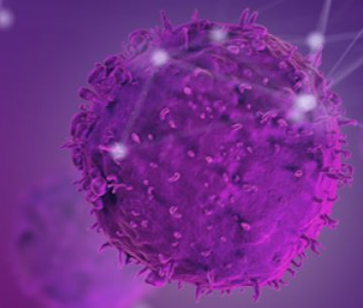
ardigen

Artificial Intelligence & Bioinformatics
for Precision Medicine

group of machine
gmum
learning research

25 November 2019



ardigen

group of machine gmum learning research

Theoretical introduction to ML in chemistry and drug design

Introduction to cheminfromatics

Drug design process

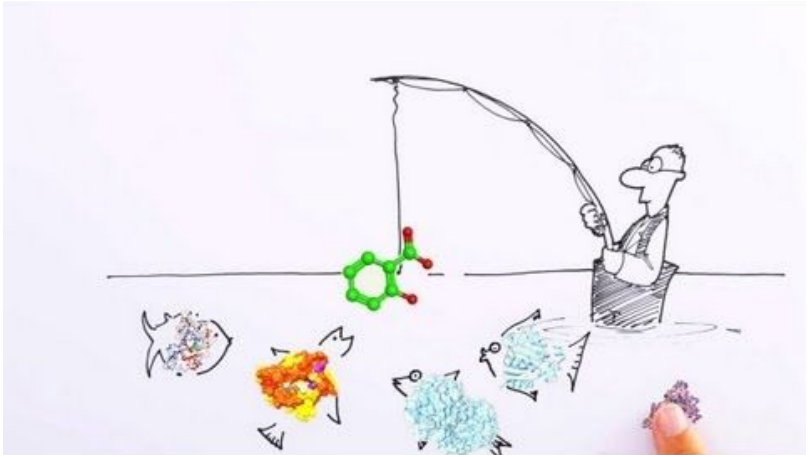
From data to drug candidates

- High cost (~$10B)
- High risk (90% fail rate from preclinical to clinical)
- Long process (~10 years)
- Requires extensive infrastructure and personnel

Target Identification

The early stages focus on finding the proper protein target, modulation of which would influence the disease of interest. Some of the basic methodologies used in that process are:

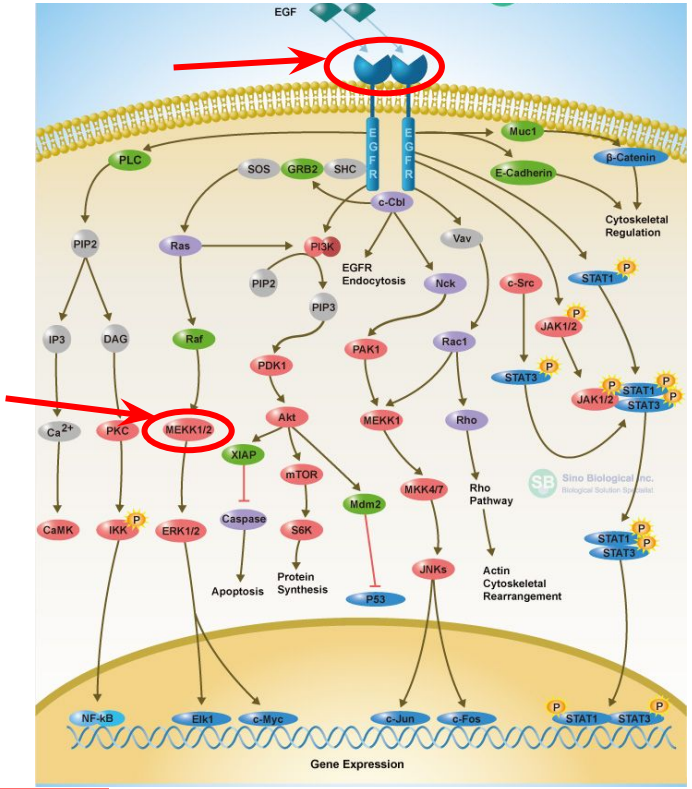
- Data mining, literature searching
- Cell-based assays
  - Knock-out, knock-in
  - Protein interaction detection
  - Expression analysis
- Protein structure acquisition/prediction
- Early docking



Target validation

Once we identify the pathway that influences the disease, we need to find which part of the pathway is a viable target for our new drug.

