Computational Drug Discovery

David Ryan Koes

4/9/2019

http://bits.csb.pitt.edu
What is a drug?

According to the Food, Drug, and Cosmetic Act (1): a substance recognized in an official pharmacopoeia or formulary (2): a substance intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease (3): a substance other than food intended to affect the structure or function of the body (4): a substance intended for use as a component of a medicine but not a device or a component, part, or accessory of a device

http://www.merriam-webster.com/dictionary/drug
What is a drug?

According to the Food, Drug, and Cosmetic Act (1) : a substance recognized in an official pharmacopoeia or formulary (2) : a substance intended for use in the diagnosis, cure, mitigation, treatment, or prevention of disease (3) : a substance other than food intended to affect the structure or function of the body (4) : a substance intended for use as a component of a medicine but not a device or a component, part, or accessory of a device

http://www.merriam-webster.com/dictionary/drug

A small molecule intended to affect the structure/function of macromolecules
THE BIOPHARMACEUTICAL RESEARCH AND DEVELOPMENT PROCESS

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

PHASE I
IND SUBMITTED

PHASE II
NDA/BLA SUBMITTED

PHASE III
FDA APPROVAL

PHASE IV
TENS HUNDREDS THOUSANDS
NUMBER OF VOLUNTEERS

POTENTIAL NEW MEDICINES

BASIC RESEARCH

DRUG DISCOVERY

PRE-CLINICAL

CLINICAL TRIALS

FDA REVIEW

POST-APPROVAL RESEARCH & MONITORING

PRE-CLINICAL PHASE I PHASE II PHASE III PHASE IV
IND SUBMITTED NDA/BLA SUBMITTED FDA APPROVAL

Source: Pharmaceutical Research and Manufacturers of America (http://phrma.org)

$2.6 BILLION

1 FDA-APPROVED MEDICINE

$2.6 BILLION

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

- Hit generation
- Lead generation
- Lead optimization
- Clinical candidate

**Steps:**
- Target and hit identification
- Hit refinement
- Lead refinement
- Regulatory development

**Cost:**
- $ for hits
- $$ for leads
- $$$ for lead optimization
- $$$$ for clinical candidate

**Techniques:**
- VHS
- LSI
- CCS

**Approaches:**
- High-throughput screening (HTS)
- High-content screening (HCS)
- Pharmacophore

**Quality versus quantity:**
- Evaluate hits and leads for their potential to reach clinical development.

**Quality measures:**
- Properties relevant for preclinical and clinical stages.

**Bioavailability:**
- Solubility, pH, absorption, distribution, metabolism, excretion (ADME).

**Virtual screening:**
- Virtual ligand screening (VLS) and virtual hit identification.

**Importance:**
- Essential for early optimization.

**Challenges:**
- High attrition rates in clinical studies.
- Importance of combining high-throughput and high-content strategies.

**Definitions:**
- Hit: A compound that demonstrates activity in a biological or biochemical assay.
- Lead: A hit that is optimized for better pharmacological properties.
- Clinical candidate: A lead optimized for preclinical and clinical development.

**Critical parameters:**
- Physicochemical properties, safety measures, ADME attributes.

**Optimization process:**
- Historical process has been largely empirical.

**Future trends:**
- More holistic approach towards drug development.

**Quality versus quantity:**
- Criteria for selecting high-potential profiles.

**Random versus targeted libraries:**
- Strategies for generating leads.

**Pharmacophore:**
- A model representing a collection of pharmacologically active compounds.

**Pharmacophore mapping:**
- Correlation between structural features and biological activity.

**Molecular modeling:**
- Computational tools for predicting properties.

**Drug discovery research:**
- The identification of potential drug molecules.

**Outline:**
- High-content lead series identification.
- Virtual hits.
- Pharmacophore.
- Quality versus quantity.

**Figure 3:**
- A diagram illustrating the drug discovery process.

**Key points:**
- Importance of hit and lead generation strategies.
- Impact of physicochemical properties on drug development.
- Strategies for improving attrition rates.

