Bologna – Industrial Chemistry Department “Toso Montanari”
02.12.2022

Drug Discovery and Cheminformatics: discovering new drugs in the Big Data era

Vincenzo Palmacci
PhD Candidate MSCA

My studies at UNIBO

BSc Industrial Chemistry

- Organic synthesis
- Physical Chemistry
- Industrial chemistry

2013-2018

Bachelor thesis
Oct 2016 – Nov 2017

MSc Bioinformatics

- Protein design
- Analysis of Biological data
- Machine Learning

2018-2020

Master thesis
Erasmus + Internship

European Molecular Biology Laboratory

BIOSAXS

Artificial neural networks for protein's shape prediction

After UNIBO: Advanced Machine Learning for Drug Discovery

Machine Learning Research

- Artificial Intelligence for Drug Discovery
- Representation Learning
- Explainable AI

COMP3D group

- Computational Drug Design
- Prediction of assay interference

2021 - 2024?

Molecular AI

- Retrosynthesis planning
- De novo molecule generation

Today's Lecture

Drug Discovery

- The Drug Discovery Pipeline
  - Early Drug Discovery

Cheminformatics

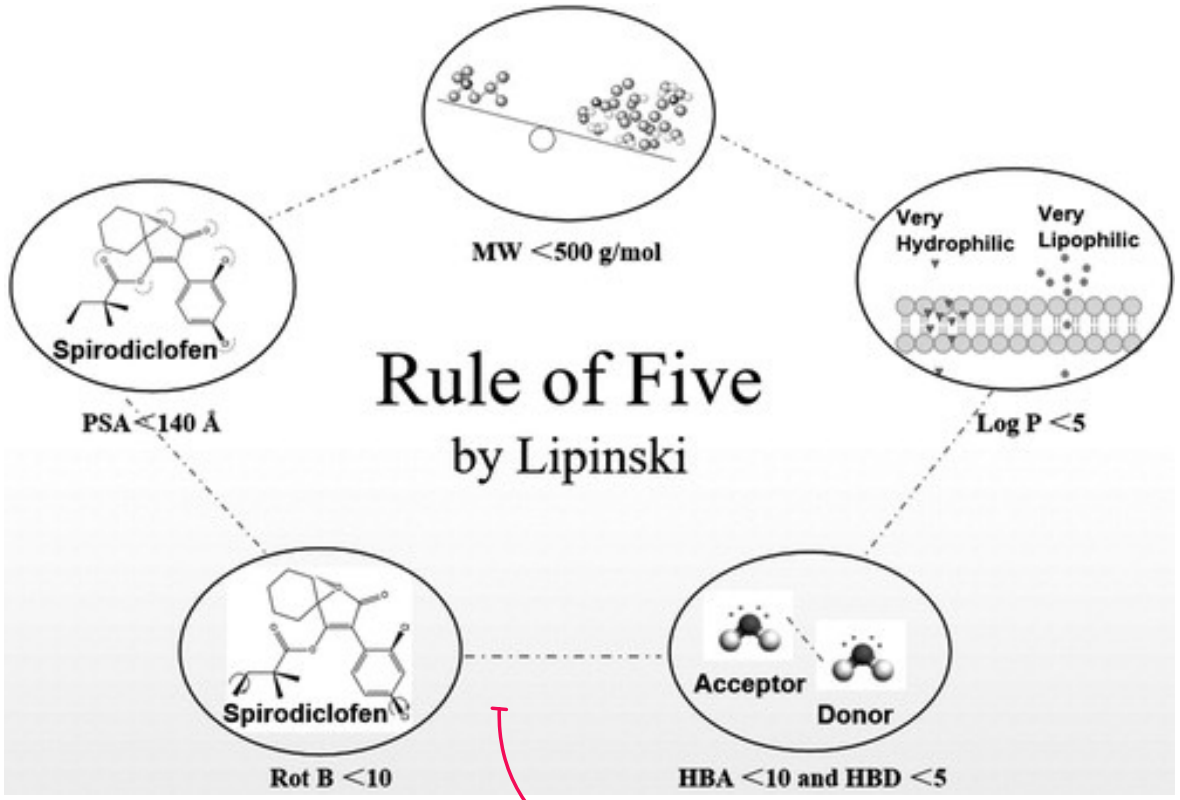
- Big Data in Chemistry
- Introduction to Cheminformatics: meaning and fundamentals
- Introduction to Machine Learning and Deep Learning
- Example: Toxicity prediction

Computational Drug Discovery

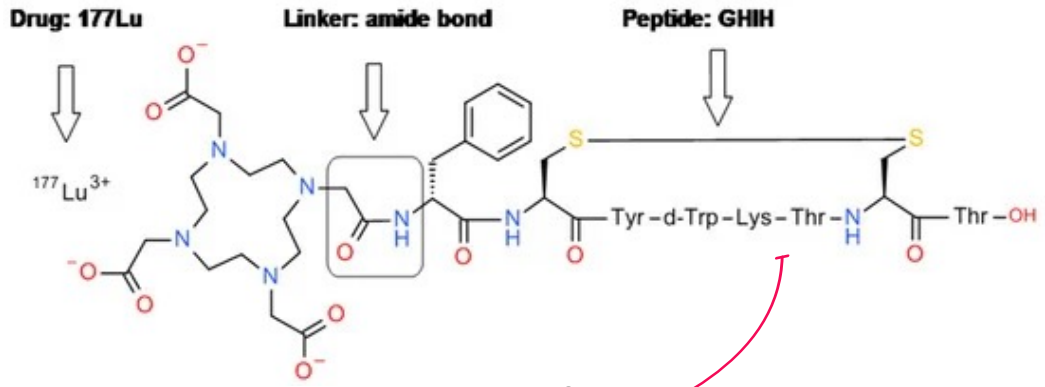
- Cool stuff: How Artificial Intelligence helps Drug Discovery

Illustration by Michele Marconi: https://www.nature.com/articles/d41586-018-05267-x

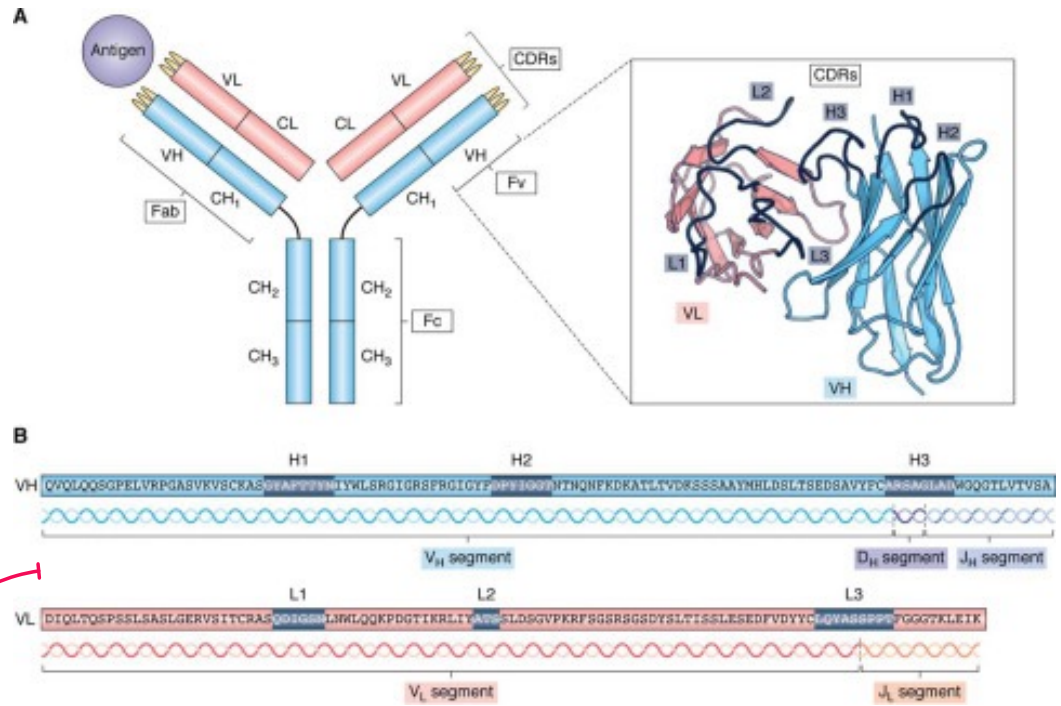
What is a drug?



