Rethinking Drug Discovery using AI

LEIDEN
DRUG DEVELOPMENT
CONFERENCE

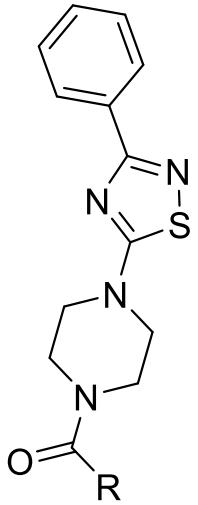
19 September 2024

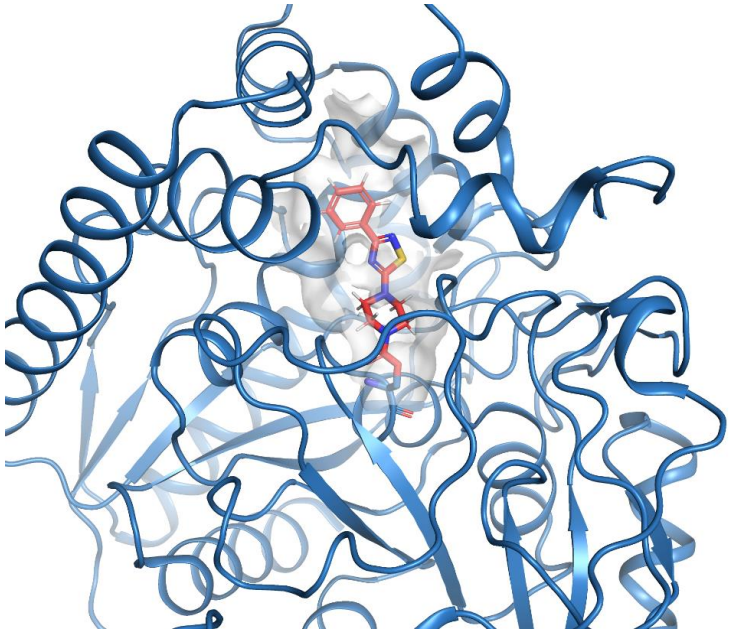
Anthe Janssen
Leiden University

Small molecule drug

- Over-simplification for today:

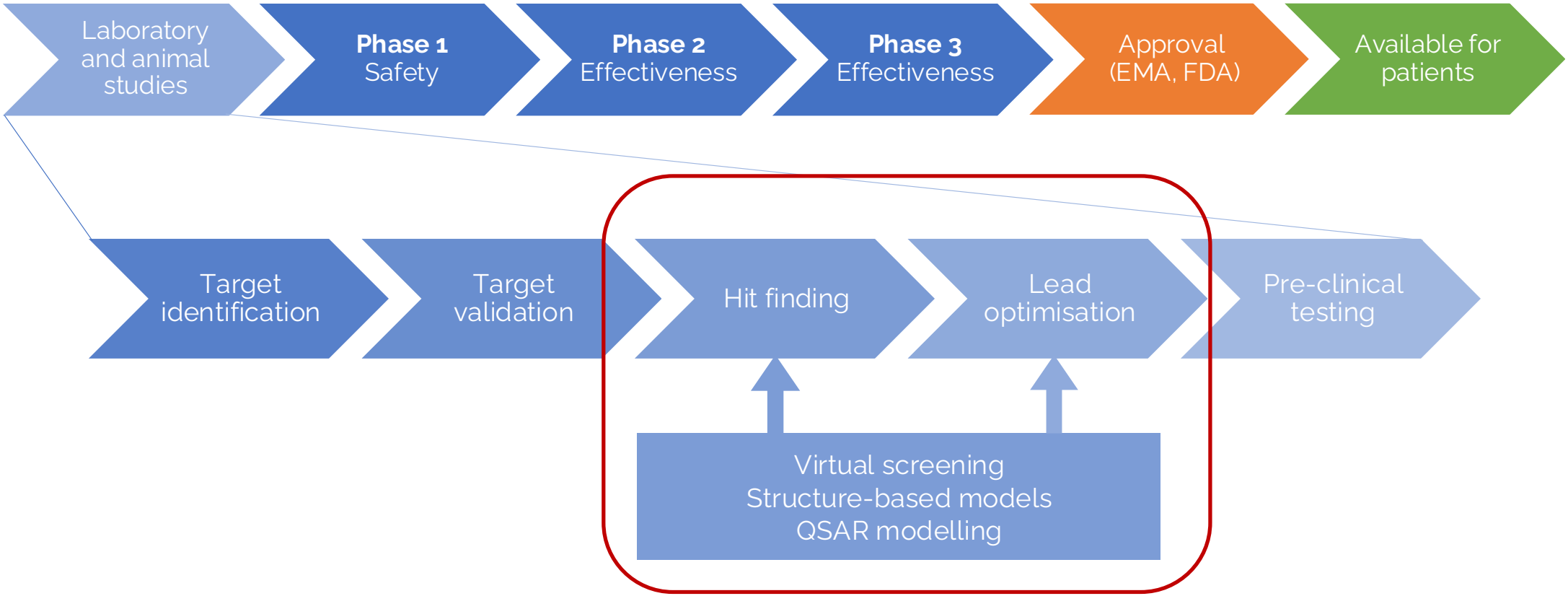
  Small molecule that binds a protein and blocks or alters the function

