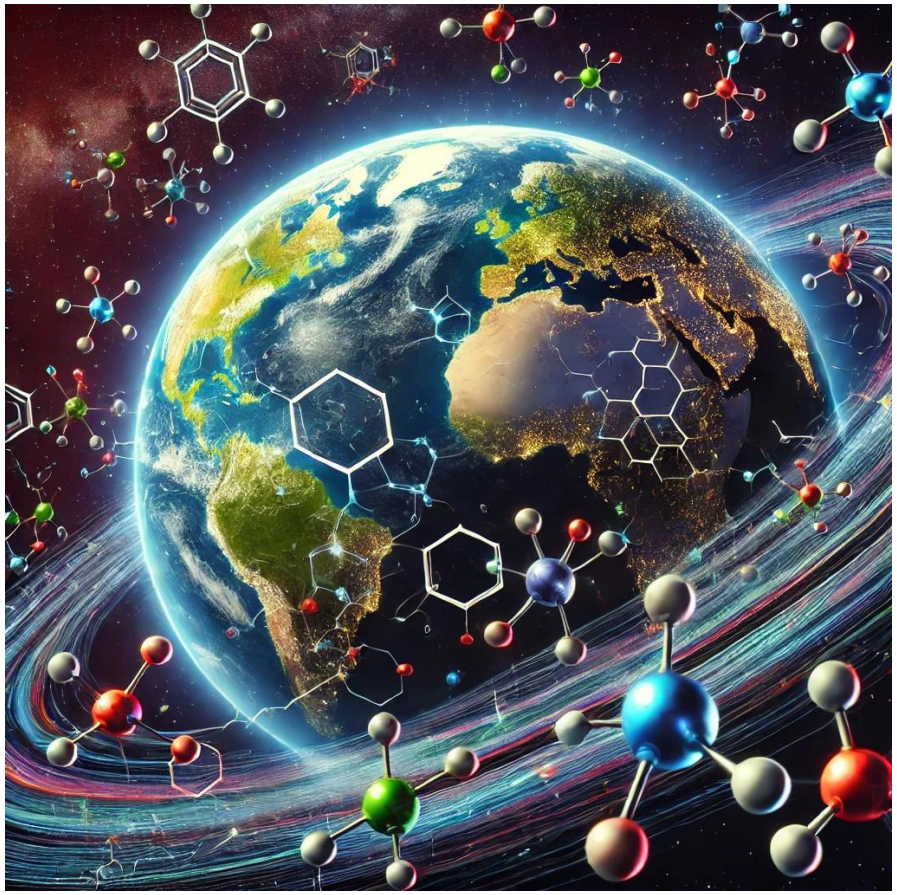
Chemography concept in chemical space exploration

Alexandre Varnek
University of Strasbourg

8ADD workshop Olomouc, 27th January 2025

Chemical Space - definition



What is « Chemical Space »

- ensemble of molecules ?
- mathematical object ?

Chemical Space - definition

Chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars

C. Lipinski and A. Hopkins, 2004

A set of molecules forms a chemical space, for which the relationships between the objects (graphs of descriptor vectors) are established

A.Varnek & I. Baskin, 2011

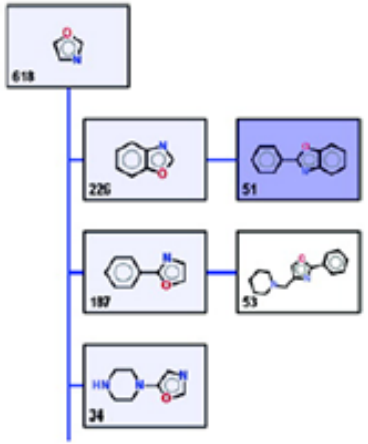
One can formulate chemical space as a mathematical and usually high-dimensional space where distances represent similarities between molecules which can be represented in the form of chemical space maps by applying various dimensionality reduction methods

J. L Reymond, 2025

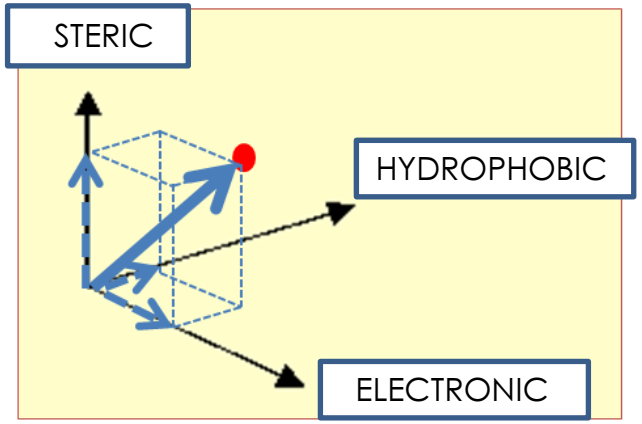
Chemical Space representations

SPACE = objects + relations between them

Molecular graphs



Vectors (descriptors, learned representations)



A. Varnek and I. Baskin *Mol. Informatics,* 2011, **30**, 20 – 32

Chemical Space representations

Graph-based chemical space

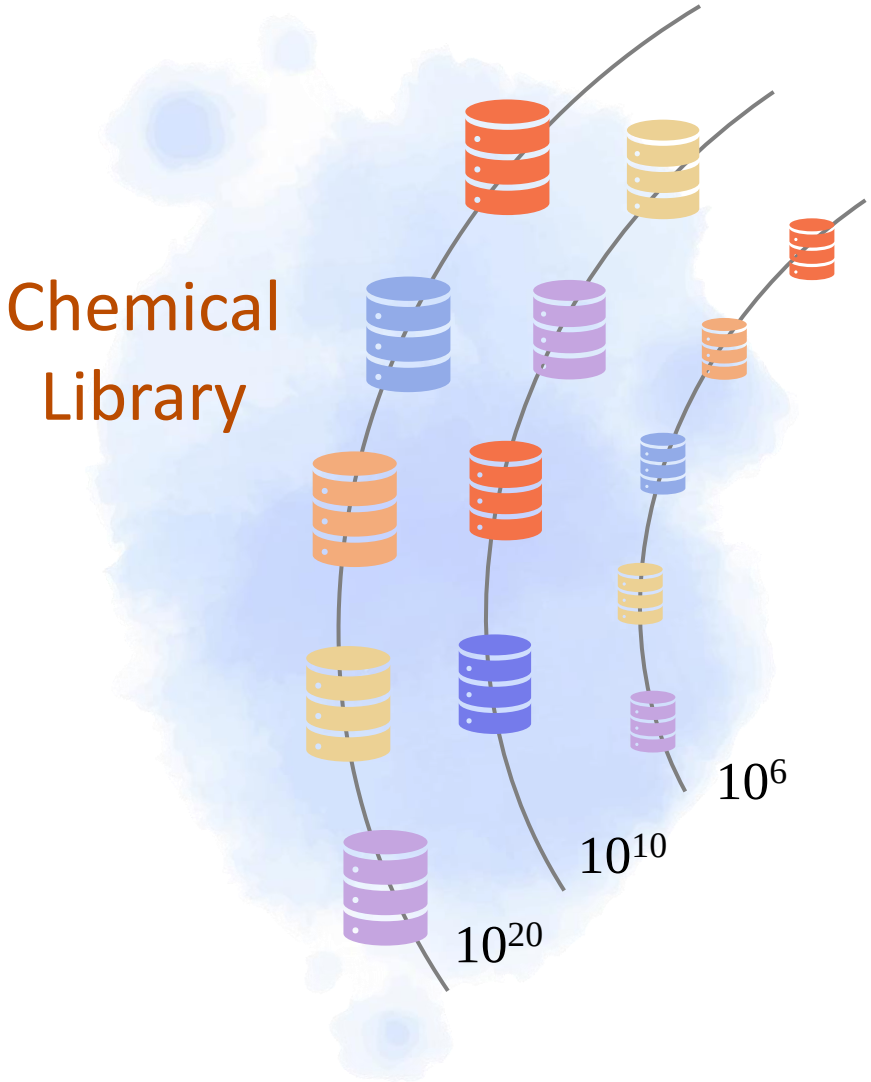
- Scaffolds analysis
- Molecular Matched Pairs
- Activity cliffs
- Chemical space networks (e.g., MMP-based)

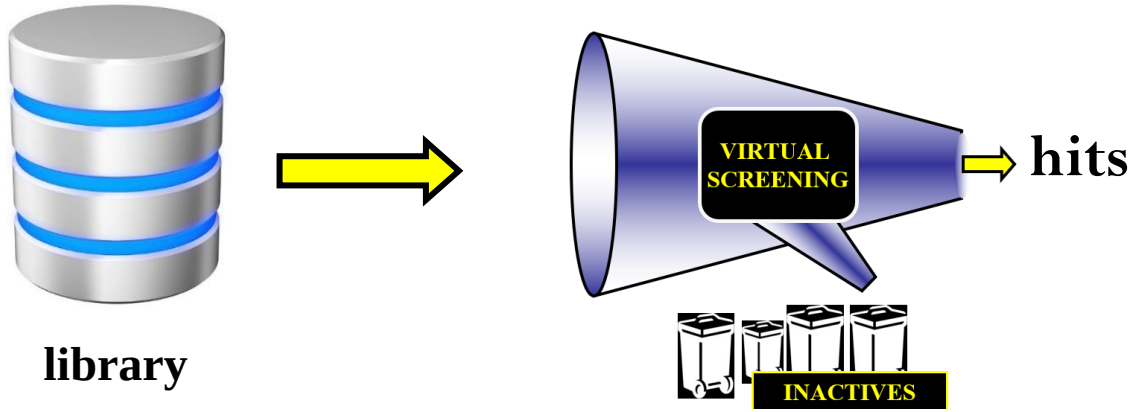
Vector-based chemical space

- Data visualization and analysis in descriptors space
- Chemical space networks (e.g., similarity-based)

In silico drug design: Big Data problem

★ 10^{33} *(P. Polishchuk et al, 2013)*

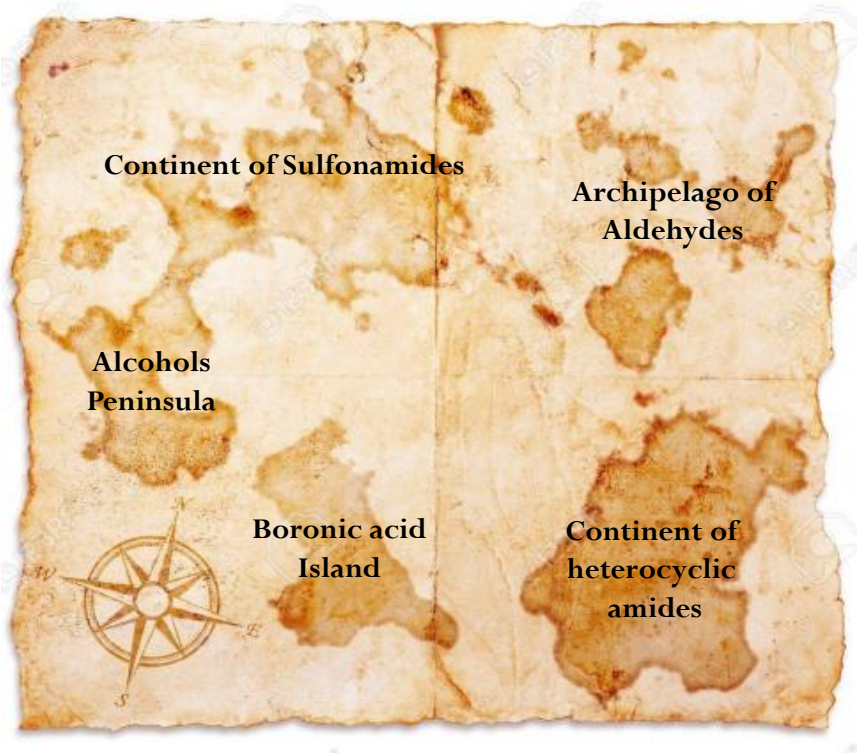




***Assumption*:** larger library increases a chance to retrieve hits in virtual screening

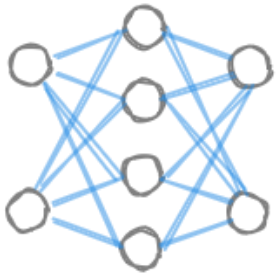
NB: both library size and chemical diversity must be considered

Chemography: efficient solution for Big Data handling

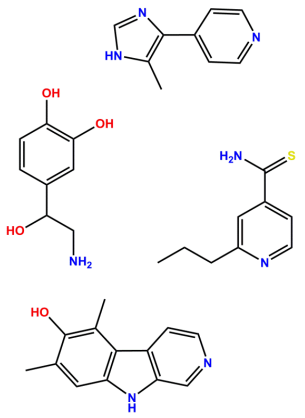


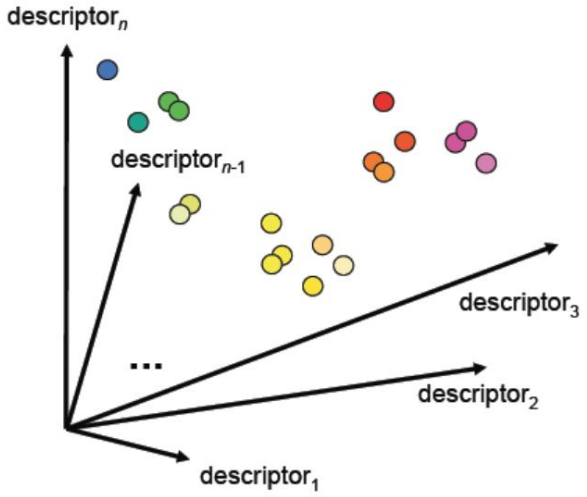
Preparation of "made-on-demand" library

Design of novel molecular entities



Data visualization: dimensionality reduction problem

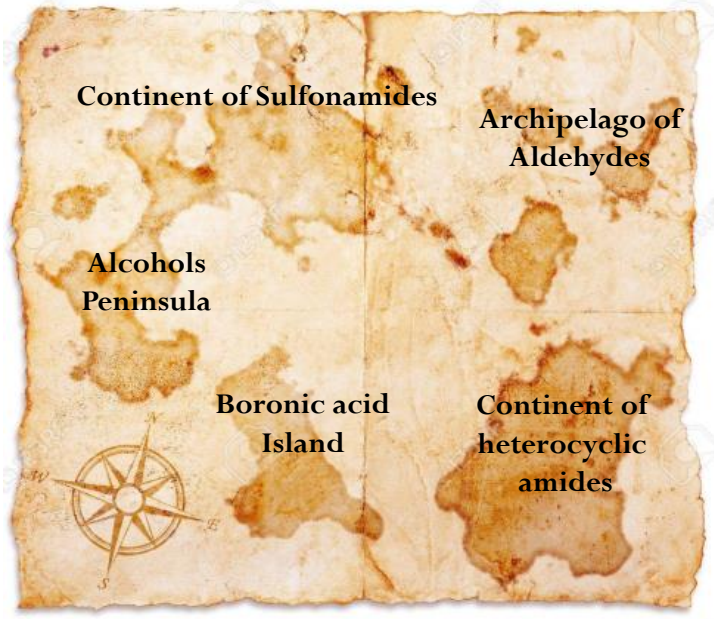




Initial chemical space

(N-dimensional)

Dimensionality reduction

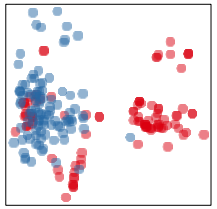


Latent chemical space

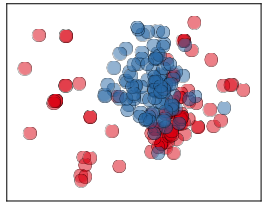
(2-dimensional)

Dimensionality reduction methods

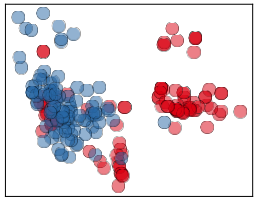
Acetylcholinesterase dataset (DUD) : 100 actives and 100 inactives



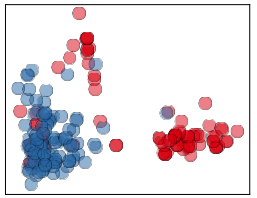
Multi-Dimensional Scaling



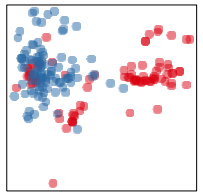
Canonical Correlation Analysis



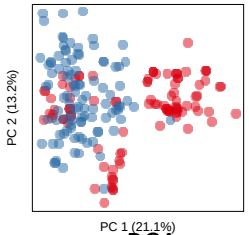
Independent Component Analysis



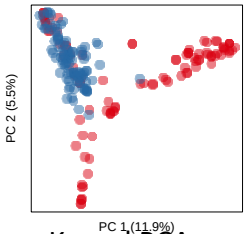
Exploratory Factor Analysis



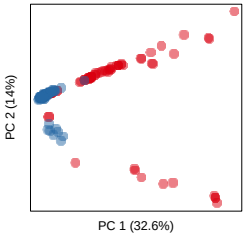
Sammon map



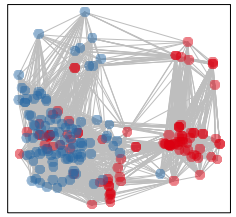
PCA



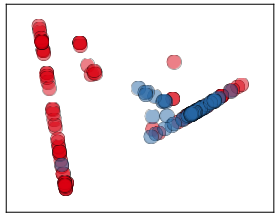
Kernel PCA (RBF kernel)



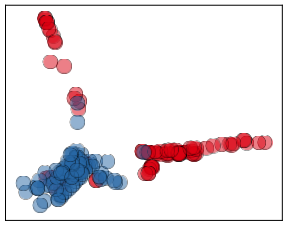
Kernel PCA (polynomial kernel)



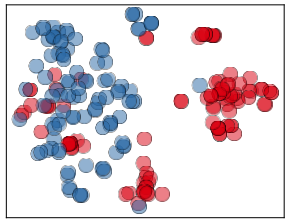
Isomap



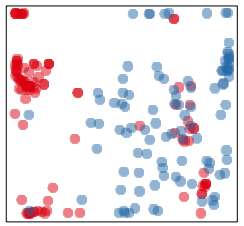
Locally Linear Embedding



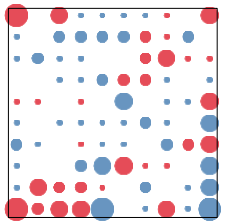
Laplacian Eigenmaps



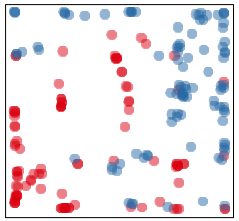
t-SNE



Autoencoder dimensionality reduction



SOM



GTM

Generative Topographic Mapping : areas of application

Conformational space analysis

Data visualisation and analysis

Ligand to Protein docking

Library comparison

Sequence space analysis

Structure-Activity modeling

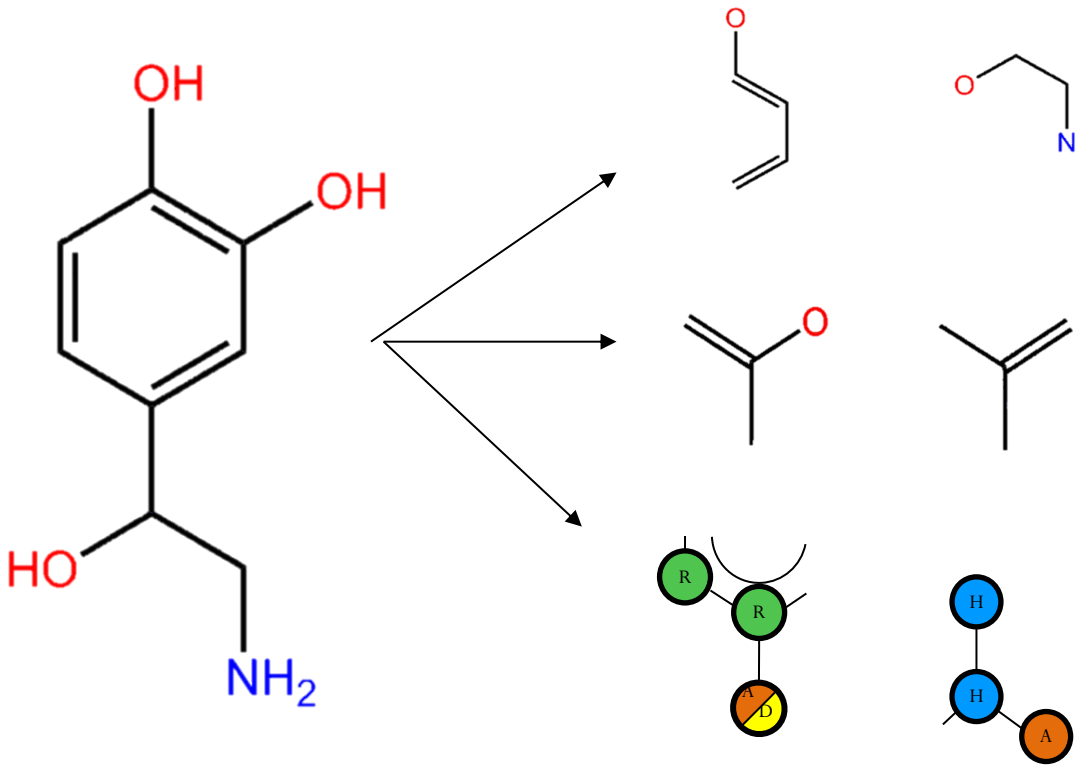
Drugs repurposing

Virtual screening

Library design

de novo design

Tunable ISIDA fragment descriptors



Journal of Computer-Aided Molecular Design (2005) 19: 693–703
DOI 10.1007/s10822-005-9008-0

Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures

A. Varnek[a,*], D. Fourches[a], F. Hoonakker[a] & V. P. Solov'ev[b]

Current Computer-Aided Drug Design, 2008, *4*, 000-000 1

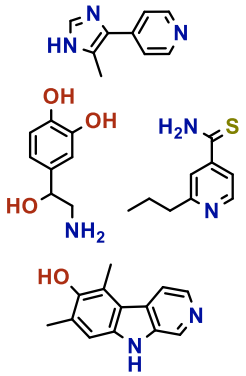
ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors

Alexandre Varnek[*,1], Denis Fourches[1], Dragos Horvath[1], Olga Klimchuk[1], Cedric Gaudin[1,2], Philippe Vayer[2], Vitaly Solov'ev[1,3], Frank Hoonakker[1], Igor V. Tetko[4] and Gilles Marcou[1]

Generative Topographic Mapping (GTM)