- Druggability
  - Protein family analysis
  - Protein sequence analysis
  - Binding pocket search
  - Fragment docking
- Tool compound design



Lead identification

Searching for a proper haystack

Now that we found our target, we need to find a compound that will bind and modulate its activity. First, we need to acquire a large library of compounds to be able to select from. To do this, we can use a range of sources:

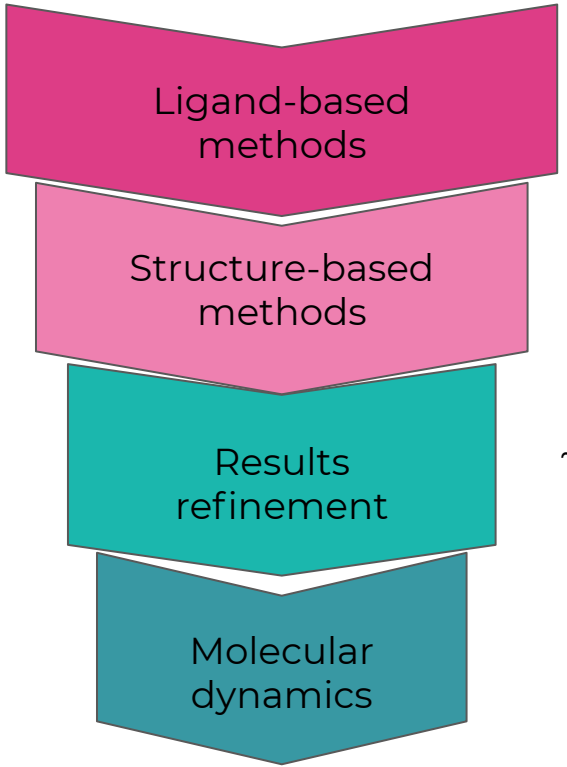
- Combinatorial libraries
  - Synthetically available - prone to bias
- Commercial vendors
  - Ensured acquisition - higher price, limited chemical space
- Brute-force compound generation (~10^{60} compounds available)
  - Broad spectrum of compounds - High computational cost, low synthetic accessibility

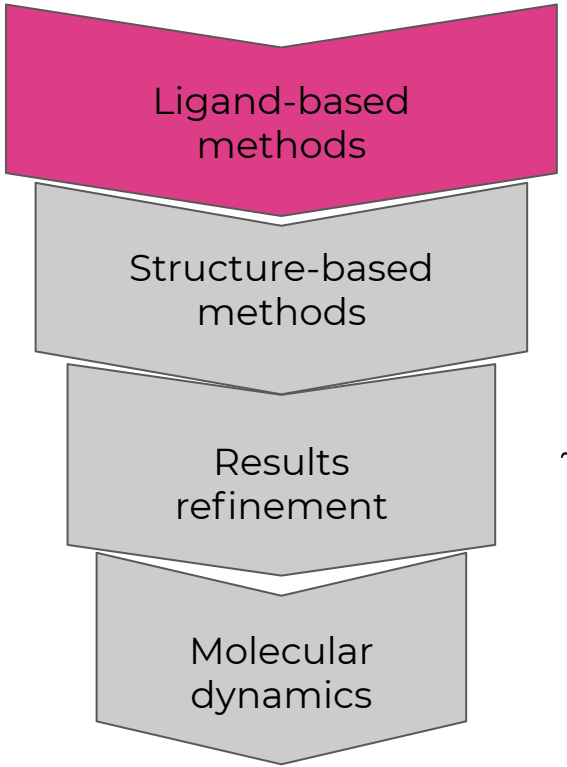
Lead Identification

Searching for a needle in a haystack

Now that we have our initial library of compounds, we need to screen it for potential positive hits, that is structures that express some activity towards our designated target.