Small Molecules



Peptides



Antibodies

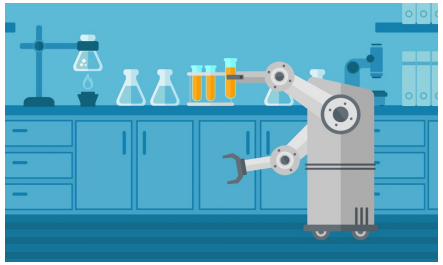
Drug Discovery: from target identification to drug approval

Start: 5 Millions molecules

End: 1 molecule

Lead discovery & optimization

- HighThroughputScreen
- Hit selection
- Lead optimization (1-2 years)



PreClinic

- *In vitro* studies
- Animal Studies

Clinical Trials

- **Phase I**: ~100 volunteers
- **Phase II:** ~300 patients
- **Phase III:** ~ 5000 patients

Approval

FDA

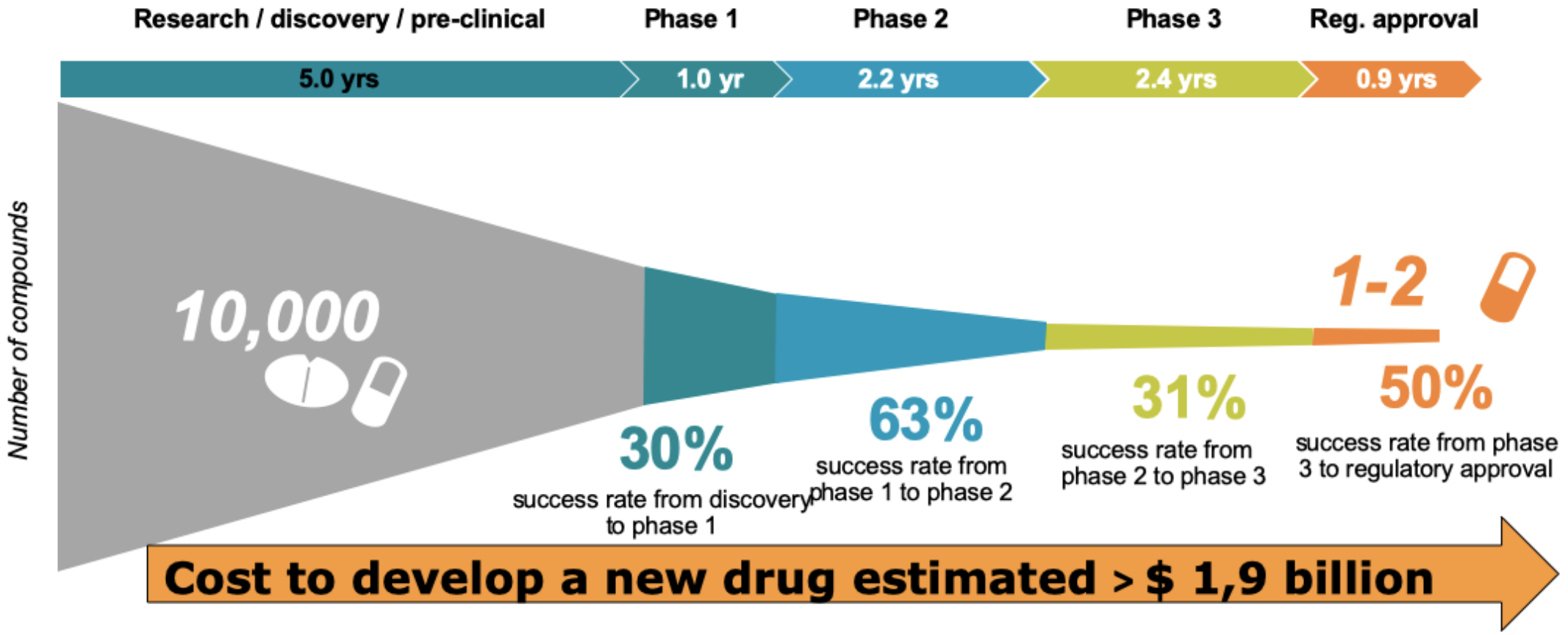
EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

0 — 4 — 6 — 8 — 10 — 12

Years

Total cost: ~ 2 Billion $

Pipeline steps success rate



Source: DiMasi et al, Journal of Health Economics, January 2016

Target identification and validation

What is a Target?

PDB ID: 1C3S

- **Biological target**: anything in a living organism that changes behaviour upon binding with other entities
- Common targets: Proteins, RNA or DNA
- Must have a role in the development of a disease

Identification and validation

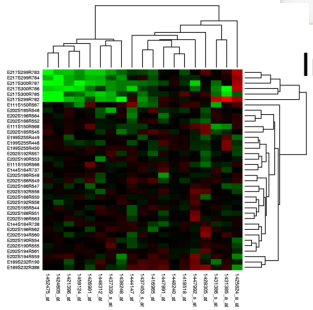
Target ID technologies

Literature and Patents search

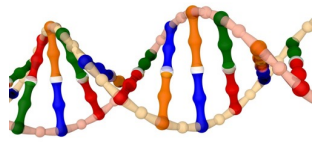
Bioinformatics

In vivo experiments

Expert knowledge

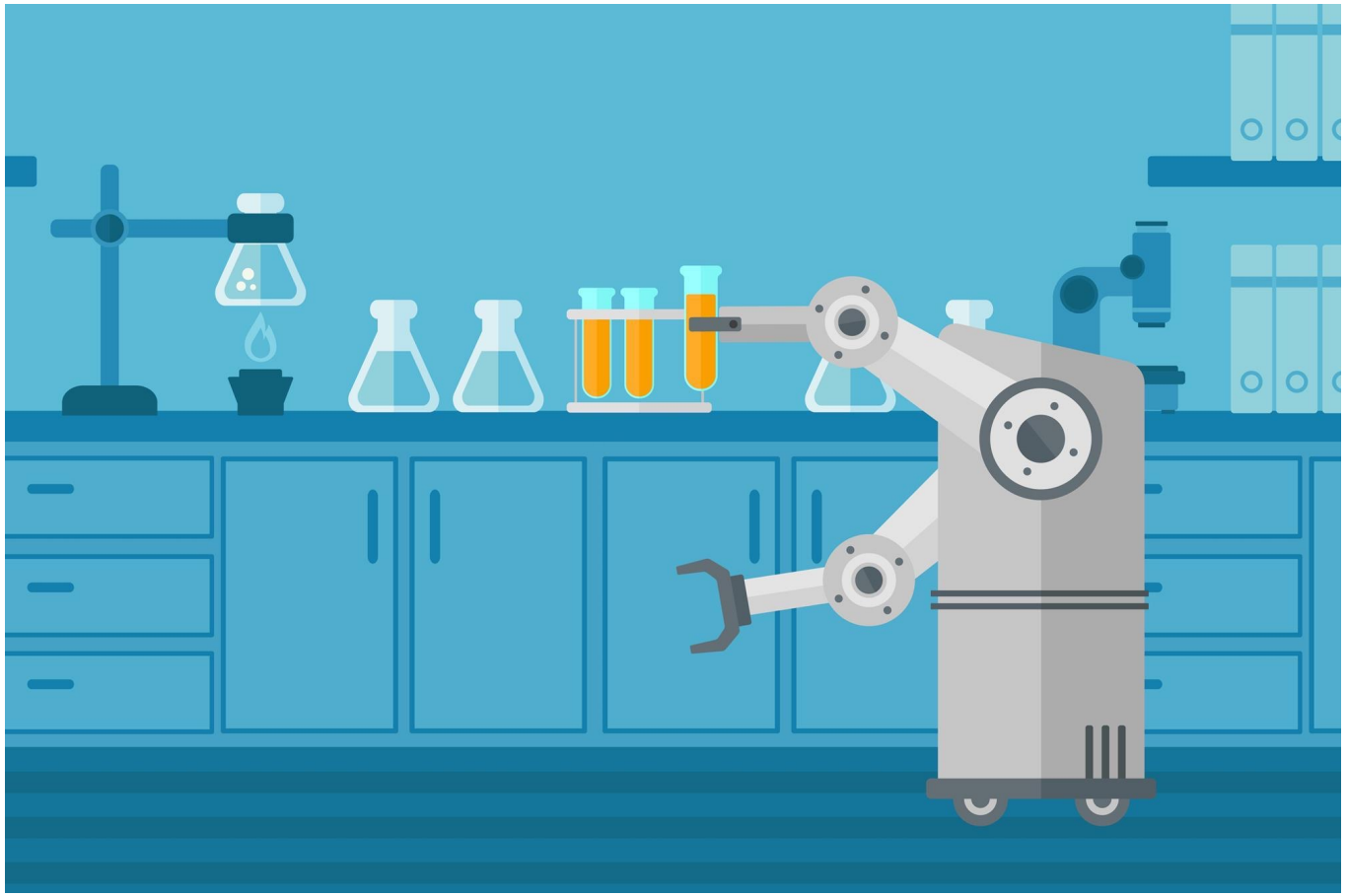


Gene expression



Gene silencing

Lead Discovery: High Throughtput screening (HTS)