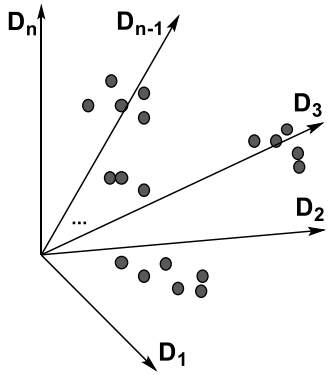
Chemical library

1



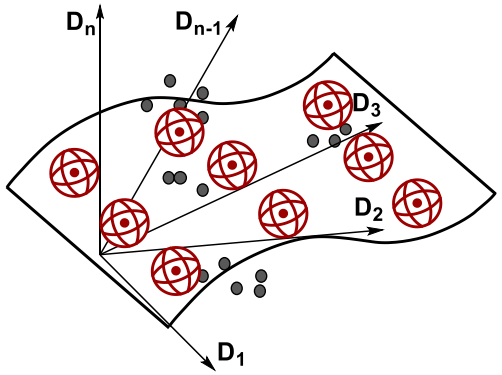
Representative frame set

2



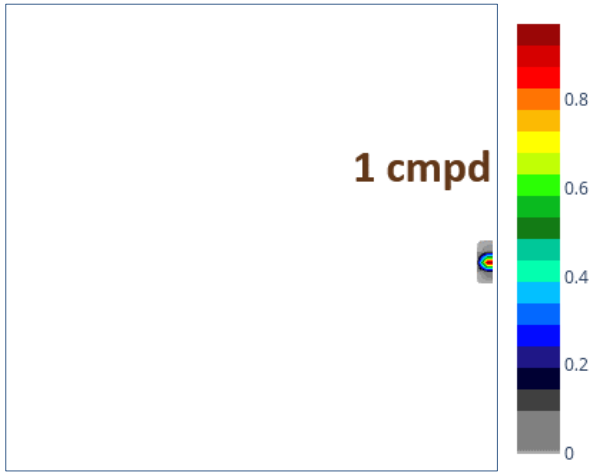
Descriptor Space

3

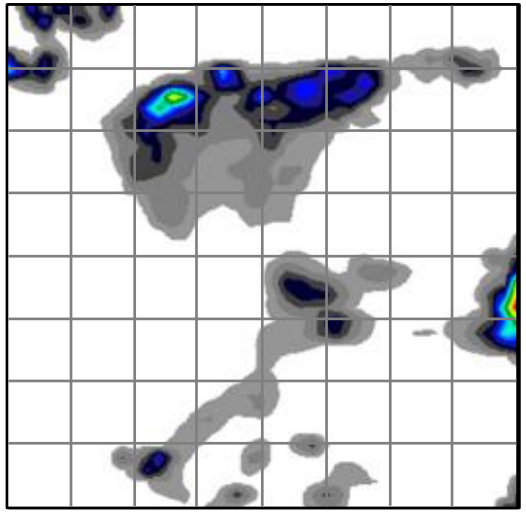


Manifold training

4



4



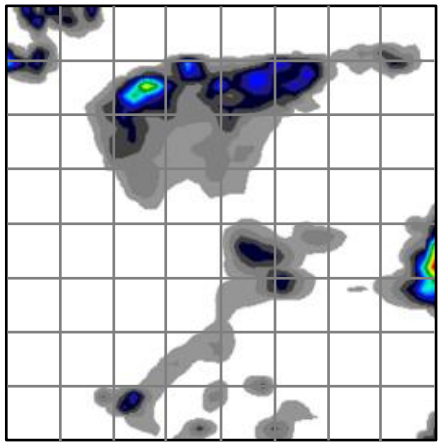
5

d_1	d_2	d_3	d_4	d_5

Data density values in the grid nodes

GTM Landscapes

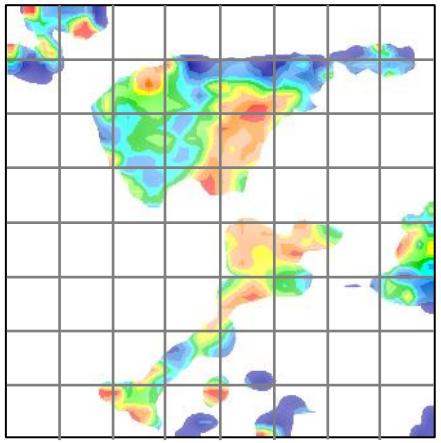
Density landscape



0 20 40 60 80

d_1	d_2	d_3	d_4	d_5

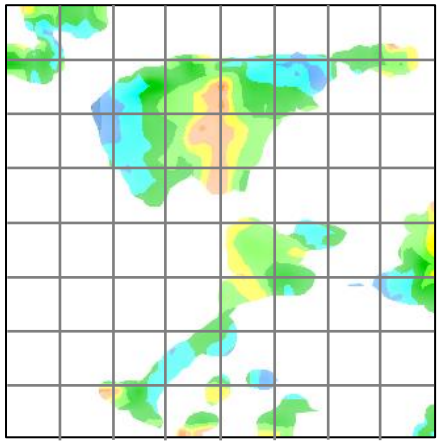
Class landscape



Inactive Active

c_1	c_2	c_3	c_4	c_5

Activity landscape

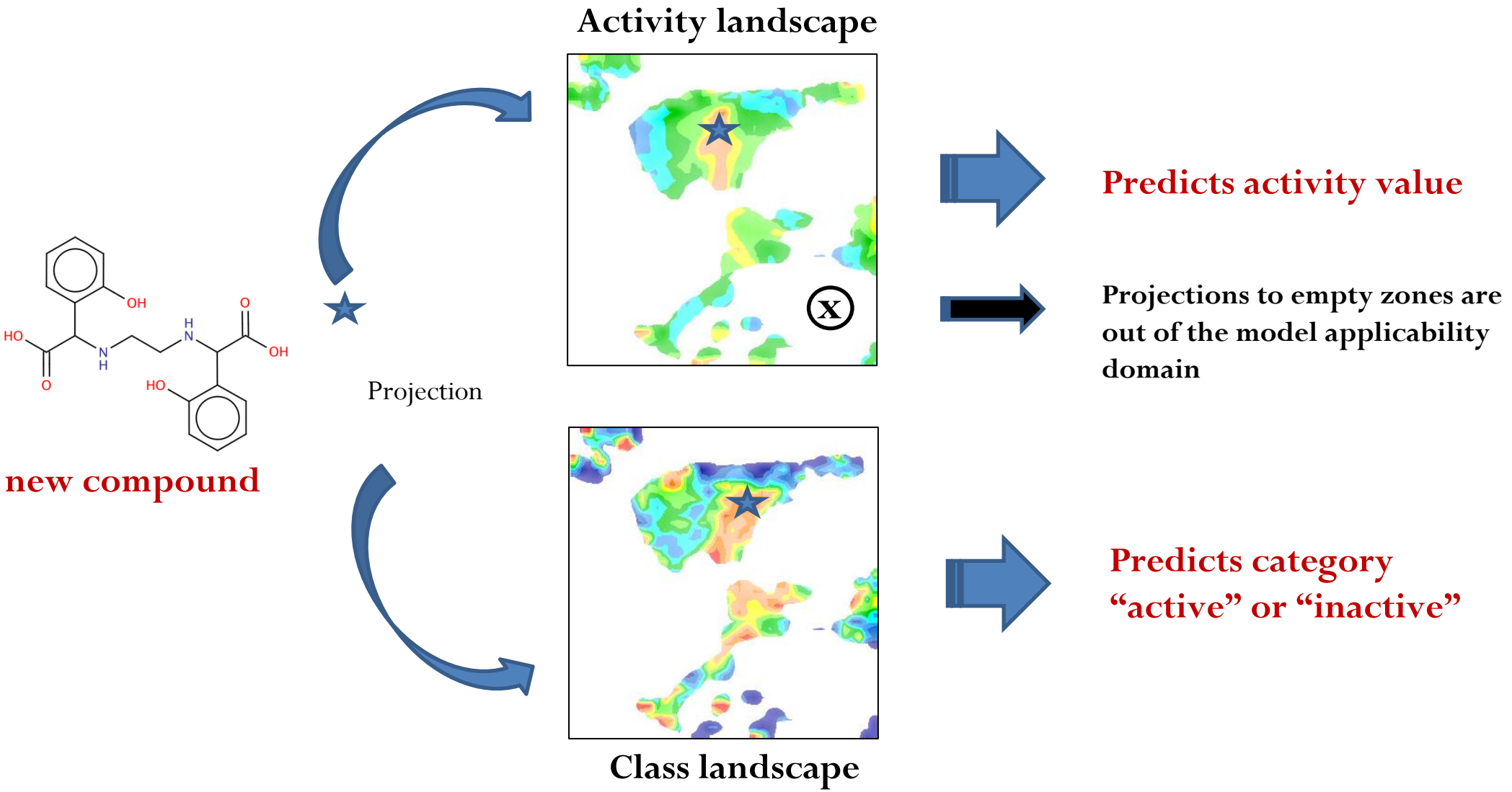


200 300 400 500 600 molecular weight

p_1	p_2	p_3	p_4	p_5

Each landscape can be encoded by a set of special GTM-descriptors characterizing either *structures* only or *structures & activity* distributions

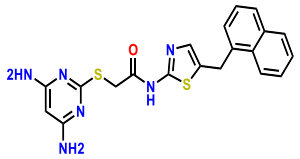
GTM Landscapes as predictive models



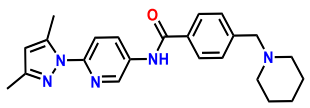
H. A. Gaspar *et al* *Mol. Informatics*, 2015, **34** (6-7), 348-356

GTM: Chemical space analysis

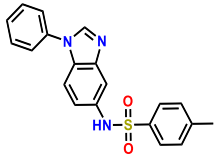
A9. 2-Aminothiazoles



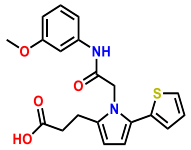
A2. Azoles

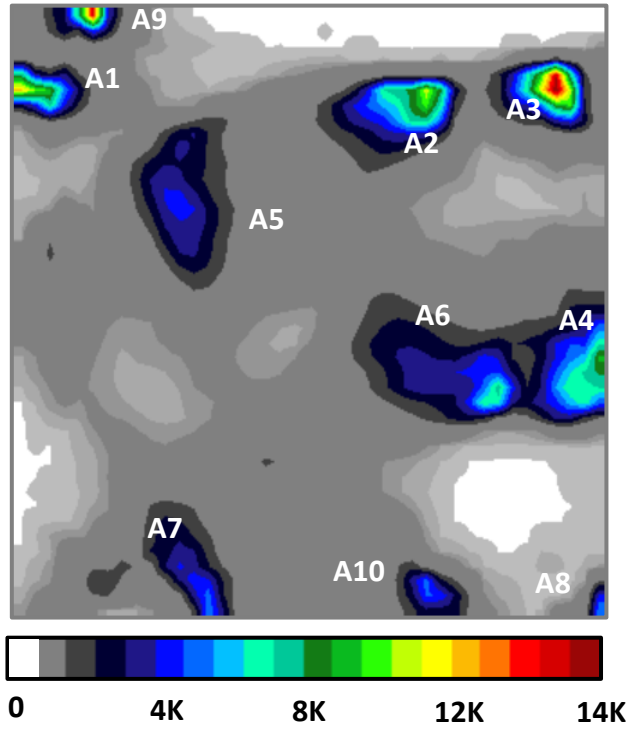


A3. Sulfonanilides

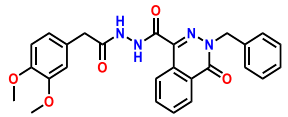


A1. Thiophenes

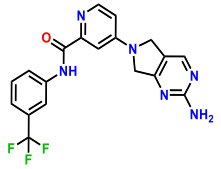




A10 Pyridazinone/oxadiazolone containing amides



A4. Halogenated N-heterocycles

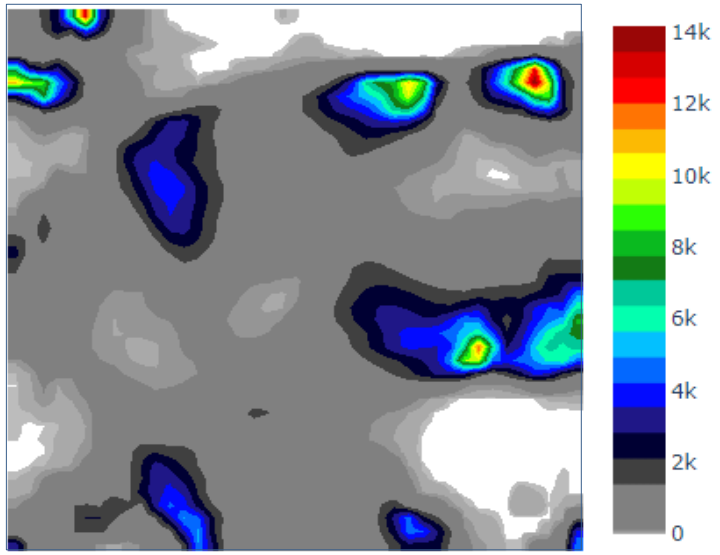


Each zone of density landscape can be associated with some chemotypes

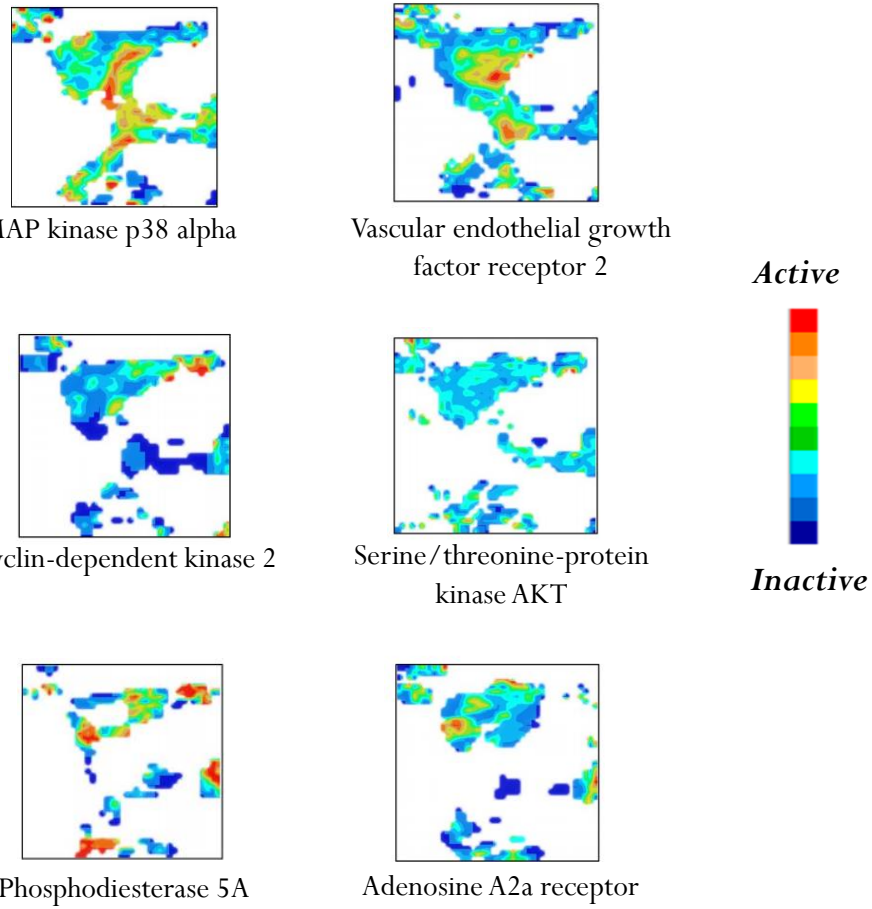
Density landscape of ChEMBL

(1.8 M cpnds , ISIDA descriptors)

GTM: density and activity landscapes for ChEMBL



Density landscape of ChEMBL
(ISIDA descriptors were used)



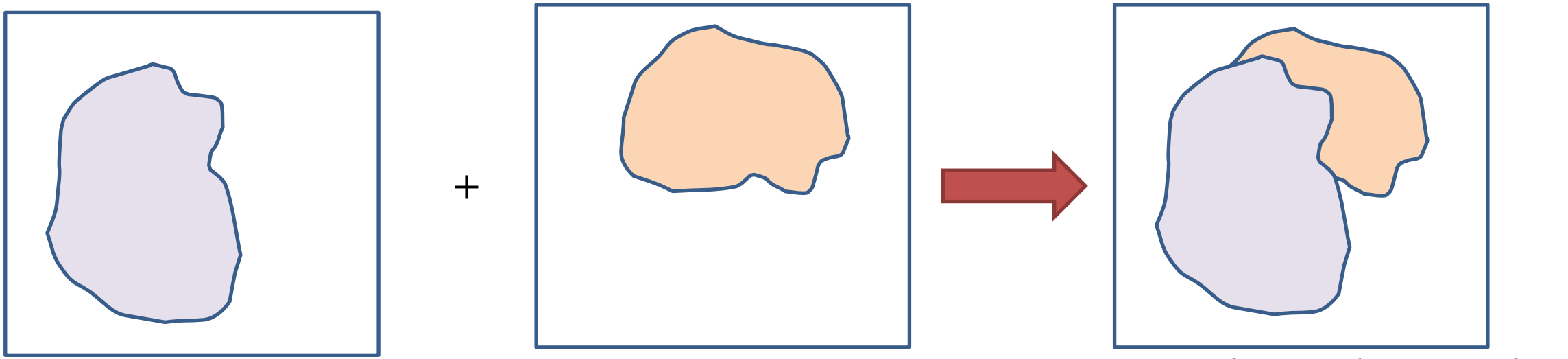
For any new compound projected on the map, one can:

- identify the most similar ChEMBL compounds,
- predict its pharmacological profile for > 700 biological activities

Casciuc et al. J. Chem. Inf. Model., 2019, **59**(1), 564-572

Pairwise libraries comparison

Task : identification of structural motifs unique for a given library



Library 1

Library 2

Extraction of unique chemotypes from the non-overlapping areas

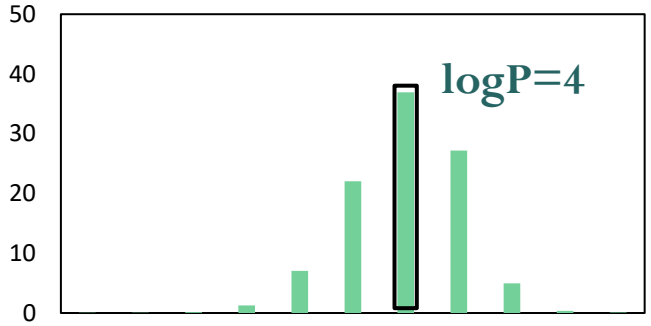
Data analysis: histograms vs chemography

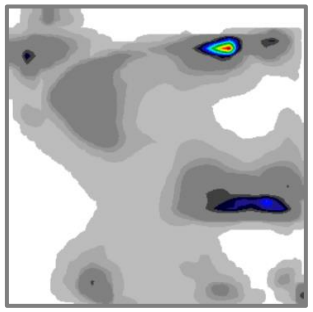
logP distribution

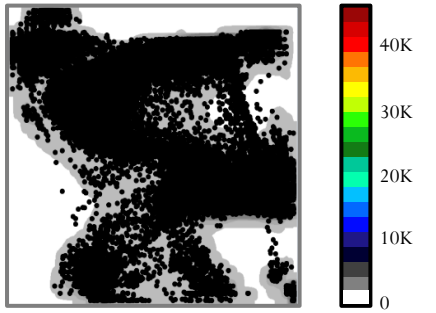
Density landscape

Projected molecules with logP=4±0.5

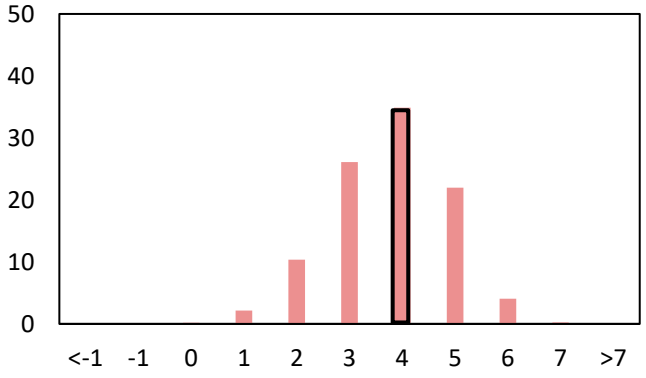
Library 1

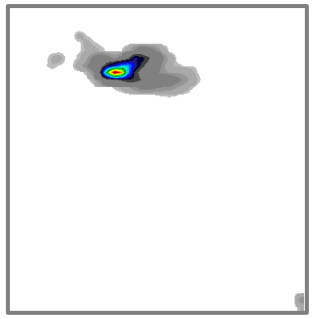


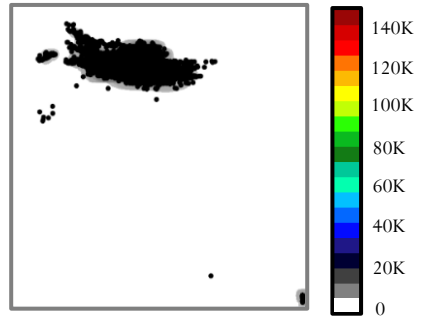




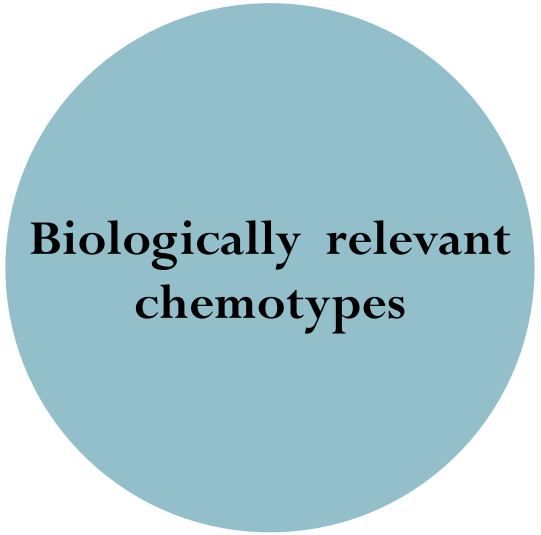
Library 2







Case study 1 : Commercial vs Biologically relevant data



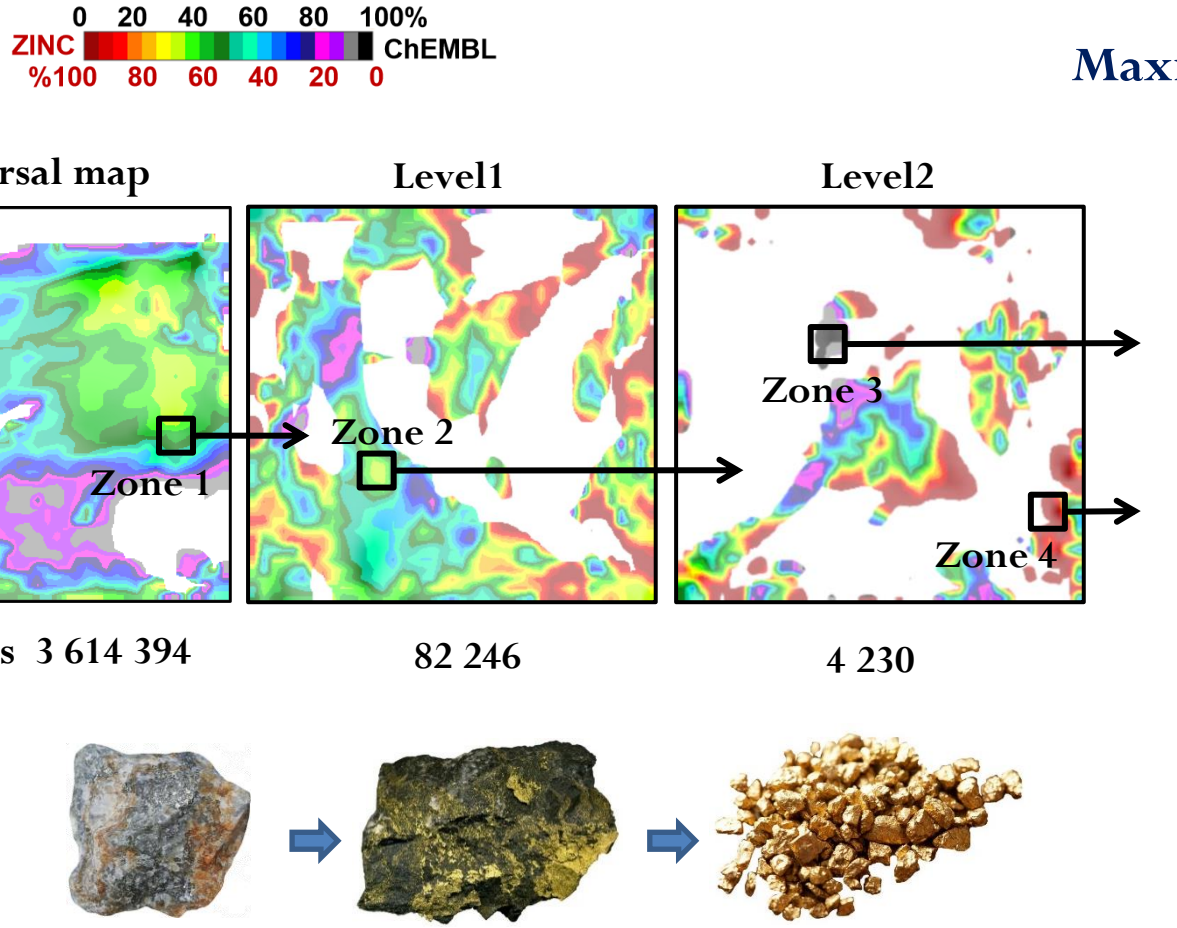
>1.3 billion cmpds

ZINC DATABASE

>1.8 M cmpds

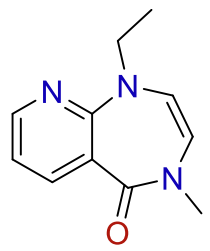
ChEMBL

Hierarchical GTM navigation of the chemical space

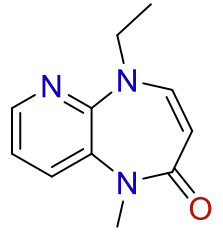


Maximum Common Substructures (MCS)