- Filtering in a sequential manner
- Each next step more computationally complex than the previous one
- May proceed to next phases after any of the steps





~ 1M compounds

~ 10k compounds

~ 1000 compounds

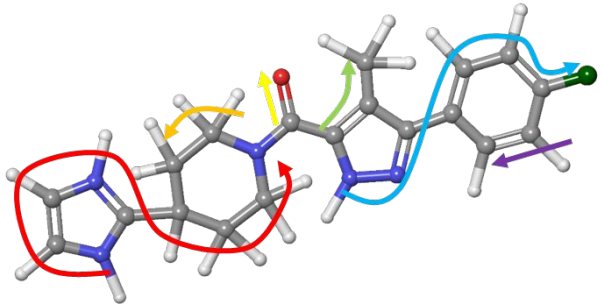
~ 10 compounds

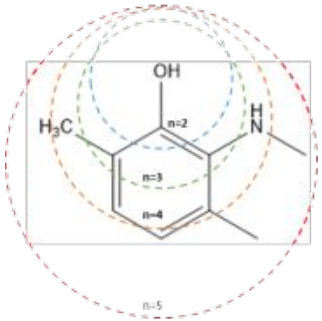
Compound representation

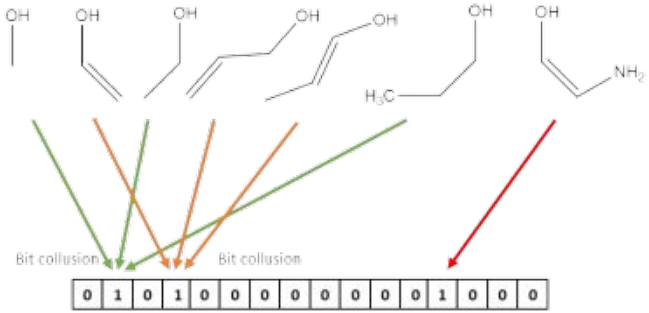
Make the computer represent

The chemical compounds, while easily understandable by us, are not much so for the computer. Thus arises the need to properly represent chemical structures in a way, that's interpretable by a machine. We can, e.g., use various representations:

- SMILES
- Structural fingerprints
- Physicochemical fingerprints
- Pharmacophores
- Spectrophores
- AI-based representations





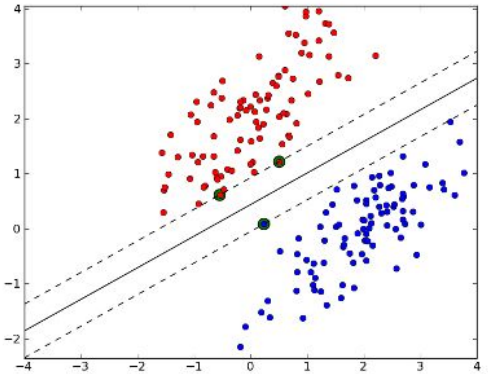


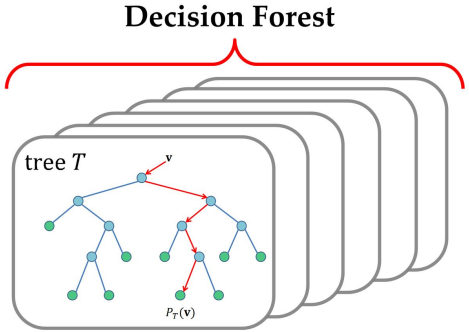
Screening methods

Make the computer understand

Once we have our representation, we can use it to screen the compounds using various criteria. We can build various models that split the compound set into subsets:

- Similarity searching
  - Very simple, not very useful
- Machine learning
  - Quite powerful
  - Multitude of methods (Support Vector Machines, Random Forests)
- Clustering
  - Allows to group compounds based on structure or other properties
- AI-based classifiers



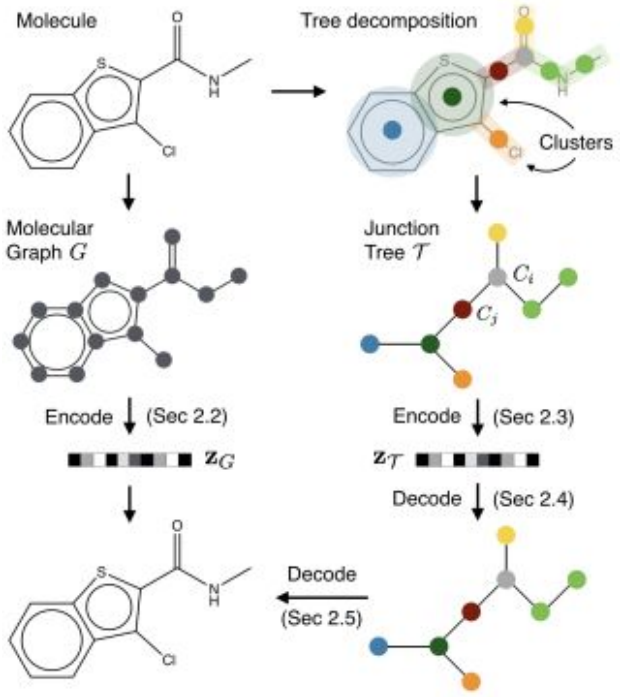


Compound representation

Teaching the computer to represent and understand

Instead of using heuristic methods with pre-defined features, we can use deep learning approaches to determine the optimal representation for a particular task.