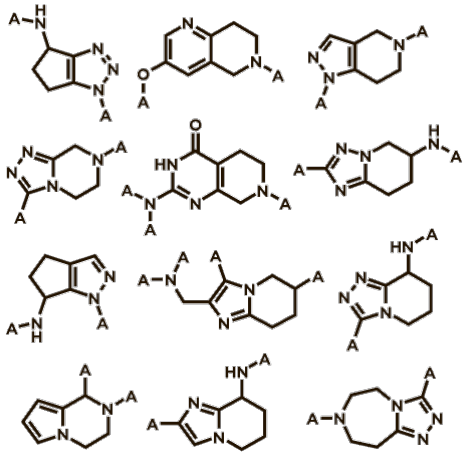
High Troughput Screening (HTS):

- Key step in the early Drug Discovery pipeline
- Main technology for hits identification
- At full capacity hundreds of thousands of compounds tested daily

High Troughput Screening advantages:

- Fully automatized workflow
- Extremely cost effective
- Suggest good starting structures that will be optimized in later steps

HTS: A closer look

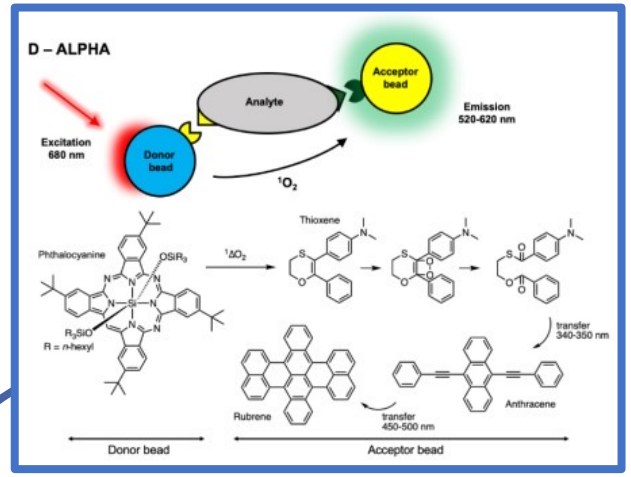


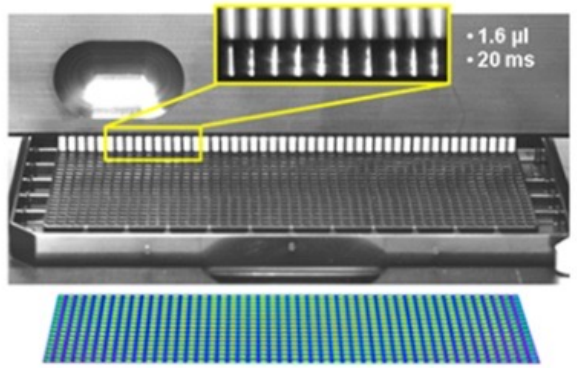
Compound Library:

Set of chemical substances to test.
Each pharma company has its own library which have been developed throughout the year

Millions of Molecules!!!

Biological assay





Hits evaluation: find the LEADS

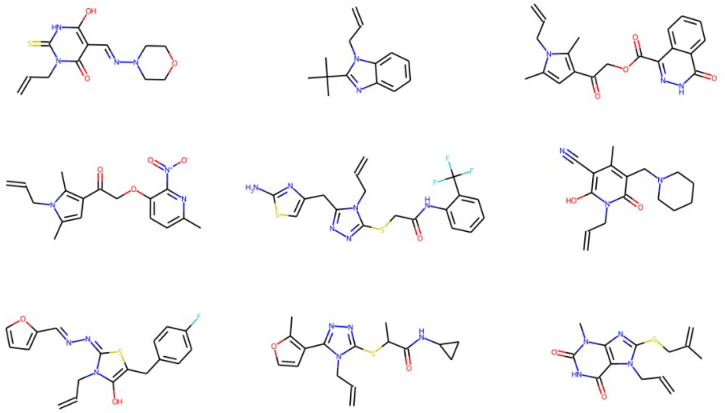
- Look at compounds with positive signal (usually ~10% of the library)
- Compare active compounds with results of older assays
- Select most promising hits

Expert knowledge

Cheminformatics

Lead compound: non-optimized chemical compound having biological activity

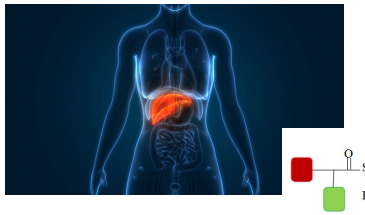
Lead Discovery: Lead optimization



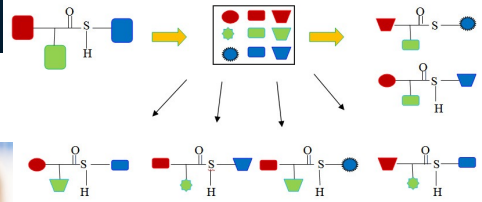
List of promising hits*

Molecules with showing bioactivity in HTS phase that passed statistical and expert-rule filters.

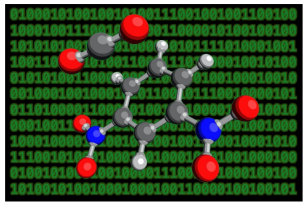
ADMET prediction



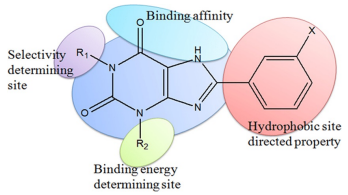
Combinatorial Chemistry



Medicinal Chemistry



Computational Chemistry



Structure Activity Relationship

Optimization

Thousands of variations generated by classical lab chemists and automated synthesis

~300 molecules

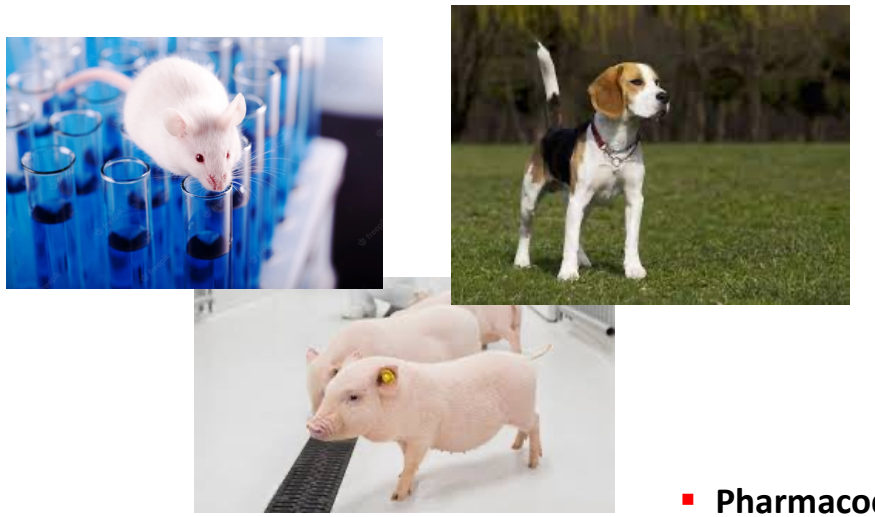
Optimized molecules

- Improved potency
- Better stability
- Increased solubility in water
- Greater selectivity

