How do Scientists Create Medicines?

An Introduction to Computational Drug Development

Rachit Mukkamala
Splash 2021
November 20th, 2021
What we’ll go over today!

1. What is Drug Development
2. Intro to the fundamental tools you need for CADD
3. Group Project to Practice what you’ve learned
Overview of Drug Development
The costs of drug discovery
Finding a needle in a haystack

Can we optimize and/or automate this process?
Computation Methods start to be used!
Computer Aided Drug Discovery Pipeline
Hit Identification Approaches: Ligand-Based

QSAR: Quantitative Structure Activity Relationship

Pharmacophore Model
Hit Identification Approaches: Structure-Based

Molecular Docking!

We will be focusing on this method today!
Lead Optimization: Higher Quality Predictions

Molecular Dynamics Simulation (MD)

Free Energy Estimation (MD)

Advanced QSAR models

ΔΔG_{binding} = ΔG_2 − ΔG_1 = ΔG_B − ΔG_A
**Lead Optimization: ADMET and Druggability**

**ADMET = Absorption, Distribution, Metabolism, Excretion, and Toxicity**

<table>
<thead>
<tr>
<th>Property</th>
<th>Model Name</th>
<th>Predicted Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absorption</td>
<td>Water solubility</td>
<td>-2.895</td>
<td>Numeric (log mol/L)</td>
</tr>
<tr>
<td>Absorption</td>
<td>Caco2 permeability</td>
<td>0.29</td>
<td>Numeric (log Papp in 10⁻⁶ cm/s)</td>
</tr>
<tr>
<td>Absorption</td>
<td>Intestinal absorption (human)</td>
<td>100</td>
<td>Numeric (% Absorbed)</td>
</tr>
<tr>
<td>Absorption</td>
<td>Skin Permeability</td>
<td>-2.736</td>
<td>Numeric (log Kp)</td>
</tr>
<tr>
<td>Absorption</td>
<td>P-glycoprotein substrate</td>
<td>Yes</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Absorption</td>
<td>P-glycoprotein inhibitor</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Absorption</td>
<td>P-glycoprotein II inhibitor</td>
<td>Yes</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Distribution</td>
<td>VDss (human)</td>
<td>-0.12</td>
<td>Numeric (log L/Kg)</td>
</tr>
<tr>
<td>Distribution</td>
<td>Fraction unbound (human)</td>
<td>0.286</td>
<td>Numeric (Fu)</td>
</tr>
<tr>
<td>Distribution</td>
<td>BBB permeability</td>
<td>-1.807</td>
<td>Numeric (log BB)</td>
</tr>
<tr>
<td>Distribution</td>
<td>CNS permeability</td>
<td>-1.372</td>
<td>Numeric (log PS)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property</th>
<th>Model Name</th>
<th>Predicted Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism</td>
<td>CYP2D6 substrate</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Metabolism</td>
<td>CYP3A4 substrate</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Metabolism</td>
<td>CYP1A2 inhibitor</td>
<td>Yes</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Metabolism</td>
<td>CYP2C19 inhibitor</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Metabolism</td>
<td>CYP3A4 inhibitor</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Excretion</td>
<td>Total Clearance</td>
<td>0.45</td>
<td>Numeric (log ml/min/kg)</td>
</tr>
<tr>
<td>Excretion</td>
<td>Renal OCT2 substrate</td>
<td>Yes</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>AMES toxicity</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Max. tolerated dose (human)</td>
<td>0.42</td>
<td>Numeric (log mg/kg/day)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>hERG I inhibitor</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>hERG II inhibitor</td>
<td>Yes</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Oral Rat Acute Toxicity (LD50)</td>
<td>2.481</td>
<td>Numeric (mol/kg)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Oral Rat Chronic Toxicity (LOAEL)</td>
<td>1.434</td>
<td>Numeric (log mg/kg_bw/day)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Hepatotoxicity</td>
<td>Yes</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Skin Sensitisation</td>
<td>No</td>
<td>Categorical (Yes/No)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Tpyriformis toxicity</td>
<td>0.285</td>
<td>Numeric (log ug/L)</td>
</tr>
<tr>
<td>Toxicity</td>
<td>Minnow toxicity</td>
<td>0.518</td>
<td>Numeric (log mM)</td>
</tr>
</tbody>
</table>
Druggability Aside: Lipinski’s Rule of Five

Rule of thumb to assess if a compound is druglike (orally active, can be absorbed, etc)

1. No more than 5 hydrogen bond donors (the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)
2. No more than 10 hydrogen bond acceptors (nitrogen or oxygen atoms)
3. A molecular mass less than 500 amu
4. An octanol-water partition coefficient (log P) that does not exceed 5

Mcule allows us to limit our virtual screen to only compounds that pass RO5 criteria
What is logP?

Log Octanol-Water Partition Coefficient

\[ \log P_{\text{oct/wat}} = \log \left( \frac{[\text{solute}]_{\text{un-ionized}}}{[\text{solute}]_{\text{un-ionized}}} \right) \]

Higher logP = Compound is more hydrophobic

Lower logP = Compound is more hydrophilic

Important for druggability! Hydrophobic compounds are harder to eliminate by the kidneys, so they are active for longer periods but also potentially more toxic!
Scaffold Hopping

- Used in both Ligand Based Screening (Hit ID) and also for Lead Optimization

- Uses a search tree to try and find new compounds in library that are similar to the reference compound (the search query)

- LBVS: Used to find new potential compounds that are similar to existing reference drug

- Lead Optimization: Search for new compounds similar to the initial hit that score better in ADMET, druggability, more unique (IP), etc.
Summary of Docking

- **Receptor** = protein that we are interested in
- **Ligand** = drug/small molecule that we are testing
- **Ligand Pose** = one particular orientation/position/shape of the ligand
  - For molecular docking, receptor generally assumed to be rigid, which is not always ideal, but simplifies calculations. The ligand, however, is free to move around.

