Bringing data science and AI/ML tools to infectious disease research

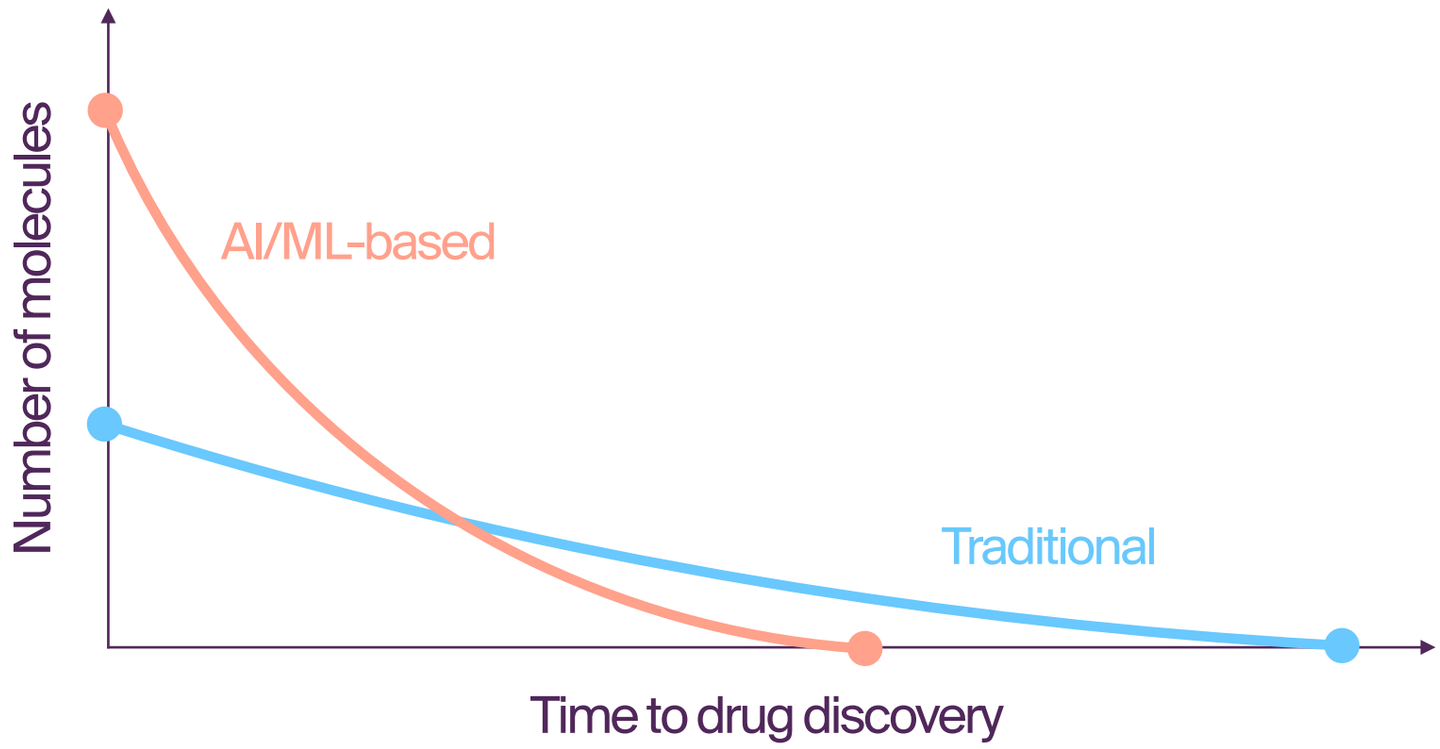
Session 2: Introduction to Supervised Machine Learning

Event Sponsors

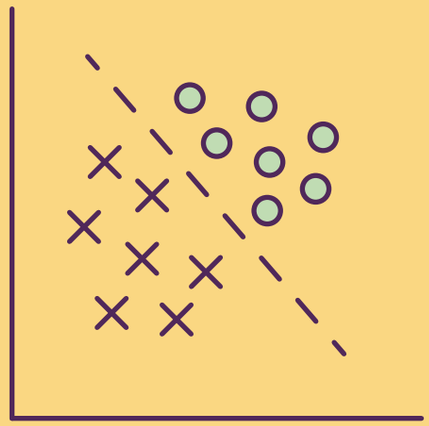
Why AI/ML for drug discovery?

– AI/ML
  Cost-effective
  Fast implementation
  Remote
  Domain agnostic

– Small molecule drugs
  Track record
  Neglected diseases
  Cheap medications
  Inspiring initiatives: MMV, DNDi, OSM…



Supervised



Labeled data
Classification
Regression

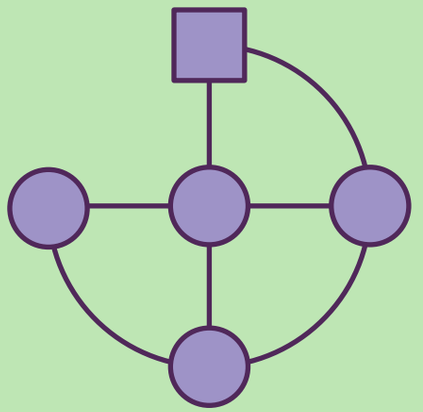
(Today)

Unsupervised

Unlabeled data
Clustering
2D projection
Similarity search

(Yesterday)

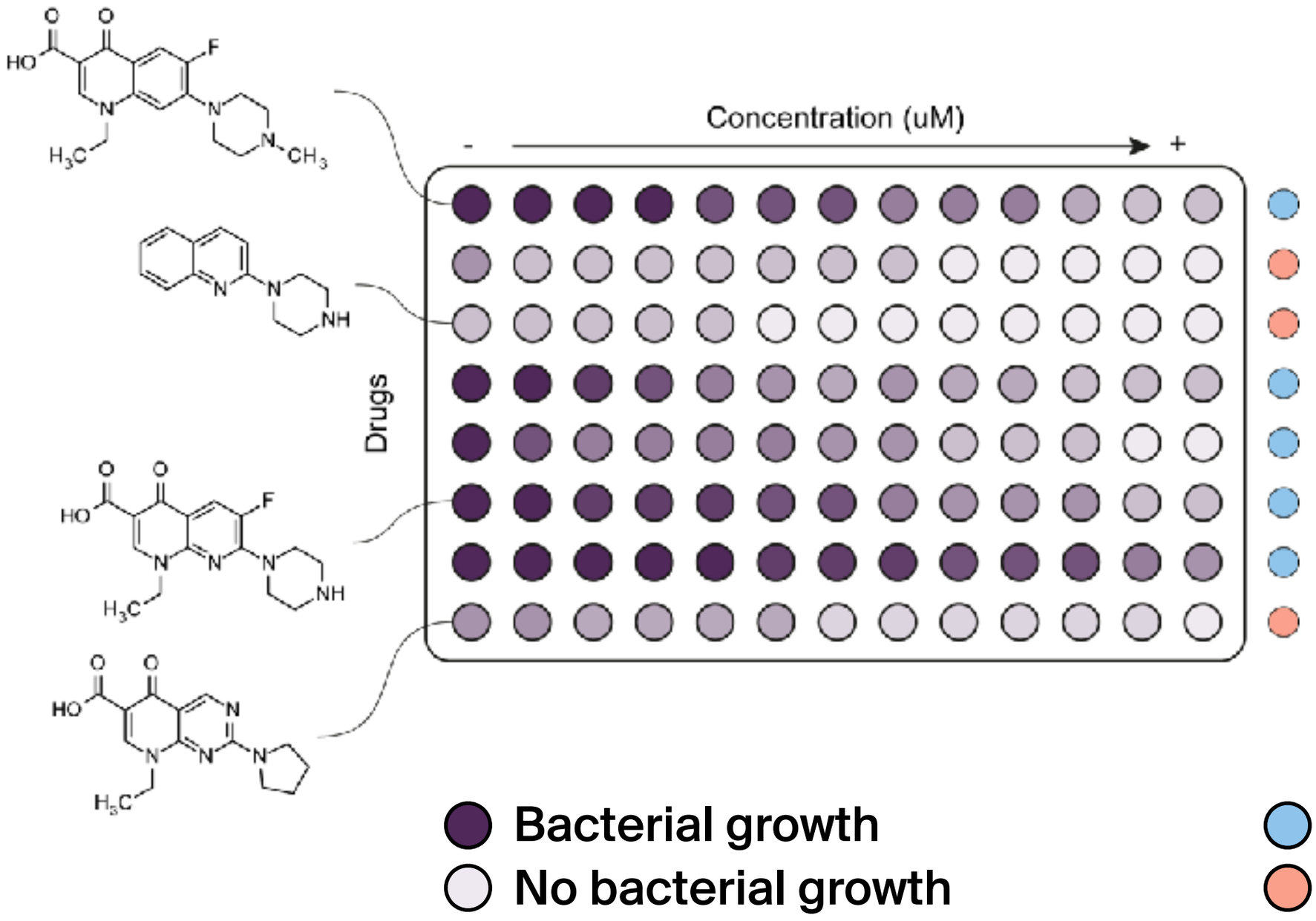
Reinforcement



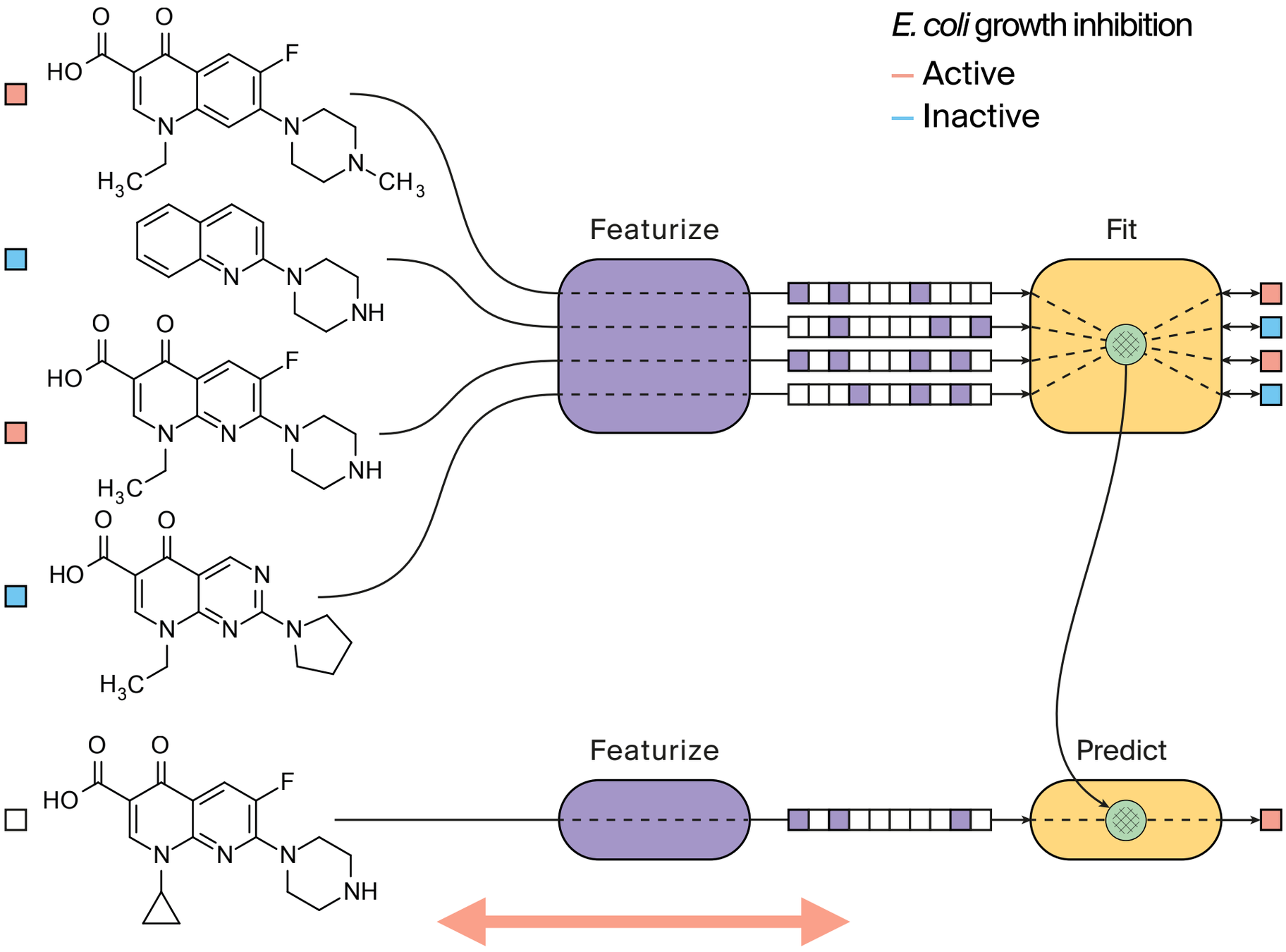
Interaction with
environment / agent
Generative models

(On Friday)

A simple supervised task in drug discovery (aka QSAR)
Antibiotic activity experiments (*E. coli*)

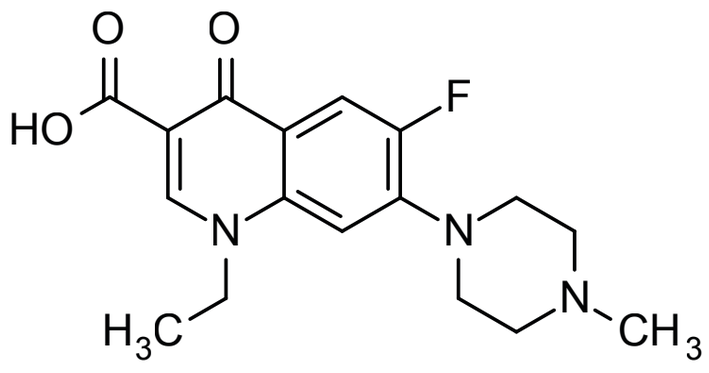


A simple supervised task in drug discovery (aka QSAR)
Antibiotic activity experiments (*E. coli*)

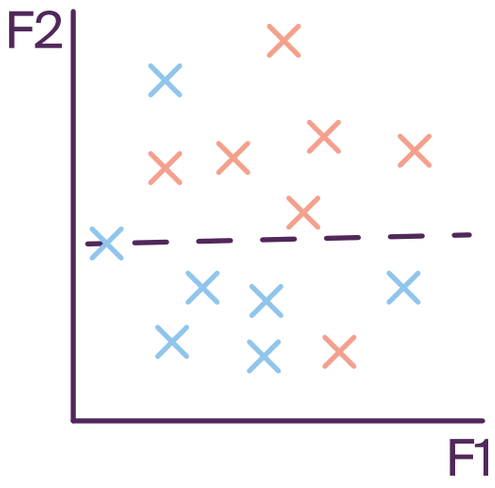


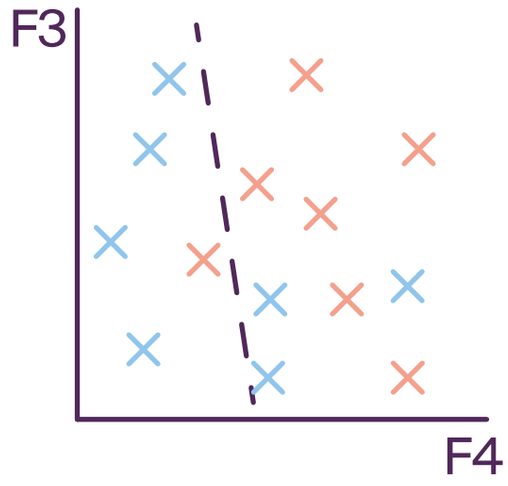
Featurization

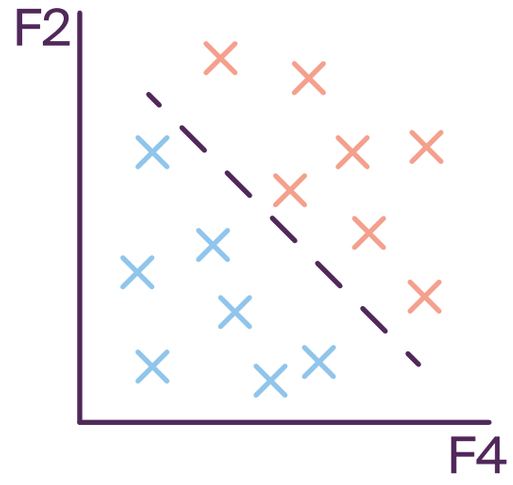
— Choose and design features that are discriminative



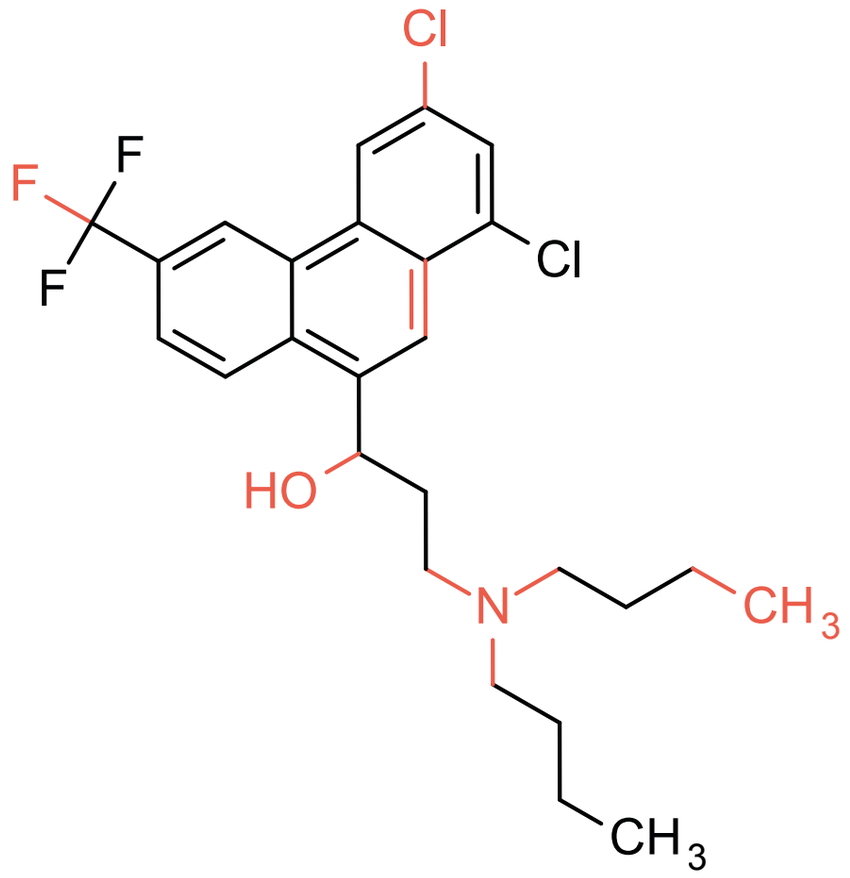
— F1: number of rings
— F2: solubility
— F3: number of fluors
— F4: molecular weight

