AtomNet

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

Izhar Wallach, Michael Dzamba, Abraham Heifets, 2015

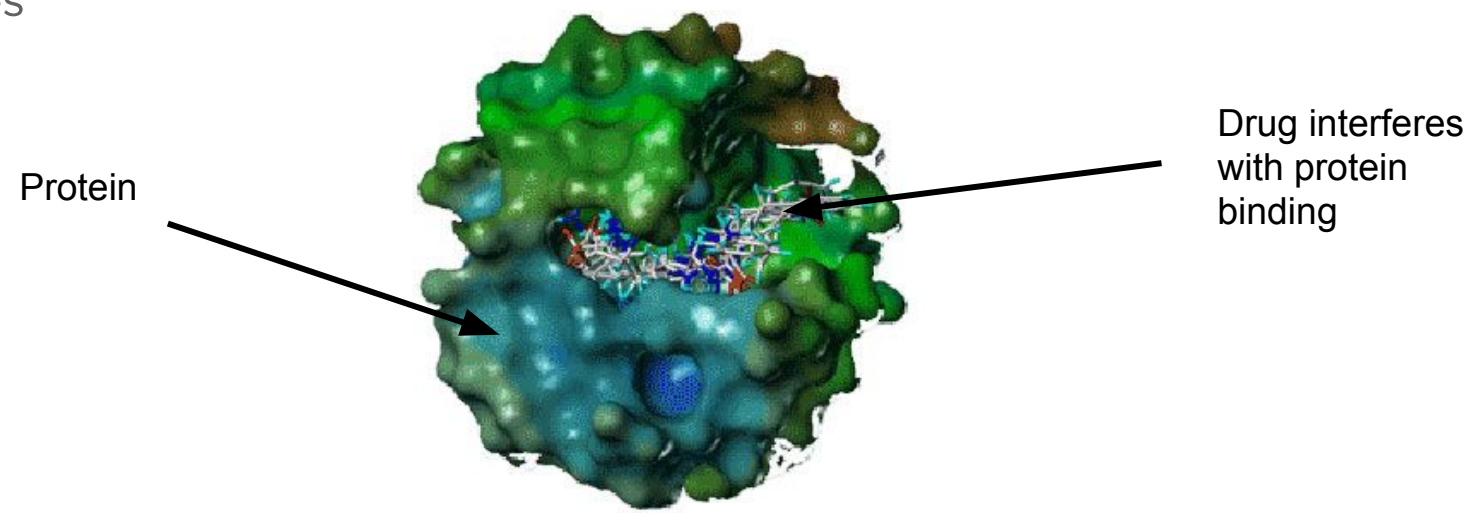
Presented by: Alex Barron & Abubakar Abid

Questions that we will explore

- What is **bioactivity prediction** and why is it important?
- What techniques existed before AtomNet, and why does it make sense to try **CNN's**?
- What's the difference between **ligand**-based learning and **structure**-based learning?
- What's the best dataset to use to test AtomNet?
- What architecture does AtomNet use?
- How well does AtomNet perform, compared to other state-of-the-art techniques?
- Do the **convolution filters** carry any interpretive significance?
- What architectural or experimental improvements exist for future work?

Bioactivity prediction tries to learn protein-ligand interactions

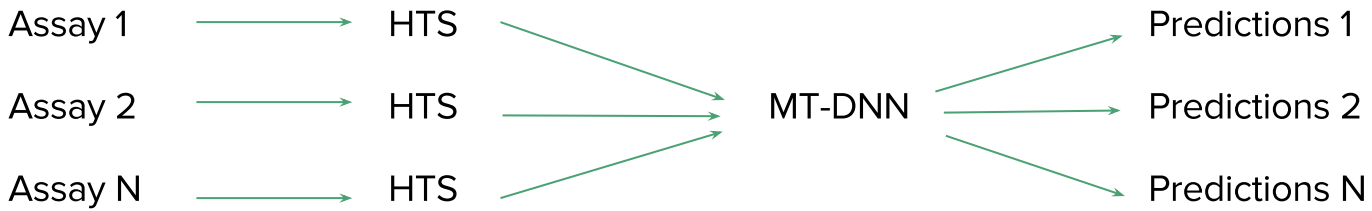
“Fundamentally, biological systems operate through the physical interaction of molecules”



If we could predict protein-ligand interaction, we could use that to design medications based only on the knowledge of a biological target

Deep learning performs better than SVMs and random forests

In 2014, research from Salakhutdinov group showed that a multi-task DNN could perform far better than existing techniques at quantitative structure-activity relationships (QSAR) competition organized by Merck, achieving an accuracy 15% higher than the next closest team.



