Université de Strasbourg · Chémoinformatique STRASBOURG · cnrs

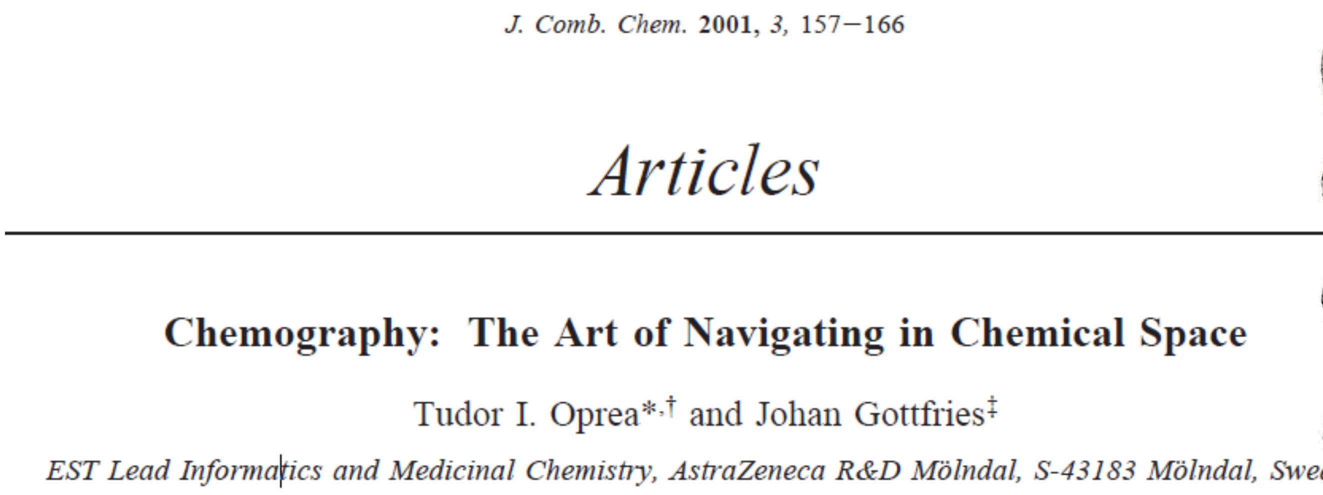
Chemography applications to drug design: from (ultra)large libraries analysis to de novo design of molecules and reactions

Dragos Horvath
Gilles Marcou
Fanny Bonachera
Alexandre Varnek
University of Strasbourg

Novartis, 9th of February 2023

Chemography

- ~10^9 compounds are physically available
- < 10^{26} structures are stored in proprietary DBs
- ~10^{33} drug-like molecules could be synthesized *



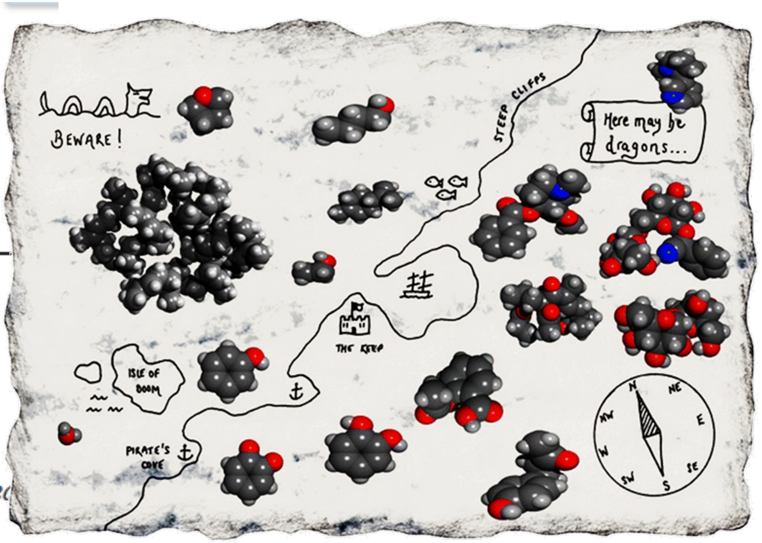
J. Comb. Chem. **2001,** *3,* 157−166

Articles

Chemography: The Art of Navigating in Chemical Space

Tudor I. Oprea[*,†] and Johan Gottfries[‡]

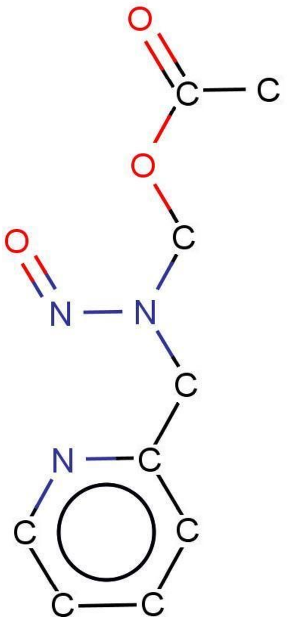
EST Lead Informatics and Medicinal Chemistry, AstraZeneca R&D Mölndal, S-43183 Mölndal, Swed



* *P. Polischuk, T. Madzidov , A. Varnek, J. Comp. Aided Mol, Des. 2013,* 27, p. 675-679

Encoding chemical structures by molecular descriptors

Molecular graph



Descriptors

Constitutional descriptors
Ring descriptors
Topological indices
Walk and path counts
Connectivity indices
Information indices
2D matrix-based descriptors
2D autocorrelations
Burden eigenvalues
P_VSA-like descriptors
ETA indices
Edge adjacency indices
Geometrical descriptors
3D matrix-based descriptors
3D autocorrelations

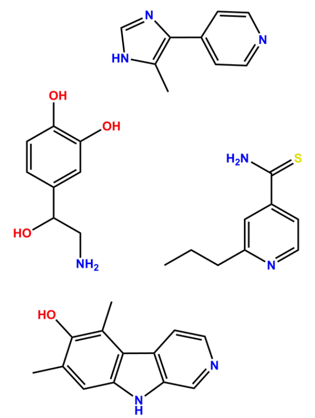
....

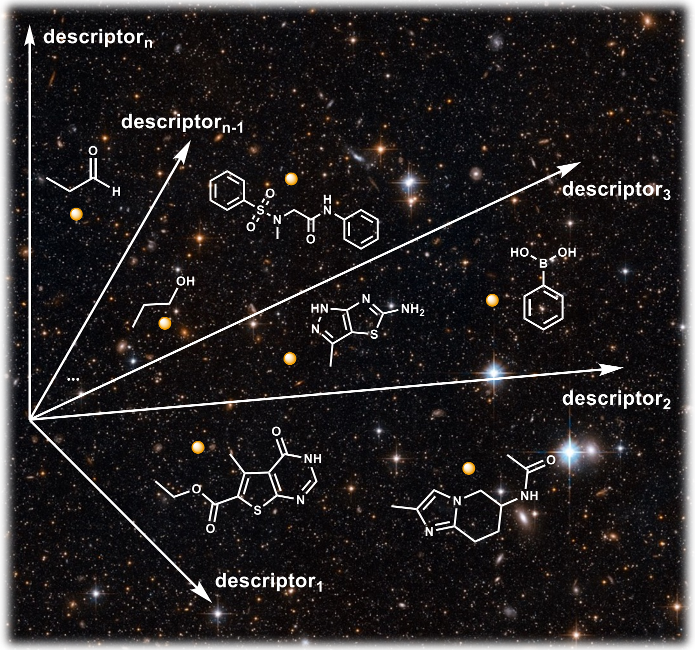
Descriptor vector

Descriptor	Value
D_1	a_1
D_2	a_2
....	...
D_i	a_i
....	...

> 5000 types of descriptors are used

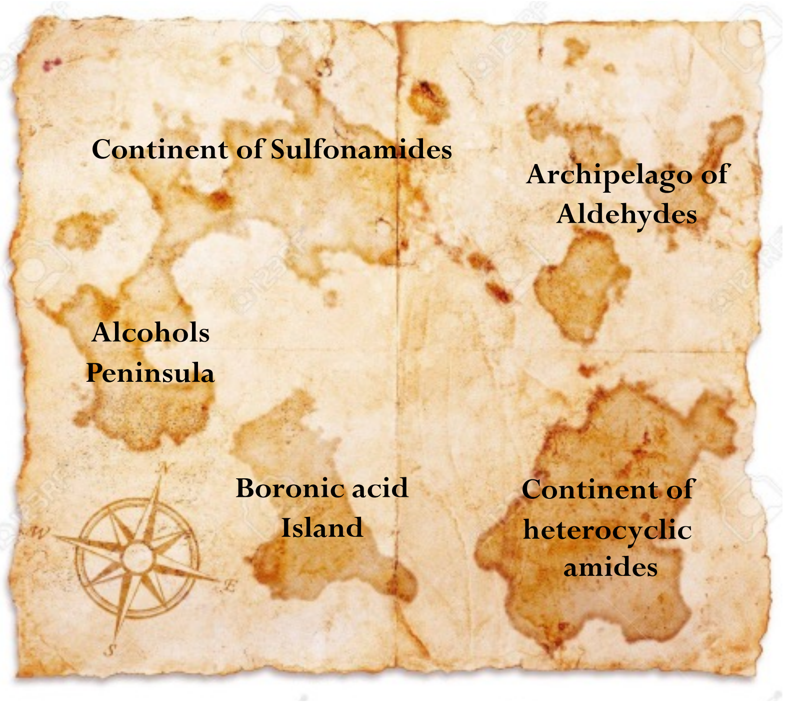
Data visualization: dimensionality reduction problem





Initial space

(N-dimensional)

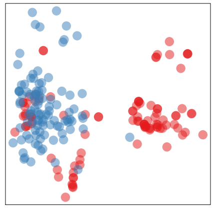


Latent space

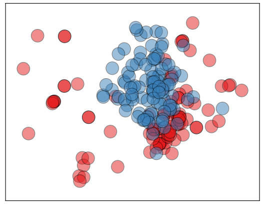
(2-dimensional)

Dimensionality reduction methods

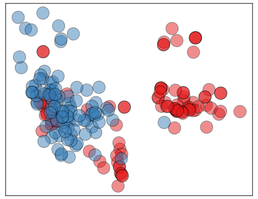
Acetylcholinesterase dataset (DUD) : 100 actives and 100 inactives



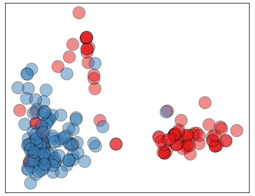
Multi-Dimensional Scaling



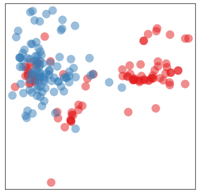
Canonical Correlation Analysis



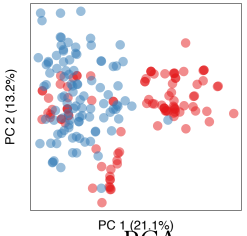
Independent Component Analysis



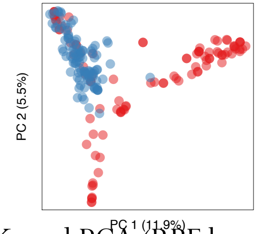
Exploratory Factor Analysis



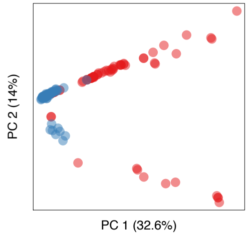
Sammon map



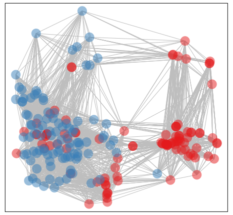
PCA



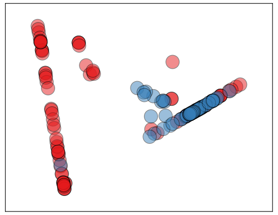
Kernel PCA (RBF kernel)



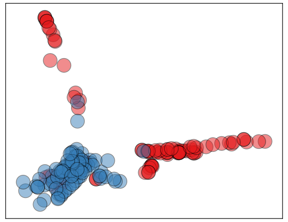
Kernel PCA (polynomial kernel)



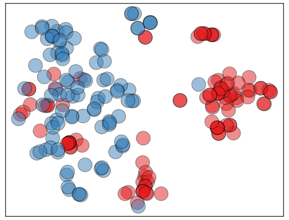
Isomap



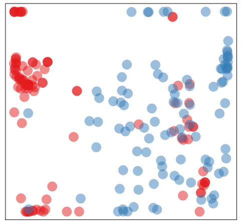
Locally Linear Embedding



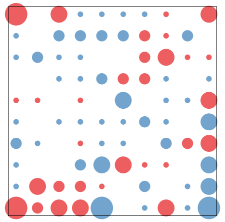
Laplacian Eigenmaps



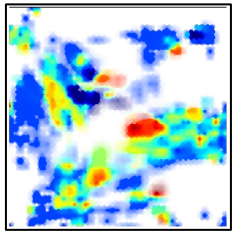
t-SNE



Autoencoder dimensionality reduction



SOM



GTM

Generative Topographic Mapping : areas of application

Conformational space analysis

Ligand to Protein docking

Sequence space analysis

Drugs repurposing

Library design

Data visualisation and analysis

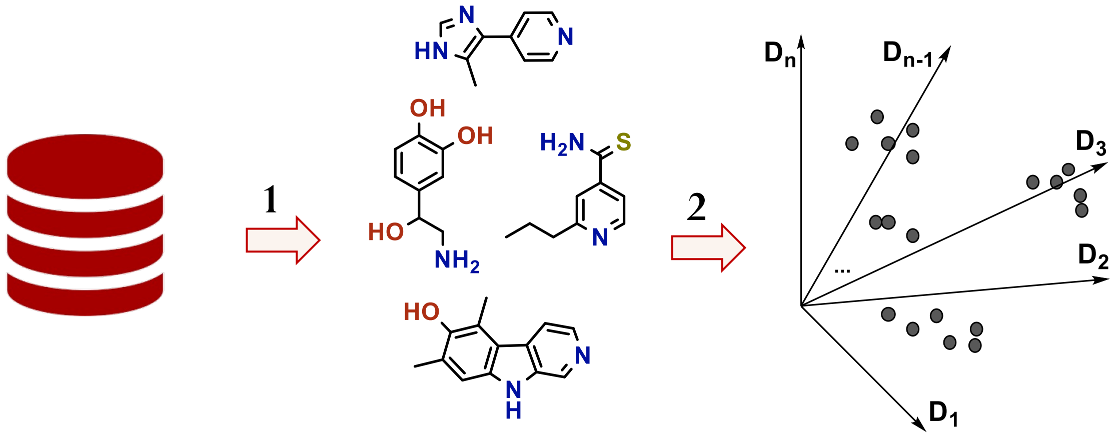
Library comparison

Structure-Activity modeling

Virtual screening

de novo design

Generative Topographic Mapping (GTM)



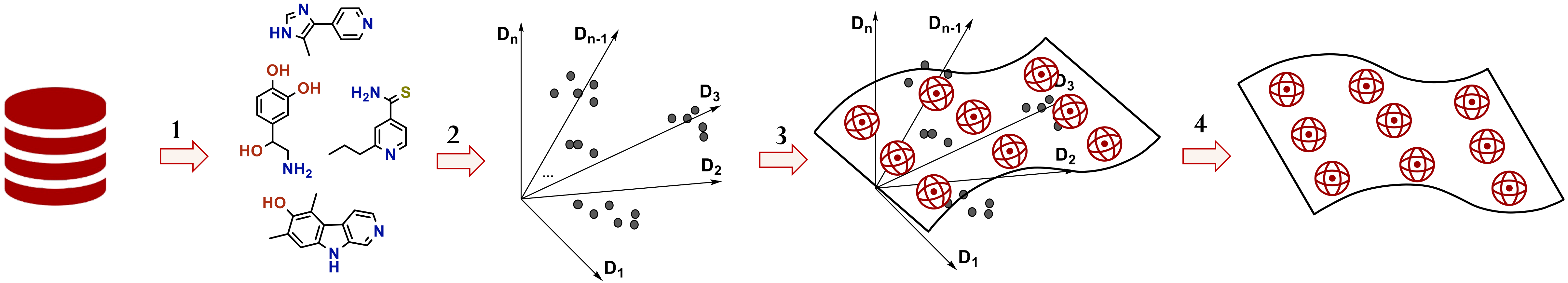
Chemical library

Representative frame set

***ISIDA* Descriptor Space**

1. Frame set selection
2. Molecules are represented in *n*-dimensional descriptor space

Generative Topographic Mapping (GTM)



Chemical library

Representative Frame set

***ISIDA* Descriptor Space**

Manifold training

Trained manifold

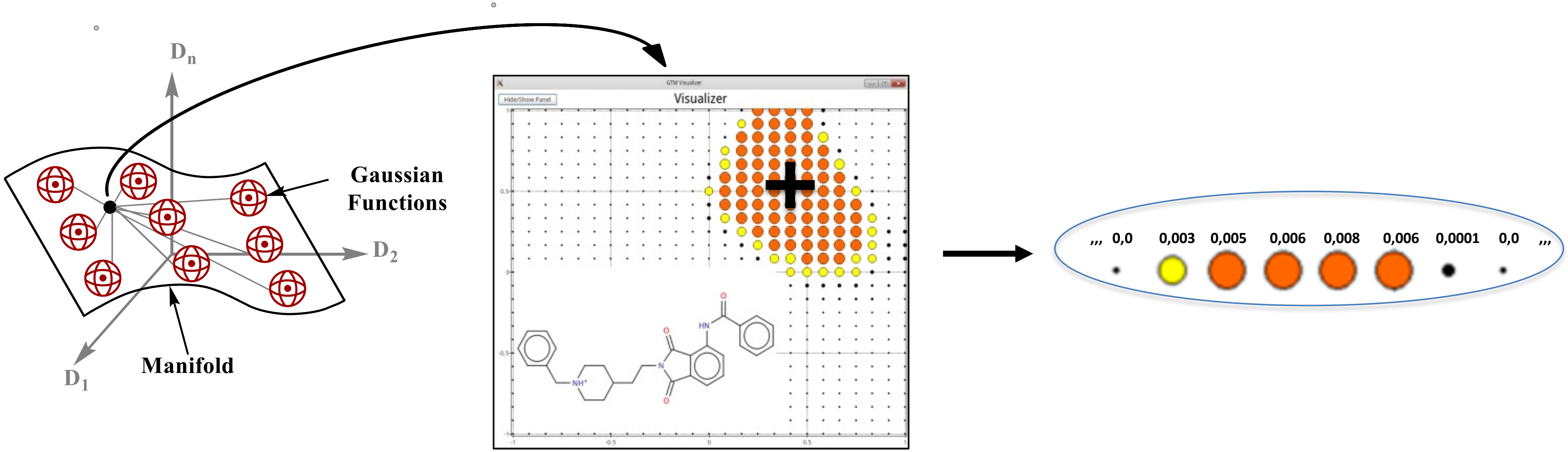
1. Frame set selection
2. Molecules are represented in *n*-dimensional descriptor space
3. A flexible 2D ***manifold*** is fitted to the data
4. Coordinates of the manifold are saved

Fuzzy nature of GTM

- GTM generates a data “probability” distribution in both initial and latent data spaces.
- *Initial space* : ensemble of Gaussian functions situated in the nodes of a grid superposed with the manifold
- *Latent space*: fuzzy projection on the nodes of flattened grid