**Impact:**
- Enhancing the efficiency of drug discovery.

**Conclusion:**
- The marriage of HTS with computational biology.

**References:**
- Literature and patents.

**Acknowledgments:**
- Contributions from the Department of Computational Biology.

**Further reading:**
- Computational and Systems Biology.
Drug Discovery

High Throughput Screening

- Hit generation
- Lead generation
- Lead optimization
- Clinical candidate

- Target and hit identification
- Hit refinement
- Lead refinement
- Regulatory development

- VHS
- LSI
- CCS

Cost:
- $ for Hit generation
- $ for Lead generation
- $$$ for Lead optimization
- $$$$ for Clinical candidate
Drug Discovery

High Throughput Screening
The State of Drug Development

New Drugs Approved

New Drug Applications

Computational Drug Discovery

Virtual Screening

existing libraries

new compounds

Hit generation

Hits

Lead generation

Leads

Lead optimization

Clinical candidate

Typical important milestones are...
Computational Drug Discovery

Virtual Screening

existing libraries

new compounds

Hit generation → Hits → Lead generation → Leads → Lead optimization → Clinical candidate

HO2C\text{R}_1\text{NH}_2 + \text{R}_2\text{CHO} + \text{R}_3\text{NC} → \text{R}_1\text{R}_2\text{R}_3
Computational Drug Discovery

Virtual Screening

existing libraries

new compounds
Kinds of Virtual Screening

ADMET

Ligand Based
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based
- dock and score
- simulation
  MM/GBSA, MM/PBSA, thermodynamic integration, free energy perturbation, Jarzynski, umbrella sampling, Monte Carlo, weighted ensemble, metadynamics…
Kinds of Virtual Screening

ADMET

Ligand Based
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based
- dock and score
- simulation

MM/GBSA, MM/PBSA, thermodynamic integration, free energy perturbation, Jarzynski, umbrella sampling, Monte Carlo, weighted ensemble, metadynamics…

Not going to cover today 😞
ADMET

- Absorption
- Distribution
- Metabolism
- Excretion
- Toxicity

Will this be a usable drug?

Screening for ADMET:
- Cytochrome P450 interaction
- Lipinski’s Rule of Five
- QSPR: Quantitative Structure Property Relationship
Kinds of Virtual Screening

ADMET

Ligand Based
- similarity to known binder
- QSAR
- pharmacophore

Receptor Based
- dock and score
Ligand Based: Similarity

Fingerprint Methods

- map molecules to a descriptor space:
  1D: molecule weight, #h-bonds, etc.
  2D: paths, bond distances between atom-pairs

- similarity is “distance” between descriptors
- for bit vectors, Tanimoto distance used

\[
T(A,B) = \frac{|A \cap B|}{|A \cup B|}
\]
Topological Fingerprints

Daylight/FP2 Fingerprints

- all paths up to 7 bonds long
- each path corresponds to bit position (hashing)
- fast similarity checking (Tanimoto)
Topological Fingerprints

Daylight/FP2 Fingerprints

- all paths up to 7 bonds long
- each path corresponds to bit position (hashing)
- fast similarity checking (Tanimoto)

O=C-N-C
O=C-O
O=N-O
N-C-C-C-C-C=0
C-C=C-C
Topological Fingerprints

ECFP4
- all substructures with diameter 4 around every atom
**Ligand Based: QSAR**

Quantitative Structure/Activity Relationships

![Chemical Structure and Vector Representation](image)

\[
f(\vec{x}) = w_1 \vec{x}_1 + w_2 \vec{x}_2 + w_3 \vec{x}_3 + \ldots + b
\]

**What's a QSAR Model?**

- **Regression model** used to estimate relationships between variables.
- Model identifies relationship between 2D chemical structures and bioactivity.

Karla Robles, Jeremy Madura, David Koes

University of Pittsburgh, Duquesne University

**QSAR Modeling: Predicting Ligand Binding Affinities and Substructures Key in Binding to SERT**
Ligand Based: Similarity

Superposition Methods
- compute “overlap” between molecules
- consider shape, electrostatics, pharmacophores

http://www.cresset-group.com/
Representing Compounds

Conformations

A single compound has many different shapes

Choices: Store sampling of explicit conformations, search for a good conformation, ignore conformations (2D only)
Ligand Based: Pharmacophore

Pharmacophore:

IUPAC: The ensemble of steric and electronic features that is necessary to ensure the optimal supra-molecular interactions with a specific biological target structure and to trigger (or to block) its biological response.

