Structure Based Drug Design

Virtual Screening

Lead Optimization

Pose Prediction

Binding Discrimination

Affinity Prediction
Protein-Ligand Scoring

AutoDock Vina

\[
\text{gauss}_1(d) = w_{\text{gauss}} e^{-(d/0.5)^2}
\]

\[
\text{gauss}_2(d) = w_{\text{gauss}} e^{-(d-3)/2}^2
\]

\[
\text{repulsion}(d) = \begin{cases} 
w_{\text{repulsion}} d^2 & d < 0 \\
0 & d \geq 0 
\end{cases}
\]

\[
\text{hydrophobic}(d) = \begin{cases} 
w_{\text{hydrophobic}} & d < 0.5 \\
0 & d > 1.5 \\
w_{\text{hydrophobic}} (1.5 - d) & \text{otherwise}
\end{cases}
\]

\[
\text{hbond}(d) = \begin{cases} 
w_{\text{hbond}} & d < -0.7 \\
0 & d > 0 \\
w_{\text{hbond}} (-\frac{10}{7}d) & \text{otherwise}
\end{cases}
\]

O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, *Journal of Computational Chemistry* 31 (2010) 455-461
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 $X$ → Model $\rightarrow y$ Prediction
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

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

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

Convolutional Neural Networks

ILSVRC top-5 error on ImageNet

https://devblogs.nvidia.com

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck
Convolutional Neural Networks

Convolution

Feature Maps

Fully Connected

Traditional NN

Dog: 0.99
Cat: 0.02
Convolutional Filters

-1 -1 -1

0 0 0

1 1 1

-1 0 1

-1 0 1

-1 0 1

-1 -1 -1

-1 8 -1

-1 -1 -1
CNNs for Protein-Ligand Scoring

- Pose Prediction
- Binding Discrimination
- Affinity Prediction
Protein-Ligand Representation

(R,G,B) pixel $\rightarrow$ (Carbon, Nitrogen, Oxygen,...) **voxel**

The only parameters for this representation are the choice of **grid resolution**, **atom density**, and **atom types**.
Training Data

Pose Prediction

4056 protein-ligand complexes
- diverse targets
- wide range of affinities
- generate poses with AutoDock Vina
- include minimized crystal pose
  - 8,688 <2Å RMSD (actives)
  - 76,743 >4Å RMSD (decoys)

Affinity Prediction

- 8,688 low RMSD poses
- assign known affinity
- regression problem
Model

2x2 Max Pooling
3x3 Convolution
2x2 Max Pooling
3x3 Convolution
2x2 Max Pooling
3x3 Convolution

Affinity
Pose
Score

48x48x48x35
24x24x24x35
24x24x24x32
12x12x12x32
12x12x12x64
6x6x6x64
6x6x6x128

Fully Connected
Pseudo-Huber Loss

Rectified Linear Unit
Softmax+Logistic Loss

48x48x48x35
24x24x24x35
24x24x24x32
12x12x12x32
12x12x12x64
6x6x6x64
6x6x6x128
Results

Trained on PDBbind refined; tested on CSAR
Beyond Scoring

Deep Dreams

https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Beyond Scoring

\[ \frac{\partial L}{\partial A} = \sum_{i \in G^A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A} \]

2Q89

More Oxygen Here

Less Oxygen Here
Minimizing Low RMSD Poses

The graph shows the distribution of poses before and after the first minimization. The x-axis represents the RMSD change, while the y-axis shows the number of poses. The graph compares the best poses (gray) with the first minimization results (blue) for better and worse poses.

- **Better Poses**: The poses that improved after the first minimization are shown in blue and are concentrated around the 0 change, indicating a significant improvement.
- **Worse Poses**: Poses that worsened are shown in blue and are spread out, with a notable increase in the number of poses with larger RMSD changes.

This visualization helps in understanding the effectiveness of the minimization process in reducing RMSD for better poses and identifying worsening trends for worse poses.
Iterative Refinement

![Graph showing RMSD change and number of poses for different iterations. The graph has a y-axis labeled '# Poses' and an x-axis labeled 'RMSD Change'. The graph compares Best, First Minimization, Second Iteration, and Third Iteration.]
Docking
vina/smina/gnina

Sampling

MCMC
MCMC
MCMC
MCMC
MCMC
...