Initial (descriptor) Space

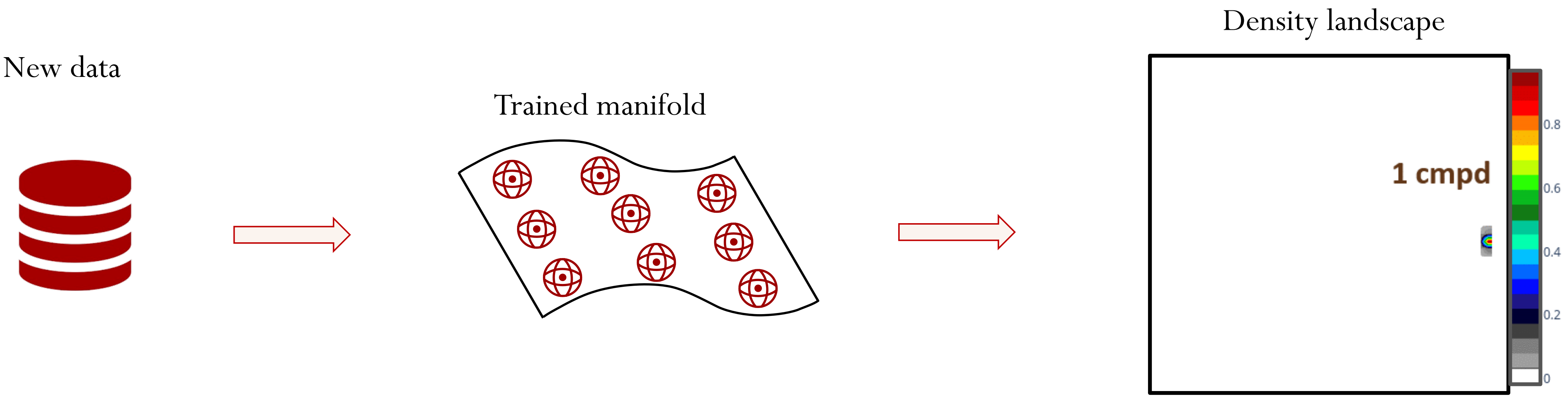
Latent space



Molecule → Responsibility (node residence « time ») vector of dimension N_{nodes}

Density landscapes

- display the compounds distribution in the chemical space
- spotting the regions that are under or overpopulated



New data

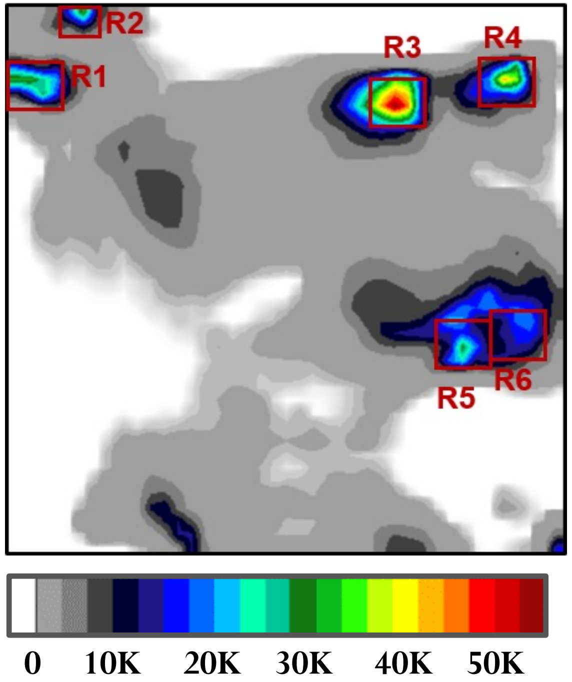
Trained manifold

Density landscape

Library → Cumulated Responsibility vector

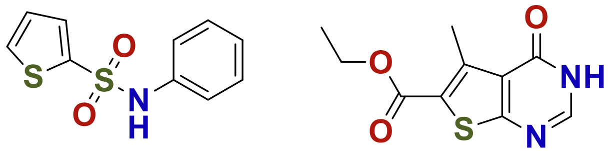
Chemical analysis with density maps

Every populated zone can be associated with some "chemotype".

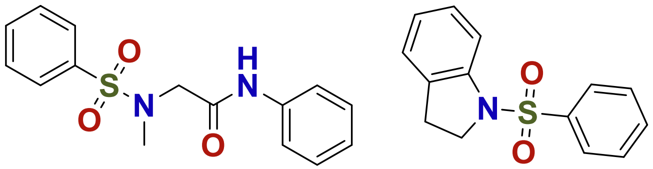


Lead-Like ZINC-In-Stock
(3.2M cmpds)

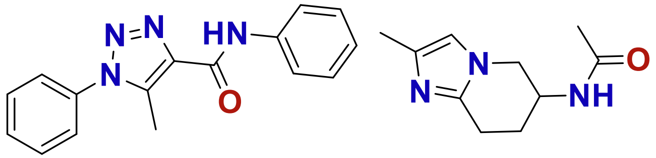
R1: Thiophenes



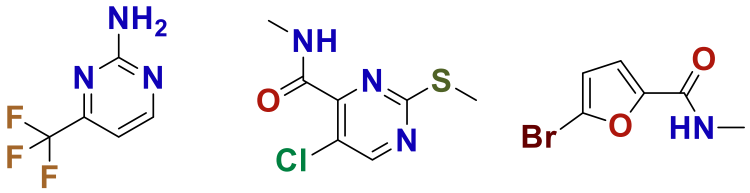
R4: Benzensulphonamides



R3: Heterocyclic amides



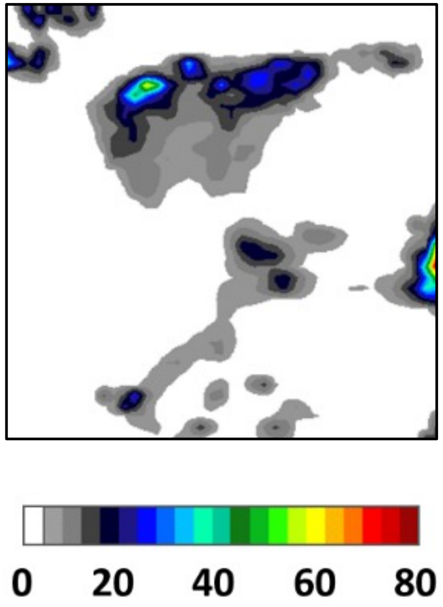
R6: Halloginated heterocycles



Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Volochnyuk, and A. Varnek, JCIM 2021 61 (1), 179-188

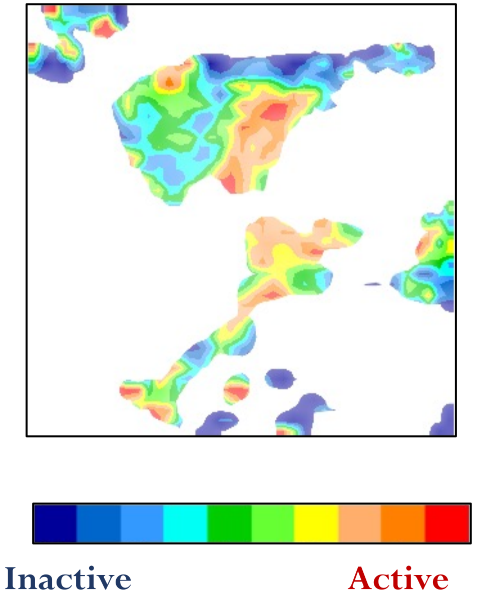
GTM Landscapes

Density landscape



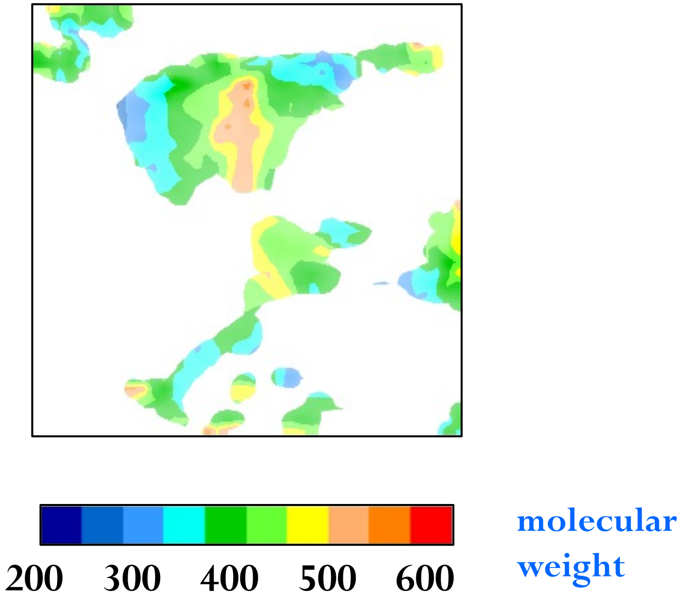
Colored according to the cumulated responsibilities in each node

Class landscape



Colored according to the resident class ratio weighted by responsibility

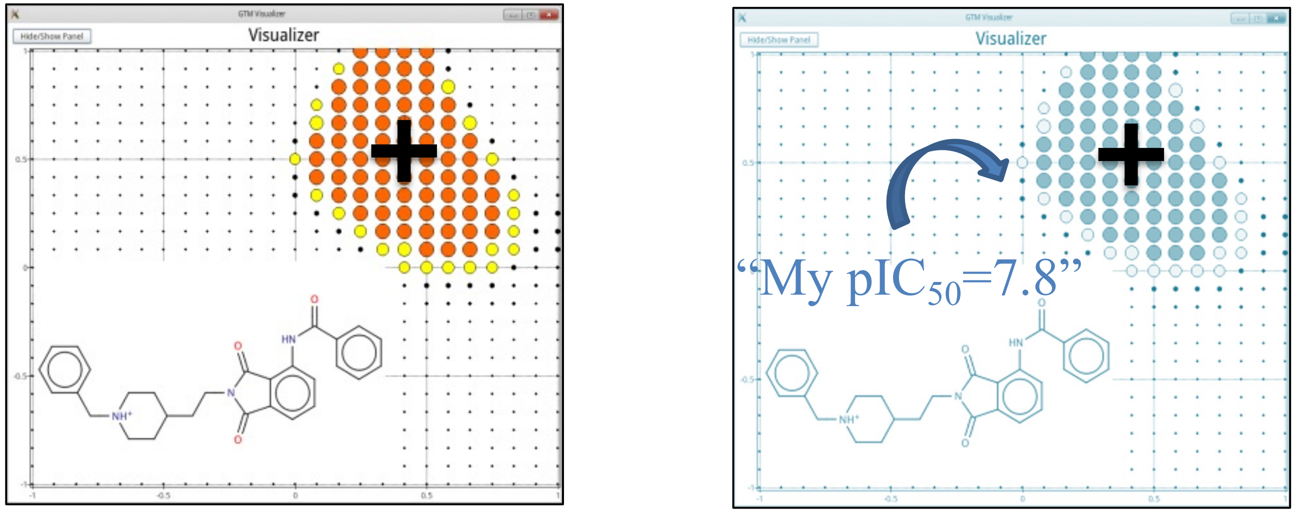
Property (activity) landscape



Colored according to the weighted average of selected activity (property)

GTM Nodes act as Knowledge Repositories…

Increment “node pIC_{50}” by R_n×7.8



“My pIC_{50}=7.8”

… and, after all “training” compounds contributed their increments, **normalize** node values by the total cumulated responsibility in there!

$$\langle P\rangle_n = \frac{\sum_m P_m R_n(m)}{\sum_m R_n(m)}$$

May want to weigh by node trustworthiness!

- Low-density nodes are not trustworthy!
- Mixed nodes (harboring residents with widely diverging properties) are not trustworthy!

for prediction, copy from node back to molecule: $P_{pred}(m') = \frac{\sum_n \langle P\rangle_n R_n(m')}{\sum_n R_n(m')} = \sum_n \langle P\rangle_n R_n(m')$

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

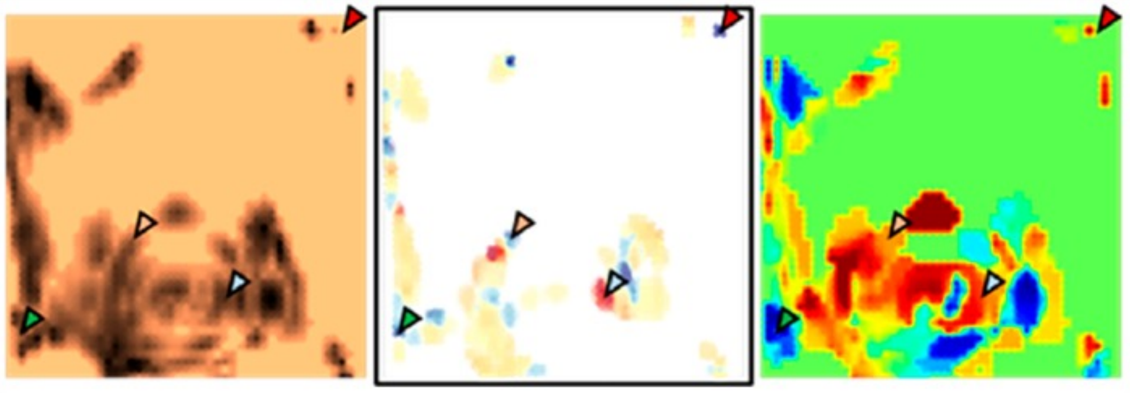
Trustworthiness, the Key to Grid-Based Map-Driven Predictive Model Enhancement and Applicability Domain Control

Dragos Horvath,* Gilles Marcou, and Alexandre Varnek

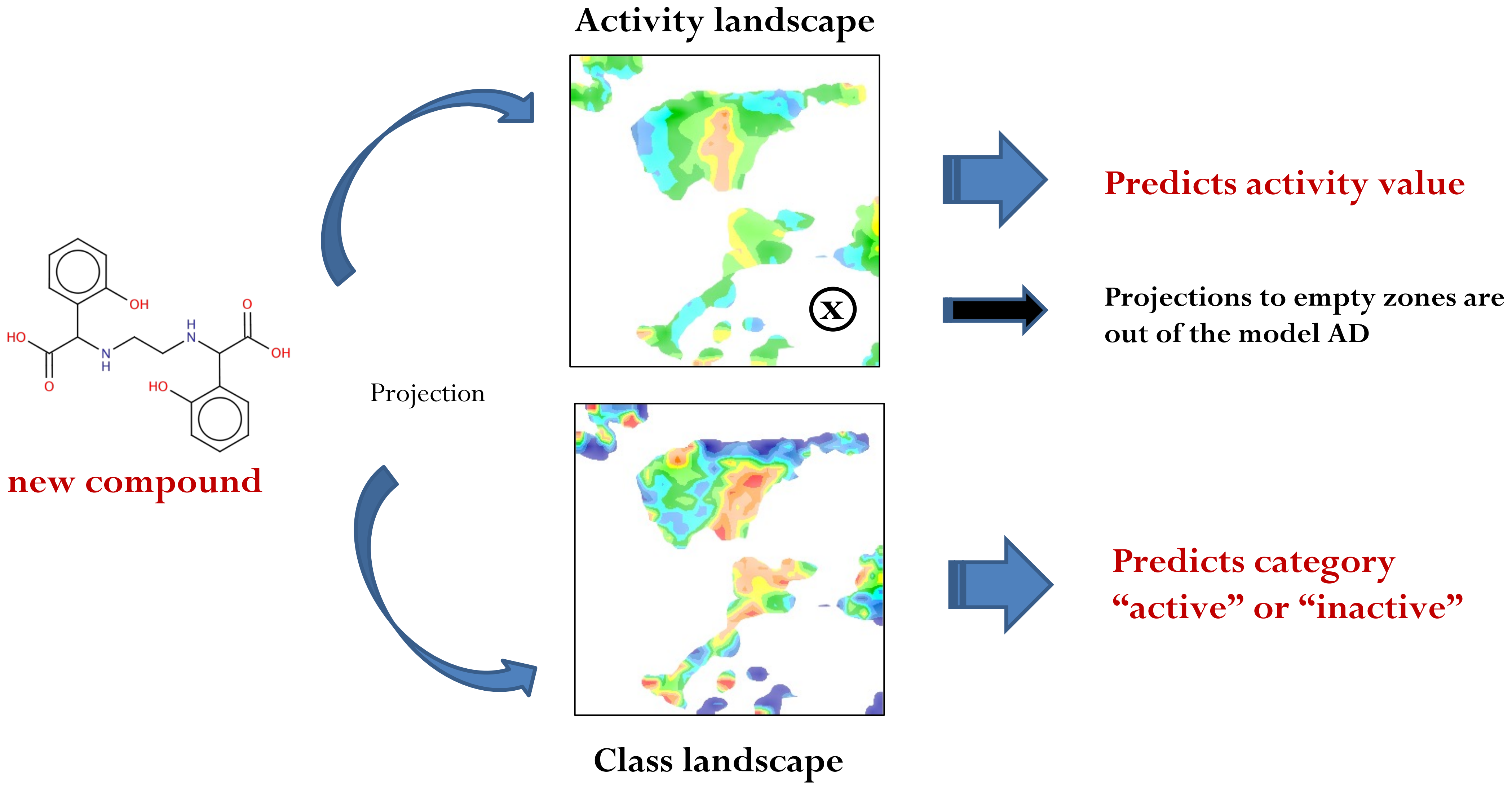
Cite This: https://dx.doi.org/10.1021/acs.jcim.0c00998 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In chemography, grid-based maps sample molecular descriptor space by injecting a set of nodes, and then linking them to some regular 2D grid representing the map. They include self-organizing maps (SOMs) and generative topographic maps (GTMs). Grid-based maps are predictive because any compound thereupon projected can "inherit" the properties of its residence node(s)—node properties themselves "inherited" from node-neighboring training set compounds. This Article proposes a formalism to define the trustworthiness of these nodes as



GTM Landscapes as predictive models



new compound

Projection

Activity landscape

Class landscape

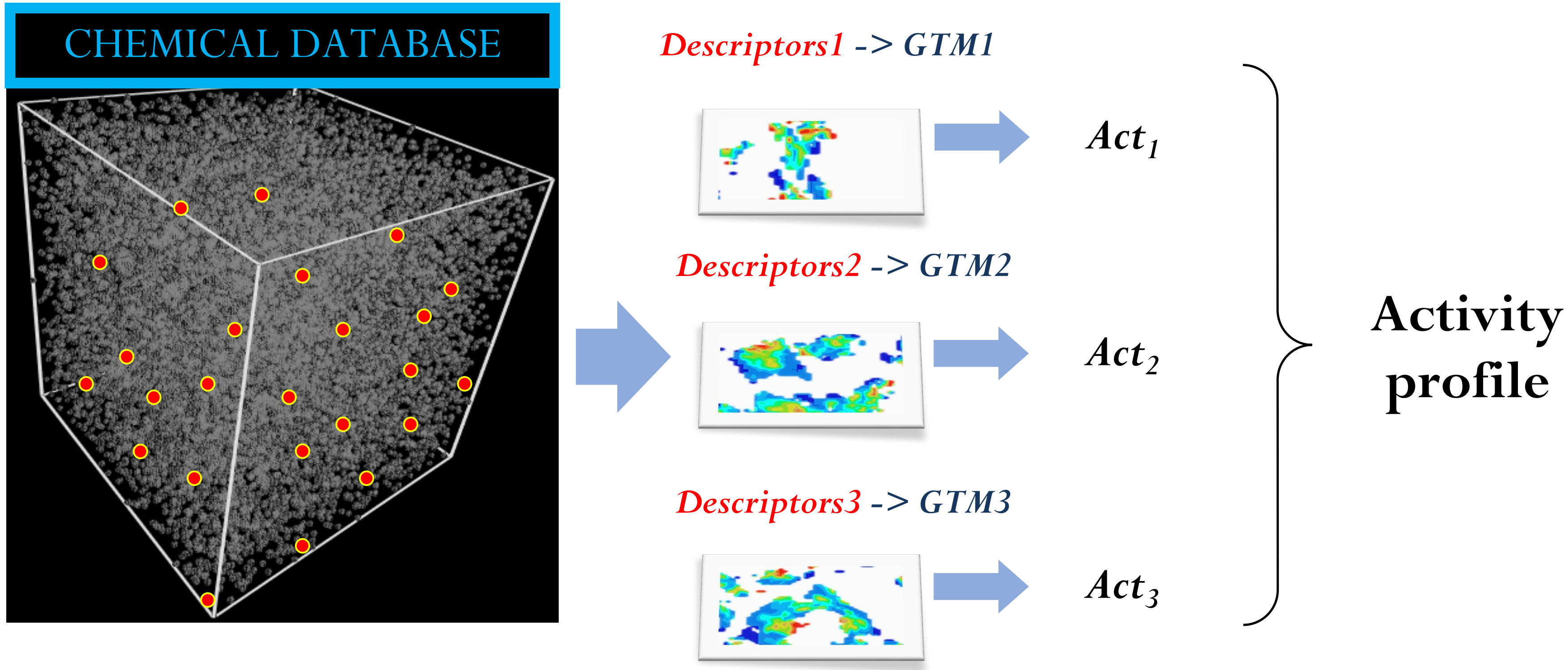
Predicts activity value

Projections to empty zones are out of the model AD

Predicts category “active” or “inactive”

