GPU-Accelerated Convolutional Neural Networks For Protein-Ligand Scoring

David Koes

GPU Technology Conference
May 8, 2017
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

<table>
<thead>
<tr>
<th>BASIC RESEARCH</th>
<th>DRUG DISCOVERY</th>
<th>PRE-CLINICAL</th>
<th>CLINICAL TRIALS</th>
<th>FDA REVIEW</th>
<th>POST-APPROVAL RESEARCH &amp; MONITORING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PHASE I</td>
<td>PHASE II</td>
<td>PHASE III</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TENS</td>
<td>HUNDREDS</td>
<td>THOUSANDS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NUMBER OF VOLUNTEERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TENS</td>
<td>HUNDREDS</td>
<td>THOUSANDS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

POTENTIAL NEW MEDICINES

1 FDA-APPROVED MEDICINE

$2.6 BILLION

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

If you stop failing so often you massively reduce the cost of drug development.
— Sir Andrew Witty
CEO, GlaxoSmithKline

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

If you stop failing so often you massively reduce the cost of drug development.
— Sir Andrew Witty, CEO, GlaxoSmithKline

$2.6 BILLION

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)
1. Does the compound do what you want it to?
2. Does the compound *not* do what you *don’t* want it to?
3. Is what you want it to do the right thing?
Protein Structures

sequence → structure → function
Protein Structures

sequence $\rightarrow$ structure $\rightarrow$ function
Structure Based Drug Design

Unlike ligand based approaches, generalizes to new targets

Requires molecular target with known structure and binding site
Structure Based Drug Design

Unlike ligand based approaches, generalizes to new targets

Requires molecular target with known structure and binding site
Structure Based Drug Design

Unlike ligand based approaches, generalizes to new targets.

Requires molecular target with known structure and binding site.
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
Protein-Ligand Scoring

AutoDock Vina

\[
gauss_1(d) = w_{\text{gauss}_1} e^{-(d/0.5)^2}
\]
\[
gauss_2(d) = w_{\text{gauss}_2} e^{-(d-3)/2^2}
\]
\[
\text{repulsion}(d) = \begin{cases} 
  w_{\text{repulsion}} d^2 & d < 0 \\
  0 & d \geq 0 
\end{cases}
\]
\[
\text{hydrophobic}(d) = \begin{cases} 
  w_{\text{hydrophobic}} & d < 0.5 \\
  0 & d > 1.5 \\
  w_{\text{hydrophobic}} (1.5 - d) & \text{otherwise}
\end{cases}
\]
\[
hbond(d) = \begin{cases} 
  w_{\text{hbond}} & d < -0.7 \\
  0 & d \geq 0 \\
  w_{\text{hbond}} (-10/7 d) & \text{otherwise}
\end{cases}
\]

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
Deep Learning
Deep Learning

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

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

ILSVRC top-5 error on ImageNet

- Convolutional Neural Networks

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

Convolution Feature Maps

Fully Connected Traditional NN

Convolution

weight 1
weight 2
weight 3

Fully-connected

weight 1
weight 2
weight 3
weight 4
weight 5

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

- Pose Prediction
- Binding Discrimination
- Affinity Prediction
CNNs for Protein-Ligand Scoring

- Pose Prediction
- Binding Discrimination
- Affinity Prediction
CNNs for Protein-Ligand Scoring

- Input representation
- Training
- Model optimization
- Visualize and Evaluation

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

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

Gaussian
Atom Types

**Ligand**

- AliphaticCarbonXSHydrophobe
- AliphaticCarbonXSNonHydrophobe
- AromaticCarbonXSHydrophobe
- AromaticCarbonXSNonHydrophobe
- Bromine
- Chlorine
- Fluorine
- Iodine
- Nitrogen
- NitrogenXSAcceptor
- NitrogenXSDonor
- NitrogenXSDonorAcceptor
- Oxygen
- OxygenXSAcceptor
- OxygenXSDonorAcceptor
- Phosphorus
- Sulfur
- SulfurAcceptor

**Receptor**

- AliphaticCarbonXSHydrophobe
- AliphaticCarbonXSNonHydrophobe
- AromaticCarbonXSHydrophobe
- AromaticCarbonXSNonHydrophobe
- Calcium
- Iron
- Magnesium
- Nitrogen
- NitrogenXSAcceptor
- NitrogenXSDonor
- NitrogenXSDonorAcceptor
- OxygenXSAcceptor
- OxygenXSDonorAcceptor
- Phosphorus
- Sulfur
- Zinc
337 protein-ligand complexes
- curated for electron density
- diverse targets
- <10µM affinity
- **generate poses** with Vina
  - 745 <2Å RMSD (actives)
  - 3251 >4Å RMSD (decoys)

12,484 protein-ligand complexes
- diverse targets
- wide range of affinities
- **generate poses** with AutoDock Vina
- include minimized crystal pose
  - 24,727 <2Å RMSD (actives)
  - 244,192 >4Å RMSD (decoys)
**Model Evaluation**

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

**PDBbind**: >80% similar targets kept in same fold
Model Training

Custom **MolGridDataLayer**

Parallelize over *atoms* to obtain a mask of atoms that overlap each grid region.
Use exclusive scan to obtain a list of atom indices from the mask.
Parallelize over *grid points*, using reduced atom list to avoid $O(N_{atoms})$ check.

