

How do Scientists Create Medicines?

An Introduction to Computational Drug Development

Rachit Mukkamala
Splash 2021
November 20th, 2021

What we'll go over today!

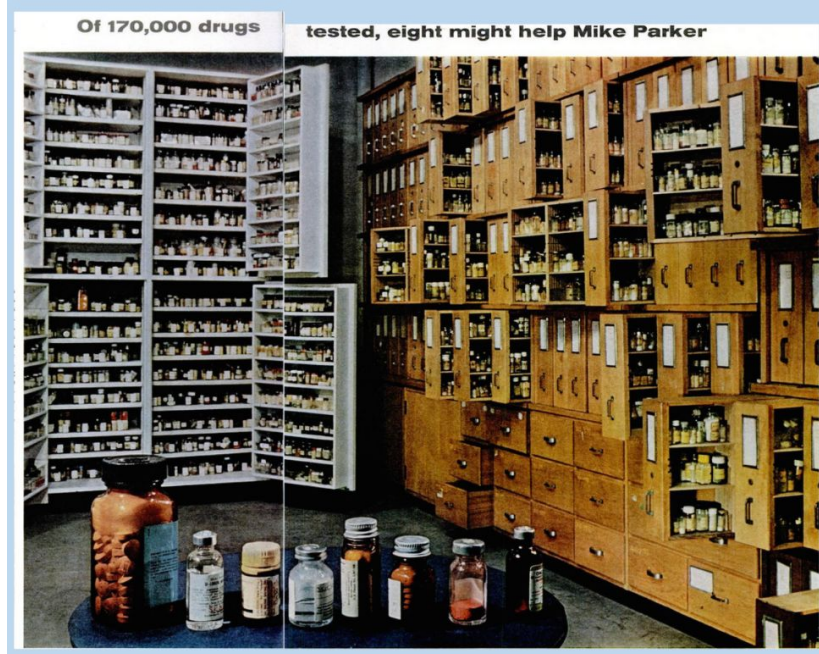
1. What is Drug Development
2. Intro to the fundamental tools you need for CADD
3. Group Project to Practice what you've learned