Common Features:
aromatic ring
hydrophobic area
positive ionizable
negative ionizable
hydrogen bond donor
hydrogen bond acceptor
Charge-Charge

\[ |F_{Q-q}| = |F_{q-Q}| = k \frac{|q|Q}{r^2} \]
Charge-Charge

2QWK.pdb

Salt Bridge
Hydrogen Bond
Hydrogen Bond
Hydrogen Bond
Hydrogen Bond
Hydrogen Bond

Distance:
D-A: 2.5Å – 3.5Å (4.0Å?)
H-A: 1.5Å – 2.5Å

Angle:
Depends on context
Hydrophobic
Hydrophobic

p53/MDM2
Hydrophobic

p53/MDM2
Hydrophobic

p53/MDM2
Aromatic

http://www.chem.ucalgary.ca/courses/351/Carey5th/Ch11/benzene-mo.jpg
Aromatic
Aromatic

Rings offset
Interplanar distance: 3.3-3.8Å
Pharmacophore Features
Efficient and Exact Pharmacophore Search

Pharmacophore
A spatial arrangement of molecular features essential for biological activity


Efficient and Exact Pharmacophore Search

Hydrophobic Features

Hydrogen Donor Feature

Hydrogen Acceptor Feature

Pharmer

Hydrogen Acceptor Feature
Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Efficient and Exact Pharmacophore Search

Pharmer

split x
  a - h

split y
  a - d
  c - d

split y
  e - f
  g - h

split x
  a - b
  c - d

split y
  e - h

a b c d e f g h
Efficient and Exact Pharmacophore Search

Pharmer

split x
a - h

split y
a - d

split y
e - h

split x
a - b
c - d
e - f
g - h
Efficient and Exact Pharmacophore Search

Pharmer
Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search

[Diagram showing a grid with labeled points a to h, split x and split y at different levels, and a tree structure with nodes labeled a-h, a-d, c-d, e-f, and g-h.]
Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Efficient and Exact Pharmacophore Search

Pharmer
Pharmer

Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Efficient and Exact Pharmacophore Search
Pharmer

Efficient and Exact Pharmacophore Search
Efficient and Exact Pharmacophore Search

![Graph showing time (s) vs. tolerance sphere radii multiplier for MOE HSP90, MOE FXIa, Pharmer HSP90, and Pharmer FXIa.](image-url)
http://pharmit.csb.pitt.edu
Kinds of Virtual Screening

ADMET

Ligand Based
  - similarity to known binder
  - QSAR
  - pharmacophore

Receptor Based
  - dock and score
Pharmacophores Aren’t Enough
Pharmacophores Aren’t Enough

.2µM

50µM

n.i.
Docking

Determine the **conformation** and **pose** of a ligand at a docking site

Challenge is to find conformation and pose with the best **score**
Two Phase Docking

1. Global Pose Estimation
2. Local Refinement

Stochastic
Minimization
Two Phase Docking

1. Global Pose Estimation

Stochastic

Minimization
Scoring Goals

Affinity Prediction
  - how well does it bind?

Inactive/Active Discrimination
  - does it bind?

Pose Prediction
  - how does it bind?
Scoring Goals

Affinity Prediction
  -how well does it bind?

Inactive/Active Discrimination
  -does it bind?

Pose Prediction
  -how does it bind?

Speed
Scoring Goals

Affinity Prediction
- how well does it bind?

Inactive/Active Discrimination
- does it bind?

Pose Prediction
- how does it bind?