*Random sample collected from the PubChem database

PreClinic: *In-vitro* and Animal studies

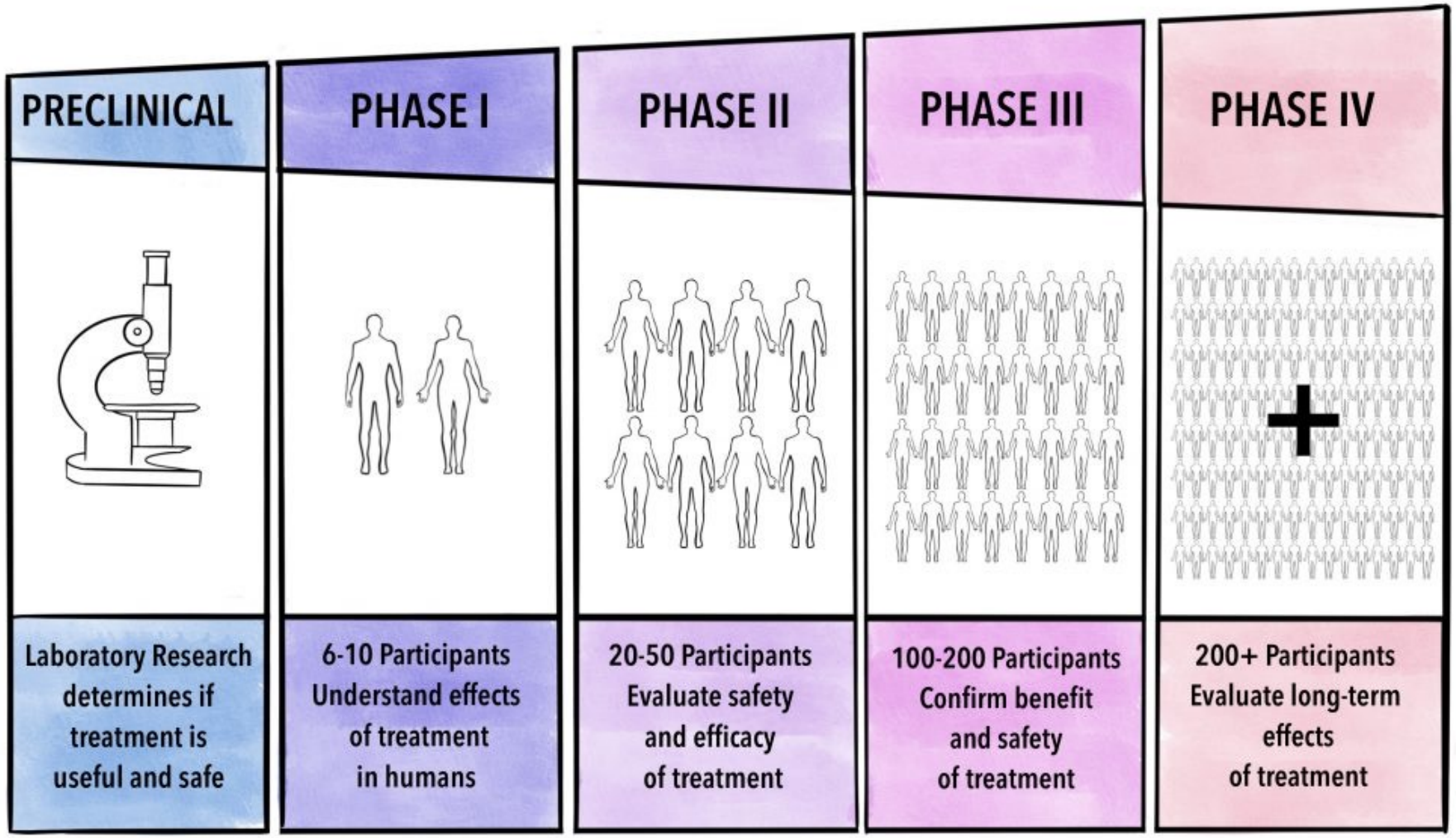
In vivo studies



In vitro studies

- **Pharmacodynamics**: What the drug does in the body?
- **Pharmacokinetics**: What the body does to the drug?
- **ADME** and **Toxicology**
- Estimate a safe starting dose for clinical trials in humans

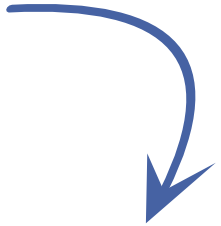
Clinical Trials: Drug safety and effectiveness



https://gene.vision/knowledge-base/clinical-trials/

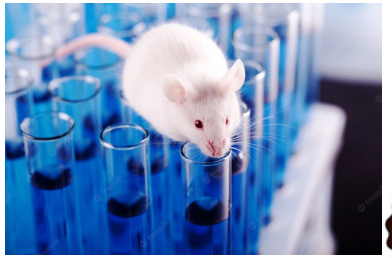
The Drug Discovery pipeline: RECAP

- Long and expensive process: 12 years and ca. 2 Billion $ invested
- Involves scientists from all the fields: Chemists (medicinal, organic, analytic, etc…), Biologists, Mathematician/Physicists
- Lot of improvement possible:
  - Speed up the discovery of new drugs
  - Increase the success rate of drug candidates



How? With **Chemoinformatics!!!**
(and Machine Learning)

“Big Data” in Chemistry



Early drug discovery

Huge amount of **noisy** data

- HTS: millions of chemicals with preliminary activity values on multiple targets.
- Lead optimization: thousands of molecules with xC50 values and ADMET proprerties

PreClinic and Clinical Trials

Few, very **reliable** data

- Toxicity studies: hundreds of compounds with know toxicity
- Pharmacodynamics: known metabolism in humans

“Big Data” in Chemistry: public data sources and databases

PubChem

https://pubchem.ncbi.nlm.nih.gov/

- Large collection of chemical informations:
  - compounds
  - biological assays
- **112** Millions compounds
- **301** Millions Bioactivity
- **42** Millions Patents

DRUGBANK

https://go.drugbank.com/

- Online database containing information on drugs and drug targets
- **15** Thousands drugs:
  - **5** K approved
  - **7** K experimental

ChEMBL

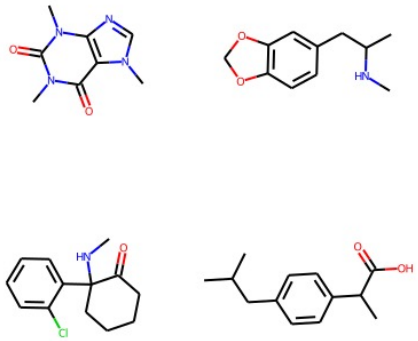
https://www.ebi.ac.uk/chembl/

- Manually curated database of bioactive molecules:
  - In vitro and in vivo assays
- **2.3** Millions Compounds
- **15** K Assays

Honorable mentions: Google Patents, Tox21

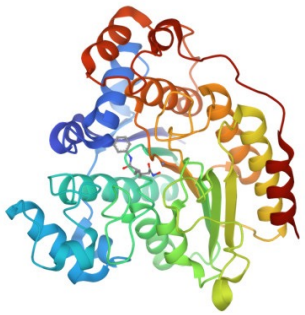
What's the deal with all those data?