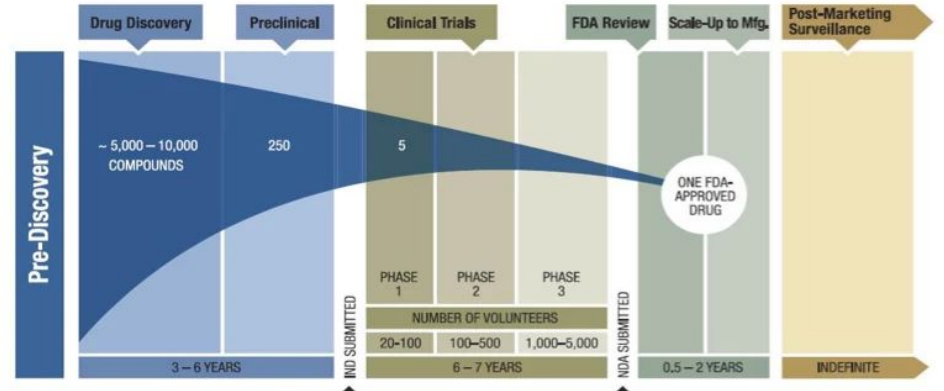


Overview of Drug Development

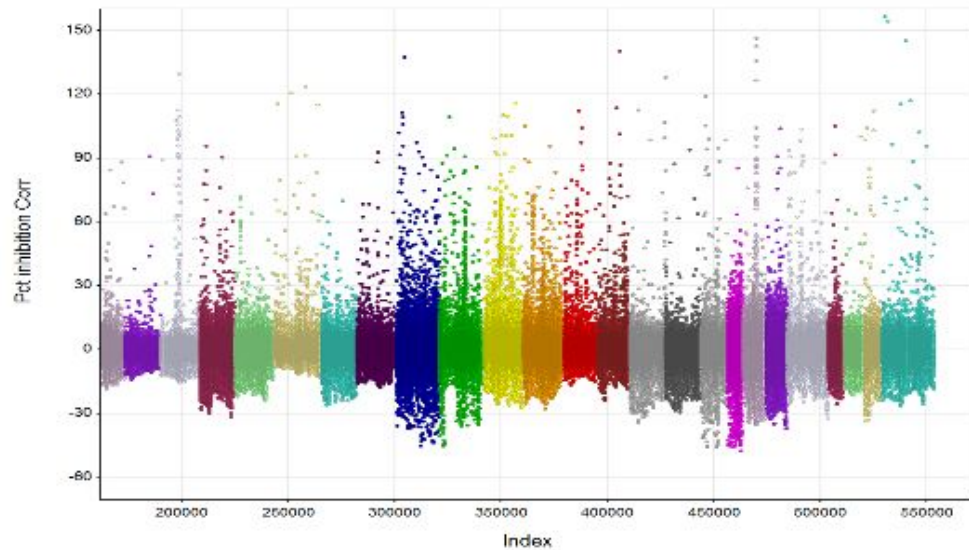
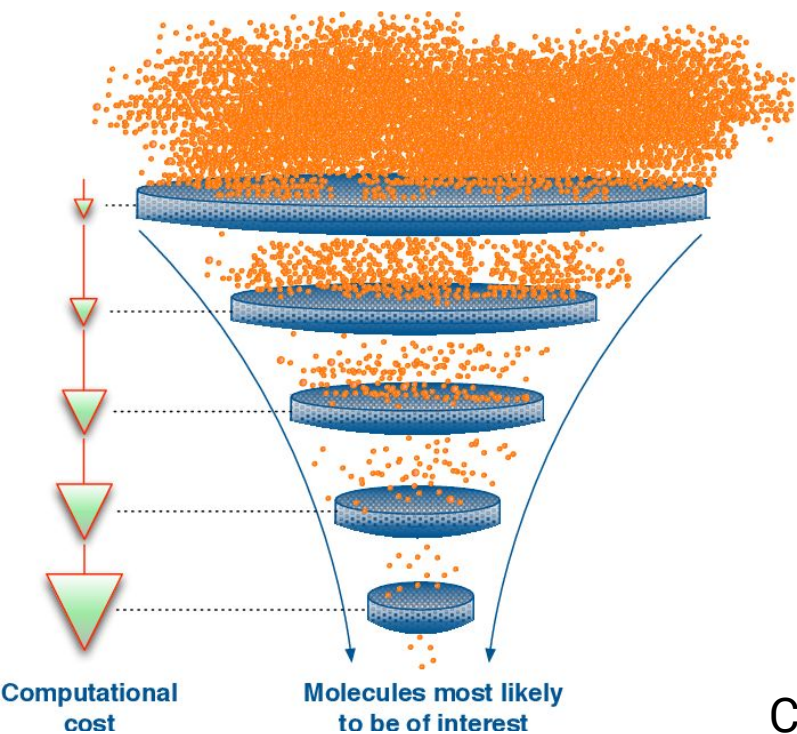
The costs of drug discovery



