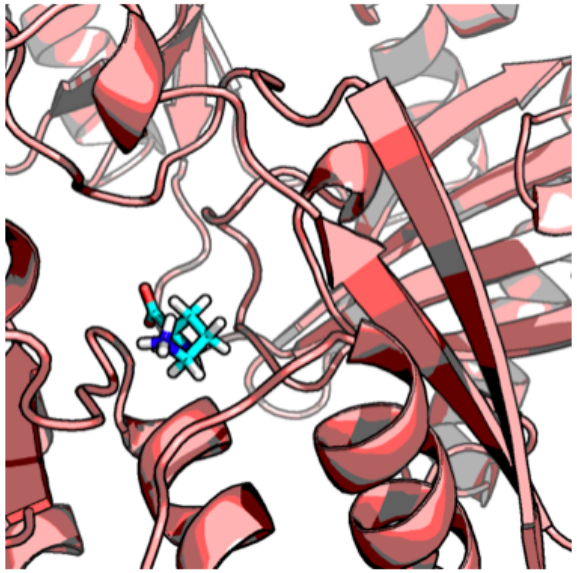
# Improvements from AtomNet

- More detailed methodology.
- Includes descriptors which better describe the binding site, includes aromaticity, protonation state etc.
- Improved visualization method enabling mutation analysis.
- Stringent model evaluation to check for overfitting.

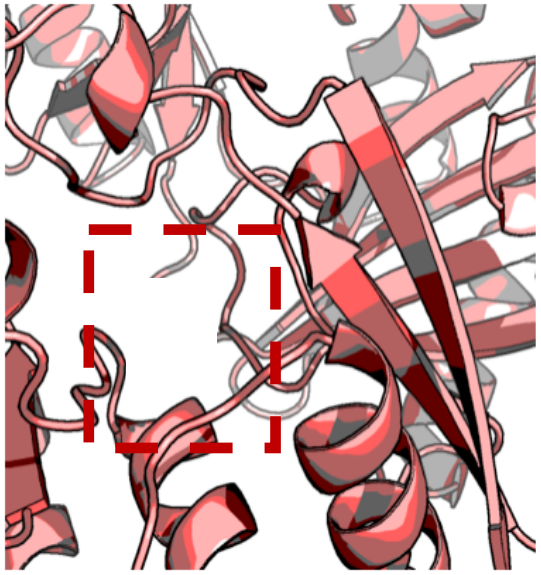


# Generating a training set

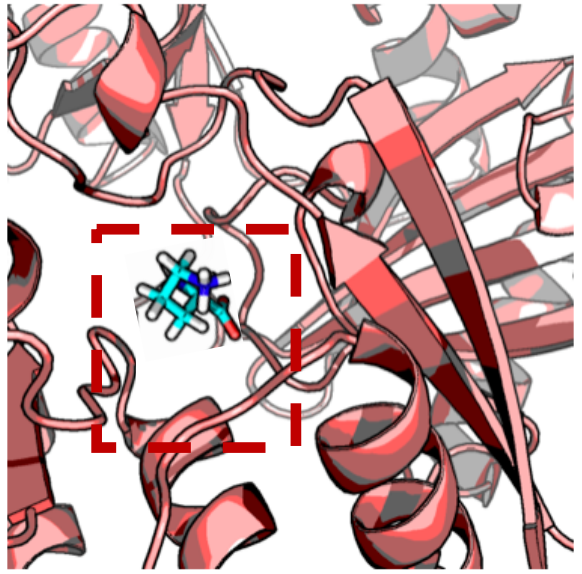
DUD-E is a dataset with 102 proteins, 200,00 ligands and over 1 million decoy molecules. However, not all co-crystal structures are available, but reference complexes are given.



The ligand on the reference receptor is removed and a box is drawn from a 8 Å around the ligand.



Ligands and decoys are docked using smina and the Autodock Vina scoring function.