ChEMBL-specific MCS

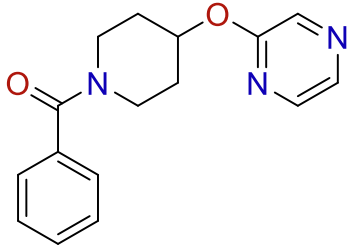


Popularity - 18

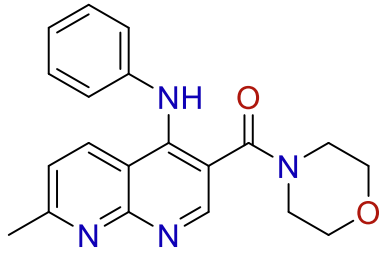


Popularity - 16

ZINC-specific MCS



Popularity - 21



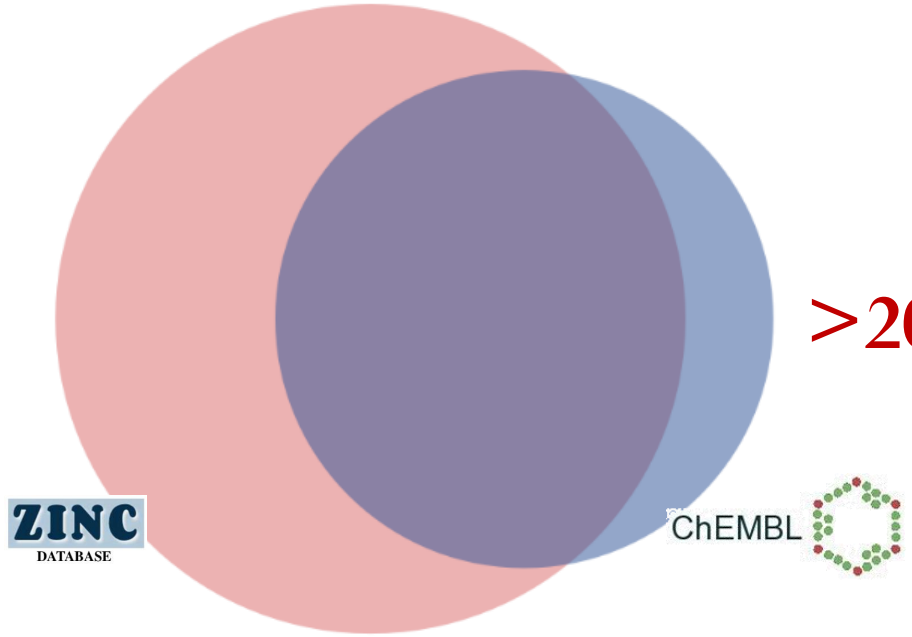
Popularity - 15

Yu. Zabolotna *et al.*, *J. Chem. Inf. Model.* 2021, 61, 1, 179–188

Commercial vs Biologically relevant data

> 40 K hierarchically related maps

> 100K chemotypes never been biologically tested



>20 K chemotypes missing in the commercial chemical space

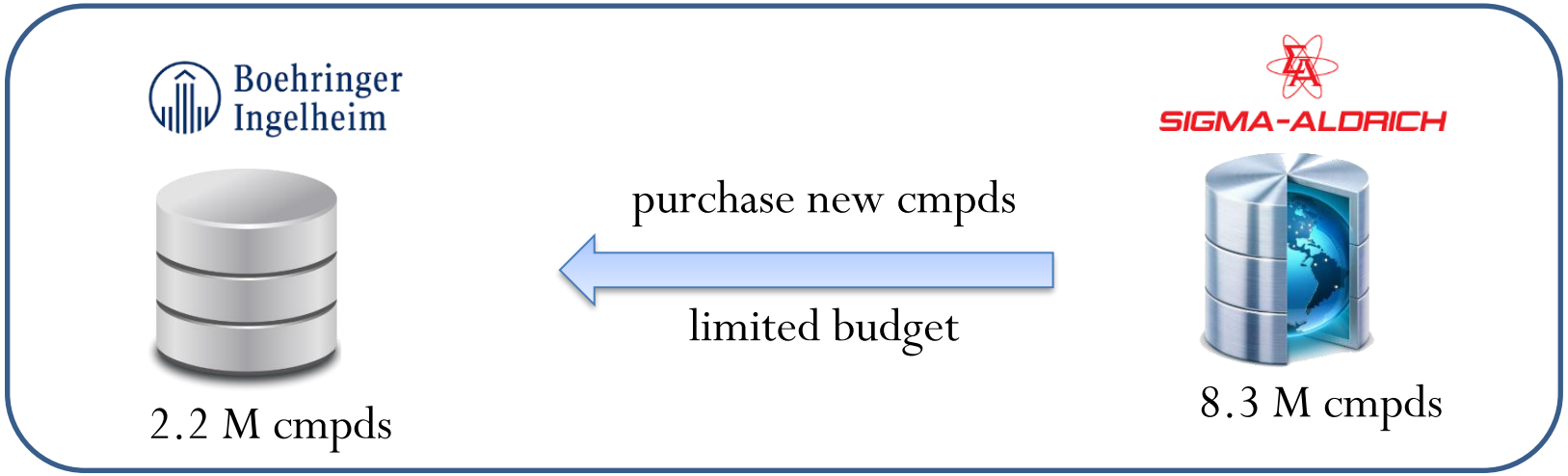
Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery

Yuliana Zabolotna, Fanny Bonachera, Dragos Horvath, Arkadii Lin, Gilles Marcou, Olga Klimchuk, and Alexandre Varnek*

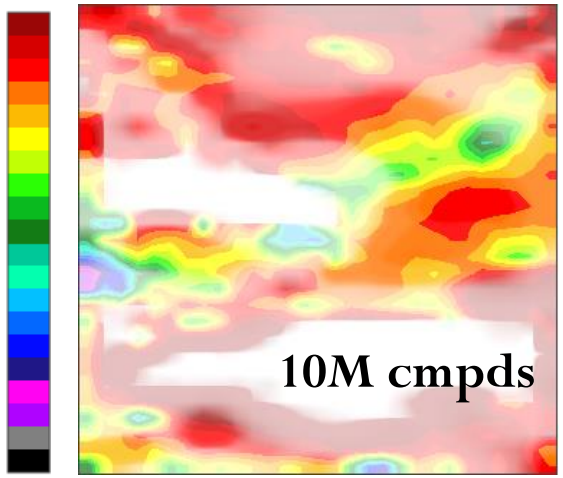
Cite This: https://doi.org/10.1021/acs.jcim.2c00509 Read Online

J. Chem. Inf. Mod., 2022

Case study 2: Proprietary Library Reshaping



Sigma-Aldrich



Boehringer

SIGMA-ALDRICH

45.5K MCS

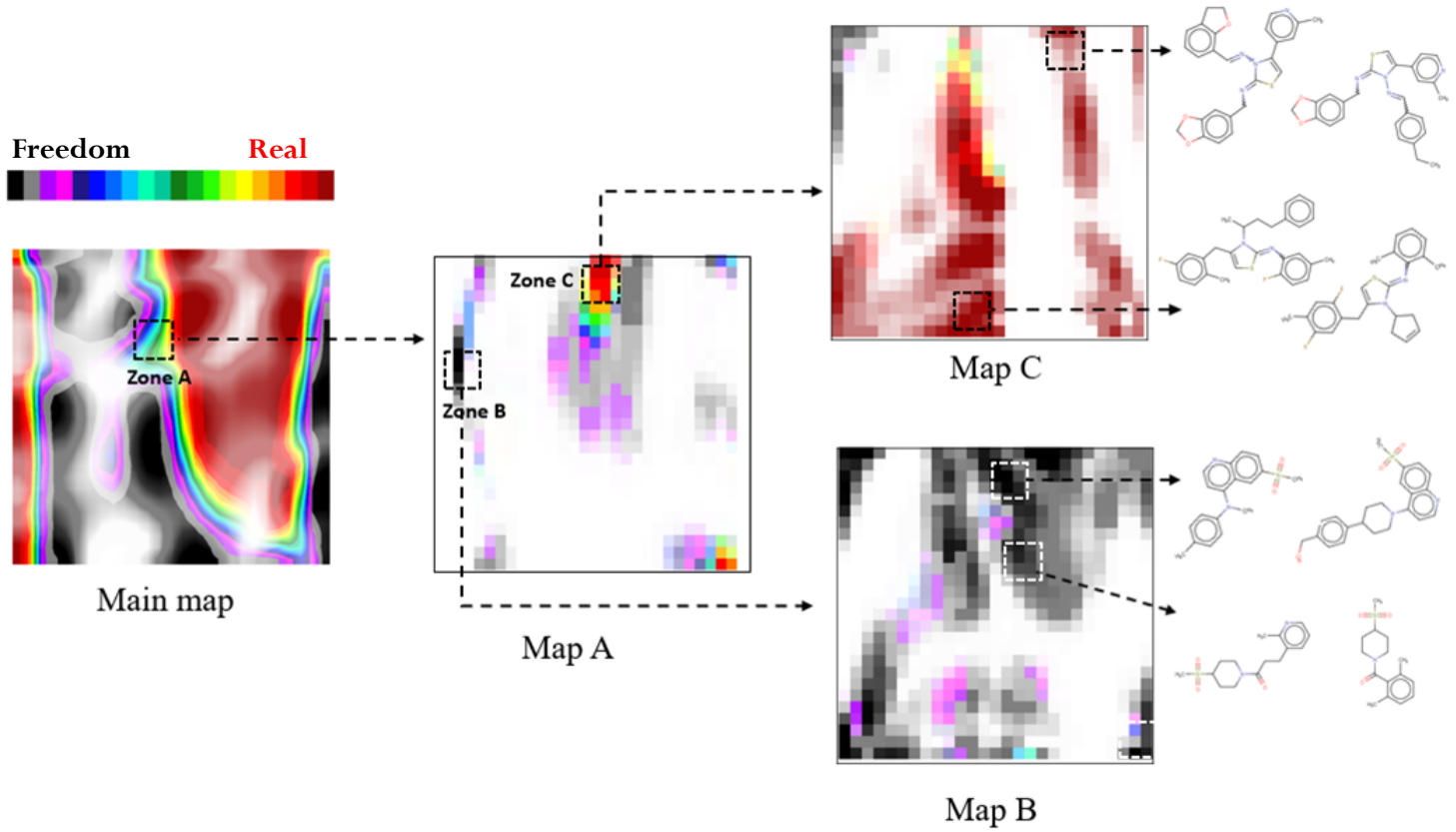
(401K cmpds)

Ro5 → GTM activity profiling → PAINS

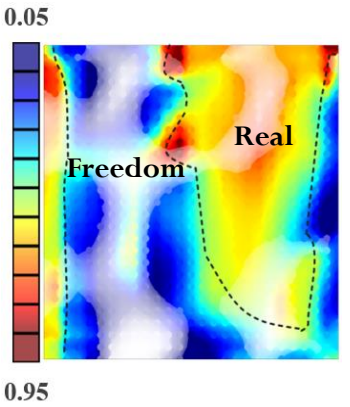
≈ 1.2K cmpds

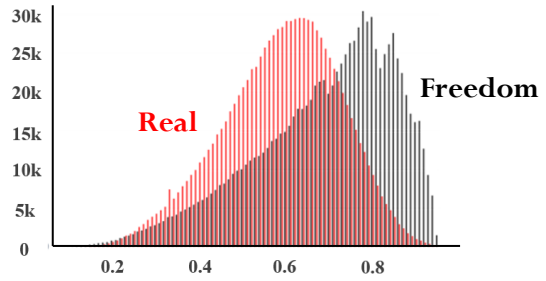
A. Lin et al. J Comput Aided Mol Des (2019), **33(3)**, 331-343

Case study 3: Freedom space (5B) vs REAL space (40B)



Quantitative Estimation of Drug-likeness (QED)



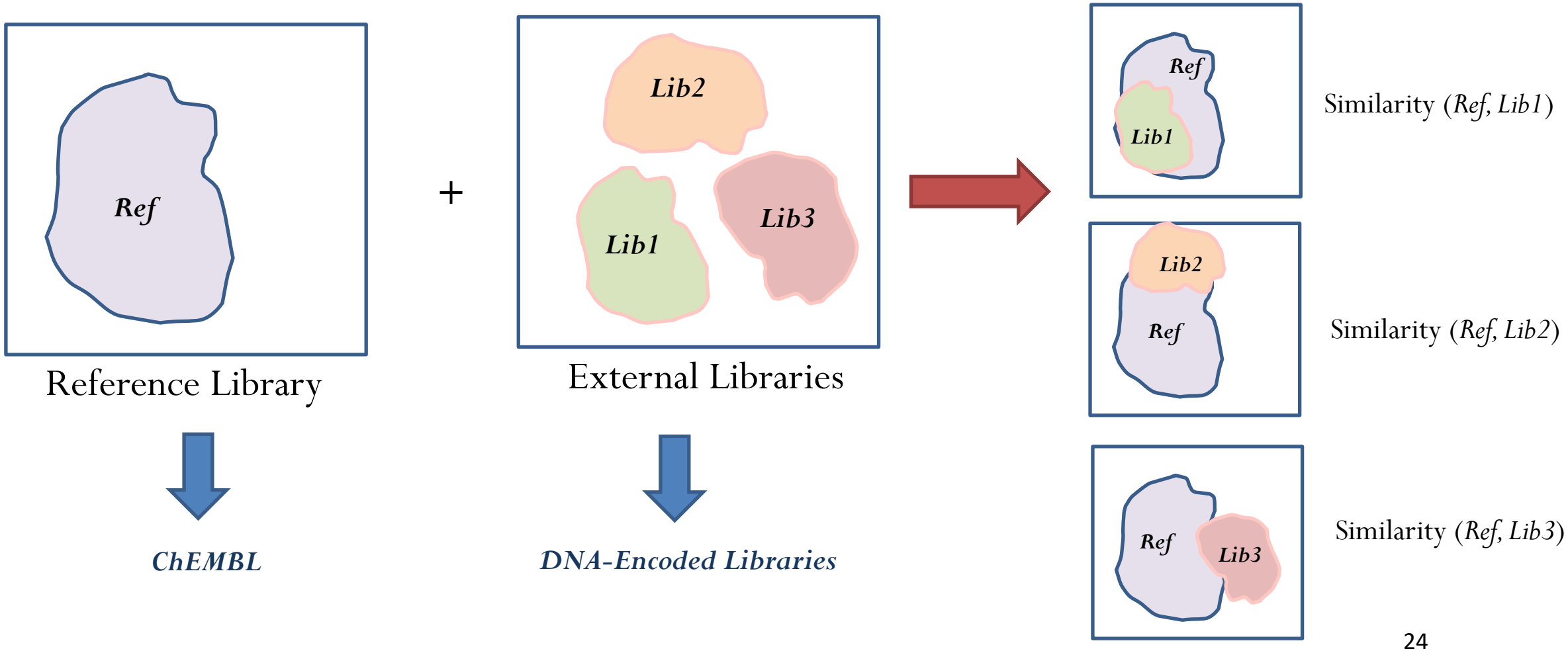


- **GTM shows very little overlap between Freedom and REAL spaces**
- **Freedom space is more enriched with drug-like compounds**

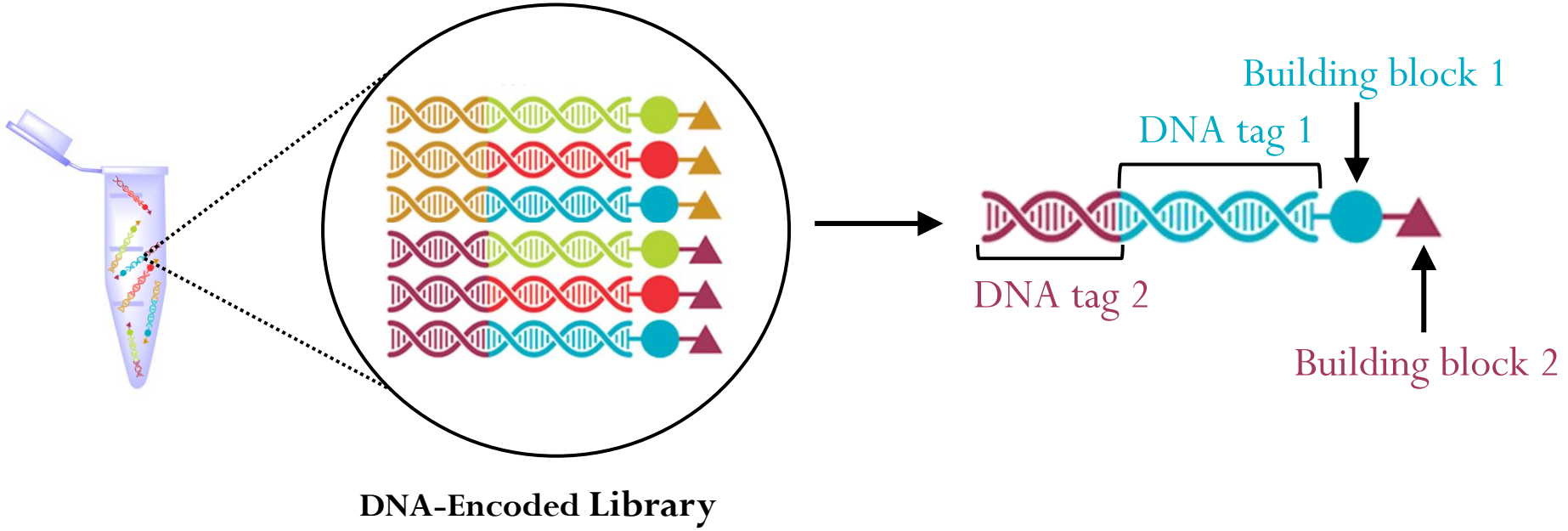
M. Protopopov et al., *Mol. Inf.* , 2024, DOI: 10.1002/minf.202400114

Multiple libraries analysis

Task : selection a library most similar to the reference one



Generation and analysis of general-purpose DELs

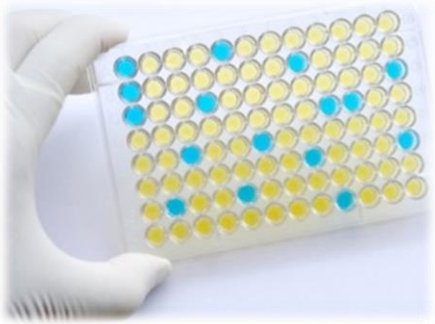


DNA-Encoded Library: combinatorial collection of small molecules covalently attached to the short DNA tag

Halford, B. How DNA-encoded libraries are revolutionizing drug discovery. *Chem. Eng. News.* **2017**, 95, 28.

DEL challenge

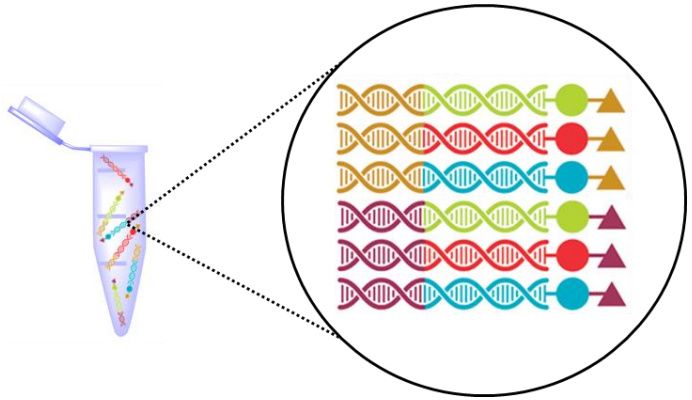
Screening libraries



Parallel screening in separate "wells"

Individual compounds may be cherry-picked

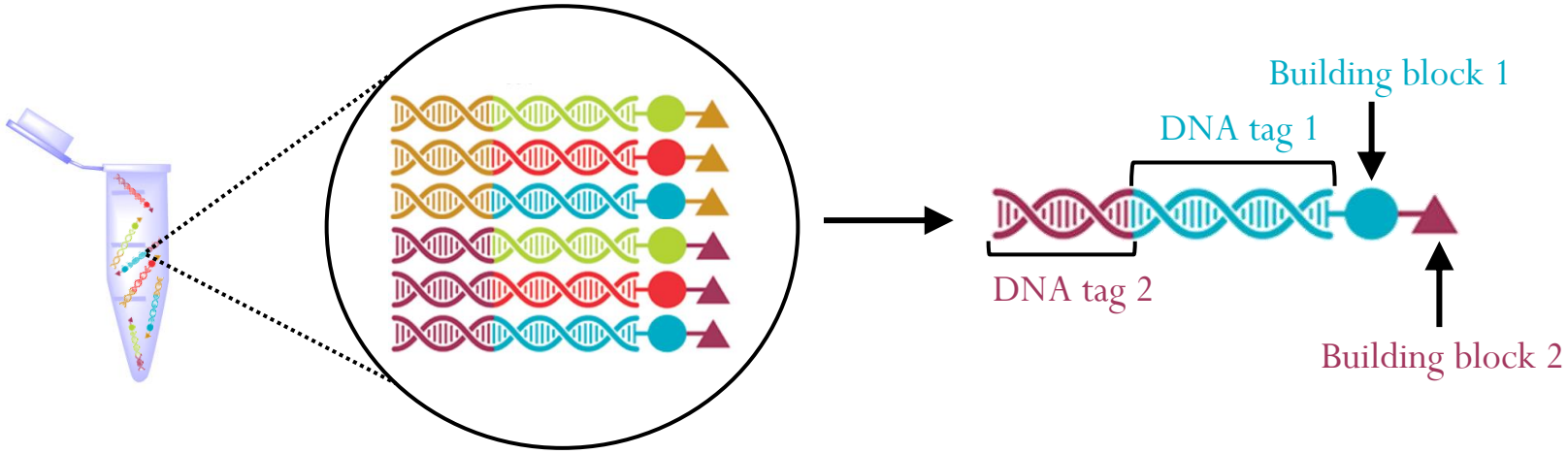
DNA-encoded libraries



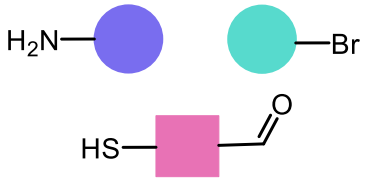
Simultaneous screening in a single tube

Entire library as an object must be considered

DNA-Encoded Libraries (DEL)



N(Building Blocks) * M(Reaction Rules)



Coupling
Amination
Heterocyclisation
……

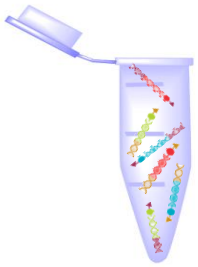
Thousands of different DELs

Huge investments for synthesis & screening:
DEL of 800M compounds ~150 000 €

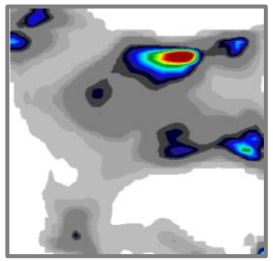
How to select an optimal DEL for a particular drug discovery task?

R. Goodnow et al Nature Reviews Drug Discovery **16**, 131–147 (2017)

GTM-based similarity assessment DEL /reference library (ChEMBL)



DEL_i



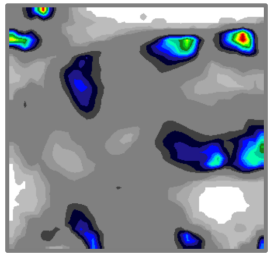
GTM (DEL_i)

Library representative descriptors vector (Φ)

ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5

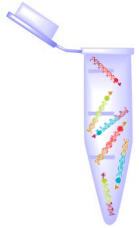
Reference library

ChEMBL



GTM (reference)

Φ_1	Φ_2	Φ_3	Φ_4	Φ_5



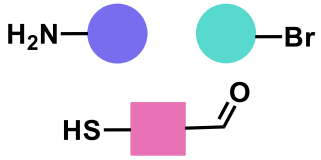
ChEMBL

ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5

Φ_1	Φ_2	Φ_3	Φ_4	Φ_5

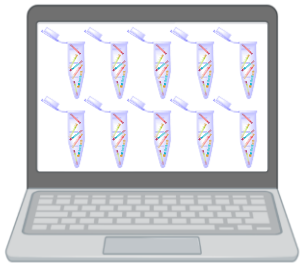
Similatity assessement
(Tanimoto coefficient)

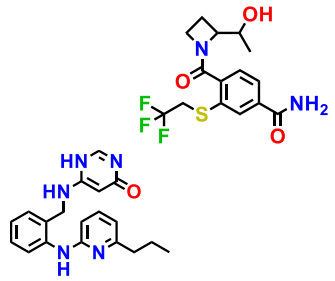
Focused Library design: case study



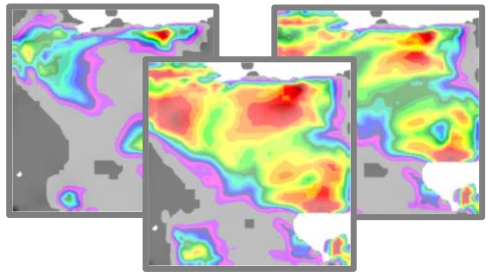
79.000 Building blocks from eMolecules

eDesigner tool

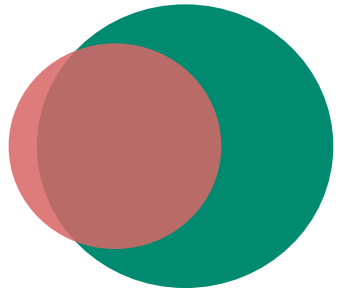




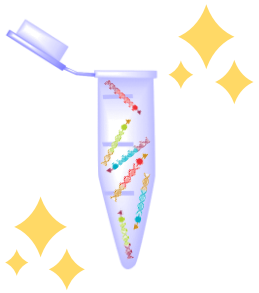
- **2500** DELs designed (size: 1M-1B)
- **2.5 B** compounds generated (1M cmpds per DEL)



2500 comparative landscapes DEL_i/reference (ChEMBL)

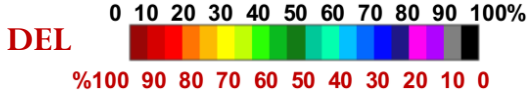


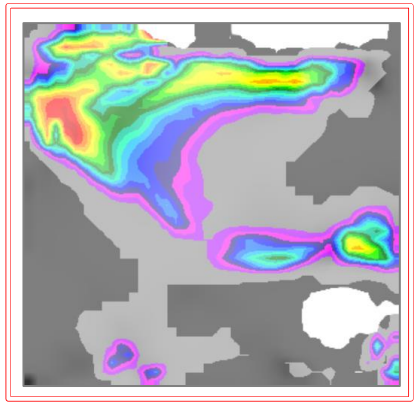
DEL_i / ChEMBL coverage score calculation for each map



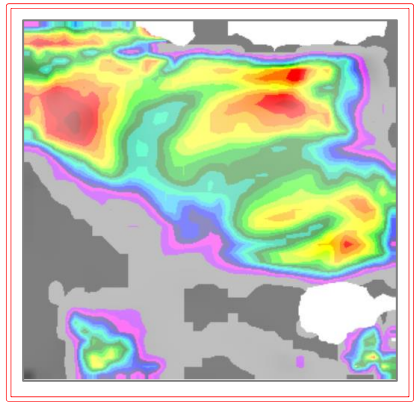
Selection of highly scored DELs

How many DELs cover ChEMBL chemical space ?