Given a set of new molecules:



30L to everyone that can name all those molecules ;)

Given a protein target:

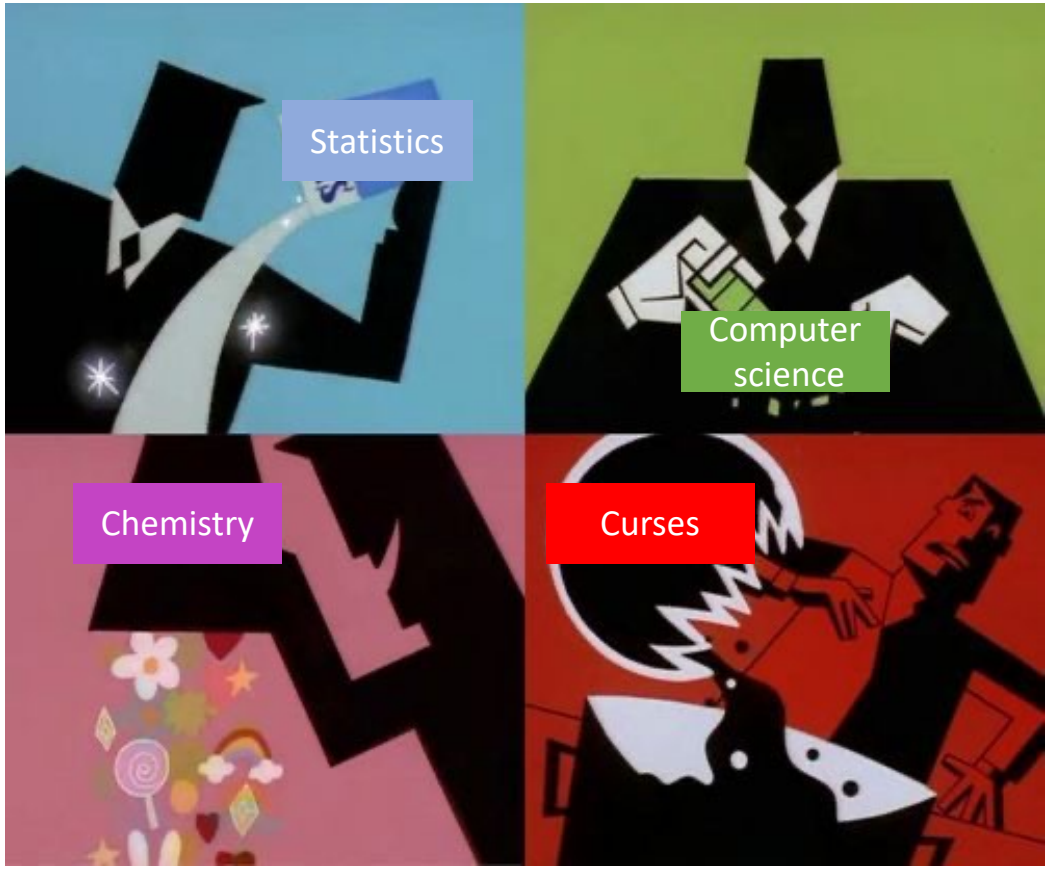


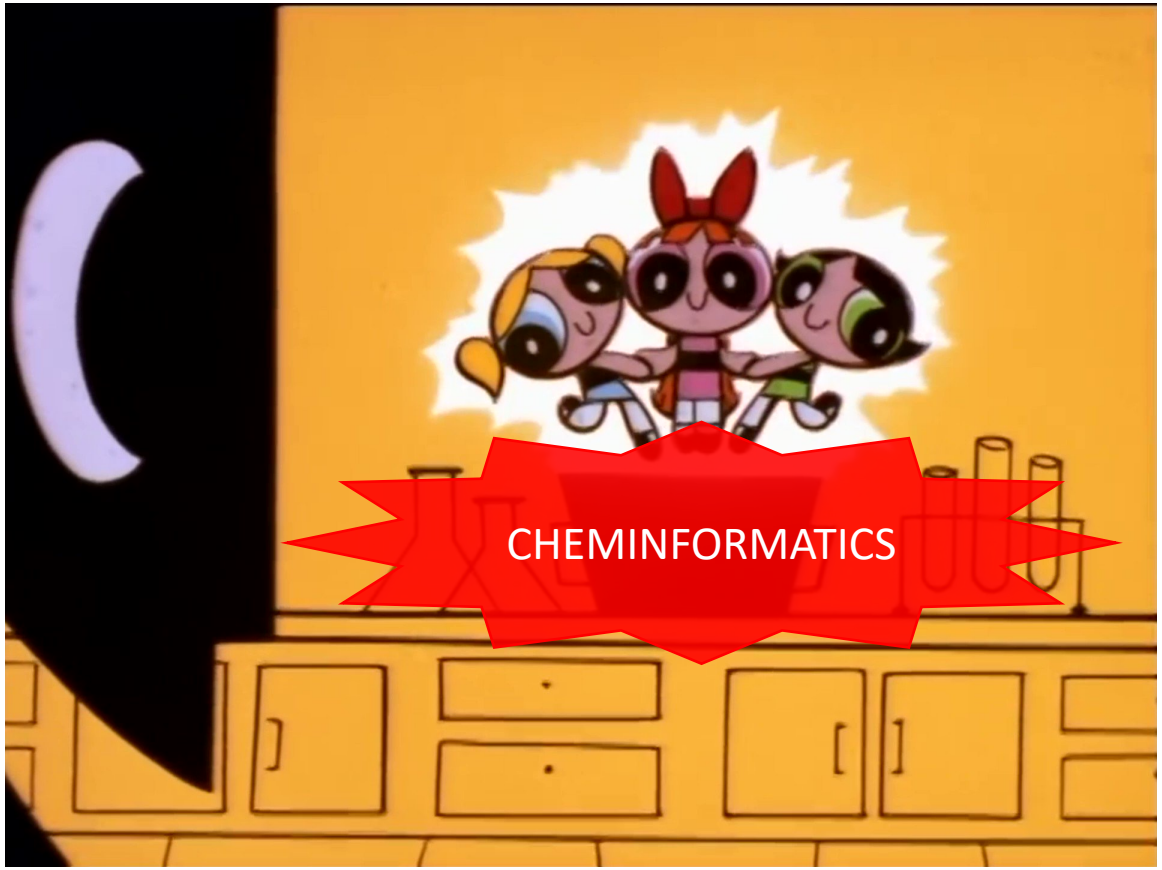
Knowledge from old experiments can be used for:

- Predict which molecules can be bioactive
- Predict which molecules can be toxic
- Predict the affinity toward a certain protein
- Generate a set of structures with high affinity

How? With **Chemoinformatics!!!**
(and Machine Learning)

Cheminformatics: Chemistry as data science





Cheminformatics: What it's done in practice?

In a nutshell: with Cheminformatics we try to exploit all the available chemical knowledge to help experimental

Chemists:

- Visualize large amounts of different compounds in an informative way (e.g.: colored by bioactivity)

  EXAMPLE: https://peter-ertl.com/molecular/rings/magicrings.html

- Develop algorithms that can use the data to predict molecular properties
- Develop algorithms that can use the data to help with synthesis (e.g.: prediction of synthetic routes)
- Find new informative ways to represent chemical compounds

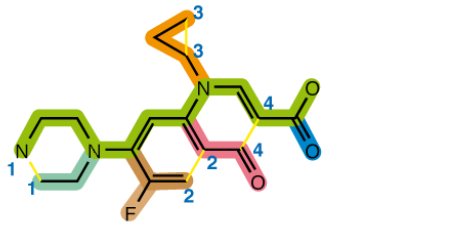
Molecular representations: SMILES, Fingerprints, 2D descriptors

Molecular representations

- Need a computer readable representation of molecules
- Molecular representations must describe:
  different types of structures (small molecules, peptides, polymers)
  with different properties (stereochemistry, valence)
- Precise representations for specific structures are needed to optimize the process of AI-driven discovery

SMILES

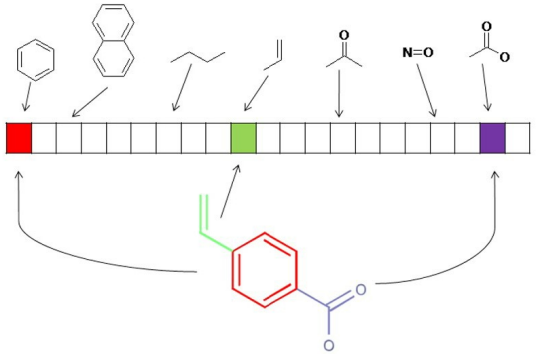
C



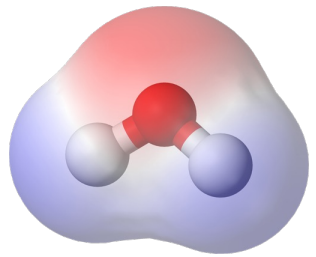
D

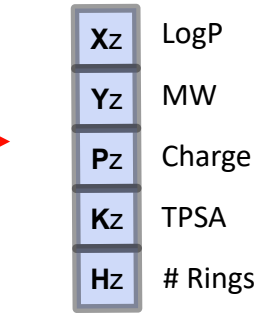
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Fingerprints