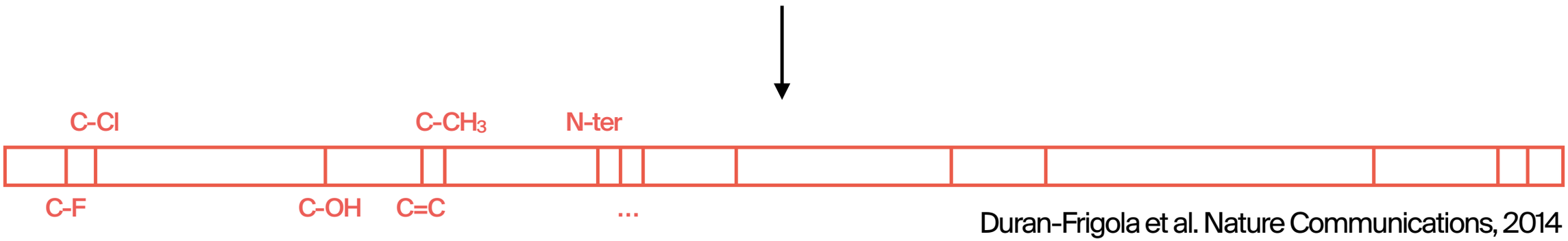






Classical chemical fingerprints

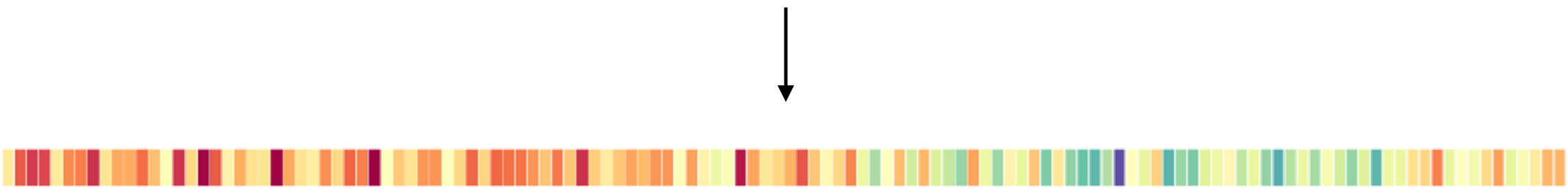




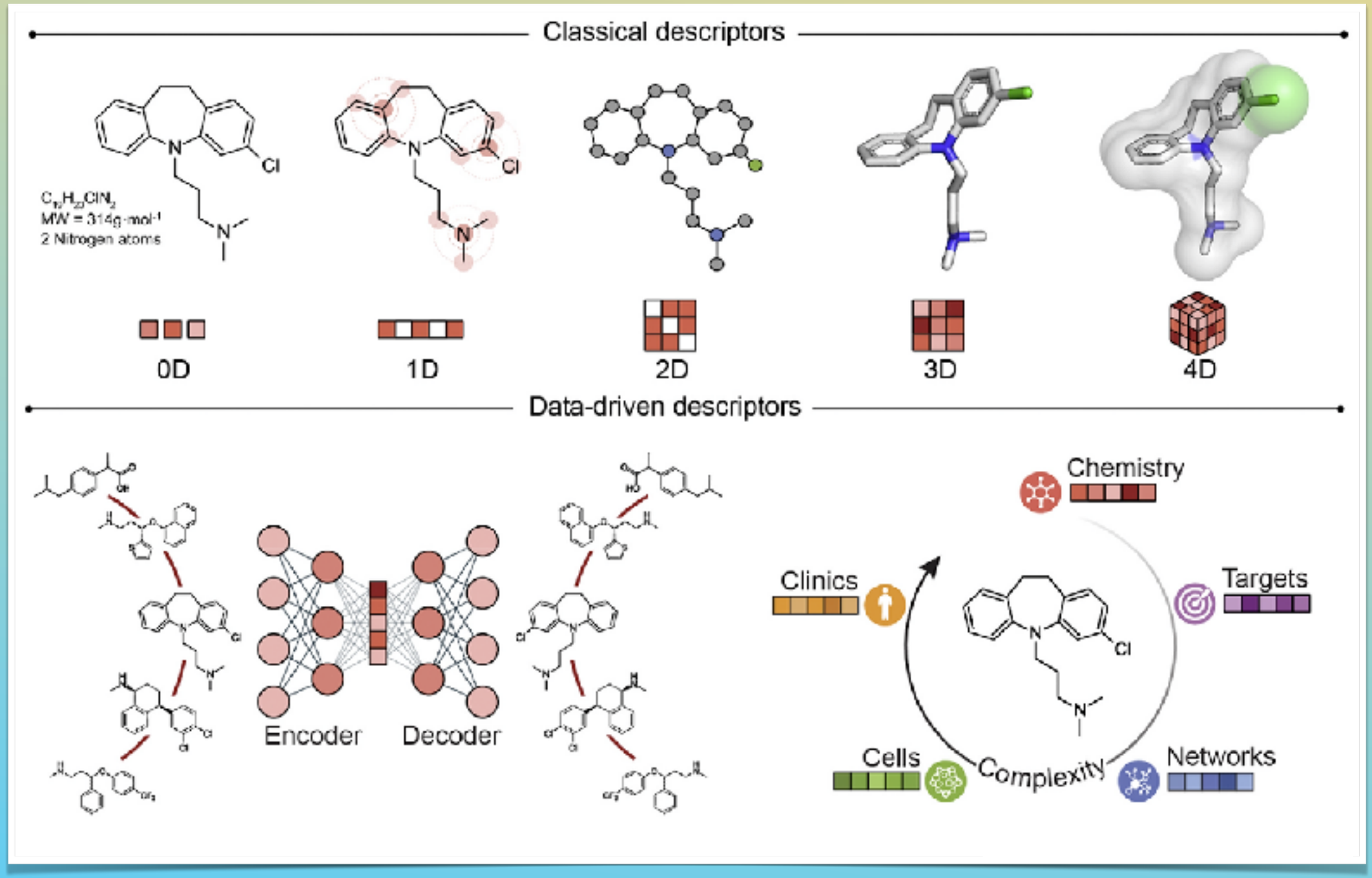
Duran-Frigola et al. Nature Communications, 2014

Chemical Checker signature

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.*



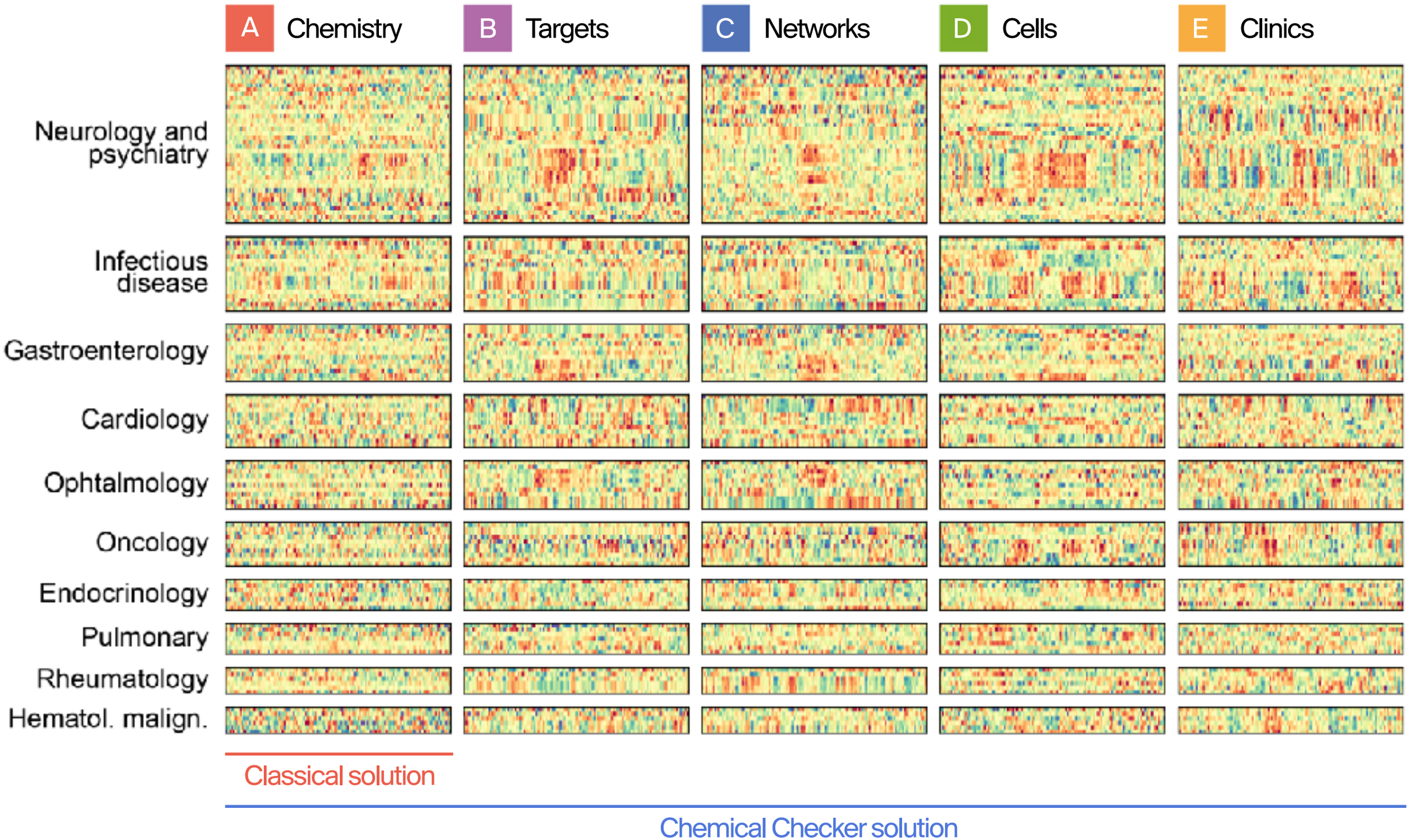
Chemical descriptors



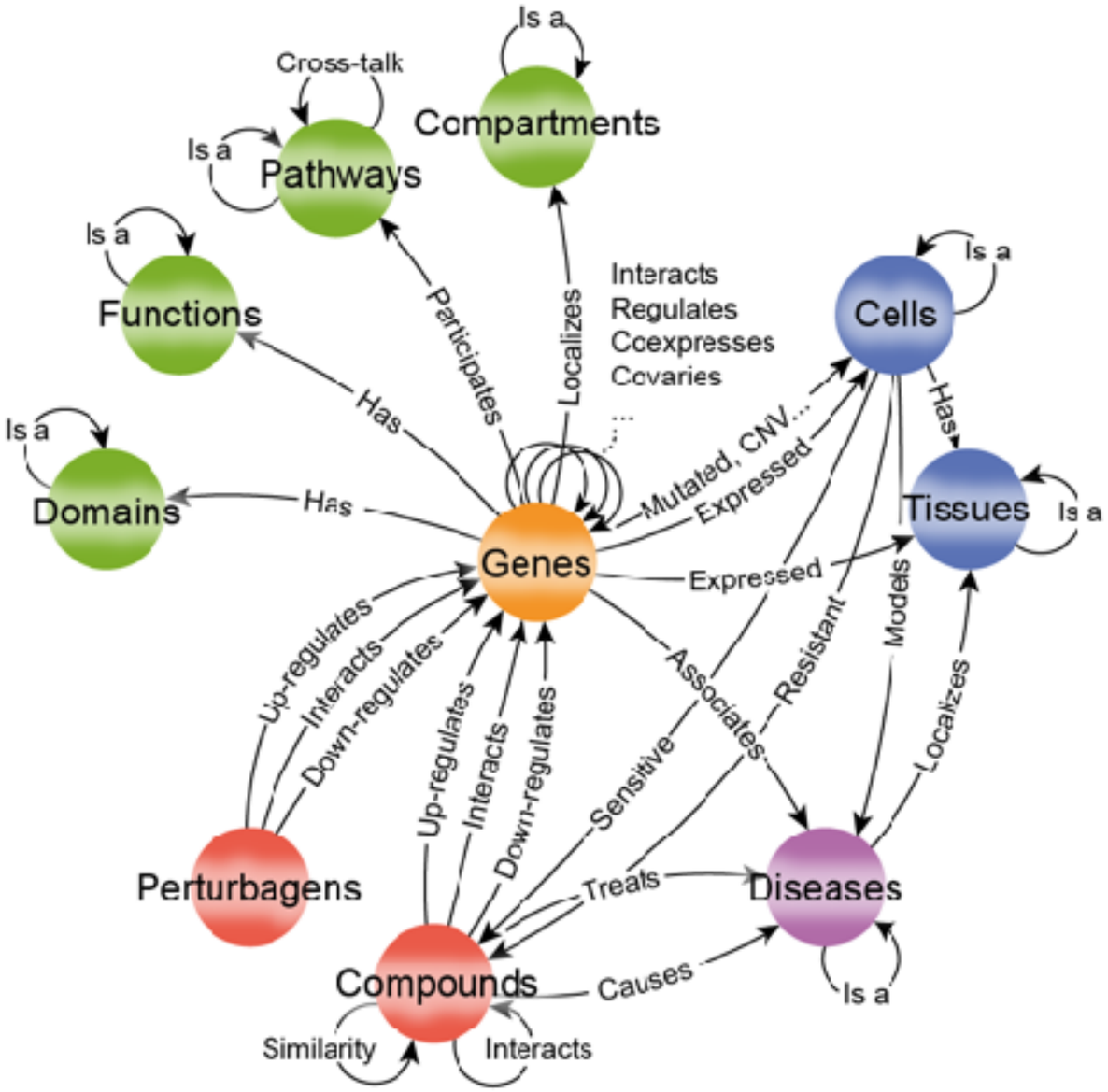
Fernandez-Torras, Curr Opin Chem Biol, 2021