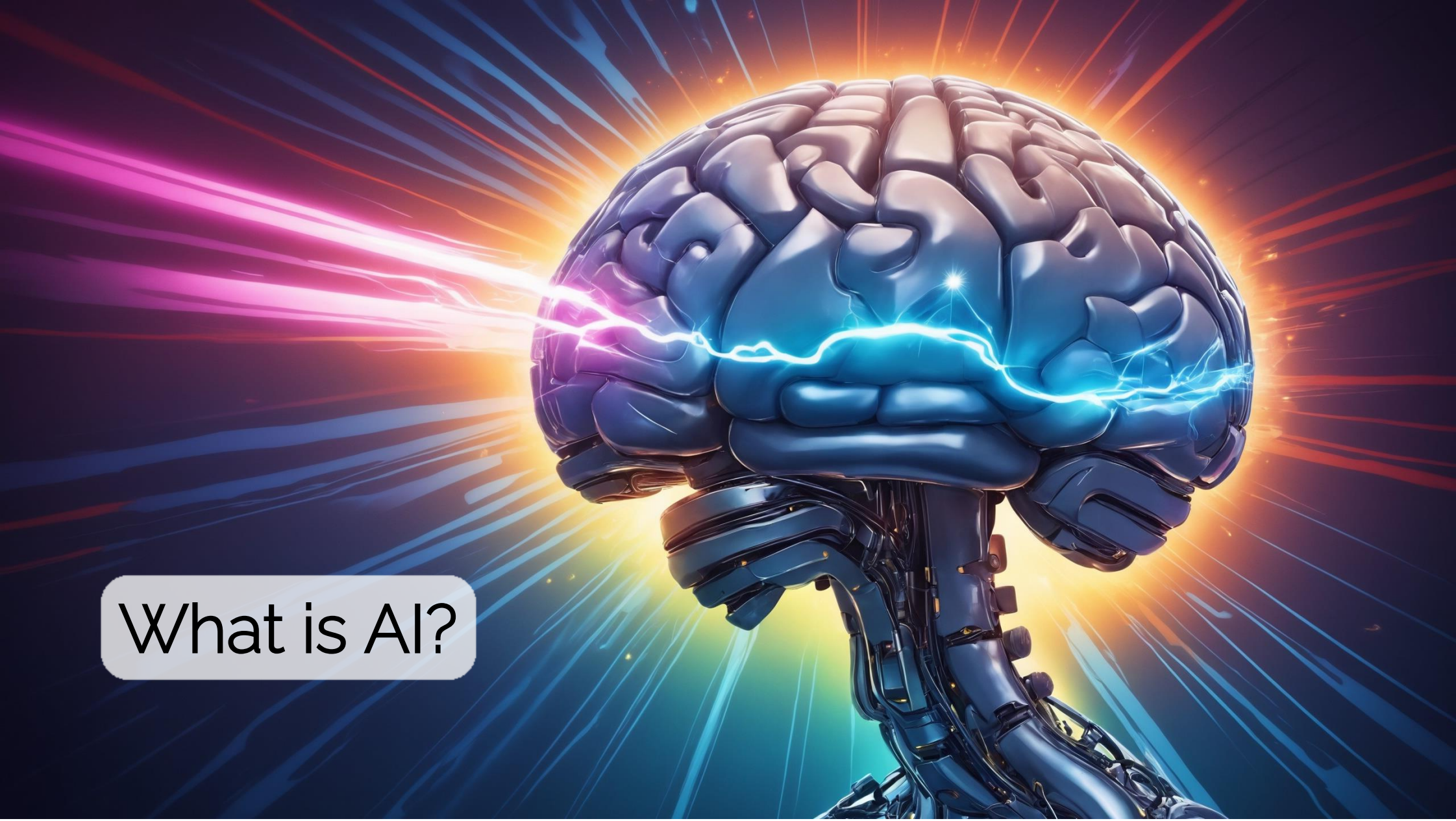




PDB: 4HBP

Chemistry & Drug Discovery




What is AI?

ChemGPT 1.0

How can I help you today?