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

GTM-based pharmacological profiling: single-task mode



CHEMICAL DATABASE

Descriptors1 -> GTM1

Act₁

Descriptors2 -> GTM2

Act₂

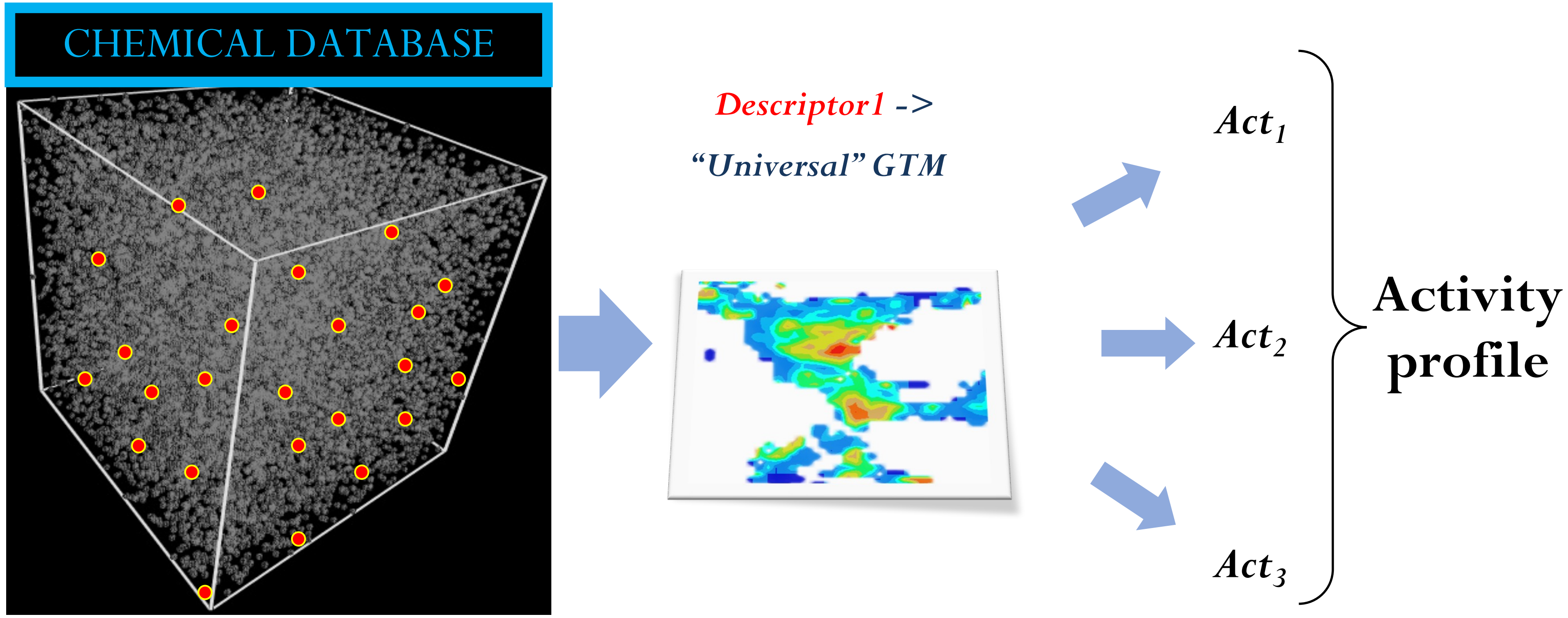
Descriptors3 -> GTM3

Act₃

Activity profile

Each GTM$_i$ predicts only one activity (*Act$_i$*)

GTM-based pharmacological profiling: multi-task mode



CHEMICAL DATABASE

Descriptor1 ->

"Universal" GTM

Act_1

Act_2

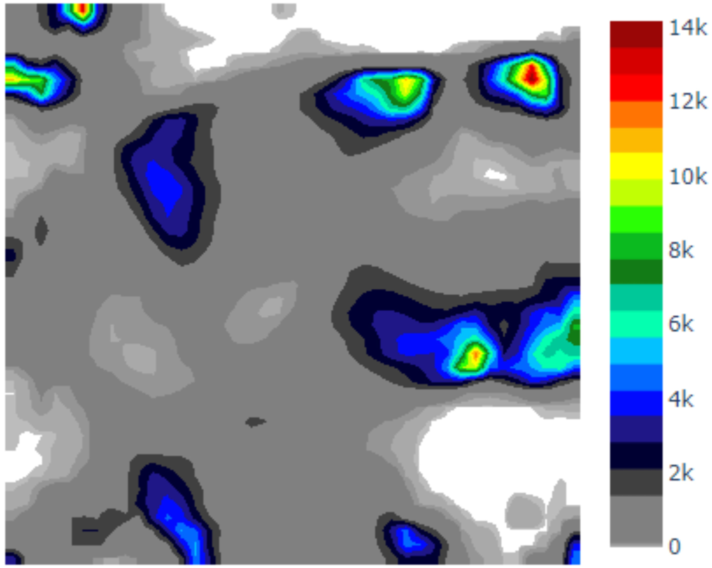
Act_3

Activity profile

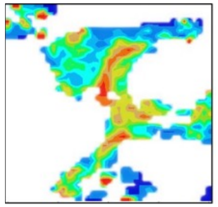
Universal GTM able to predict simultaneously several Act_i

« Universal » map

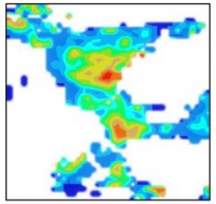
- Defines a frame of biological relevant chemical space (ChEMBL database)
- Based on ISIDA descriptors tuned with respect to the modelled activities
- Predicts of > 700 biological activities



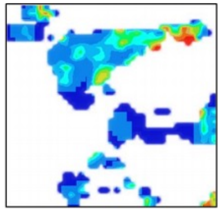
Density landscape of the ChEMBL database (1.7 M cmds)



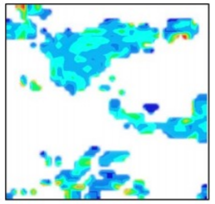
MAP kinase p38 alpha



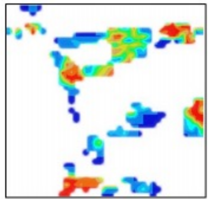
Vascular endothelial growth factor receptor 2



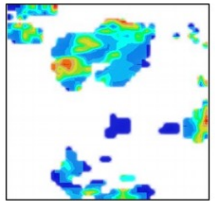
Cyclin-dependent kinase 2



Serine/threonine-protein kinase AKT



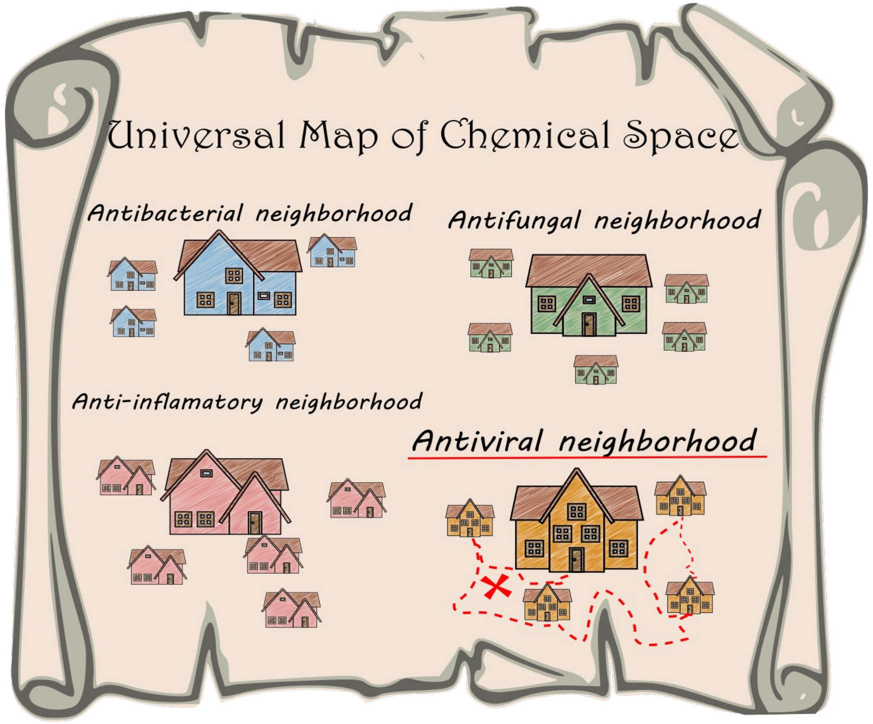
Phosphodiesterase 5A



Adenosine A2a receptor

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

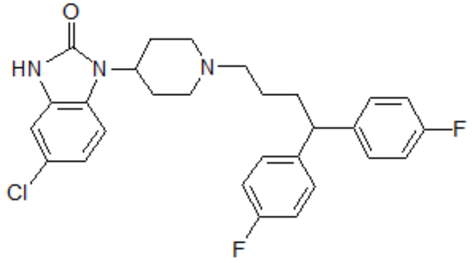
« Universal » map of Chemical Space


Universal Map of Chemical Space
Antibacterial neighborhood
Antifungal neighborhood
Anti-inflamatory neighborhood
Antiviral neighborhood

A map of a chemical space is expected:

- to accommodate the variety of known chemotypes;
- to distinguish between different activity classes;
- to separate actives and inactives within a given activity class
- to be *neighborhood behaviour (NB)* compliant, e.g., molecules grouped together are expected to display similar activities

One single descriptors space may not be sufficient !

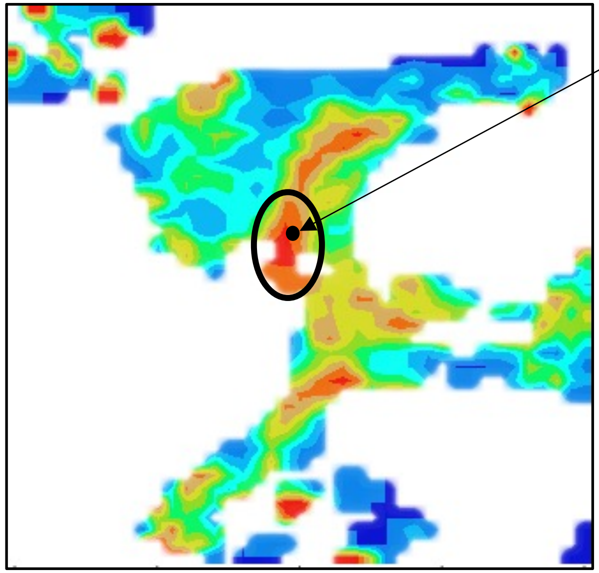


Experiment: *Active*

Predictions: *Active*

Experiment: *Active*

Predictions: *Inactive*

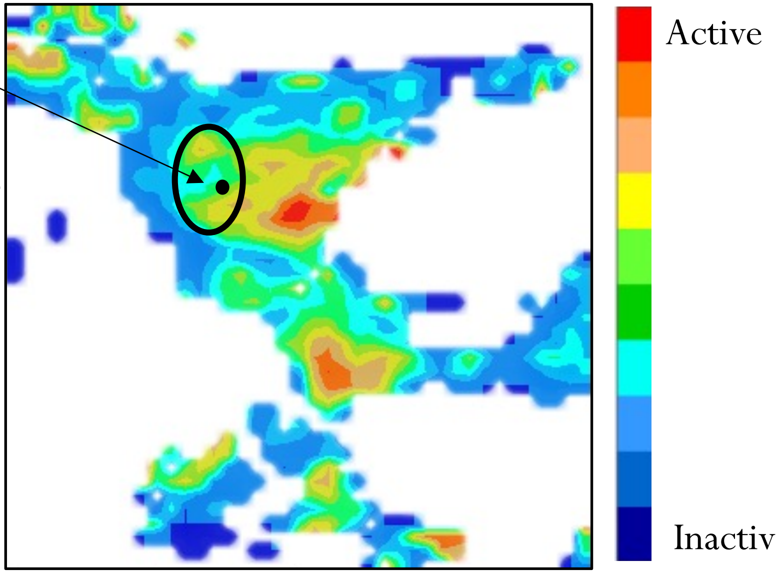


Adenosine A2a receptor

97%

42 %

Probability to be active

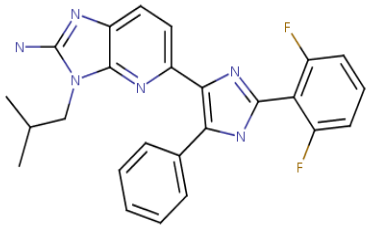


MAP kinase p38 alpha

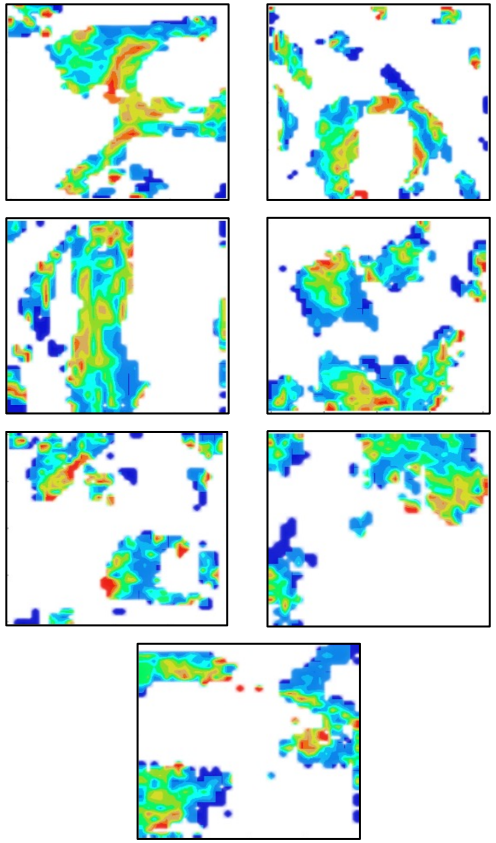
One descriptor space may not be sufficient to correctly separate actives/inactives for all targets

Chemical multiverse: ensemble of several optimal descriptor spaces

New molecule



7 Universal maps



GTM class landscapes for Adenosine A2a receptor

Individual predictions

Active	Active
Inactive	Active
Active	Active

Out of AD

Consensus prediction

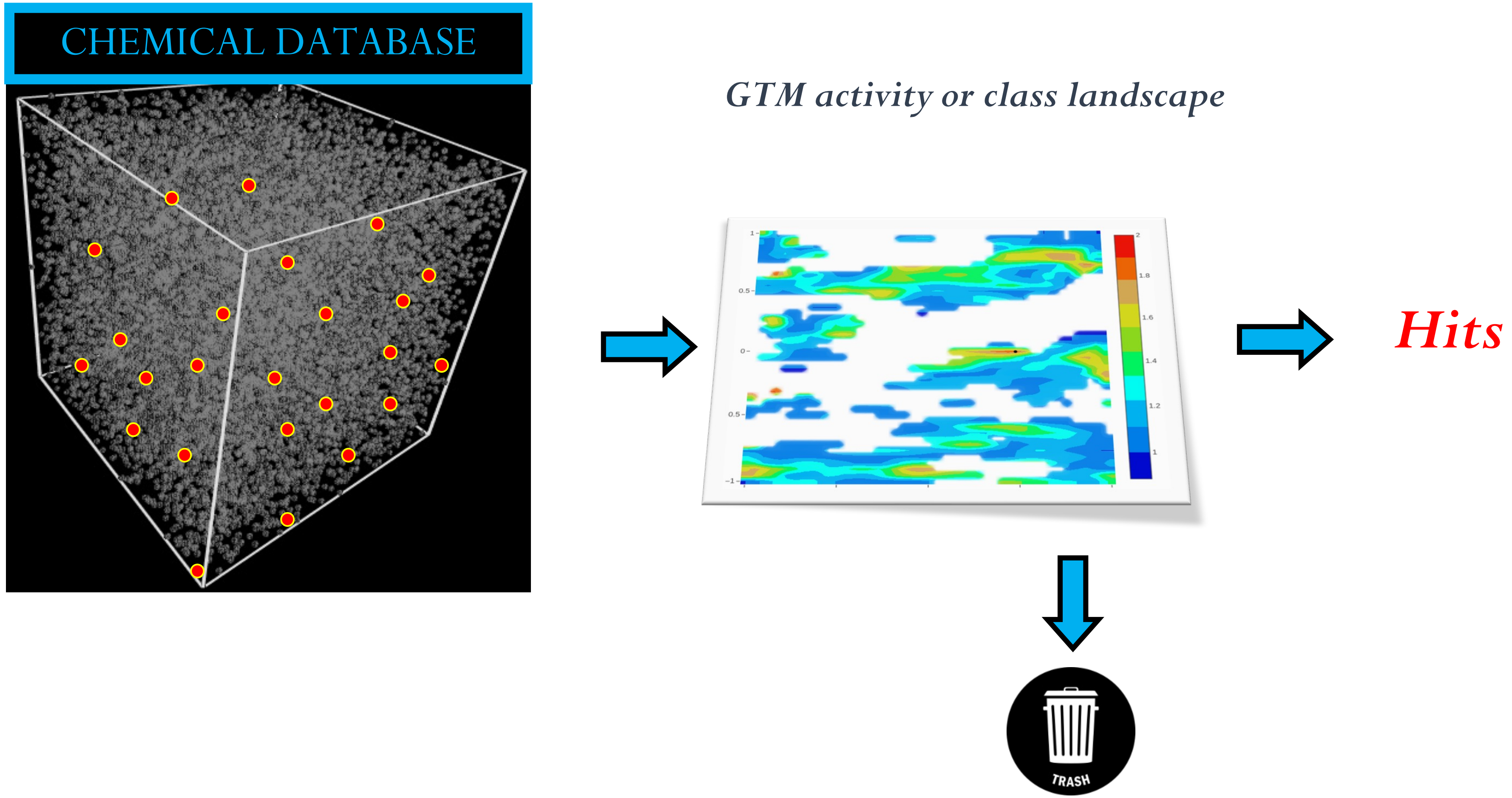
Active

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

GTM: applications

- Virtual screening
- Analysis of large chemical collections
- Drug resistance analysis
- AI-driven design of new molecules and reactions