2D descriptors

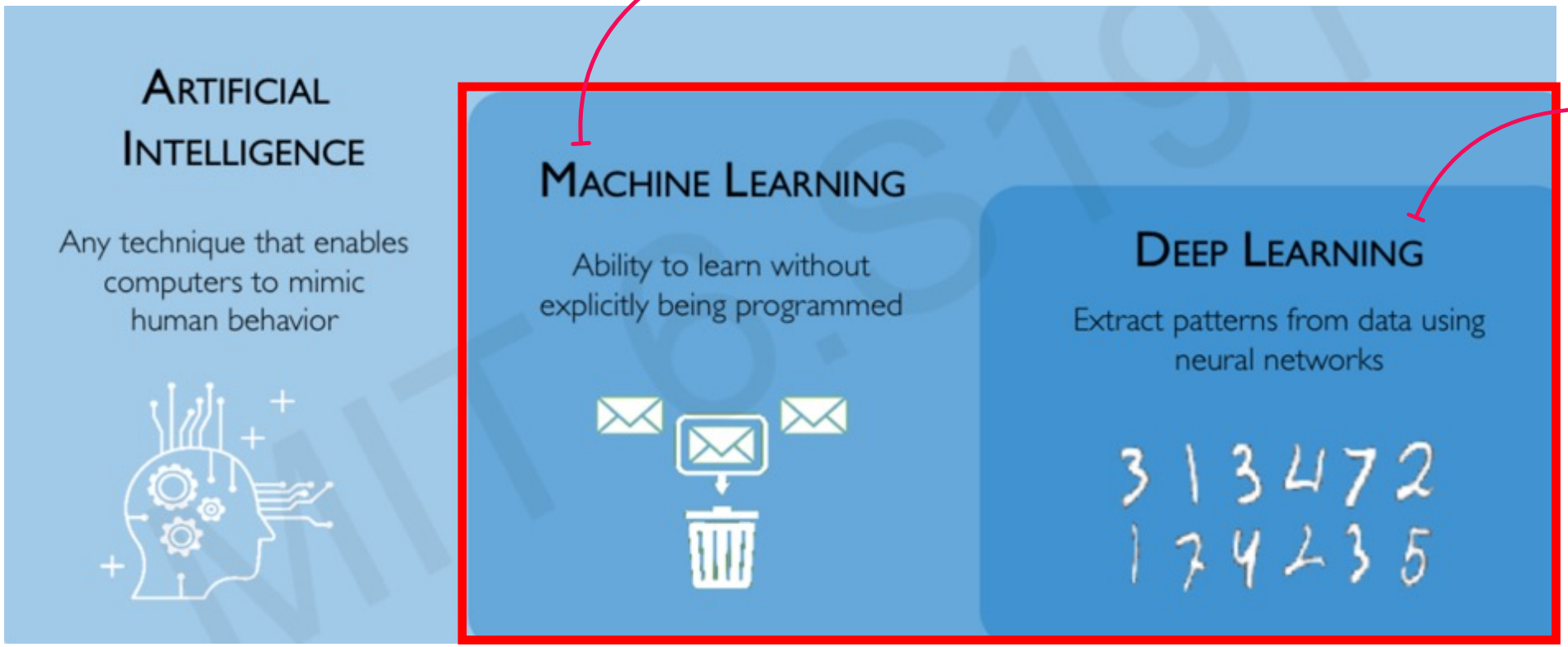


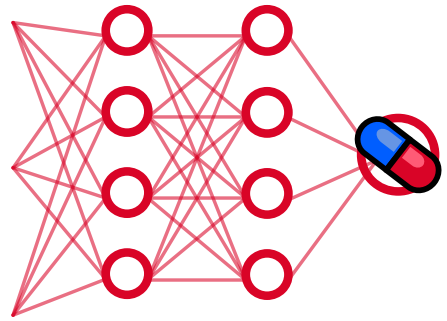


Properties vector

A hint of Machine Learning and Deep Learning

- Extract common rules that can explain a dataset.
- Rules: set of characteristics that are connected to a property





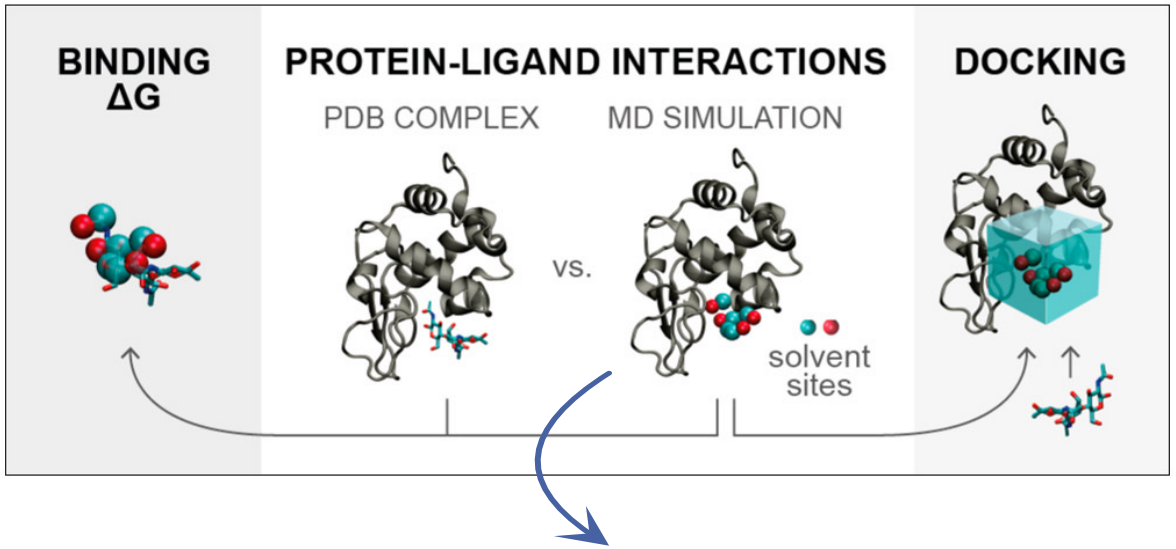
- Special type of algorithms called Artificial Neural Networks

This is a very complex and wide topic. If someone is interested in it I can give a couple of examples at the blackboard

Why not using standard Computational Chemistry?

Note: Results from Computational Chemistry experiments can be used as good starting points for modern approaches

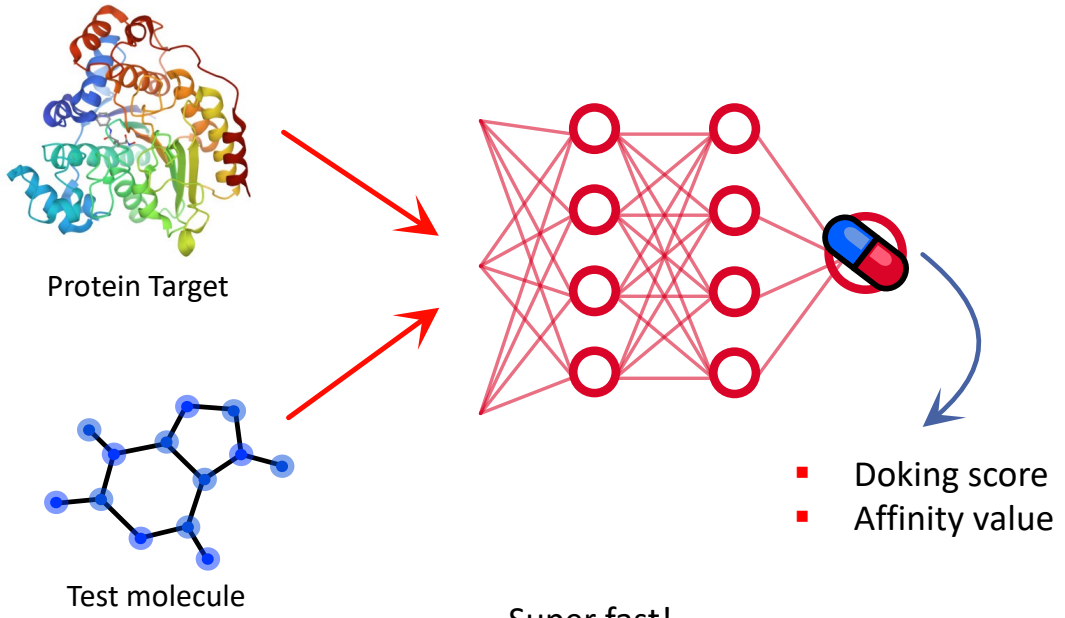
Classic computational chemistry approach



Extremely slow!