- Variational Autoencoders (VAE)
  - Learn the proper representation
  - Build a “latent space” - chemical space
- Recurrent Neural Networks
  - Use SMILES representation
  - Learns the proper sequence of atoms
- Graph Convolutional Networks

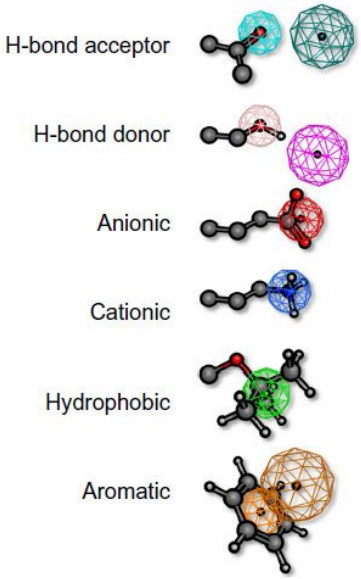


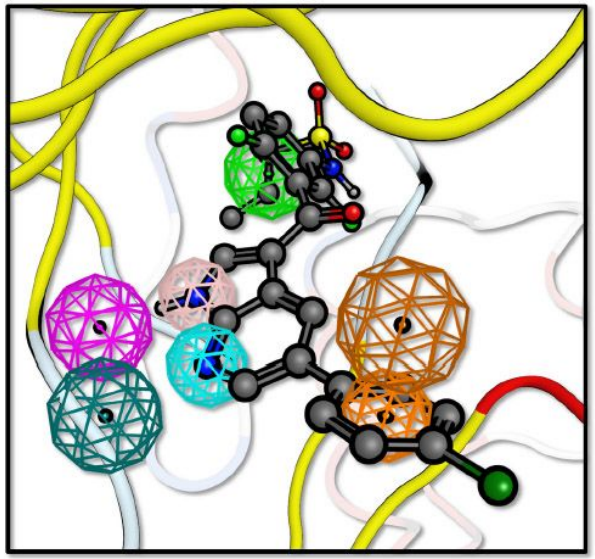
Other methods of compound representation

Pharmacophores

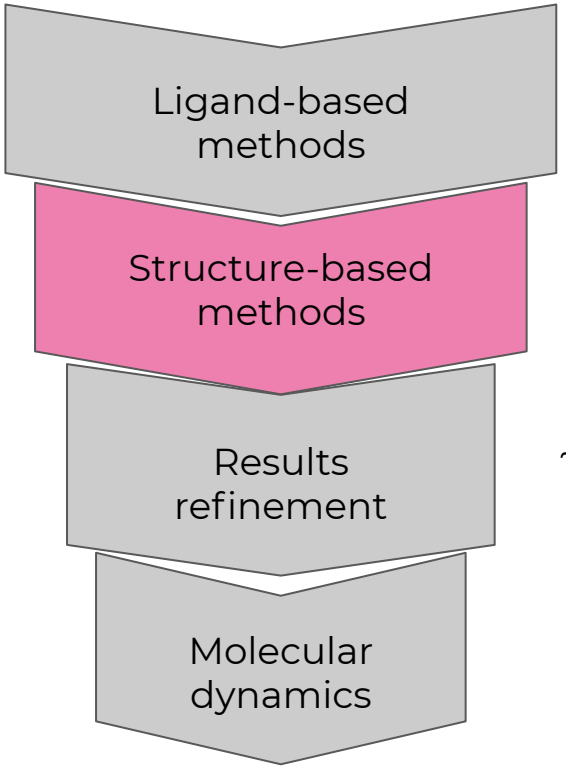
Since it is believed that the shape of the compound is one of the factors involved in its activity, we can try to encode the shape, or at least the position and correlations of the important parts of the molecule.

