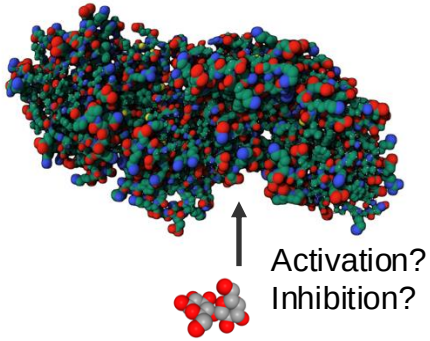
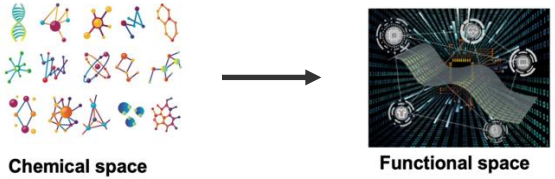
Heinrich Heine
Universität Düsseldorf

Why do we want to make predictions for molecules?

Motivation

- Molecules are the building blocks of life
  - Understanding their functions and interactions is relevant for many different natural sciences
    - Biology
    - Biochemistry
    - Medicine
    - Environmental Sciences
  - It has also many import industrial applications
    - Designing new materials with desired properties
    - Designing new pathways of chemical reactions for the synthesis of different substances
      - Drugs
      - Biofuels
      - Chemicals



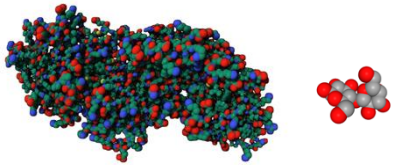


Small molecule images: pubchem.ncbi.nlm.nih.gov; Protein image: https://www.rcsb.org/3d-view ; Material design: https://doi.org/10.3390/ma15051811

Experimental validation of small molecule-protein interaction

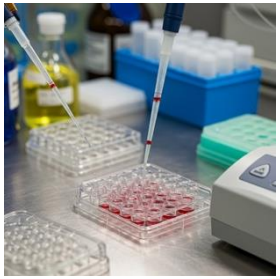
Step 1: Molecule Acquisition

Timeline: 1 – 4+ weeks



Step 2: Assay Selection & Optimization:

Timeline: 1 – 4 weeks



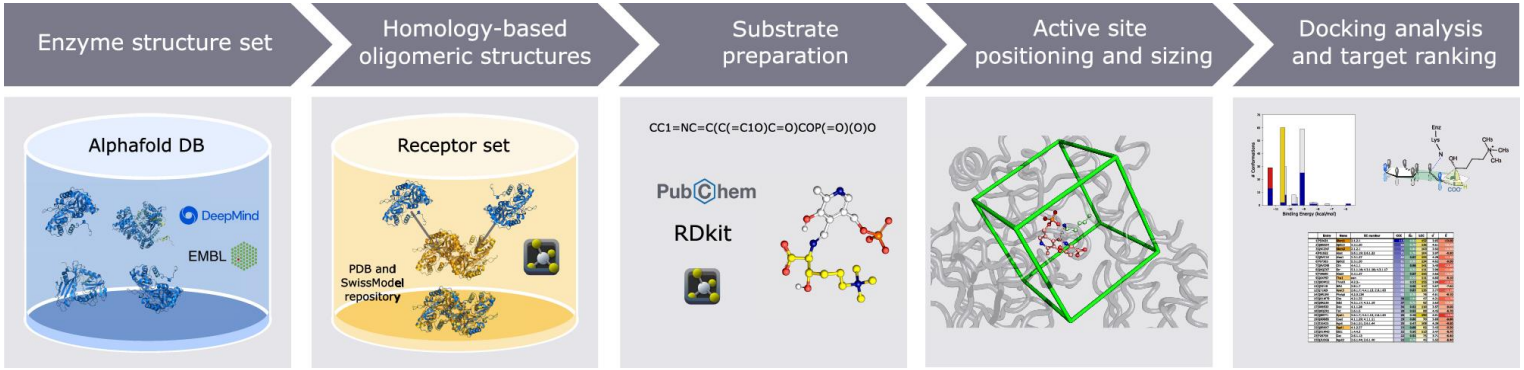
Step 3: Experiment execution:

Timeline: 1 – 2 weeks

Step 4: Results Analysis and Interpretation:

Timeline: Up to 2 weeks

Computational molecular docking analysis



- In molecular docking simulations, we try to place the small molecule into the protein in many different ways by rotating, flipping, and shifting it to see how it fits.
- For each position, the program calculates a score that tells us how good the fit is.