Days, even weeks depending on the initial conditions

Machine Learning approach



Super fast!

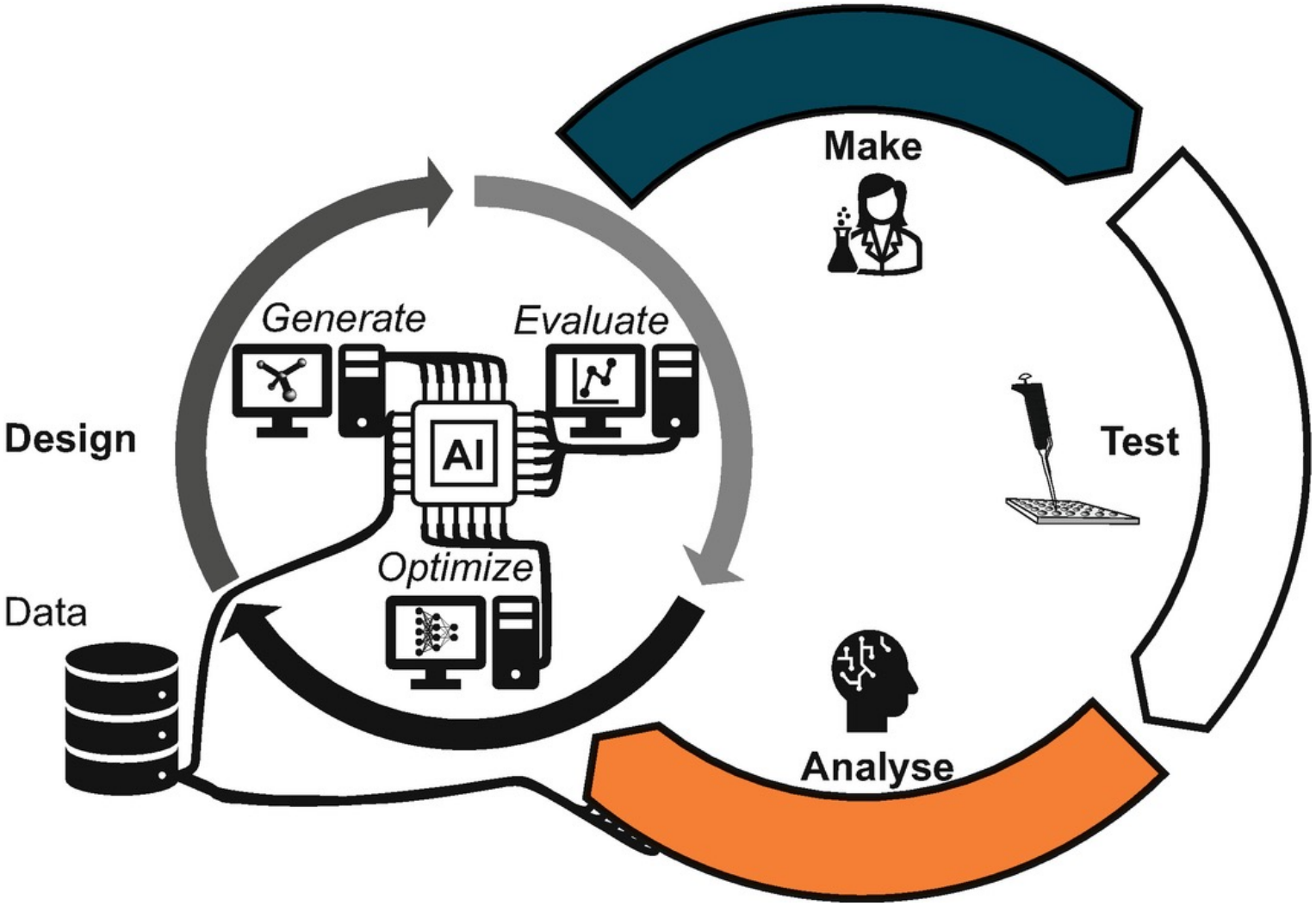
After training results for new molecules can be obtained in fractions of seconds

Arcon et al, Molecular Dynamics in Mixed Solvents Reveals Protein–Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions

Cheminformatics: RECAP

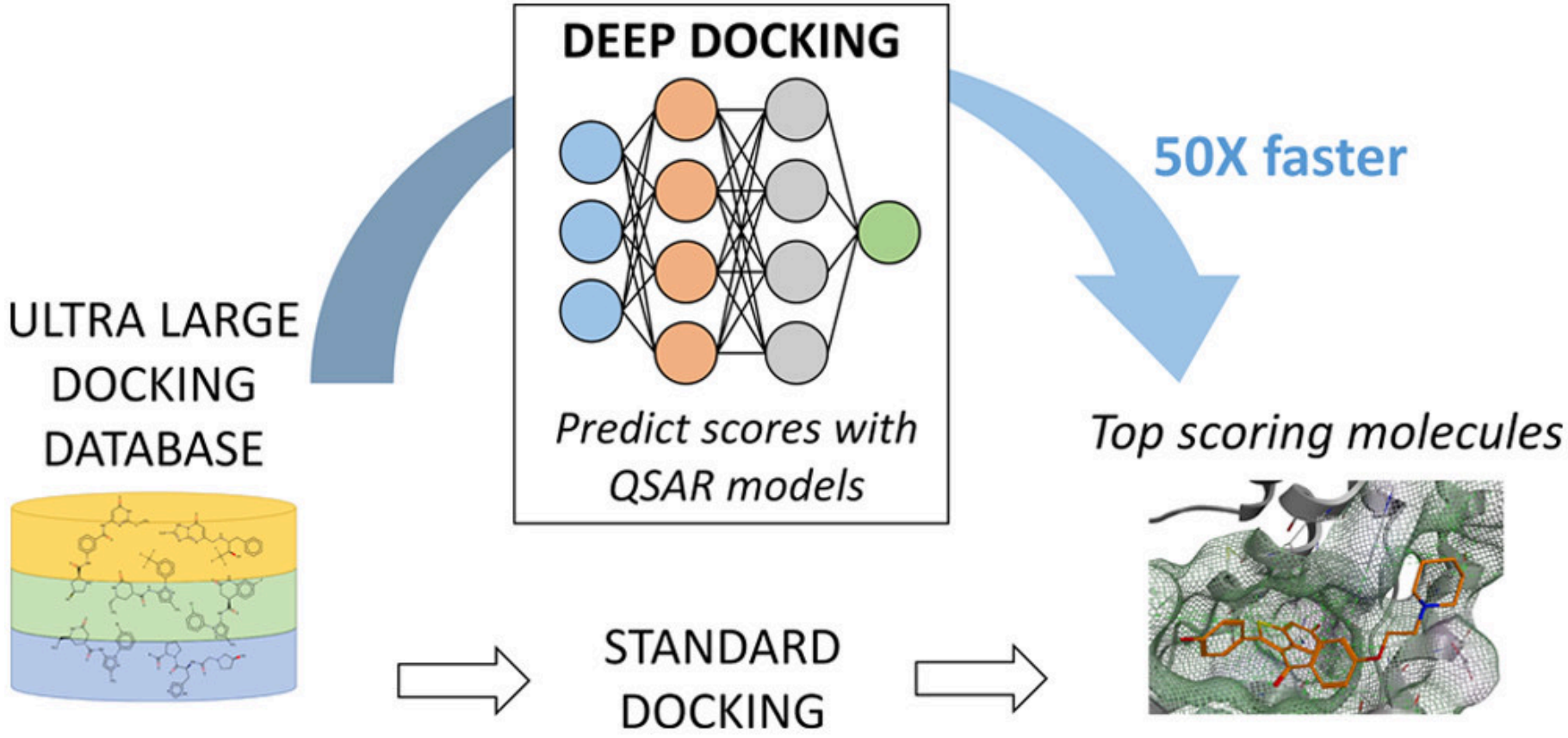
- Goal: helping experimentalists with data science
- Very requested professional profile
- Requires knowledge from different fields: Chemistry, Data Science, Computer Science, lots of Patience
- Still room for improvement and many unsolved problems: great research field!
- Requires meaningful molecular representations
- Use of Machine Learning and Deep Learning: nowadays the two most looked forward technologies in Life Science