- Requires proper 3D structure of the compound
- Mapping in 3D is difficult
- Easy to overfit/underfit





From: Dove Medical Press



~ 1M compounds

~ 10k compounds

~ 1000 compounds

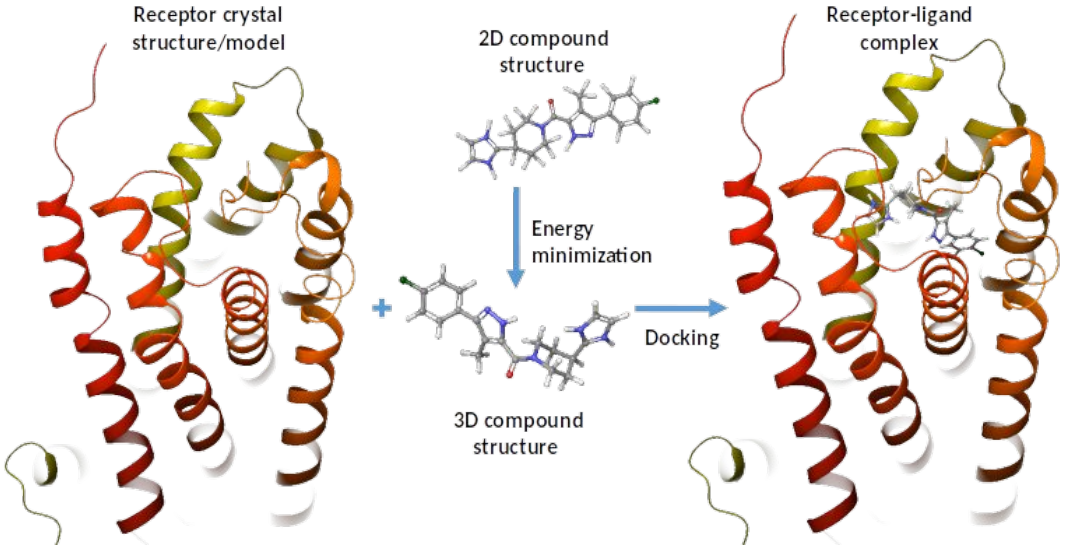
~ 10 compounds

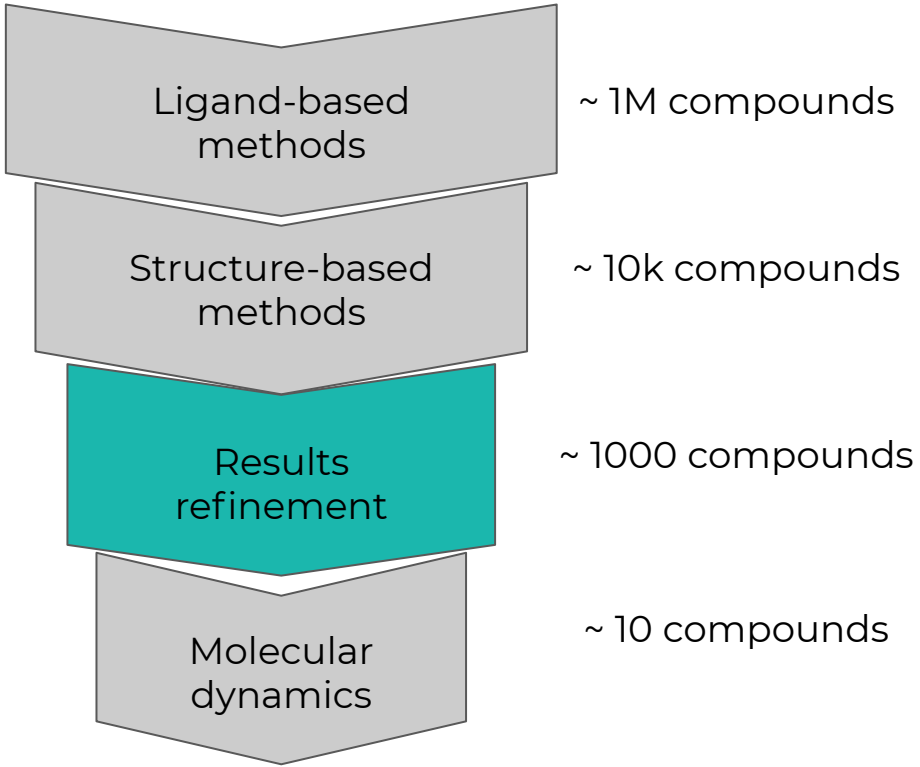
Compound docking

Fitting the key to the lock

Having the 3D structure of the target receptor and the structure of a molecule, we can “fit” the compound inside the binding pocket of the protein.