MT-DNNs have now been generalized to larger databases, but they come with certain limitations because they are **ligand-based approaches**

AtomNet proposes a structural (not ligand-based) approach

Ligand Approach

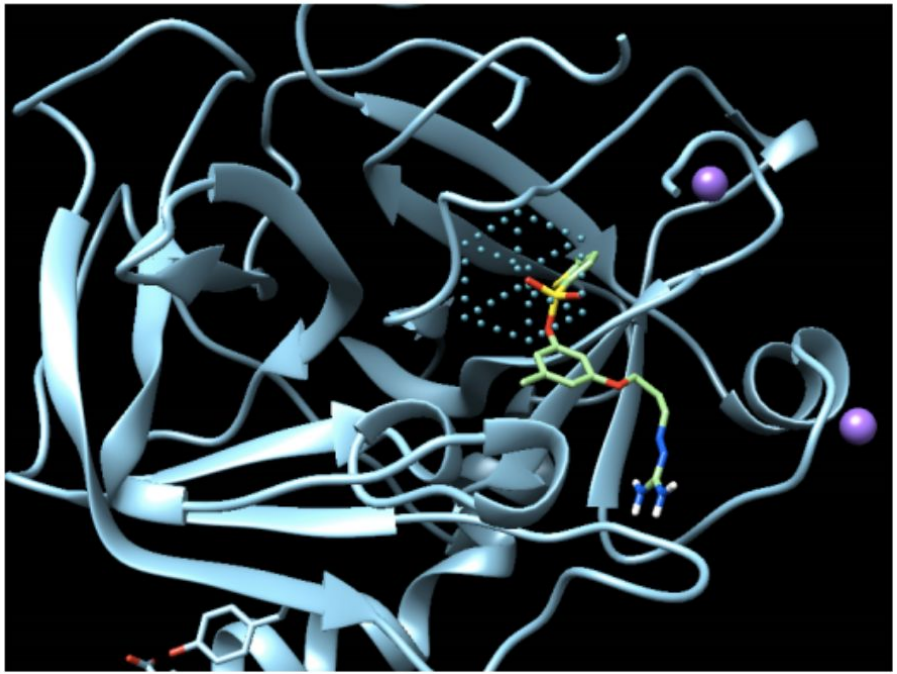
- Uses as input prior data from high-throughput screens/assays
- Discovers relationships between ligand (but not protein target) structure and affinity
- Input features are ‘molecular fingerprints’ (ECFP) of the ligand

Structural Approach

- Requires information about the binding site in the protein
- Is more resistant to artifacts and biases, which may be present in a set of assays
- Input features are 3D ‘snapshots’ of target/ligand binding site

“In practice, this creates a paradoxical dynamic – [ligand-based] models offer the most help precisely for those targets which least require it”

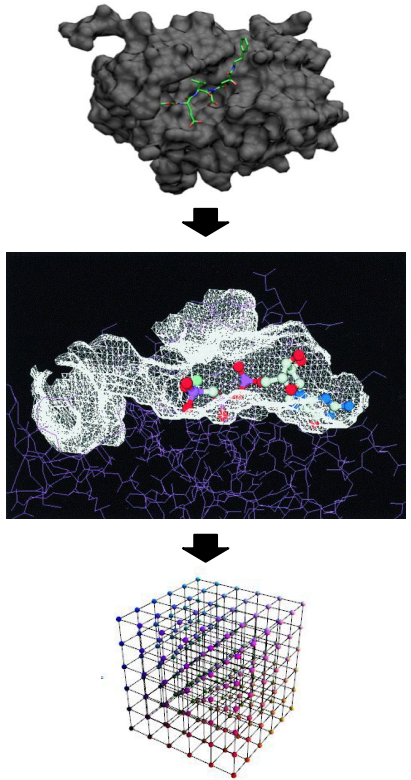
AtomNet proposes a 3D-CNN architecture. Why?