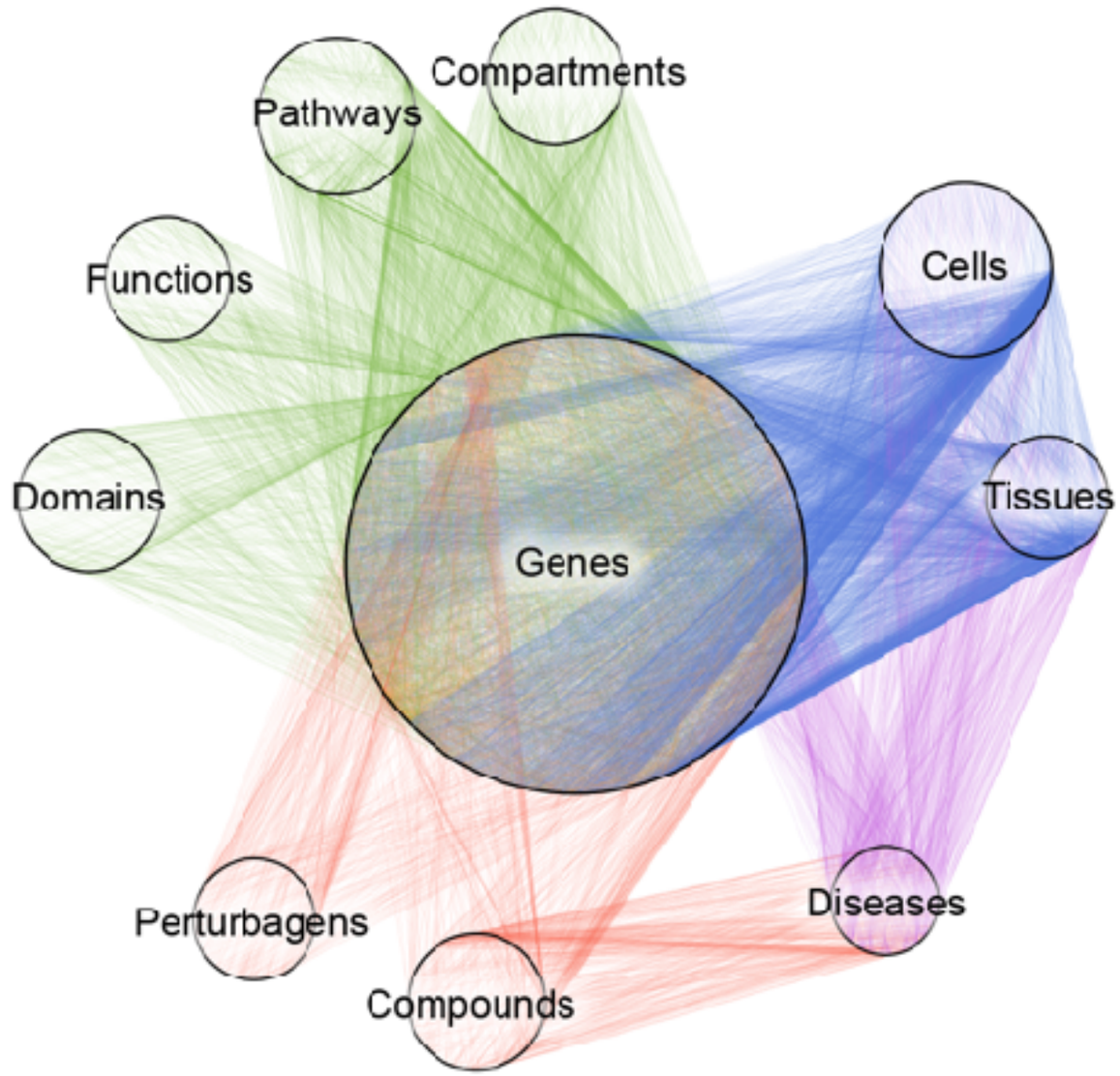
Chemical Checker signatures

Duran-Frigola et al, Nature Biotechnology, 2020



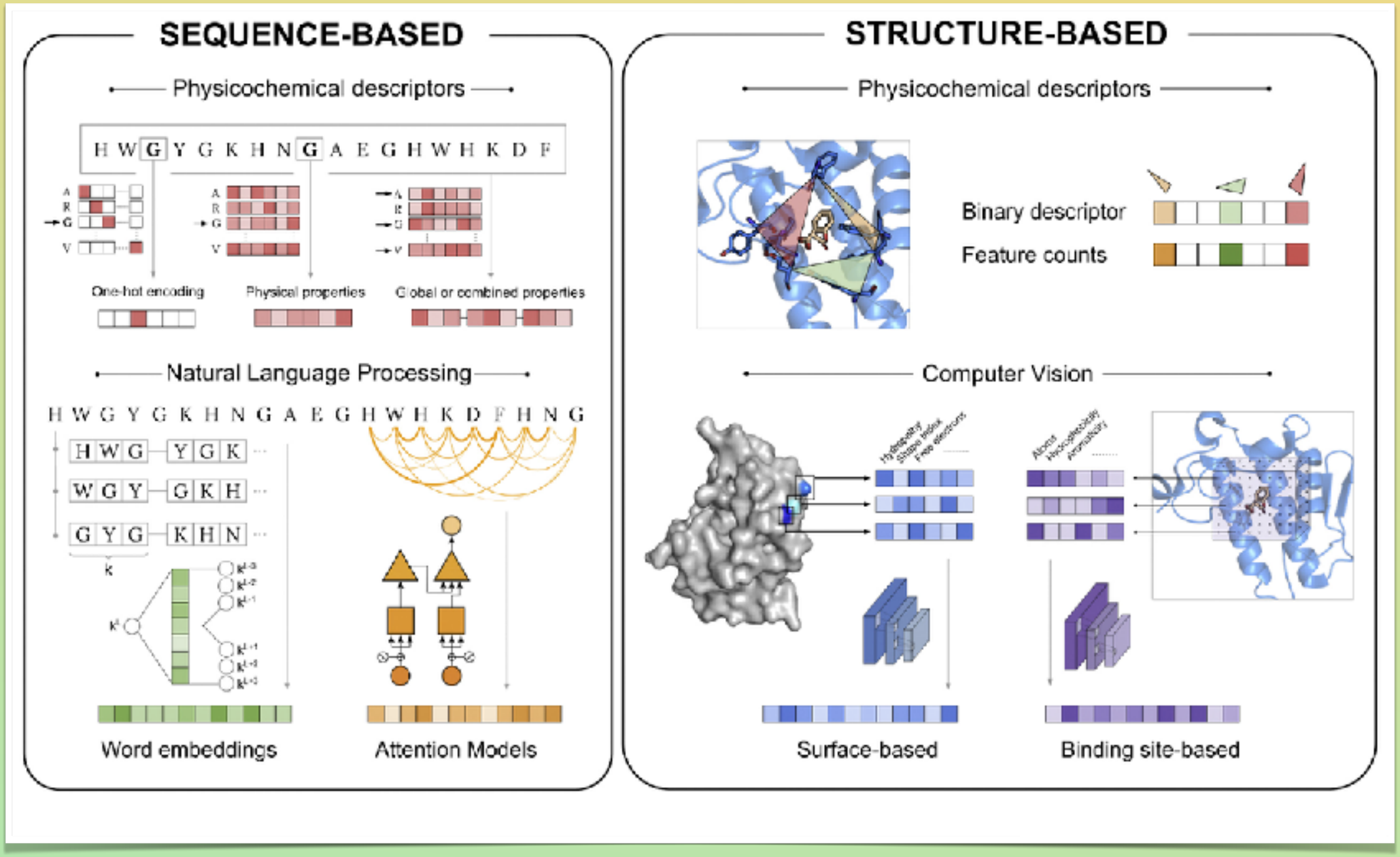
Biomedical knowledge graph





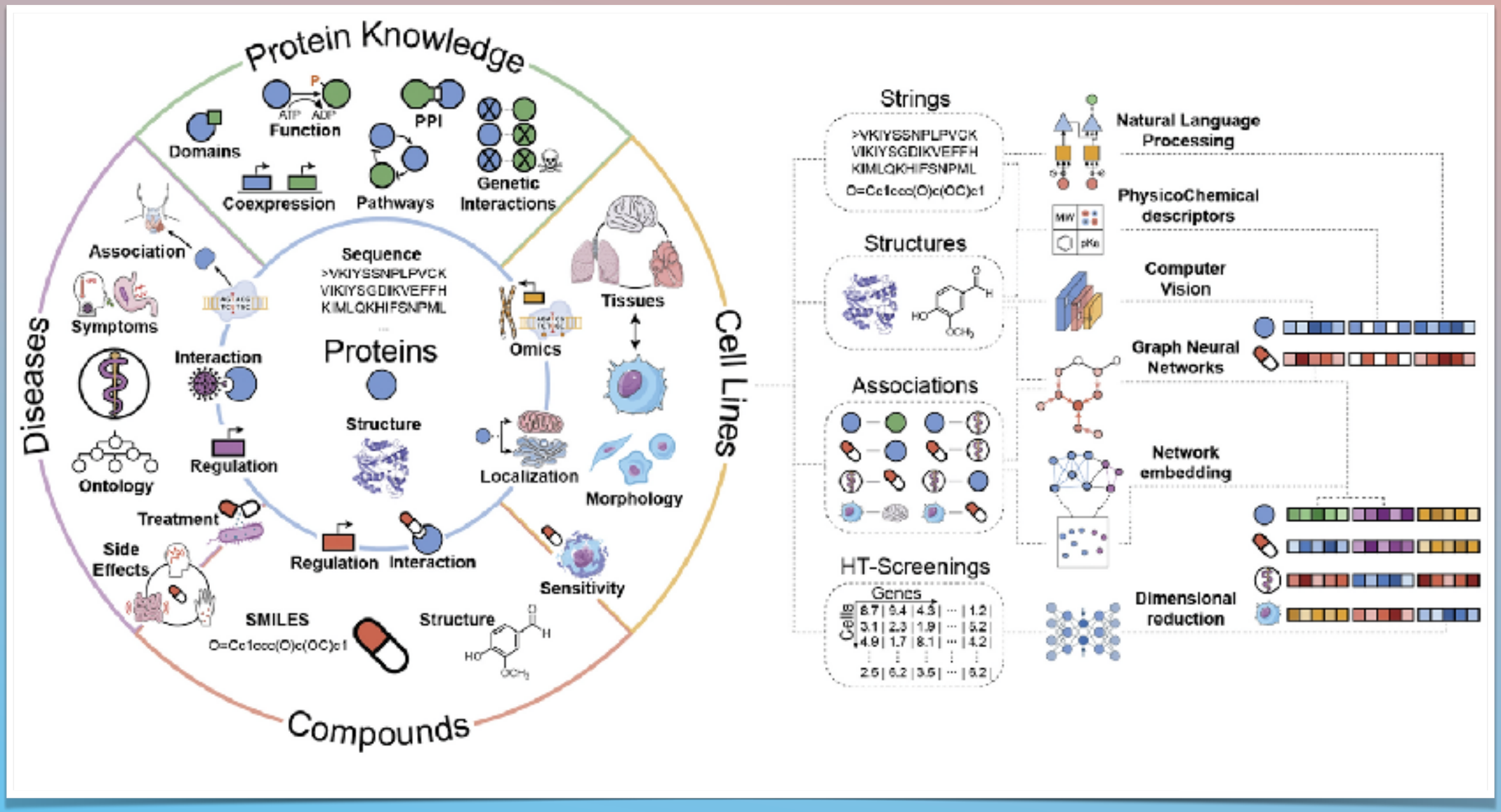
Duran-Frigola et al. WIREs Advanced Reviews, 2019
Fernandez-Torras et al, Nature Communications, 2022