- Requires 3D structure of the protein
- Decent accuracy
- Allows to filter out false positives
- Relatively cheap



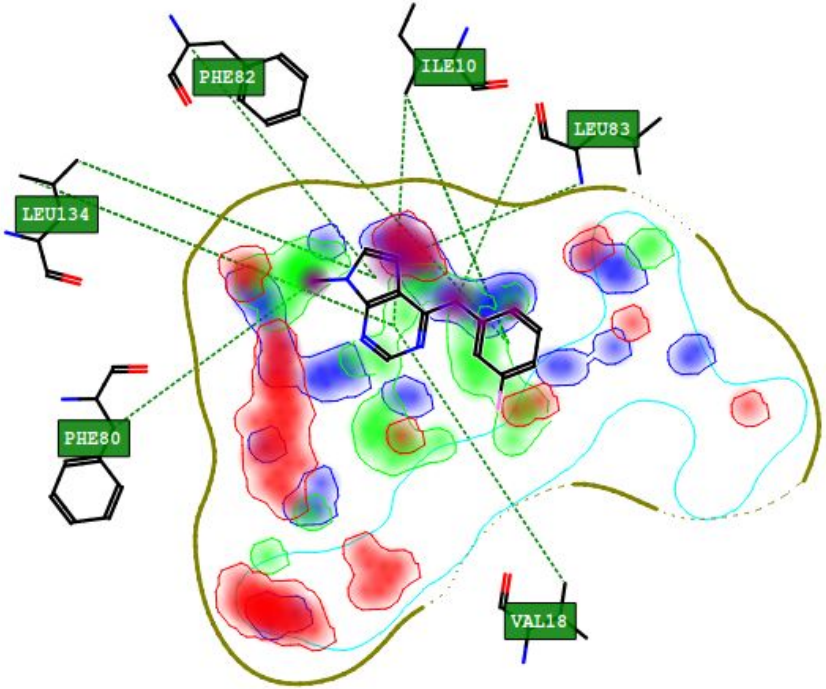

Ligand-based methods
~ 1M compounds
Structure-based methods
~ 10k compounds
Results refinement
~ 1000 compounds
Molecular dynamics
~ 10 compounds

Interaction fingerprints

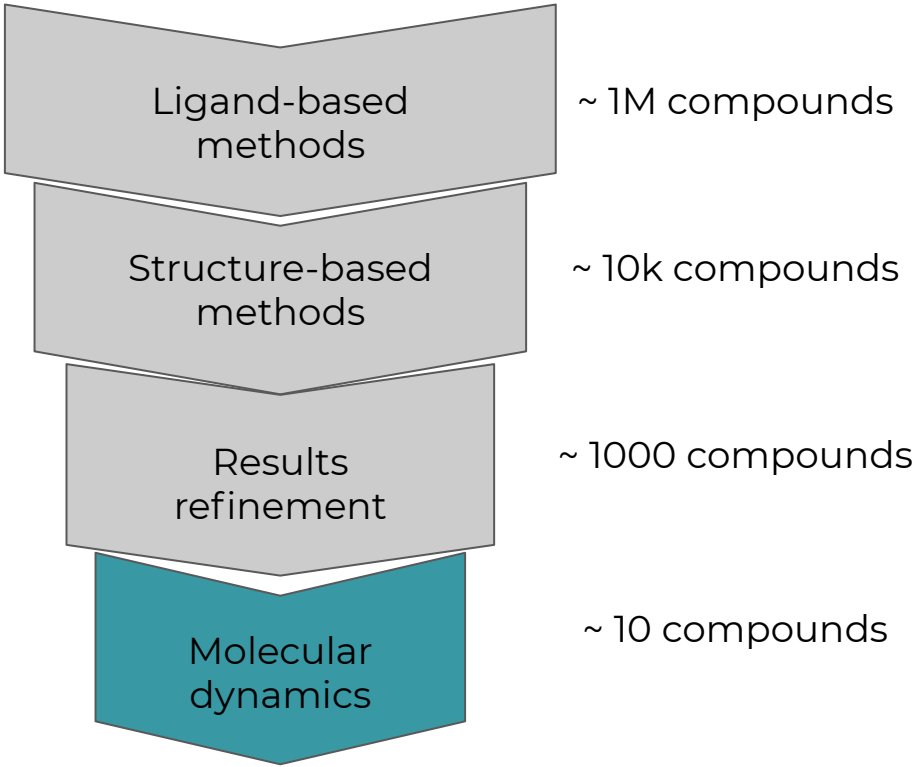
Does the molecule push the buttons

Once our compound is docked we can check, whether it interacts properly with the protein.

