AI/ML for prediction of biological properties of molecules

Module 0. Introduction to drug discovery

Gemma Turon & Miquel Duran-Frigola
Ersilia Open Source Initiative (www.ersilia.io)
18th - 27th of September, 2023

Ersilia

Welcome and introduction to the course

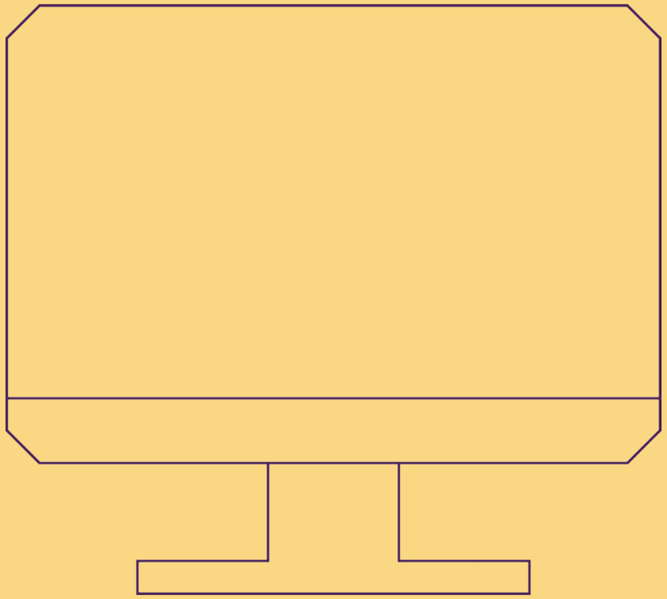
Ersilia

We are a small and young tech non-profit aimed at reducing inequality in global health.

We equip laboratories in LMICs with artificial intelligence tools for infectious disease research.

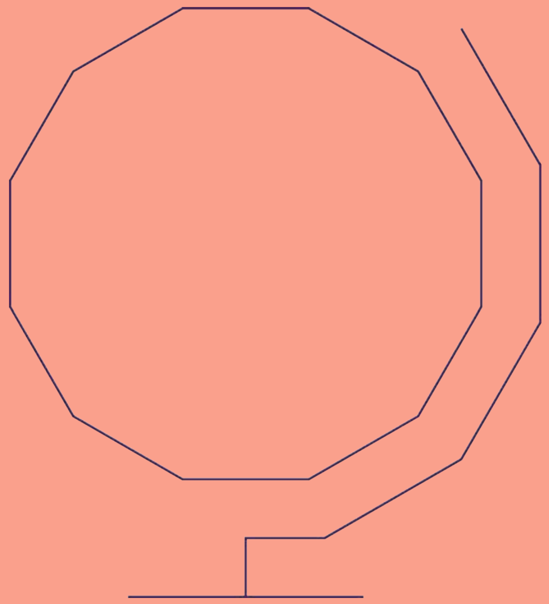
Free & open source

Public code
Open access
No patents



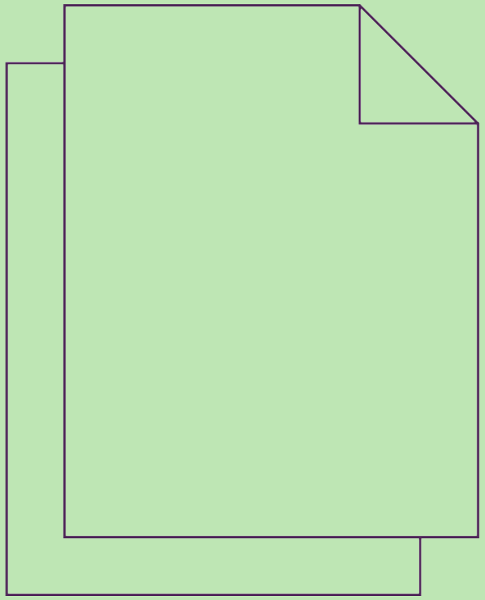
In-country research

Local scientific leadership



Sustainable collaborations

Capacity building
Low-resource tools



Course facilitators

Gemma Turon, PhD
Molecular biology & biomedicine
gemma@ersilia.io
@TuronGemma

Miquel Duran-Frigola, PhD
Computational biology & chemistry
miquel@ersilia.io
@mduranfrigola

Let’s get started!

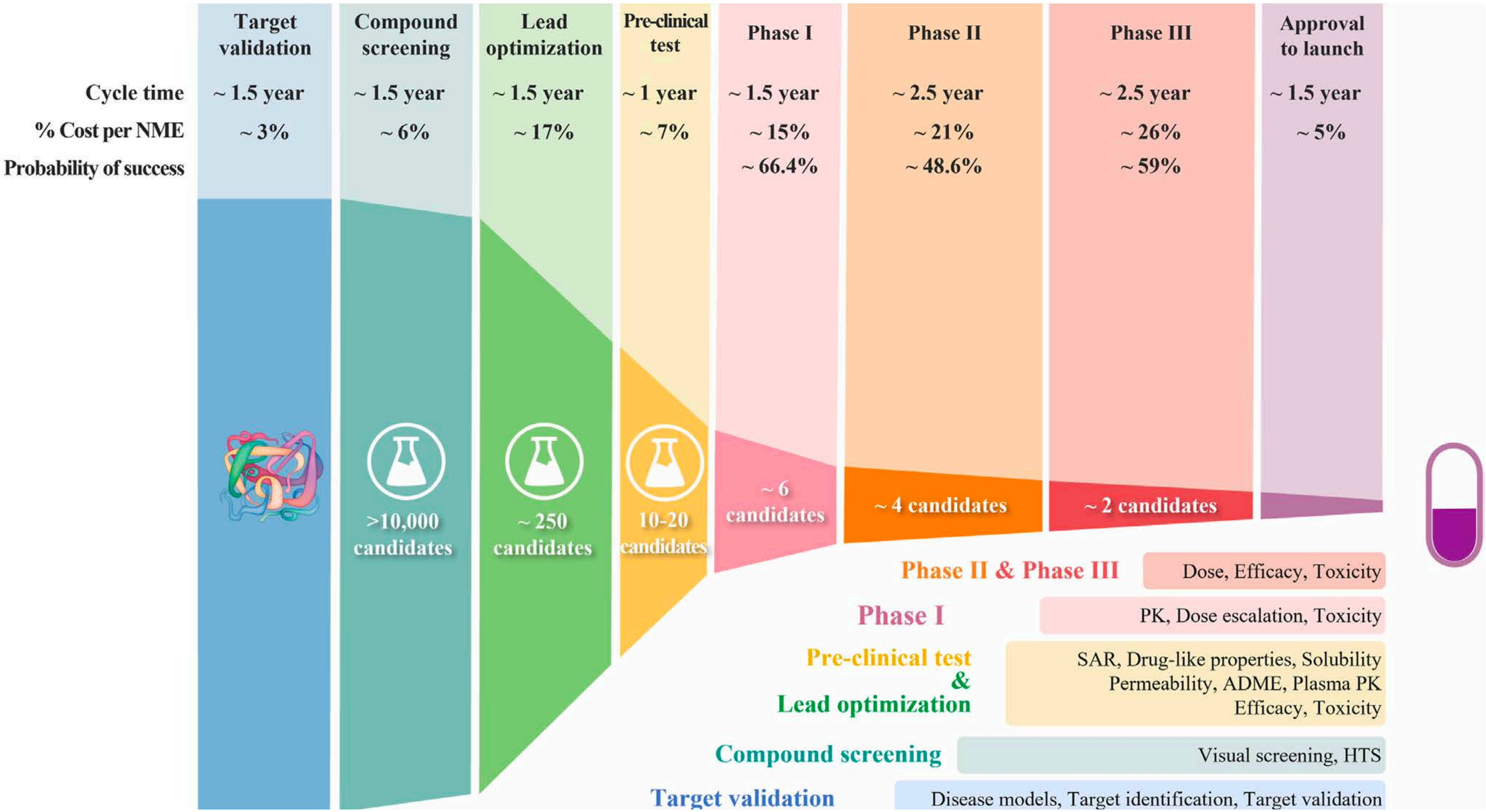
Go to menti.org and introduce this code: 89 87 55 7