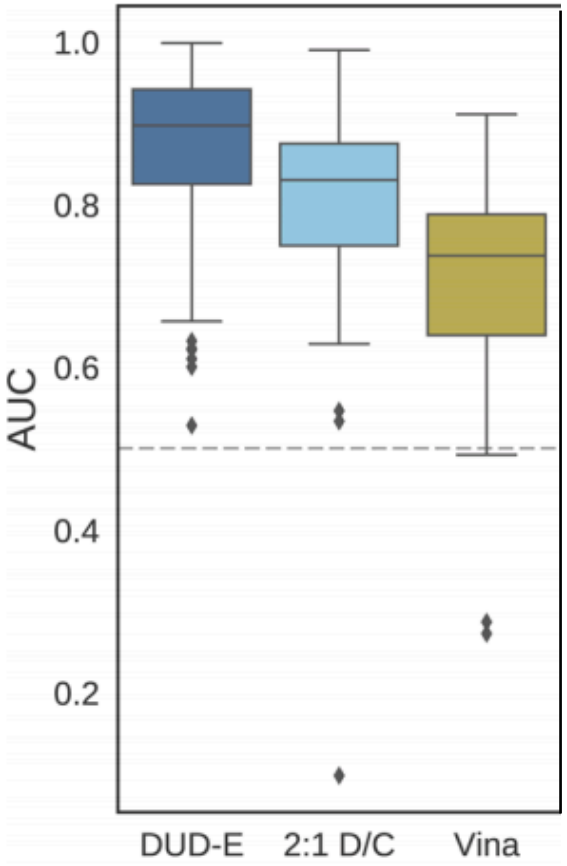


# About Autodock Vina and smina

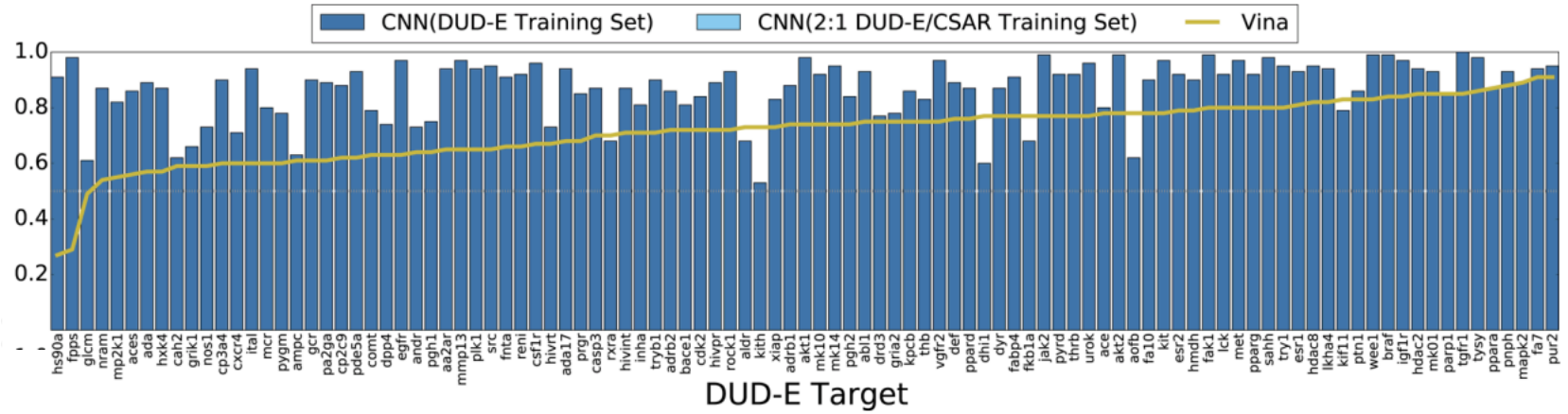
- Given a bounded box, smina will run a docking algorithm to generate protein-ligand conformations.
- Each structure is then given a score, where the score is constructed from pairwise interactions including sterics, hydrogen bonding etc. These parameters are optimized using the PDBbind dataset.
- Assumptions
  - Receptor is rigid.
  - Protonation state of atoms remain unchanged.