Image copied from: https://doi.org/10.1038/s41467-024-47466-3

Methods to determine molecule properties

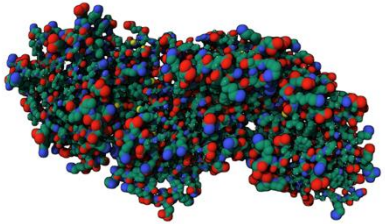
- Experiments 
  - ⊕ Advantage: Usually very accurate
  - ⊖ Disadvantages: Often time-consuming, difficult, and expensive

- Non-ML computational methods: 
  Simulate the behavior of molecules using theoretical principles, empirical data, and statistical analysis
  - ⊕ Advantages: Speed, Costs, Accessibility
  - ⊖ Disadvantages: Accuracy, Complexity Limits, Data Requirement

- Machine Learning models: 
  - ⊕ Advantages: Handling Complexity, Detecting Patterns, Speed, Flexibility
  - ⊖ Disadvantages: Accuracy, Data Dependency, Overfitting

Examples for predicting molecule properties

- Protein property prediction:

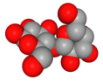


Optimal pH

Effect of mutations

...

- Small molecule property prediction:



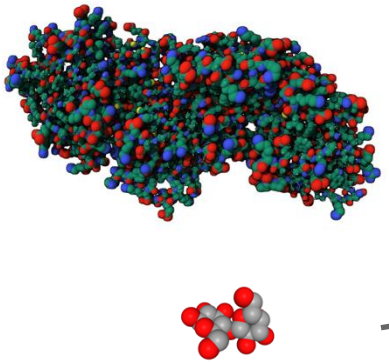
Toxicity

...

Small molecule image: pubchem.ncbi.nlm.nih.gov/compound/6134#section=3D-Conformer; Protein image generated with: https://www.rcsb.org/3d-view

Examples for predicting molecule properties (2)

- Protein-small molecule interaction predictions:



Binding? Yes/No

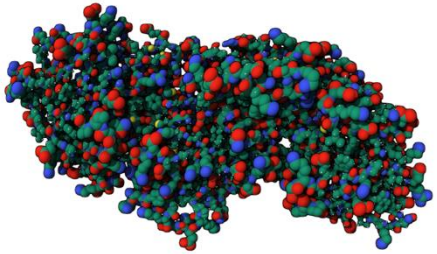
Inhibition?

...

Small molecule image: pubchem.ncbi.nlm.nih.gov/compound/6134#section=3D-Conformer; Protein image generated with: https://www.rcsb.org/3d-view

On what kind of molecules will we focus?

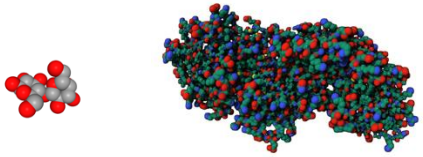
- We will mostly focus on biomolecules:
  - Micromolecules (Small Molecules):
    - Vitamins
    - Amino Acids
    - Monosaccharides
    - Nucleotides
  - Macromolecules
    - **Proteins**
    - Nucleic Acids (DNA and RNA)
    - Polysaccharides



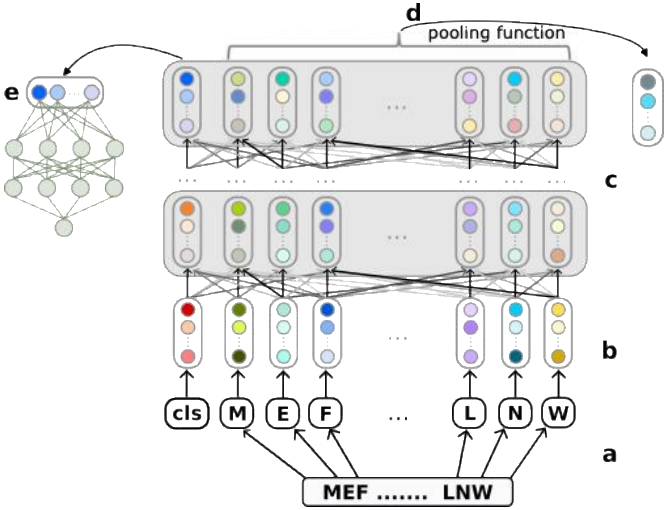
Small molecule image: pubchem.ncbi.nlm.nih.gov/compound/6134#section=3D-Conformer; Protein image generated with: https://www.rcsb.org/3d-view