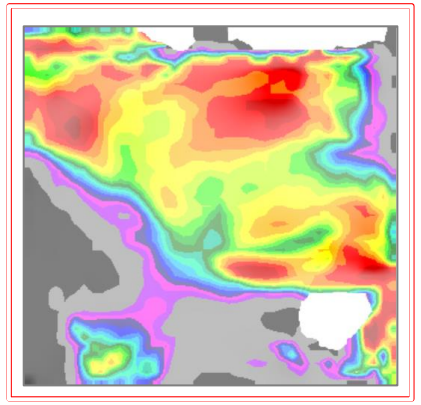




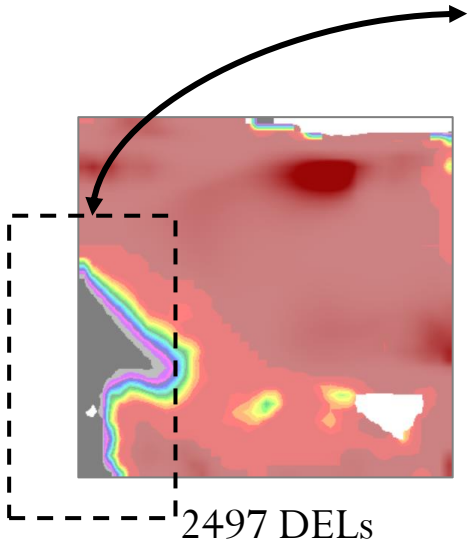
the best DEL

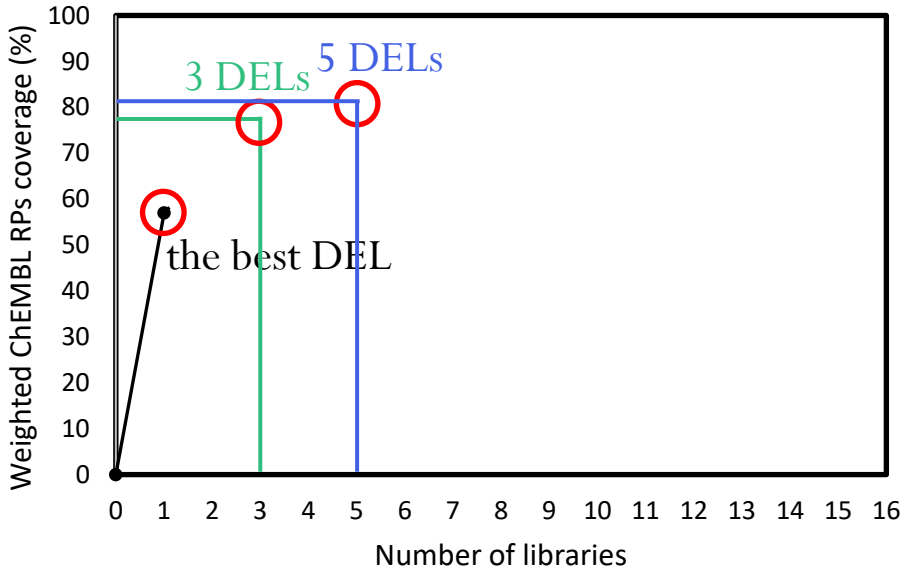


3 DELs



5 DELs



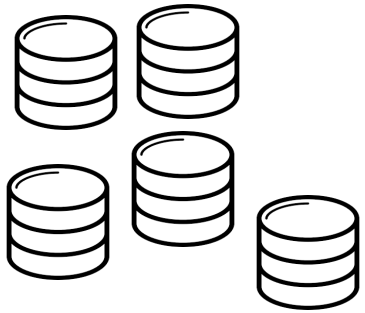


3 “platinum” DELs cover >80% of ChEMBL chemical space

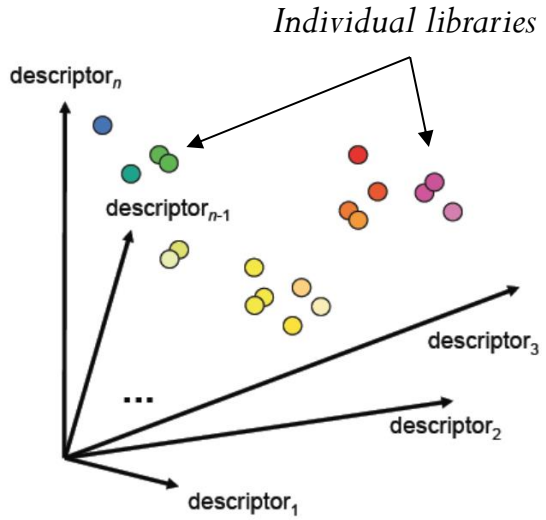
R. Pikalyova et al. Mol. Inf. 2022, 41, 2100289.

Chemical Library Space

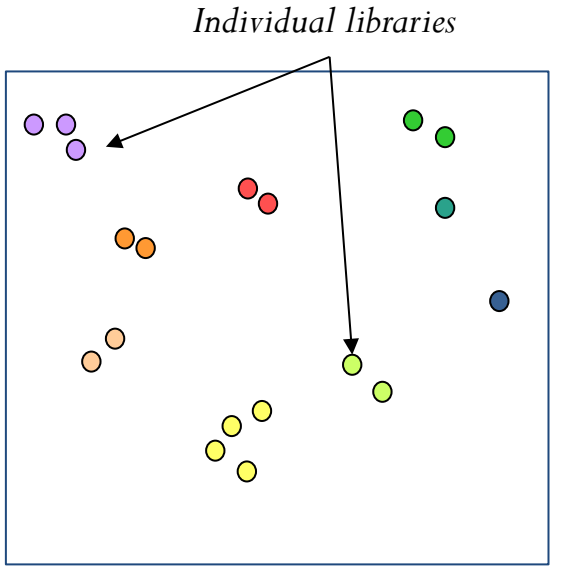
- **A GTM encodes a chemical library as a vector calculated from the property/activity landscape**
- **Ensemble of vectors can be used to build a meta-GTM (*μGTM*) where each data point represents a library**



Collection of chemical libraries



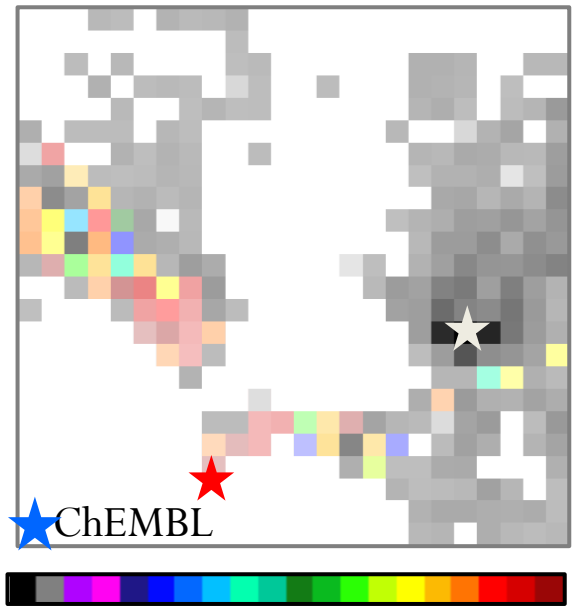
Initial library space
(N-dimensional)



μGTM

µGTM of the DEL space

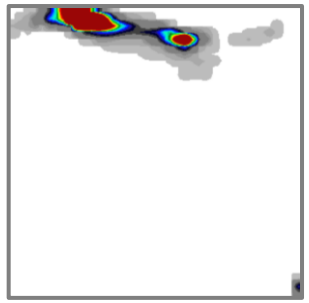
Close /remote to ChEMBL

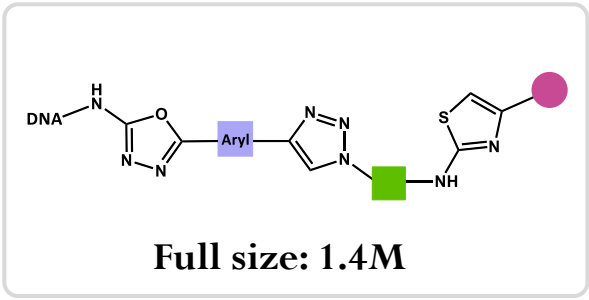


Remaining DELs

100 DELs best overlapping with ChEMBL

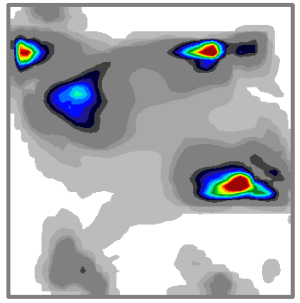
Density landscapes

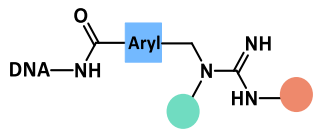




Full size: 1.4M

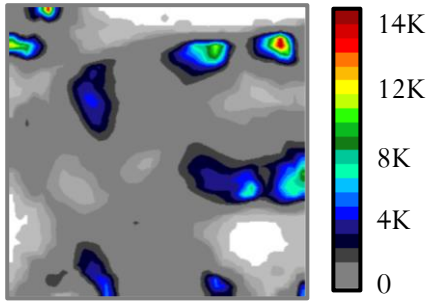
remote





Full size: 95.6M

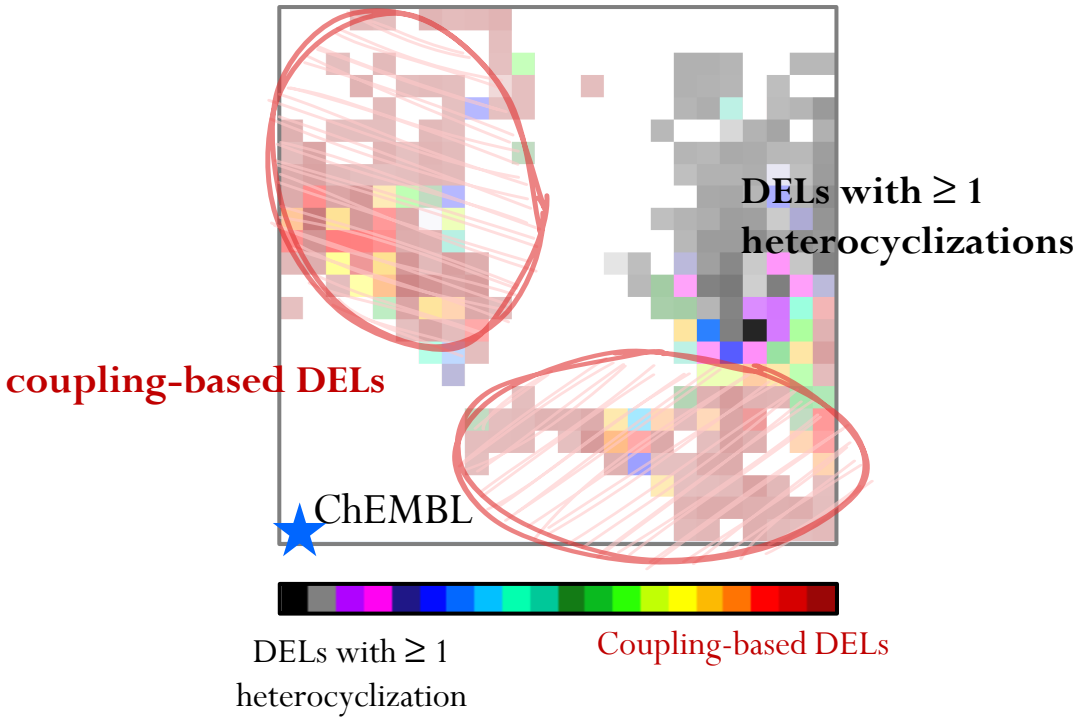
closest

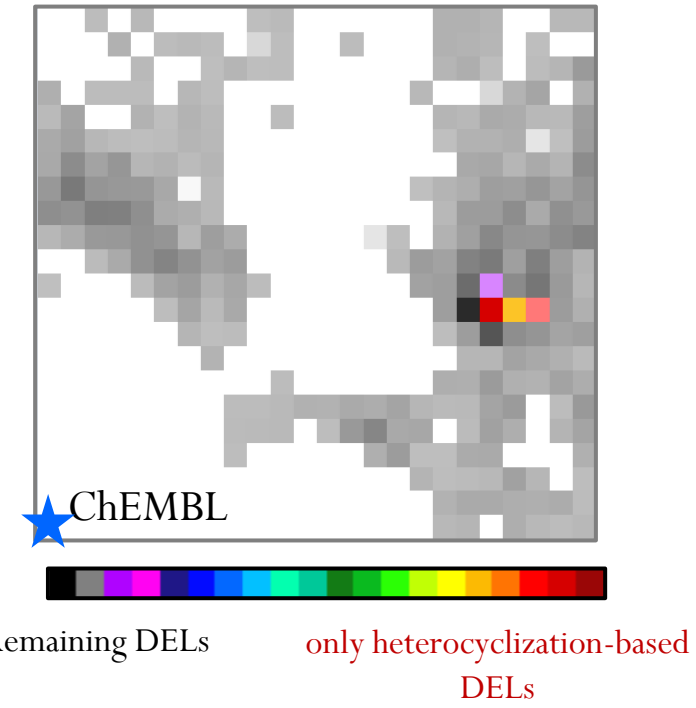


Reference (ChEMBL)

R. Pikalyova et al. J Chem Inf Model. **2023** 63 (17), 5571-5582

μGTM: DEL reaction types

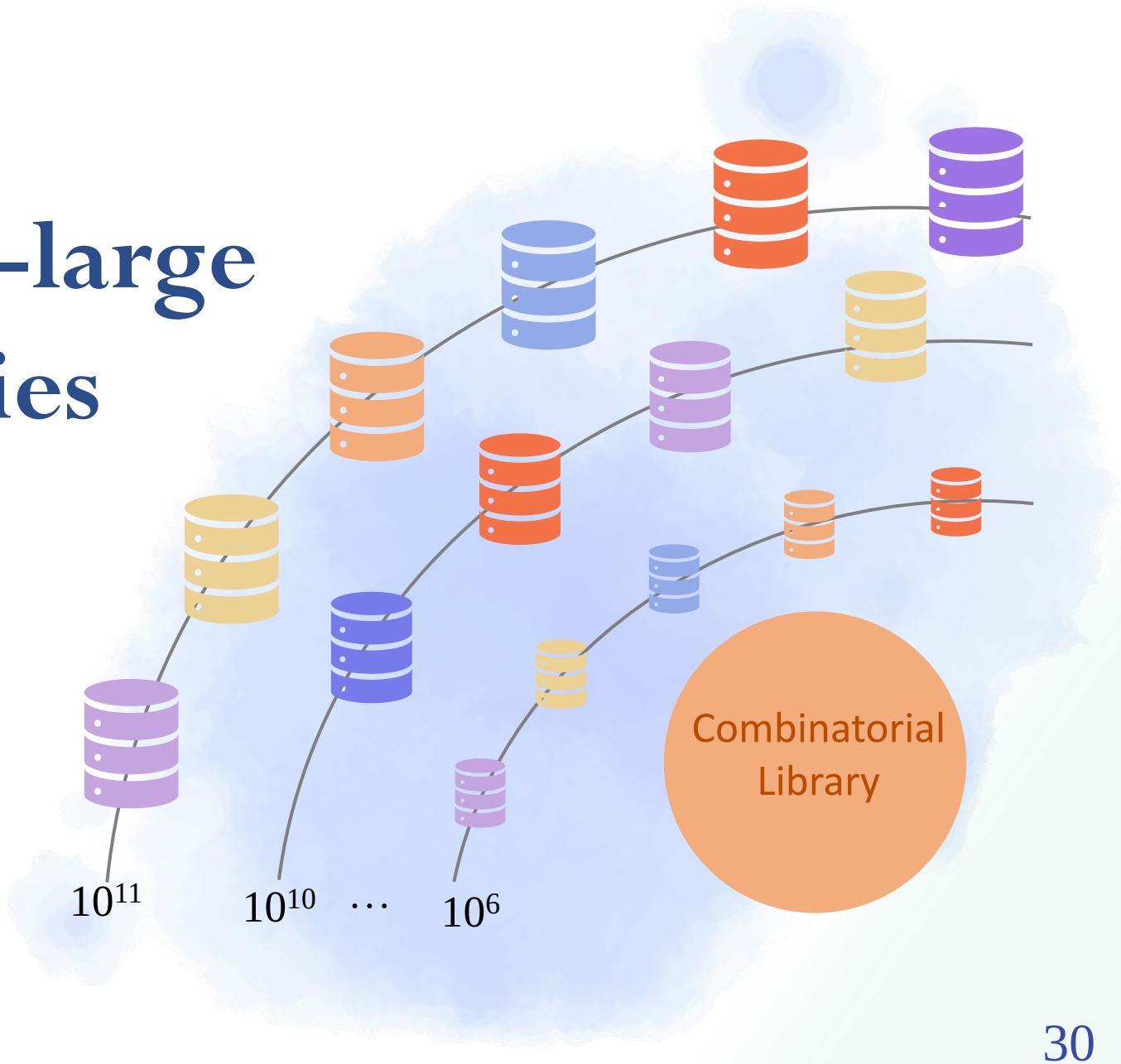




- **Most of the DEL space is covered by coupling-based libraries**
- **Very few DELs that are purely heterocyclization-based**
- **The coupling-based DELs are more similar to ChEMBL than the heterocyclization ones.**

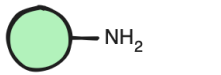
R. Pikalyova et al. J Chem Inf Model. **2023** 63 (17), 5571-5582

Cartography of ultra-large combinatorial libraries

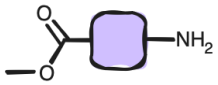


Pipeline of GTM construction for combinatorial library

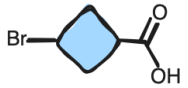
Building blocks **Reactions**



Rxn 11

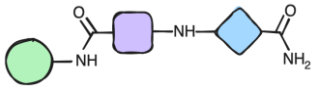


Rxn 25

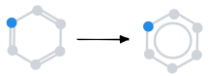


Rxn 18

1. Enumeration

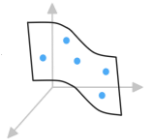


2. Standardization

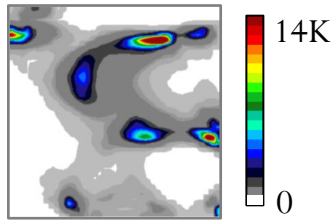


3. Descriptor calculation

4. Projection on the map



GTM

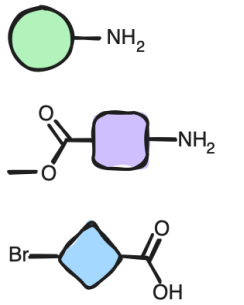


Time consuming for large libraries

1M cmpds (MFP) -12h/1 CPU

CoLiNN – Neural Network for GTM preparation without structure enumeration

Building blocks **Reactions**

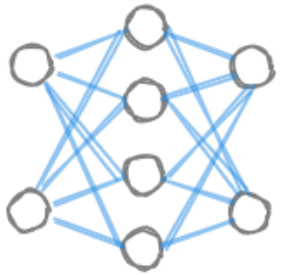


Rxn 11

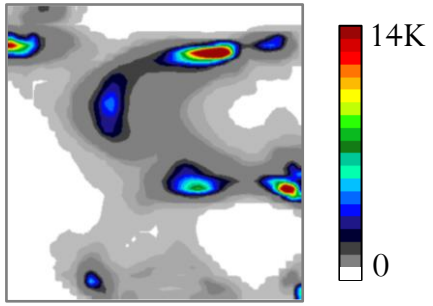
Rxn 25

Rxn 18

CoLiNN



GTM



1M cmpds (MFP) -1 min/1 CPU

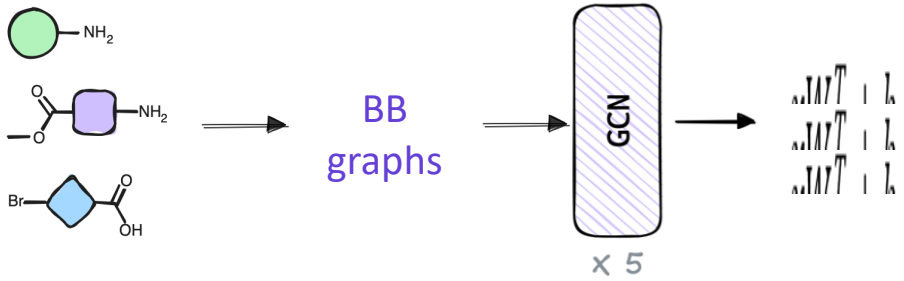
CoLiNN skips the enumeration step, thus accelerating GTM construction

Pikalyova R., T. Akhmetshin et al., et al. ChemRxiv , 2024, DOI: 10.26434/chemrxiv-2024-qh3bn

CoLiNN – Neural Network for GTM preparation without structure enumeration

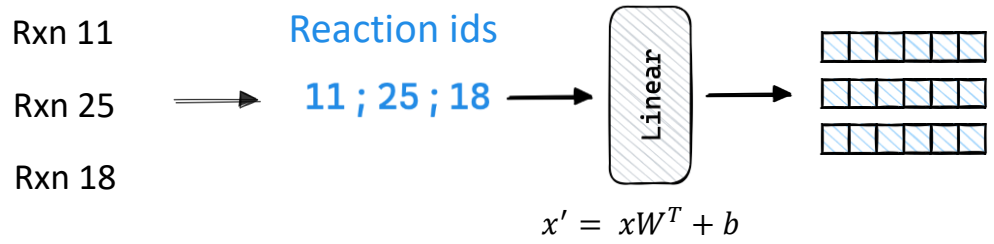
1. Building Block Embedding

Building blocks:

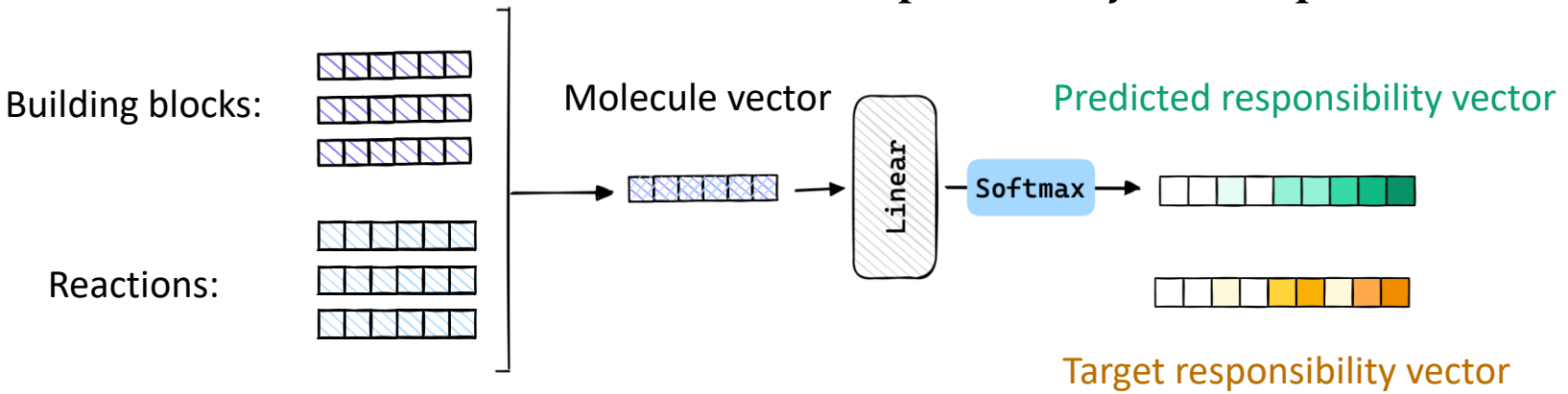


2. Reaction Embedding

Reactions:



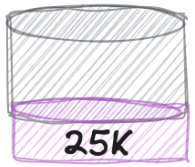
3. Responsibility vector prediction



Loss: Kullback-Leibler divergence

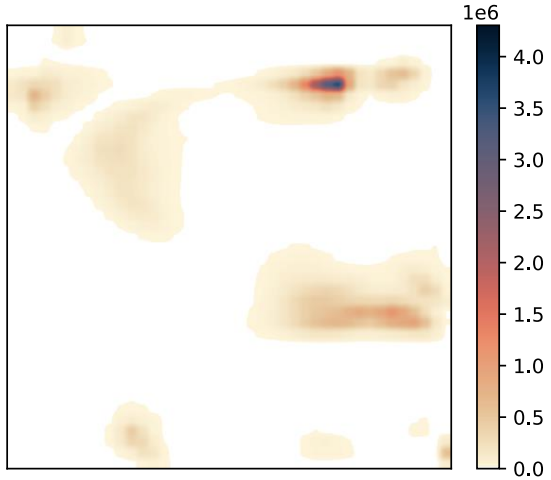
Pikalyova R., T. Akhmetshin et al., et al. ChemRxiv (2024).