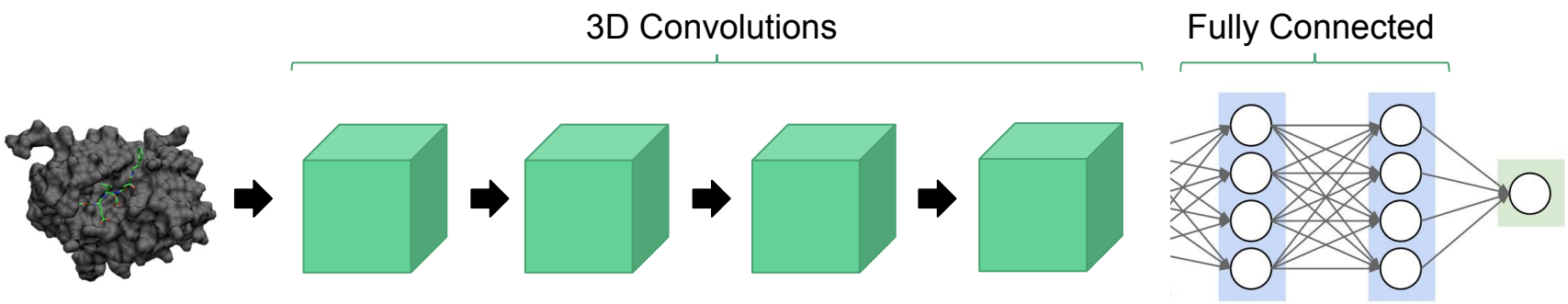
- Molecular interactions are generally **local** in 3D space
- Higher-order features (e.g. hydrogen bonds) are built **hierarchically** from simpler elements
- Higher-order features are **spatially invariant** and can be described with the same type of simpler features

AtomNet uses a structure based input representation

- Binding site located using annotated bound ligands in the scPDB database
- Coordinates shifted to 3D cartesian with center-of-mass of binding site as origin
- Multiple poses (orientations) of the small molecule and binding site sampled
- Values cropped to 20x20x20 Å grid and translated features to 1 Å spacing
- Each grid cell holds a value representing a structural feature



3D convolutions extract local features



Kernel		5x5x5	3x3x3	3x3x3	3x3x3		
Num Filters/ Hidden Size		128	256	256	256	1024	1024	1

Architecture trained end to end

- AdaDelta SGD
- Mini-batches of 768 examples
- One week of training

“No attempt was made to optimize meta-parameters”

What dataset to use to test AtomNet's proposed architecture?

- A standardized dataset that can be used to bench Atomnet against other predictors

- An internal dataset that ensures no overlaps between molecules in the training set and test set

- A 'challenge set' that consists of **topologically similar** negatives discovered through experiment

Directory of Useful Decoys Enhanced (**DUDE**) consists of 102 targets, each with ~224 binders, and *property matched, topologically dissimilar* 50 decoys (negatives) for each positive

ChEMBL-20 PMD, a dataset that was prepared similar to DUDE, but ligands among the test, training, and validation had to be a certain distance apart

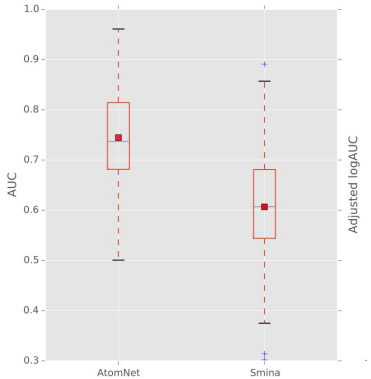
EVI, a dataset of experimentally verified inactives. This consisted of 290 targets, with about the same number of binders and 5 decoys per binder

AtomNet outperforms state of the art commercial and open source docking algorithms