Design a drug
to cure all types of cancer

Plan an itinerary
from Leiden to the PAC symposium

Help me pick
an outfit that looks good in the lab

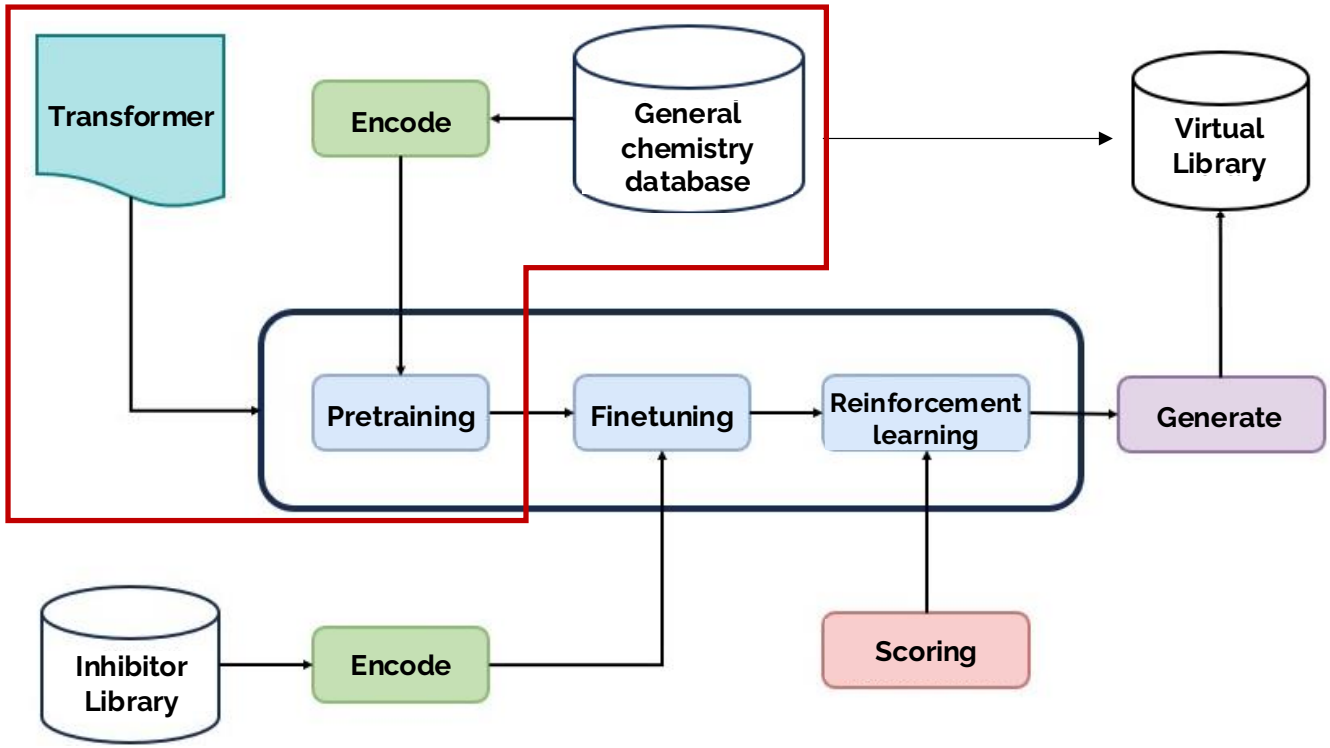
Suggest some codenames
for this cool new compound I made

ChemGPT

- Aim is to somehow generate the perfect molecule

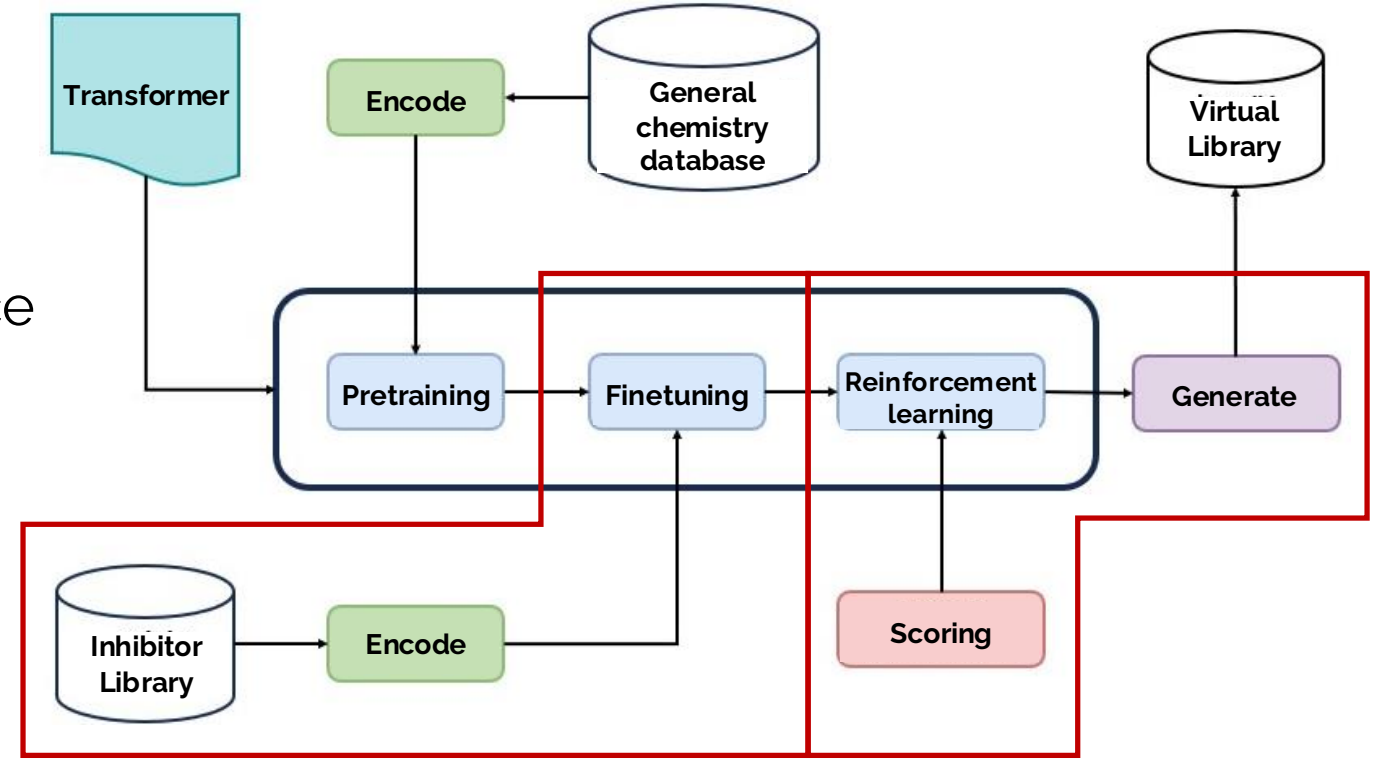
ChemGPT

- Train language model for ‘all chemistry’