Computer aided Drug Discovery



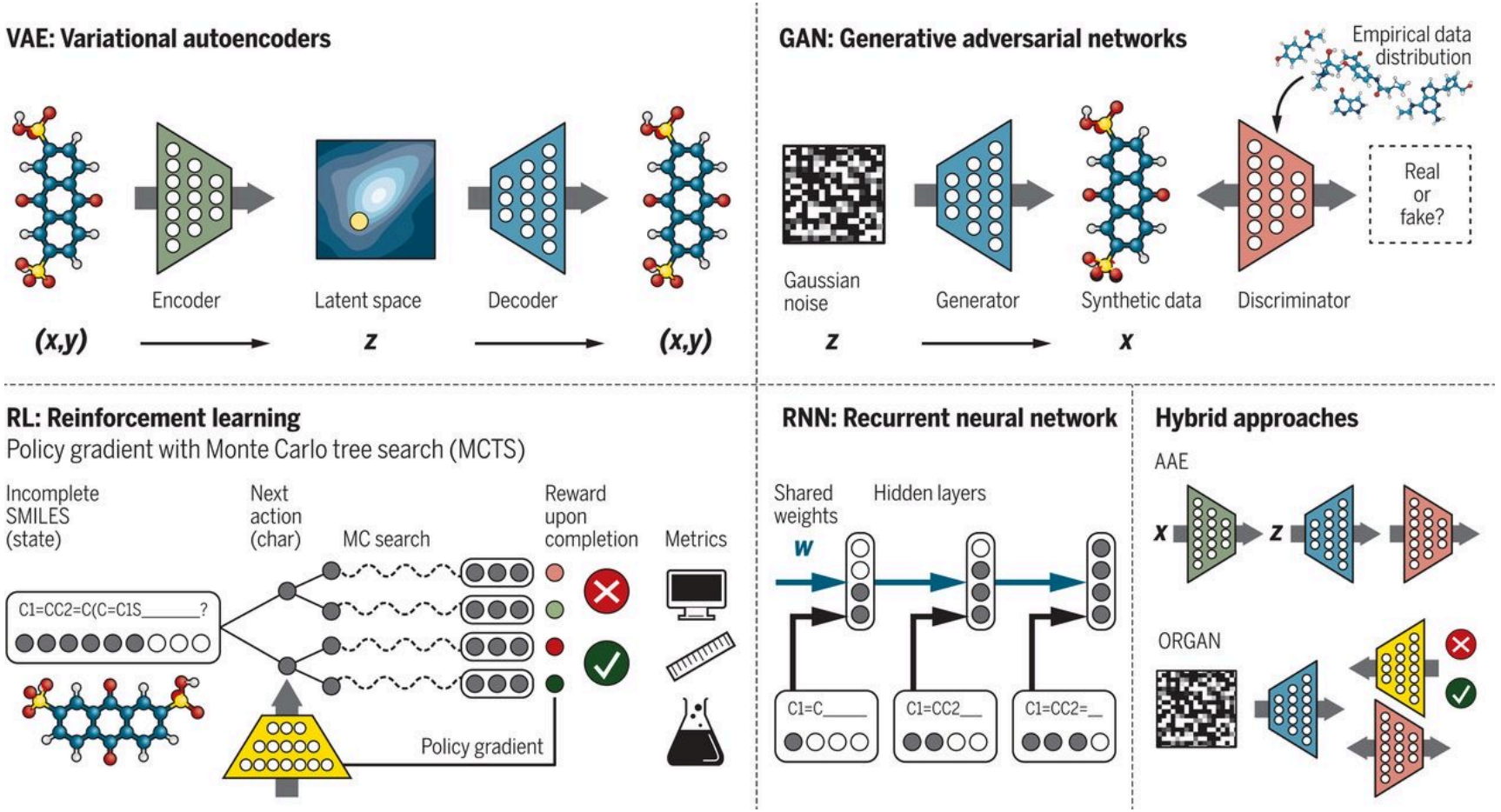
Thomas et al, Applications of Artificial Intelligence in Drug Design: Opportunities and Challenges

Molecular docking



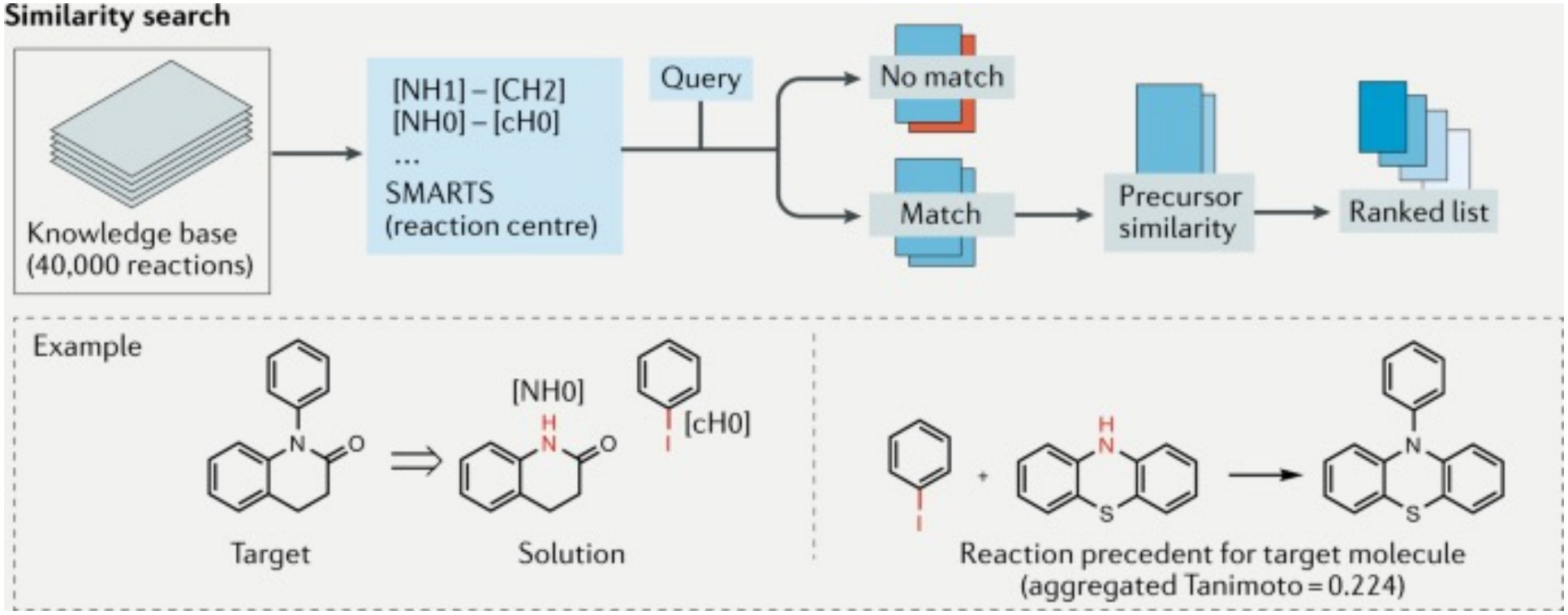
Gentile et al, Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery

De novo molecules generation



Aleshri et al, Deep Learning and Knowledge-Based Methods for Computer Aided Molecular Design - Toward a Unified Approach: State-of-the-Art and Future Directions

Retrosynthesis planning and synthesis prediction



Filipa de Almeida et al, Synthetic organic chemistry driven by artificial intelligence

Thanks!

https://github.com/bayer-science-for-a-better-life

https://ai-dd.eu/

https://comp3d.univie.ac.at/

Take home message

- Drug Discovery is very time expensive and money consuming but opens up many opportunities to have an impact on society with science
- Standard laboratory Chemistry in 2022 requires computational approaches to be efficient
- Data Science tools allows a better understanding of Chemistry in the Drug Discovery context
- My suggestion: have a look to Deep Learning and Machine Learning
- If you want to do research choose a topic that fascinates you and have fun with it
- Cheminformatics starter pack: RdKit (https://www.rdkit.org/), Python (https://www.python.org/)