Protein-Ligand Scoring with Convolutional Neural Networks

David Koes
@david_koes

ACS
August 23, 2017
Structure Based Drug Design

Virtual Screening

Pose Prediction

Binding Discrimination

Lead Optimization

Affinity Prediction
Structure Based Drug Design

Virtual Screening

Lead Optimization

Pose Prediction

Binding Discrimination

Affinity Prediction
Protein-Ligand Scoring

AutoDock Vina

\[ \text{gauss}_1(d) = w_{\text{gauss}_1}e^{-(d/0.5)^2} \]
\[ \text{gauss}_2(d) = w_{\text{gauss}_2}e^{-(d-3)/2} \]
\[ \text{repulsion}(d) = \begin{cases} 
  w_{\text{repulsion}}d^2 & d < 0 \\
  0 & d \geq 0 
\end{cases} \]
\[ \text{hydrophobic}(d) = \begin{cases} 
  w_{\text{hydrophobic}} & d < 0.5 \\
  0 & d > 1.5 \\
  w_{\text{hydrophobic}}(1.5 - d) & \text{otherwise} 
\end{cases} \]
\[ \text{hbond}(d) = \begin{cases} 
  w_{\text{hbond}} & d < -0.7 \\
  0 & d > 0 \\
  w_{\text{hbond}}(-\frac{10}{7}d) & \text{otherwise} 
\end{cases} \]

State of the Art

Pose Prediction  Binding Discrimination  Affinity Prediction

Can we do better?

Accurate pose prediction, binding discrimination, **and** affinity prediction without sacrificing performance?
Can we do better?

Accurate pose prediction, binding discrimination, and affinity prediction without sacrificing performance?

**Key Idea:** Leverage “big data”
- 231,655,275 bioactivities in PubChem
- 125,526 structures in the PDB
- 16,179 annotated complexes in PDBbind
Machine Learning

Features $X \rightarrow$ Model $\rightarrow y$ Prediction
Neural Networks

\[
\text{output} = \sigma \left( \sum_i w_i x_i + b \right)
\]
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

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

$$\frac{\partial L}{\partial w^l_{jk}} = a_{k}^{l-1} \delta^l_j$$ and $$\frac{\partial L}{\partial b^l_j} = \delta^l_j$$
Image Recognition

ILSVRC top-5 error on ImageNet

Convolutional Neural Networks

https://devblogs.nvidia.com
Convolutional Neural Networks

Convolution

Convolution Feature Maps

Fully Connected Traditional NN

Dog: 0.99
Cat: 0.02
CNNs for Protein-Ligand Scoring

- 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.
Atom Density

\[
A(d, r) = \begin{cases} 
  e^{-\frac{2d^2}{r^2}} & 0 \leq d < r \\
  \frac{4}{e^2r^2} d^2 - \frac{12}{e^2} d + \frac{9}{e^2} & r \leq d < 1.5r \\
  0 & d \geq 1.5r 
\end{cases}
\]

Gaussian
Atom Types

**Ligand**
- Aliphatic Carbon XSHydrophobe
- Aliphatic Carbon XSNonHydrophobe
- Aromatic Carbon XSHydrophobe
- Aromatic Carbon XSNonHydrophobe
- Bromine
- Chlorine
- Fluorine
- Iodine
- Nitrogen
- Nitrogen XSAcceptor
- Nitrogen XSDonor
- Nitrogen XSDonorAcceptor
- Oxygen
- Oxygen XSAcceptor
- Oxygen XSDonorAcceptor
- Phosphorus
- Sulfur
- Sulfur Acceptor

**Receptor**
- Aliphatic Carbon XSHydrophobe
- Aliphatic Carbon XSNonHydrophobe
- Aromatic Carbon XSHydrophobe
- Aromatic Carbon XSNonHydrophobe
- Calcium
- Iron
- Magnesium
- Nitrogen
- Nitrogen XSAcceptor
- Nitrogen XSDonor
- Nitrogen XSDonorAcceptor
- Oxygen XSAcceptor
- Oxygen XSDonorAcceptor
- Phosphorus
- Sulfur
- Zinc
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
  - 8,688 <2Å RMSD (actives)
  - 76,743 >4Å RMSD (decoys)
Data Augmentation

![Graph showing AUC over training iterations with different augmentation strategies.]