CoLiNN – Neural Network for GTM preparation without structure enumeration

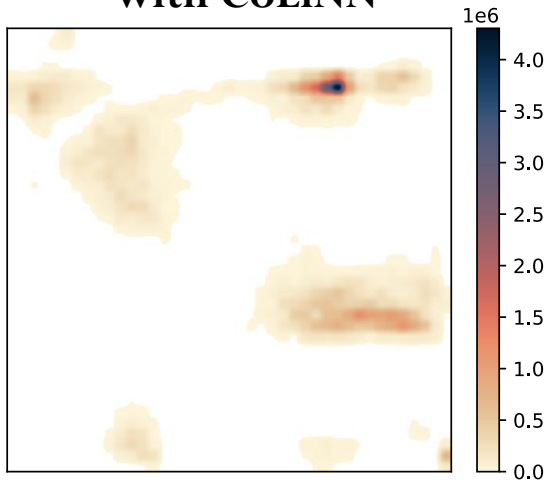


a set of 25K compounds is sufficient to train a model accurately predicting GTM for 90M DEL

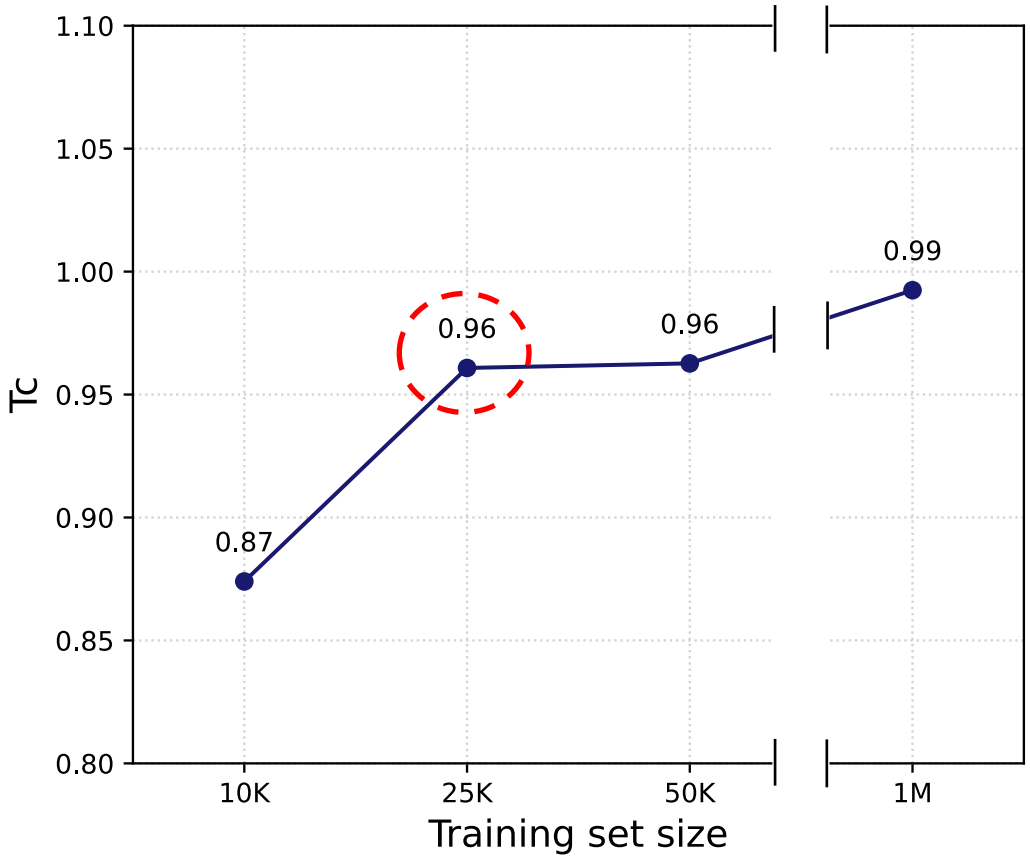
Data density of 90M enumerated structures



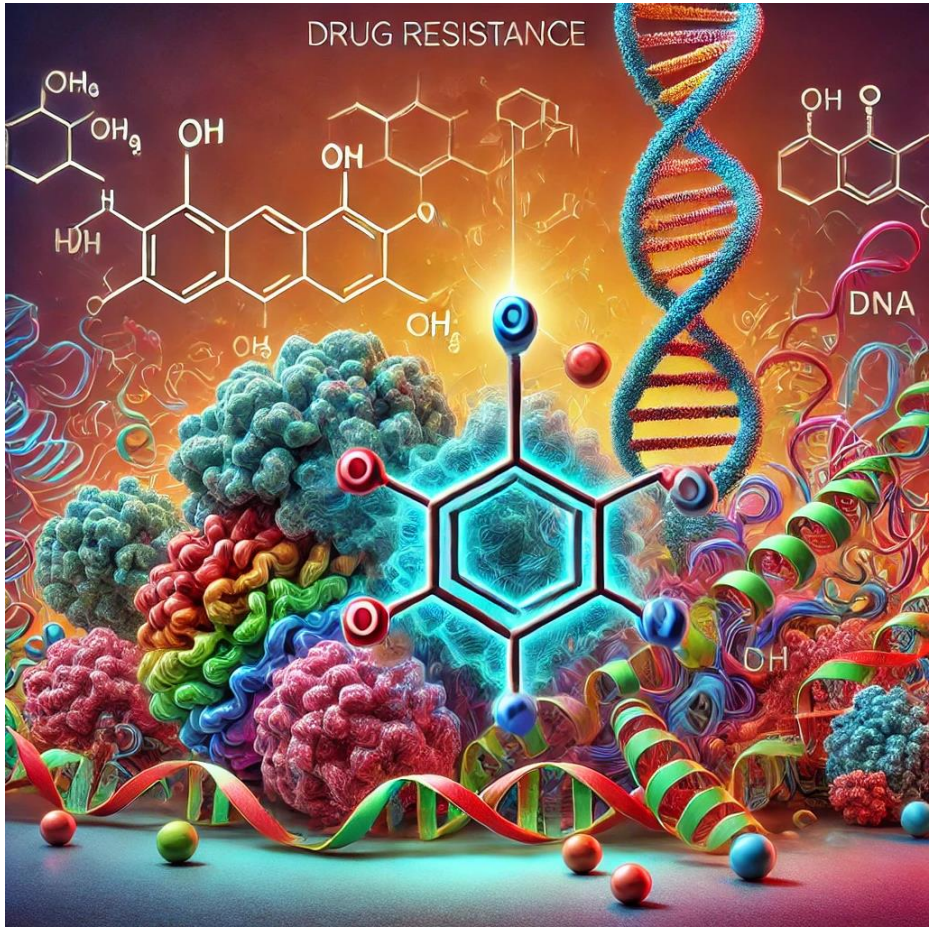
Data density predicted with CoLiNN



Model performance as a function of training set size



Pikalyova R., T. Akhmetshin et al., et al. ChemRxiv (2024).



Drug resistance:

Cartography analysis of proteins and nucleic acids mutations

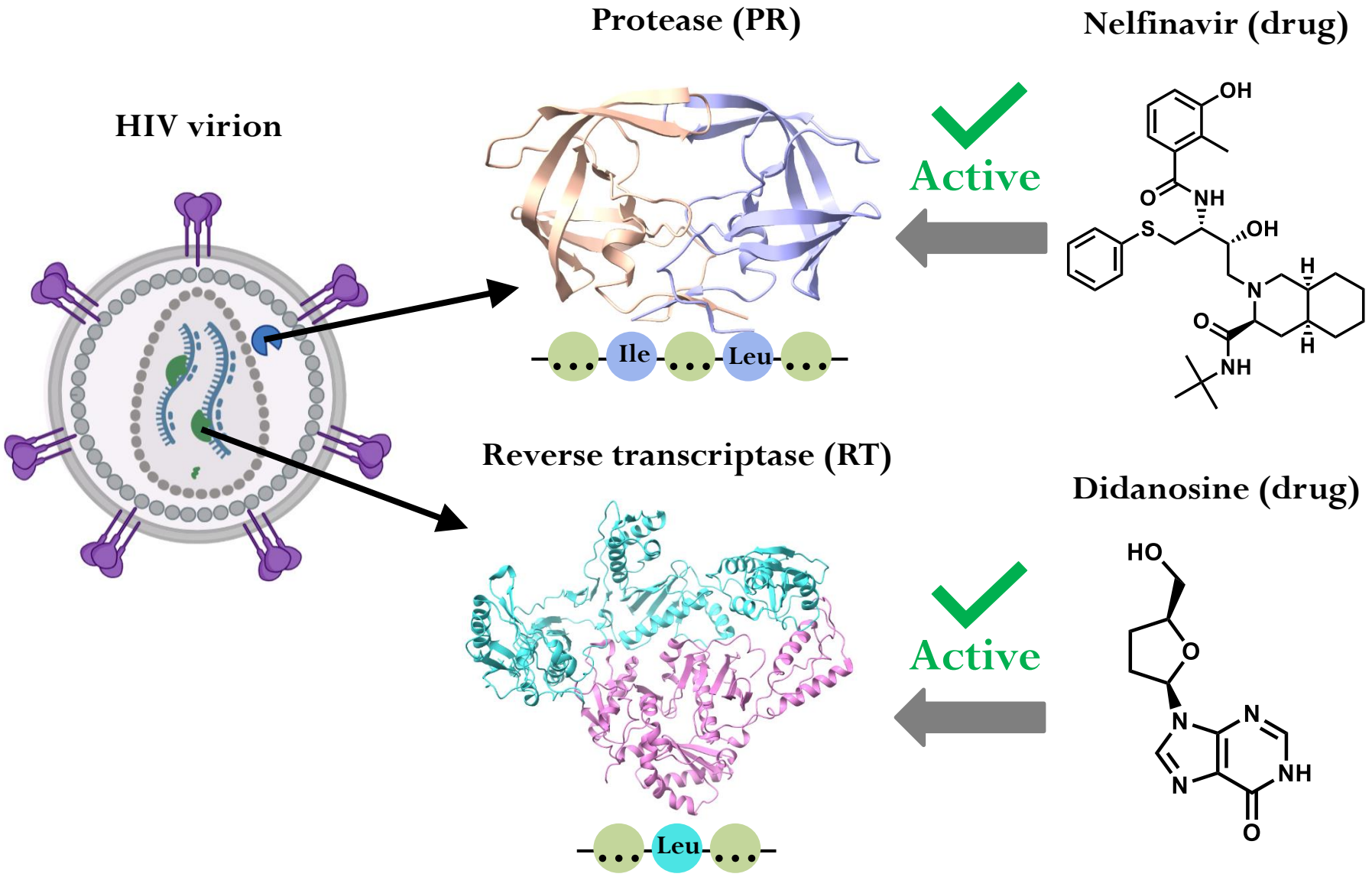
- **Human Immunodeficiency Virus (HIV)**
- ***Staphylococcus aureus* bacteria**

HIV drug resistance

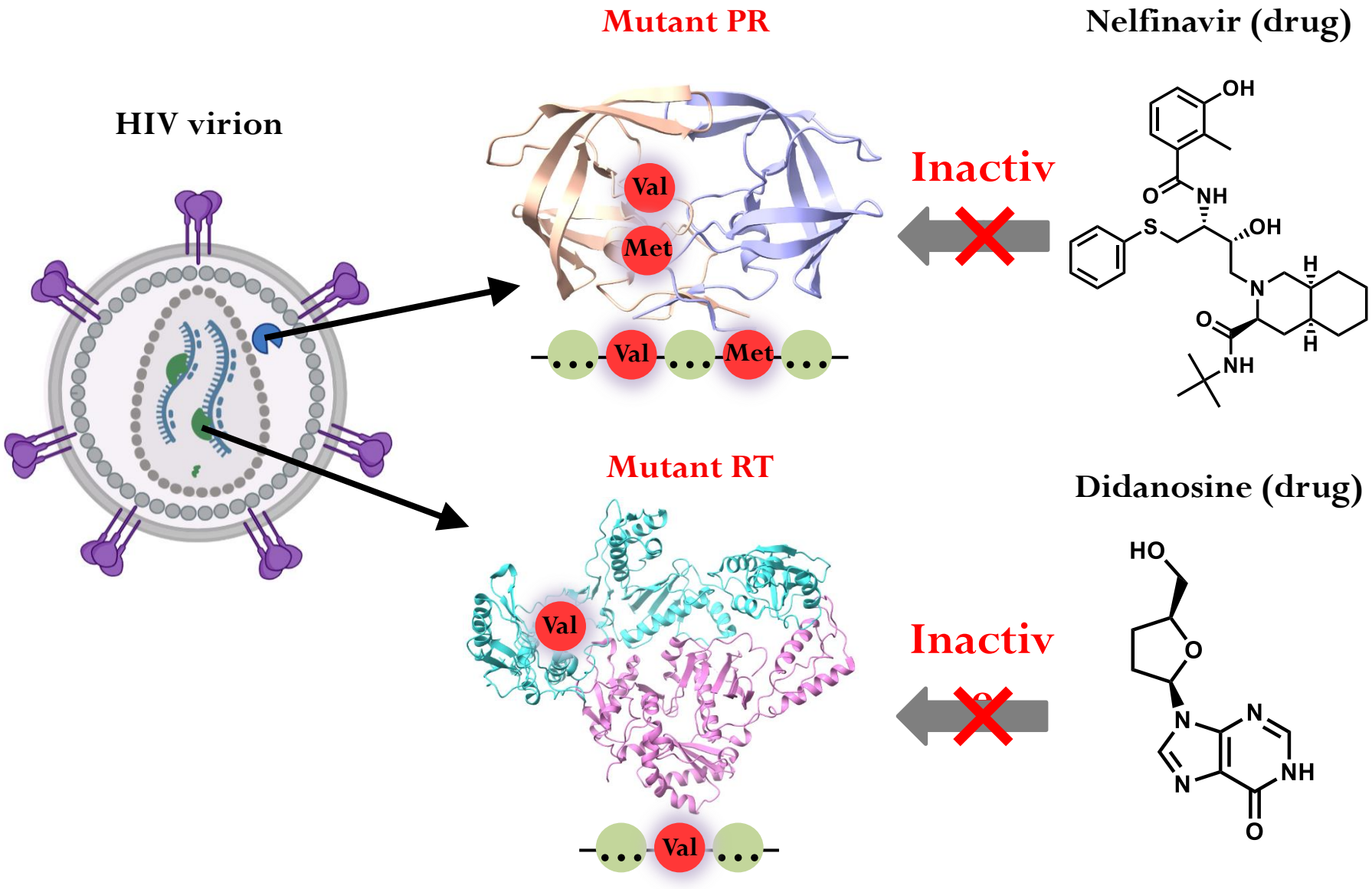
- 38 million people infected with HIV
- Highly effective therapy is available
- Increasing emergence of drug resistance

World Health Organization, *HIV Drug Resistance Report 2019*, **2019**.

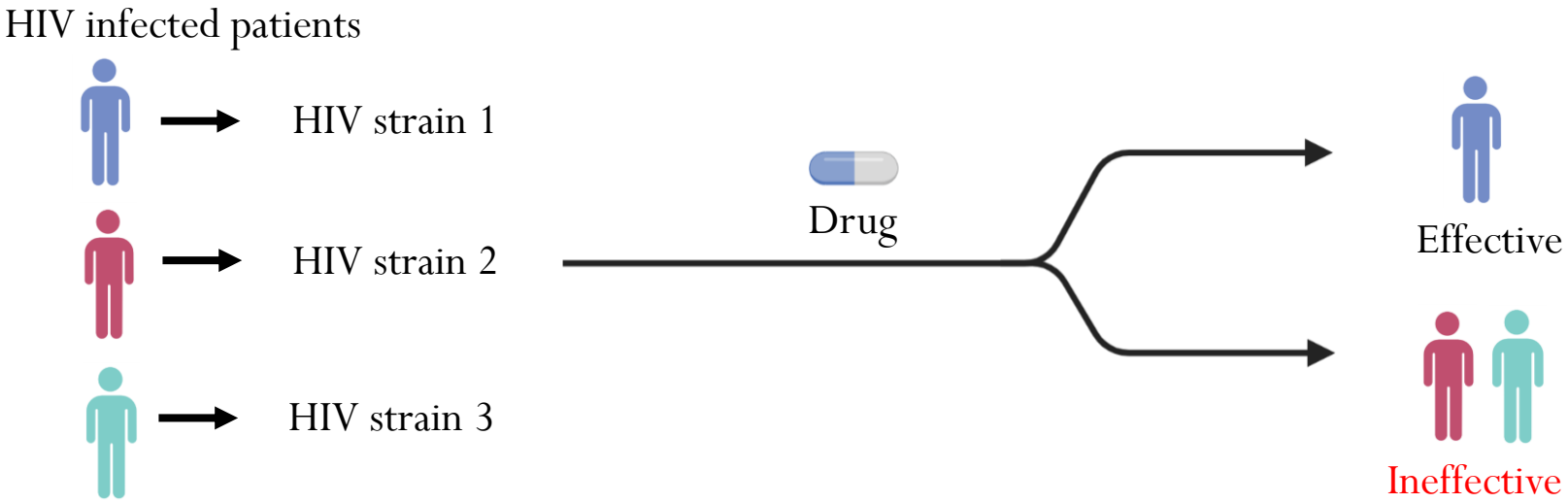
HIV drug resistance



HIV drug resistance

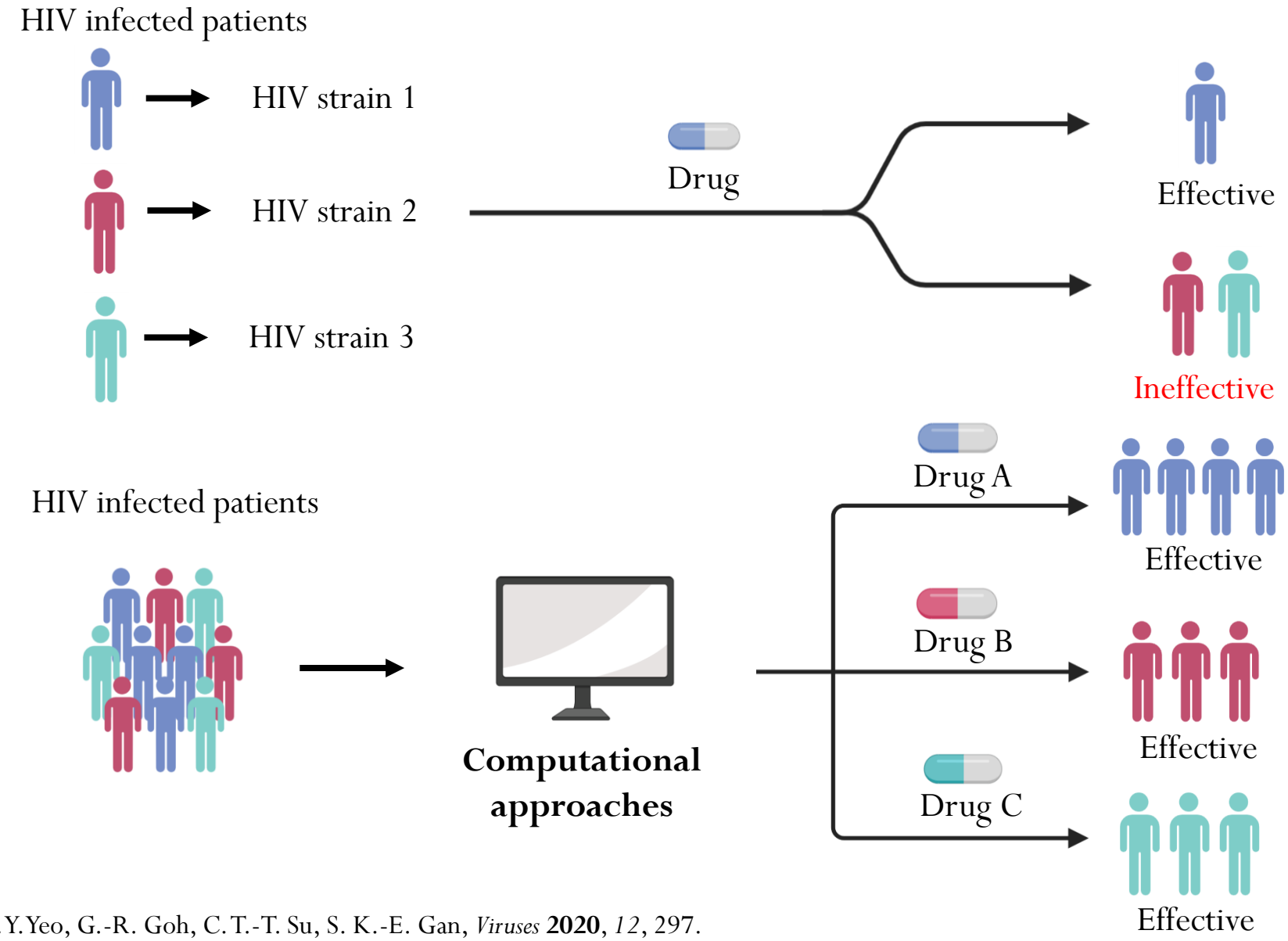


HIV drug resistance



J.Y.Yeo, G.-R. Goh, C.T.-T. Su, S. K.-E. Gan, *Viruses* **2020**, *12*, 297.

HIV drug resistance



J.Y.Yeo, G.-R. Goh, C.T.-T. Su, S. K.-E. Gan, *Viruses* **2020**, *12*, 297.

