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

Izhar Wallach, Michael Dzamba, Abraham Heifets
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
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 3, 256*3 3, 256*3 3, 256*3 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
Comparing ROC Curves

True positive rate vs False positive rate

- Worthless
- Good
- Excellent

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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AtomNet AUC Mean</th>
<th>AtomNet AUC Median</th>
<th>Smina AUC Mean</th>
<th>Smina AUC Median</th>
<th>Adjusted logAUC Mean</th>
<th>Adjusted logAUC Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChEMBL-20 PMD</td>
<td>0.781</td>
<td>0.792</td>
<td>0.552</td>
<td>0.544</td>
<td>0.317</td>
<td>0.328</td>
</tr>
<tr>
<td>DUDE-30</td>
<td>0.855</td>
<td>0.875</td>
<td>0.7</td>
<td>0.694</td>
<td>0.321</td>
<td>0.355</td>
</tr>
<tr>
<td>DUDE-102</td>
<td>0.895</td>
<td>0.915</td>
<td>0.696</td>
<td>0.707</td>
<td>0.385</td>
<td>0.38</td>
</tr>
<tr>
<td>ChEMBL-20 inactives</td>
<td>0.745</td>
<td>0.737</td>
<td>0.607</td>
<td>0.607</td>
<td>0.145</td>
<td>0.133</td>
</tr>
</tbody>
</table>

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

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

Patterns in the molecule (Note – all substructures!):

Hashing function

uses atom type and bond type info

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
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.
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?

\[ h_1 = \sigma\left(\sum_{i} w_{i1}x_i + b_1\right) \]
What are Neural Networks?

\[ h_1 = \sigma \left( \sum_{i}^{25} w_{i1} x_i + b_1 \right) \]
Convolutional Neural Networks

Local connectivity in 2D space. Similar to human vision.
Convolutional Neural Networks

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

Convolutional Neural Networks

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

Convolutional Neural Networks

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

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 steric, hydrogen bonding etc. These parameters are optimized using the PDBbind dataset.
• Assumptions
  • Receptor is rigid.
  • Protonation state of atoms remain unchanged.
Results

Outperforms Vina on 90% of the targets.
Independent Test Sets

• How much is our model overfitting our data?

Visualization

Delete single ligand atoms

Delete ligand fragments

Delete single residues

Average

Score
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.