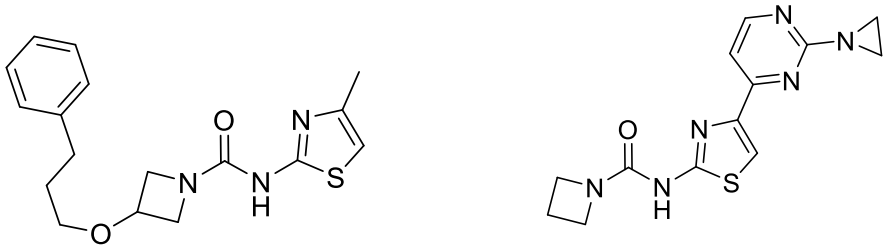
ChemGPT

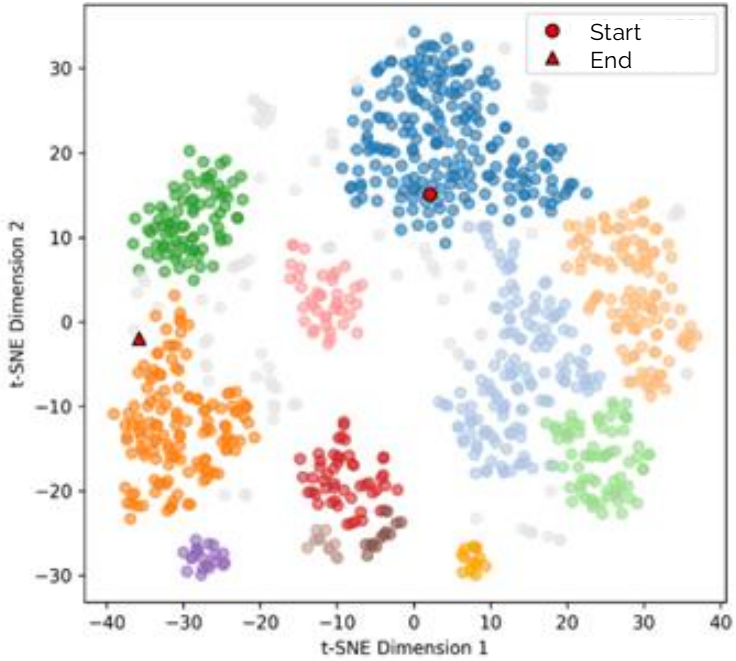
- Train language model for ‘all chemistry’
- Learn protein preference
- Generate and rank new molecules *in silico*



ChemGPT

- Use it like auto-complete
- Starting molecule ⟶ Variations





- 1000s of variations can be ‘made’ in minutes

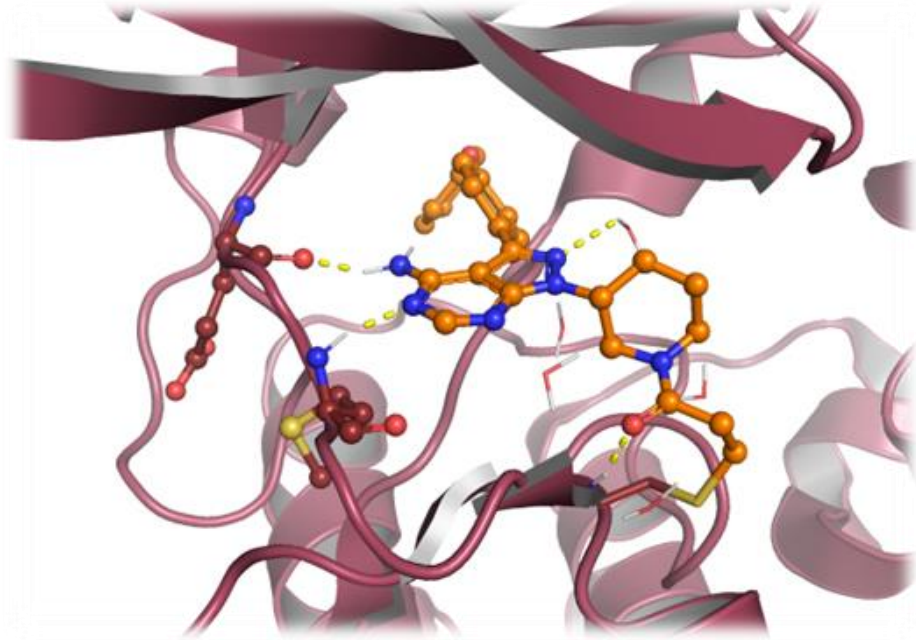
Is it any good?

- Well . . . maybe
- Downsides:
  - Look-a-likes
  - Synthesisability issues
  - Stability issues
- How to know it really binds the target?

Protein affinity prediction

- How well does a compound bind its target?