Universal maps: application to virtual screening

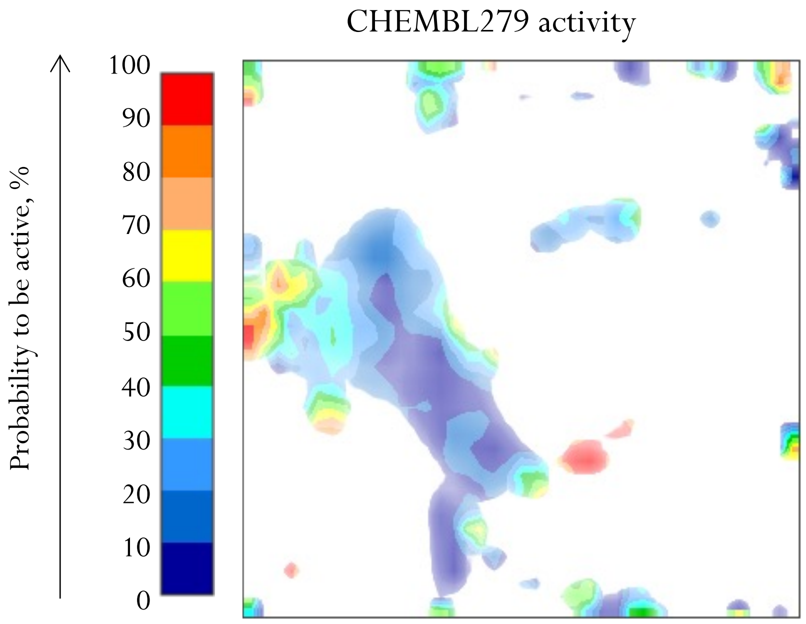


CHEMICAL DATABASE

GTM activity or class landscape

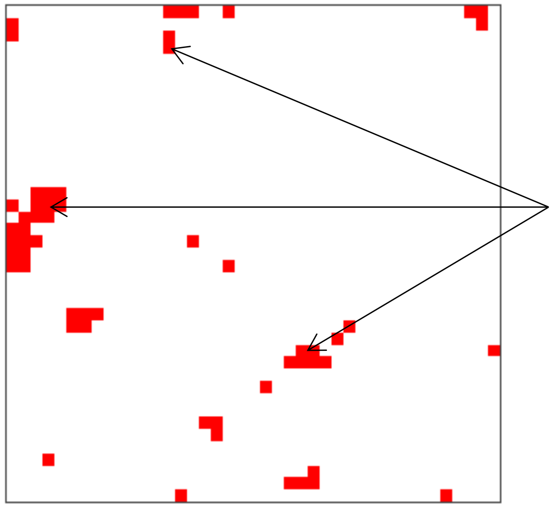
Hits

Constrained Screening: Zones of Interest



User’s query

Probability to be active is > 50 %

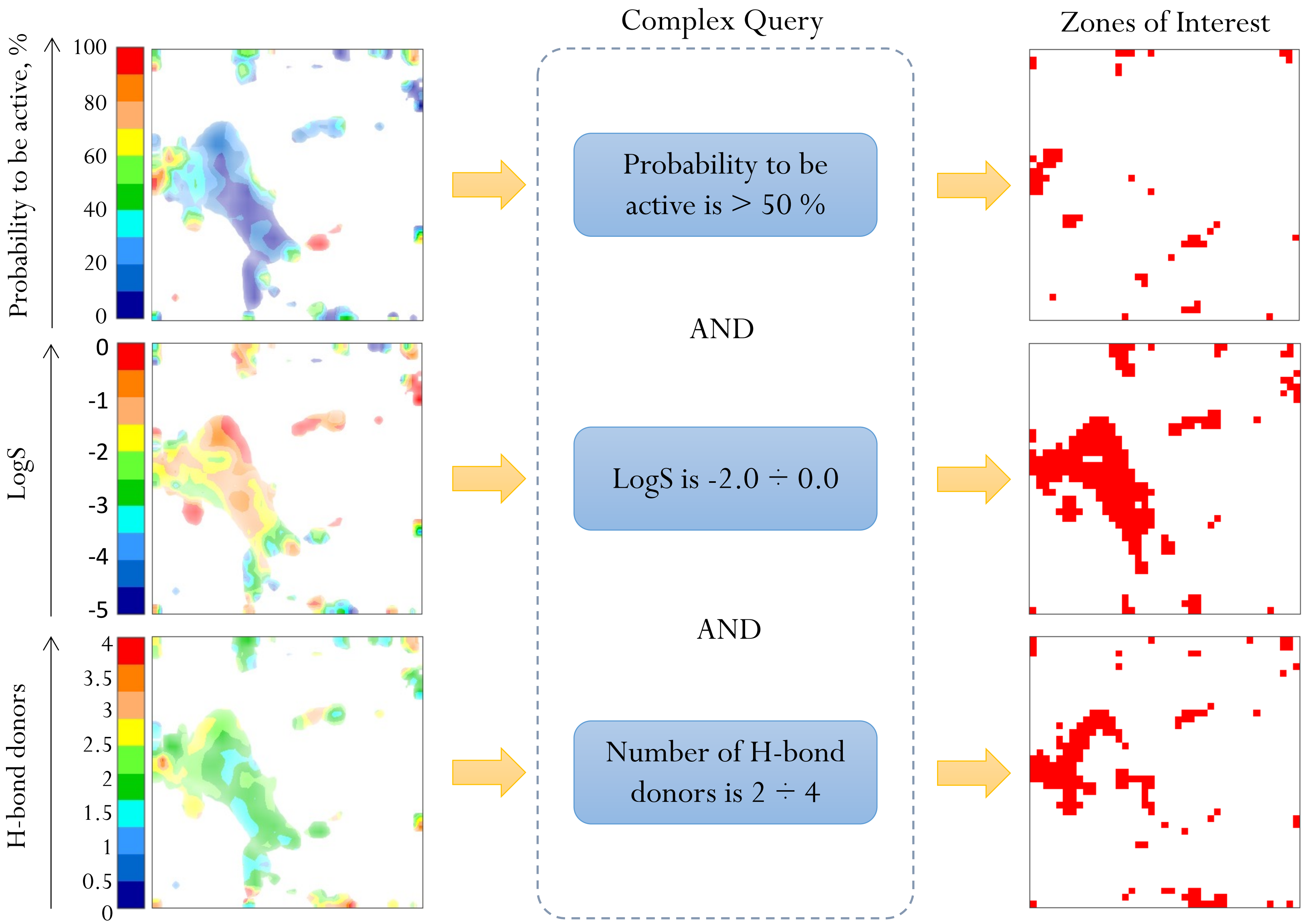


Zones satisfying the query

CHEMBL279 is Vascular endothelial growth factor receptor 2.

Gaspar HA et al. (*2015*) *J Chem Inf Model 55:84–94.*

Constrained Screening



Complex Query

Probability to be active is > 50 %

AND

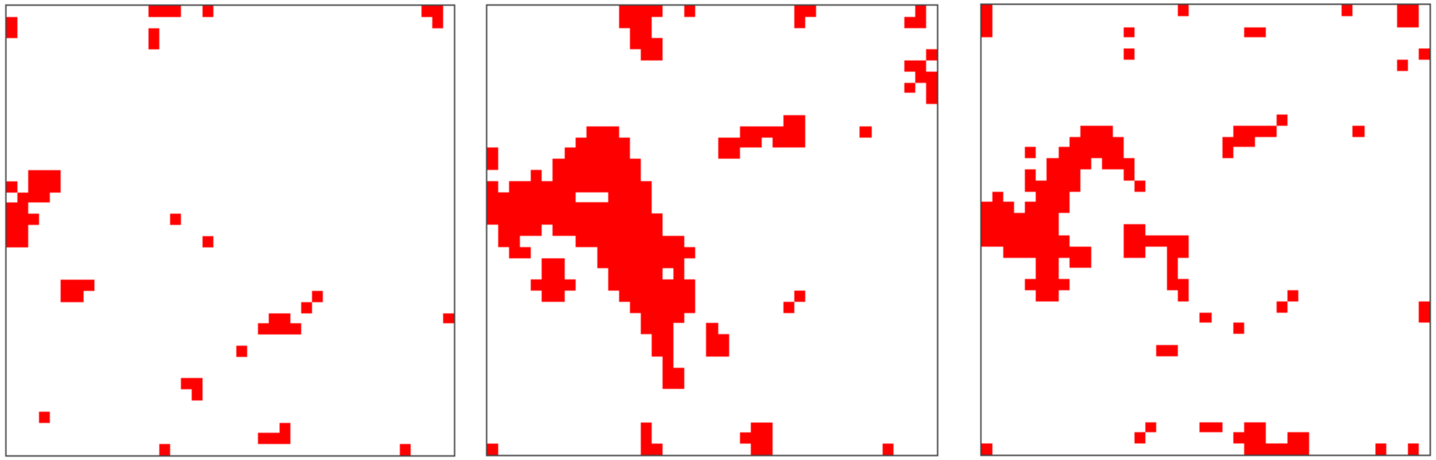
LogS is -2.0 ÷ 0.0

AND

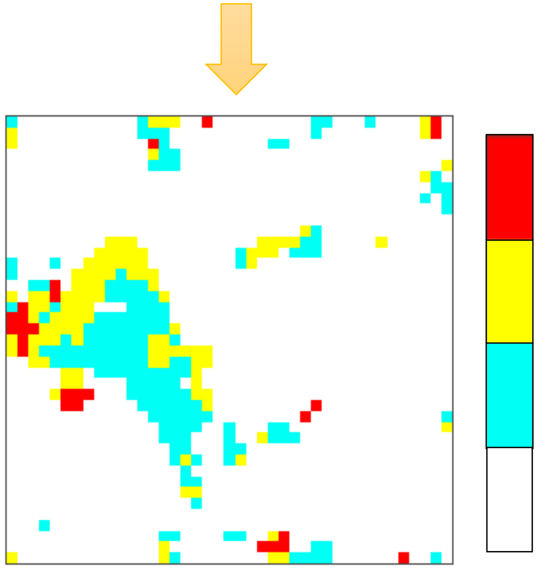
Number of H-bond donors is 2 ÷ 4

Zones of Interest

Constrained Screening



Zones of Interest



🟥 All queries are satisfied

🟨 2/3 queries are satisfied

🟦 1/3 queries are satisfied

⬜ Undesirable area

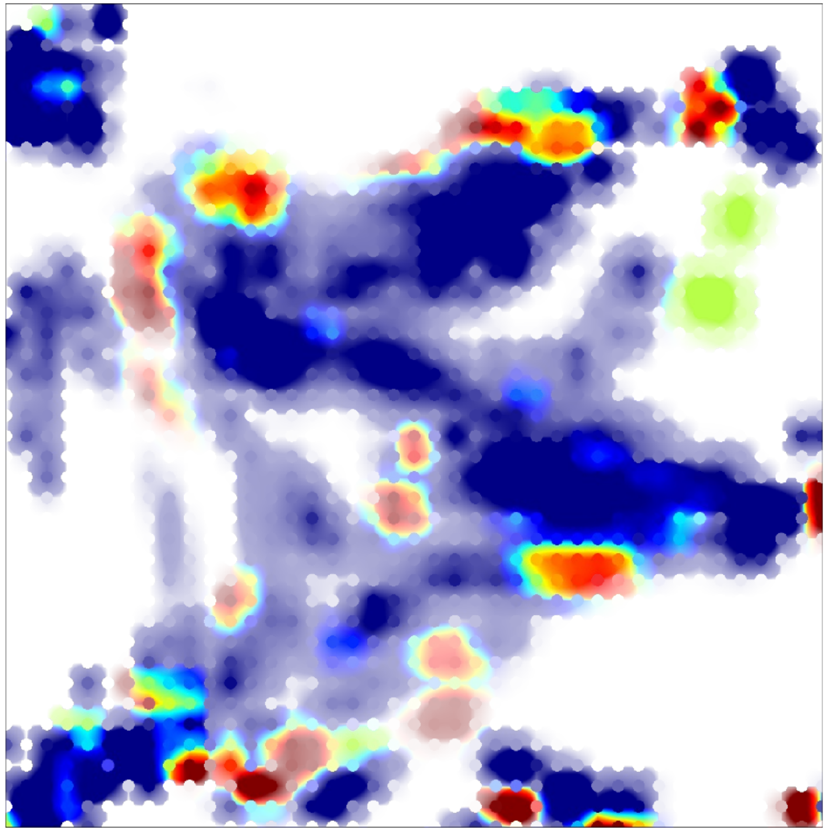
Query Landscape

Discovery of new SARS-Cov2 agents

- **SARS-COV Relevant Antiviral Space** covers *alpha* (269 molecules) and *beta* (1308) genus of CoVs
- The data for SARS-COV2 was not available at the very beginning of the COVID19 pandemic.

SARS and MERS 🟦 beta

🟥 alpha



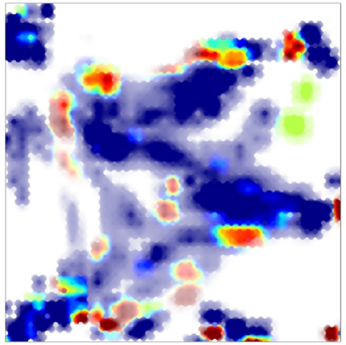
D. Horvath et al. *Molecular Informatics,* 2020, 39(12), 2000080

Discovery of new SARS-CoV-2 M^{pro} inhibitors Space

ZINC

DATABASE

>1.3 billion cmpds

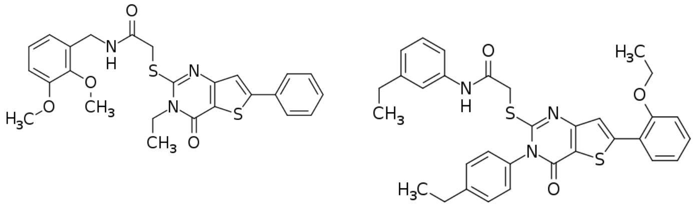


574 hits

Docking to SARS-CoV-2 M^{pro}

10 hits

experiment



Confirmed M^{pro} inhibitors

D. Horvath et al. *Molecular Informatics*, 2020, **39**(12), 2000080

M.Yu. Zakharova et al *Frontiers in Pharmacology*, 2021, **12**, 773198

Comparative analysis of (ultra)large chemical libraries

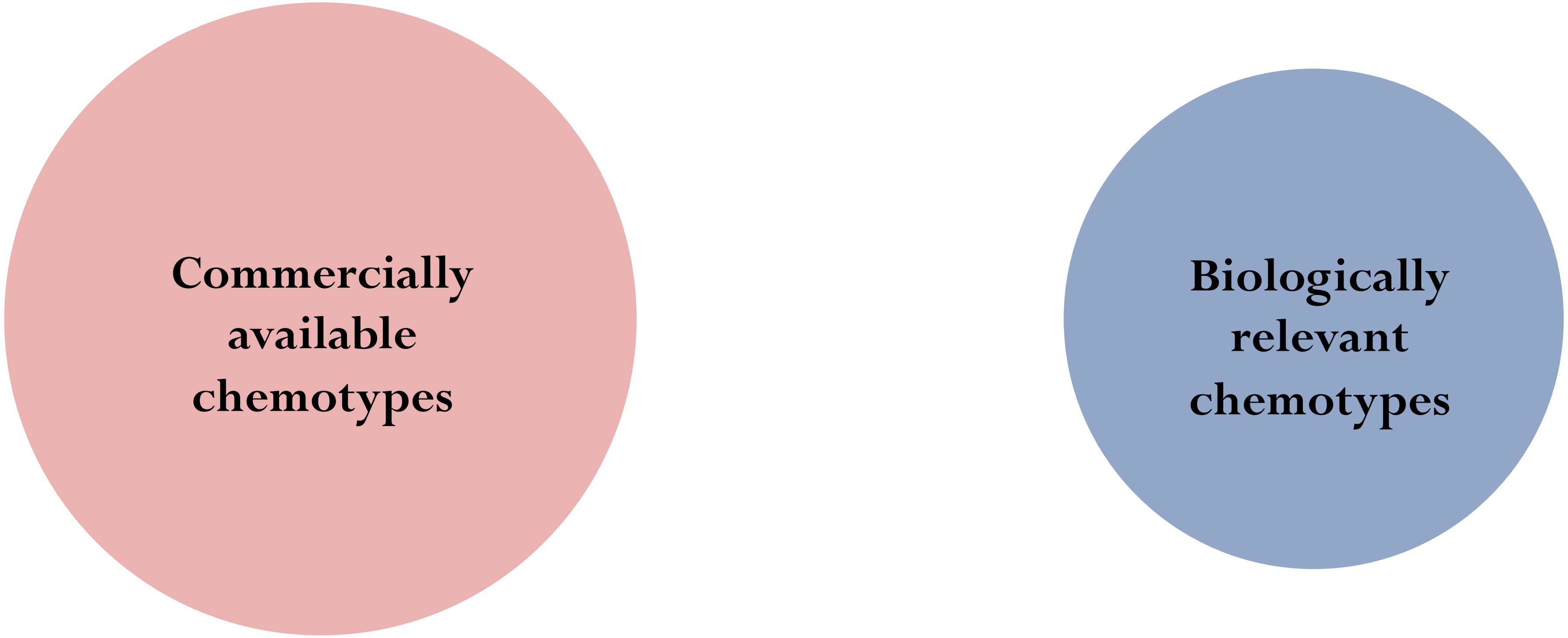
Case studies:

- ChEMBL / ZINC (1B structures)
- Proprietary collection reshaping
- Selection of optimal DELs

Commercial vs Biologically relevant data

ZINC DATABASE

ChEMBL



Commercially available chemotypes

Biologically relevant chemotypes

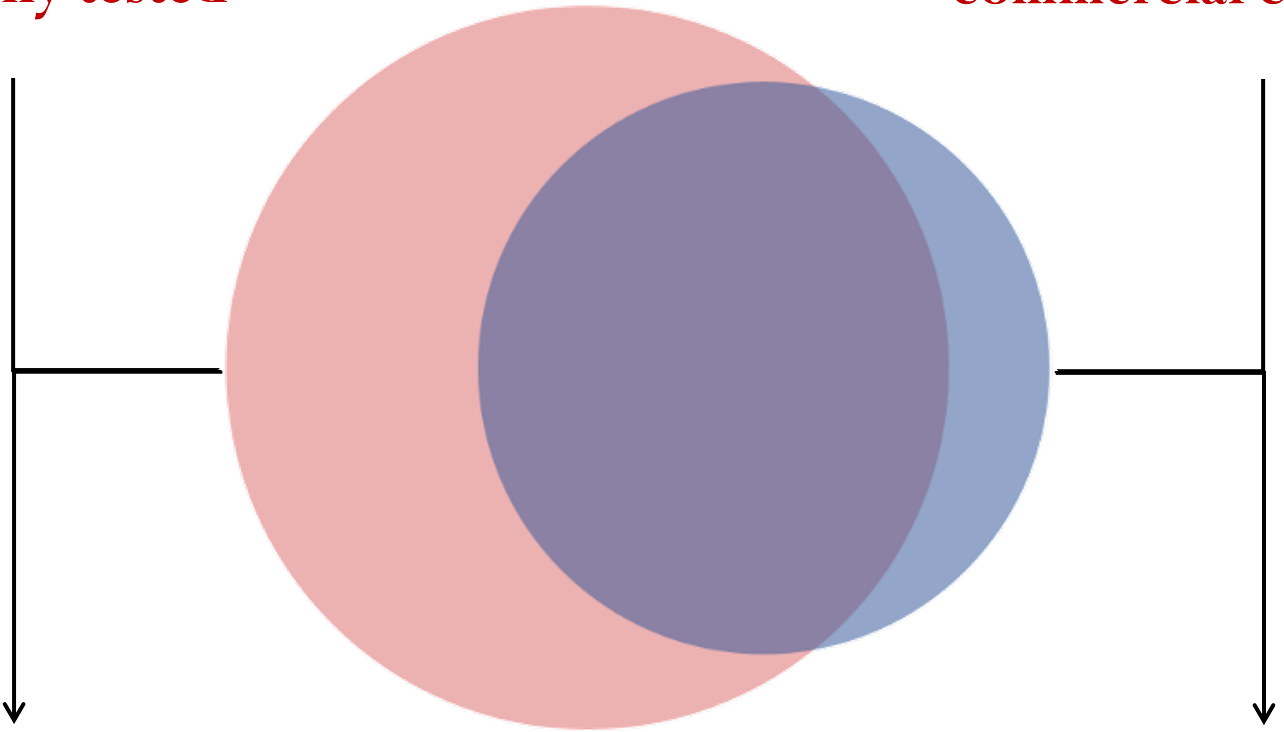
>1.3 billion cmpds

>1.8 M cmpds

Commercial vs Biologically relevant data

Chemotypes never been biologically tested

Chemotypes missing in the commercial chemical space

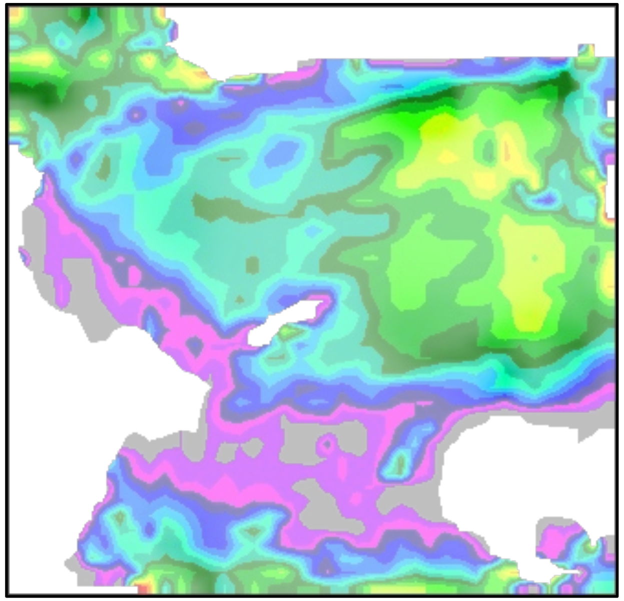


Enhancement of screening libraries

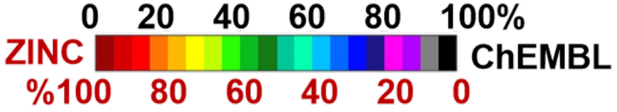
Biologically biased commercial libraries enhancement