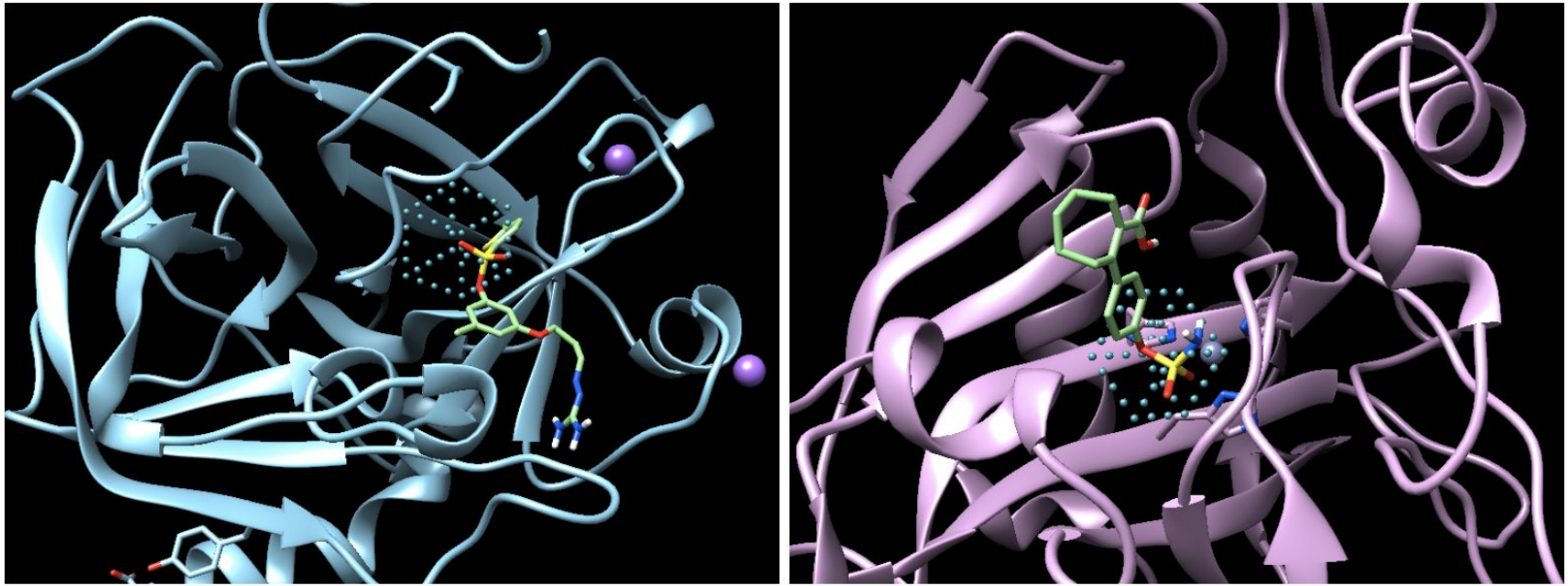
Baseline Approach

- **Smina** - linear regression with carefully chosen structural features, open source
- **Dock3.7** - linear
- **No explicit comparison to DNN methods**

		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



Convolutional filters learn complex chemical features

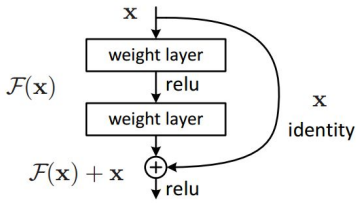


Two main avenues for future work

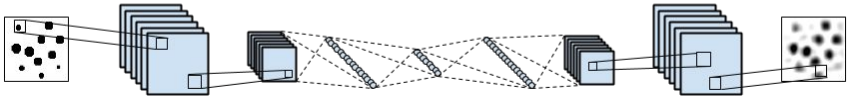
- Improve model architecture
- Improve datasets

Possible architectural improvements

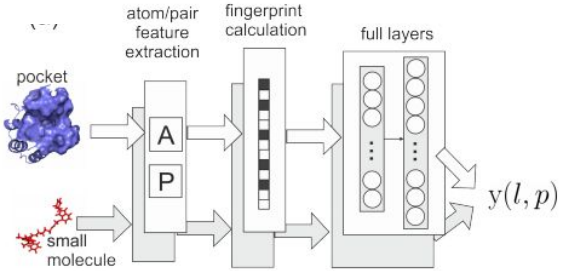
- Residual networks



- Down sampling with stride 2 layers and convolutional auto-encoders



- Separate embeddings for target protein and small molecule using SVD or GloVe word vectors



Possible dataset improvements

- Large unsupervised datasets available for embedding approach
  E.g. PubChem, DrugBank, ChEMBL
- Combine negative sampling from existing structural datasets (such as PDBBind) with DUDE targets to remove bias from artificially generated decoys