Drug Discovery and Development Timeline



Finding a needle in a haystack

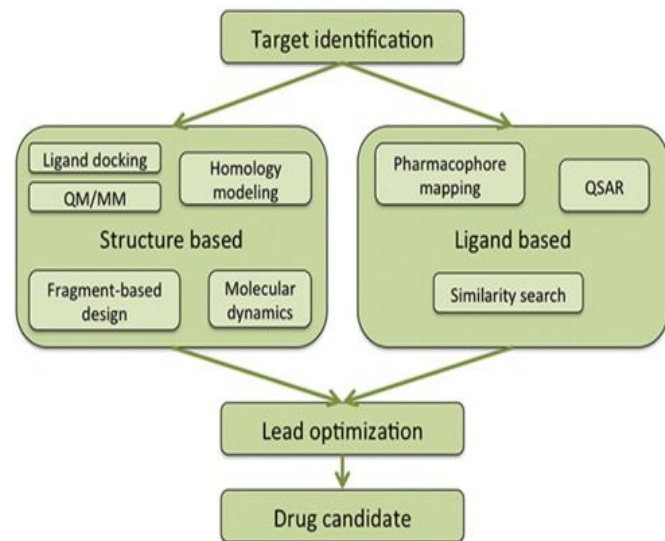
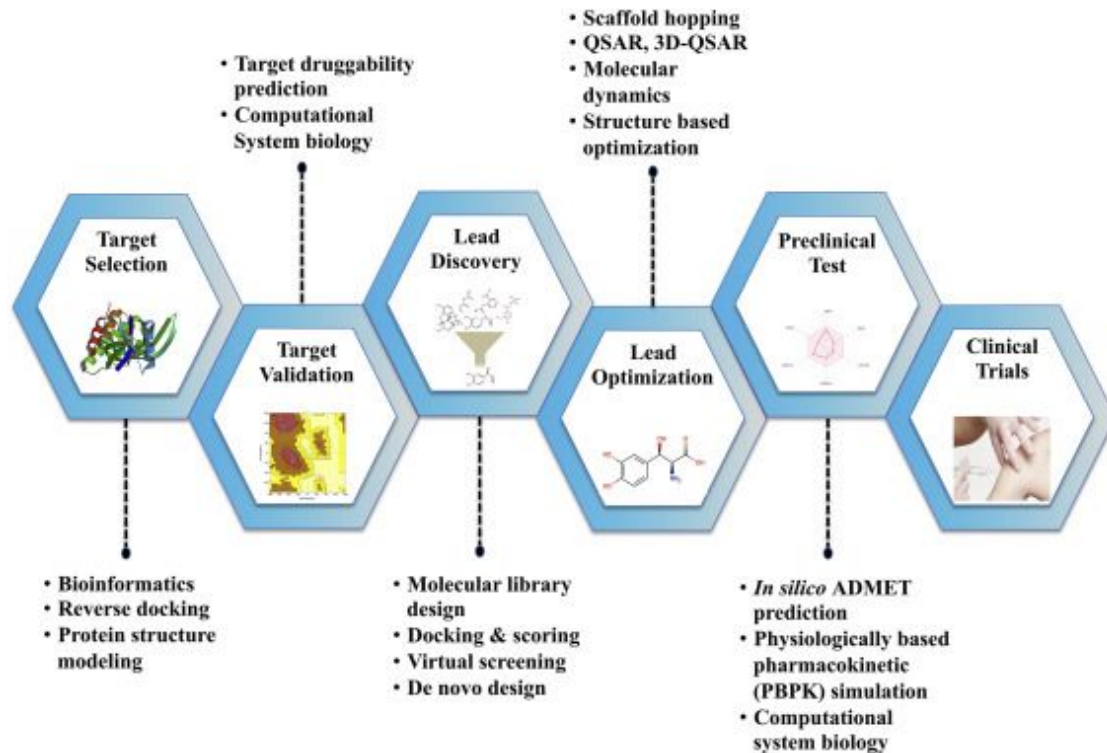


Can we optimize and/or automate this process?

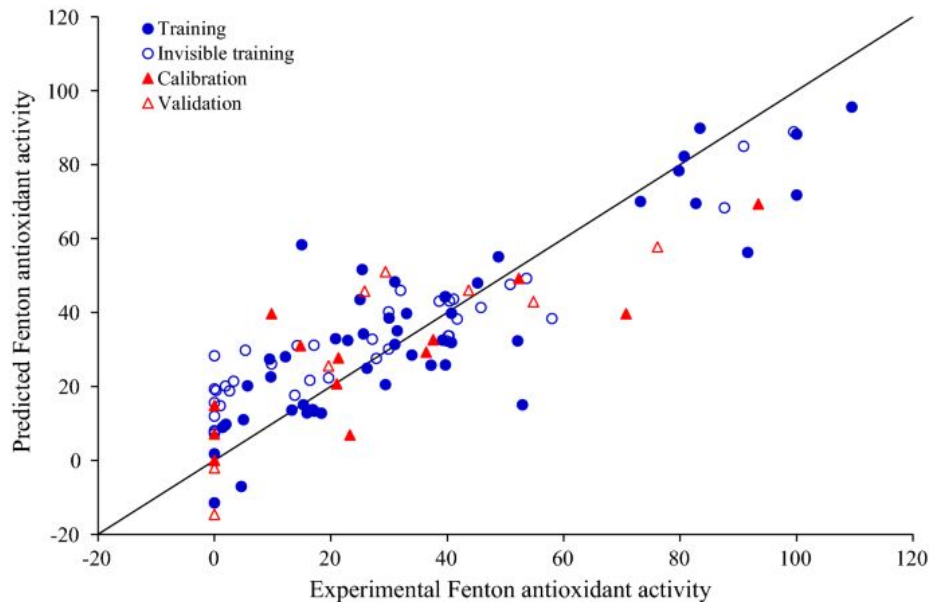
Computation Methods start to be used!