HIV drug resistance: data

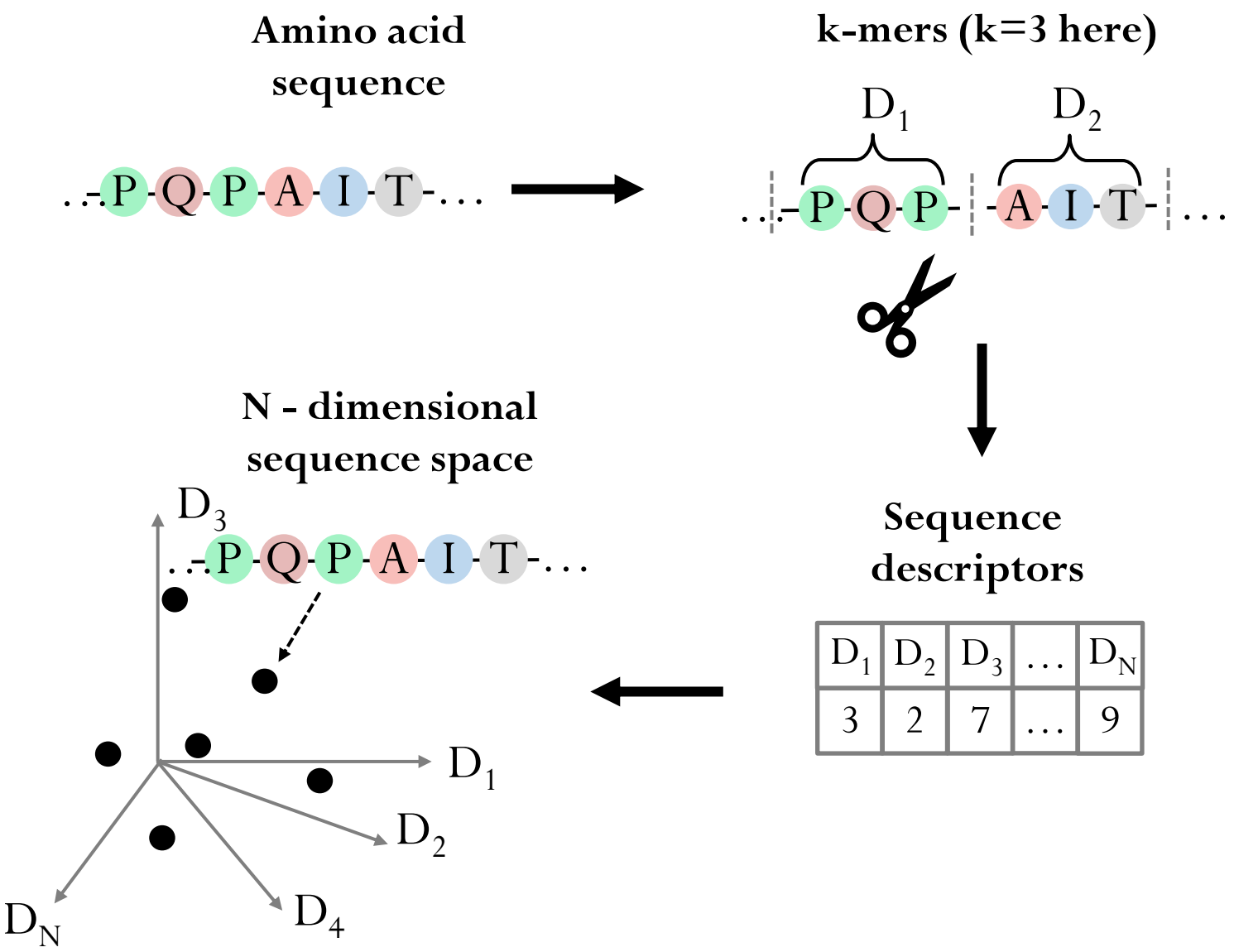
Stanford HIV database

4324 HIV protease (PR) and reverse transcriptase (RT) amino acid sequences and associated resistance profiles to 6 PR and 8 RT inhibitors

Sequence	***Drug resistance profile***				
	Indinavir	**Darunavir**	**Nelfinavir**	**Lopinavir**	**…**
. . . P N I W K T – . . .	Resistant	Susceptible	Resistant	Susceptible	…
. . . P L I T K T – . . .	Resistant	Susceptible	Resistant	Resistant	…
. . . P Q I T K T – . . .	Susceptible	Susceptible	Resistant	Susceptible	…

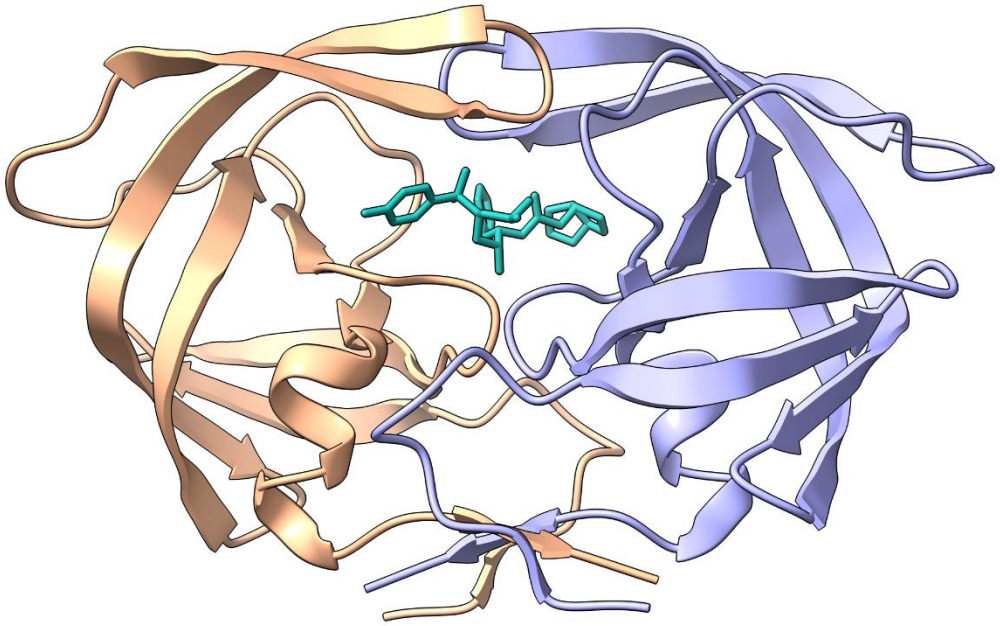
HIV Drug Resistance Database. https://hivdb.stanford.edu/. Datasets updated on 03.02.2021.

Sequence descriptors



HIV-1 protease drug resistance

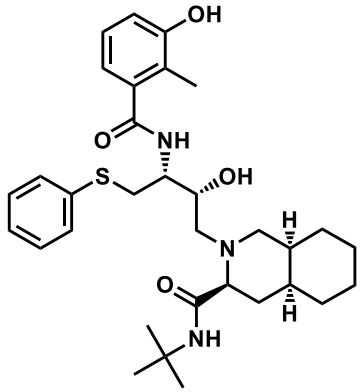
HIV-1 Protease



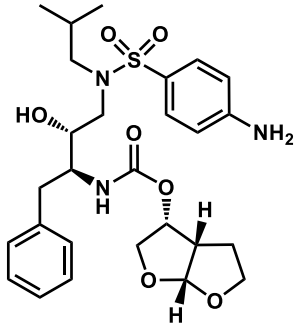
- Homodimer
- 99 amino acid residues per monomer
- 8 approved protease inhibitors

Protease inhibitors:

Nelfinavir

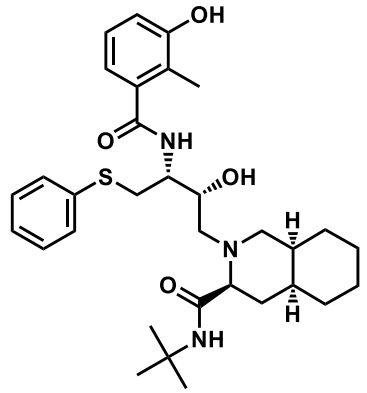


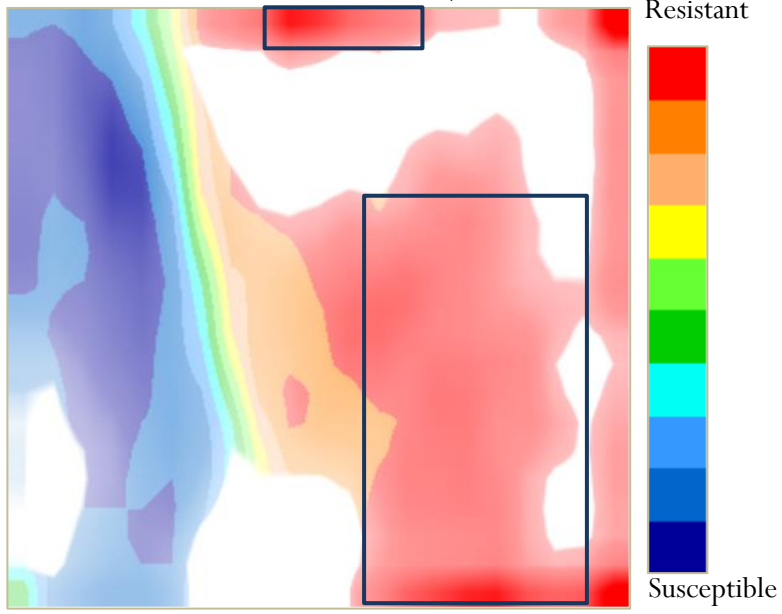
Darunavir



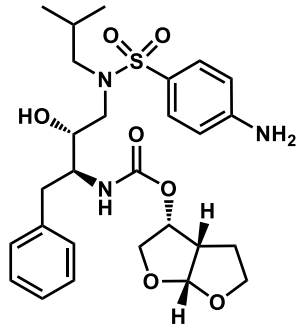
Resistance landscapes for protease inhibitors

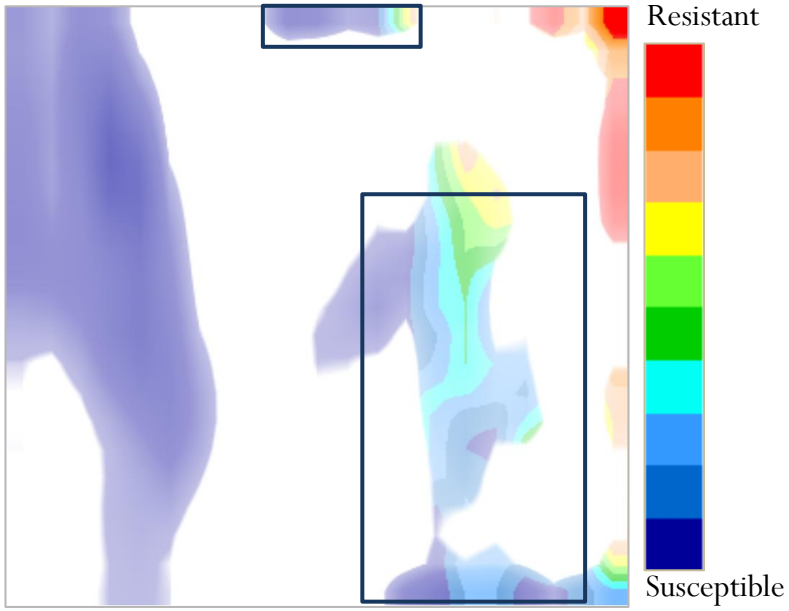
Nelfinavir





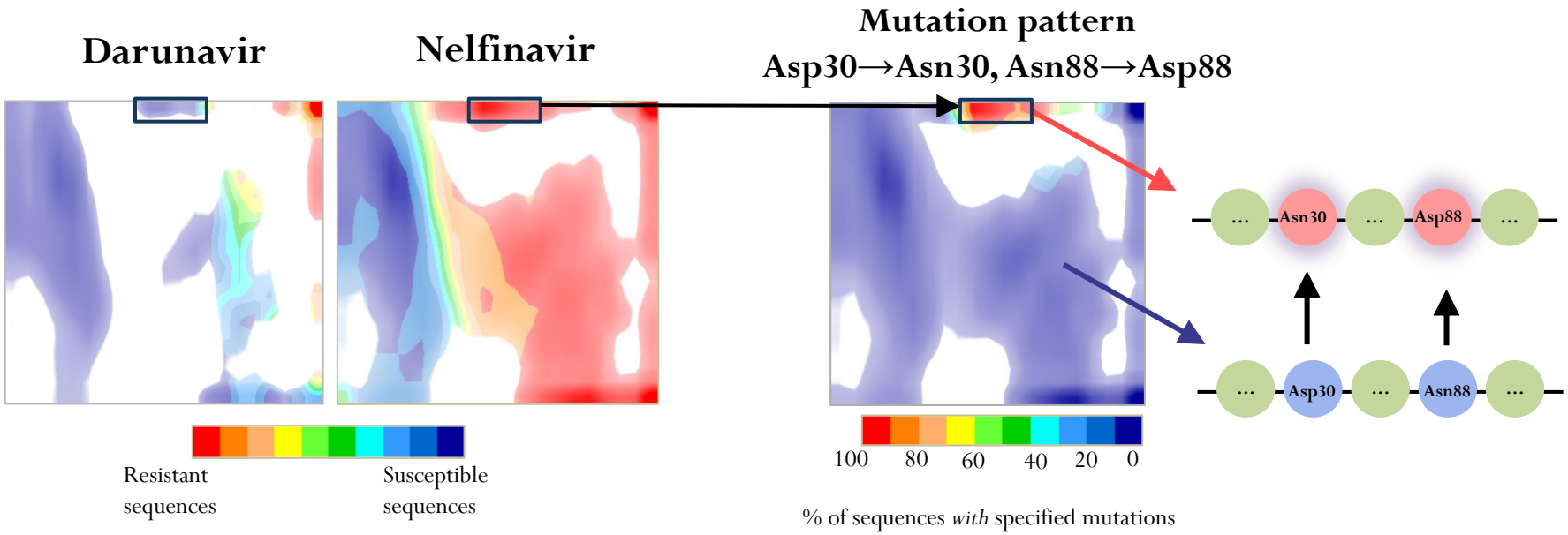
Darunavir





Pikalyova, K. et al. *Bioinformatics*, 2022, **38** (8), 2307–2314

Resistance-determining mutation patterns in protease

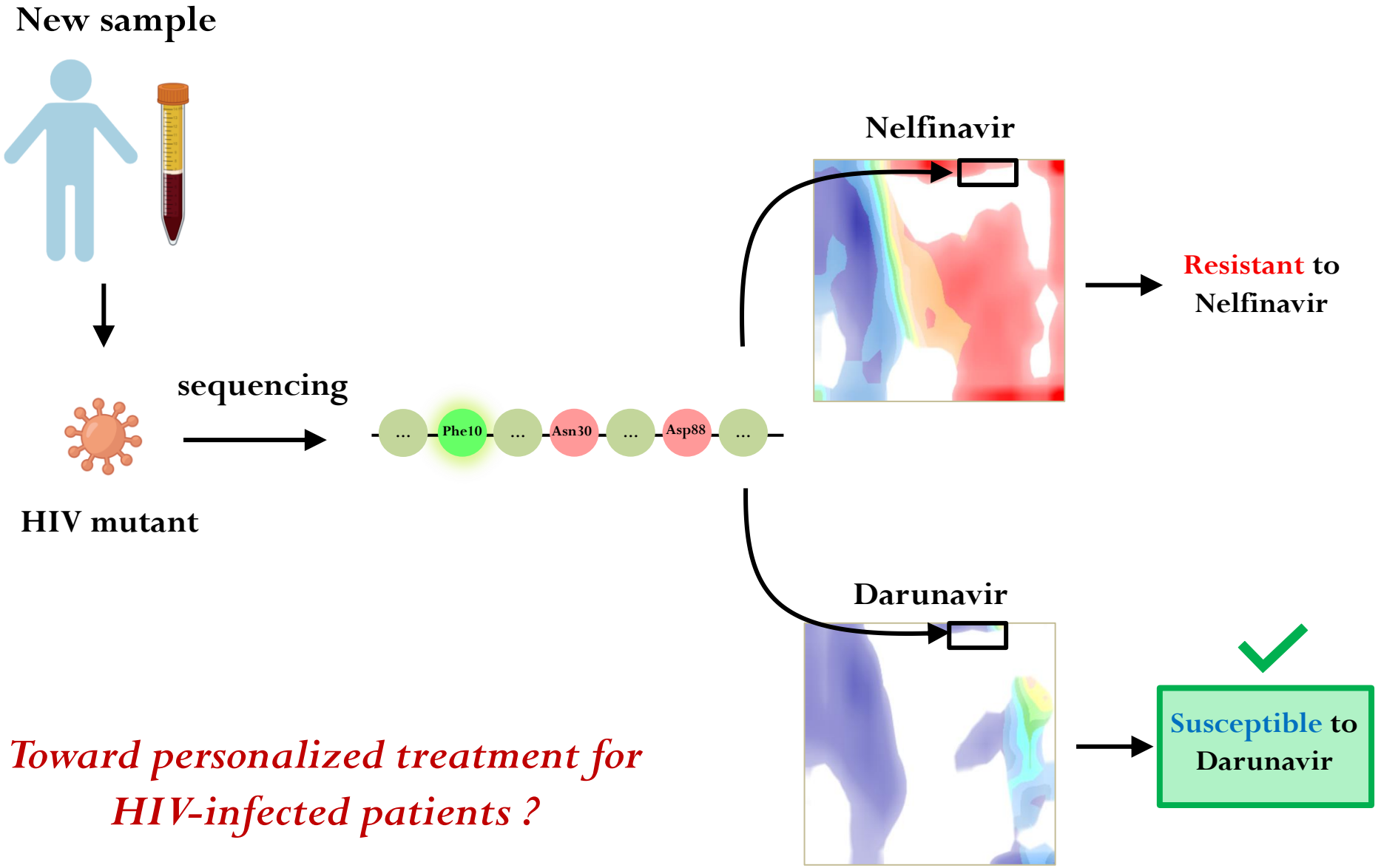


Asp30→Asn30, Asn88→Asp88

Mutation pattern defining resistance to *nelfinavir*

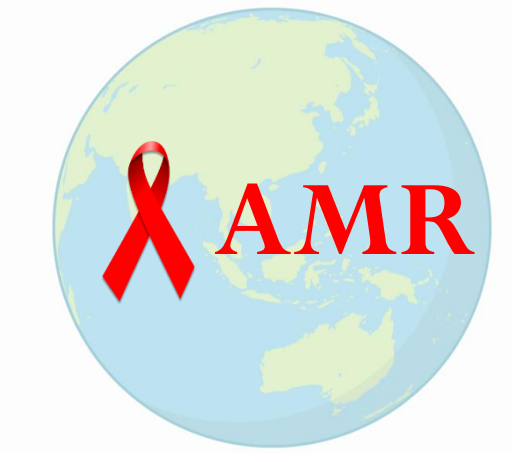
Pikalyova, K. et al. *Bioinformatics*, 2022, **38** (8), 2307–2314

Predicting drug resistance for emergent strains



Toward personalized treatment for HIV-infected patients ?

Antimicrobial resistance: *Staphylococcus aureus*

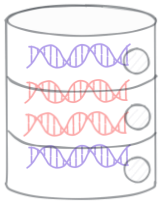


- Bacterial antimicrobial resistance (AMR) responsible for 1.27 million global deaths
- AMR could result in US$ 1 trillion additional healthcare costs by 2050 (the World Bank estimates)
- Increasing emergence of antibiotic resistance

World Health Organization, *Antimicrobial Resistance Report 2019*, **2019**.

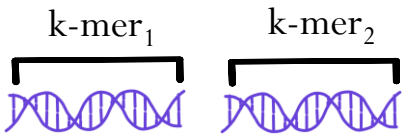
Tackling “fat data” challenge in genomic data

5773 Susceptible/Resistant for 11 antibiotics
S. aureus genomes from PATRIC DB



≈ 2.8 million base pairs long

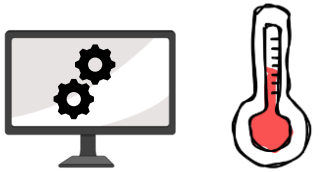
Calculation of k-mers (k=10, 15, 30, 90)



Issues:

Tens of millions of k-mers

48 M descriptors (31-mers)



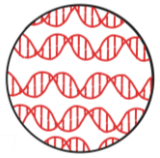
Challenging memory-wise

Kernel PCA
300 PCs

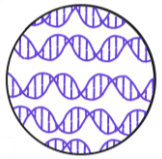
- GTM
- Random Forest
- SVM
- Gradient Boosting

BA > 0.75

Resistant

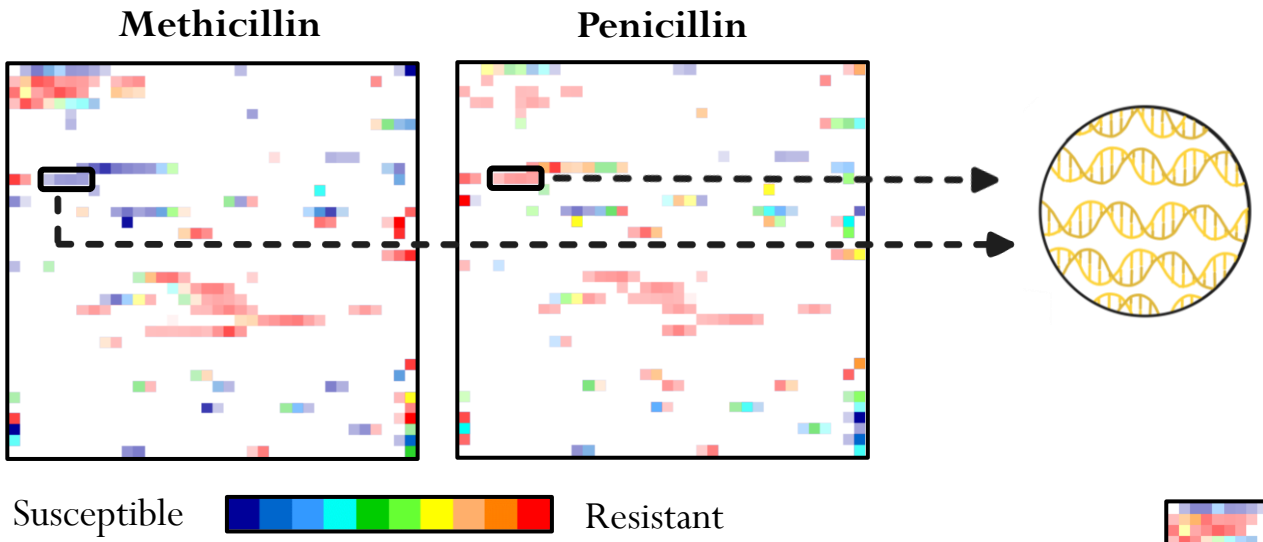


Susceptible

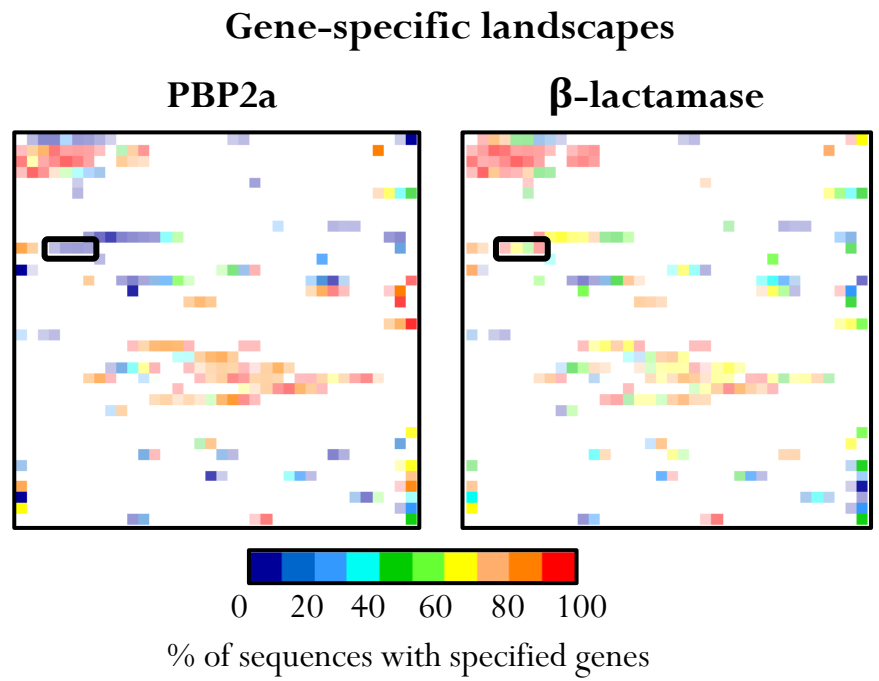


Pikalyova, K. et al. *Molecular Informatics*. 2024, 43:e202300263.

Gene-specific landscapes linking genes with drug resistance



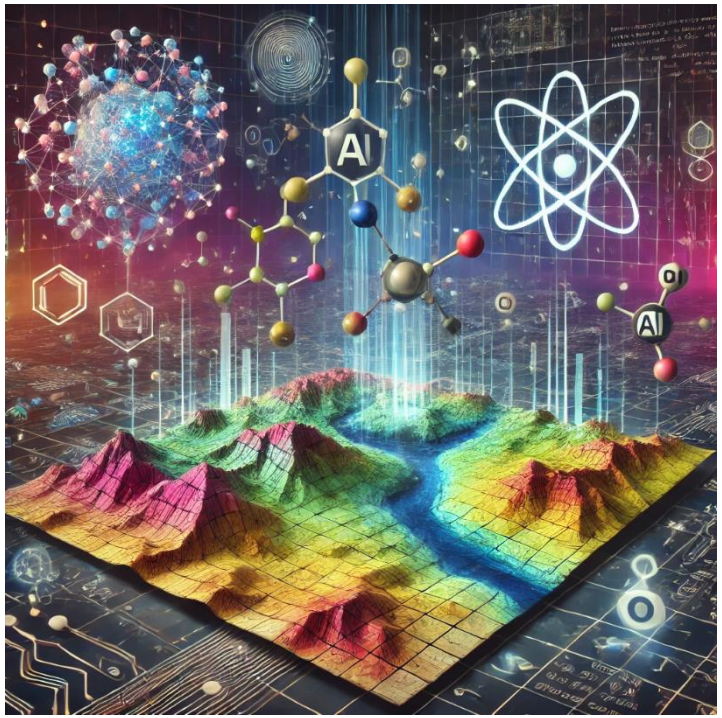
Which genes define resistance to penicillin?



Possible reason of the penicillin resistance is the presence of β-lactamase is some isolates

GTM landscapes links resistance prediction and associated genes

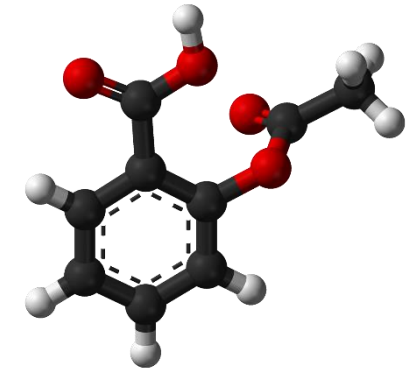
Pikalyova, K. et al. *Molecular Informatics*. 2024, 43:e202300263.