GTM class landscape for library comparison

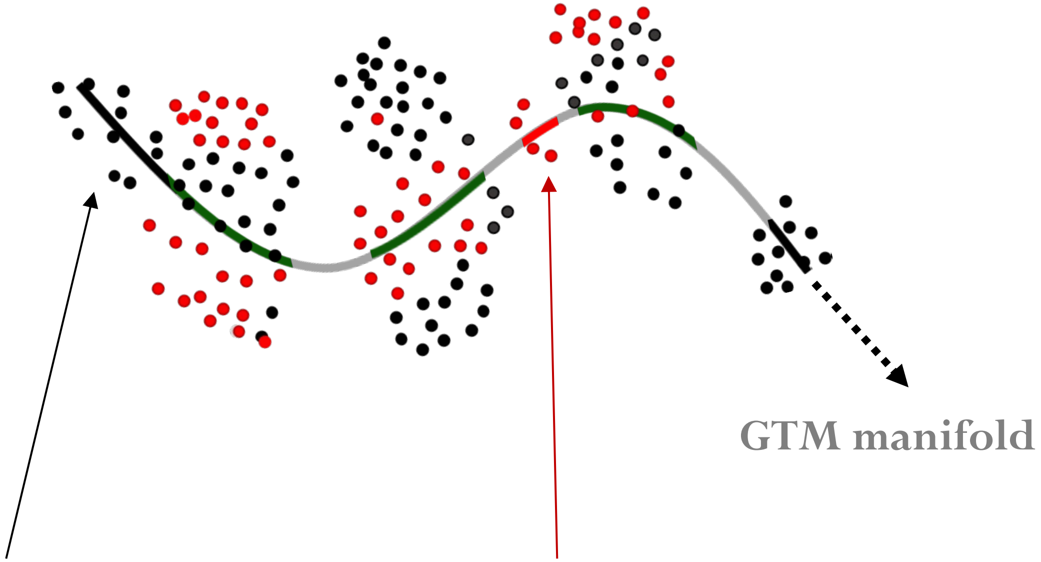
Subset of Fragment-like cmpds



#compounds 3 614 394



« Global » GTM manifold in low resolution map doesn't separate the classes

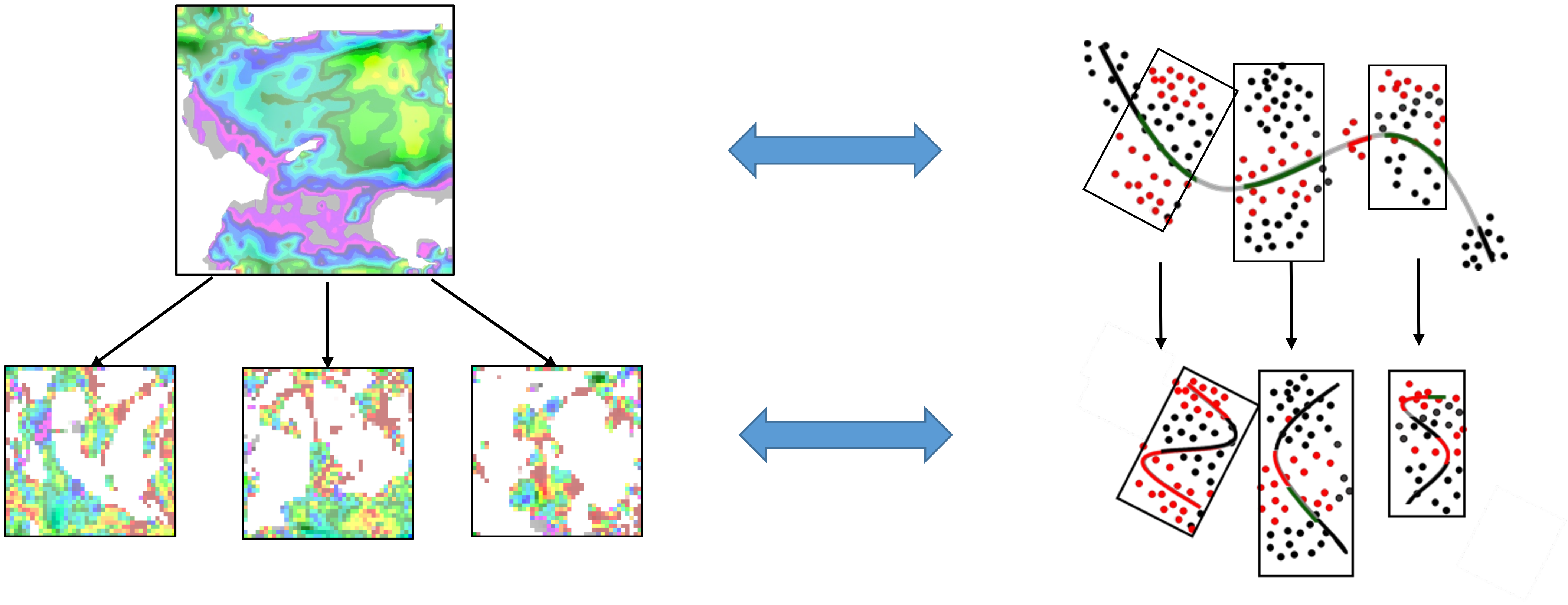


ChEMBL-specific zone

ZINC-specific zone

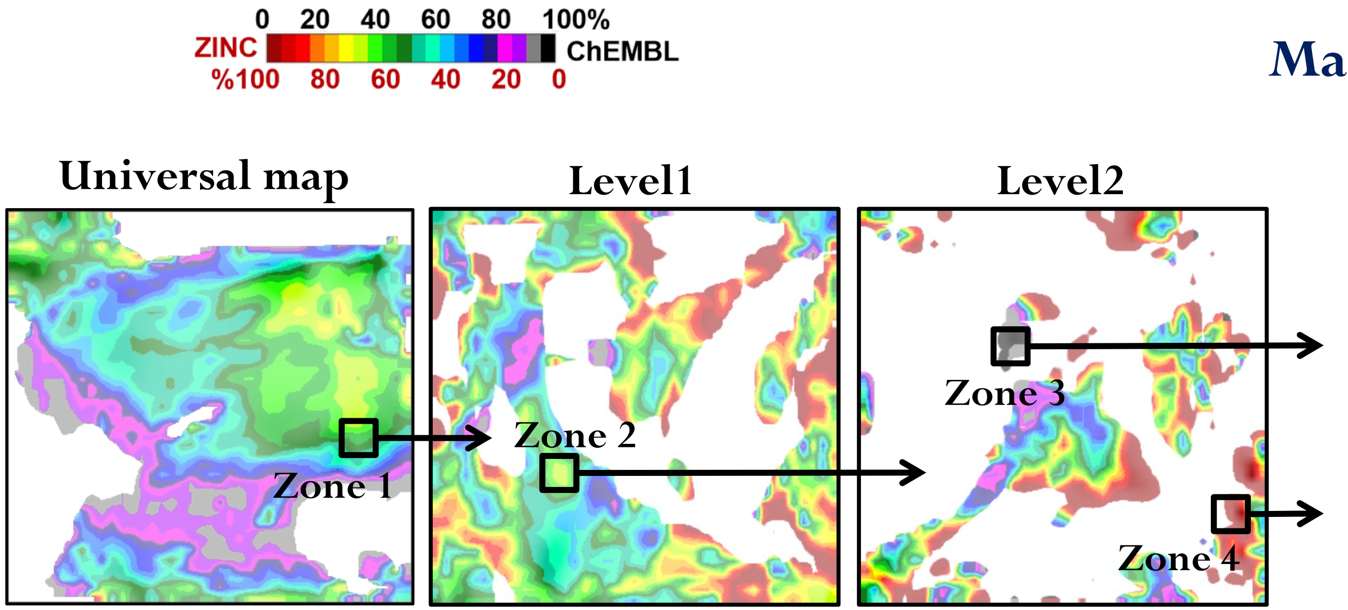
Low resolution of the map doesn't allow to identify the library-specific zones

Hierarchical GTM (Zooming)



New higher resolution maps better separate the library members

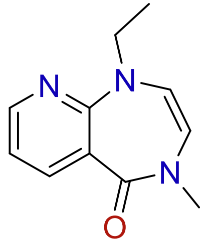
Hierarchical GTM navigation of the chemical space



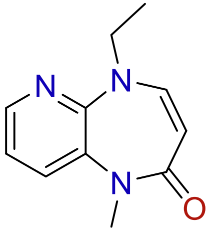
#compounds 3 614 394 | 82 246 | 4 230

Maximum Common Substructures (MCS)

ChEMBL-specific MCS

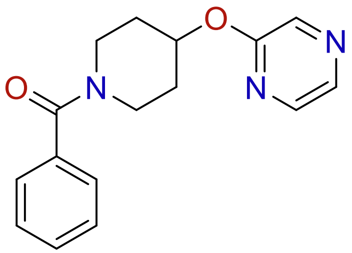


Popularity - 18

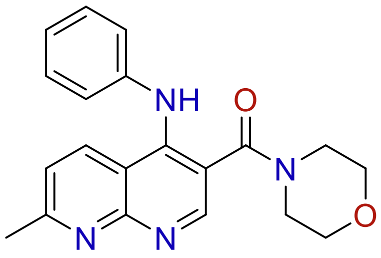


Popularity - 16

ZINC-specific MCS



Popularity - 21



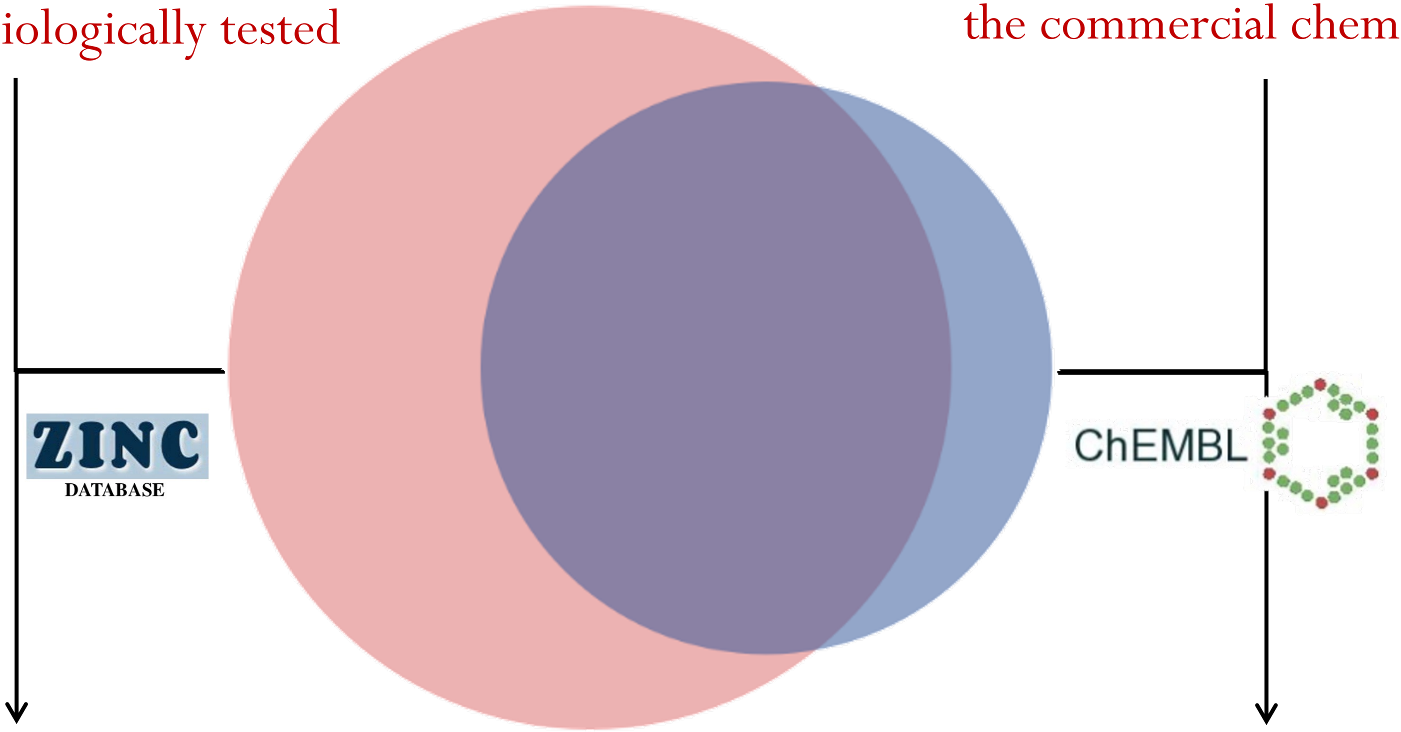
Popularity - 15

Yu. Zabolotna *et al.*, « *Searching for hidden treasures* », *J. Chem. Inf. Model.* 2021, 61, 1, 179–188

Commercial vs Biologically relevant data

> 100K chemotypes
never been biologically tested

>20 K chemotypes missing from
the commercial chemical space



Enhancement of screening libraries

Biorelevance-biased commercial library enhencement

Chemical Library Enrichment

Boehringer Ingelheim

SIGMA-ALDRICH

2.2 M cmps

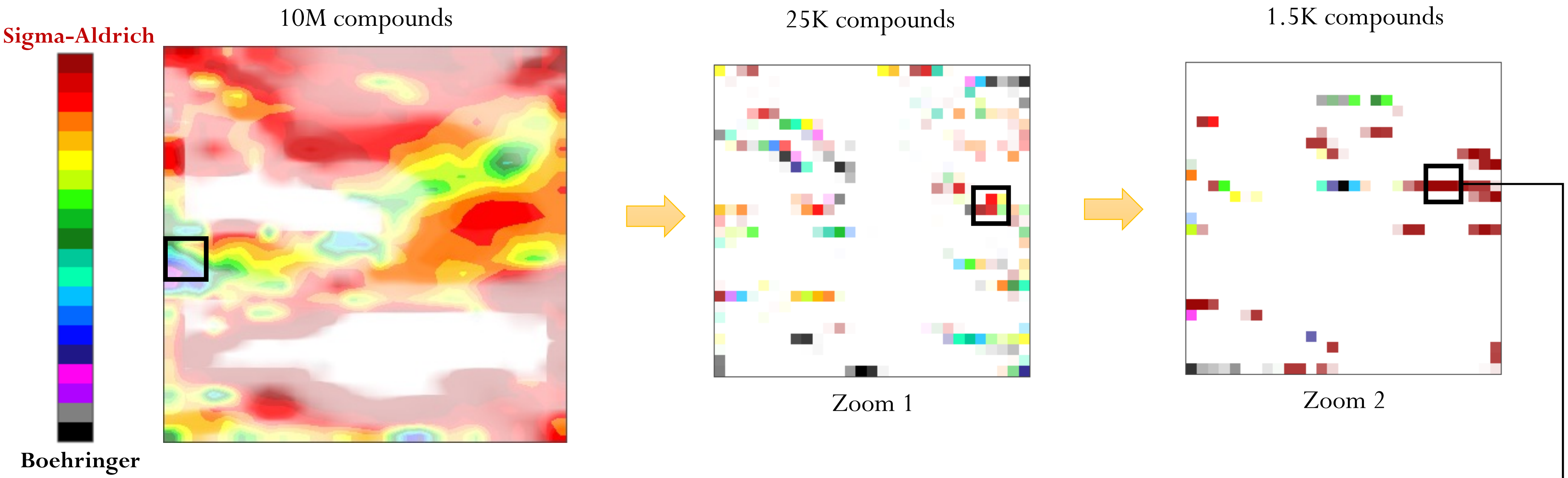
BI intended to diversify its library by purchasing compounds from the Aldrich-Market Select (AMS) Database

8.3 M cmps

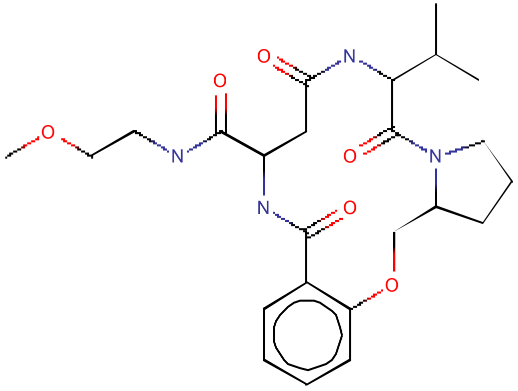
***Goal:* selection of the AMS compounds with *new* scaffolds/substructures**

PhD project of Arkadii Lin

Chemical Library Enrichment

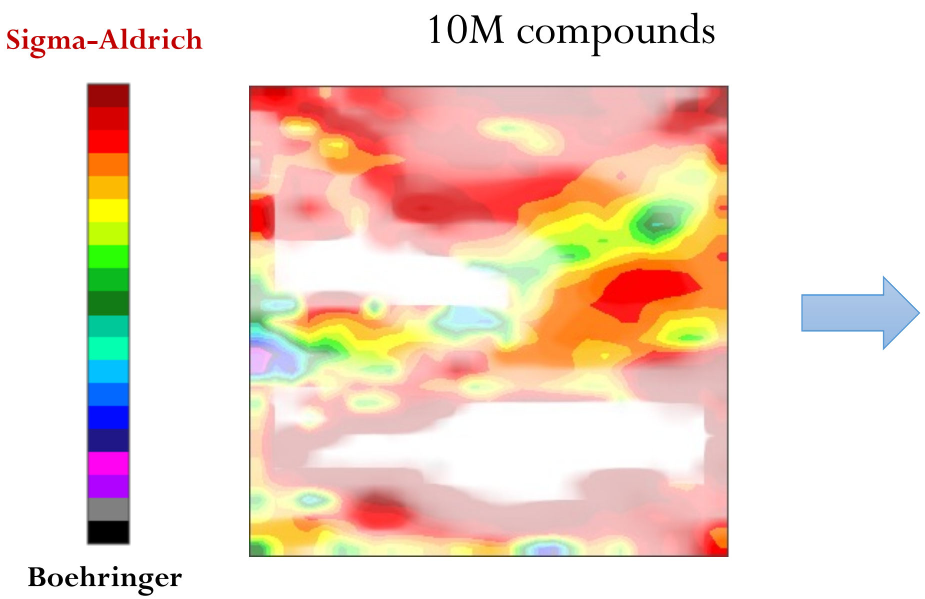


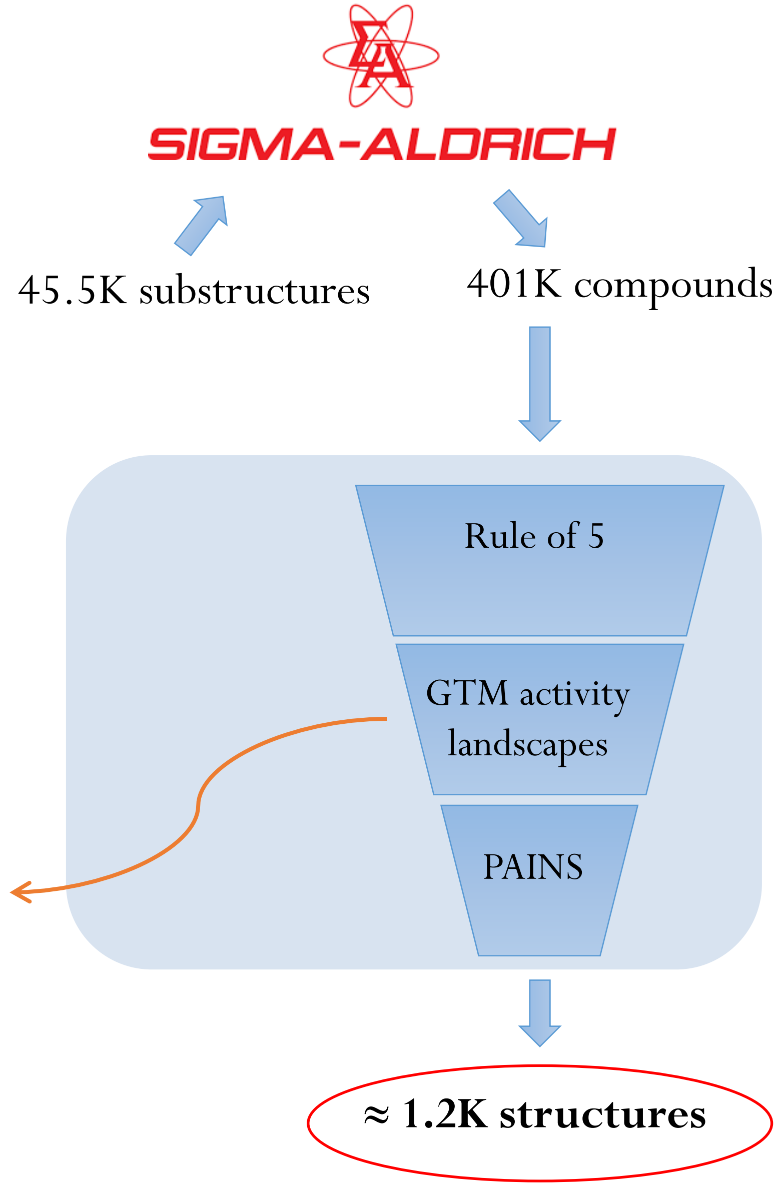
MCS extraction from the zones populated by AMS