Protein descriptors



Fernandez-Torras, Curr Opin Chem Biol, 2021

Holistic protein descriptors



Fernandez-Torras, Curr Opin Chem Biol, 2021

ALPHAGO

TRIBECA FILM FESTIVAL 2017

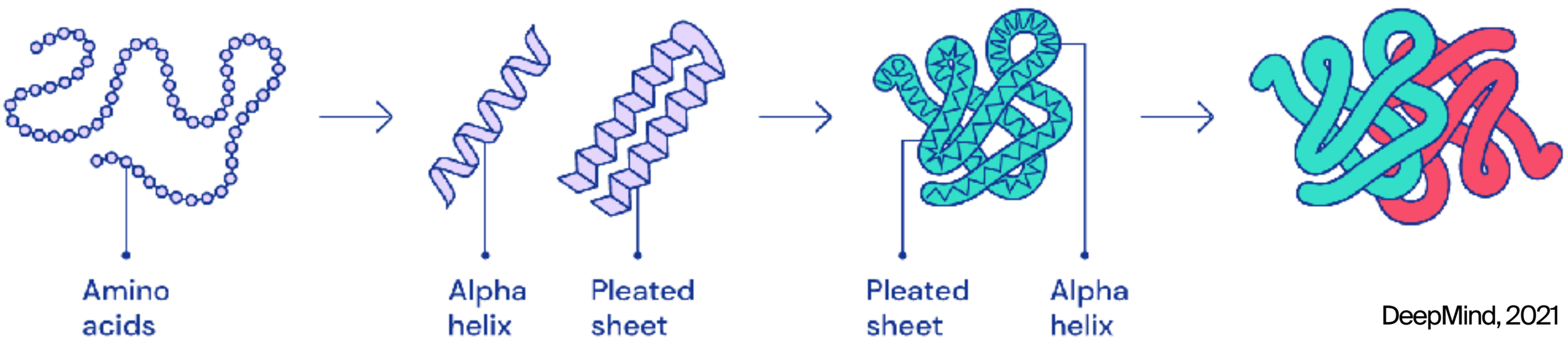
BFI LONDON FILM FESTIVAL OFFICIAL SELECTION 2017

WINNER TRAVERSE CITY FILM FESTIVAL 2017

OFFICIAL SELECTION MILL VALLEY FILM FESTIVAL 2017

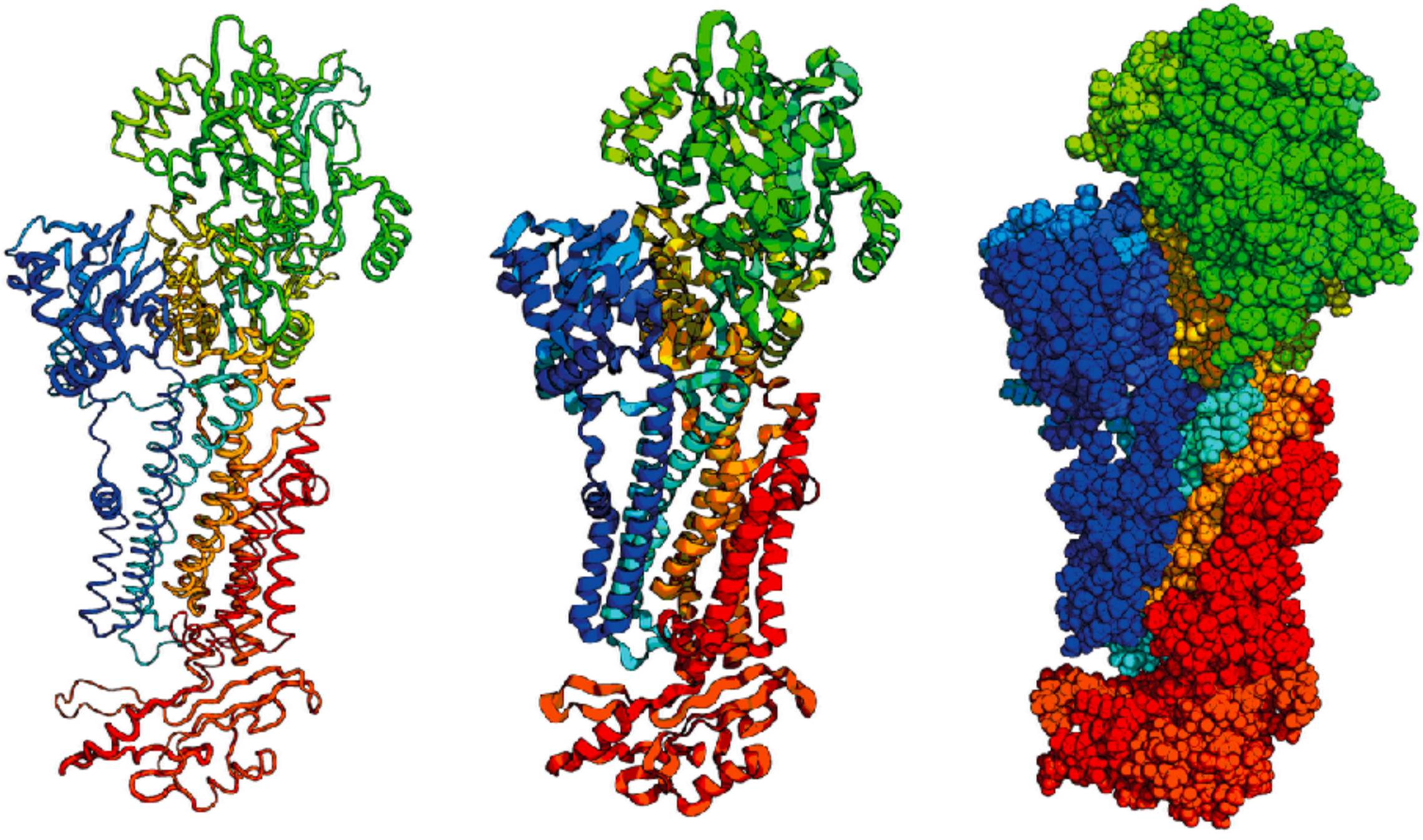
Official Selection WARSAW FILM FESTIVAL

VIFF Official Selection 2017

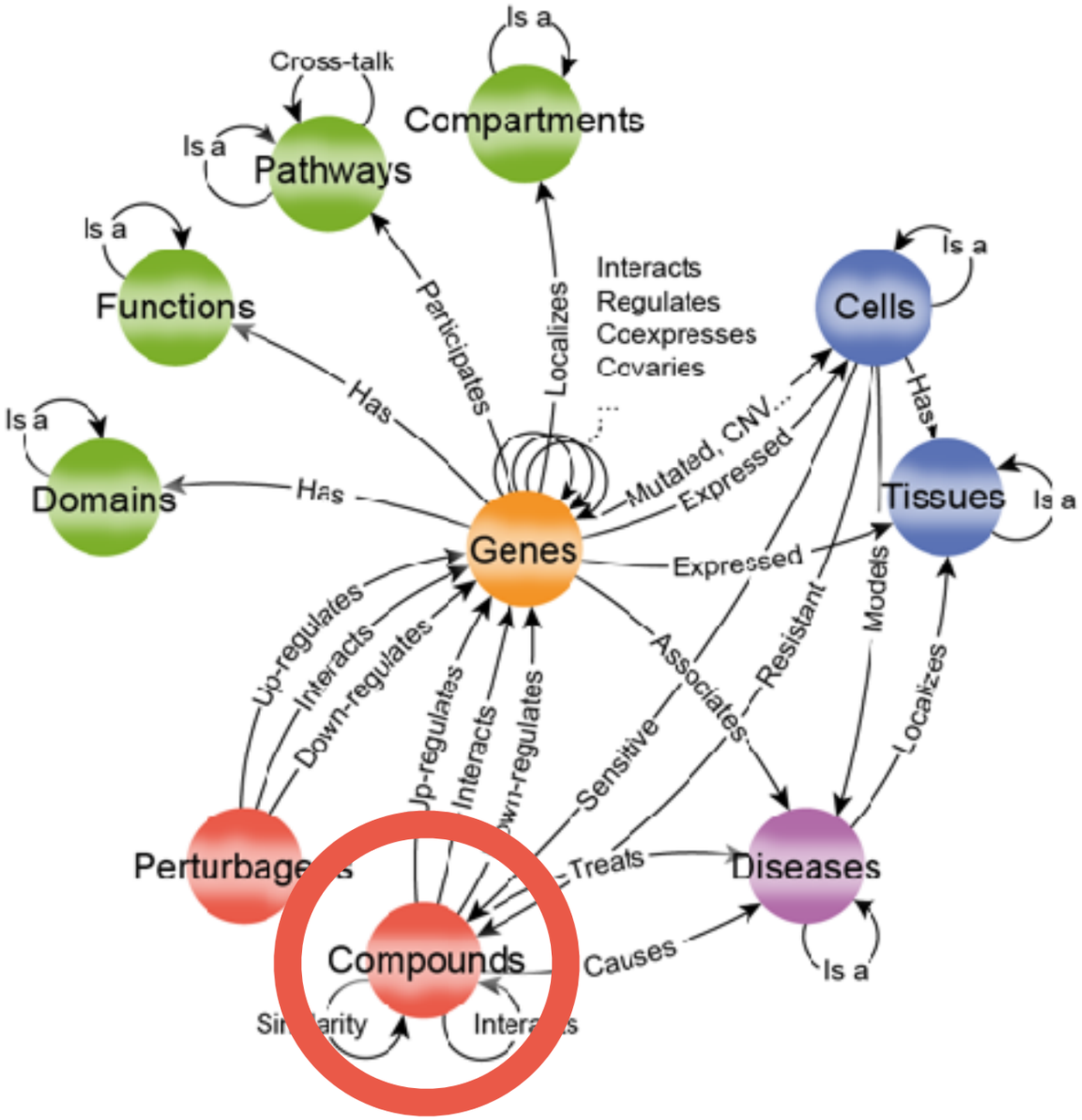


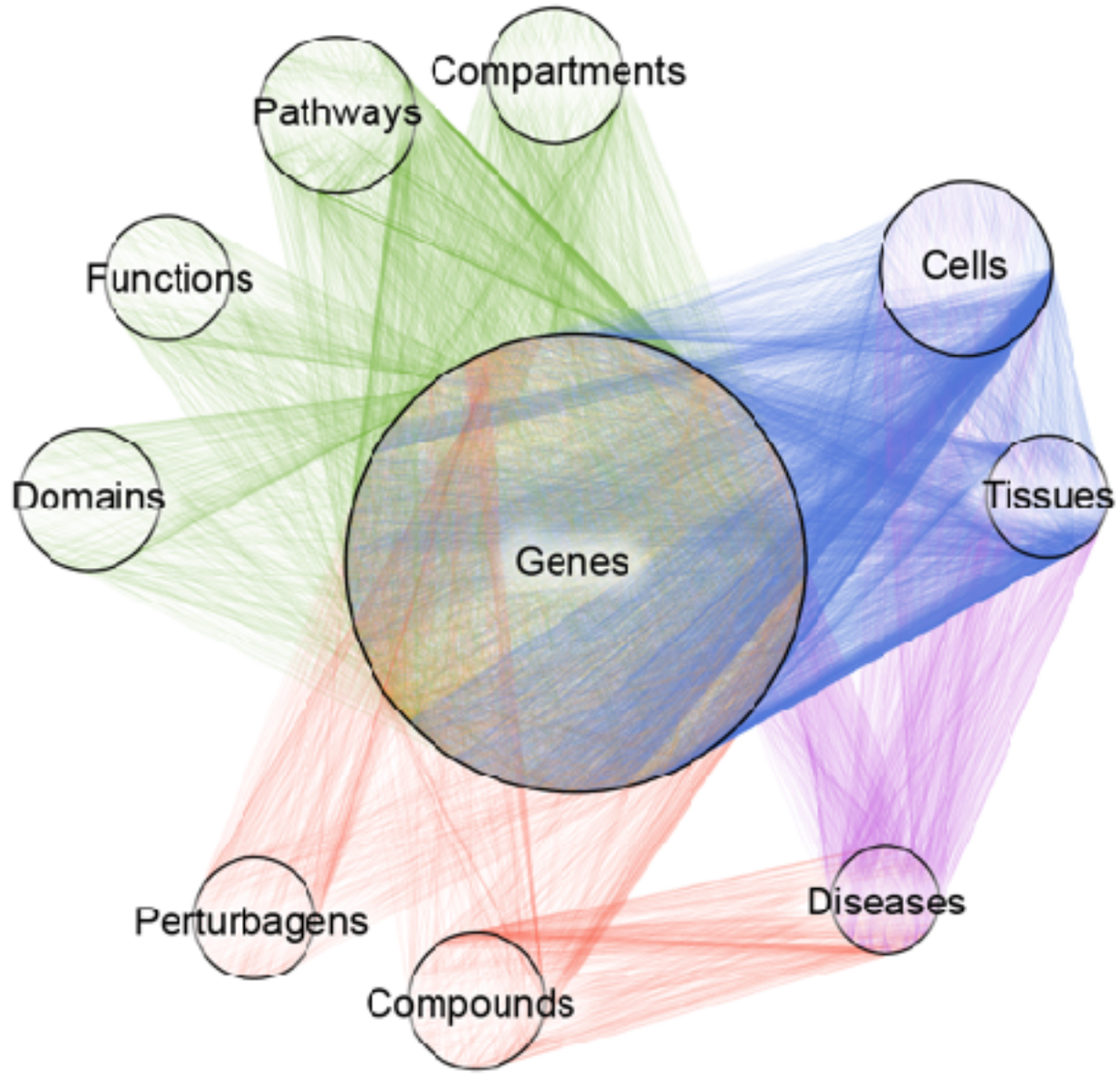
DeepMind, 2021

PfATP4 target of *P. falciparum*

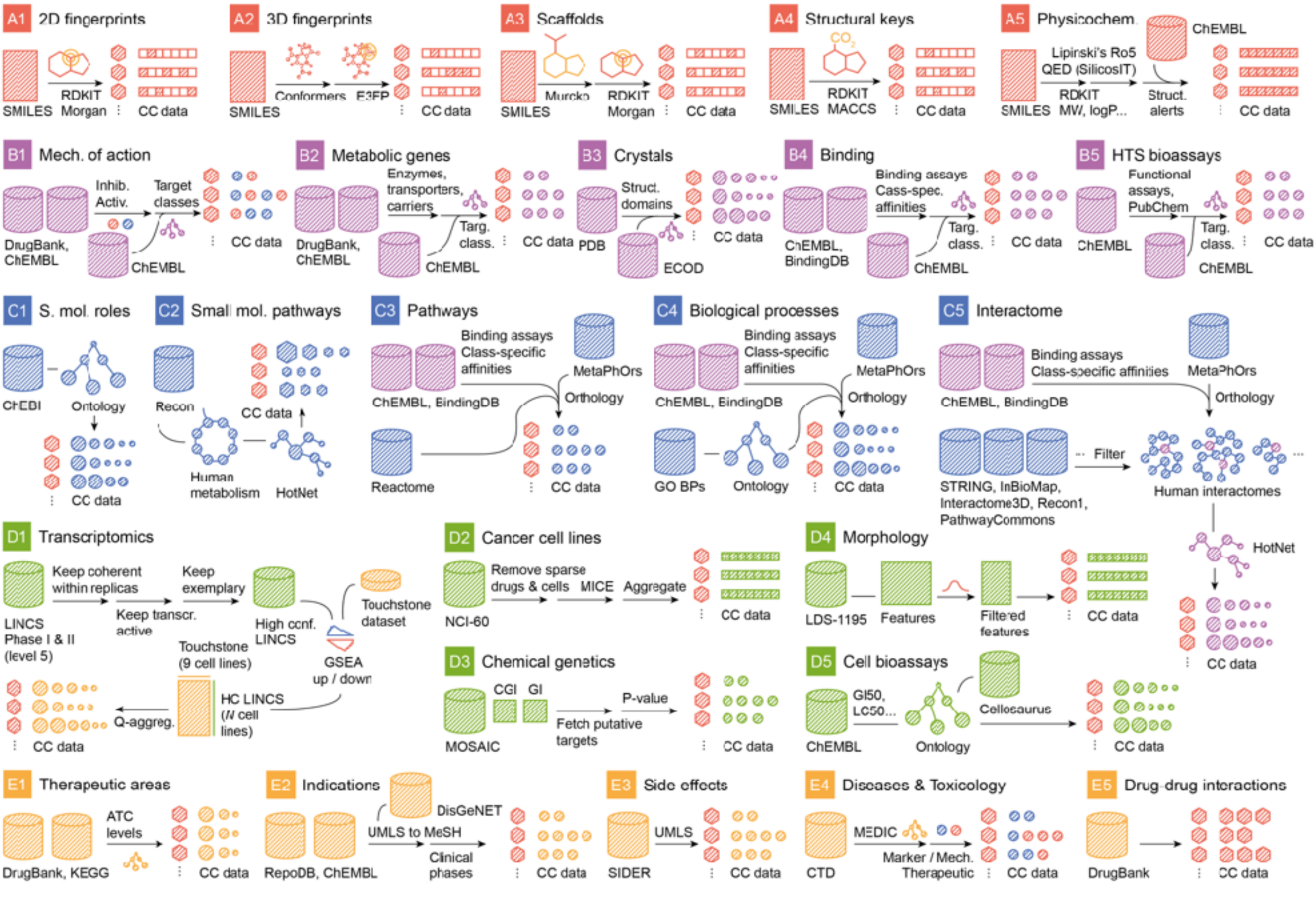


Back to compounds!





Duran-Frigola et al. WIREs Advanced Reviews, 2019
Fernandez-Torras et al, Nature Communications, 2022


A1 2D fingerprints
SMILES
RDKIT Morgan
CC data
A2 3D fingerprints
SMILES
Confomers
E3FP
CC data
A3 Scaffolds
SMILES
Murcko
RDKIT Morgan
CC data
A4 Structural keys
SMILES
RDKIT MACCS
CC data
A5 Physicochem.
SMILES
Lipinski's Ro5
QED (SilicosIT)
RDKIT MW, logP...
ChEMBL
Struct. alerts
CC data
B1 Mech. of action
DrugBank, ChEMBL
Inhib. Activ.
Target classes
ChEMBL
CC data
B2 Metabolic genes
DrugBank, ChEMBL
Enzymes, transporters, carriers
Targ. class.
ChEMBL
CC data
B3 Crystals
PDB
Struct. domains
ECOD
CC data
B4 Binding
ChEMBL, BindingDB
Binding assays Cass-spec. affinities
Targ. class.
ChEMBL
CC data
B5 HTS bioassays
ChEMBL
Functional assays, PubChem
Targ. class.
ChEMBL
CC data
C1 S. mol. roles
ChEBI
Ontology
CC data
C2 Small mol. pathways
Recon
Human metabolism
HotNet
CC data
C3 Pathways
ChEMBL, BindingDB
Binding assays Class-specific affinities
MetaPhOrs
Orthology
Reactome
CC data
C4 Biological processes
ChEMBL, BindingDB
Binding assays Class-specific affinities
MetaPhOrs
Orthology
GO BPs
Ontology
CC data
C5 Interactome
ChEMBL, BindingDB
Binding assays Class-specific affinities
MetaPhOrs
Orthology
STRING, InBioMap, Interactome3D, Recon1, PathwayCommons
Filter
Human interactomes
HotNet
CC data
D1 Transcriptomics
LINCS Phase I & II (level 5)
Keep coherent within replicas
Keep transcr. active
Keep exemplary
High conf. LINCS
Touchstone dataset
Touchstone (9 cell lines)
GSEA up / down
HC LINCS (N cell lines)
Q-aggreg.
CC data
D2 Cancer cell lines
NCI-60
Remove sparse drugs & cells
MICE
Aggregate
CC data
D3 Chemical genetics
MOSAIC
CGI
GI
Fetch putative targets
P-value
CC data
D4 Morphology
LDS-1195
Features
Filtered features
CC data
D5 Cell bioassays
ChEMBL
GI50, LC50...
Ontology
Cellosaurus
CC data
E1 Therapeutic areas
DrugBank, KEGG
ATC levels
CC data
E2 Indications
RepoDB, ChEMBL
UMLS to MeSH
DisGeNET
Clinical phases
CC data
E3 Side effects
SIDER
UMLS
CC data
E4 Diseases & Toxicology
CTD
MEDIC
Marker / Mech.
Therapeutic
CC data
E5 Drug-drug interactions
DrugBank
CC data

The bioactive chemical space, organised

Chemistry	2D fingerprints	3D fingerprints	Scaffolds	Structural keys	Physico-chemistry
Targets	Mechanism of action	Metabolic genes	Crystals	Binding	HTS bioassays
Networks	Small mol. roles	Small mol. pathways	Signaling pathways	Biological processes	Interactome
Cells	Gene expression	Cancer cell lines	Chemical genetics	Morphology	Cell bioassays
Clinics	Therapeutic areas	Indications	Side effects	Diseases and toxicology	Drug-drug interactions

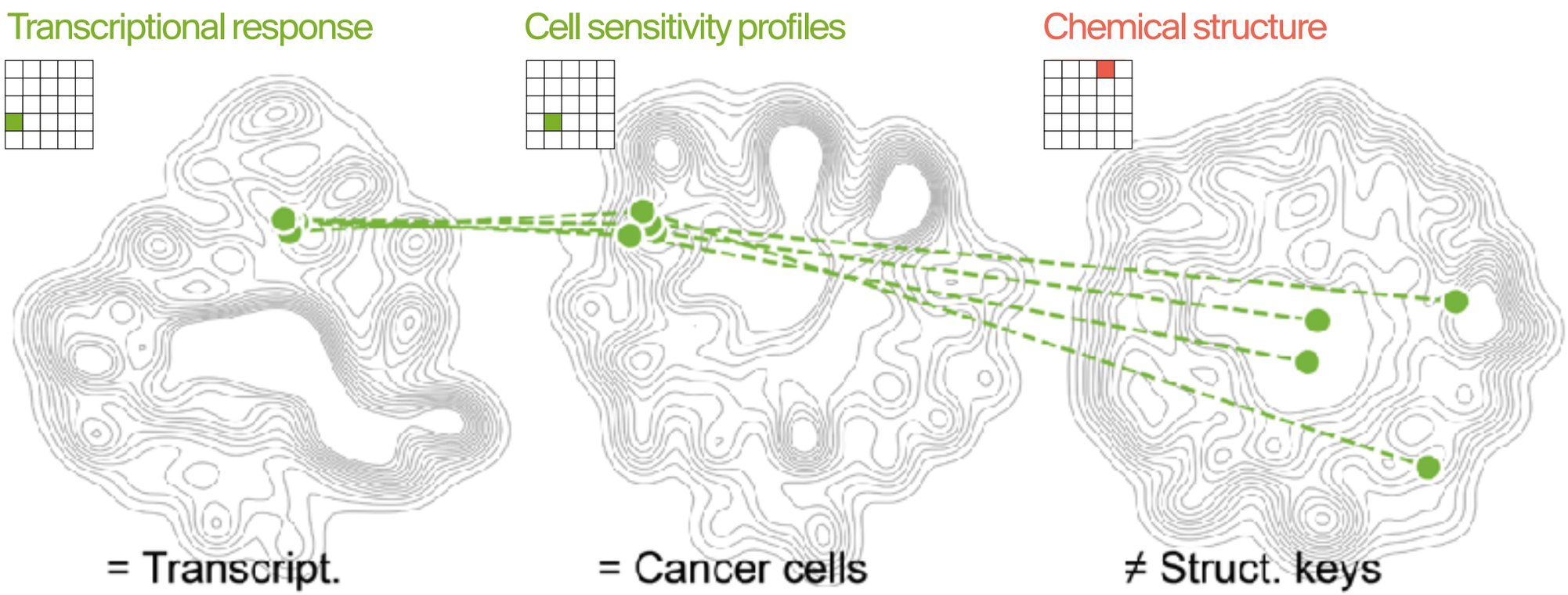
- 800k bioactive molecules
- 25 data types, from chemistry to the clinics
- The major small molecule databases are integrated
- chemicalchecker.org
- bioactivitysignatures.org