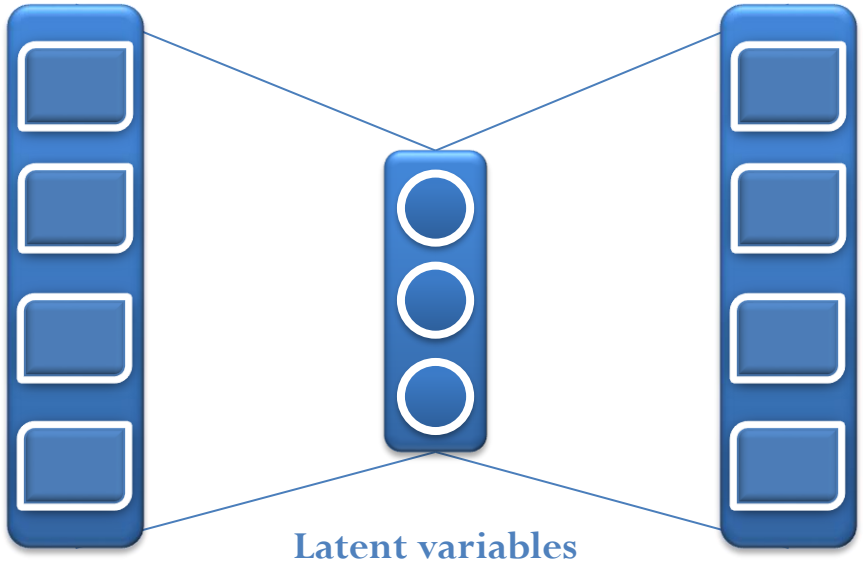
Chemography-guided generation of novel entities using AI tools

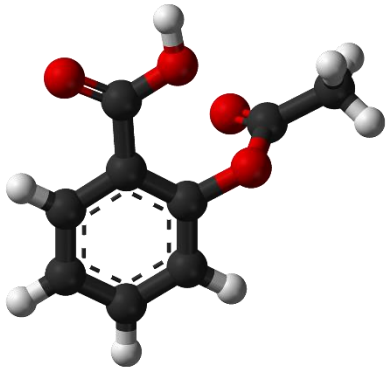
- Molecules
- Reactions
- Peptides

Autoencoder performing SMILES reconstruction



O=C(C)Oc1ccccc1C(=O)O

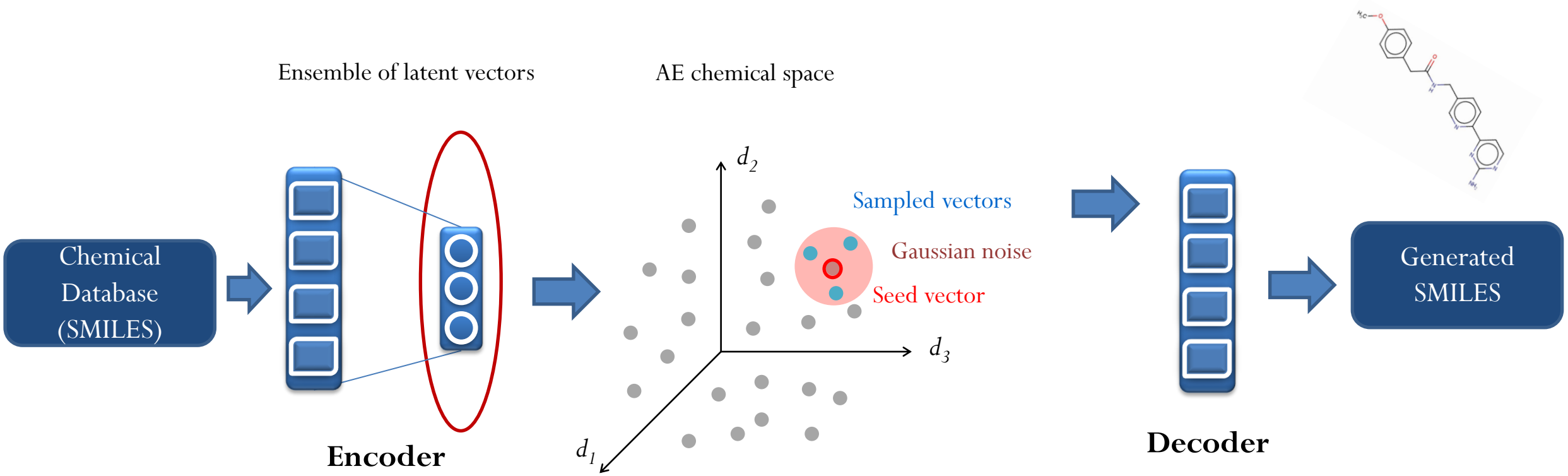




O=C(C)Oc1ccccc1C(=O)O

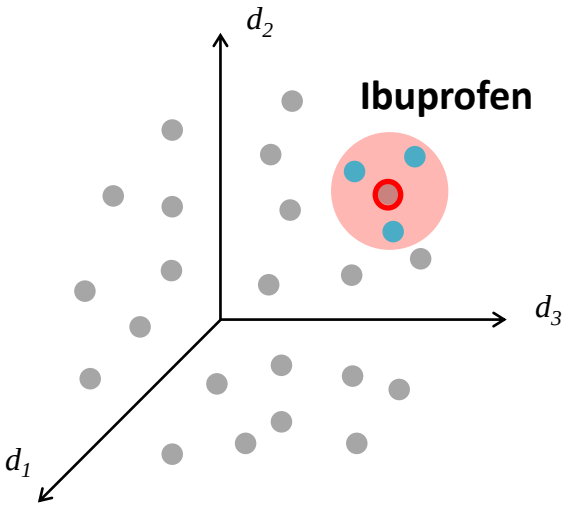
Chemical structure → **Real numbers** → **Chemical structure**

AutoEncoder: sampling using a seed vector

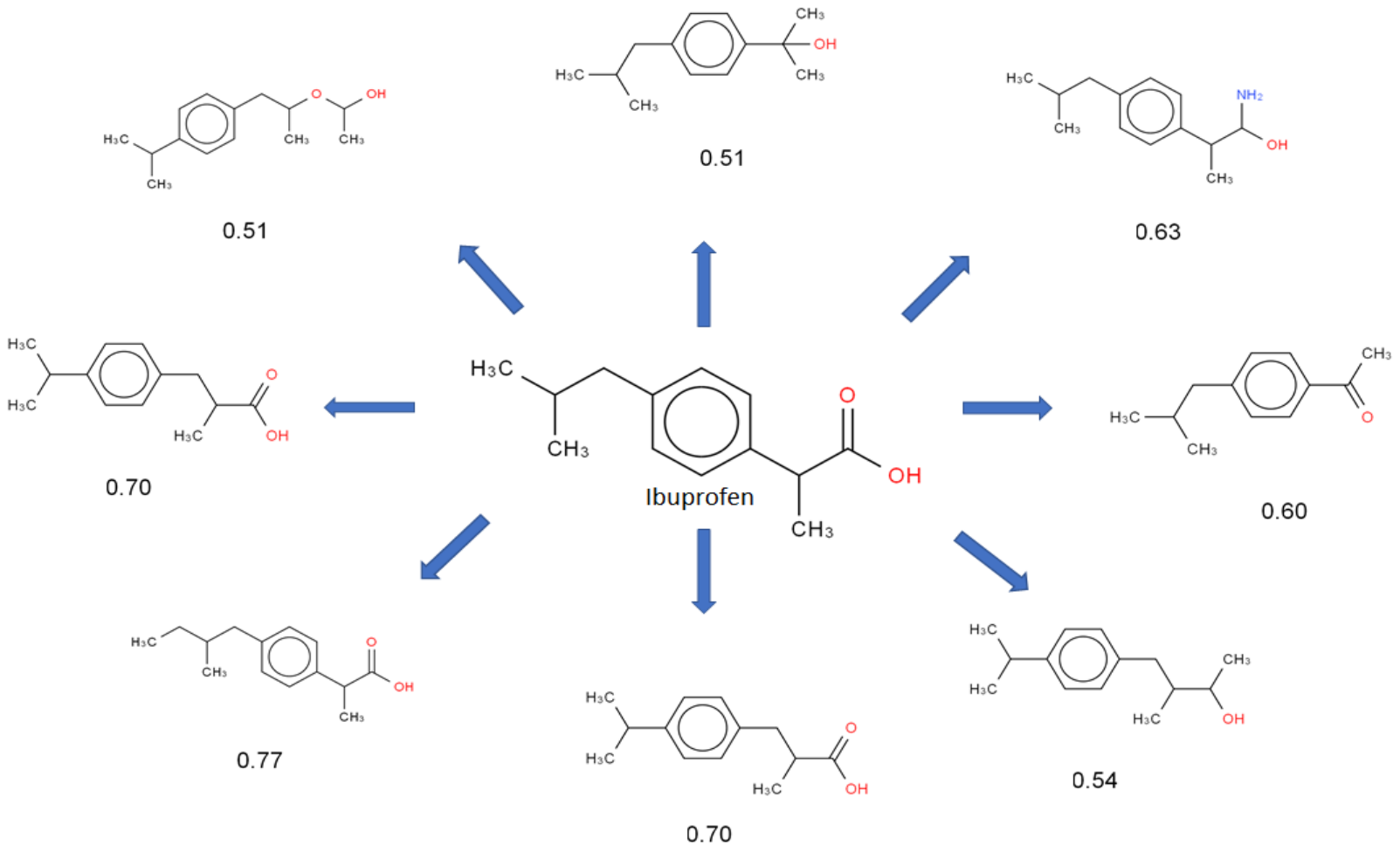


Goal: to identify a seed vector from which valid structures possessing a given activity can be generated

AutoEncoder: example of sampled structures

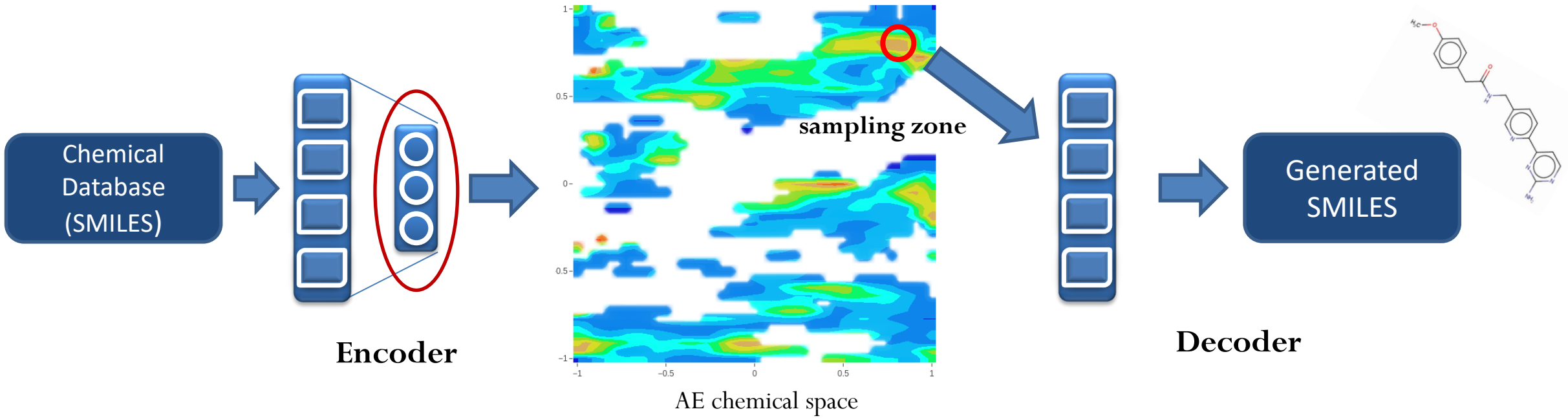


Decoder



The numbers correspond to the Tanimoto similarity

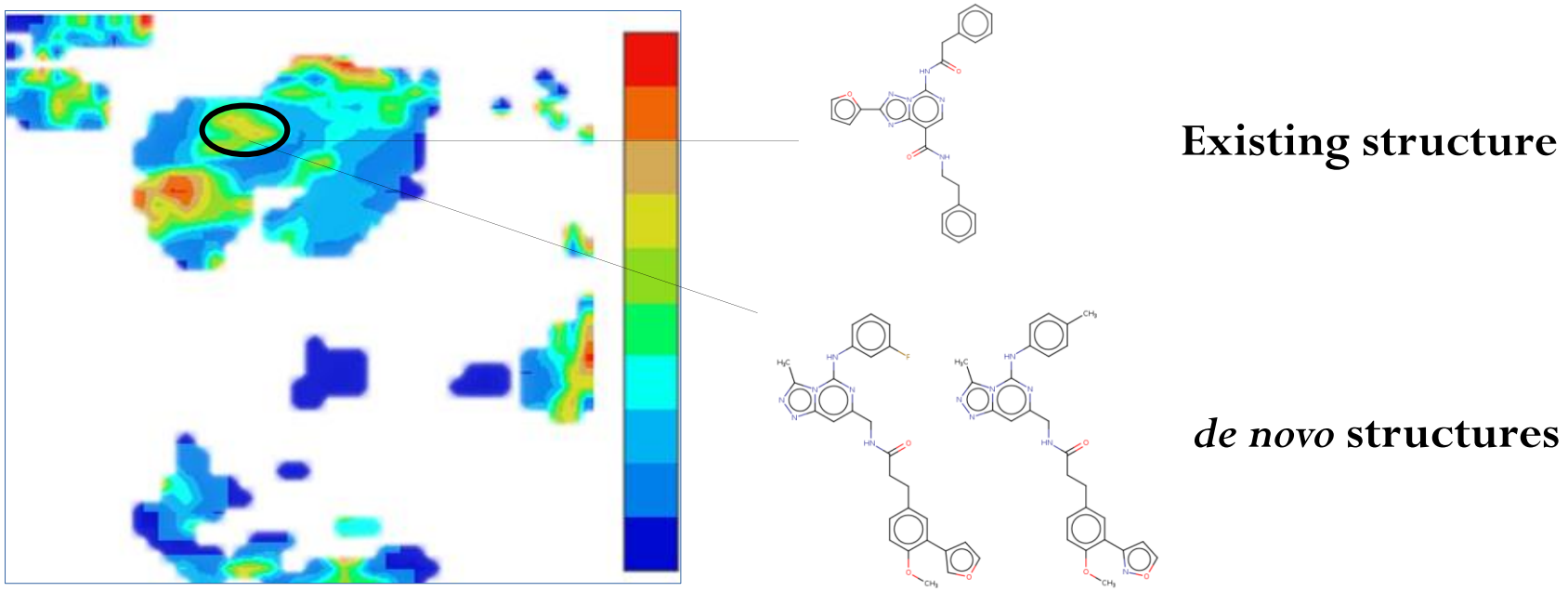
Chemography-guided molecule generation



GTM identifies a zone in the latent space from which "useful" structures are sampled. Such zone is detected either by the human or by computer algorithm

B. Sattarov et al. J. Chem. Inf. Model., 2019, 59(3), 1182-1196

Case study: Generation of inhibitors of A2a receptor



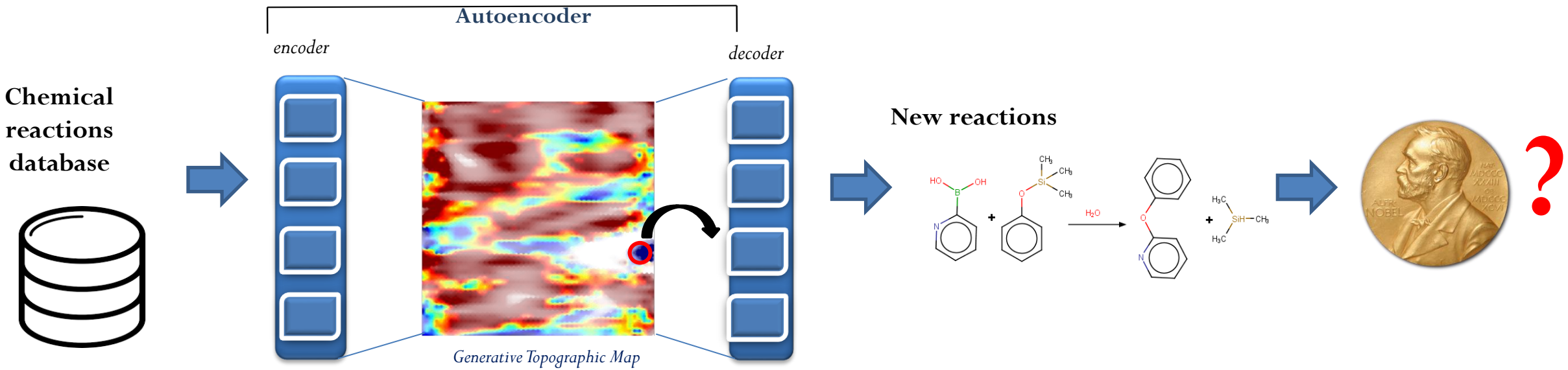
Existing structure

***de novo* structures**

- **Generated structures are enriched with new scaffolds**
- **According to docking experiments they are efficiently able to bind A2a**

B. Sattarov et al. J. Chem. Inf. Model., 2019, 59(3), 1182-1196

AI-driven design of new chemical transformations

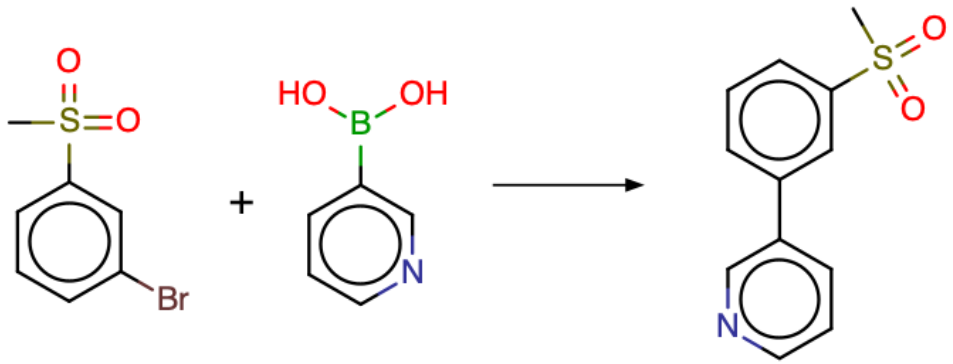


Reactions: methodological problems:

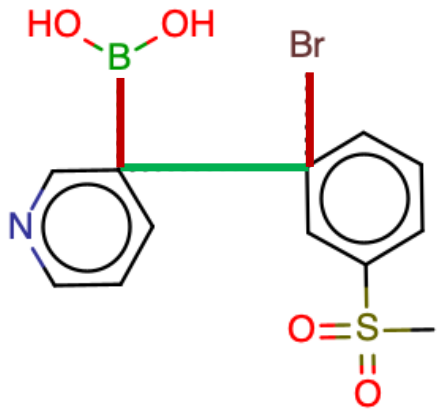
- Complexity: reaction equation contains several molecular graphs of two types – reactants and products
- Reaction novelty detection and reaction feasibility assessment are not well established

Complexity reduction: *Condensed Graph of Reaction (CGR)*

Reaction Equation



CGR



Dynamic bonds
(formed or broken)

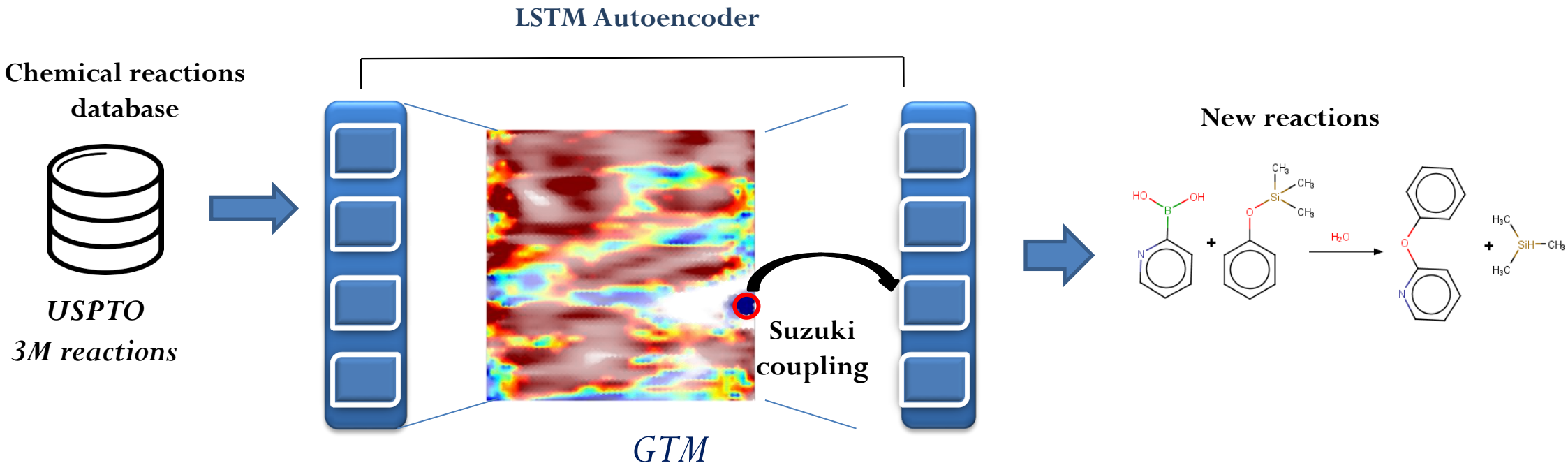
OB(O)C1:C:N:C:C:C1.O=S(=O)(C)C1:C:C(Br):C:C:C1>>O=S(=O)(C)C1:C:C:C:C(:C1)C2:C:N:C:C:C2

OB(O)[->.]C1(:C:N:C:C:C1)[.>-]C2([->.]Br):C:C:C:C(:C2)S(=O)(C)=O

CGR SMILES is almost twice shorter than conventional reaction SMILES

A. Varnek *et al.*, J. Computer-Aided Molecular Design, 2005, **19**, 693-703

AI-driven design of new Suzuki-like reactions

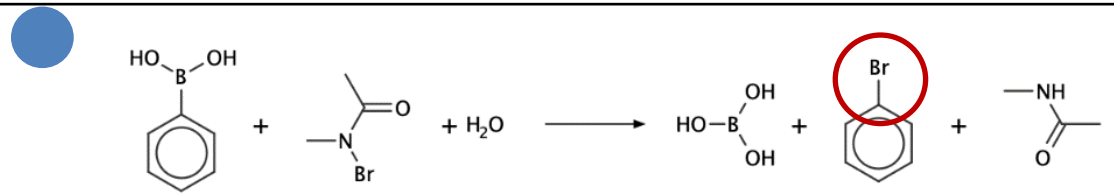
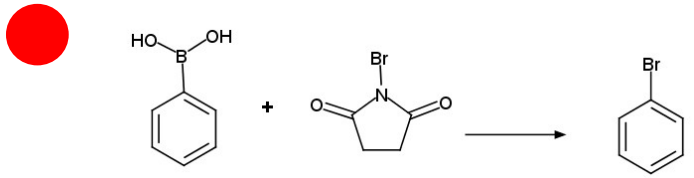
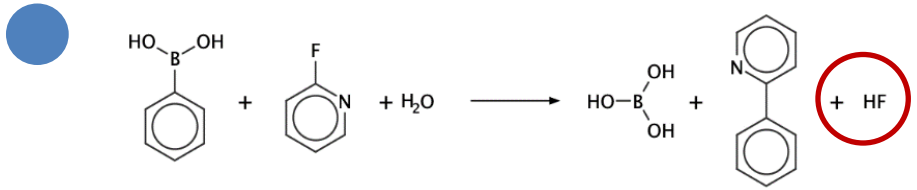
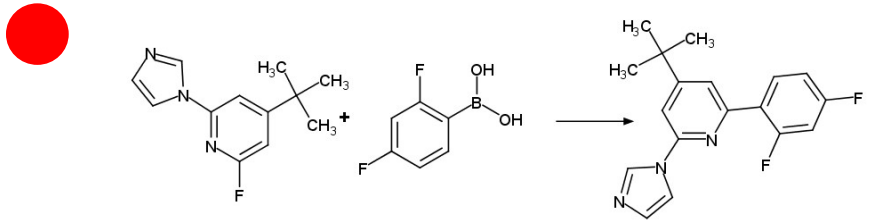


- 13 new (with respect to the training data) Suzuki-like reactions have been detected
- 5 of them have been found in SciFinder

W. Bort *et al.*, *Nature Scientific Reports*, 2021, **11**, 3178

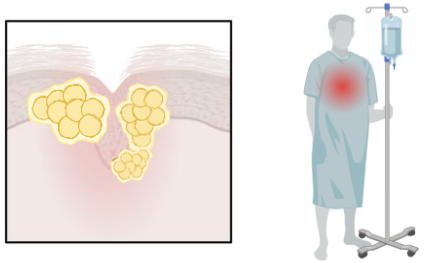
AI-driven design of reactions: experimental validation

- Reaction generated by AI
- Reaction found in SciFinder

coupling	Reaction	
C-Br		Exotic coupling
		
C-C		Exotic leaving group
		

W. Bort *et al.*, *Nature Scientific Reports,* 2021, **11**, 3178

In silico design of novel peptides against Methicillin Resistant *Staphylococcus Aureus* (MRSA)



Diverse range of infections
(from mild skin infections to life-threatening conditions, e.g. sepsis)

Duerden BI. *Eye (Lond)*. **2012**. 26.