Speed

Approximations:
Rigid or semi-rigid receptor
Implicit water model
Scoring Types

Force-field based
inter- and intra- molecular forces
van der Waals, electrostatic, torsional

Empirical
parameterized function is fit to binding energy data

Knowledge based
scoring function based on known structure, not physical principles

Consensus
### Force Field Scoring

#### G-Score

\[ E_{vdW} + E_{H-bond} = \sum \sum_{prot \text{ lig}} \left[ \left( \frac{A_{ij}}{d_{ij}^6} - \frac{B_{ij}}{d_{ij}^{12}} \right) + \left( E_{dr} + E_{ww} \right) - \left( E_{vw} + E_{ew} \right) \right] \]

\[ E_{vdW} + E_{torsion} = \sum_{lig} \frac{1}{2} \left[ 1 + \frac{n}{|\omega|} \cos(n|\omega|) \right] \]

#### D-Score

\[ E_{vdW} + E_{electrostatic} = \sum \sum_{prot \text{ lig}} \left[ \left( \frac{A_{ij}}{d_{ij}^6} + \frac{B_{ij}}{d_{ij}^{12}} \right) + 332.0 \frac{q_i q_j}{d_{ij}} \right] \]

#### Gold

\[ E_{vdW} + E_{electrostatic} = \sum \sum_{prot \text{ lig}} \left[ \left( \frac{A_{ij}}{d_{ij}^6} + \frac{B_{ij}}{d_{ij}^{12}} \right) + 332.0 \frac{q_i q_j}{d_{ij}} \right] \]

#### AutoDock

\[ E_{vdW} + E_{H-bond} + E_{electrostatic} = \sum \sum_{prot \text{ lig}} \left[ \left( \frac{A_{ij}}{d_{ij}^6} - \frac{B_{ij}}{d_{ij}^{12}} \right) + \left( E(t) \times \left( \frac{C_{ij}}{d_{ij}^{12}} - \frac{D_{ij}}{d_{ij}^{10}} \right) \right) + 332.0 \frac{q_i q_j}{d_{ij}} \right] \]

\[ E(t) = \text{angular weight factor} \]

#### DOCK (v4.0)

\[ E_{vdW} + E_{electrostatic} = \sum \sum_{prot \text{ lig}} \left[ \left( \frac{A_{ij}}{d_{ij}^6} + \frac{B_{ij}}{d_{ij}^{12}} \right) + 332.0 \frac{q_i q_j}{d_{ij}} \right] \]
Dock 4.0

Coulomb’s Law
q: partial charges
D: dielectric constant

\[ E = \sum_{i=1}^{\text{lig}} \sum_{j=1}^{\text{rec}} \left( \frac{A_{ij}}{r_{ij}^a} - \frac{B_{ij}}{r_{ij}^b} + 332 \frac{q_i q_j}{D r_{ij}} \right) \]

van der Waals
a = 12, b = 6
Lennard-Jones potential
## Empirical Scoring

<table>
<thead>
<tr>
<th>Method</th>
<th>Functional form</th>
</tr>
</thead>
</table>
| LUDI       | \[
\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta \alpha) + 
\]
\[
\Delta G_{\text{hydrophobic}} \sum_{\text{hydrophobic}} A_{\text{hydrophobic}} + \Delta G_{\text{rotor}} N_{\text{rotor}} + \Delta G_0
\]
\[
A_{\text{hydrophobic}} = \text{molecular surface area}
\]

| F-Score    | \[
\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{aromatic}} \sum_{\text{aromatic}} f(\Delta R, \Delta \alpha)
\]
\[
+ \Delta G_{\text{contact}} \sum_{\text{contact}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{rotor}} N_{\text{rotor}} + \Delta G_0
\]

| Chem-Score | \[
\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{metal}} \sum_{\text{metal}} f(\Delta R, \Delta \alpha) + 
\]
\[
\Delta G_{\text{liro}} \sum_{\text{liro}} f(\Delta R) + \Delta G_{\text{rotor}} \sum_{\text{rotor}} f\left(P_{nl}, P'_{nl}\right) + \Delta G_0
\]
# Empirical Scoring