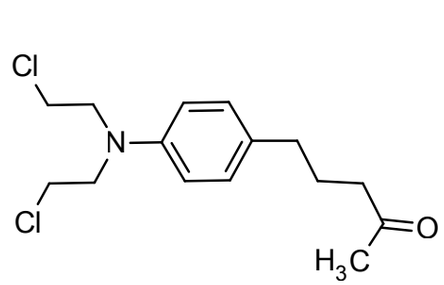
Chemical Checker

Duran-Frigola et al. Nat Biotech, 2020

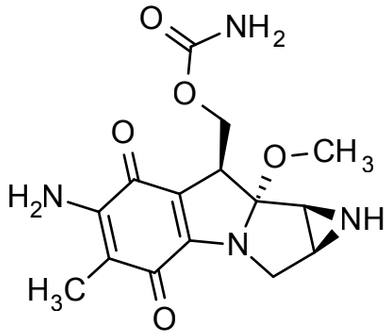
🏆 CTD2-pancancer DREAM challenge, 2020

Complex similarity searches

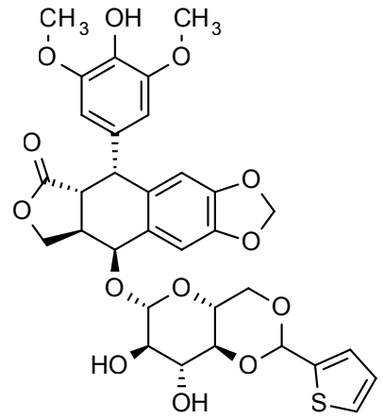




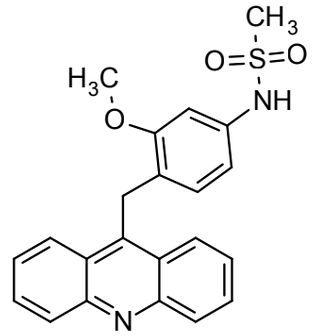
Chlorambucil



Mitomycin

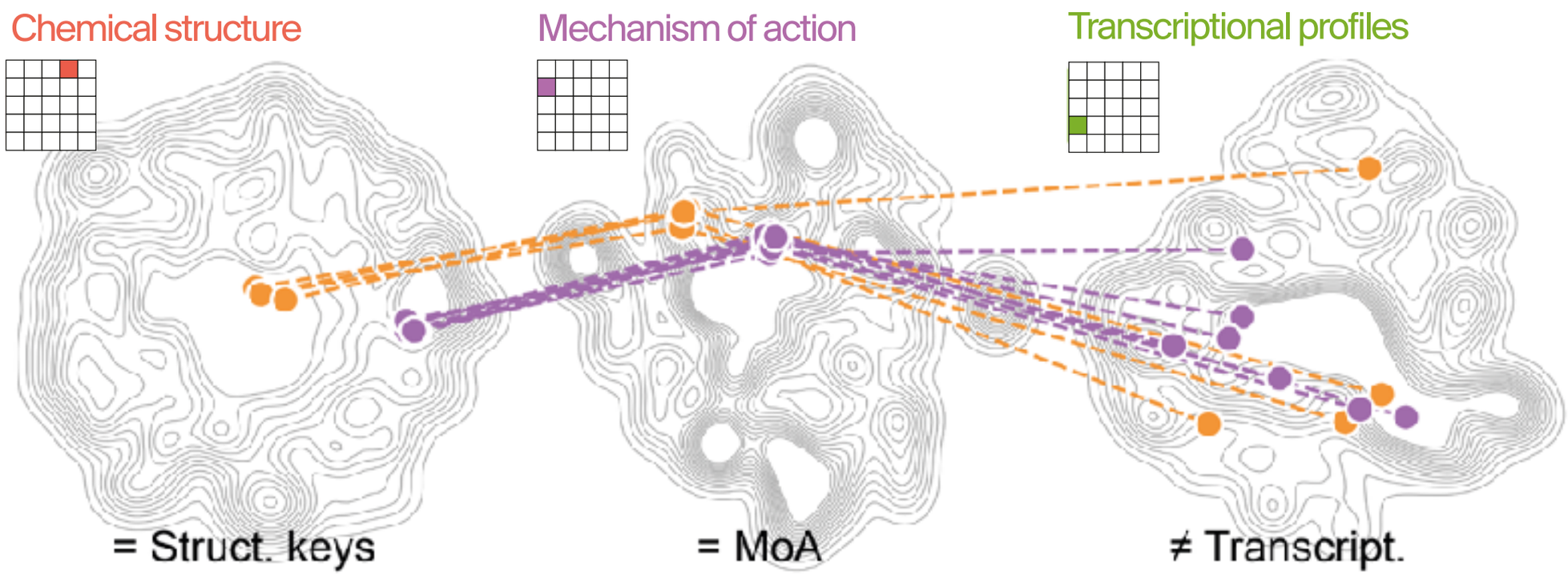


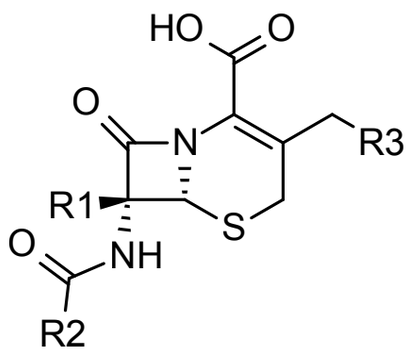
Teniposide



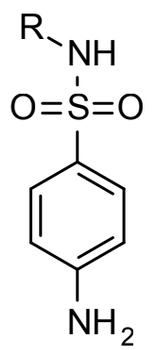
New molecule

Complex similarity searches





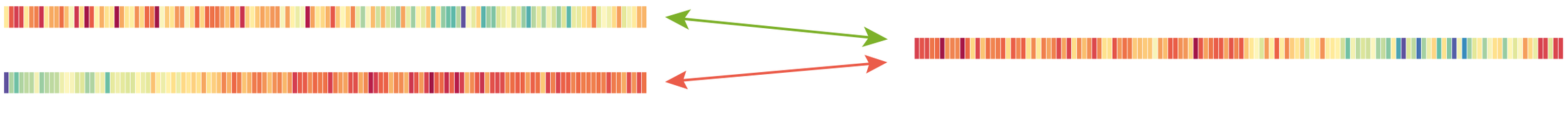
Cephalosporins



Sulphonamides

Chemical mimetics of biological drugs

> 10,000 compounds considered

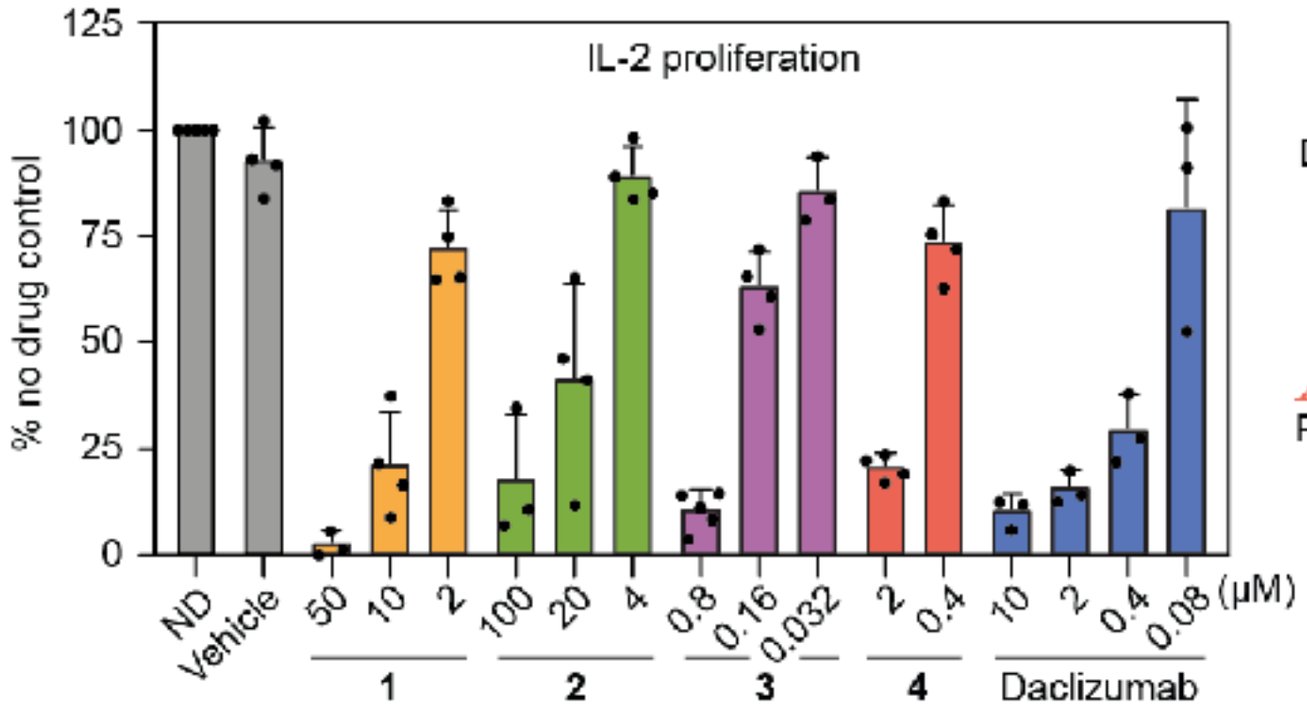


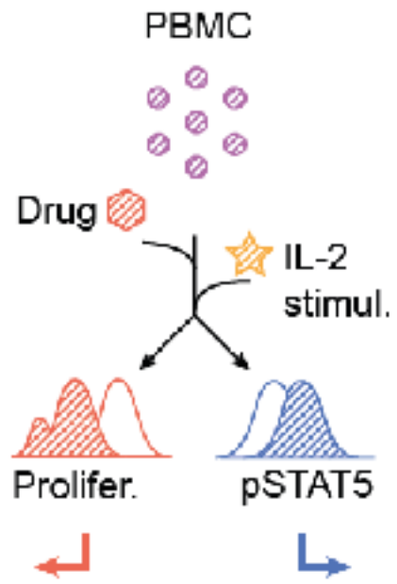
Transcriptional effects of compounds, measured in cell lines. Known and inferred targets of the compounds, together with their signalling pathways.

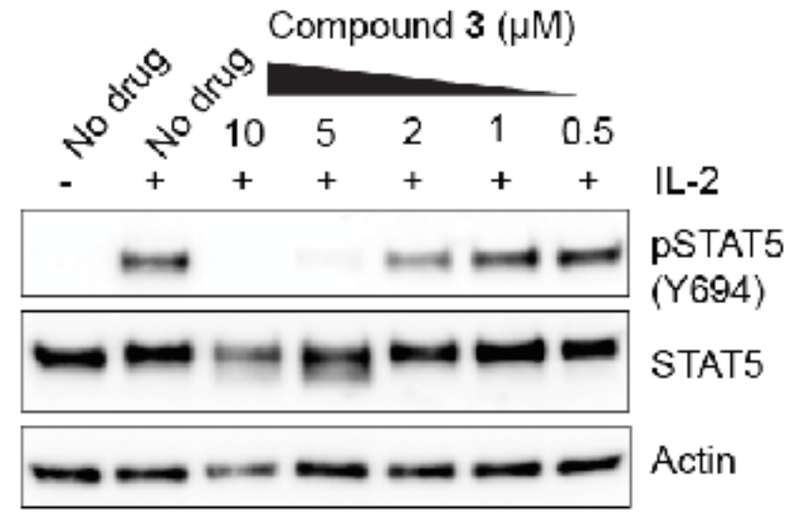
IL-2 receptor antibody (daclizumab)

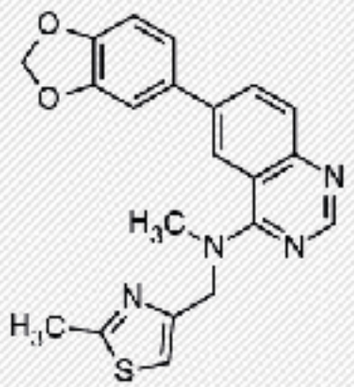
Transcriptional effects of the antibody drug, measured in cell lines. Known targets of the antibody drug, together with their signalling pathways.

In vitro experiments

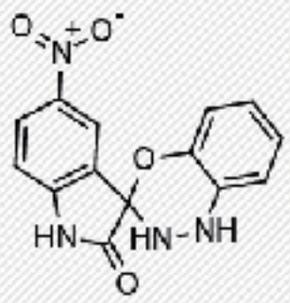




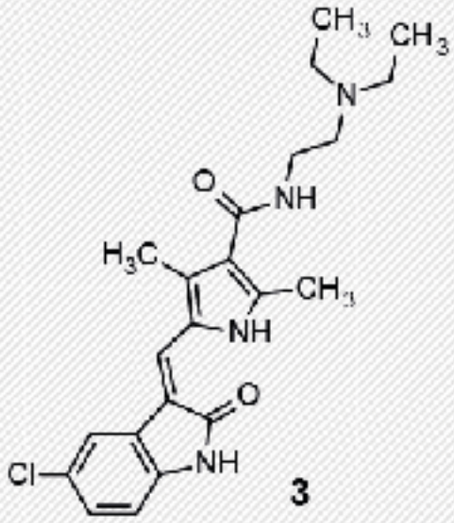




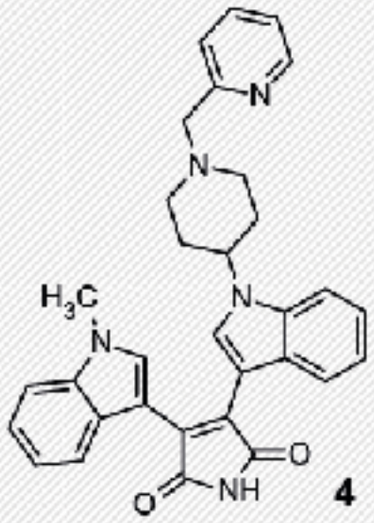
1



2



3



4

...14 of 19 tested worked*

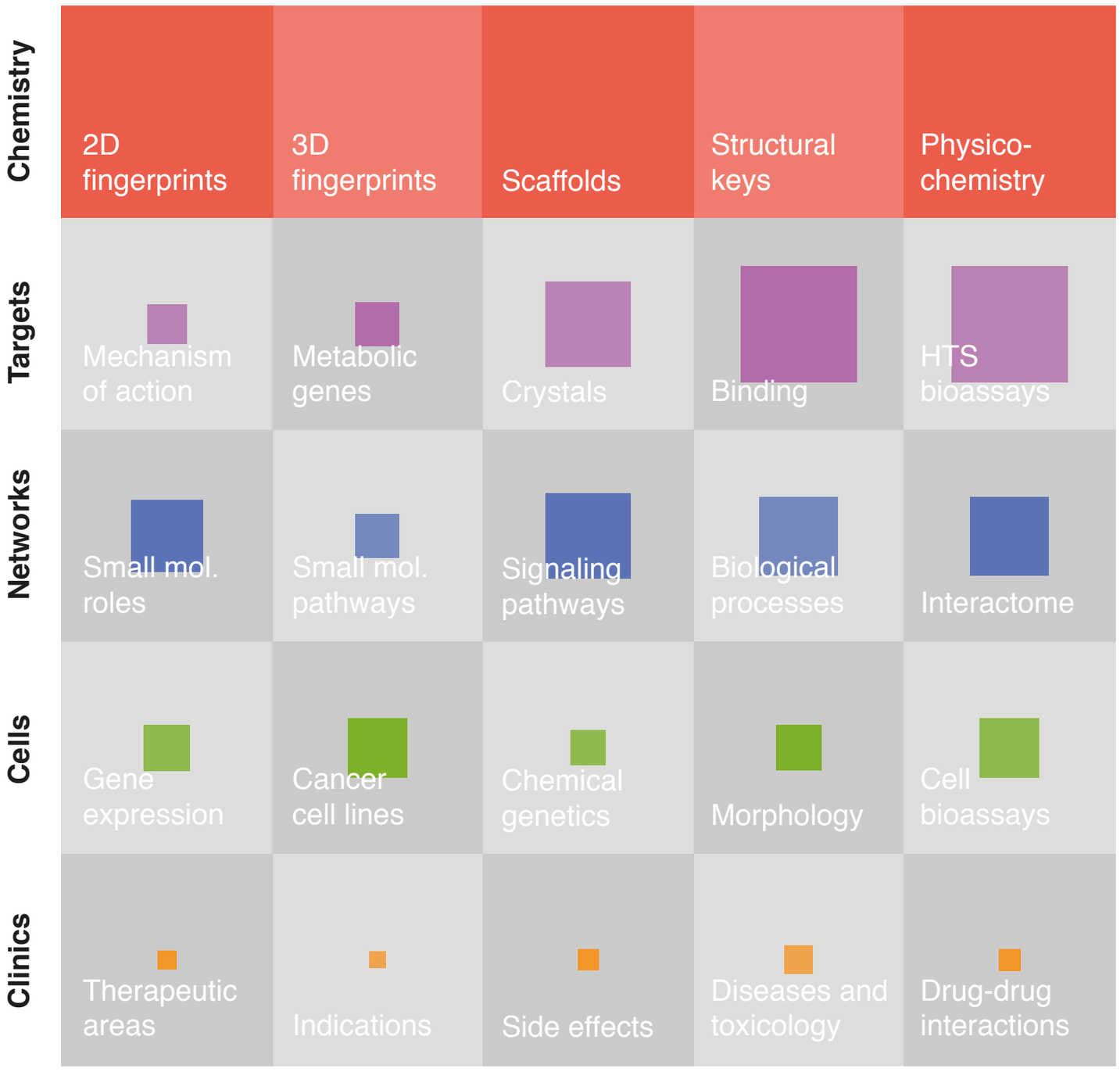
Event Fund

Chemical space

Chemical space


SEPTENTRIO
SCYTHICVS OCEANVS
EOVS OCEANVS
ORIENS
INDICVS OCEANVS
MARE RVBRVM
AETHIOPICVM MARE
ATLANTICVM MARE
OCCIDENS
BRITANNICVS OCEANVS
Britannia insula
EVROPA
ASIA
AFRICA
SCYTHIA
SERES
INDIA
PERSAE
ARIANE
SYRIA
ARMENIA
SARMATIA
GERMANIA
GALLIA
GALLIA TOGATA
HISPANIA
ITALIA
PALVS MAEOTIS
PONTVS EVXINVS
Caspium Mare
MARE NOSTRVM
LIBYCVM MARE
FRETVM
AEGYPTII
CYRENAICA
AFRICA MINOR
Garamantae
Arabicum Mare
Arabia Eudaemon
Persicum Mare
Mesopotamia
HESPERII AETHIOPES
AETHIOPES
Taprobane insula

An honest view of the Chemical Checker

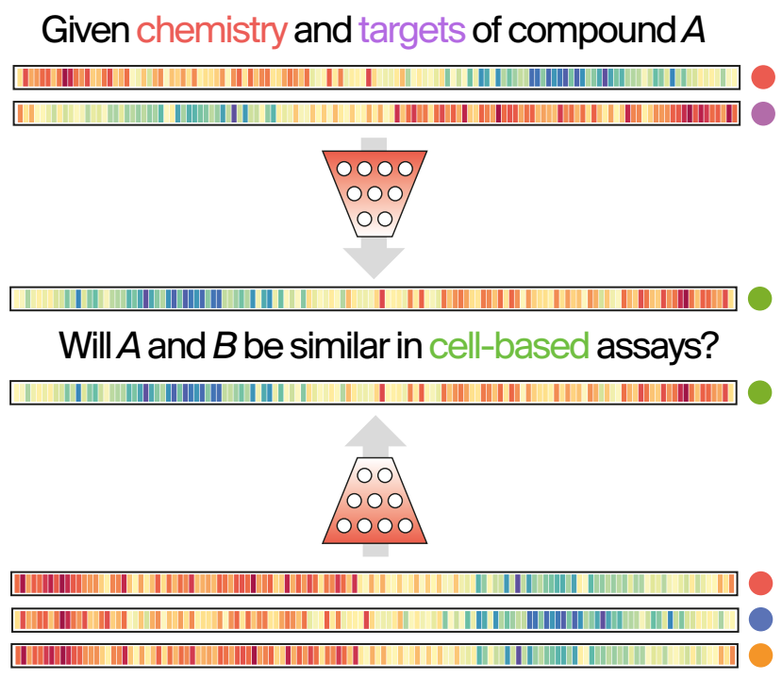


Bertoni*, Duran-Frigola* et al. Nat Commun, 2021

An honest view of the Chemical Checker

Chemistry	2D fingerprints	3D fingerprints	Scaffolds	Structural keys	Physico-chemistry
Targets	Mechanism of action	Metabolic genes	Crystals	Binding	HTS bioassays
Networks	Small mol. roles	Small mol. pathways	Signaling pathways	Biological processes	Interactome
Cells	Gene expression	Cancer cell lines	Chemical genetics	Morphology	Cell bioassays
Clinics	Therapeutic areas	Indications	Side effects	Diseases and toxicology	Drug-drug interactions

– Siamese neural nets 🤓

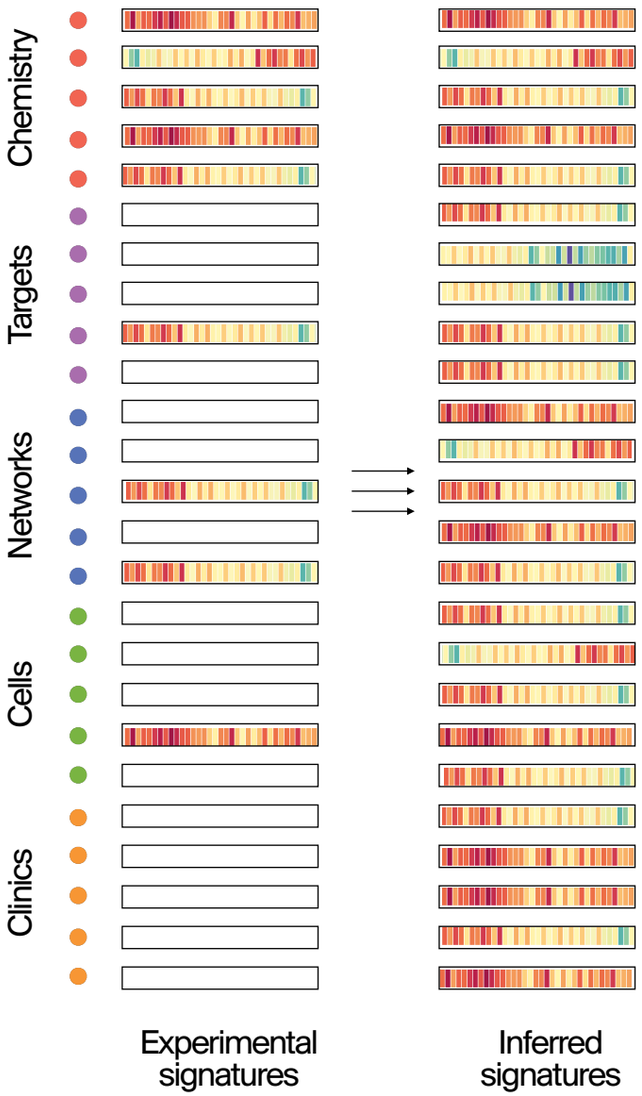


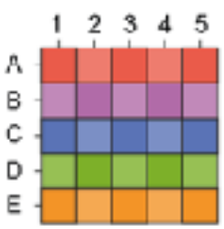
– Trained over a week in a supercomputing center