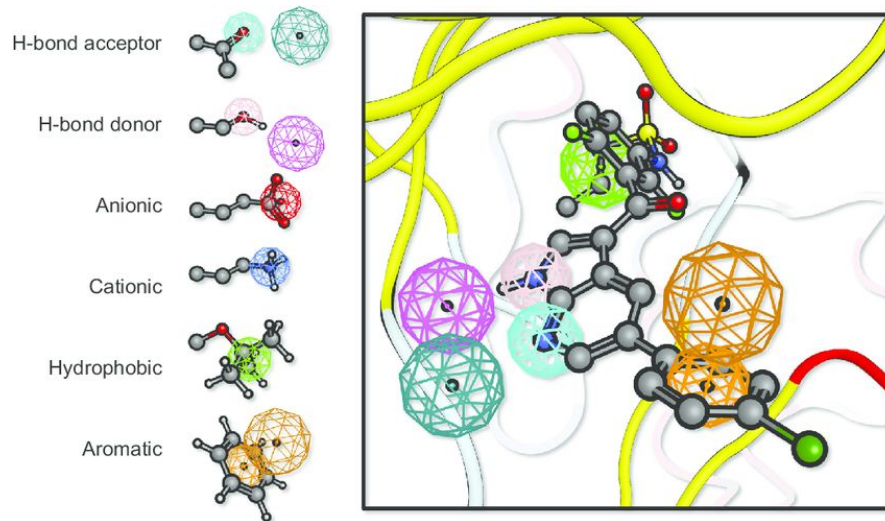
Computer Aided Drug Discovery Pipeline



Hit Identification Approaches: Ligand-Based



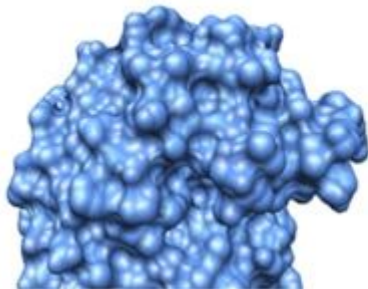
QSAR: Quantitative Structure Activity Relationship