<table>
<thead>
<tr>
<th>Method</th>
<th>Functional Form</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LUDI</strong></td>
<td>$\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{hydrophobic}} \sum_{\text{hydrophobic}} A_{\text{hydrophobic}} + \Delta G_{\text{rotor}} N_{\text{rotor}} + \Delta G_0$</td>
</tr>
<tr>
<td><strong>F-Score</strong></td>
<td>$\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{ionic}} \sum_{\text{ionic}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{aromatic}} \sum_{\text{aromatic}} f(\Delta R, \Delta \alpha)$ $+ \Delta G_{\text{contact}} \sum_{\text{contact}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{rotor}} N_{\text{rotor}} + \Delta G_0$</td>
</tr>
<tr>
<td><strong>Chem-Score</strong></td>
<td>$\Delta G_{\text{bind}} = \Delta G_{\text{H-bond}} \sum_{\text{H-bond}} f(\Delta R, \Delta \alpha) + \Delta G_{\text{metal}} \sum_{\text{metal}} f(\Delta R, \Delta \alpha)$ $+ \Delta G_{\text{lipo}} \sum_{\text{lipo}} f(\Delta R)$ $+ \Delta G_{\text{rotor}} \sum_{\text{rotor}} f(P_n, P_n') + \Delta G_0$</td>
</tr>
</tbody>
</table>
## AutoDock Vina

The AutoDock Vina scoring function is a combination of several terms, each of which is weighted differently. The function is used to predict the binding affinity of a ligand to a receptor. Here are the terms and their definitions:

### Gauss Terms
- \( \text{gauss}_1(d) = w_{\text{gauss}_1} e^{-(d/0.5)^2} \)
- \( \text{gauss}_2(d) = w_{\text{gauss}_2} e^{-(d-3)/2} \)

### Repulsion Term
- \( \text{repulsion}(d) = \begin{cases} \ w_{\text{repulsion}} d^2 & \text{if } d < 0 \\ \ 0 & \text{if } d \geq 0 \end{cases} \)

### Hydrophobic Term
- \( \text{hydrophobic}(d) = \begin{cases} \ w_{\text{hydrophobic}} & \text{if } d < 0.5 \\ \ 0 & \text{if } d \geq 1.5 \\ \ w_{\text{hydrophobic}} (1.5 - d) & \text{otherwise} \end{cases} \)

### Hydrogen Bond Term
- \( \text{hbond}(d) = \begin{cases} \ w_{\text{hbond}} & \text{if } d < -0.7 \\ \ 0 & \text{if } d > 0 \\ \ w_{\text{hbond}} (-\frac{10}{7} d) & \text{otherwise} \end{cases} \)

### Weights

<table>
<thead>
<tr>
<th>Weight</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.0356</td>
<td>gauss_1</td>
</tr>
<tr>
<td>-0.00516</td>
<td>gauss_2</td>
</tr>
<tr>
<td>0.840</td>
<td>Repulsion</td>
</tr>
<tr>
<td>-0.0351</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>-0.587</td>
<td>Hydrogen bonding</td>
</tr>
<tr>
<td>0.0585</td>
<td>( N_{\text{rot}} )</td>
</tr>
</tbody>
</table>

### Score vs. Surface Distance

The graph on the right shows the score (x-axis) against surface distance (A, y-axis) for different terms: steric, steric + hydrophobic, and steric + H-bond.
## Knowledge Based

### Functional form

<table>
<thead>
<tr>
<th>PMF</th>
<th>Parametrized pairwise potential PMF score:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[ PMF = \sum_{prot} \sum_{lig} A_{ij}(d_{ij}) \cdot A_{ij}(d_{ij}) = -k_B T \ln \left[ f^{ij}<em>\text{Vol corr}(r) \frac{\rho</em>{ij}^\text{seg}(r)}{\rho_{ij}^\text{bulk}} \right] ]</td>
</tr>
</tbody>
</table>

where \( k_B \) is the Boltzmann constant, \( f^{ij}_\text{Vol corr}(r) \) is a ligand volume correction factor and \( \rho_{ij}^\text{seg}(r) / \rho_{ij}^\text{bulk} \) indicates a radial distribution function for a protein atom \( i \) and a ligand atom \( j \).

| DrugScore (v1.2) | \[ \Delta W = \gamma \sum_{prot} \sum_{lig} \Delta W_{ij}(r) + (1 - \gamma) \times \left[ \sum_{lig} \Delta W_i(SAS, SAS_0) + \sum_{prot} \Delta W_j(SAS, SAS_0) \right] \] |