Objectives of the course

- Understanding the uses and limitations of AI for drug discovery
- Learning how to apply AI to your ongoing research
- Getting the basic coding skills for data science
- Collecting tools and resources for further work

The drug discovery pipeline and the promise of AI

Drug discovery pipeline

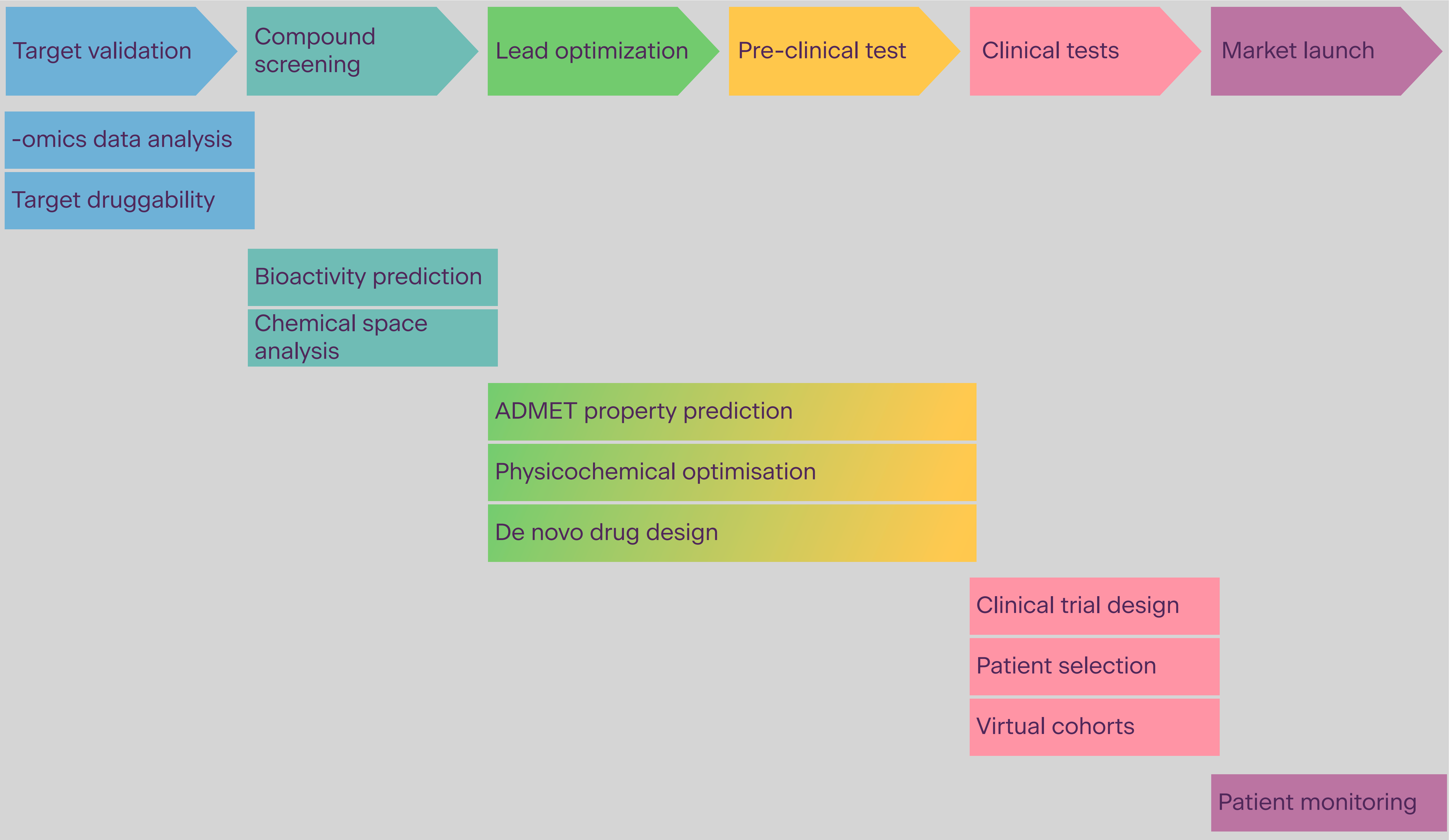


	Target validation	Compound screening	Lead optimization	Pre-clinical test	Phase I	Phase II	Phase III	Approval to launch
Cycle time	~ 1.5 year	~ 1.5 year	~ 1.5 year	~ 1 year	~ 1.5 year	~ 2.5 year	~ 2.5 year	~ 1.5 year
% Cost per NME	~ 3%	~ 6%	~ 17%	~ 7%	~ 15%	~ 21%	~ 26%	~ 5%
Probability of success					~ 66.4%	~ 48.6%	~ 59%	
		>10,000 candidates	~ 250 candidates	10-20 candidates	~ 6 candidates	~ 4 candidates	~ 2 candidates	

Phase II & Phase III	Dose, Efficacy, Toxicity
Phase I	PK, Dose escalation, Toxicity
Pre-clinical test & Lead optimization	SAR, Drug-like properties, Solubility Permeability, ADME, Plasma PK Efficacy, Toxicity
Compound screening	Visual screening, HTS
Target validation	Disease models, Target identification, Target validation