Pharmacophore Model

Hit Identification Approaches: Structure-Based

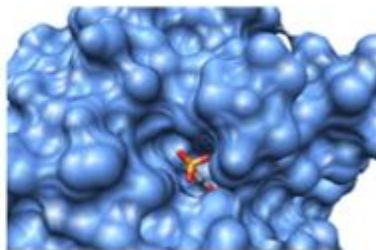
Target



Ligand

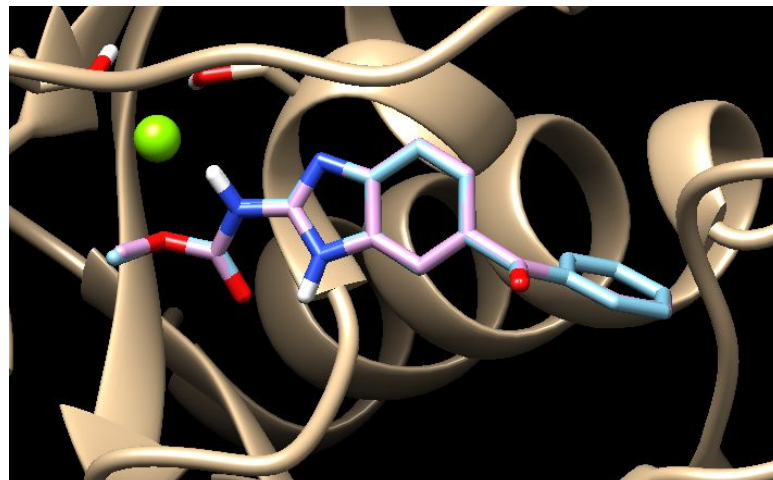
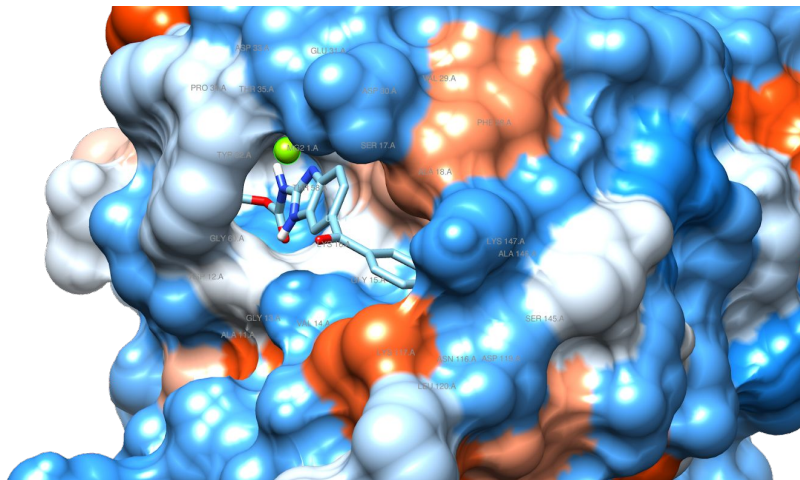


Molecular Docking



Molecular Docking!

We will be focusing on this method today!

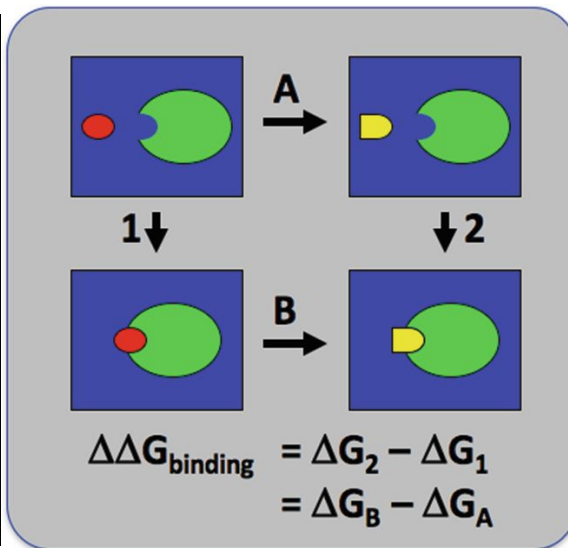


Lead Optimization: Higher Quality Predictions

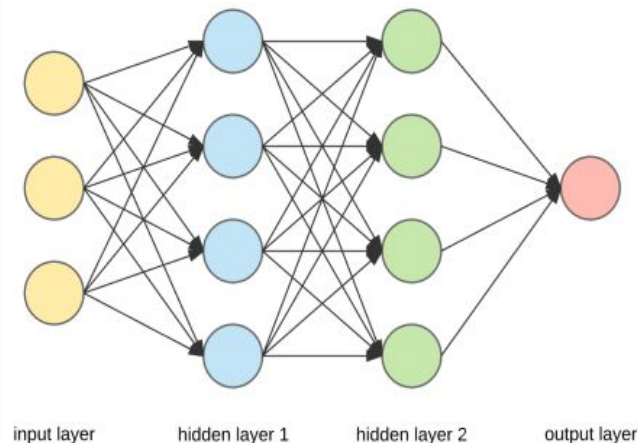
Molecular Dynamics Simulation (MD)



Free Energy Estimation (MD)



Advanced QSAR models



Lead Optimization: ADMET and Druggability

ADMET = **A**bsorption, **D**istribution, **M**etabolism, **E**xcretion, and **T**oxicity