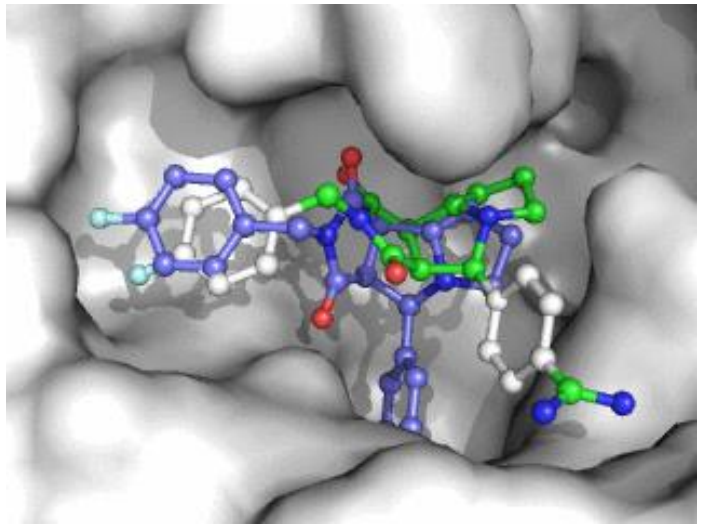
- Specific protein family: kinases
- Extremely important for oncology
- 80 kinase targeting drugs on the market
- ~~Combined ~$60 bln in 2023~~
- Lifesaving treatments



Ibrutinib bound to BTK, PDB: 5P9J

3D complex: docking

- Inhibitor should bind into active site
- Physics-based optimisation: docking
- AI-based tool: DiffDock