Sun et al, Acta Pharmaceutica Sinica B, 2022

AI in the drug discovery pipeline



Target validation

Compound screening

Lead optimization

Pre-clinical test

Clinical tests

Market launch

-omics data analysis

Target druggability

Bioactivity prediction

Chemical space analysis

ADMET property prediction

Physicochemical optimisation

De novo drug design

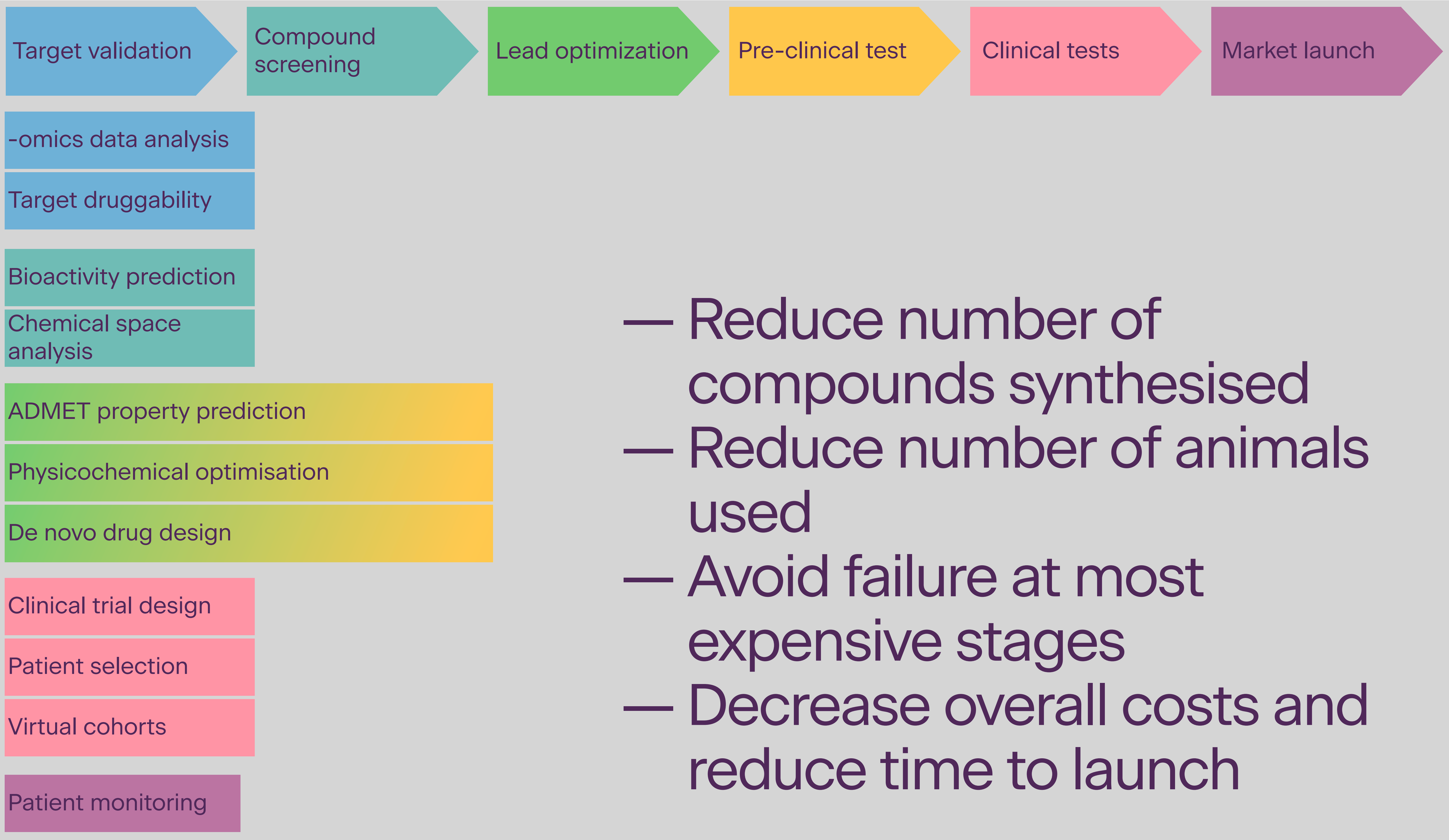
Clinical trial design

Patient selection

Virtual cohorts

Patient monitoring

AI in the drug discovery pipeline



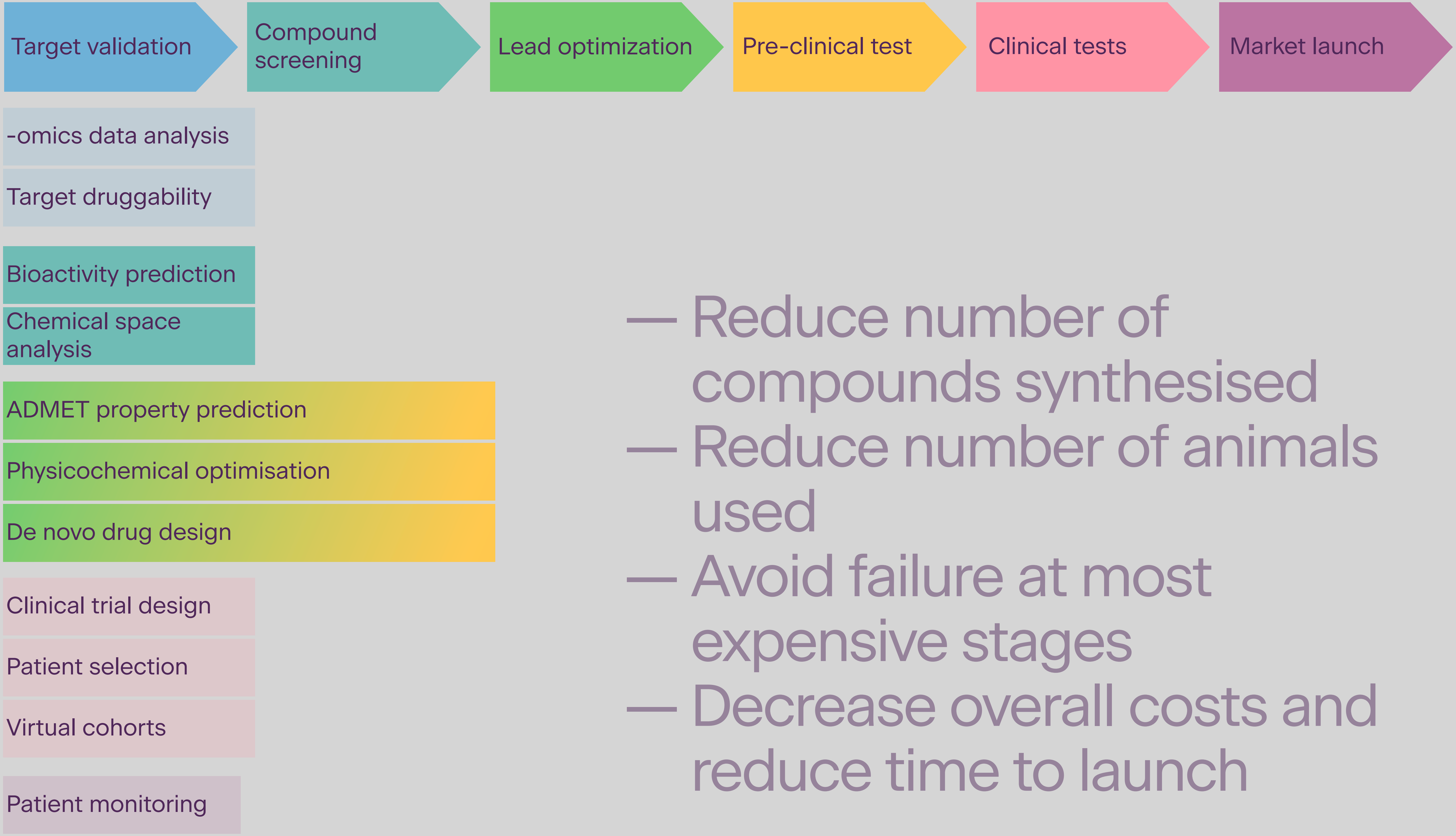
— Reduce number of compounds synthesised