- **AUC**
- **Training Iterations**

Legend:
- train - no augmentation
- train - random rotation & translation
- test - random rotation & translation
- test - no augmentation
Data Augmentation

AUC vs training iterations

- train - no augmentation
- train - random rotation & translation
- test - random rotation & translation
- test - no augmentation
Model Evaluation

**CSAR:** >90% similar targets kept in same fold

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

![AUC Plot](image)
Pose Prediction (CSAR)
Pose Prediction (CSAR)

inter-target ranking

intra-target ranking
Pose Prediction (PDBbind)

**inter-target ranking**

- **True Positive Rate**
  - CNN (AUC=0.94)
  - Vina (AUC=0.62)

- **False Positive Rate**
  - 0.0 0.2 0.4 0.6 0.8 1.0

**intra-target ranking**

- **% Low-RMSD Poses**
  - Top-1: Random, CNN, Vina
  - Top-3: Random, CNN, Vina
  - Top-5: Random, CNN, Vina
Visualization

masking

gradients

layer-wise relevance
Visualizing Enzymes

PDB ID: 1YZ3
Wild Type $K_i$: 1.55 nM
E219A $K_i$: 1375 nM
D267A $K_i$: 999 nM

PDB ID: 3C3U
Wild Type $K_i$: 5.9 nM
L308A $K_i$: 2800 nM

PDB ID: 2DOR
Wild Type $K_i$: $2.9 \times 10^4$ nM
N193A $K_i$: $1.16 \times 10^7$ nM
Pose Sensitivity

Partially Aligned Poses
Beyond Scoring

data 48^3

unit1_pool

unit1_conv1 32 x 24^3

unit2_pool

unit2_conv1 64 x 12^3

unit3_pool

unit3_conv1 128 x 6^3

output_fc 2

output

label

loss
Beyond Scoring

\[
\delta^l = \left( (w^{l+1})^T \delta^{l+1} \right) \odot \sigma' (z^l)
\]

\[
\frac{\partial L}{\partial w^l_{jk}} = a^{-1} \delta^l_j \quad \text{and} \quad \frac{\partial L}{\partial b^l_j} = \delta^l_j
\]
Beyond Scoring

\[
\delta^l_{\text{unit}} = \left( (w^{l+1})^T \delta^{l+1} \right) \bigotimes \sigma'(z^l)
\]

\[
\frac{\partial L}{\partial w^l_{jk}} = a^{l-1} \delta^l_j \quad \text{and} \quad \frac{\partial L}{\partial b^l_j} = \delta^l_j
\]

output

loss

data

48\times3

unit1\_pool

unit1\_conv1
32 \times 24\times3

\sigma'(z^l)

output\_fc
2

unit3\_conv1
128 \times 6\times3

label

unit1\_pool
Beyond Scoring

2Q89

More Oxygen Here

Less Oxygen Here
Beyond Scoring

\[ \frac{\partial L}{\partial A_{\{x,y,z\}}} = \sum_j \frac{\partial L}{\partial x_j} \frac{\partial x_j}{\partial \text{dist}_{A,x_j}} \frac{\partial \text{dist}_{A,x_j}}{\partial A_{\{x,y,z\}}} \]

More Oxygen Here

Less Oxygen Here

2Q89
Minimizing Low RMSD Poses

The graph illustrates the distribution of RMSD change in poses, with the x-axis representing RMSD change and the y-axis showing the number of poses. The graph is divided into two categories: 'Better' and 'Worse', with the bars showing the number of poses that are closer to a best pose (gray) and the first minimization attempt (blue). The figure visually compares how many poses are improved or worsened during the minimization process.
Iterative Refinement

![Graph showing the change in RMSD with number of poses, highlighting the difference between Best and First Minimization.](image-url)
Iterative Refinement

RMSD Change

# Poses

-4 -3 -2 -1 0 1 2 3 4

Best
First Minimization
Second Iteration
Iterative Refinement

![Graph showing iterative refinement with histogram of poses and RMSD Change]
Minimizing Random Poses

![Graph showing RMSD Change vs. # Poses for Best and Vina methods.]