- Filter by positive/negative interactions
- Build models based on interaction profiles
- Provide some insight into compounds functionality



From: moldiscovery.com

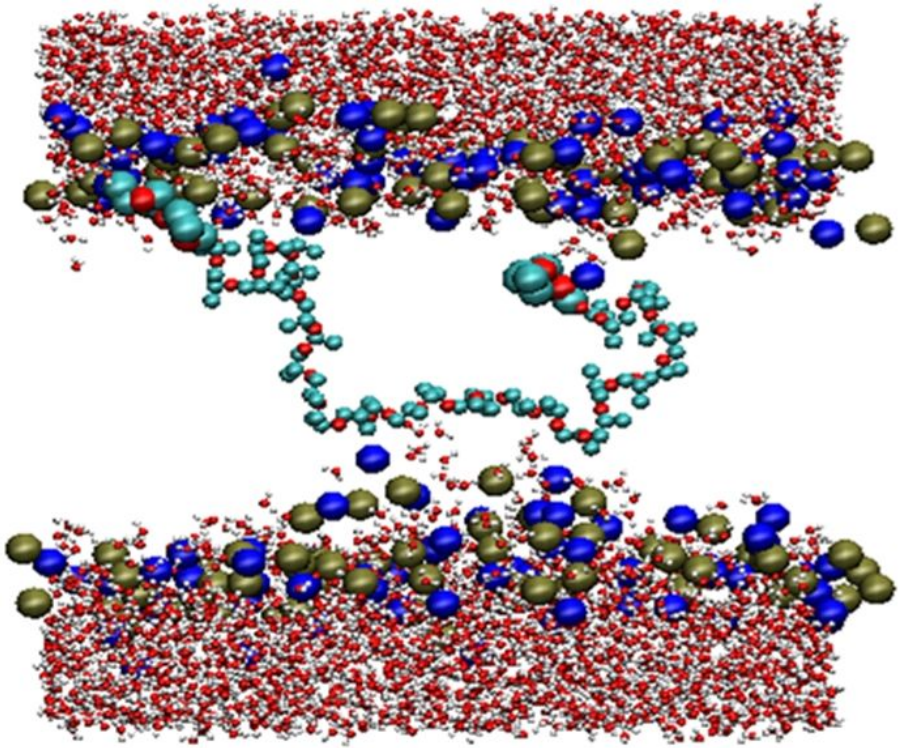

Ligand-based methods
~ 1M compounds
Structure-based methods
~ 10k compounds
Results refinement
~ 1000 compounds
Molecular dynamics
~ 10 compounds

Molecular dynamics

A more intimate look into molecular systems

To ensure that our compound is active, we may employ the molecular dynamics methodology. It is a very costly approach, as all forces between all atoms in a system must be computed on each time step.

- Extremely expensive
- Provides insight into:
  - Functionality
  - Stability
  - Pharmacokinetics
  - Structural changes