Bertoni*, Duran-Frigola* et al. Nat Commun, 2021

Large-scale inference of bioactivity signatures

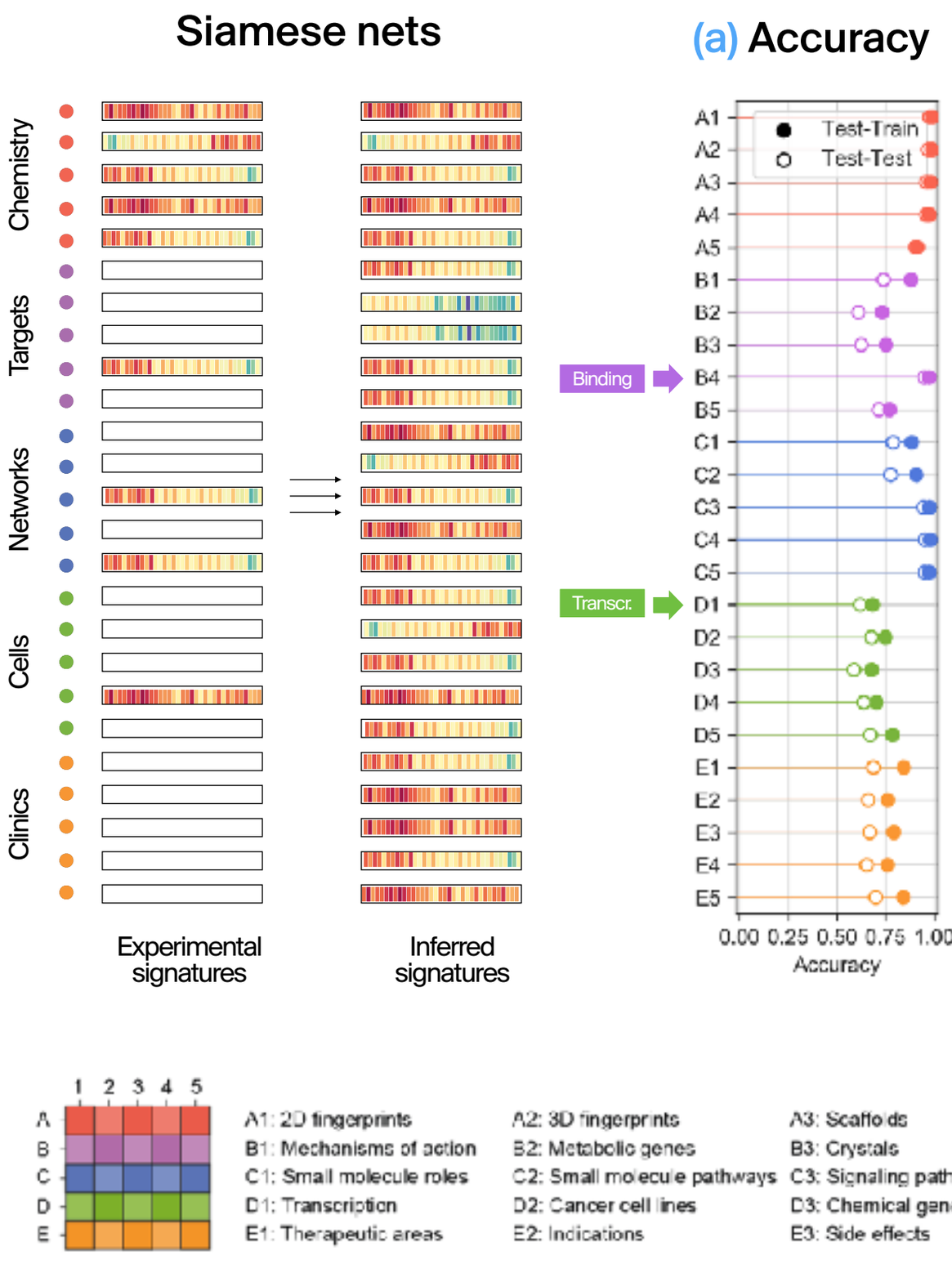
Siamese nets





A1: 2D fingerprints	A2: 3D fingerprints	A3: Scaffolds	A4: Structural keys	A5: Physicochemistry
B1: Mechanisms of action	B2: Metabolic genes	B3: Crystals	B4: Binding	B5: HTS bioassays
C1: Small molecule roles	C2: Small molecule pathways	C3: Signaling pathways	C4: Biological processes	C5: Interactome
D1: Transcription	D2: Cancer cell lines	D3: Chemical genetics	D4: Morphology	D5: Cell bioassays
E1: Therapeutic areas	E2: Indications	E3: Side effects	E4: Diseases & toxicology	E5: Drug-drug interactions

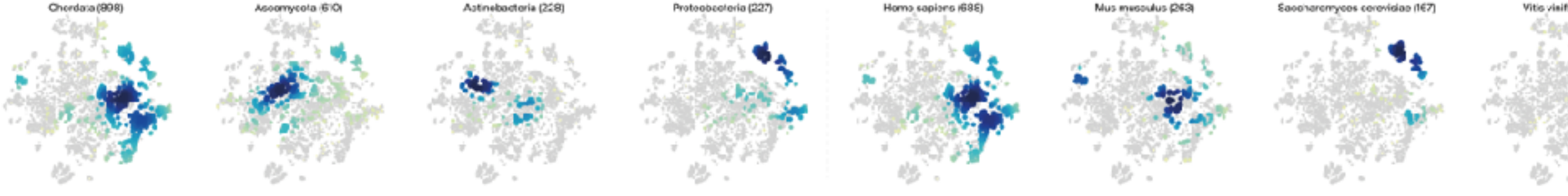
Large-scale inference of bioactivity signatures



A1: 2D fingerprints	A2: 3D fingerprints	A3: Scaffolds	A4: Structural keys	A5: Physicochemistry
B1: Mechanisms of action	B2: Metabolic genes	B3: Crystals	B4: Binding	B5: HTS bioassays
C1: Small molecule roles	C2: Small molecule pathways	C3: Signaling pathways	C4: Biological processes	C5: Interactome
D1: Transcription	D2: Cancer cell lines	D3: Chemical genetics	D4: Morphology	D5: Cell bioassays
E1: Therapeutic areas	E2: Indications	E3: Side effects	E4: Diseases & toxicology	E5: Drug-drug interactions

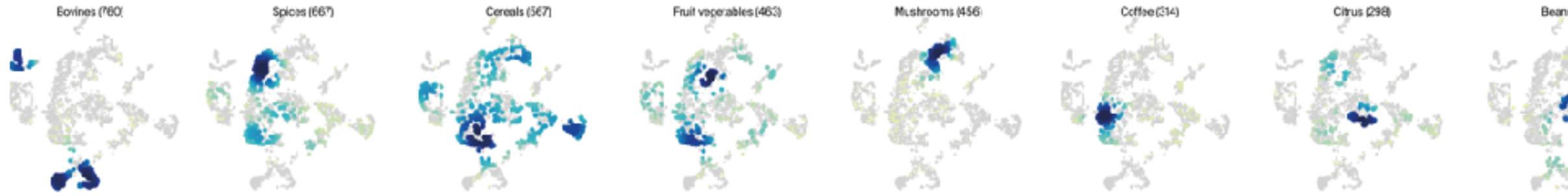
— Drug molecules

— Species metabolites

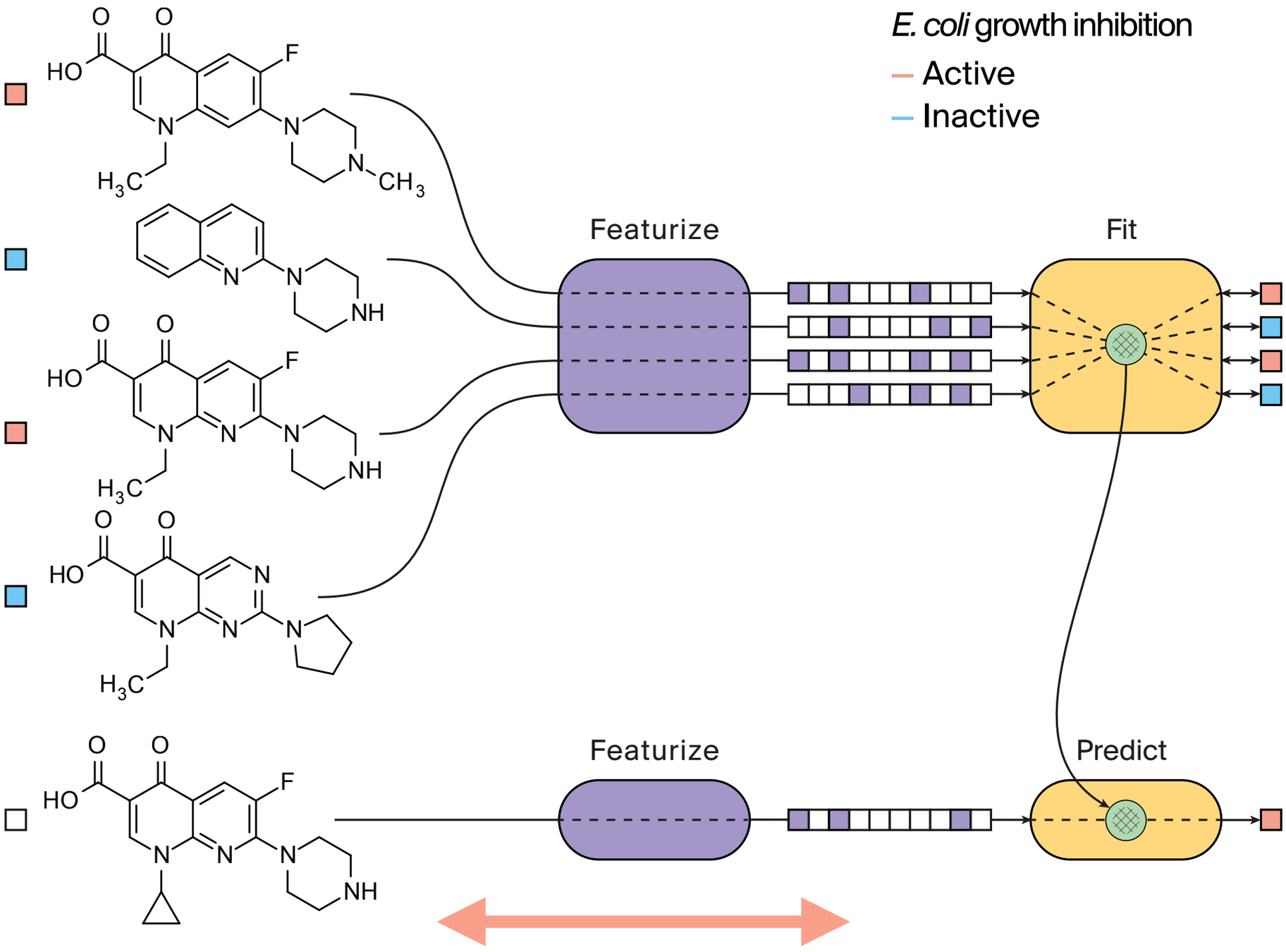


— Tissues and fluids

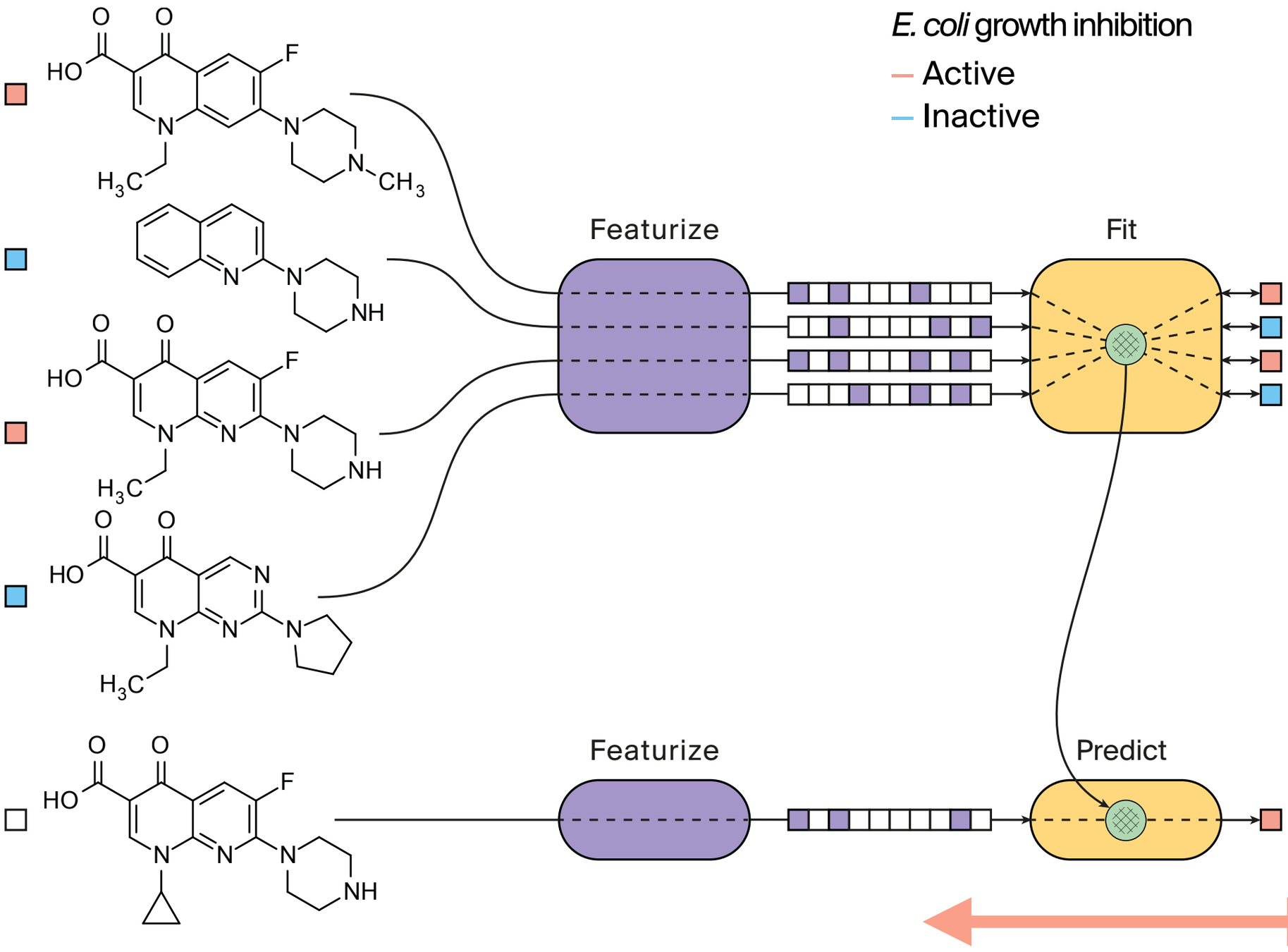
— Food ingredients



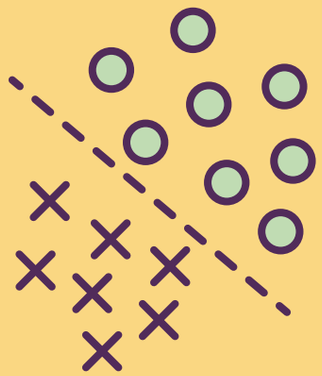
A simple supervised task in drug discovery (aka QSAR)
Antibiotic activity experiments (*E. coli*)



A simple supervised task in drug discovery (aka QSAR)
Antibiotic activity experiments (*E. coli*)

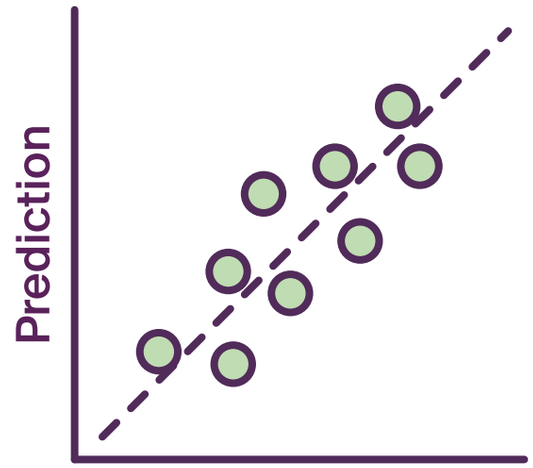


Classification



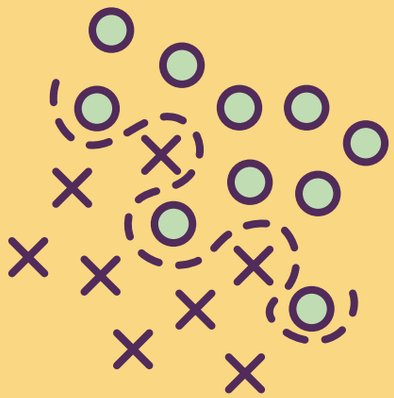
Prediction task:
Active = 1
Inactive = 0