— Reduce number of animals used

— Avoid failure at most expensive stages

— Decrease overall costs and reduce time to launch

AI in the drug discovery pipeline



Target validation

Compound screening

Lead optimization

Pre-clinical test

Clinical tests

Market launch

-omics data analysis

Target druggability

Bioactivity prediction

Chemical space analysis

ADMET property prediction

Physicochemical optimisation

De novo drug design

Clinical trial design

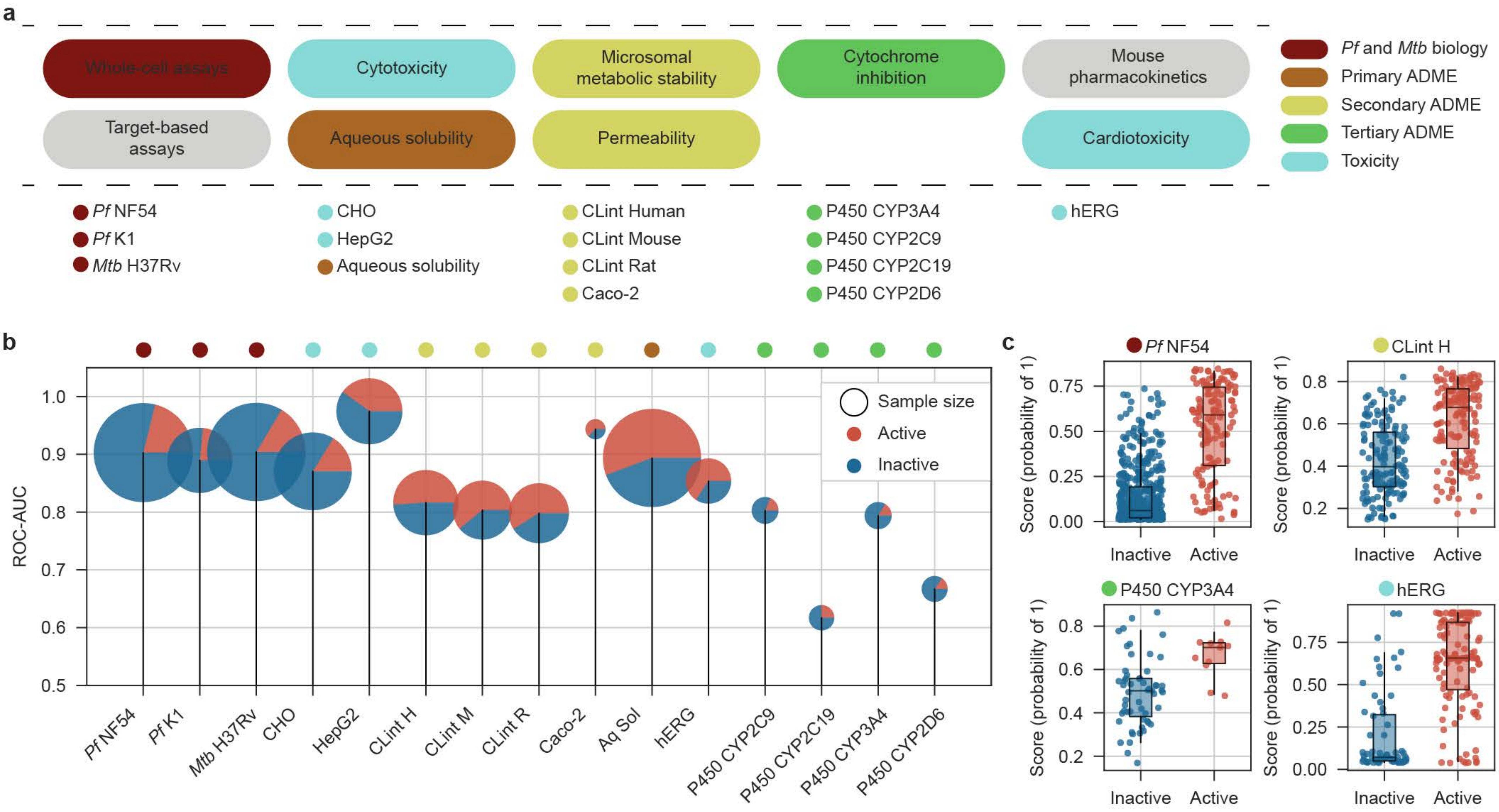
Patient selection

Virtual cohorts

Patient monitoring

- Reduce number of compounds synthesised
- Reduce number of animals used
- Avoid failure at most expensive stages
- Decrease overall costs and reduce time to launch

An example of a virtual screening cascade

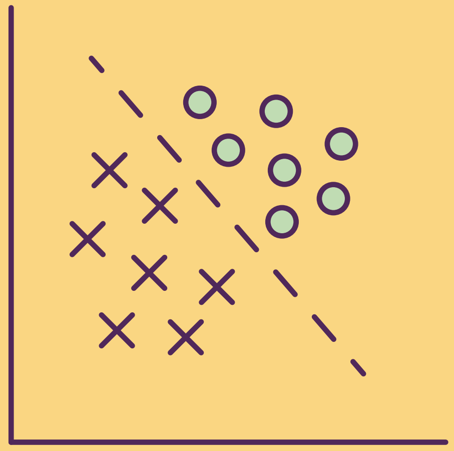


Turon*, Hlozek* et al, Nat Commun, 2023

Basic AI concepts for molecular modelling

Supervised

Labeled data
Classification
Regression

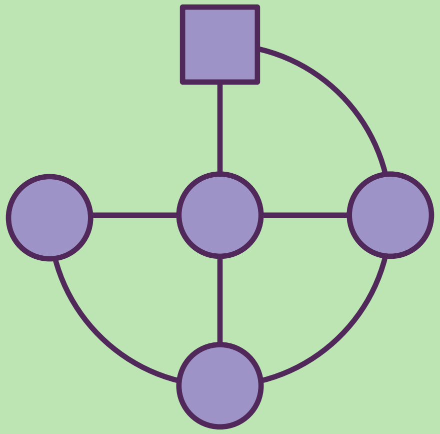


Unsupervised

Unlabelled data
Clustering
2D projection
Similarity search

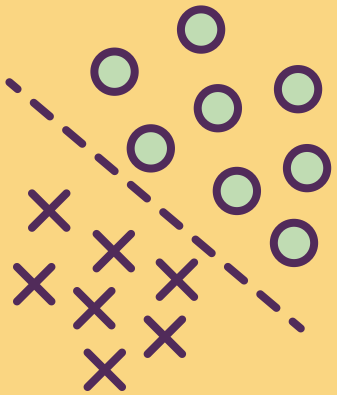
Reinforcement

Interaction with environment & agent
Generative models



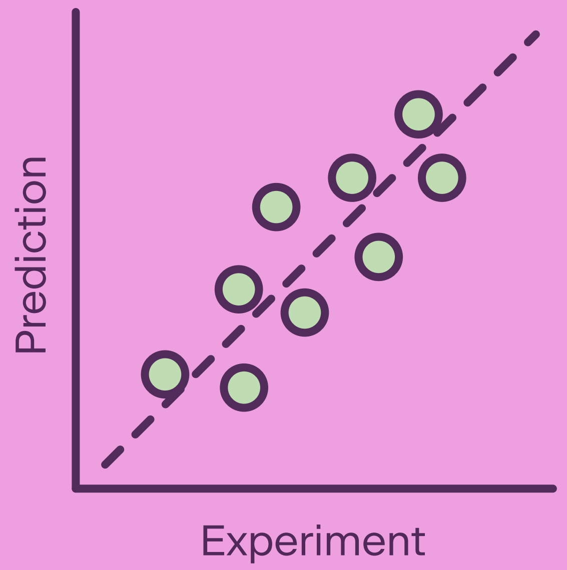
Supervised machine learning

Classification



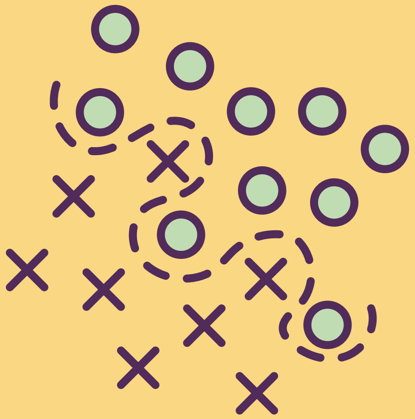
Prediction task:
Active = 1
Inactive = 0