Property	Model Name	Predicted Value	Unit
Absorption	Water solubility	-2.895	Numeric (log mol/L)
Absorption	Caco2 permeability	0.29	Numeric (log Papp in 10 ⁻⁸ cm/s)
Absorption	Intestinal absorption (human)	100	Numeric (% Absorbed)
Absorption	Skin Permeability	-2.735	Numeric (log Kp)
Absorption	P-glycoprotein substrate	Yes	Categorical (Yes/No)
Absorption	P-glycoprotein I inhibitor	No	Categorical (Yes/No)
Absorption	P-glycoprotein II inhibitor	Yes	Categorical (Yes/No)
Distribution	VDss (human)	-0.12	Numeric (log L/kg)
Distribution	Fraction unbound (human)	0.285	Numeric (Fu)
Distribution	BBB permeability	-1.807	Numeric (log BB)
Distribution	CNS permeability	-1.372	Numeric (log PS)
Metabolism	CYP2D6 substrate	No	Categorical (Yes/No)
Metabolism	CYP3A4 substrate	Yes	Categorical (Yes/No)
Metabolism	CYP1A2 inhibitor	Yes	Categorical (Yes/No)
Metabolism	CYP2C19 inhibitor	No	Categorical (Yes/No)

Metabolism	CYP2D6 inhibitor	No	Categorical (Yes/No)
Metabolism	CYP3A4 inhibitor	No	Categorical (Yes/No)
Excretion	Total Clearance	0.45	Numeric (log ml/min/kg)
Excretion	Renal OCT2 substrate	Yes	Categorical (Yes/No)
Toxicity	AMES toxicity	No	Categorical (Yes/No)
Toxicity	Max. tolerated dose (human)	0.42	Numeric (log mg/kg/day)
Toxicity	hERG I inhibitor	No	Categorical (Yes/No)
Toxicity	hERG II inhibitor	Yes	Categorical (Yes/No)
Toxicity	Oral Rat Acute Toxicity (LD50)	2.481	Numeric (mol/kg)
Toxicity	Oral Rat Chronic Toxicity (LOAEL)	1.434	Numeric (log mg/kg_bw/day)
Toxicity	Hepatotoxicity	Yes	Categorical (Yes/No)
Toxicity	Skin Sensitisation	No	Categorical (Yes/No)
Toxicity	<i>T.Pyiformis</i> toxicity	0.285	Numeric (log ug/L)
Toxicity	Minnow toxicity	0.518	Numeric (log mM)

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

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

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.

Molecular Docking and SBDD

Summary of Docking

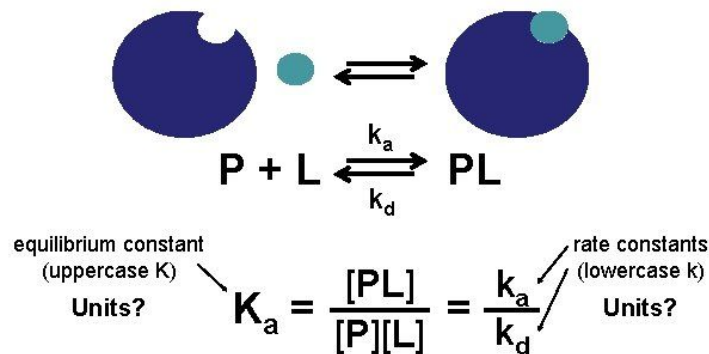
- **Receptor** = protein that we are interested in
- **Ligand** = drug/small molecule that we are testing
- **Ligand Pose** = one particular orientation/position/shape of the ligand
- *For molecular docking, receptor generally assumed to be rigid, which is not always ideal, but simplifies calculations.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

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

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



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

Units: **kcal/mol**, kJ/mol
(autodock reports in kcal/mol)

How do we calculate ΔG using software?