Regression

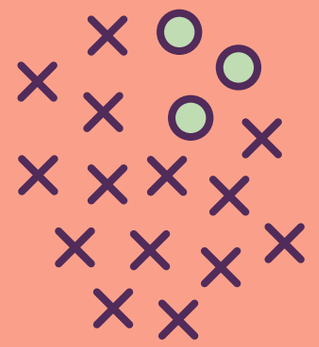


Prediction task:
IC50 value

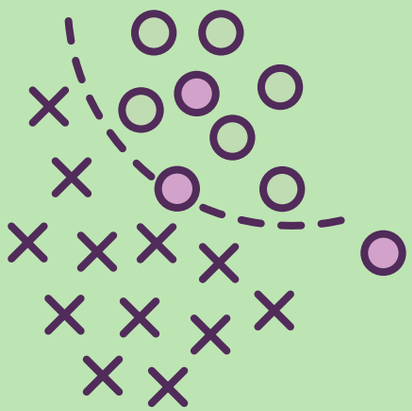
Overfitting



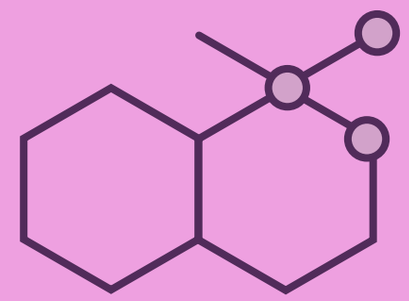
Imbalance



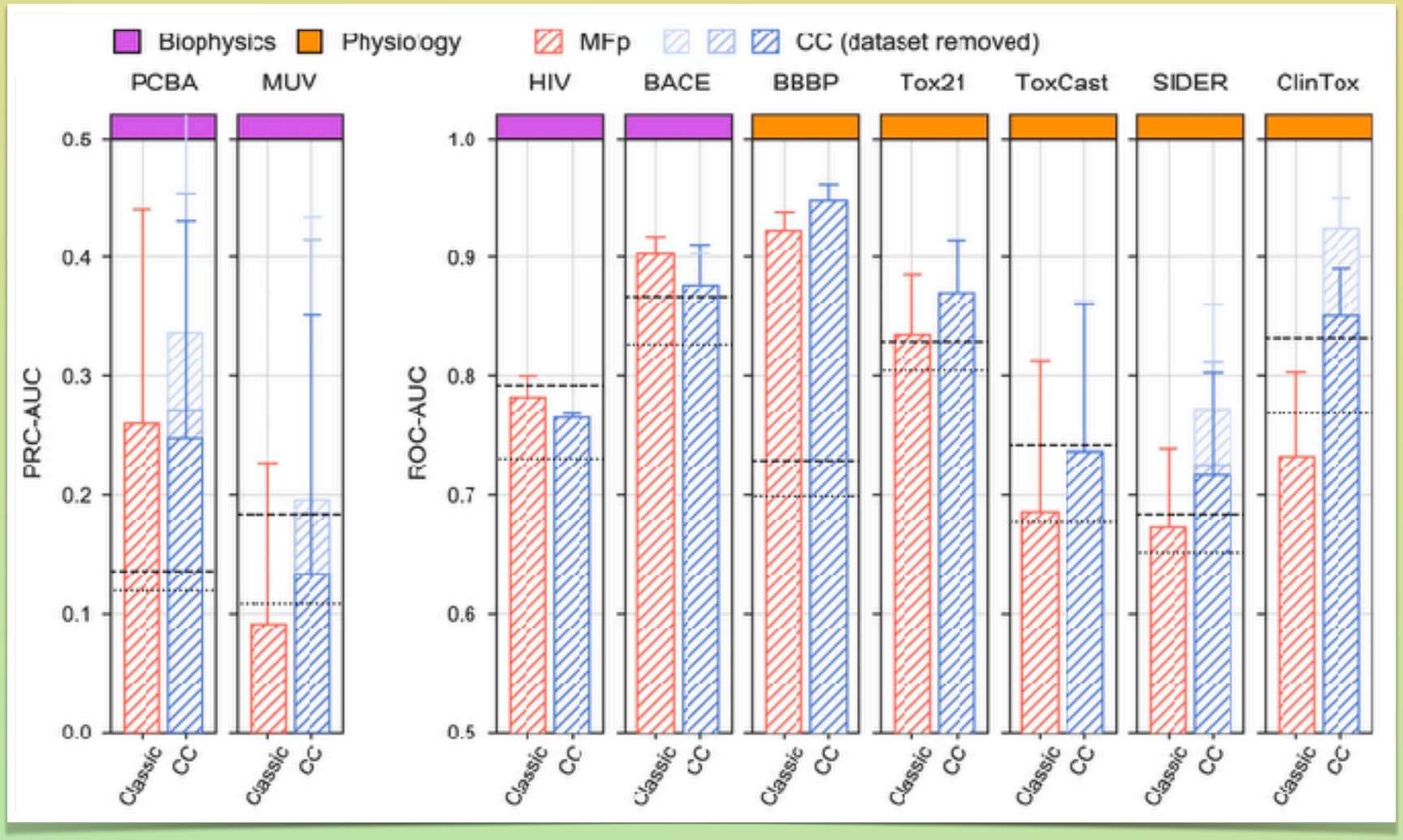
Confidence



Interpretability



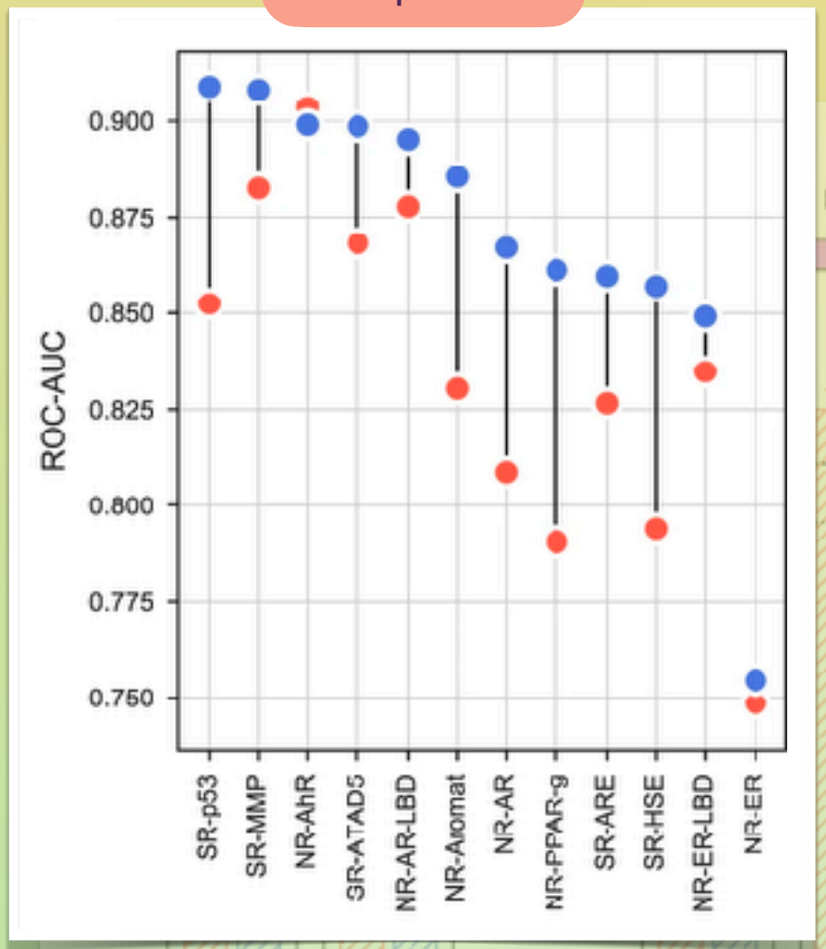
Automated AI/ML pipelines for pharmacology
State-of-the-art performance

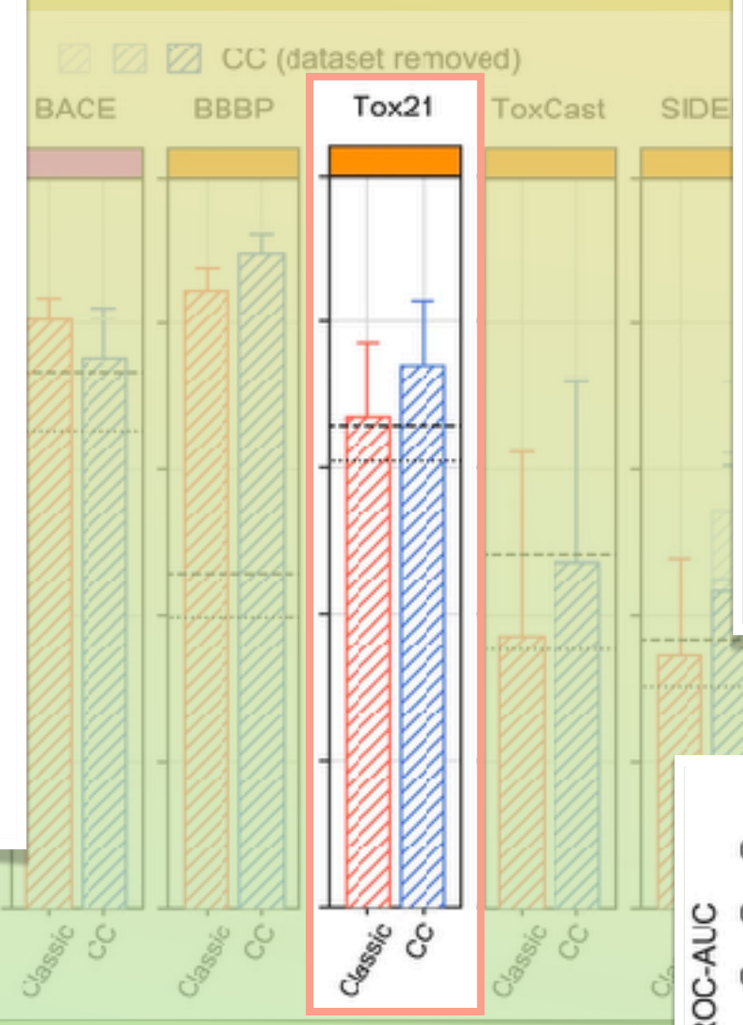


Bertoni*, Duran-Frigola* et al, Nat Commun, 2021
Ersilia's tool: ZairaChem

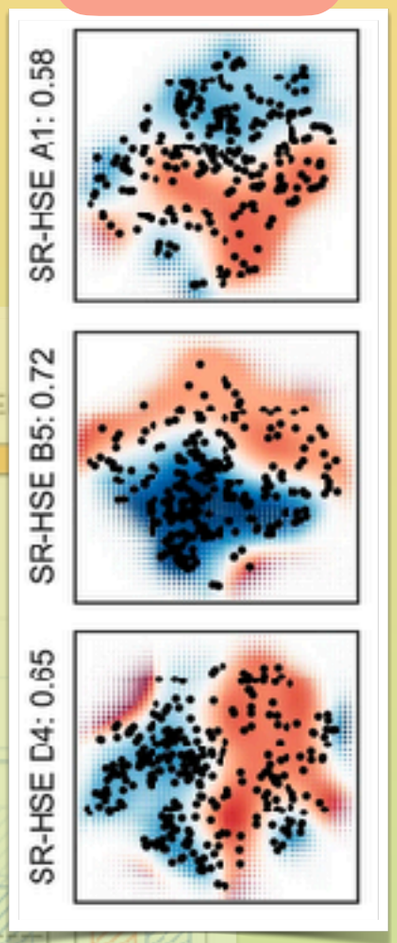
Performance of automated AI/ML

Multiple tasks

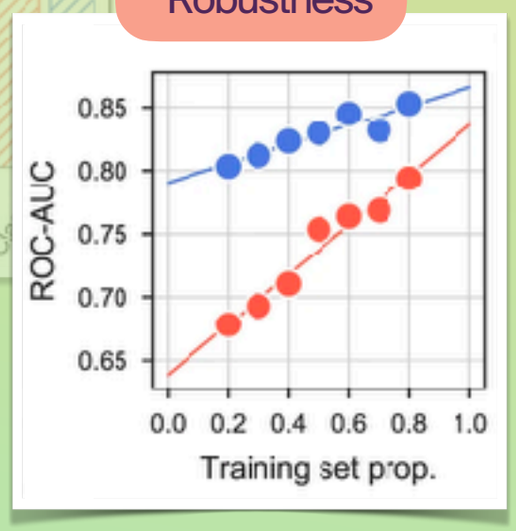




Visualization



Robustness



Bertoni*, Duran-Frigola* et al, Nat Commun, 2021
Ersilia's tool: ZairaChem

A generic H3D screening cascade

~13,000 molecules

Whole cell assays

P.falciparum IC50 (NF54/K1)
M.tuberculosis MIC90 (H37Rv)
ESKAPE pathogens IC50

Enzymology

Enzyme inhibition IC50 (PI4KB, PKG)

↓

Cytotoxicity

Human Cell Lines IC50 (CHO / HepG2)

↓

ADMET primary & secondary screens

Solubility
Microsomal clearance
Permeability

↓

ADMET tertiary screens

CYP P450 inhibition
hERG Blockade

~100 molecules

- A ML-based virtual screening cascade modelled after a generic experimental screening cascade
- Specific to the H3D chemical space where there is sufficient data
- Now deployed and available for use at H3D

Event Fund

Learning from few data points

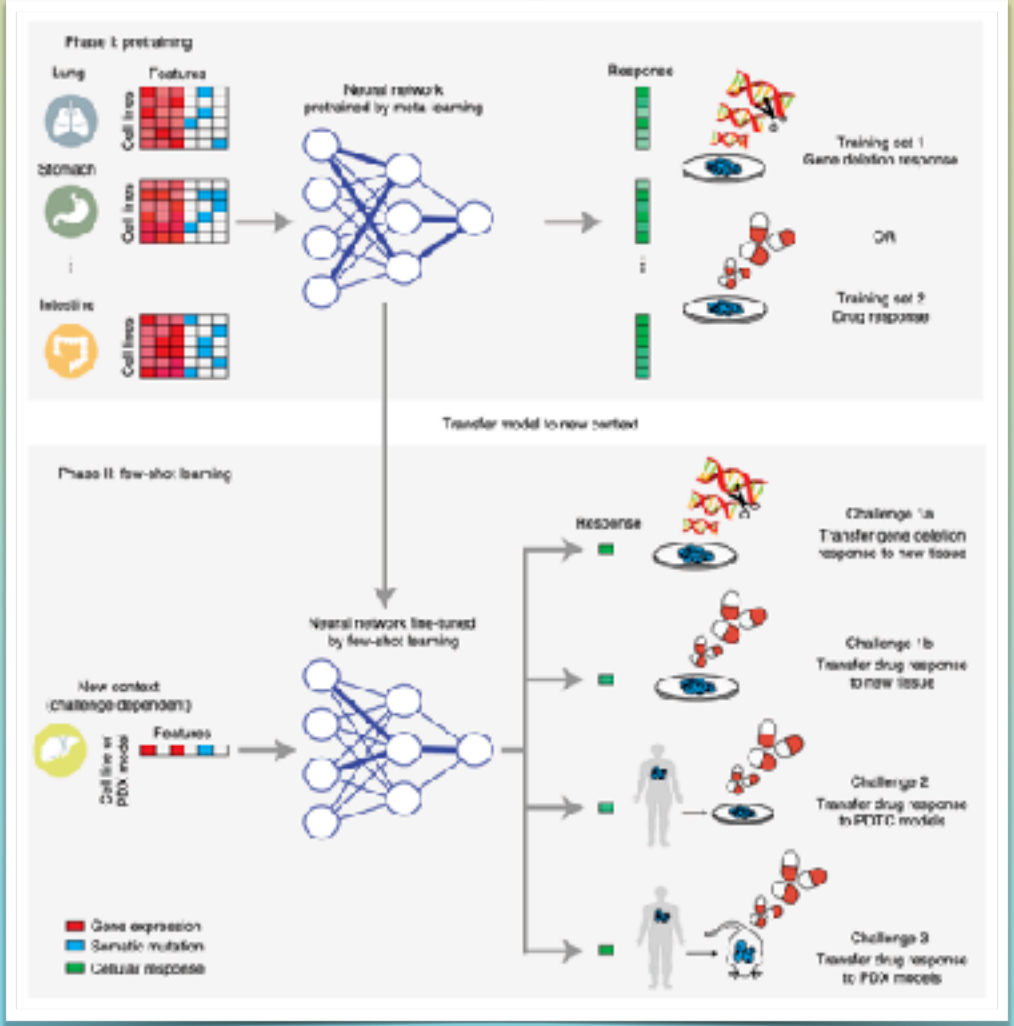
nature cancer

ANALYSIS

Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients

Jianzhu Ma, Samson H. Fong, Yunan Luo, Christopher J. Bakkenist, John Paul Shen, Soufiane Mourragui, Lodewyk F. A. Wessels, Marc Hafner, Roded Sharan, Jian Peng and Trey Ideker

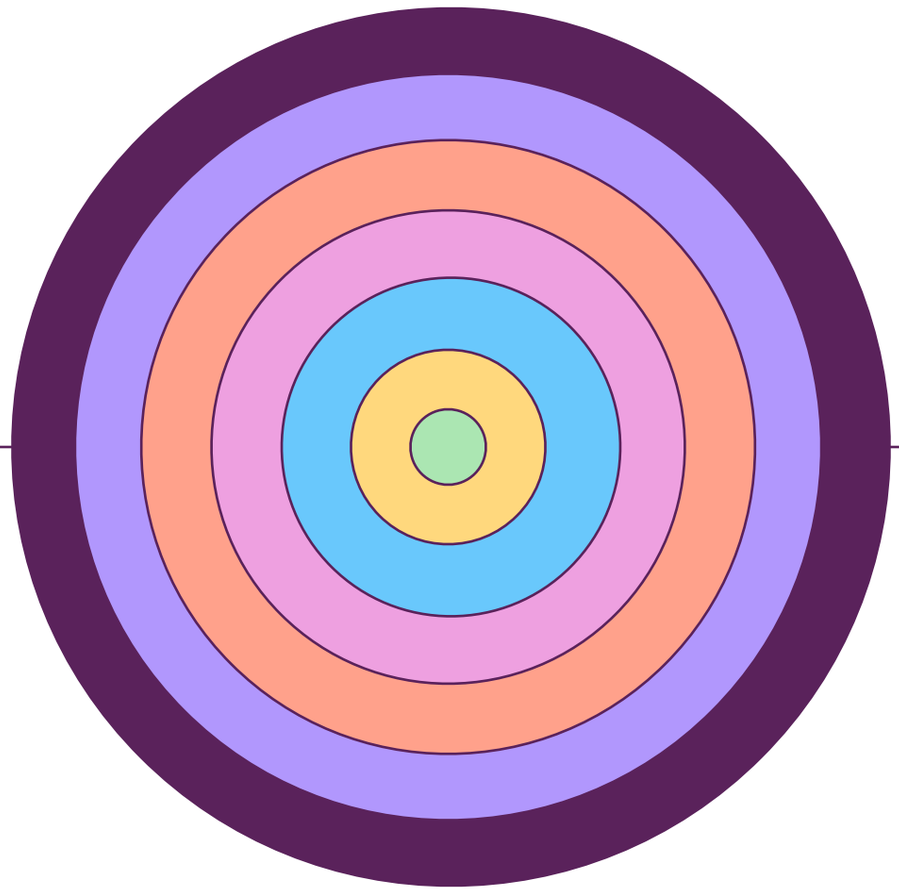
Cell-line screens create expansive datasets for learning predictive markers of drug response, but these models do not readily translate to the clinic with its diverse contexts and limited data. In the present study, we apply a recently developed technique, few-shot machine learning, to train a versatile neural network model in cell lines that can be tuned to new contexts using few additional samples. The model quickly adapts when switching among different tissue types and in moving from cell-line models to clinical contexts, including patient-derived tumor cells and patient-derived xenografts. It can also be interpreted to identify the molecular features most important to a drug response, highlighting critical roles for RB1 and SMAD4 in the response to CDK inhibition and RNF8 and CHD4 in the response to ATM inhibition. The few-shot learning framework provides a bridge from the many samples surveyed in high-throughput screens (n-of-many) to the distinctive contexts of individual patients (n-of-one).