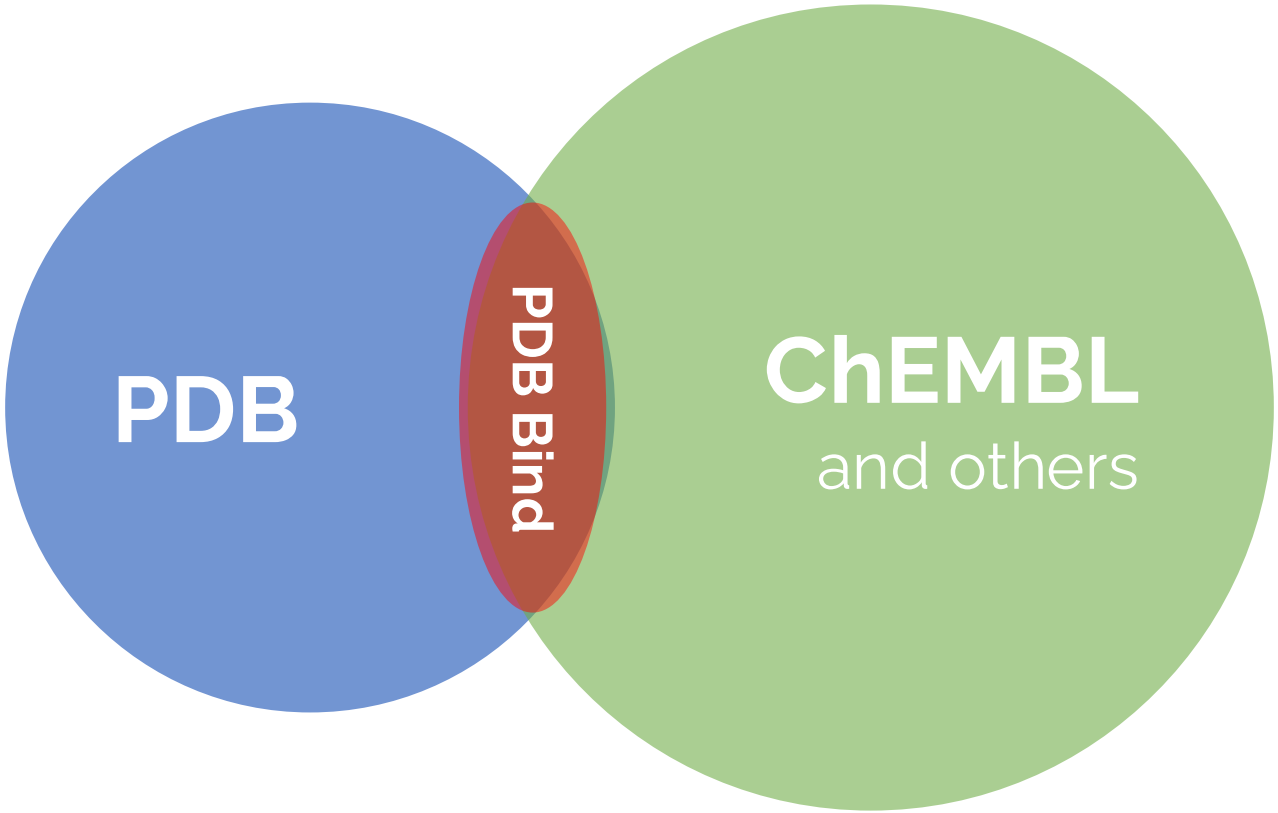
https://arxiv.org/abs/2210.01776

Two steps of docking

- The *pose finding* of docking is good

- Scoring is the main challenge

The problem: Data



Two steps of docking

- The *pose finding* of docking is good

- Scoring is the main challenge
- Deep learning needs more data

Two steps of docking

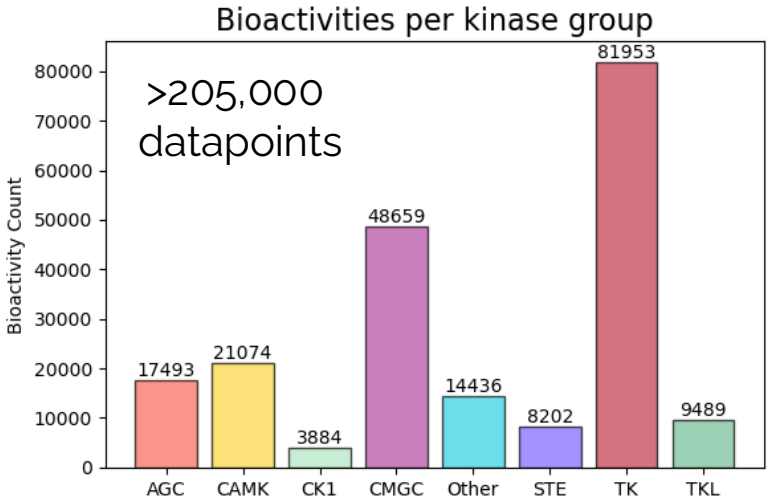
- The *pose finding* of docking is good

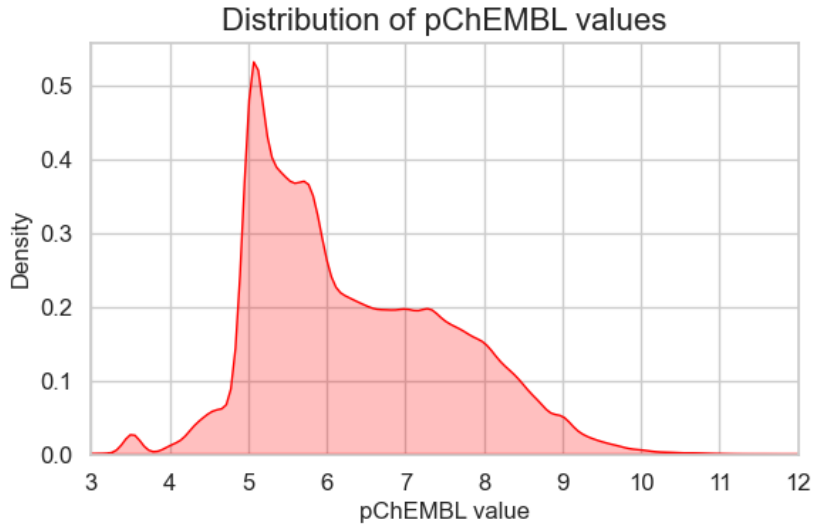
- Scoring is the main challenge
- Deep learning needs more data

PDB + ChEMBL + docking = “predicted PDB bind”

Step 1: Gather literature data

- Will function as 'ground truth' for AI to learn from
- Large datasets available: Papyrus





Béquignon *et al.* Journal of Cheminformatics **15** (2023)
ebi.ac.uk/chembl/
pubchem.ncbi.nlm.nih.gov/