Trott et al. J Comput Chem. 31 455-461 (2010)

# Results



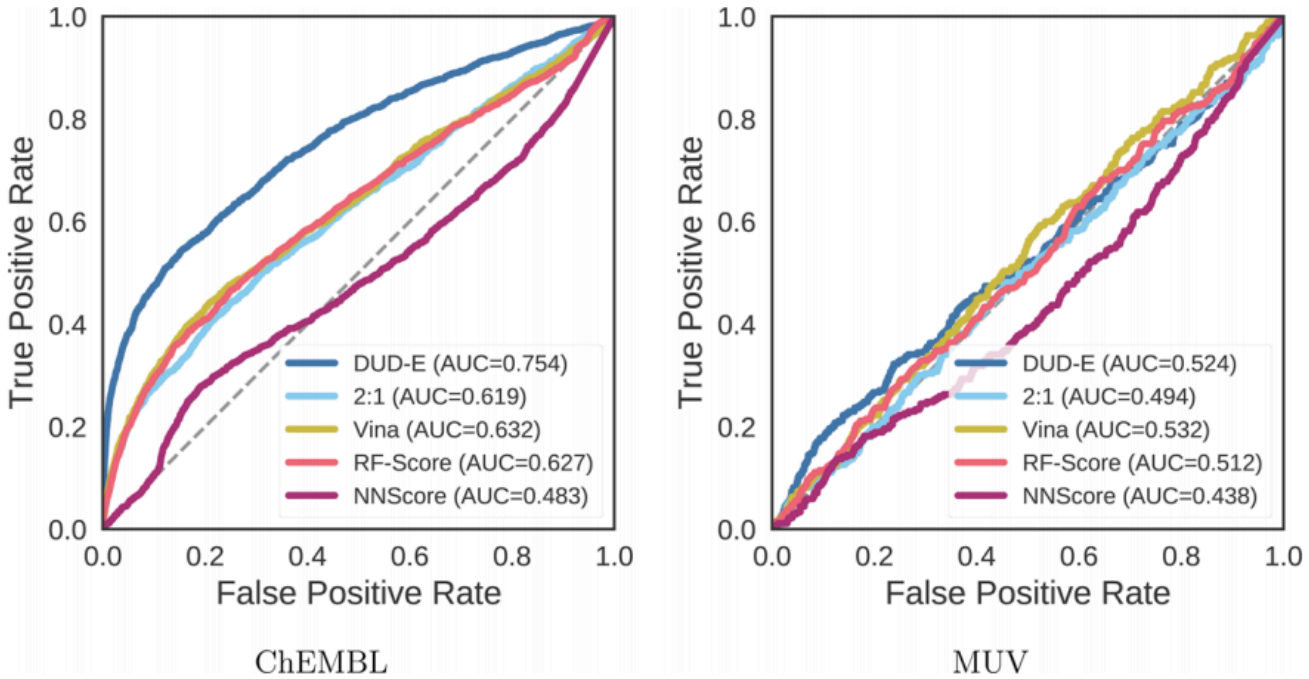

Outperforms Vina on 90% of the targets.




Ragoza M. et al. J. Chem. Inf. Model. 57 942−957 (2017)