For example, consider subgrid 5:

| Atom mask: | 1 1 0 0 0 |
| Exclusive scan: | 0 1 2 2 2 |
| Final indices: | 0 1 |
Data Augmentation

![Graph showing AUC over training iterations with different augmentation methods.](image)

- `train - no augmentation`
- `train - random rotation & translation`
- `test - random rotation & translation`
- `test - no augmentation`
Data Augmentation

![Graph showing AUC vs. training iterations with different augmentation methods.]
Model Optimization

Atom Types
- Vina (34)
- element-only (18)
- ligand-protein (2)

Atom Density Type
- Boolean
- Gaussian

Radius Multiple

Resolution

Pooling

Depth

Width

Fully Connected Layers

Data

Unit 1_pool

Unit 1_conv1 128 x 24 x 3

Unit 2_pool

Unit 2_conv1 256 x 12 x 3

Unit 3_pool

Unit 3_conv1 512 x 6 x 3

Unit 4_pool

Unit 4_conv1 1024 x 3 x 3

Unit 5_pool

Unit 5_conv1 2048 x 1 x 3

Output_fc 2

Output

Loss
Model Optimization

Cross Validation AUC vs. training time (ms) per iteration

Parameter Evaluated:
- atom types
- pooling type
- depth
- radius multiple
- fully connected layer
- resolution
- atom density type
- width

Reference models:
- model 1
- model 2
- model 3
Model Optimization

Cross Validation AUC vs training time (ms) per iteration

- Parameter Evaluated:
  - atom types
  - pooling type
  - depth
  - fully connected layer
  - radius multiple
  - resolution
  - atom density type
  - width

- Reference models:
  - reference model 1
  - reference model 2
  - reference model 3
Cross-Validation Evaluation
Pose Prediction (CSAR)

- CNN (AUC=0.82)
- Vina (AUC=0.64)
Pose Prediction (CSAR)

inter-target ranking

intra-target ranking

**Graphs:**
- **Left:** ROC curve for inter-target ranking.
- **Right:** Bar graph for intra-target ranking with 30% Random, 80% CNN, and 90% Vina for Top-1, Top-3, and Top-5 ranks.

**Legend:**
- CNN (AUC=0.82)
- Vina (AUC=0.64)
Pose Prediction (PDBbind)

- **True Positive Rate vs. False Positive Rate**
  - CNN (AUC=0.930)
  - Vina (AUC=0.633)

- **% Low-RMSD Poses**
  - Top-1
  - Top-3
  - Top-5
  - Comparison:
    - Random
    - Vina
    - CNN
Pose Prediction (PDBbind)

**inter-target ranking**

**intra-target ranking**
Visualization

Delete single ligand atoms

Delete ligand fragments

Delete single residues

Average

Score
Examples

Partially Aligned Poses
Beyond Scoring
Beyond Scoring

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

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

Data: \( 48^3 \)

Label

Output

Loss
Beyond Scoring

\[
\delta^l_{\text{unit}} = ((w^l + 1)T \delta^{l+1}) \odot \sigma'(z) \\
\frac{\partial L}{\partial w^l_{jk}} = a_k^{l-1} \delta^l_j \text{ and } \frac{\partial L}{\partial b^l_j} = \delta^l_j
\]

output

loss
Beyond Scoring

2Q89

Less Oxygen Here

More Oxygen Here
Beyond Scoring

\[
\frac{\partial L}{\partial A} = \sum_{i \in G_A} \left( \text{data} \frac{\partial L}{\partial G_i} \right) \frac{\partial D}{\partial A} 
\]

2Q89

More Oxygen Here

Less Oxygen Here
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

Lead Optimization

Iterative Training

Pose Generation

Iterative Training

Virtual Screening

Lead Optimization

Graph:

- CNN (AUC=0.930)
- Vina (AUC=0.633)
Acknowledgements

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

National Institute of General Medical Sciences
R01GM108340
Questions?

Binding Determination

Affinity Prediction

Relevance Propagation
Questions?

Binding Determination

Affinity Prediction

Relevance Propagation

Data (48^3)

Label

32 x 24^3

Unit 1 Pool

Unit 2 Pool

Unit 2 Conv1 64 x 12^3

Unit 3 Pool

Unit 3 Conv1 128 x 6^3

Output FC 2

Output

Loss

Multi-pose CNN (AUC=0.864)

Single-pose CNN (AUC=0.851)

Vina (AUC=0.683)

RF-Score (AUC=0.611)

NNScore (AUC=0.582)

False Positive Rate

True Positive Rate

0.2

0.4

0.6

0.8

1.0

0.2

0.4

0.6

0.8

1.0

Experimental Affinity

CNN

R=0.687
RMS=2.186