Regression

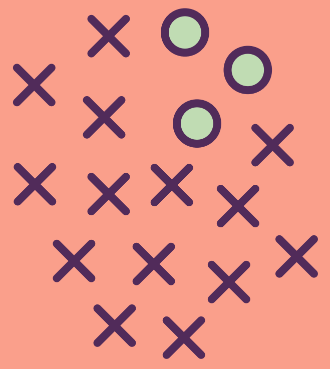


Prediction task:
IC50 value

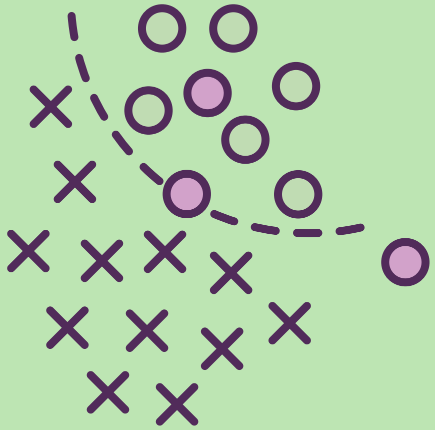
Overfitting



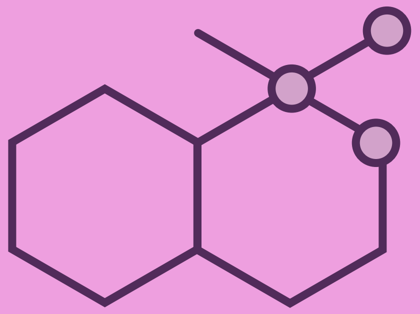
Imbalance



Confidence



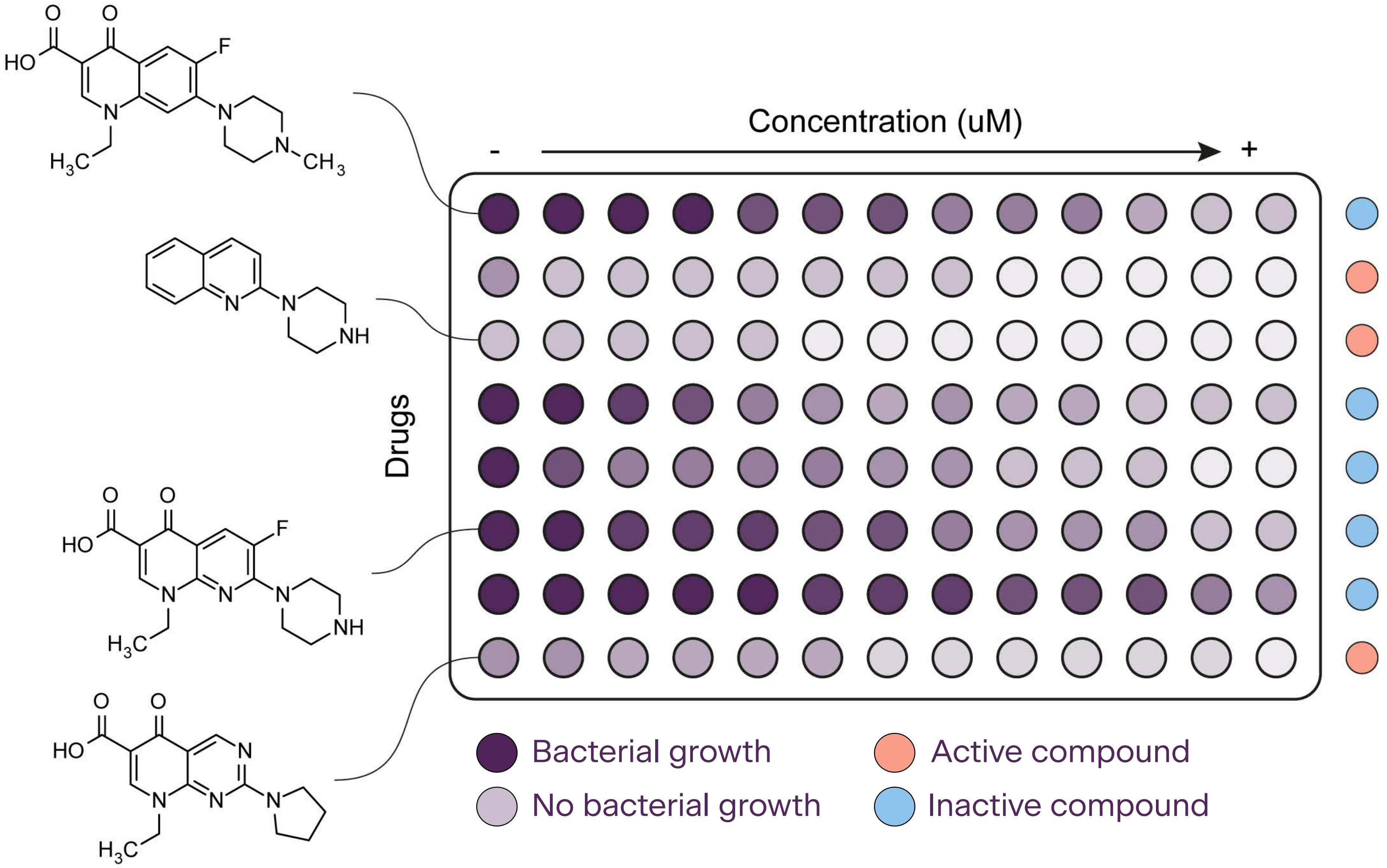
Interpretability



Basic components of an AI/ML model for bioactivity prediction

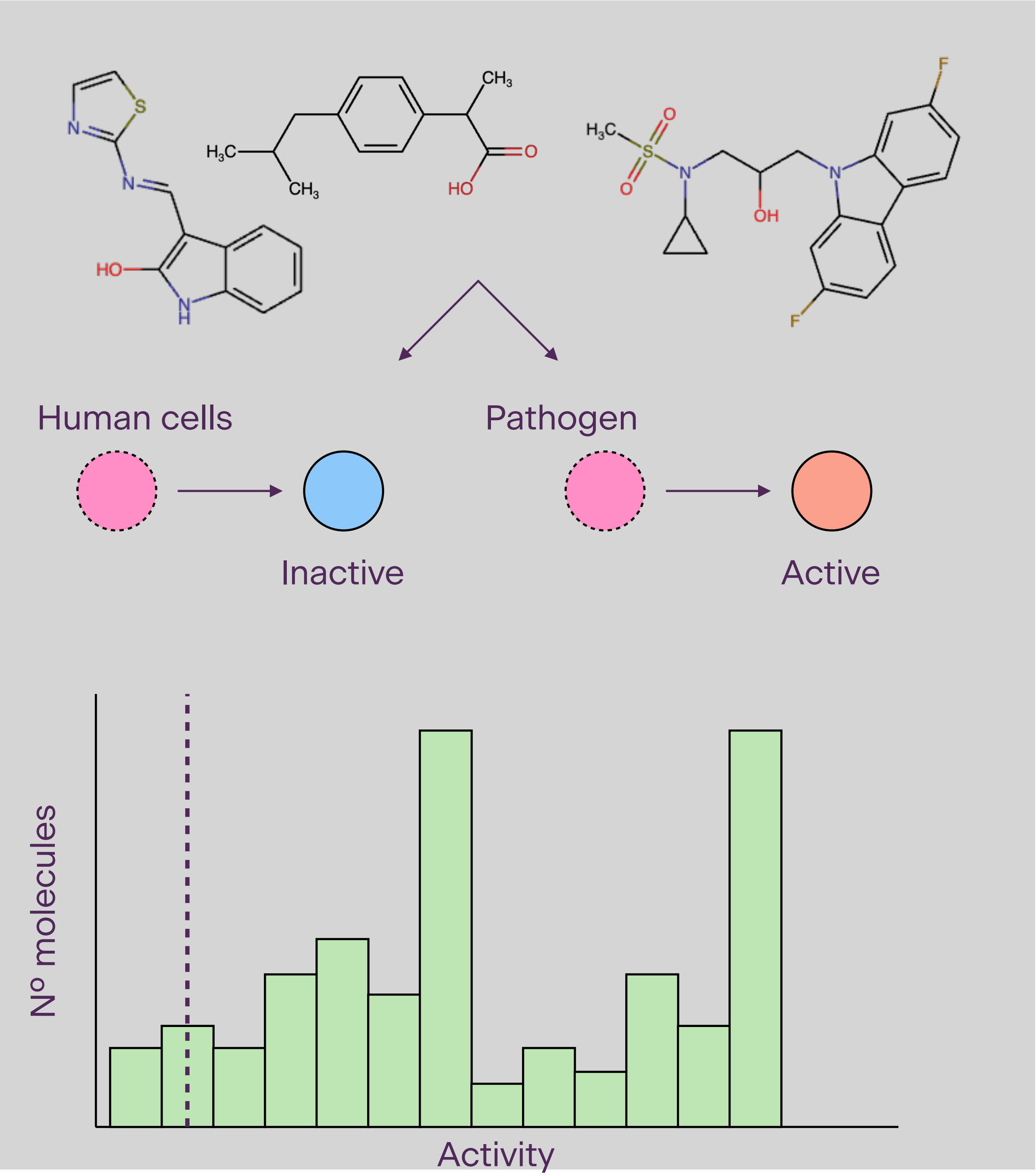
Go to menti.org and introduce this code: 8771 8241

Understanding our training dataset



Experimental data:
— Conditions
— Units of measurement
— Activity range

Understanding our training set



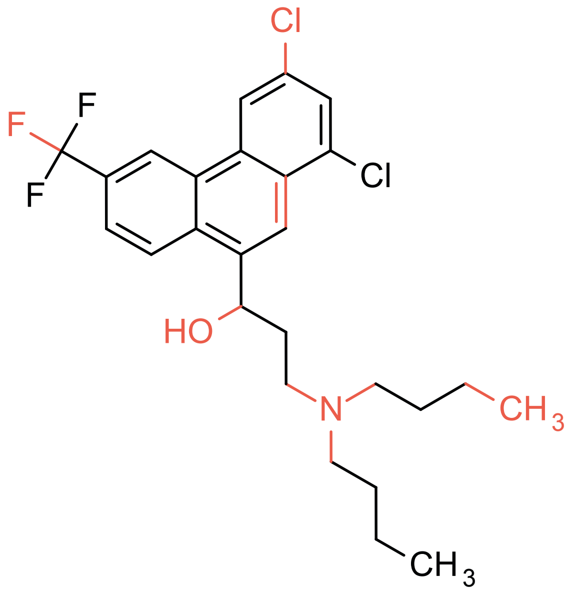
A few considerations:

— Do we want active or inactive molecules?

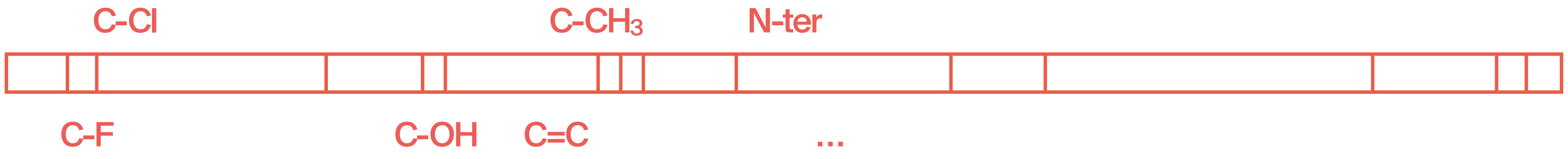
— What are good thresholds of bioactivity?

— Are there experimental limits?

Gathering information about our molecules



Calculated properties
— Molecular weight
— LogP or LogD
— Hydrogen bonds
— pKa
— Topological surface area (TPSA)
— ...