New substructure from Sigma-Aldrich

Chemical Library Enrichment




SIGMA-ALDRICH

45.5K substructures

401K compounds

Rule of 5

GTM activity landscapes

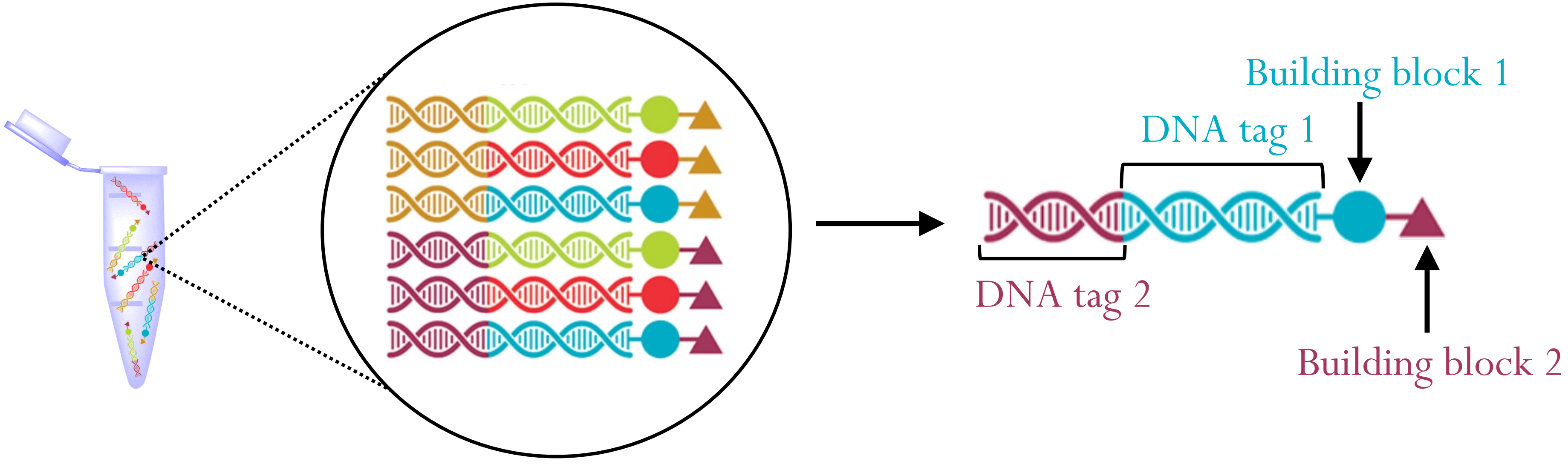
PAINS

≈ **1.2K structures**

Compounds potentially active against, at least, one out of 749 ChEMBL targets were selected

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

Generation and analysis of general-purpose DELs



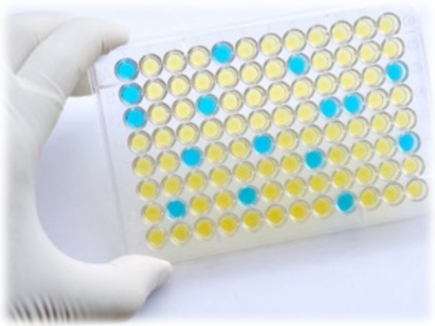
DNA-Encoded Library

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

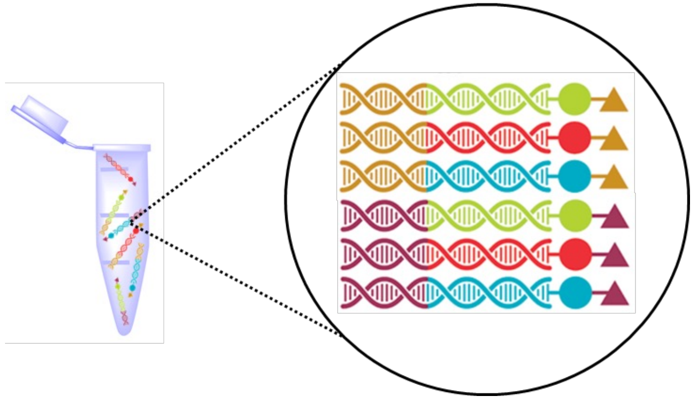
Screening libraries



Parallel screening in separate "wells"

Individual compounds may be cherry-picked

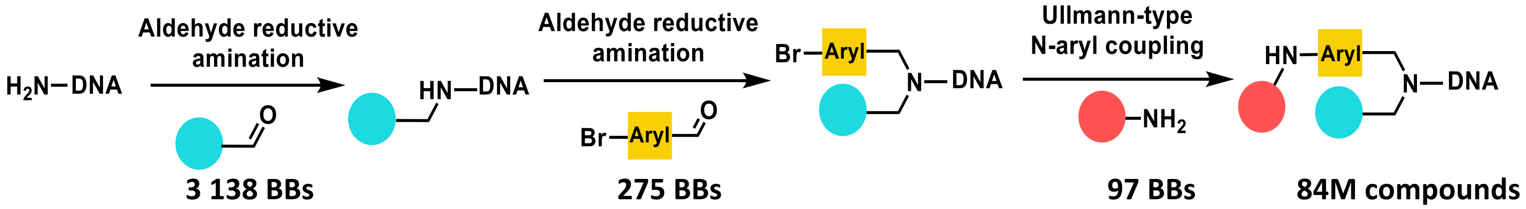
DNA-encoded libraries

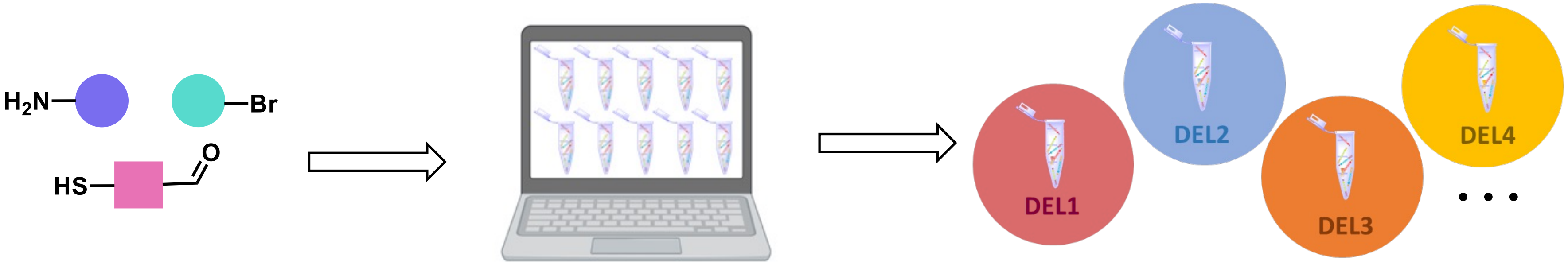


Simultaneous screening in a single tube

Entire library as an object must be considered

Selection of an “optimal” DEL


H2N—DNA
Aldehyde reductive amination
3 138 BBs
HN—DNA
Aldehyde reductive amination
Br—Aryl
275 BBs
Br—Aryl
N—DNA
Ullmann-type N-aryl coupling
NH2
97 BBs
HN—Aryl
N—DNA
84M compounds

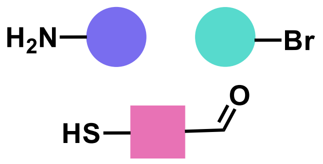


Commercially available BBs

Thousands of DELs containing billions of molecules

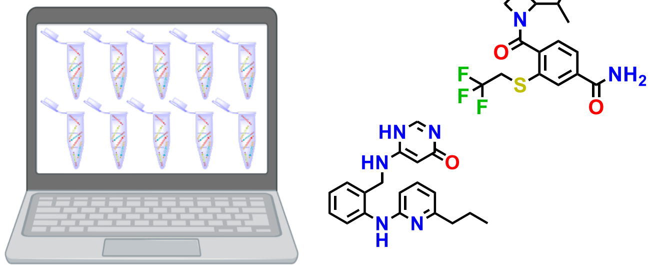
How to select an “optimal” DEL for a given task (e.g., primary screening) ?

Selection of DEL the best covering a reference library (ChEMBL) chemical space

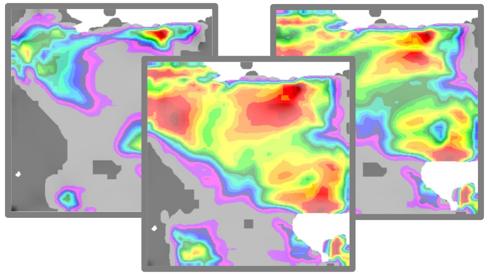


79.000 Building blocks from eMolecules

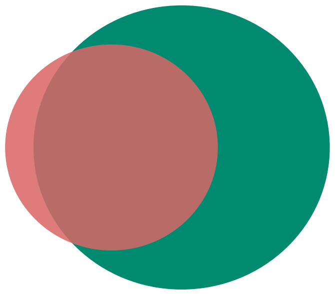
eDesigner tool



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



2500 comparative landscapes
DEL_i/ChEMBL

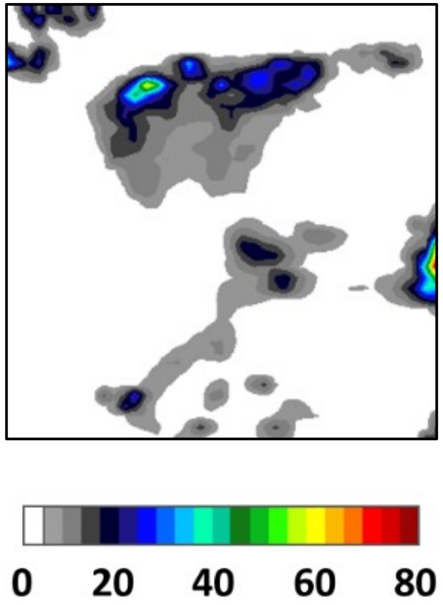


DEL_i / ChEMBL coverage score calculation for each map

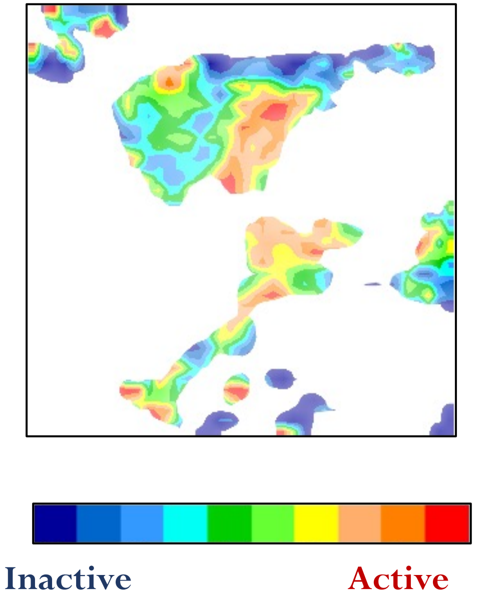
Selection of highly scoring DELs

Encoding a library by a vector using GTM landscapes

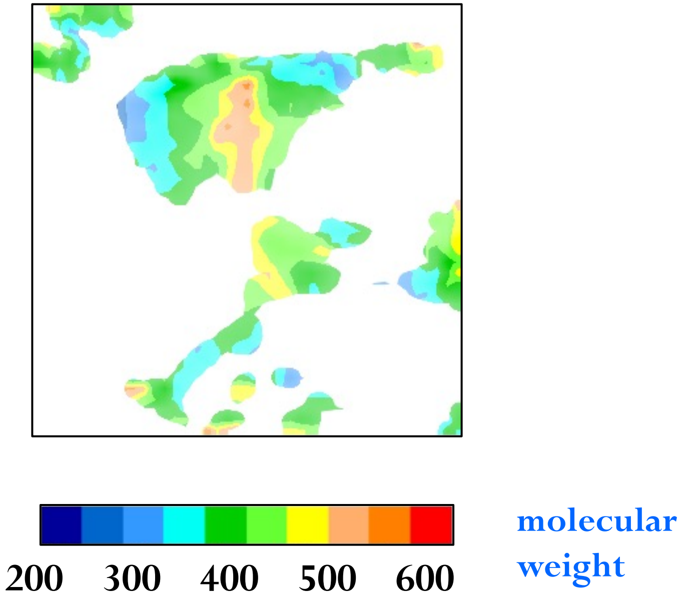
Density landscape



Class landscape

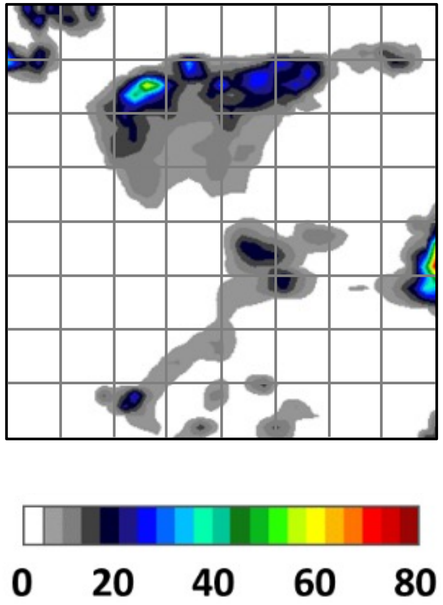


Property (activity) landscape



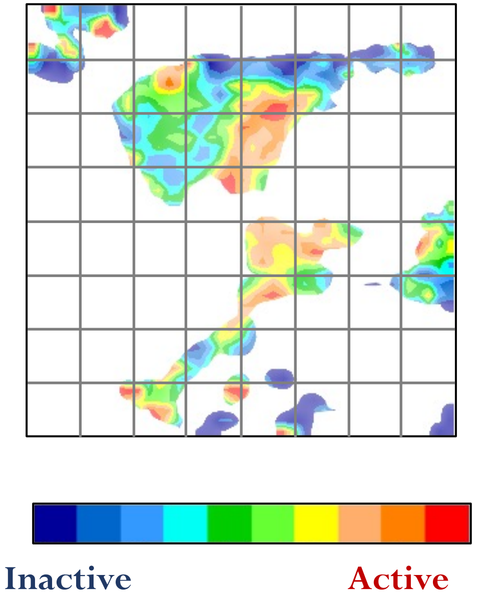
Encoding a library by a vector using GTM landscapes

Density landscape



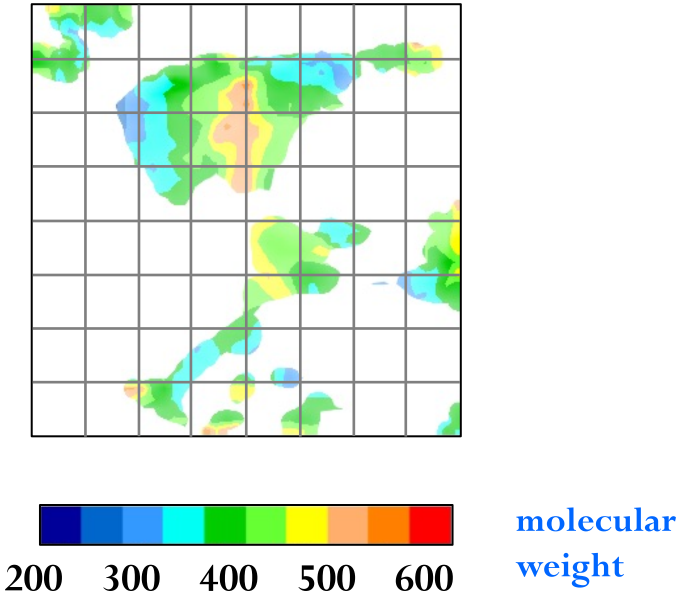
Cumulated Responsibility Vector (**CRV**)

Class landscape



Class Modulated Responsibility Vector (**cCRV**)

Property (activity) landscape



Property Modulated Responsibility Vector (**pCRV**)

GTM-based metrics of chemical libraries similarity

Library encoding

1. **Responsibility Patterns (RP),** e.g. GTM “address labels” obtained from a discretized, coarse responsibility vector

   $\vec{\boldsymbol{R}}$=*(12:0.003 36:0.51 37:0.48 77:0.007)* → **RP=/36:5/37:5/**

2. **Cumulated Responsibility vectors**
3. **Property-modulated vectors**
4. **Library (class)-modulated vectors**

Metric

Coverage of *Lib_1* by *Lib_2*

$$RPcov\ (Lib1, Lib2) = \frac{N_{common}(RP_1\ \cap RP_2)}{N_{total}(RP_1)}$$

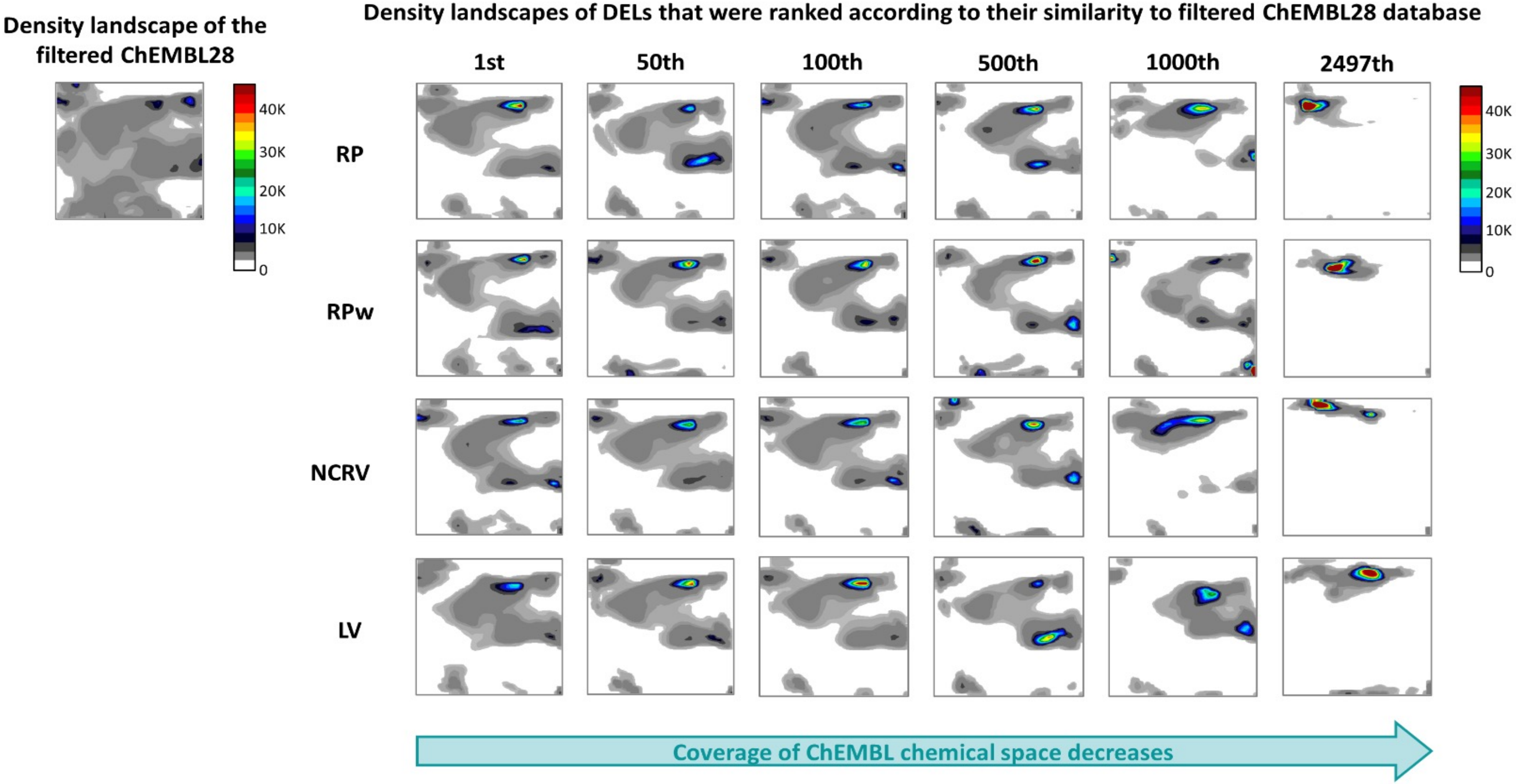
Pairwise *Lib1 / Lib2*:

- Tanimoto coefficient $(Vect_1 / Vect_2)$

Ensemble of libraries:

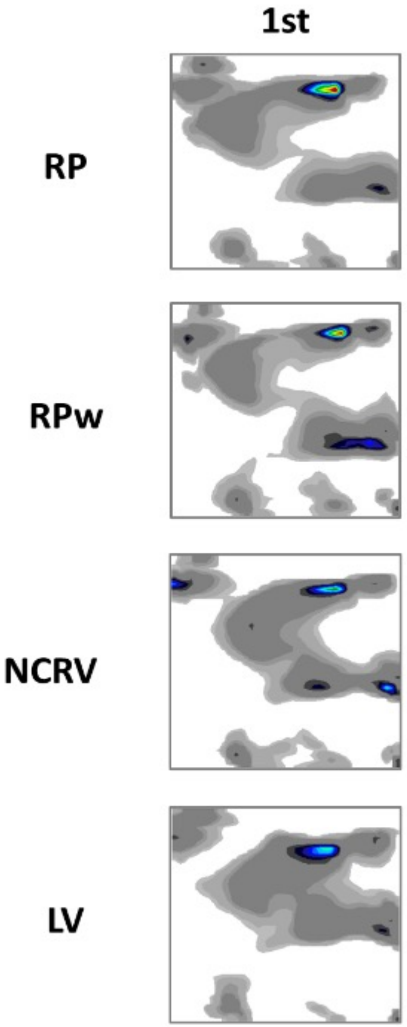
- meta-GTM built on $\{Vect_i\}$

ChEMBL28 / DEL similarity



Irrespective of the metric, common density (overlap) is a key factor defining inter-library similarity

DELs with the highest similarity to ChEMBL



RP:

Aldehyde reductive amination

Aldehyde reductive amination

Ulman-type N-aryl coupling

RPw:

Aldehyde reductive amination

Migita thioether synthesis

Guanidinilation

NCRV:

Aldehyde reductive amination

Carboxylic acid/amine condensation

Migita thioether synthesis

LV:

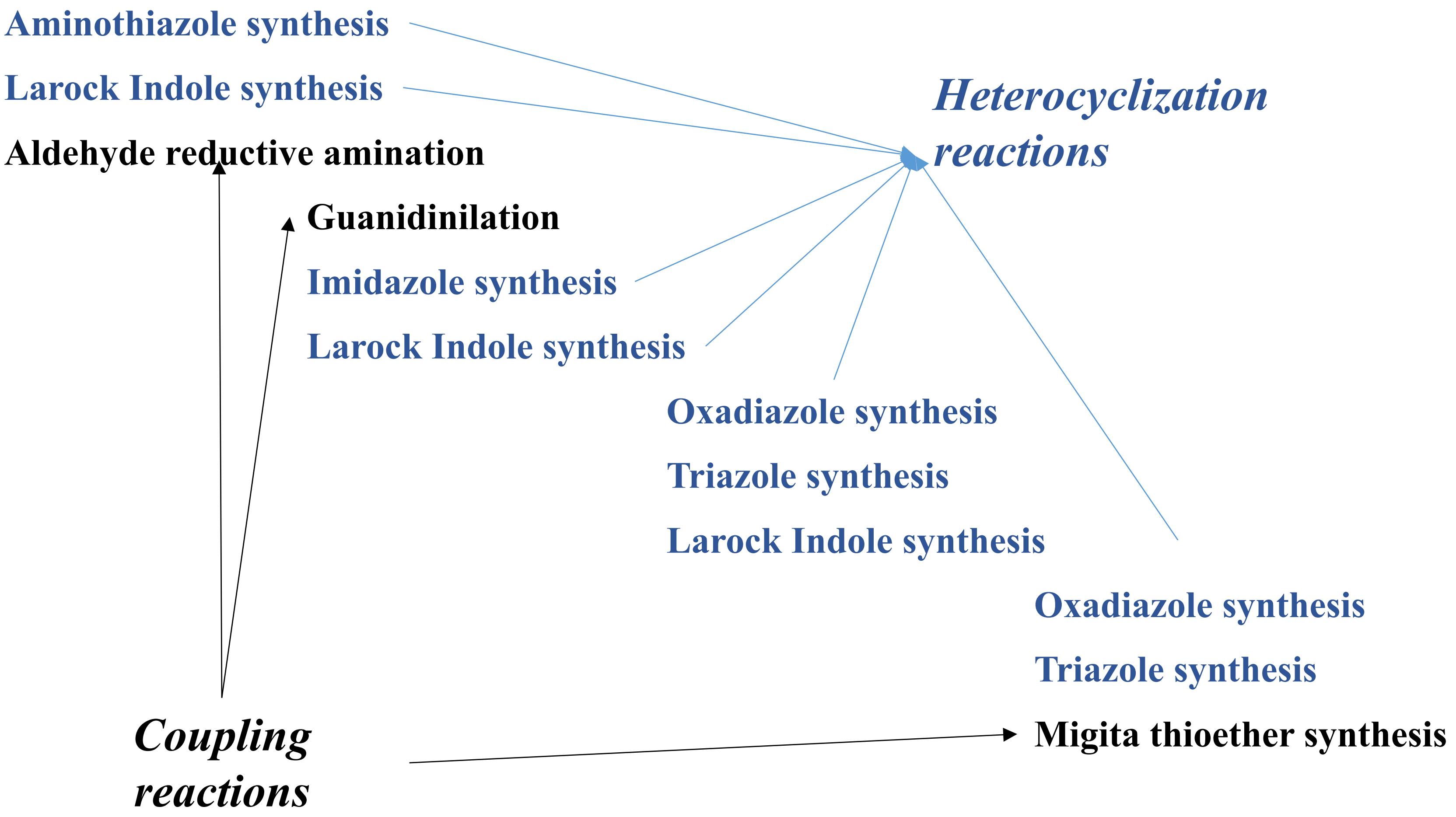
Aldehyde reductive amination

Carboxylic acid/amine condensation

Ulman-type N-aryl coupling

Only robust coupling reactions

DELs with the lowest similarity to ChEMBL



Aminothiazole synthesis

Larock Indole synthesis

Aldehyde reductive amination

Guanidinilation

Imidazole synthesis

Larock Indole synthesis

Oxadiazole synthesis

Triazole synthesis

Larock Indole synthesis

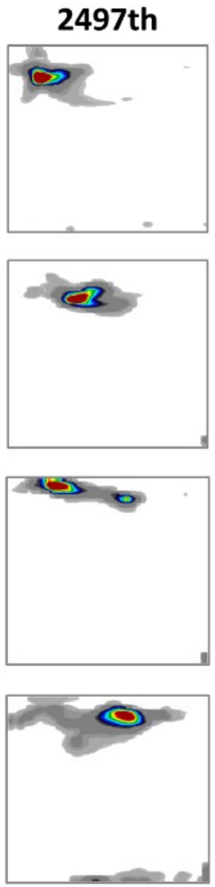
Oxadiazole synthesis

Triazole synthesis

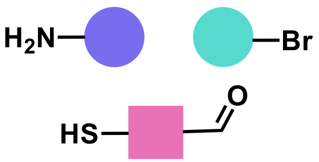
Migita thioether synthesis

Heterocyclization reactions

Coupling reactions

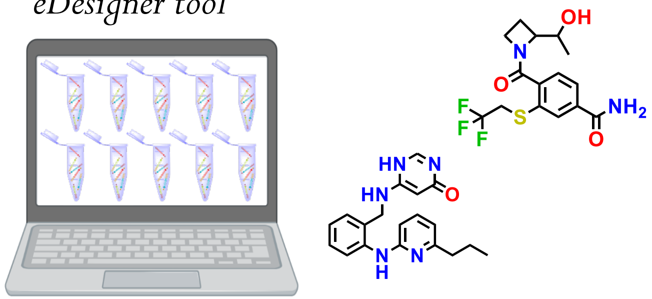


Selection of DEL the best covering a reference library (ChEMBL) chemical space

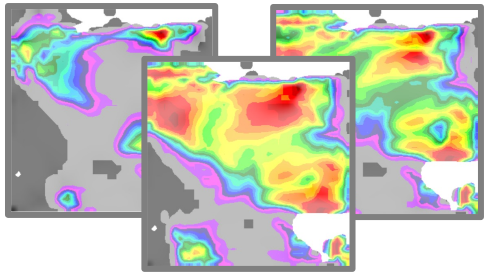


79.000 Building blocks from eMolecules

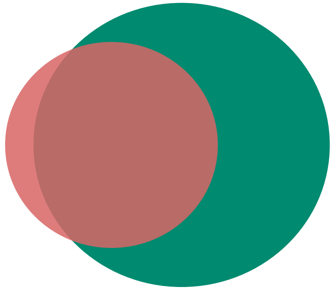
eDesigner tool



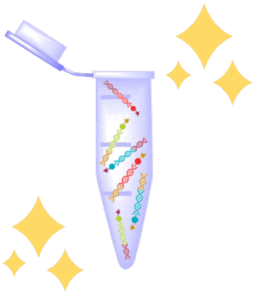
2500 DELs designed (size: 1M-1B)
2.5B compounds generated
(1M compounds per DEL)



2500 comparative landscapes
DEL_i/ChEMBL

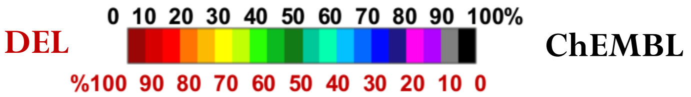


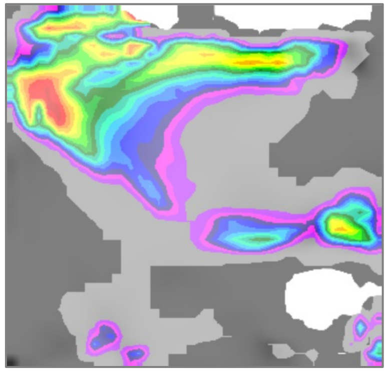
DEL_i / ChEMBL coverage score calculation for each map



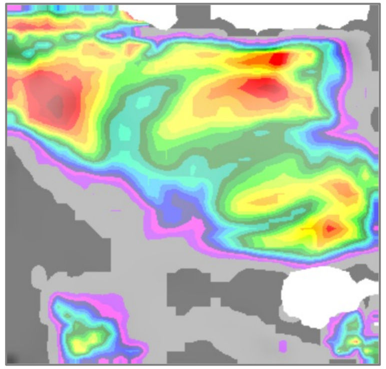
Selection of highly scored DELs

Selected DELs vs ChEMBL

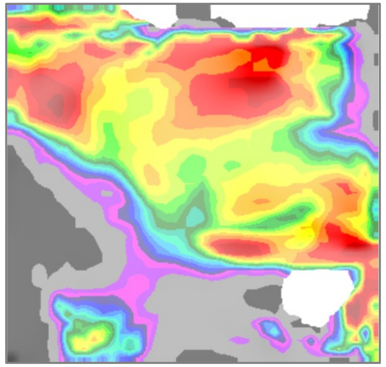




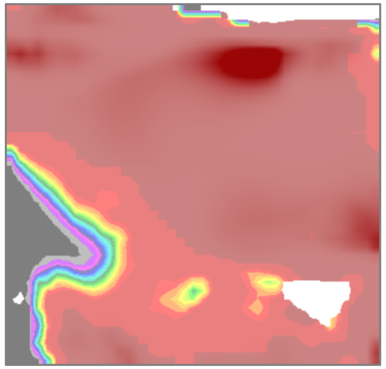
the best DEL



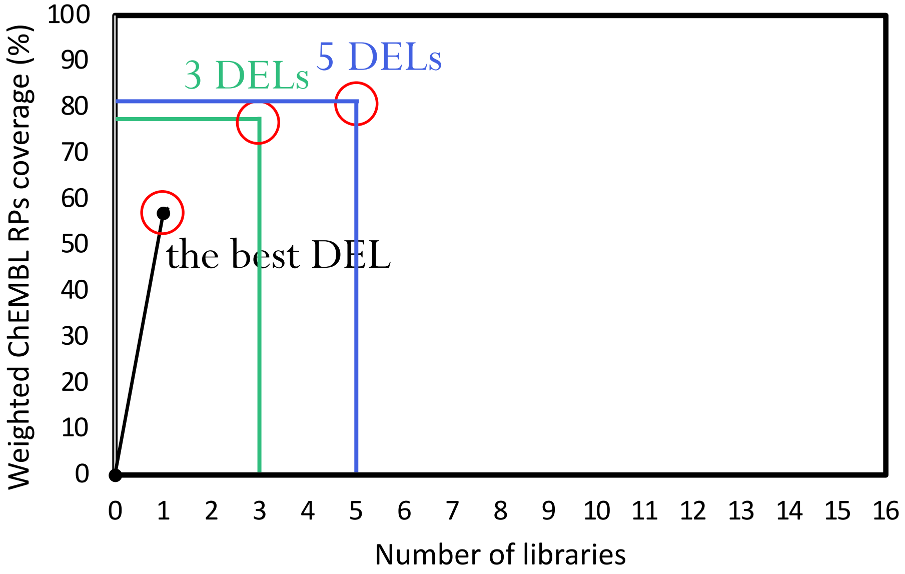
3 DELs



5 DELs



2500 DELs

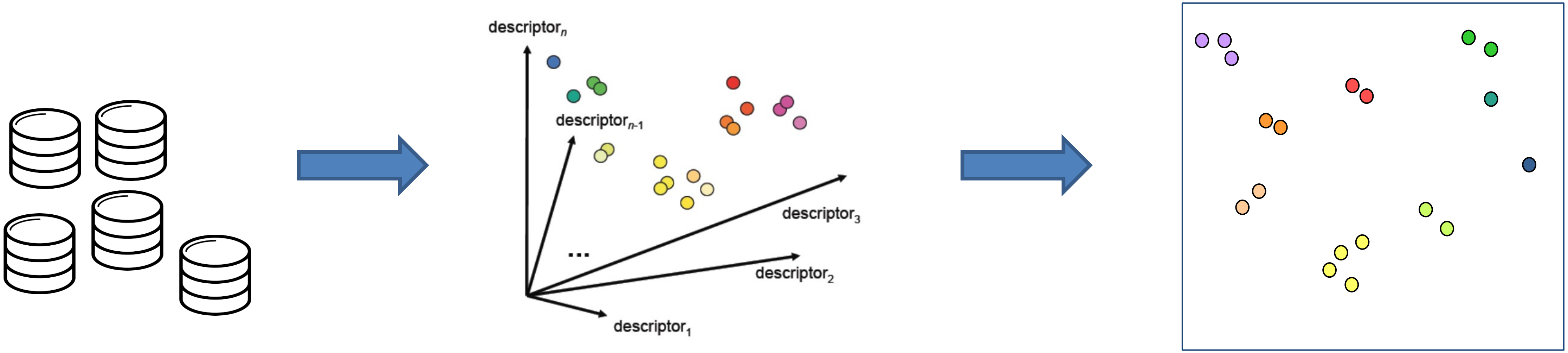


3 “platinum” DELs cover >80% of ChEMBL chemical space

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

Meta-GTM: a compact visualization of library space

- GTM encodes a chemical library as a vector of descriptors (cumulated responsibilities, property or class modulated responsibilities) calculated from related landscapes
- This vectors can be used to build a meta-GTM where each data point represents a library



Initial library space

(N-dimensional)

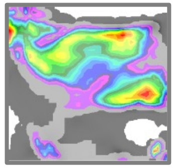
Meta-GTM

H.A. Gaspar et al., *J. Chem. Inf. Model.*, **2015**, *55* (1), 84–94

Meta-GTM built on reference library-modulated descriptors

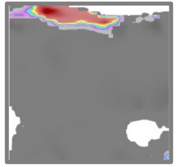
Vectors calculated for DEL/ChEMBL class landscapes

DEL1/ChEMBL



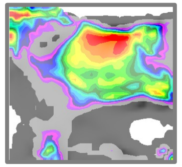
Vector 1

DEL2/



Vector 2

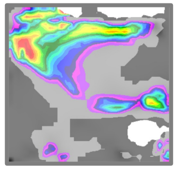
DEL3/



Vector 3

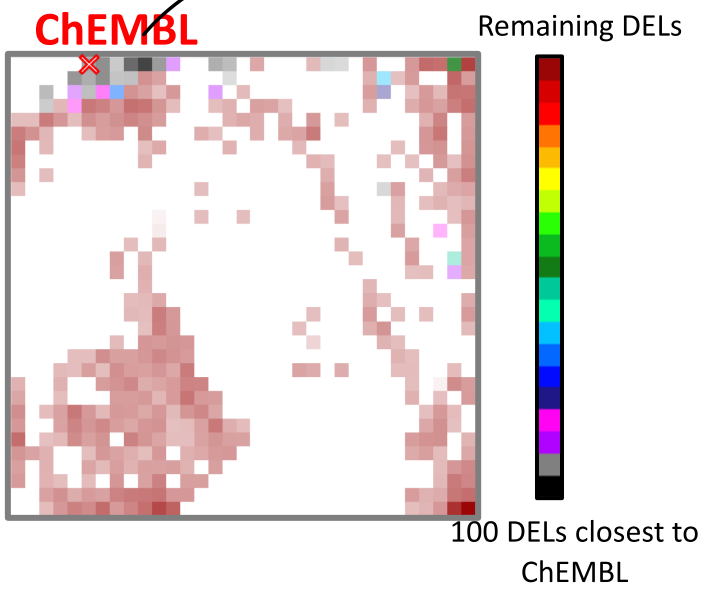
....