- **Goal**: Find the ligand pose which has the strongest binding affinity/energy to the protein.
What is $\Delta G$ - Free Energy

- Change in free energy determines whether a process is thermodynamically favorable (spontaneous)
- $\Delta G < 0 \rightarrow$ spontaneous, $\Delta G > 0 \rightarrow$ non spontaneous
- Protein-Ligand binding can be thought of as a reaction
- If a drug has a smaller, more negative $\Delta G$, it is more favored to bind to protein.
- We can use $\Delta G$ to score/compare our compounds

The association constant ($K_a$) provides a measure of affinity between protein & ligand

$$K_a = \frac{[PL]}{[P][L]} = \frac{k_a}{k_d}$$

Equilibrium constant (uppercase K)

Rate constants (lowercase k)

$\Delta G^o = -RT \ln K = -RT \ln \left( \frac{[Protein + Ligand]}{[Protein][Ligand]} \right)$

Units: $\text{kcal/mol, kJ/mol}$

(Autodock reports in kcal/mol)
How do we calculate $\Delta G$ using software?

$$
\Delta G^0 = -RT \ln K = -RT \ln \frac{[\text{Protein} + \text{Ligand}]}{[\text{Protein}][\text{Ligand}]}
$$

- Technically, we would need to run this experiment in vitro, and calculate starting/ending concentrations.

- **Molecular Docking software** uses force fields to estimate the binding free energy (combination of different potential functions)

- The goal of docking software is to find the lowest possible $\Delta G$ for a protein-ligand combo
Breaking down the generic docking algorithm

How to find the lowest $\Delta G$ for a protein and ligand? We could brute force test every single possible ligand pose, but this would take way too long.

Instead, modern docking algorithms will either use optimization algorithms (genetic algorithm, gradient descent) or randomly sample various poses/positions. These will give a good-enough estimate.
Mcule uses Autodock Vina, which uses a modified form of gradient descent to minimize $\Delta G$.

**Gradient Descent Steps:**

1. Pick a starting pose for the ligand
2. Calculate $\Delta G$ and the gradient of $\Delta G$ (the slope of $\Delta G$ in every coordinate direction)
3. Move down the gradient (the slope) towards a $\Delta G$ minimum.

**What are the axes/dimensions?**
- Each dimension represents a different parameter of the ligand (x,y,z position, bond angle(s), torsional angles, distance from protein/amino acids)
Interpreting Autodock Vina Docking Results

- When running Vina on computer, the 10 best poses with the lowest $\Delta G$ are reported. Mcule only reports the best pose though.
- We can visualize pose, see what amino acids ligand is interacting with.
- We can use docking scores to pick out the best ligands for further processing

Things to keep in mind:
1. Docking scores are only estimates. They often vary after a repeated docking. They also do not account for solvent, receptor flexing, etc. The gold standard for binding free energy calculation is molecular dynamics.
2. Docking is best for rapidly screening a big library of compounds, not too worried about accuracy.
3. Because docking scores vary, don’t only focus on the best pose/ligand. Pick the top 5-10 or so.
4. Just because a compound binds strongly doesn’t mean that it will do what you want
File/Data Types and other Resources
PDB: Protein Data Bank

- www.rcsb.org
- Contains structure files for **thousands** of proteins (XYZ atom coords)
- File Format: .pdb file, can be inputted into Mcule
SMILES String Format

- **SMILES** = format that allows molecules to be represented as a string of text
- **EX:** C1=CC=C2C(=C1)NC=N2 represents:

- Basic logic: C is carbon, N is nitrogen, etc., single bonds and hydrogens are implied, double bonds are =
- Lots of software can generate a SMILES string for you, which you can paste into many other places (tox check, pubchem, etc.)
pkCSM: ADMET Checker/Calculator

- pkCSM is an web server which takes in a SMILES string as input and calculates most major ADMET properties and other basic properties (logP)
- Very, very easy and straightforward to use. Save URL to save the results (or screenshot)
Demo of MCULE!

Case study of Beta-Lactamase
Group Project!!!
Overview of Group Project

Objectives:

a. Pick a disease/problem/challenge in biology you are interested in
b. Identify a protein target that is important/related to your problem
c. Use structure-based drug design to identify a hit compound
d. Use toxicity check, druggability check, etc, to optimize your best hits and create lead compounds
e. Present your lead compounds!
Overview of Group Project

Logistics:

- You will be assigned into groups of 5, and put into breakout rooms
- You will have 30 mins to work on your mini-project (hopefully lol)
- I will be rotating between breakout rooms to help y’all out. If you have questions/need help from me, send me a message directly.
- Since MCULE free has limits, you can pool your accounts together
- Split up/delegate tasks to get done in time!!
- I highly recommend creating a shared google doc so that you can keep track of tasks and info
Overview of Group Project

Presentation:

- If possible, try and make a tiny Google slides with a few pics/info
- The presentation should be pretty short (~2 mins). Suggested content:
  1. Name of disease/problem and basic biology background
  2. Name of your protein target
  3. Pictures of 3 hits and their dock scores
  4. Pictures of your leads and their dock scores
Overview of Group Project

Helpful Resources:

- PDB (www.rcsb.org)
- OMIM - online mendelian inheritance in man, great resource to learn about the molecular basis of genetic diseases and the proteins involved (www.omim.org)
- And of course, Mcule (www.mcule.com). For Mcule account, just put your school name for organization.