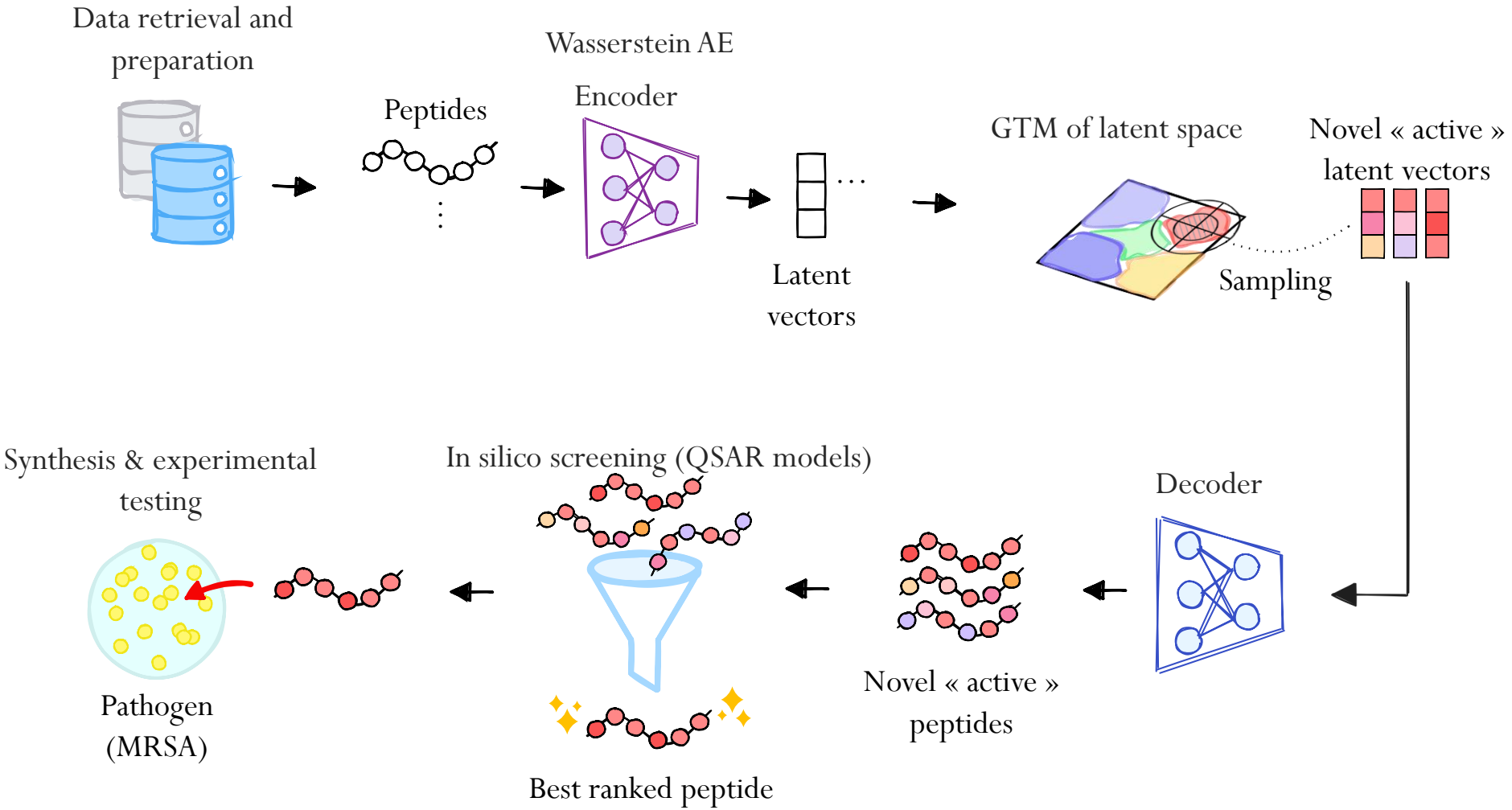
People with MRSA are 64% more likely to die as opposed to its non-resistant variant

World Health Organization (2018). *Antimicrobial resistance fact sheet*.

Formation of biofilms
(more resistant at conventional antibiotics dosages)

Choi, V. et al. *Nat Rev Microbiol* 21, 555–572 (**2023**).

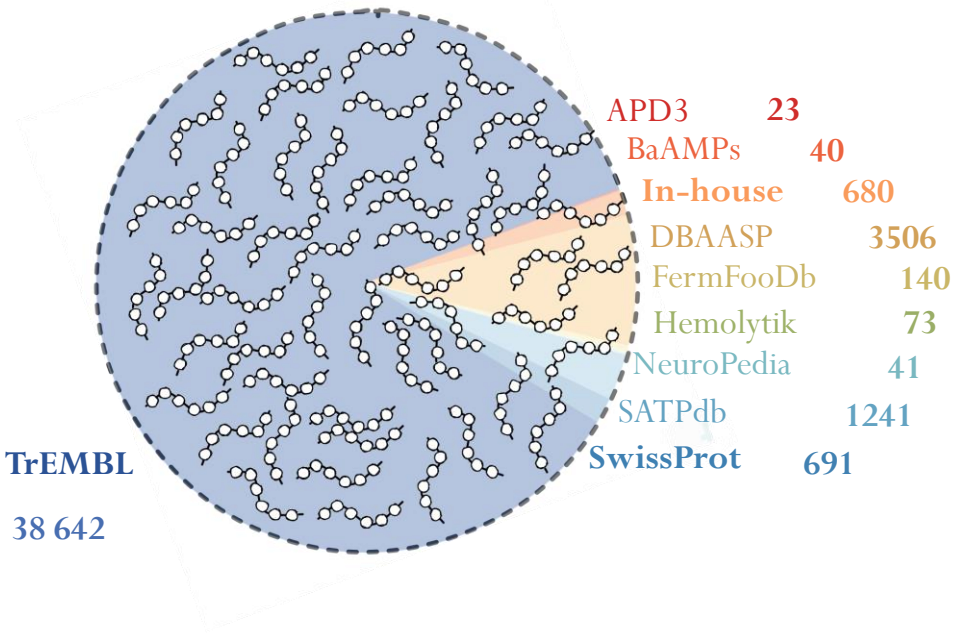
Design of antimicrobial peptides: generation workflow



Pikalyova K., et al., et al. BioRxiv , 2024, DOI: 10.1101/2024.11.17.622654

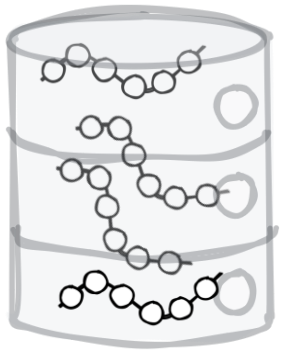
Training set composition

45 K of training peptides



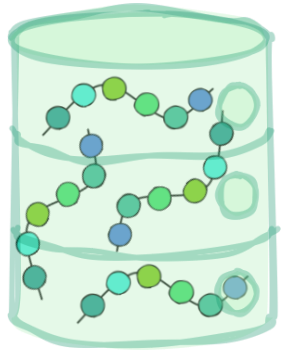
≈

In-house dataset with activity labels against MRSA (680 peptides)



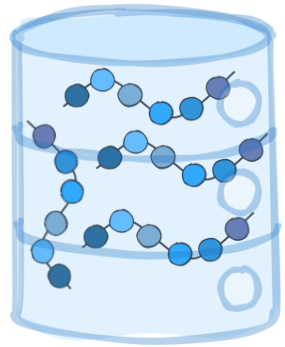
+

Bioactive peptides from public DBs



+

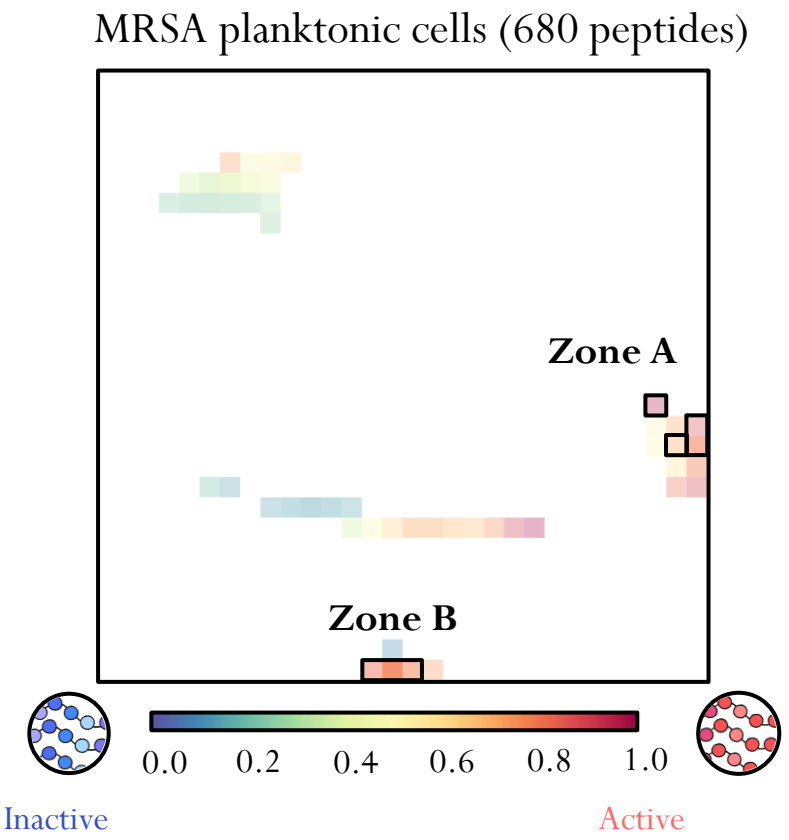
UniProt
(TrEMBL+SwissProt)

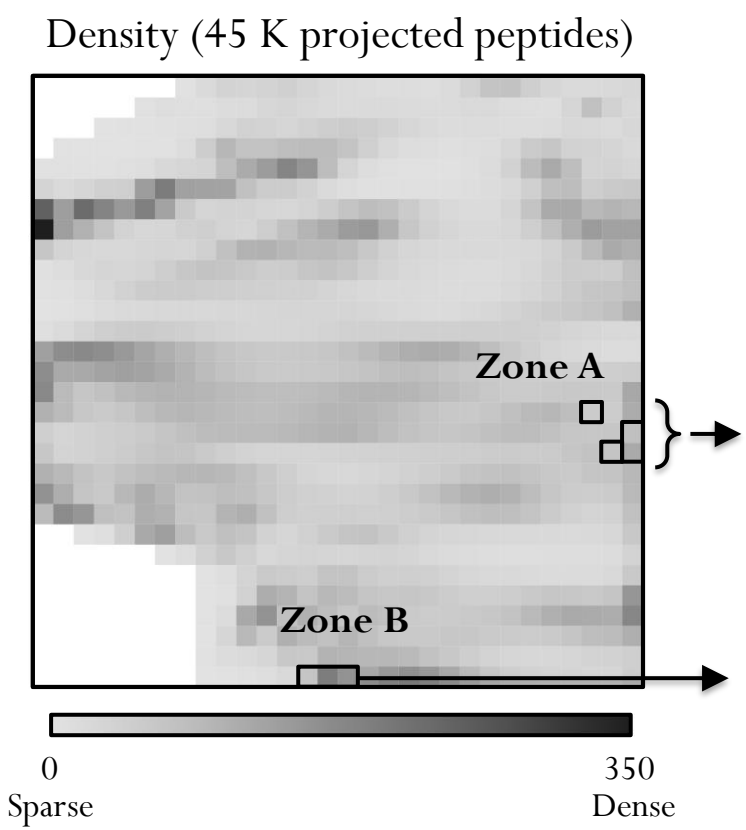


Inclusion criteria:

- 10-14 amino acids long
- Only natural amino acids
- Only linear peptides

Cartography-based design of novel peptides





Sampling

298.6 K peptides

Screening with QSAR models

1014 peptides

Selection for exp. tests

8 peptides

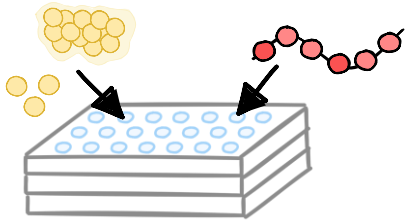
Experiments

Pikalyova K., et al., et al. BioRxiv , 2024, DOI: 10.1101/2024.11.17.622654

Experimental results

THE UNIVERSITY OF BRITISH COLUMBIA

Synthesis & experimental testing of peptides against MRSA biofilms and planktonic cells



Screening results

8/8 peptides more active against MRSA biofilms compared to control

5/8 peptides active against MRSA planktonic cells compared to control

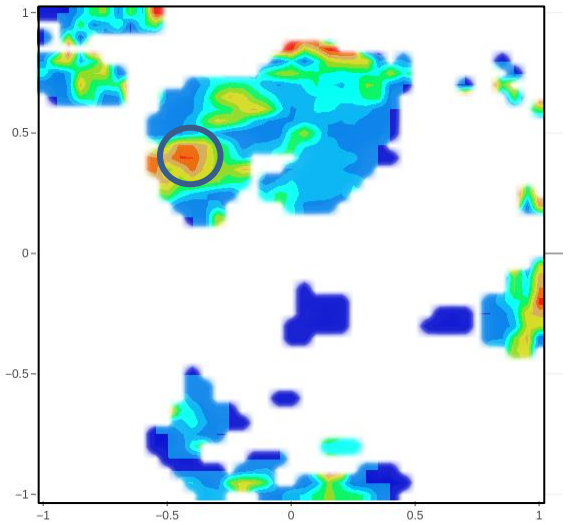
Most active peptide is 10-fold more active against MRSA biofilms well-studied control peptide IDR-1018

Control

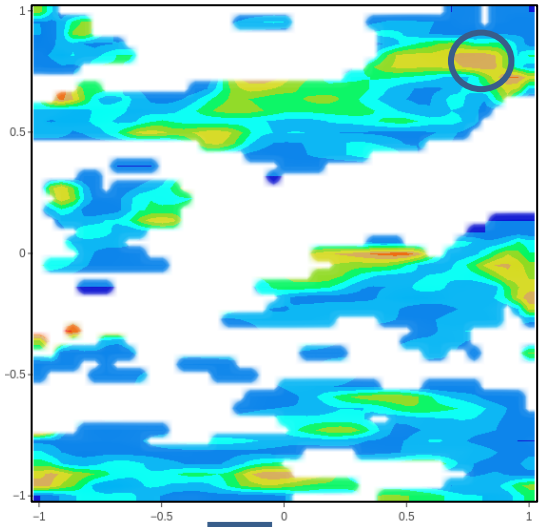
Pikalyova K., et al., et al. BioRxiv , 2024, DOI: 10.1101/2024.11.17.622654

AutoEncoder vs Molecular descriptors space

Molecular descriptors space



Autoencoder latent space



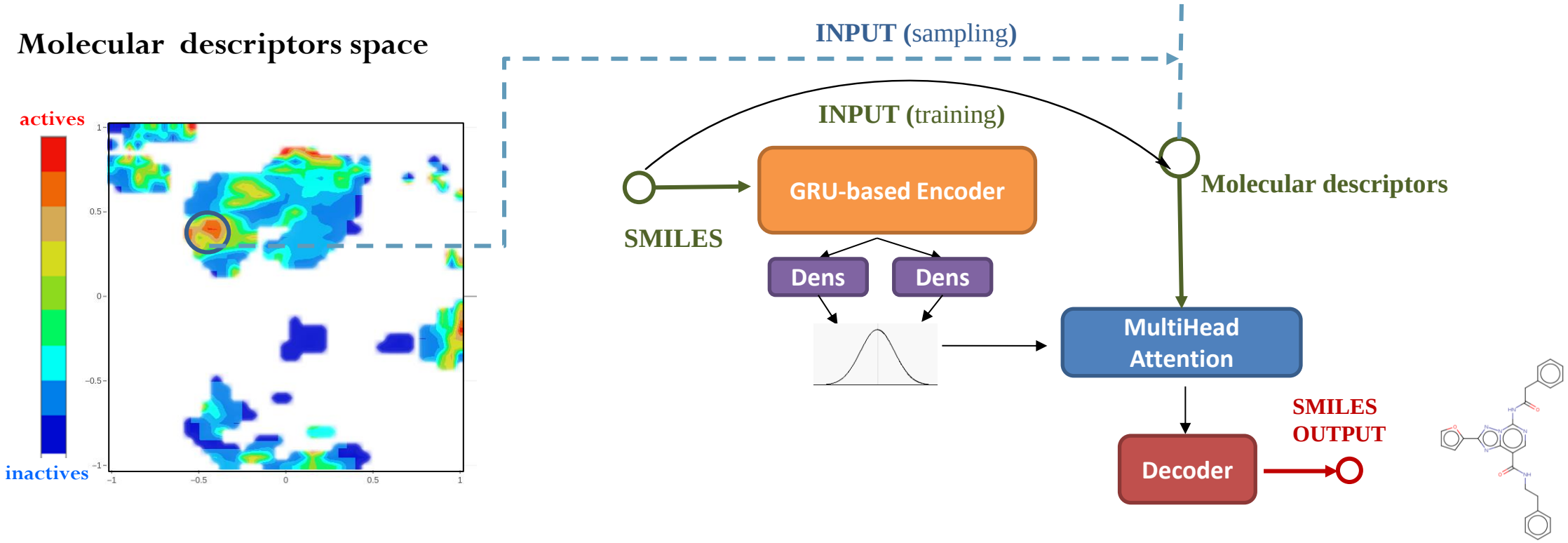
GTM Class landscapes for A2a-receptors binders (1303 actives and 3618 inactives)

- activity prediction
- no structures generation

- activity prediction
- generation of new structures

Goal: development of deep-learning architecture able to generate structures with desired activities using *any* descriptor space

Sampling from any descriptor space: *inverse-QSAR task*



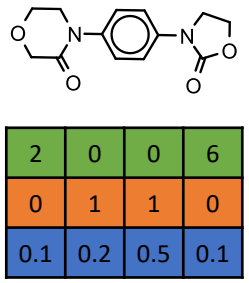
Attention-based Conditional Variational Autoencoder (ACoVAE) - deep-learning architecture able to generate structures from any descriptor space

W. Bort *et al.* *Chem. Inf. Model.*, 2022, **62**, 22, 5471–5484

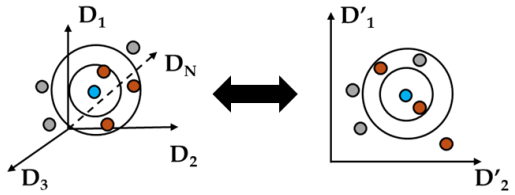

GTM
T-SNE
U-MAP
Why GTM ?
UMAP
PCA

Neighborhood preservation benchmark: GTM *vs* t-SNE, UMAP and PCA

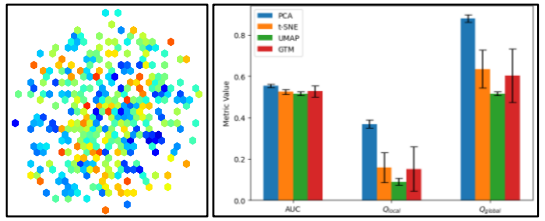
Congeneric series
103 subsets extracted from ChEMBL



Descriptor sets
Morgan fingerprints, MACCS keys, ChemDist



Dimensionality reduction
optimization of hyperparameters for neighborhood preservation



Comparative analysis
of neighborhood preservation according to 18 metrics

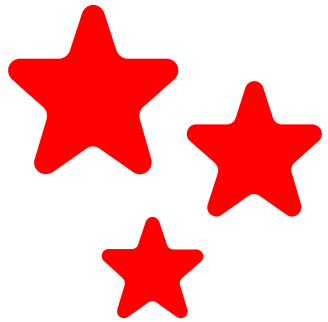
	PCA	t-SNE	UMAP	GTM
Neighborhood preservation (NP)	☒	☑	☑	☑
Out-of-sample NP	☑	☒	☒	☑

A. Orlov et al. *Molecular Informatics*, 2024, https://doi.org/10.1002/minf.202400265

Why GTM ?

GTM considers data probability distribution which makes it suitable for Big Data analysis

GTM manifold can be built on a small representative subset and it can accommodate new data



GTM landscapes can be used in regression or classification tasks

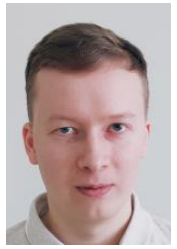
A data collection can be encoded by a vector which enables fast chemical libraries' comparison

Coupling of GTM with the AI technologies facilitates *de novo design*

Teams

Chém Informatique
STRASBOURG

ICReDD
Institute for Chemical Reaction Design and Discovery
HOKKAIDO UNIVERSITY

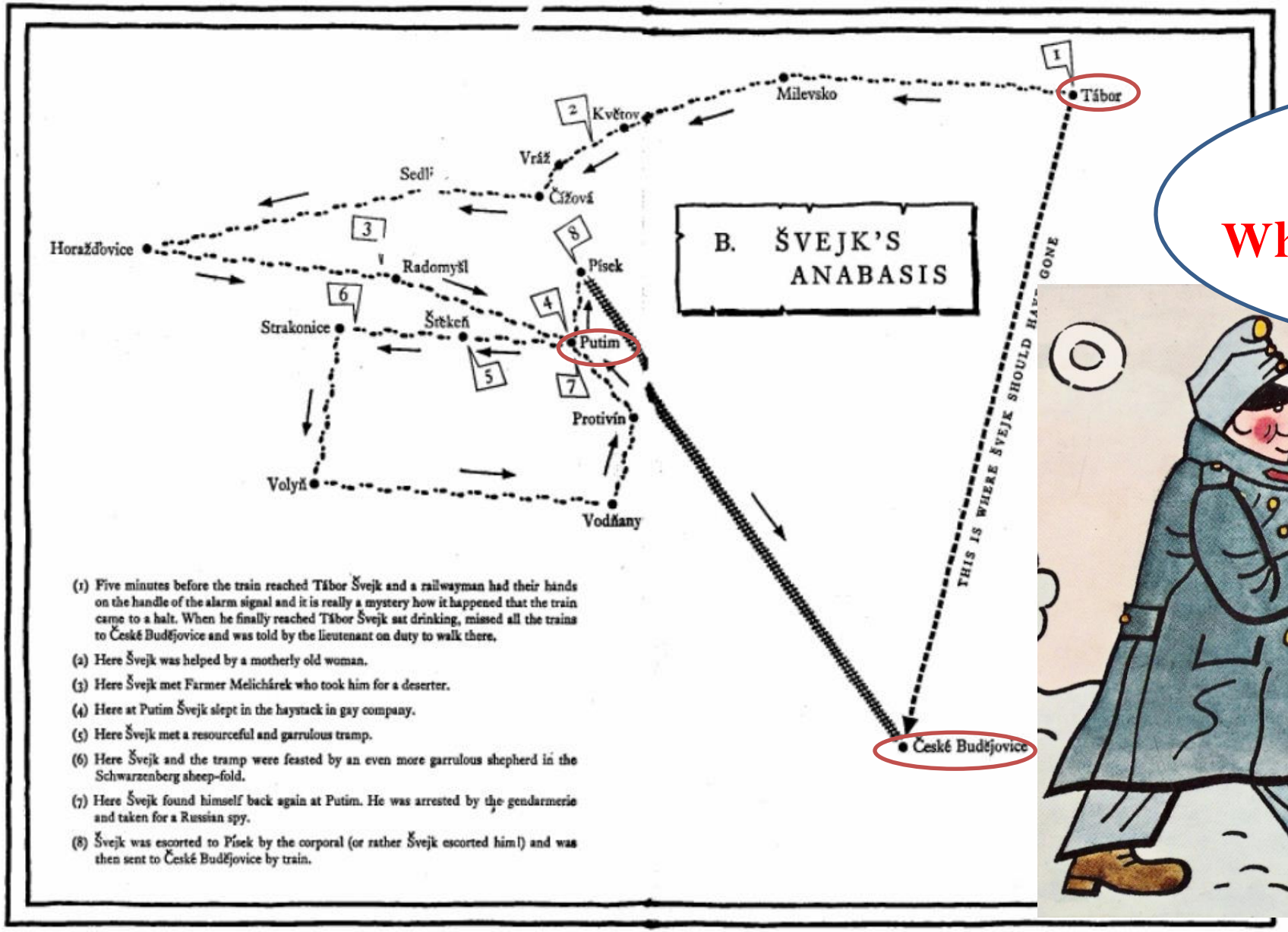


Collaboration

- MSC-DN Marie-Curie *AiChemist*
- ITN Marie-Curie *BigChem*
- ITN Marie Curie *TubInTrain*
- Institute of Organic Chemistry, Kiev, Ukraine

- Eli Lilly
- SANOFI
- Enamine
- eMolecules
- Novalix
- Janssen Pharmaceutical
- TotalEnergy
- SOLVAY

Cartography is an efficient way to explore a (chemical) space



Why didn't I use GTM ???