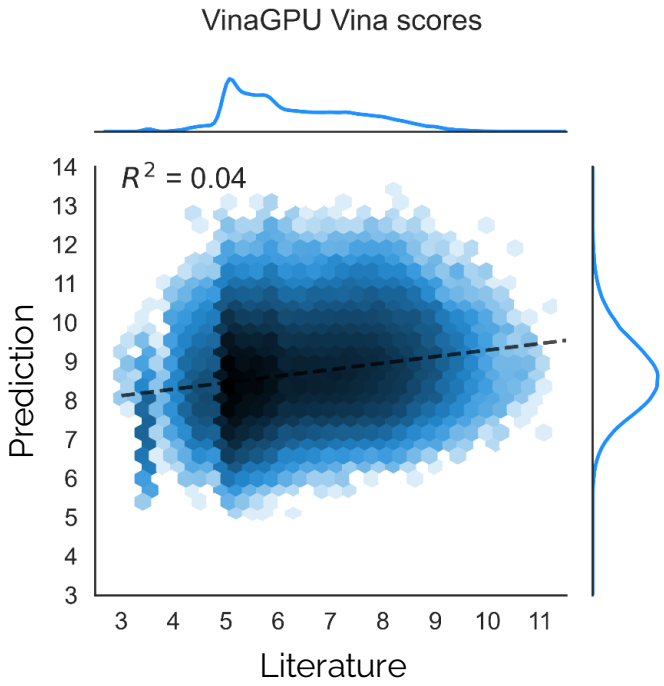
Step 1: Gather literature data

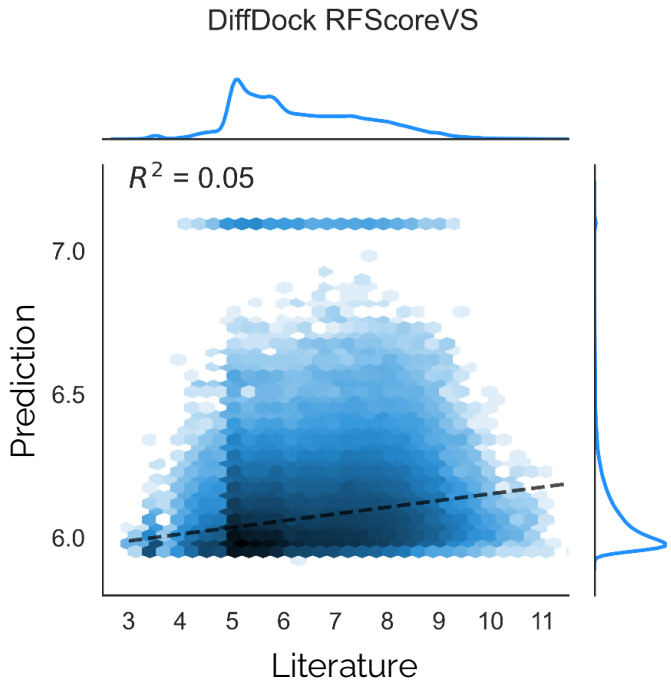
- Protein structures from KLIFS (kinase subset of PDB)
- Extremely well annotated

https://doi.org/10.1093/nar/gkaa895

Predict affinity

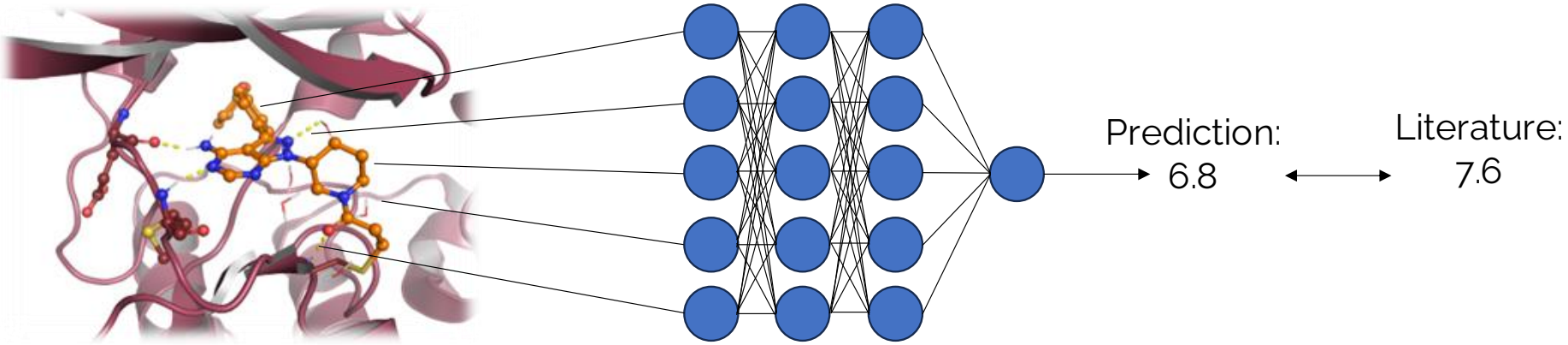
- Current ‘state-of-the-art’:





Better model needed!

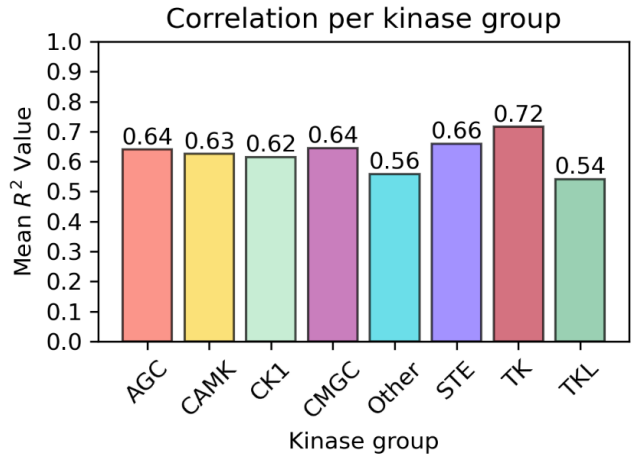
- With all this data we can do Machine Learning

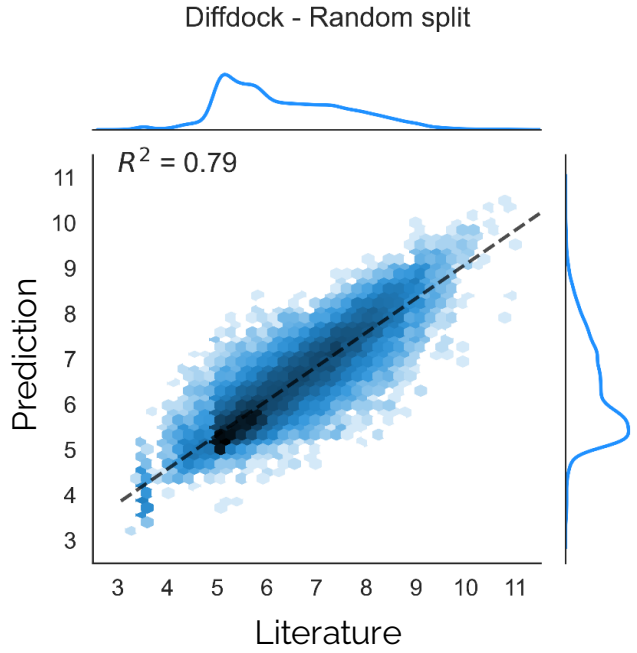


Ibrutinib bound to BTK, PDB: 5P9J

Can it do better?