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

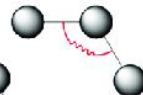
- Technically, we would need to run this experiment in vitro, and calculate starting/ending concentrations.
- **Molecular Docking software** uses force fields to estimate the binding free energy (combination of different potential fxns)
- The goal of docking software is to **find the lowest possible ΔG for a protein-ligand combo**

$$U(R) = \sum_{bonds} k_r (r - r_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} k_\phi (1 + \cos[n\phi - \gamma]) + \sum_{impropers} k_\omega (\omega - \omega_{eq})^2 + \sum_{i < j}^{atoms} \epsilon_{ij} \left[\left(\frac{r_m}{r_{ij}} \right)^{12} - 2 \left(\frac{r_m}{r_{ij}} \right)^6 \right] + \sum_{i < j}^{atoms} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

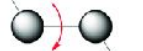
bond



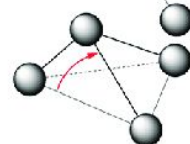
angle



dihedral



improper



van der Waals



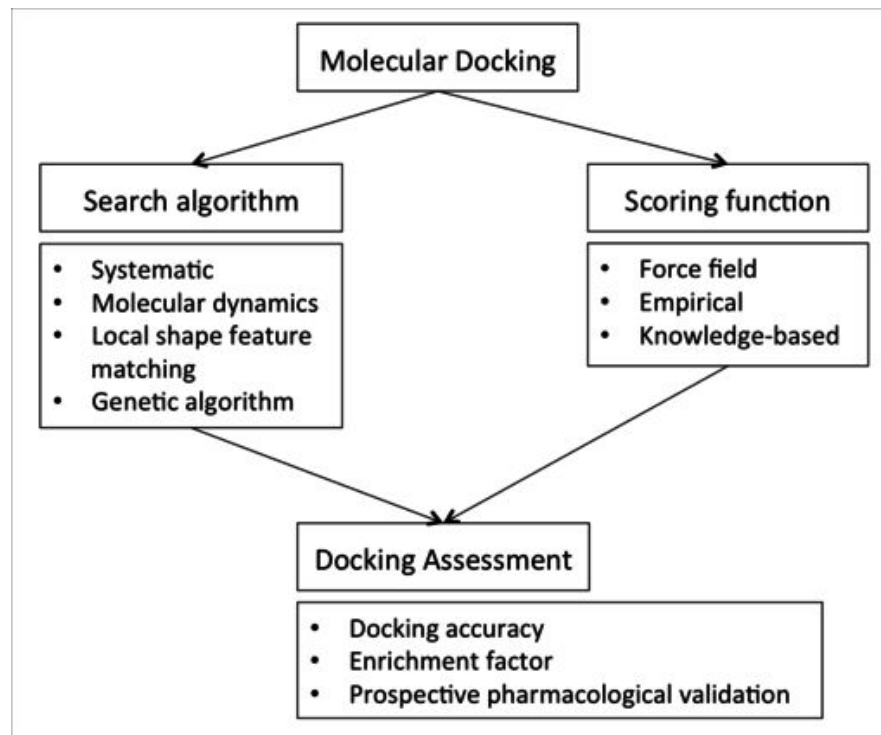
electrostatic



Breaking down the generic docking algorithm

How to find the lowest ΔG 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

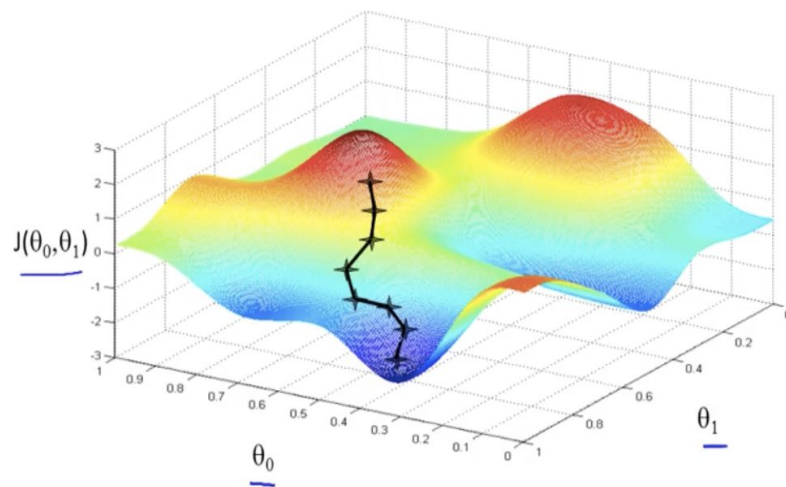


Autodock Vina - Gradient Descent

Mcule uses Autodock Vina, which uses a modified form of gradient descent to minimize ΔG .