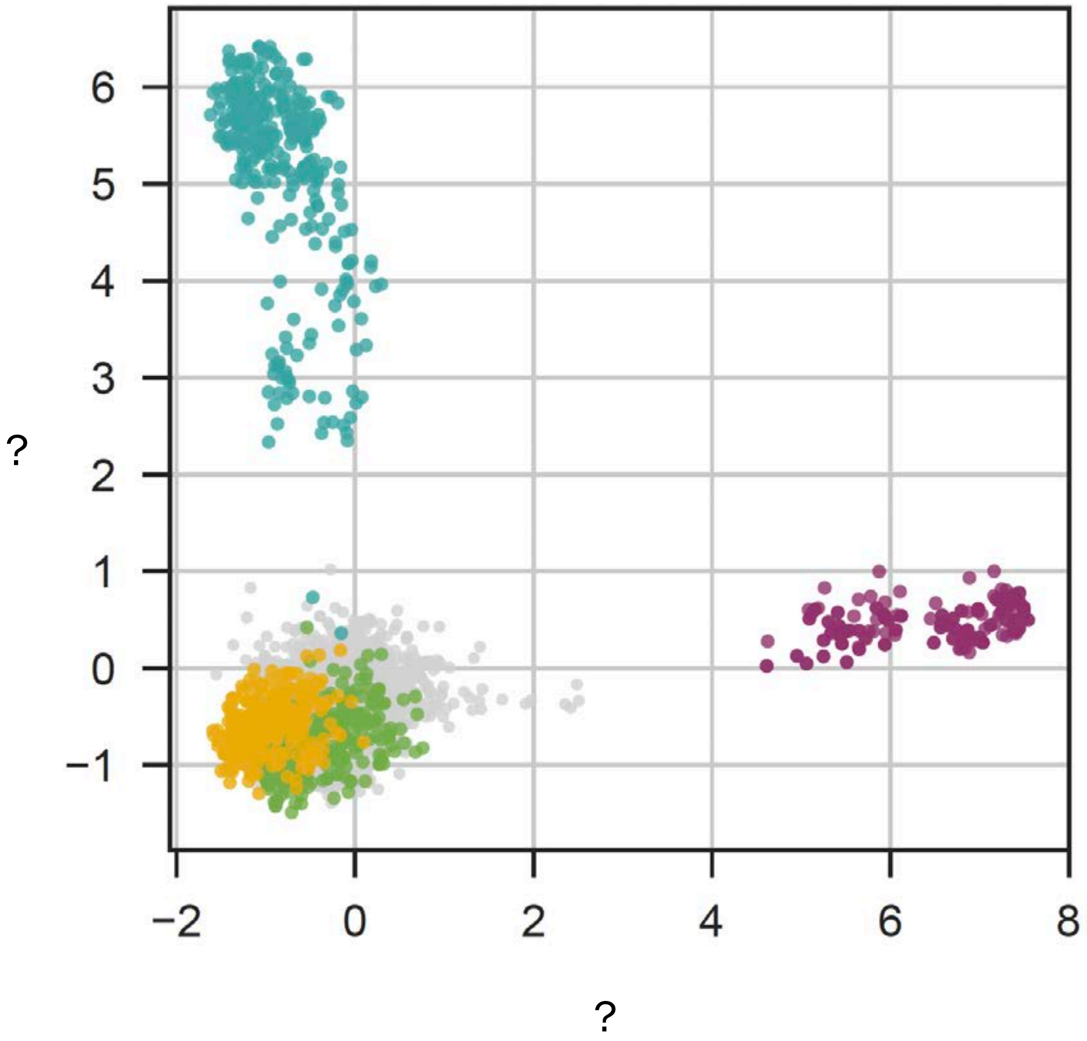


Halofantrine belongs to the class of organic compounds known as phenanthrenes and derivatives. These are polycyclic compounds containing a phenanthrene moiety, which is a tricyclic aromatic compound with three non-linearly fused benzene. Halofantrine is a synthetic antimalarial which acts as a blood schizonticide. It is effective against multi drug resistant (including mefloquine resistant) P. falciparum malaria. The mechanism of action of Halofantrine may be similar to that of chloroquine, quinine, and mefloquine; by forming toxic complexes with ferritoporphyrin IX that damage the membrane of the parasite. It appears to inhibit polymerisation of heme molecules (by the parasite enzyme 'heme polymerase'), resulting in the parasite being poisoned by its own waste. Halofantrine has been shown to preferentially block open and inactivated HERG channels leading to some degree of cardiotoxicity. Side effects include coughing noisy, rattling, troubled breathing, loss of appetite, aches and pain in joints, indigestion, and skin itching or rash, et cetera, et cetera.

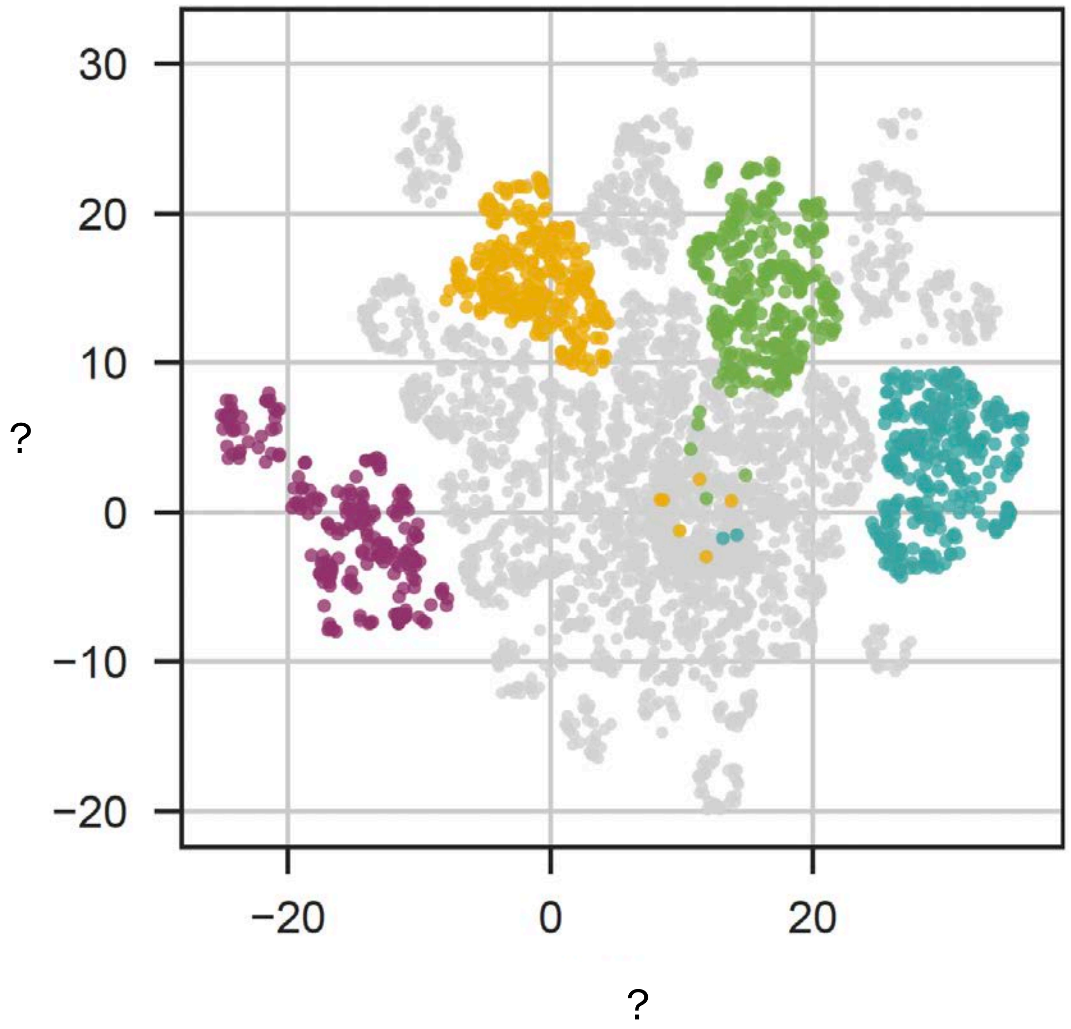
Duran-Frigola et al. Nat Commun, 2014
Duran-Frigola et al. Nat Biotech, 2020

Visualising the chemical space

Principal component analysis (PCA)