\( SAS = \) Solvent accessible surface area terms, \( W_{ij} = \) distance dependent pairwise potential

| SMoG | \[ G = \sum_{ij} g_{ij} \Delta_{ij}; \quad \Delta_{ij} = \begin{cases} 0 & (i, j \text{ more than 5 Å}) \\ 1 & (i, j \text{ within 5 Å}) \end{cases}; \quad g_{ij} = -kT \log \left[ \frac{p_{ij}}{\bar{p}} \right]; \] |

\( p_{ij} \) an \( \bar{p} \) are interatomic and averaged interactomic interactions
RF-Score

Pairwise Distance Counts (<12Å)

<table>
<thead>
<tr>
<th>Protein</th>
<th>C</th>
<th>N</th>
<th>O</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Br</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Random Forest
**RF-Score Output**

\[ R = 0.776 \text{ on independent test set (195 complexes)} \]

<table>
<thead>
<tr>
<th>Scoring function</th>
<th>( R )</th>
<th>( Rs )</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF-Score</td>
<td>0.776</td>
<td>0.762</td>
<td>1.58</td>
</tr>
<tr>
<td>X-Score::HMScore</td>
<td>0.644</td>
<td>0.705</td>
<td>1.83</td>
</tr>
<tr>
<td>DrugScore\textsuperscript{CSD}</td>
<td>0.569</td>
<td>0.627</td>
<td>1.96</td>
</tr>
<tr>
<td>SYBYL::ChemScore</td>
<td>0.555</td>
<td>0.585</td>
<td>1.98</td>
</tr>
<tr>
<td>DS::PLP1</td>
<td>0.545</td>
<td>0.588</td>
<td>2</td>
</tr>
<tr>
<td>GOLD::ASP</td>
<td>0.534</td>
<td>0.577</td>
<td>2.02</td>
</tr>
<tr>
<td>SYBYL::G-Score</td>
<td>0.492</td>
<td>0.536</td>
<td>2.08</td>
</tr>
<tr>
<td>DS::LUDI3</td>
<td>0.487</td>
<td>0.478</td>
<td>2.09</td>
</tr>
<tr>
<td>DS::LigScore2</td>
<td>0.464</td>
<td>0.507</td>
<td>2.12</td>
</tr>
<tr>
<td>GlideScore-XP</td>
<td>0.457</td>
<td>0.435</td>
<td>2.14</td>
</tr>
<tr>
<td>DS::PMF</td>
<td>0.445</td>
<td>0.448</td>
<td>2.14</td>
</tr>
<tr>
<td>GOLD::ChemScore</td>
<td>0.441</td>
<td>0.452</td>
<td>2.15</td>
</tr>
<tr>
<td>SYBYL::D-Score</td>
<td>0.392</td>
<td>0.447</td>
<td>2.19</td>
</tr>
<tr>
<td>DS::Jain</td>
<td>0.316</td>
<td>0.346</td>
<td>2.24</td>
</tr>
<tr>
<td>GOLD::GoldScore</td>
<td>0.295</td>
<td>0.322</td>
<td>2.29</td>
</tr>
<tr>
<td>SYBYL::PMF-Score</td>
<td>0.268</td>
<td>0.273</td>
<td>2.29</td>
</tr>
<tr>
<td>SYBYL::F-Score</td>
<td>0.216</td>
<td>0.243</td>
<td>2.35</td>
</tr>
</tbody>
</table>

**RMSE = 1.58**
RF-Score Output

$R = 0.776$ on independent test set (195 complexes)

RMSE = 1.58

$R = 0.46; \ RMSE = 1.6$
Scoring

Ideally, score would equal affinity – but this is an unsolved problem.

http://www.csardock.org/
Scoring

Ideally, score would equal affinity – but this is an unsolved problem.

\[ R^2 = 0.28 \]
\[ \text{RMSE} = 1.9 \]

http://www.csardock.org/
Scoring

Ideally, score would equal affinity – but this is an unsolved problem.

\[ R^2 = 0.58 \]
\[ \text{RMSE} = 1.51 \]

http://www.csardock.org/
Scoring

Ideally, score would equal affinity – but this is an unsolved problem.