Gradient Descent Steps:

1. Pick a starting pose for the ligand
2. Calculate ΔG and the gradient of ΔG (the slope of ΔG in every coordinate direction)
3. Move down the gradient (the slope) towards a ΔG 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 ΔG 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**

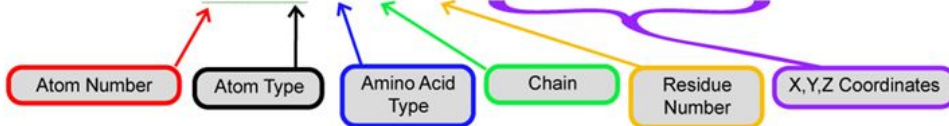


File/Data Types and other Resources

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 Mucle

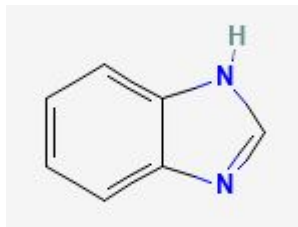
```
ATOM 1132 NH1 ARG A 149 31.814 -31.597 16.995
ATOM 1133 NH2 ARG A 149 32.203 -32.934 18.816
ATOM 1134 N ASN A 150 29.346 -24.359 18.812
ATOM 1135 CA ASN A 150 28.480 -23.190 18.933
ATOM 1136 C ASN A 150 28.606 -22.168 17.808
ATOM 1137 O ASN A 150 27.803 -21.276 17.678
ATOM 1138 CB ASN A 150 28.732 -22.524 20.282
ATOM 1139 CG ASN A 150 28.284 -23.389 21.447
ATOM 1140 OD1 ASN A 150 27.205 -23.981 21.430
ATOM 1141 ND2 ASN A 150 29.110 -23.463 22.466
ATOM 1142 N LEU A 151 29.629 -22.313 16.996
ATOM 1143 CA LEU A 151 29.868 -21.415 15.894
ATOM 1144 C LEU A 151 29.953 -22.205 14.597
ATOM 1145 O LEU A 151 30.149 -23.422 14.614
ATOM 1146 CB LEU A 151 31.208 -20.735 16.100
ATOM 1147 CG LEU A 151 31.436 -19.884 17.337
ATOM 1148 CD1 LEU A 151 32.846 -19.333 17.256
```



A screenshot of the RCSB PDB website. The header includes the RCSB PDB logo, the tagline "184202 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education", a search bar, and navigation links like "Advanced Search" and "Browse Applications". Below the header is a navigation bar with "Developers: Join the RCSB PDB Team" and "Explore Open Positions". The main content area features a "Welcome" message, a "Deposit" button, a "Search" button, a "Visualize" button, and an "Analyze" button. A section titled "A Structural View of Biology" provides information about the database's resources. To the right, there is a "November Molecule of the Month" section featuring a 3D molecular model.

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

- <http://biosig.unimelb.edu.au/pkcsm/>
- pkCSM is a 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

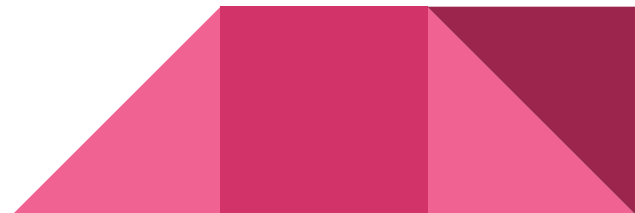
Logistics:

- 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)
 - PubChem - database containing a bunch of compounds and their data (www.pubchem.ncbi.nlm.nih.gov)
 - And of course, Mcule (www.mcule.com). For Mcule account, just put your school name for organization.
- 