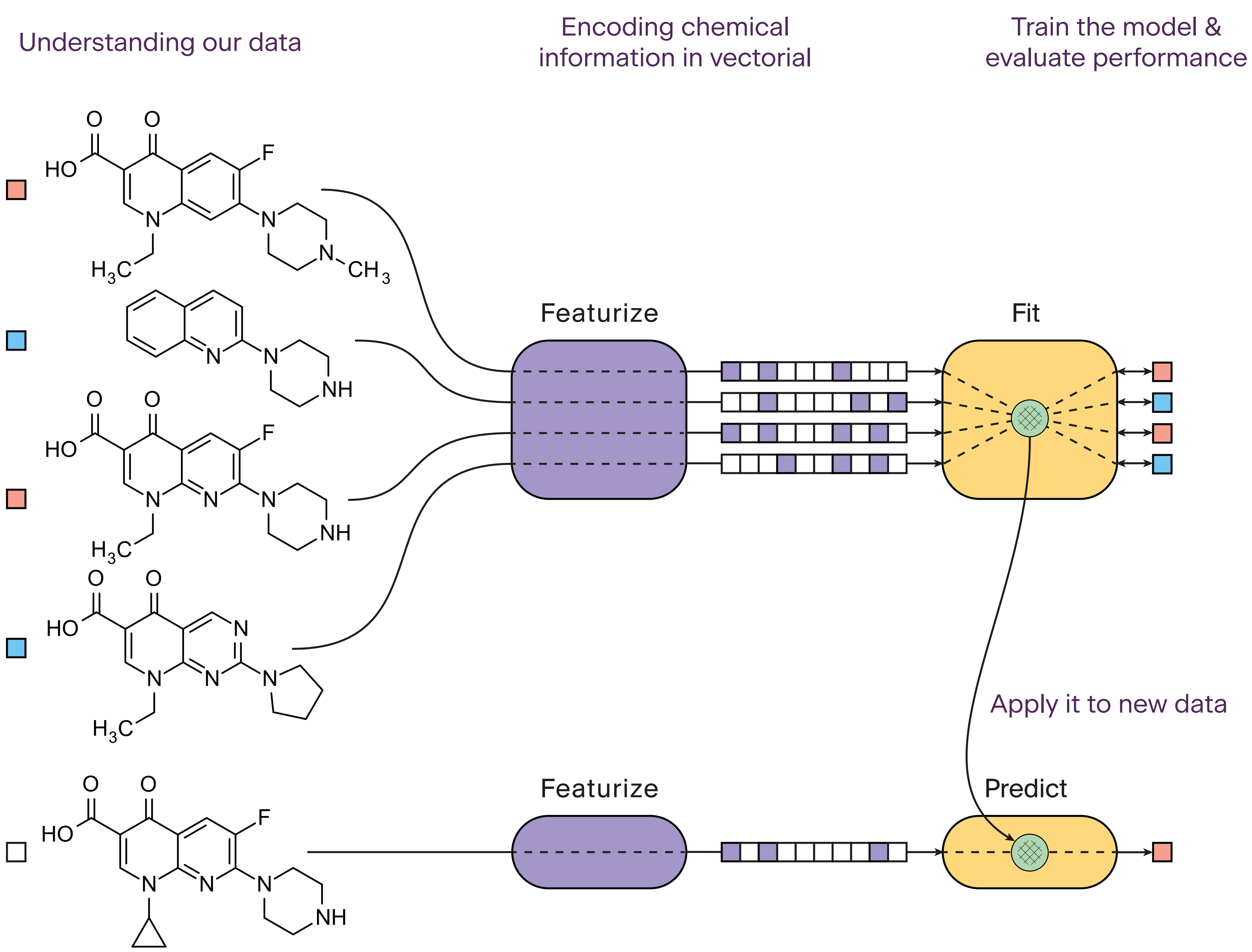


Uniform Manifold Approximation and Projection (UMAP)



Turon*, Hlozek* et al, Nat Commun, 2023

Building a AI model



Keywords

- Fingerprint
- Supervised
- Unsupervised
- PCA
- UMAP
- Embeddings
- Overfitting
- Imbalance
- Interpretability/ explainability/ transparency
- Reinforcement
- SMILES
- IC50/EC50
- Confidence
- Outlier
- Applicability domain
- Chemical space
- Cross-validation

Course overview

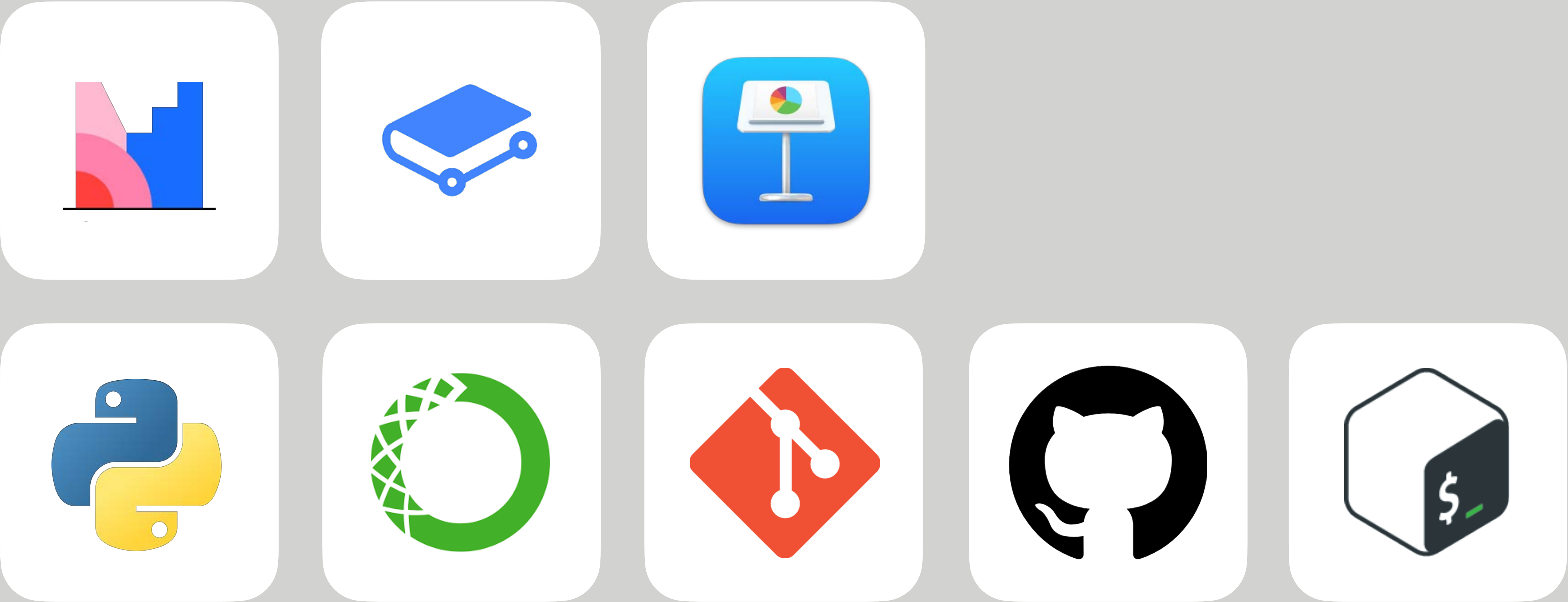
Course overview

— Module 1. Using AI models for drug discovery
— Module 2. Setting up our computational environment
— Module 3. The Ersilia Model Hub
— Module 4. Introduction to AI model training and performance evaluation

General schedule

9:00 - 10:00: good morning & setting up
10:00 - 12:30: course part 1
12:30 - 13:30: lunch break
13:30 - 16:00: course part 2
16:00 - 17:00: wrap up & good bye

Get familiar with the these tools!



Any questions?

https://ersilia.io
hello@ersilia.io
@ersiliaio