THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

BASIC RESEARCH  DRUG DISCOVERY  PRE-CLINICAL  CLINICAL TRIALS  FDA REVIEW  POST-APPROVAL RESEARCH & MONITORING

PHASE I  PHASE II  PHASE III  PHASE IV

IND SUBMITTED  NDA/BLA SUBMITTED  FDA APPROVAL

TENS  HUNDREDS  THOUSANDS  NUMBER OF VOLUNTEERS

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

BASIC RESEARCH

DRUG DISCOVERY

PRE-CLINICAL

CLINICAL TRIALS

PHASE I

PHASE II

PHASE III

PHASE IV

IND SUBMITTED

NDA/BLA SUBMITTED

FDA APPROVAL

TENS

HUNDREDS

THOUSANDS

NUMBER OF VOLUNTEERS

POTENTIAL NEW MEDICINES

1 FDA-APPROVED MEDICINE

$2.6 BILLION

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)
Structure Based Drug Design

Pose Prediction  Binding Discrimination  Affinity Prediction

Virtual Screening  Lead Optimization
Structure Based Drug Design

Pose Prediction  Binding Discrimination  Affinity Prediction

Virtual Screening  Lead Optimization
Drug Discovery Funnel

http://pharmit.csb.pitt.edu
Drug Discovery Funnel

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

Protein-Ligand Scoring

Model

Pose Prediction
Binding Discrimination
Affinity 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
Deep Learning
Deep Learning

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

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

Convolution Feature Maps

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
Protein-Ligand Representation

(R,G,B) pixel
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.
Why Grids?

Cons

• coordinate frame dependent
• pairwise interactions not explicit

Pros

• clear spatial relationships
• amazingly parallel
• easy to interpret
Data Augmentation

![Graphs showing the improvement of Pose AUC and Affinity RMSE over iterations for both augmented and non-augmented data.](image)

- **Pose AUC** improves steadily with iterations, with augmented data showing a higher AUC compared to non-augmented data.
- **Affinity RMSE** decreases significantly over iterations, with augmented data demonstrating a lower RMSE compared to non-augmented data.

The graphs illustrate the effectiveness of data augmentation in enhancing model performance.
Data Augmentation
PDBbind 2016 refined set
- **4056** protein-ligand complexes
- diverse targets
- wide range of affinities
- generate poses with AutoDock Vina
- include minimized crystal pose

Target sequence similarity < 0.5
AND
Ligand similarity < 0.9
Optimized Models

Default2018

HiRes Pose

HiRes Affinity

Percent Targets with Low RMSD Top Pose

Pearson R

Time (ms)

0.50 0.52 0.54 0.56 0.58 0.60

65 70 75 80

8x8x8 Ave Pooling 2x2x2 Ave Pooling 2x2x2 Max Pooling 4x4x4 Max Pooling 8x8x8 Ave Pooling 4x4x4 Max Pooling 2x2x2 Max Pooling

3x3x3 Convolution 3x3x3 Convolution 3x3x3 Convolution 3x3x3 Convolution 3x3x3 Convolution 4x4x4 Convolution

48x48x48x28 24x24x24x35 24x24x24x32 12x12x12x32 12x12x12x64 6x6x6x64 6x6x6x128

Fully Connected

Softmax+Logistic Loss

L2 Loss

Rectified Linear Unit

Exponential Linear Unit

5x5x5 Convolution

HiRes Affinity

Default2018
Pose Results

Redocked Pose

Percent Targets with Low RMSD Top Pose

Default2017  Default2018  HiRes Affinity  HiRes Pose  Vina
Pose Results

Crossdocked Pose

Percent Targets with Low RMSD Top Pose

Default2017  Default2018  HiRes Affinity  HiRes Pose  Vina
Affinity Results

HiRes Affinity
Spearman = 0.598, RMSE = 1.714

Default 2018
Spearman = 0.570, RMSE = 1.686

Vina
Spearman = 0.473, RMSE = 1.887
Affinity Results

Clustering Split

Spearman = 0.570, RMSE = 1.686

Random Split

Spearman = 0.690, RMSE = 1.496

PDBbind Core Set

Spearman = 0.789, RMSE = 1.336
Virtual Screening

Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data

Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane

© Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3UB, U.K.
© Structural Genomics Consortium, University of Oxford, Oxford OX3 7DQ, U.K.
© Department of Chemistry, University of Oxford, Oxford OX1 3TA, U.K.
© Diamond Light Source Ltd., Oxford OX1 2DE, U.K.
© Department of Engineering, University of Oxford, Oxford OX1 3PJ, U.K.
© Alan Turing Institute, London NW1 2DB, U.K.