$R^2 = 0.58$
$RMSE = 1.51$

http://www.csardock.org/
Scoring 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 PubCher
- 125,526 structures in the PDB
- 16,179 annotated complexes in PDBbind
Machine Learning

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

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

\text{step} \hspace{1cm} \text{sigmoid} \hspace{1cm} \text{ReLU}
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

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

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

- airplane
- automobile
- bird
- cat
- deer
- dog
- frog
- horse
- ship
- truck

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

CNN

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.
Optimized Models

- Default2017
- Default2018
- HiRes Pose
- HiRes Affinity

Percent Targets with Low RMSD Top Pose vs Pearson R

- 48x48 Ave Pooling
- 2x2x2 Ave Pooling
- 3x3x3 Convolution
- 1x1x1 Convolution
- 2x2x2 Max Pooling
- 3x3x3 Max Pooling
- 4x4x4 Max Pooling

- Rectified Linear Unit
- Exponential Linear Unit
- Softmax + Logistic Loss
- Pseudo-Huber Loss

- HiRes Pose
- HiRes Affinity

Default2018

- 48x48 Ave Pooling
- 2x2x2 Ave Pooling
- 3x3x3 Convolution
- 1x1x1 Convolution
- 2x2x2 Max Pooling
- 3x3x3 Max Pooling
- 4x4x4 Max Pooling

- Rectified Linear Unit
- Exponential Linear Unit
- Softmax + Logistic Loss
- Pseudo-Huber Loss
Pose Results

Redocked Pose

Percent Targets with Low RMSD Top Pose

Default2017  Default2018  HiRes Affinity  HiRes Pose  Vina

0.5  0.6  0.7  0.8
Pose Results

Crossdockected 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

HiRes Affinity

Spearman = 0.598, RMSE = 1.714

Default 2018

Spearman Correlation

Clumped

Random

Spearman = 0.70, RMSE = 1.65

Spearman = 0.55, RMSE = 1.55

Vina

Spearman = 0.473, RMSE = 1.887
Beyond Scoring
Beyond Scoring
Beyond Scoring
Beyond Scoring

Deep Dreams

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

2Q89

More Oxygen Here

Less Oxygen Here
Beyond Scoring

2Q89

More Oxygen Here

Less Oxygen Here

$$\frac{\partial L}{\partial A} = \sum_{i \in G_A^{\text{data}}} \frac{\partial L}{\partial G_i} \frac{\partial G_i}{\partial D} \frac{\partial D}{\partial A}$$

unit_pool

label
Minimizing Low RMSD Poses

![Graph showing the number of poses](#)

- **better**
- **RMSD Change**
- **worse**

Legend:
- Grey: Best
- Blue: First Minimization
Iterative Refinement

![Graph showing RMSD change and number of poses. The x-axis represents RMSD change, and the y-axis represents the number of poses. Two distributions are shown: one for the best poses and another for the first minimization. The graph illustrates the distribution of poses across different RMSD change values.]
Iterative Refinement

![Graph showing Iterative Refinement](image-url)
Iterative Refinement

The diagram shows the RMSD (Root Mean Square Deviation) change for different iterations and poses. The x-axis represents the RMSD change, while the y-axis shows the number of poses. Thehistogram includes four categories: Best, First Minimization, Second Iteration, and Third Iteration, each represented by different colors.
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 (v2), last revised 6 Jan 2017 (this version, v3))

AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery
Izhari 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 Lusco†, Gianluca Poliastro, and Pierre Balle‡
†School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland
‡Department of Computer Science, University of California, Irvine, Irvine, California 92697, United States
DOI: 10.1021/ci400187y
Publication Date (Web): June 24, 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)

Protein–Ligand Scoring with Convolutional Neural Networks
Matthew Ragan‡, Joshua Hochuli†, Elisa Iordachita, Jocelyn Sunseri, and David Ryan Koepsel†
†Department of Neuroscience, ‡Department of Computer Science, Department of Biological Sciences, and Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, United States
DOI: 10.1021/acs.jcim.6b00740
Publication Date (Web): April 3, 2017
Copyright © 2017 American Chemical Society