From: comp-physics-lincoln.com

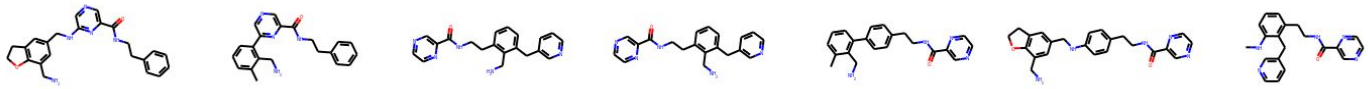
Compound selection

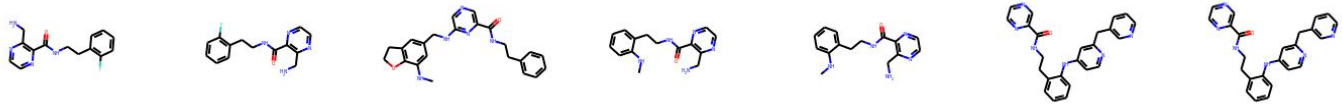
Picking the needle

Lead optimization

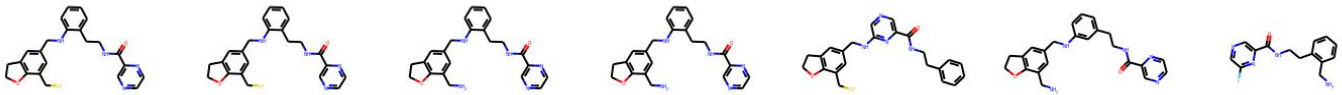
Sharpening the needle

- Activity





- Selectivity



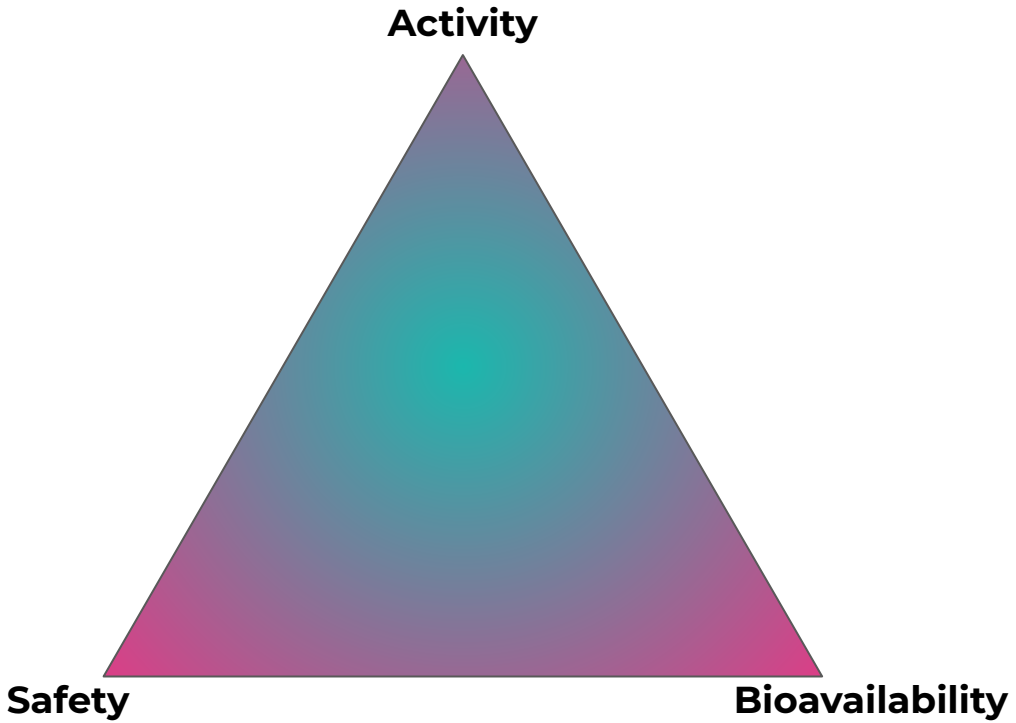
ADME/Tox

- Adsorption
  - Compound must be bioavailable
- Distribution
  - Compound must reach the designated target
- Metabolism
  - Compound must be stable
- Excretion
  - Compound must not leave residual waste
- Toxicity
  - Compound must be safe

Lead optimization

How to achieve perfection

- Expert-based
  - Medicinal chemist bias
- Brute-force
  - Time/trash
- Combinatorial libraries
  - Time/bias
- **Variational Auto Encoders (VAE)**
  - **Structure divergence**
- **Reinforcement Learning**
  - **Hard to implement**



AI-based methods

Fine-tuned chemical space exploration

- VAE
  - Build an abstract chemical space using known compounds
  - Sample the space for new compounds
  - Use previously built classifiers to filter

- Reinforcement learning
  - Teach the algorithm to build molecules from scratch
  - Apply reward for compounds with good properties
  - Gradient ascent towards the best molecule

Pre-clinical trials

Pre-clinical and clinical trials

While testing on live organisms, the AI-based methods can help in analysing the data, finding relations between responders and non-responders, determining the proper dosage, and in a number of other tasks.