AUC = 0.93
AUC = 0.92
AUC = 0.91
AUC = 0.44
AUC = 0.019
Virtual Screening

In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey*  Universität Hamburg, ZBH - Center for Bioinformatics, Research Group for Computational Molecular Design, Bundesstraße 43, 20146 Hamburg, Germany

Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening

Preprint submitted on 24.03.2019, 15:39 and posted on 25.03.2019, 12:58 by Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J. Dickson, José S. Duca, Viktor Hornak, David R. Koes, Tom Kurtzman
Virtual Screening

In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening

Jochen Sieg, Florian Flachsenberg, and Matthias Rarey

Universität Hamburg, ZBH - Center for Bioinformatics, Research Group for Computational Molecular Design, Bundesstraße 43, 20146 Hamburg, Germany

Hidden Bias in the DUD-E Dataset Leads to Misleading Performance of Deep Learning in Structure-Based Virtual Screening

Preprint submitted on 24.03.2019, 15:39 and posted on 25.03.2019, 12:58 by Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J. Dickson, José S. Duca, Viktor Hornak, David R. Koes, Tom Kurtzman

AUC

0.0

0.2

0.4

0.6

0.8

1.0

CNN

Vina
Beyond Scoring

2x2x2 Max Pooling
3x3x3 Convolution
2x2x2 Max Pooling
3x3x3 Convolution
2x2x2 Max Pooling
3x3x3 Convolution
12x12x12x32
24x24x24x32
24x24x24x35
48x48x48x35

Fully Connected

Pseudo-Huber Loss
Softmax+Logistic Loss

Affinity
Pose Score
Beyond Scoring

2x2x2 Max Pooling

3x3x3 Convolution

2x2x2 Max Pooling

3x3x3 Convolution

2x2x2 Max Pooling

3x3x3 Convolution

6x6x6x64

6x6x6x128

Fully Connected

Affinity

Pose

Score

Pseudo-Huber Loss

Softmax+Logistic Loss
Beyond Scoring

2x2x2 Max Pooling

3x3x3 Convolution

2x2x2 Max Pooling

3x3x3 Convolution

2x2x2 Max Pooling

3x3x3 Convolution

Fully Connected

Affinity

Pose

Score

Beyond Scoring

2x2x2 Max Pooling

3x3x3 Convolution

2x2x2 Max Pooling

3x3x3 Convolution

2x2x2 Max Pooling

3x3x3 Convolution

Fully Connected

Affinity

Pose

Score
Beyond Scoring

\[ \frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A} \]
Beyond Scoring

2Q89

More Oxygen Here

Less Oxygen Here

\[
\frac{\partial L}{\partial A} = \sum_{i \in G_A} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}
\]
Minimizing Low RMSD Poses

![Graph showing the distribution of poses with RMSD change. The y-axis represents the number of poses and the x-axis represents the RMSD change. The graph compares the best poses with the first minimization.](image)

- Better poses are on the left side of the graph, indicating a lower RMSD change.
- Worse poses are on the right side of the graph, indicating a higher RMSD change.
Iterative Refinement
Iterative Refinement

![Histogram](image)

- **Best**
- **First Minimization**
- **Second Iteration**
Iterative Refinement

![Graph showing Iterative Refinement](image-url)
Protein–Ligand Scoring with Convolutional Neural Networks

Matthew Ragoza, Joshua Hochuli, Elisa Idrisbo, Jocelyn Sunseri, and David Ryan Koes
1 Department of Neuroscience, 2Department of Computer Science, 3Department of Biological Sciences, and 4Department of Computer and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States
5 Department of Computer Science, The College of New Jersey, Ewing, New Jersey 08628, United States

DOI: 10.1021/acs.jcim.6b00740
Publication Date (Web): April 3, 2017
Copyright © 2017 American Chemical Society

Ligand Pose Optimization with Atomic Grid–Based Convolutional Neural Networks
Matthew Ragoza, Lillian Turner, David Ryan Koes
(Submitted on 20 Oct 2017)

Visualizing convolutional neural network protein-ligand scoring
Joshua Hochuli, Alec Heilbling, Tamar Skaist, Matthew Ragoza, David Ryan Koes

Convolutional neural network scoring and minimization in the D3R 2017 community challenge
Jocelyn Sunseri, Jonathan E. King, Paul G. Franzoa, and David Ryan Koes
libmolgrid

providing support for:
- balanced, randomized, stratified batches
- temporal and spatial recurrences
- generation of tensors from molecular input data, and not just grids either!
libmolgrid

providing support for:
- balanced, randomized, stratified batches
- temporal and spatial recurrences
- generation of tensors from molecular input data, and not just grids either!
libmolgrid

Caffe Training

PyTorch Training

Keras Training

GPU Performance

GPU Memory Utilization

e = molgrid.ExampleProvider(balanced=True, shuffle=True)
e.populate('examples.txt')
gmaker = molgrid.GridMaker()
batch = e.next_batch(batch_size)
gmaker.forward(batch, input_tensor, random_translation=6, random_rotation=True)
Acknowledgements

Jocelyn Sunseri

Matt Ragoza
Tomohide Masuda

Paul Francoeur

Department of Computational and Systems Biology
Anatomy of a deep learning paper

- Strong empirical results
- Post hoc theoretical explanation

github.com/gnina
http://bits.csb.pitt.edu
@david_koes
Generative Modeling
Discriminative Model