- Better poses are represented by a peak at lower RMSD values.
- Worse poses spread out across a wider range of RMSD values.

RMSD Change

# Poses
Minimizing Random Poses

![RMSD Change Distribution](image)

- **Best**
- **Vina**
- **Initial CNN**

- **better**
- **worse**
Minimizing Random Poses

![Graph showing RMSD Change and number of poses, with categories for better and worse performance.](image)

- Best
- Vina
- Initial CNN
- Final CNN
The Future

Pose Selection

Iterative Training

Pose Generation

Iterative Training

Compound Generation

Virtual Screening

Lead Optimization
The Future

Pose Selection  →  Pose Generation  →  Compound Generation

Virtual Screening  

Iterative Training

Graph: True Positive Rate vs. False Positive Rate

- CNN (AUC=0.930)
- Vina (AUC=0.633)
Related Work

MolecuLeNet: A continuous–filter convolutional neural network for modeling quantum interactions
Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, Klaus–Robert Müller
(Submitted on 26 Jun 2017)

Automatic chemical design using a data–driven continuous representation of molecules
Rafael Gómez–Bombarelli, David Duvenaud, José Miguel Hernández–Lobato, Jorge Aguilera–Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, Alán Aspuru–Guzik
(Submitted on 7 Oct 2016 (v1), last revised 6 Jan 2017 (this version, v2))

AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure–based Drug Discovery
Izhar Wallach, Michael Dzamba, Abraham Heifets
(Submitted on 10 Oct 2015)

ANI–1: An extensible neural network potential with DFT accuracy at force field computational cost
Justin S. Smith, Olexandr Isayev, Adrian E. Roitberg
(Submitted on 27 Oct 2016 (v1), last revised 6 Feb 2017 (this version, v4))

Convolutional Networks on Graphs for Learning Molecular Fingerprints
David Duvenaud, Dougal Maclaurin, Jorge Aguilera–Iparraguirre, Rafael Gómez–Bombarelli, Timothy Hirzel, Alán Aspuru–Guzik, Ryan P. Adams
(Submitted on 30 Sep 2015 (v2), last revised 3 Nov 2015 (this version, v2))

Atomic Convolutional Networks for Predicting Protein–Ligand Binding Affinity
Joseph Gomes, Bharath Ramsundar, Evan N. Feinberg, Vijay S. Pande
(Submitted on 30 Mar 2017)

Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug–Like Molecules
Alessandro Lusci†, Gianluca Polastritt, and Pierre Baldi†
1 School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
2 Department of Computer Science, University of California, Irvine, Irvine, California 92697, United States
DOI: 10.1021/ci400187y
Publication Date (Web): June 04, 2013

Low Data Drug Discovery with One–shot Learning
Han Altae–Tran, Bharath Ramsundar, Aneesh S. Pappu, Vijay Pande
(Submitted on 10 Nov 2016)

Massively Multitask Networks for Drug Discovery
Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, Vijay Pande
(Submitted on 6 Feb 2015)
OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.

STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS/DEEP LEARNING!

SIX MONTHS LATER: WOW, THIS PROBLEM IS REALLY HARD. YOU DON'T SAY.
Acknowledgements

**Group Members**
- Jocelyn Sunseri
- Matt Ragoza
- Josh Hochuli
- Pulkit Mittal
- Alec Helbling
- Tamar Skaist
- Christopher Dunstan

**Department of Computational and Systems Biology**

**National Institute of General Medical Sciences**
R01GM108340

**NVIDIA**