Ma et al, Nature Cancer, 2021

Our goal: to provide ready-to-use AI/ML models

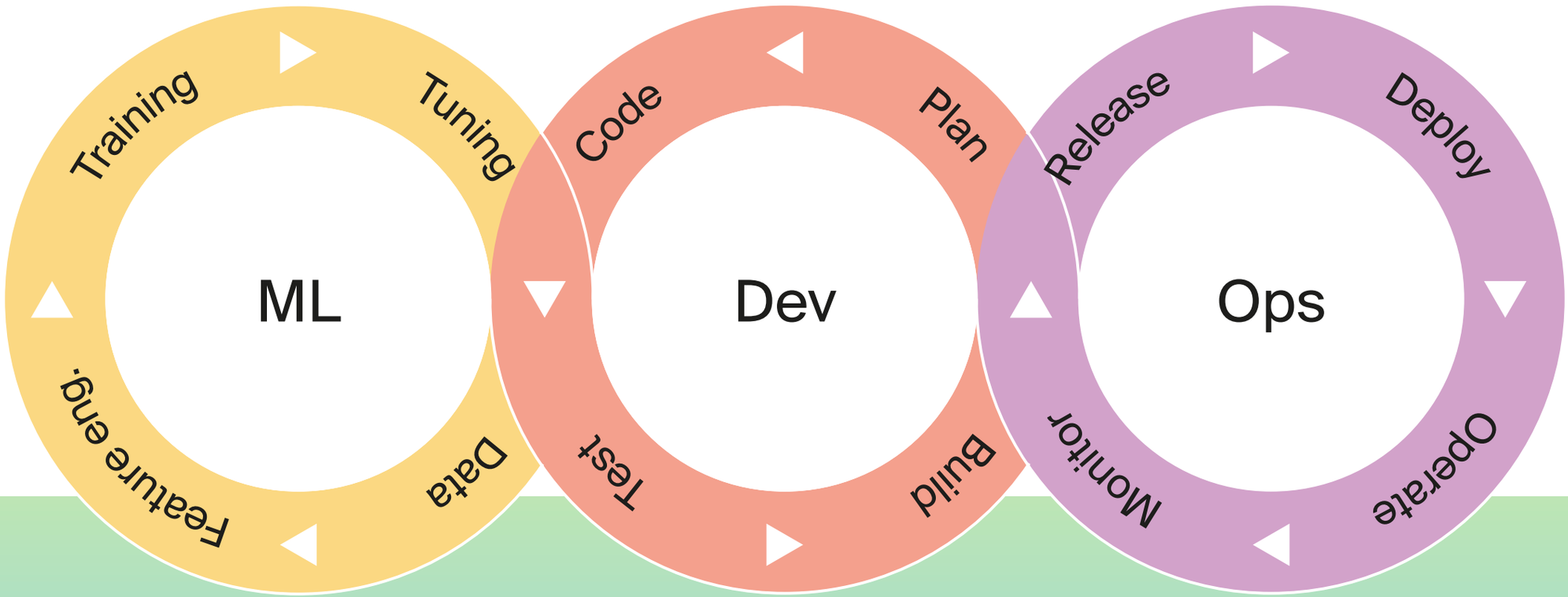
Input
e.g. a drug



Output
e.g. targets

- Raw data
- Processed data
- Model training
- Testing and tuning
- Validation
- External validation
- Deployment

Continuous delivery and automation pipelines



– Trainining mode: once or continuously updated
– Inference mode: black-box or interpretable
– Privacy: prevent reverse engineering
– Interoperability: easy distribution of AI/ML assets

Event Fund

AI/ML from the literature

Ersilia ‘bundles’ a good model developed by others


Cell

Article

A Deep Learning Approach to Antibiotic Discovery

SUMMARY

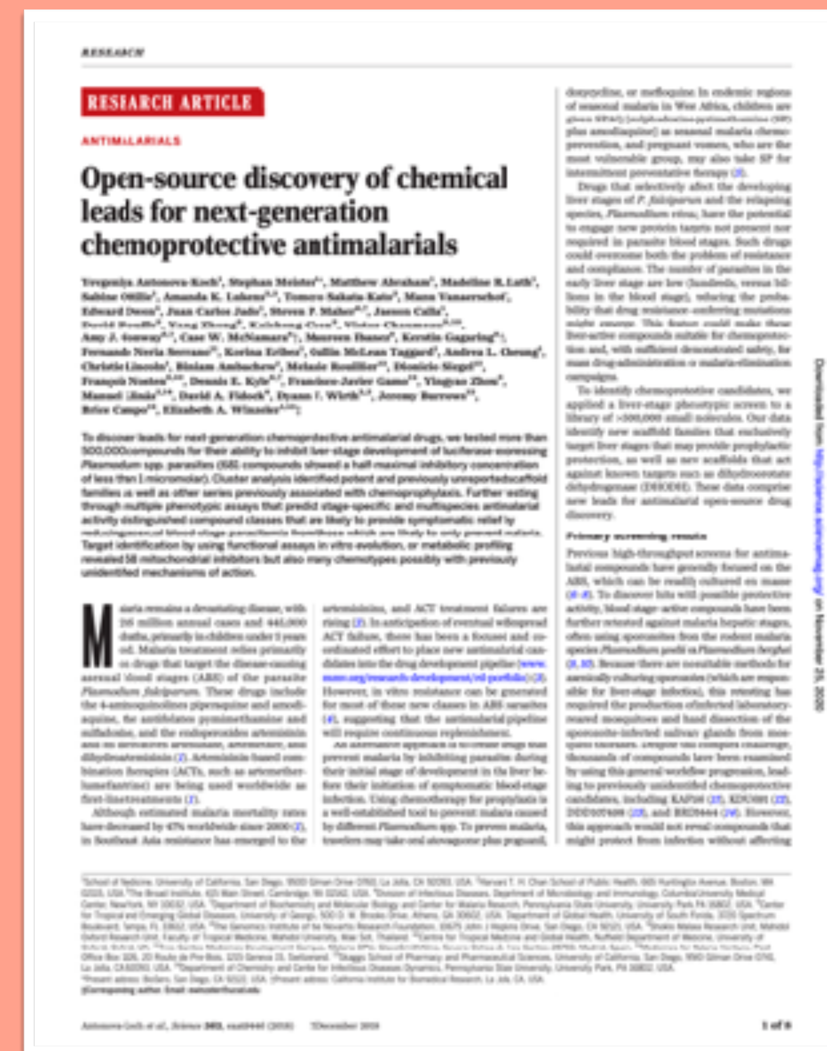
INTRODUCTION

Antibiotic activity *E.coli*

Stokes et al, 2020

In-house AI/ML

Ersilia ‘trains’ a model based on good data


RESEARCH

RESEARCH ARTICLE

ANTIMALARIALS

Open-source discovery of chemical leads for next-generation chemoprotective antimalarials

Chemoprotective antimalarials

Antonova-Koch et al, 2018

Event Fund

AI/ML from the literature

Ersilia ‘bundles’ a good model developed by others

Cell

Article

A Deep Learning Approach to Antibiotic Discovery

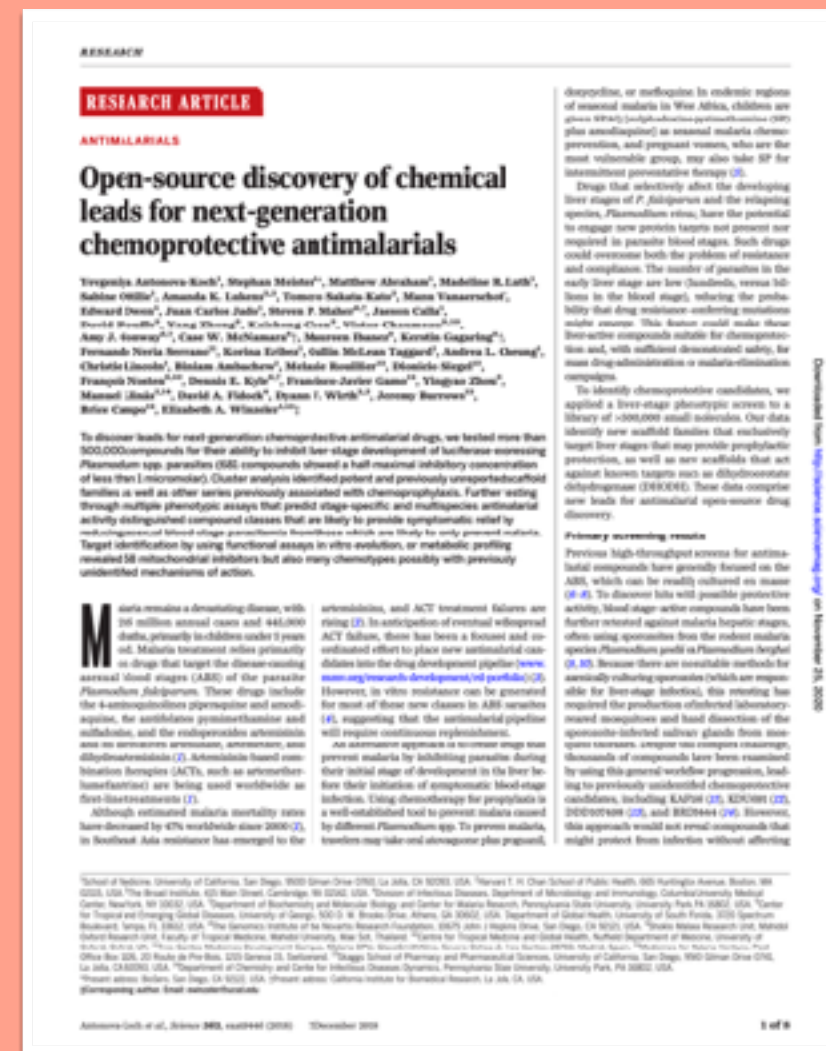
Antibiotic activity *E.coli*

Stokes et al, 2020

Halicin | Active!

In-house AI/ML

Ersilia ‘trains’ a model based on good data



RESEARCH

RESEARCH ARTICLE

ANTIMALARIALS

Open-source discovery of chemical leads for next-generation chemoprotective antimalarials

Chemoprotective antimalarials

Antonova-Koch et al, 2018

Event Fund

AI/ML from the literature

Ersilia 'bundles' a good model developed by others


Cell

Article

A Deep Learning Approach to Antibiotic Discovery

SUMMARY

INTRODUCTION

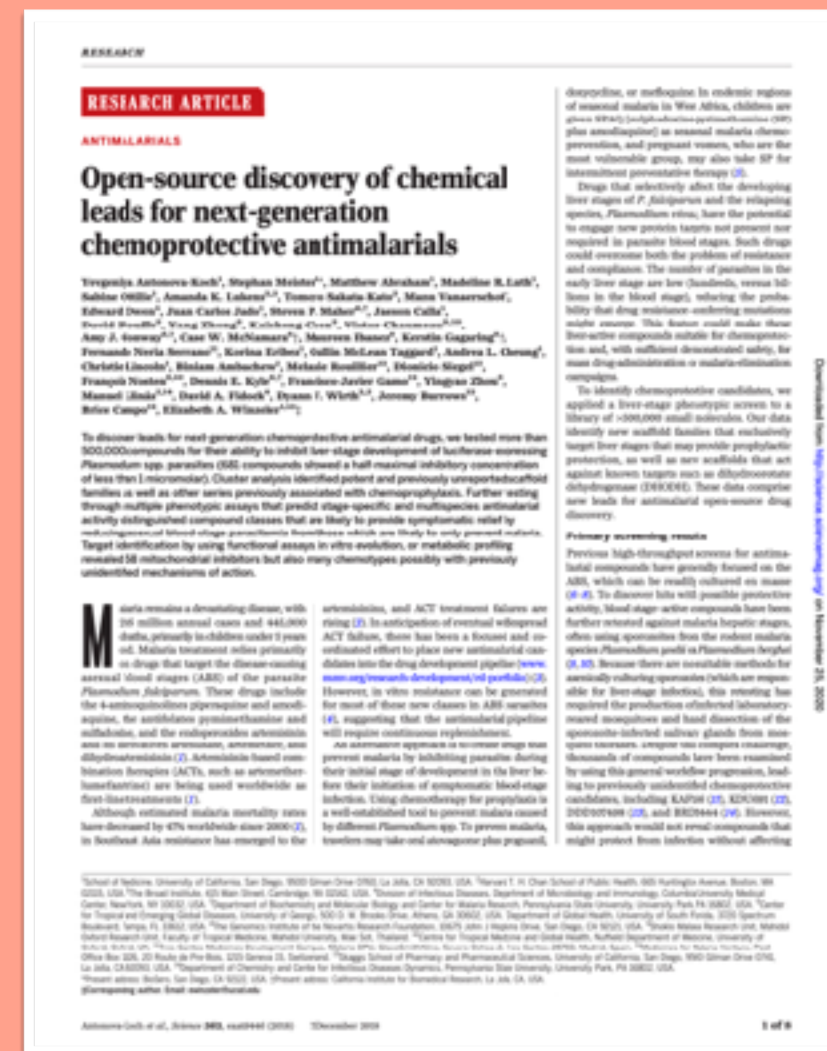
Antibiotic activity *E.coli*

Stokes et al, 2020

Halicin | Active!

In-house AI/ML

Ersilia 'trains' a model based on good data


RESEARCH

RESEARCH ARTICLE

ANTIMALARIALS

Open-source discovery of chemical leads for next-generation chemoprotective antimalarials

Chemoprotective antimalarials

Antonova-Koch et al, 2018

Atovaquone | Active!

AI/ML in collaboration

Ersilia ‘trains’ a model based on partner’s data

Natural product activity

University of Buea

AI/ML in collaboration

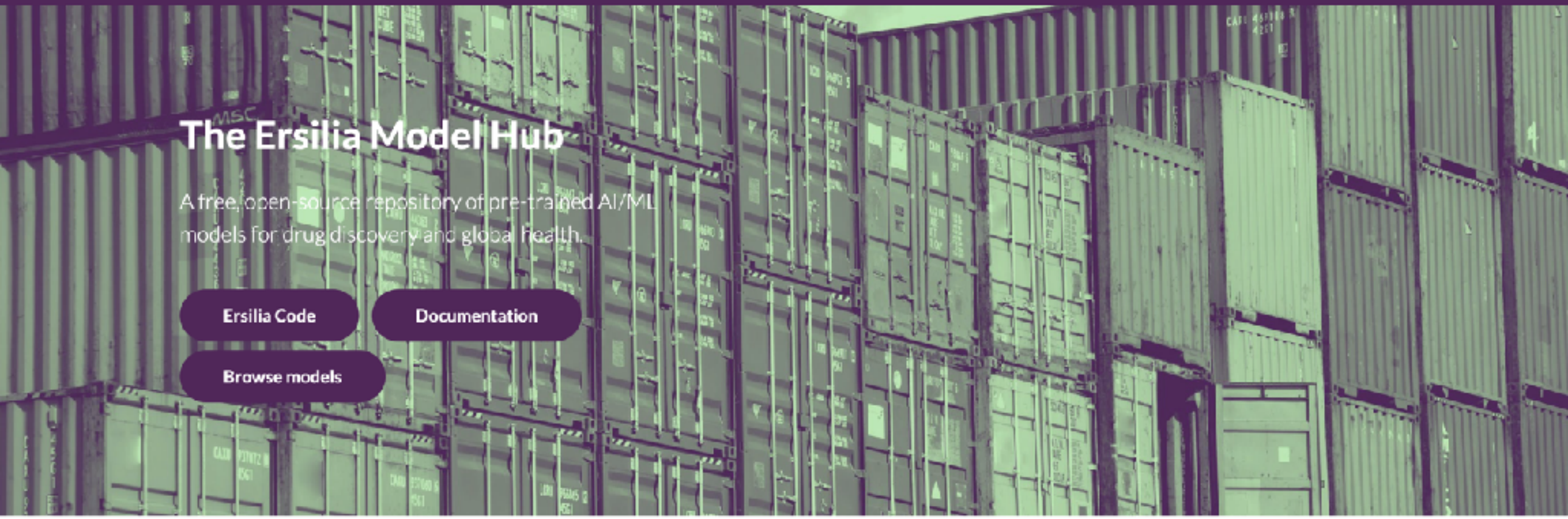
Ersilia ‘trains’ a model based on partner’s data

Natural product activity

University of Buea

Humimycin A

Active!



The Ersilia Model Hub

A free, open-source repository of pre-trained AI/ML models for drug discovery and global health.

Ersilia Code | Documentation | Browse models

A hundred models in the Ersilia Model Hub

Check the latest additions to the Ersilia Model Hub! We systematically look for AI/ML models and datasets in the scientific literature and incorporate them in our platform.

Classification of carcinogenic human metabolites

Predicts six properties using an ensemble machine learning method to predict carcinogenicity

Toxicity

SMILES transformer descriptor

Molecular fingerprint based on SMILES representations for low-data drug discovery

Vector

SARS-CoV inhibition

This model was developed to support the early efforts in the identification of novel drugs against SARS-CoV2. It pred...

Antiviral activity

More models

convert them into a readable .csv file.

bioassay-db

Code

chemical libraries as well as generative models

chem-sampler

Code

molecule bioactivity signatures.

chemical-checker

Code

End-to-end record linkage

A fully automated pipeline for fuzzy matching of medical datasets.

cidrz-e2e-linkage

Code

Helper tools for the ESTHER record linkage course

Helper methods to simplify record linkage in the context of the ESTHER course.

estherlinkage

Code

Griddify

Griddify high-dimensional tabular data for easy visualization and deep learning

griddify

Code

A data lake of pre-calculated predictions based on the Ersilia Model Hub

Isaura is a data lake of pre-calculated descriptors based on the Ersilia Model Hub

isaura

Code

On-Disk X and Y matrices

Manage X and Y matrices on disk for scalable machine learning

ondisk-xy

Code

Chemical space visualization

Visualization of the chemical space based on TMAP trees

tmap-chemistry

Code

convert them to a readable .csv file.

bioassay-db

Code

chemical libraries as well as generative models

chem-sampler

Code

molecule bioactivity signatures.

chemical-checker

Code

End-to-end record linkage

A fully automated pipeline for fuzzy matching of medical datasets.

cidrz-e2e-linkage

Code

Helper tools for the ESTHER record linkage course

Helper methods to simplify record linkage in the context of the ESTHER course.

estherlinkage

Code

Griddify

Griddify high-dimensional tabular data for easy visualization and deep learning

griddify

Code

A data lake of pre-calculated predictions based on the Ersilia Model Hub

Isaura is a data lake of pre-calculated latent descriptors based on the Ersilia Model Hub

isaura

Code

On-disk X and Y matrices

Binary X and Y matrices on disk for scalable machine learning

ondisk-xy

Code

Chemical space visualization

Visualization of the chemical space based on TMAP trees

tmap-chemistry

Code

Welcome to the Ersilia Model Hub!

Broad spectrum antimicrobial activity

Type to search model...

Tags

- Tox21
- Toxicity
- MoleculeNet
- Grover
- Graph Transformer

Output

- Antibiotic activity
- Toxicity
- Synthetic accessibility
- Antiviral activity
- Target

Mode

- Pretrained
- Retrained
- In-house
- Online

License

Carcinogenic potential of metabolites and small molecules

eos1579 metabokiller

Carcinogenicity is a result of several potential effects on cells. This model predicts the carcinogenic potential of a small molecule based on their potential to induce cellular proliferation, genomic instability, oxidative stress, anti-apoptotic responses and epigenetic alterations.
Metabokiller uses the Chemical Checker signaturizer to featurize the molecules, and the Lime package to provide interpretable results.
Using Metabokiller, the authors screened a panel of human metabolites and experimentally demonstrated two of the predicted carcinogenic metabolites induced carcinogenic transformations in yeast and human cells.

Molecular maps based on broadly learned knowledge-based representations

eos6m4j bidd-molmap

Descriptor-based or fingerprint-based molecular maps (images) are created. Typically, the goal is to use these images as inputs for an image-based deep learning model such as a convolutional neural network

SMILES transformer descriptor

eos2lm8 smiles-transformer

Molecular fingerprint based on natural language processing. It converts SMILES into fingerprints using an unsupervised model pre-trained on a very large SMILES dataset. The transformer is particularly well-suited for low-data drug discovery

Similarity search in the ZINC database

eos54c7 | zinc-similarity

Look for 100 nearest neighbors, according to ECFP4 Tanimoto similarity, in the ZINC database.
This model posts queries to an external online server

Similarity search in ChEMBL

eos2a9n | chembl-similarity

Given a molecule, this model looks for its 100 nearest neighbors in the ChEMBL database, according to ECFP4 Tanimoto similarity.
This model is redirected to ChEMBL web server directly, so molecules are submitted to an external server.

See more

Can't find what you are looking for?

If your are looking for a specific model, please reach out to us. We will be happy to help!

Contact us

Ersilia

About | Request model | Backlog | Volunteer

Take-home messages

— Supervised AI/ML for activity prediction
— Importance of chemical descriptors
— Data-driven descriptors with the Chemical Checker
— Challenges of supervised AI/ML methods
— Pre-trained models in the Ersilia Model Hub
— Use Slack!
— Add terms to the Glossary!
— Learn more in the hands-on sessions!

Bringing data science and AI/ML tools to infectious disease research

Session 2: Introduction to Supervised Machine Learning

Event Sponsors