Features $X$ \rightarrow Prediction $y$
Generative Model

Features X
Generative Model

Features $\mathbf{X}$
Generative Model

$y ? \rightarrow \text{Features } X$
Generative Adversarial Networks

True Examples

Generator

Discriminator

Loss
Is this a real dog picture?
Generative Adversarial Networks

Ian Goodfellow @goodfellow_ian · 2h
4.5 years of GAN progress on face generation. arxiv.org/abs/1406.2661
arxiv.org/abs/1812.04948

https://thispersondoesnotexist.com
Generative Models

Generative models approximate a data distribution directly. They can map samples from one distribution (noise or input data) to realistic samples from an output distribution of interest.

noise sample

generated receptor & ligand grid
Autoencoding

Encoder

Generator

Latent Space

L2 Loss
Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli† # ID, Jennifer N. Wei‡ # ID, David Duvenaud¶ # ID, José Miguel Hernández-Lobato§ # ID, Benjamín Sánchez-Lengeling†, Dennis Sheberla‡ ID, Jorge Aguilera-Iparraguirre†, Timothy D. Hirzel†, Ryan P. Adams¶ ID, and Alán Aspuru-Guzik* ‡ ID

SMILES input

ENCODER
Neural Network

CONTINUOUS
MOLECULAR
REPRESENTATION
(Latent Space)

DECODER
Neural Network

SMILES output

L2 Loss
Model

molgrid
12x12x12Å
0.5Å resolution
19 channels

3x3x3 conv + LReLU
32 filters

3x3x3 conv + LReLU
32 filters

2x2x2 ave pool

3x3x3 conv + LReLU
64 filters

3x3x3 conv + LReLU
64 filters

ave pool
cnv
128
cnv
128

1024

2x2x2 nearest-neighbor upsample

3x3x3 conv + LReLU
64 filters

3x3x3 conv + LReLU
64 filters

2x2x2 nearest-neighbor upsample

3x3x3 conv + LReLU
32 filters

3x3x3 conv + LReLU
19 filters

1024
Variational Autoencoding Examples

2BES

Atom Fitting

VAE
## Variational Autoencoding Examples

<table>
<thead>
<tr>
<th>PDB</th>
<th>True structure</th>
<th>True density</th>
<th>Gen. density</th>
<th>Fit density</th>
<th>Fit structure</th>
<th>Gen. L2 distance</th>
<th>Fit L2 distance</th>
<th>Fit RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3h78</td>
<td><img src="image1" alt="True structure" /></td>
<td><img src="image2" alt="True density" /></td>
<td><img src="image3" alt="Gen. density" /></td>
<td><img src="image4" alt="Fit density" /></td>
<td><img src="image5" alt="Fit structure" /></td>
<td>9.4053</td>
<td>8.3141</td>
<td>0.6160</td>
</tr>
<tr>
<td>4jx9</td>
<td><img src="image6" alt="True structure" /></td>
<td><img src="image7" alt="True density" /></td>
<td><img src="image8" alt="Gen. density" /></td>
<td><img src="image9" alt="Fit density" /></td>
<td><img src="image10" alt="Fit structure" /></td>
<td>13.8545</td>
<td>9.7198</td>
<td>0.8820</td>
</tr>
<tr>
<td>3jgp</td>
<td><img src="image11" alt="True structure" /></td>
<td><img src="image12" alt="True density" /></td>
<td><img src="image13" alt="Gen. density" /></td>
<td><img src="image14" alt="Fit density" /></td>
<td><img src="image15" alt="Fit structure" /></td>
<td>14.8525</td>
<td>12.5245</td>
<td>1.2066</td>
</tr>
<tr>
<td>4cwf</td>
<td><img src="image16" alt="True structure" /></td>
<td><img src="image17" alt="True density" /></td>
<td><img src="image18" alt="Gen. density" /></td>
<td><img src="image19" alt="Fit density" /></td>
<td><img src="image20" alt="Fit structure" /></td>
<td>11.4730</td>
<td>9.0564</td>
<td>0.6725</td>
</tr>
</tbody>
</table>
Context Encoding

http://people.eecs.berkeley.edu/~pathak/context_encoder/
Context Encoding

receptor grid

Generator

generated ligand grid

GAN loss
Conditioning on the Receptor
Conditioning on the Receptor
Context Encoding with Fully Convolutional Network

1m5w

Generated

Fit Densities

Fit Atoms
Context Encoding with Fully Convolutional Network

1m5w

Generated  Fit Densities  Fit Atoms
Context Encoding with Fully Convolutional Network

3bxg

Generated

Fit Densities

Fit Atoms
Context Encoding with Fully Convolutional Network

3bxg

Generated
Fit Densities
Fit Atoms
Context Encoding with Fully Convolutional Network

3ebp

Generated  Fit Densities  Fit Atoms
Context Encoding with Fully Convolutional Network

3ebp

Generated  Fit Densities  Fit Atoms