Course content (1)

- Biological / chemical background

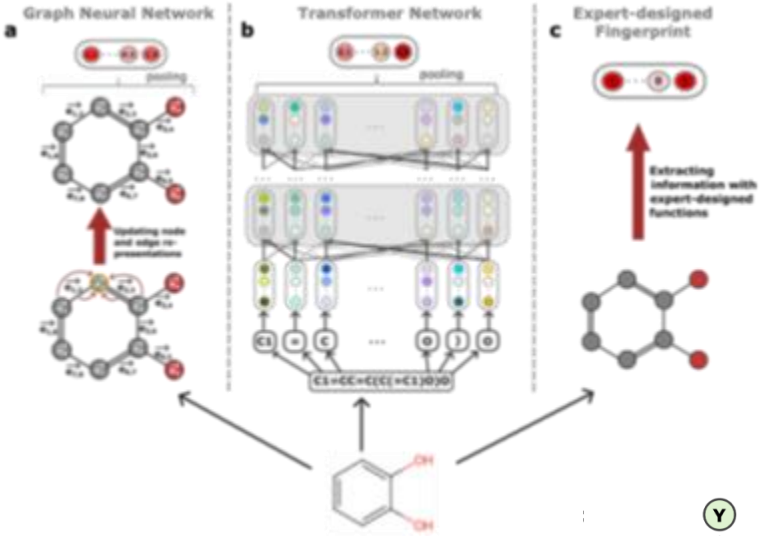


- Transformer Networks
  - Architecture
  - Training
    - Self-Supervised
    - Supervised: Fine-tuning pre-trained models
- Transformer Networks for proteins (large molecules)

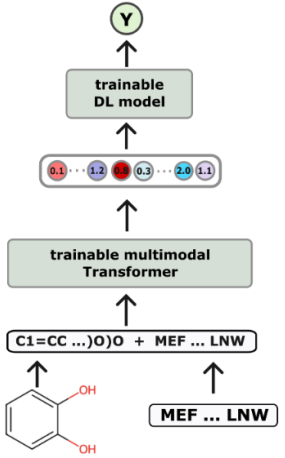


Course content (2)

- Encoding small molecules
  - Transformer networks
  - Graph neural networks (GNNs)
  - Traditional methods

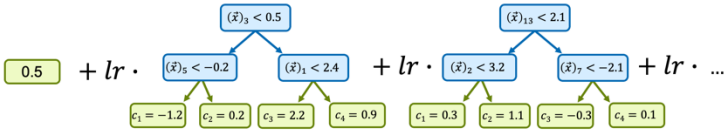


- Multimodal Transformer Networks:
  - Applying a single Transformer to multiple types of input data
  - Predicting interactions between proteins and small molecule

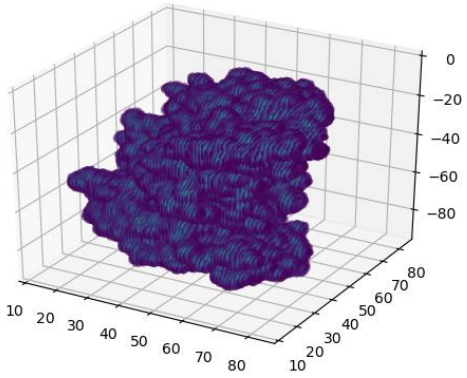


Course content (3)

- Visualizing what Transformer Networks learn
- Gradient Boosting decision tree models

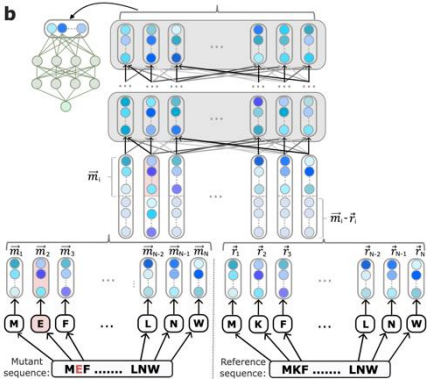


- Utilizing protein 3D structures to predict protein function
  - Protein structure prediction methods
  - Encoding protein structure with
    - Graph Neural Networks
    - 3D Convolutional Neural Networks



Course content (4)

- Protein engineering methods



- Challenges and best practices
- Optimizing large language models

google-research/
tuning_playbook

A playbook for systematically maximizing the performance of deep learning models.

14 Contributors | 10 Issues | 9 Discussions | 28k Stars | 2k Forks