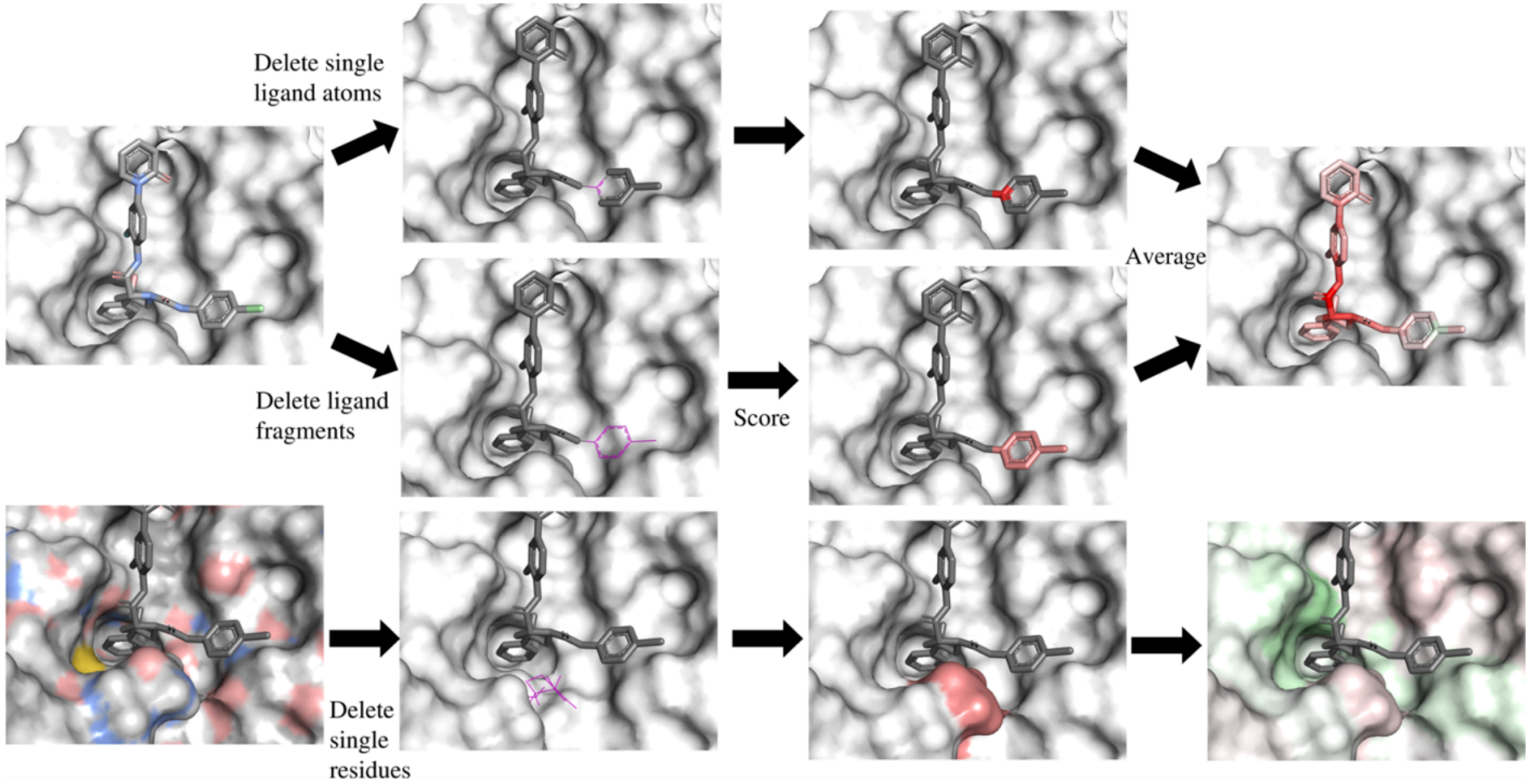
# Independent Test Sets

- How much is our model overfitting our data?




Ragoza M. et al. J. Chem. Inf. Model. 57 942−957 (2017)

# Visualization




Ragoza M. et al. J. Chem. Inf. Model. 57 942−957 (2017)

# Limitations

- Ligand screening takes Autodock Vina prediction of ligand conformation as the “truth”.
- Assumption that receptors only have one binding site.
- Co-crystal structures are not always readily available.
- Entropy and enthalpy not fully encapsulated within the descriptors.
- Feature maps are potentially interpretable.
- Paper also carried out pose prediction, but the performance was around the same as Vina.