N (50) independent Monte Carlo chains
Scored with grid-accelerated Vina
Best identified pose retained

Refinement

Vina

CNN

Rescoring

CNN
pose
affinity
D3R Results
Grand Challenge 3

Grand Challenge 3 - CatS_stage2

Affinity Ranking - Kendall's Tau
Grand Challenge 3

Grand Challenge 3 - JAK2_SC2

Affinity Ranking - Kendall’s Tau

Kendall’s Tau

Receipt ID

Green circle indicates your predictions (requires login)
Grand Challenge 3 - TIE2
Affinity Ranking - Kendall's Tau

Green circle indicates your predictions (requires login)
Grand Challenge 3

Grand Challenge 3 - VEGFR2

Affinity Ranking - Kendall's Tau

Kendall's Tau

Receipt ID

Green circle indicates your predictions (requires login)
## Grand Challenge 3

### Spearman Correlation

|       | cnn_docked_affinity | cnn_rescore_affinity | cnn_docked_scoring | cnn_rescore_scoring | vina |
|-------|---------------------|----------------------|--------------------|---------------------|------|
| cat   | 0.0701              | 0.154                | -0.0351            | 0.178               | 0.179|
| p38a  | -0.0784             | -0.116               | -0.329             | -0.305              | -0.0631|
| vegfr2| 0.366               | 0.484                | 0.434              | 0.448               | 0.414|
| jak2  | 0.428               | 0.338                | 0.39               | 0.27                | 0.106|
| jak2_sub3 | 0.68             | 0.369               | -0.372             | 0.159               | -0.633|
| tie2  | 0.648               | 0.835                | 0.136              | -0.078              | 0.561|
| abl1  | 0.634               | 0.745                | 0.005              | 0.182               | 0.713|
GC3: Pose Prediction

Pose RMSDs (Å) - Compound: Average over all - Pose 1

Green bar indicates your predictions (requires login)
GC3: Pose Prediction

Cathepsin Phase 1

- Cross-docking
  - # Compounds < RMSD vs RMSD
  - Best Sampled
  - CNN affinity
  - Vina
  - CNN pose score

Cathepsin Phase 1b

- Redocking
  - # Compounds < RMSD vs RMSD
  - Best Sampled
  - CNN affinity
  - Vina
  - CNN pose score
Grand Challenge 1

Affinity Prediction (Regression)
HSP90

Kendall Tau

Rank

Target LASSO  QSAR SMARTS  Target Balanced NN  QSAR RDKit  Balanced CNN2  Reduced NN  Target Reduced NN  Balanced CNN1  QSAR ECFP6  Balanced NN  Balanced Linear

Vina

affinity rescore
Grand Challenge 2

Affinity Ranking (Stage 2) - Kendall's Tau

Kendall's Tau

Receipt ID

Green circle indicates your predictions (requires login)
Future Plans

Train CNN for docking
- iteratively train on docked poses
- train on cross-docked poses
- fully integrate CNN scoring into search

Continue to improve model/training parameters

Next Grand Challenge
- Finish fully automated predictions early
- Make automated+human insight submission
Acknowledgements

Group Members
Jocelyn Sunseri
Jonathan King
Paul Francoeur
Matt Ragoza
Josh Hochuli
Pulkit Mittal
Alec Helbling
Gibran Biswas
Sharanya Bandla
Faiha Khan
Lily Turner

Jocelyn Sunseri
Josh Hochuli
Matt Ragoza
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.

github.com/gnina
http://bits.csb.pitt.edu
@david_koes
Prospective Case Study: TIGIT
Can we block TIGIT/PVR interaction with a small molecule?
10 diverse compounds selected for screening
• top ranked by Vina
• top ranked by CNN

| Name      | CNN Affinity | CNN Score  | Vina      |
|-----------|--------------|------------|-----------|
| Compound 1| 7.69807      | 0.994763   | 85.95     |
| Compound 2| 5.57909      | 0.0180277  | -8.12632  |
| Compound 3| 6.73692      | 0.0624742  | -9.81935  |
| Compound 4| 6.87897      | 0.953488   | -3.81378  |
| Compound 5| 6.32813      | 0.209807   | -8.60293  |
| Compound 6| 5.689        | 0.0437     | -8.991    |
| Compound 7| 4.368        | 0.022      | -9.34722  |
| Compound 8| 4.81         | 0.072      | -6.81787  |
| Compound 9| 5.22         | 0.032      | -6.264    |
| Compound 10| 6.67        | 0.361      | 6.1053    |
Results

TIGIT:CD155 Interaction

Negative Control
Positive Control

1 2 3 4 5 6 7 8 9 10

Test Compound # (all at 100 uM except #2 and #3 at 50 uM)

% Activity

EC$_{50}$ = 11.6uM

Fold induction

PD-1:PD-L1 Interaction

IC$_{50}$ > 100 µM
16% inhibition at 100 µM

% Activity

IC$_{50}$ = 14 µM

Cellular assay
But...
Filter Visualization