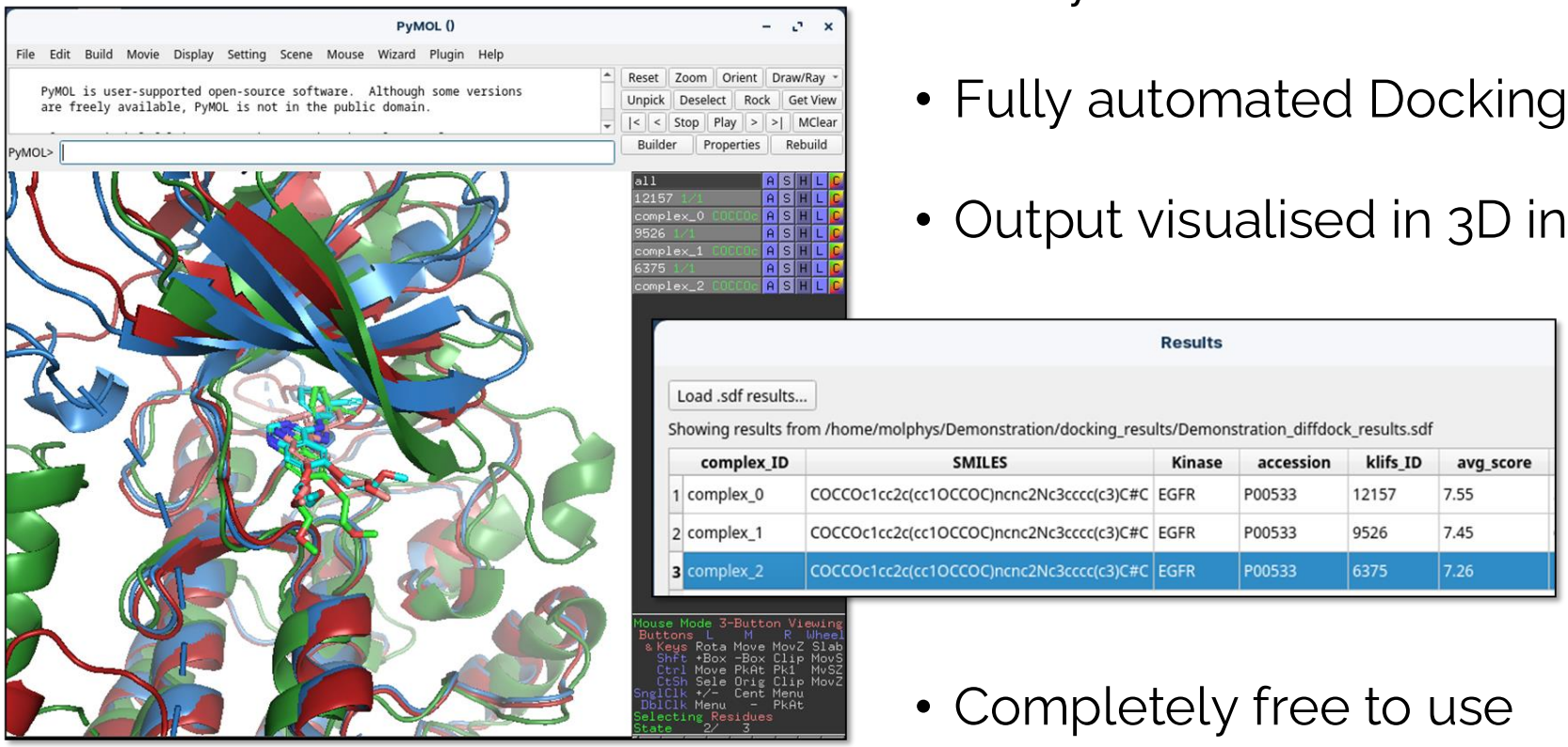
- Yes!
- R^2 > 0.6 for 40% of the kinases





https://chemrxiv.org/engage/chemrxiv/article-details/660568d39138d231618b4e50

KinaseDocker²



	complex_ID	SMILES	Kinase	accession	klifs_ID	avg_score
1	complex_0	COCCOc1cc2c(cc1OCCOC)ncnc2Nc3cccc(c3)C#C	EGFR	P00533	12157	7.55
2	complex_1	COCCOc1cc2c(cc1OCCOC)ncnc2Nc3cccc(c3)C#C	EGFR	P00533	9526	7.45
3	complex_2	COCCOc1cc2c(cc1OCCOC)ncnc2Nc3cccc(c3)C#C	EGFR	P00533	6375	7.26

- Easily installed
- Fully automated Docking and AI prediction
- Output visualised in 3D in PyMOL
- Completely free to use

https://chemrxiv.org/engage/chemrxiv/article-details/660568d39138d231618b4e50

What's next?

- Better, faster models
- Combine with compound generators

- Nobody expected ChatGPT, maybe ChemGPT is closer than we think?

Acknowledgements

Molecular Physiology

- Jordy Schifferstein
- Patrick Rehorst
- Frans ter Brake
- Prof. Mario van der Stelt

Computational Drug Discovery

- Andrius Bernatavicius
- Prof. Gerard van Westen

Paper on ChemRxiv

Code on GitHub