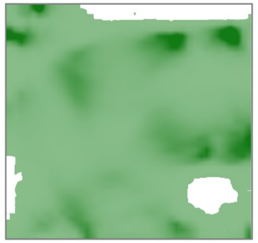
DEL2497



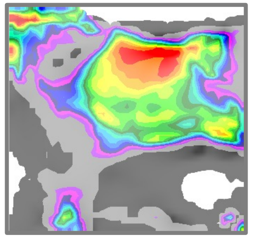
Vector 2497

Meta-GTM

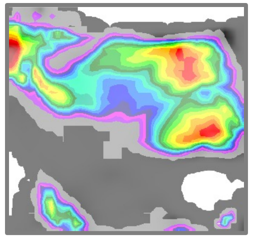




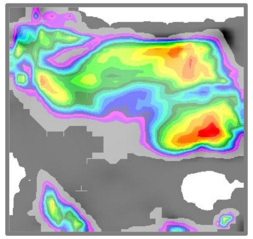
ChEMBL



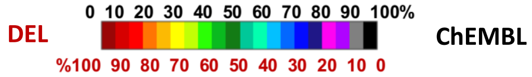
DEL845



DEL309

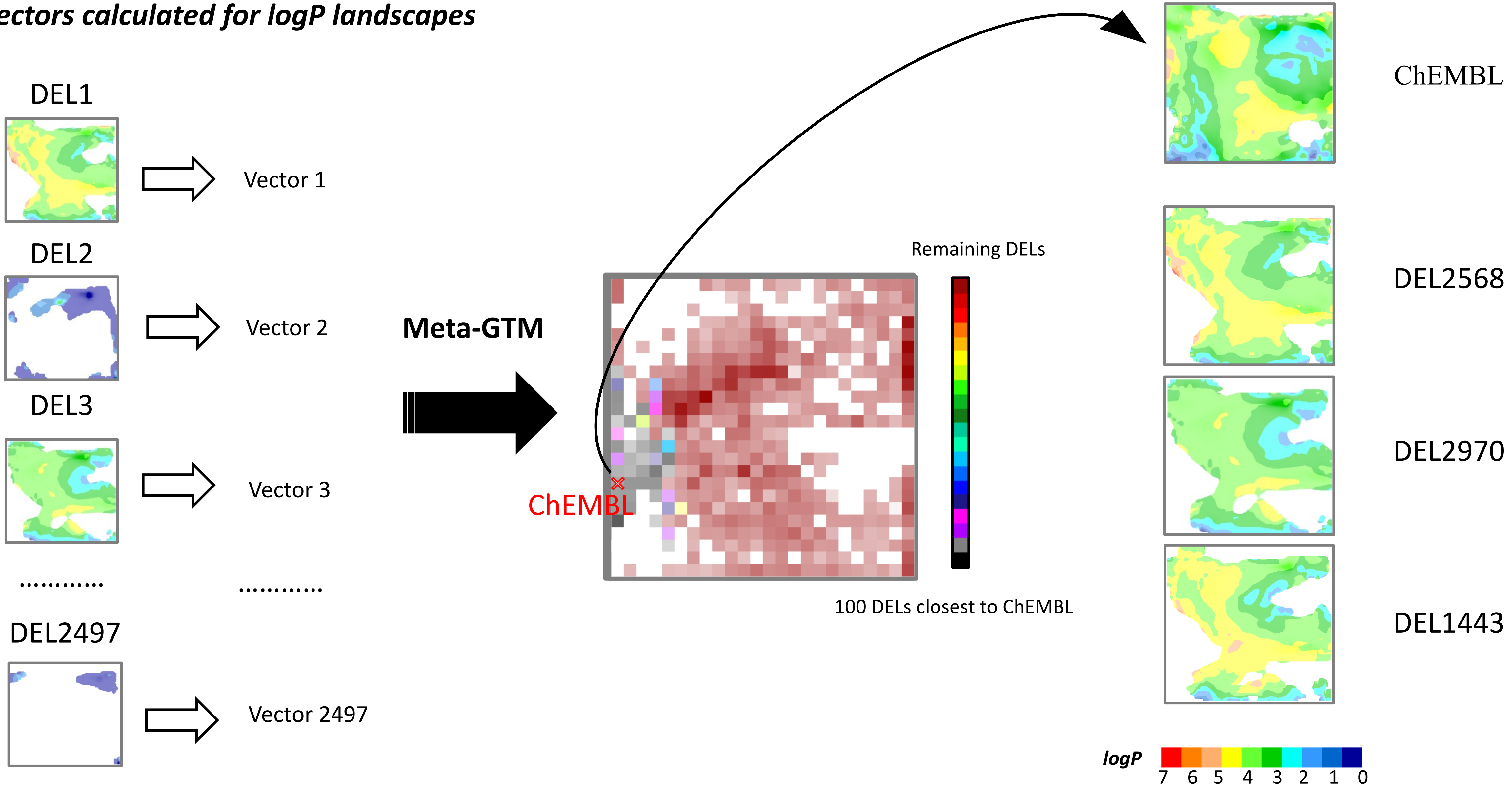


DEL1108



Meta-GTM built on property-modulated descriptors

Vectors calculated for logP landscapes



DEL1 → Vector 1

DEL2 → Vector 2

DEL3 → Vector 3

………… …………

DEL2497 → Vector 2497

Meta-GTM

Remaining DELs

ChEMBL

100 DELs closest to ChEMBL

ChEMBL

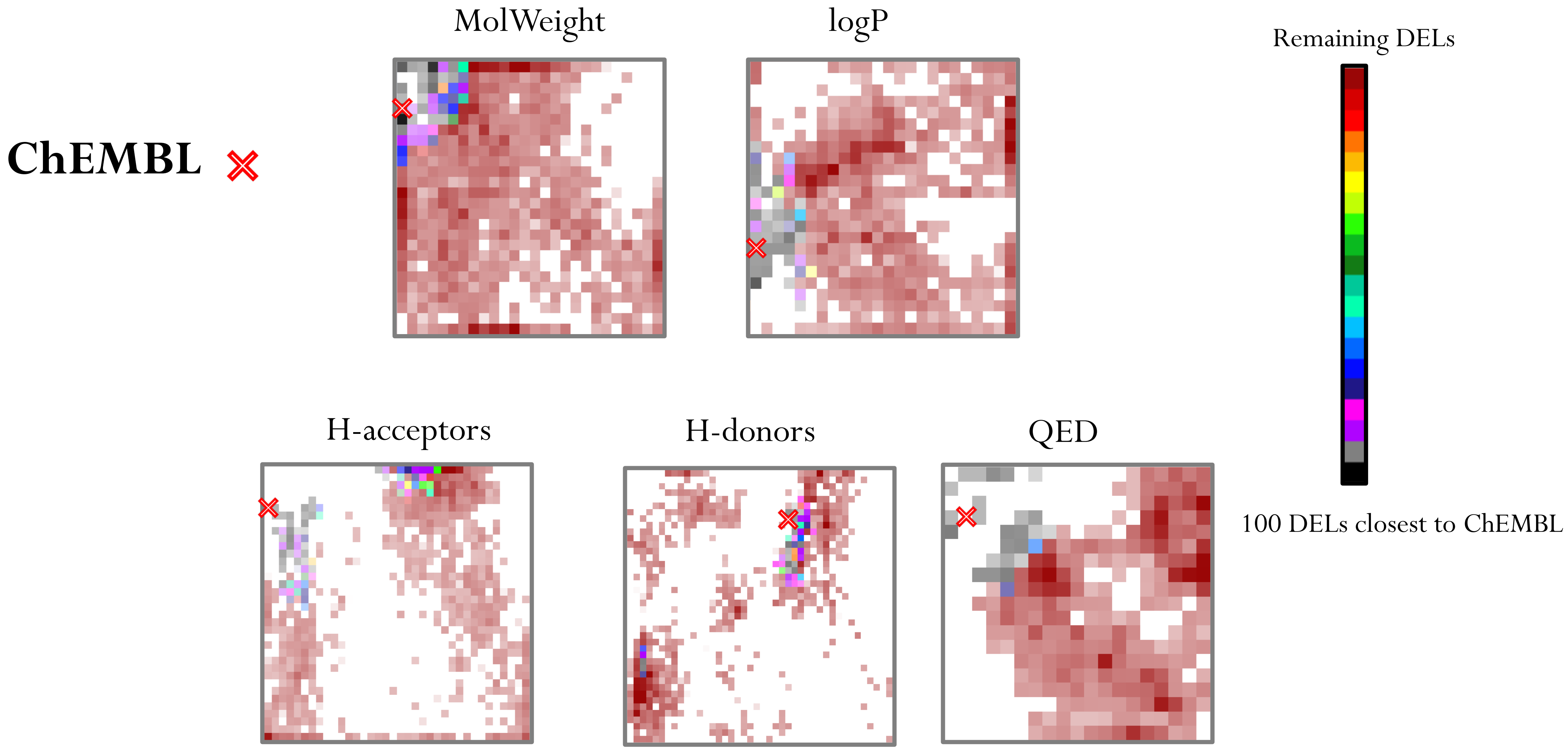
DEL2568

DEL2970

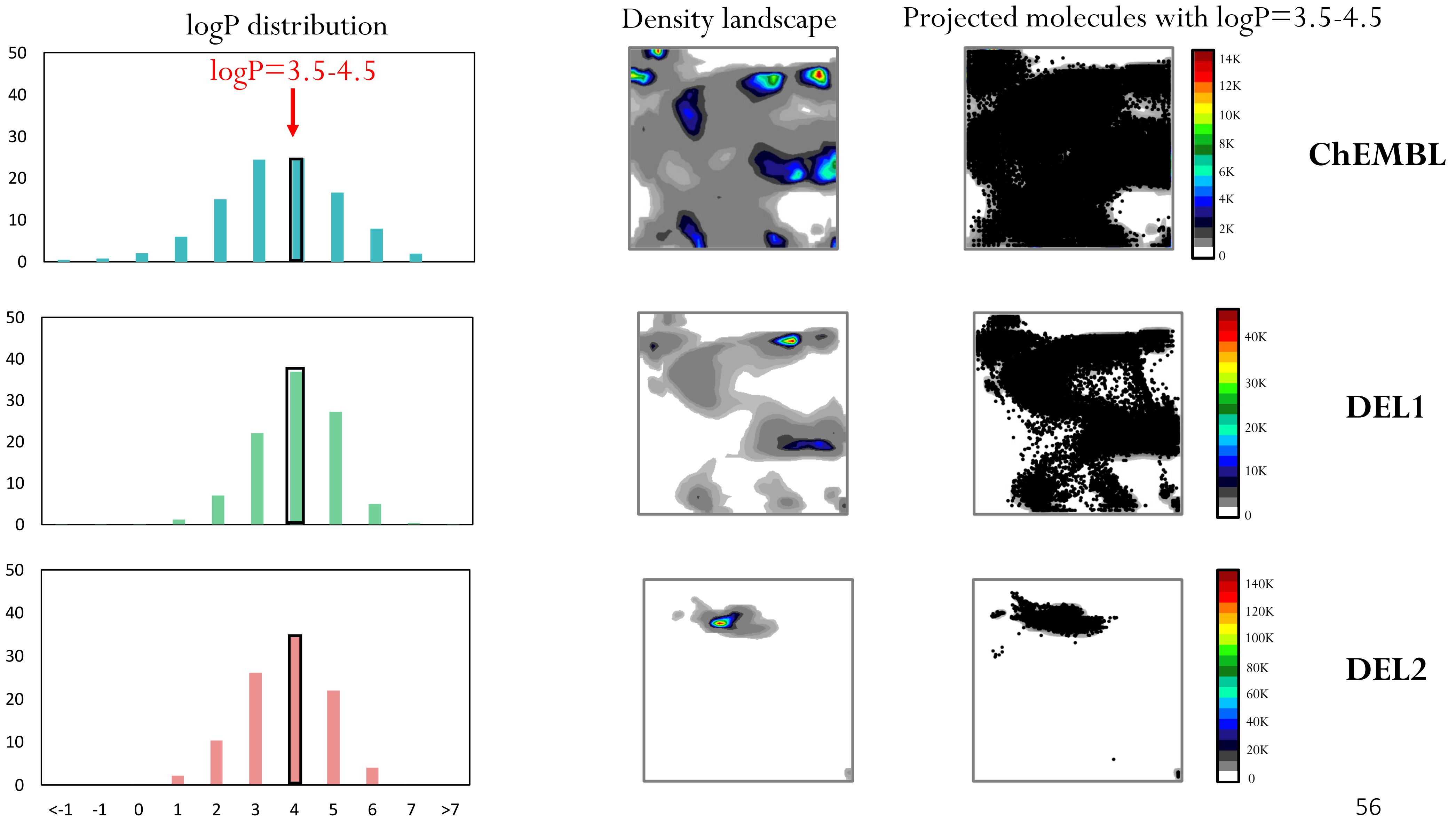
DEL1443

logP 7 6 5 4 3 2 1 0

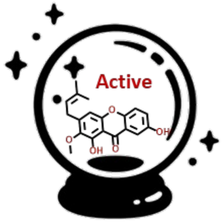
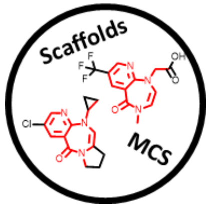
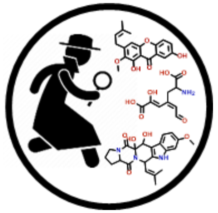
Meta-GTM built on property-modulated descriptors



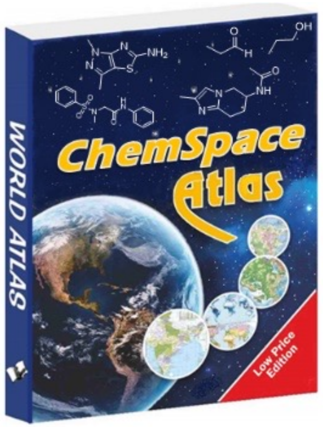
Linear vs chemographic property distribution



Online tools for Big Data analysis

Big Data compatible		Data visualization	Properties analysis	Activity analysis	QSAR/ QSPR	Structural analysis	Analogue search
							
10^3	**PUMA** (PCA)	✓	✓	✓	✗	✗	✓
10^5	**ChemMaps** (PCA)	✓	✓	✗	✓ Toxicity	✗	✓
10^7	**tMAP** (TMAP)	✓	✓	✓	✗	✗	✗
10^{10}	**InfiniSee** (Ftrees)	✗	✗	✗	✗	✓	✓

Chemical space size



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, **62**, 4537–4548

Big Data compatible

Data visualization

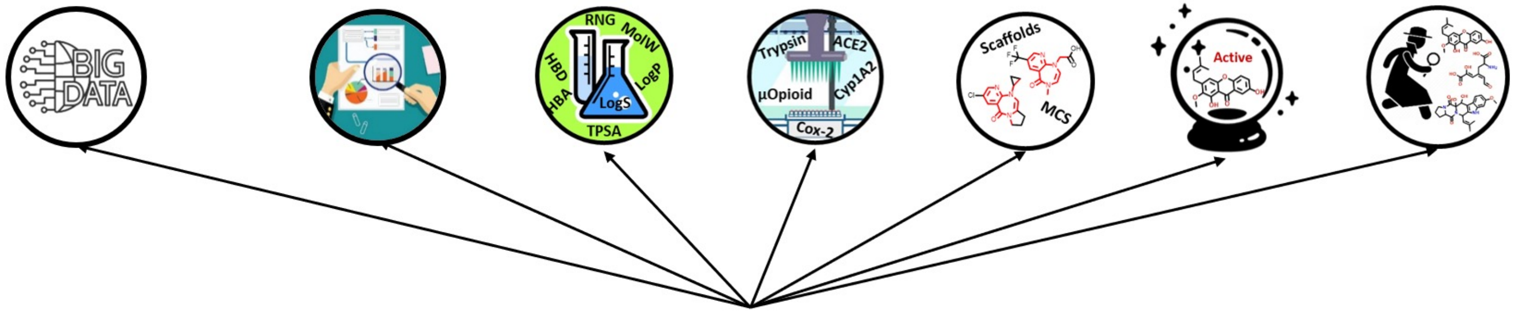
Properties analysis

Activity analysis

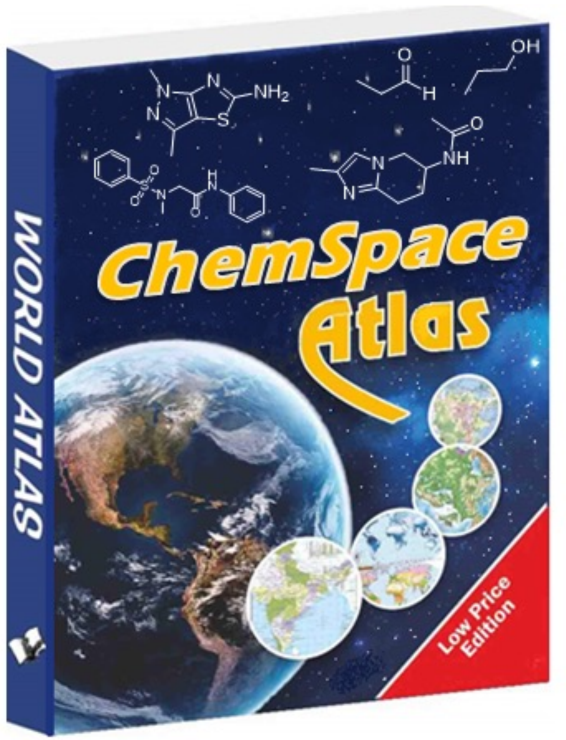
Structural analysis

QSAR/ QSPR

Analogues search



ChemSpace Atlas tool



Main features

- polyvalent tool based on the GTM Universal Maps
- accommodates > 1.5 billion compounds
- assembles > 40.000 hierarchically related maps of different scale

Main options

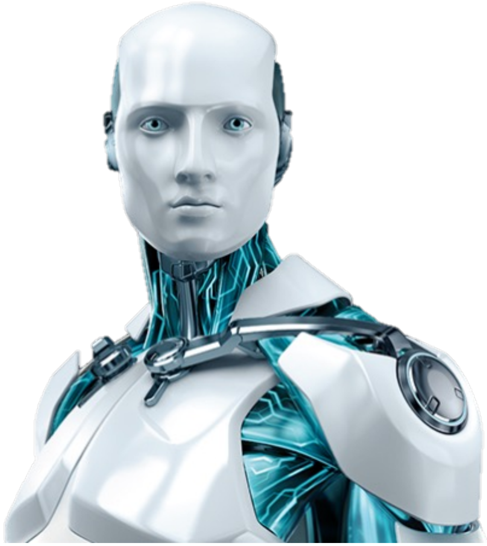
- Data visualization, search, subsets selection
- Automated extraction of Maximal Common Substructures
- Scaffold analysis
- Projection of new compounds
- Pharmacological profiling with respect to >700 biological targets

Libraries

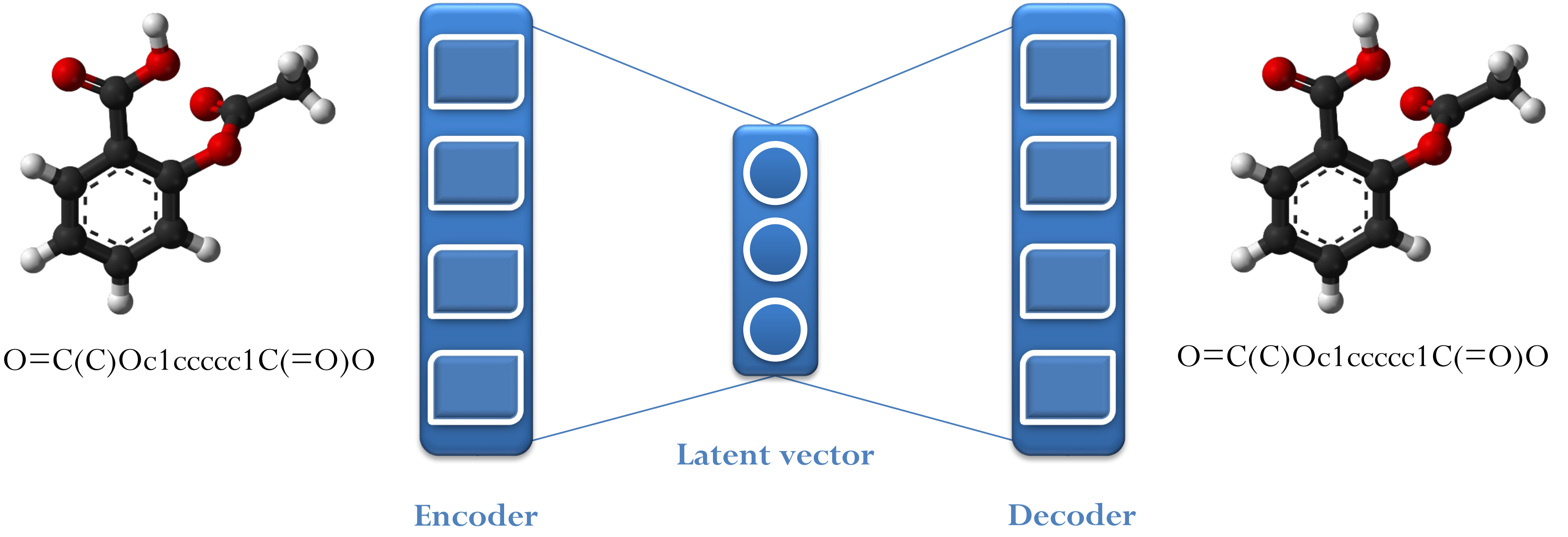
- Screening Compounds
- Natural products and their analogues

Yu. Zabolotna et al. J. Chem. Inf. Mod., 2022, **62**, 4537–4548

De novo design of biological active molecules using Artificial Intelligence tools



Autoencoder performing SMILES reconstruction



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

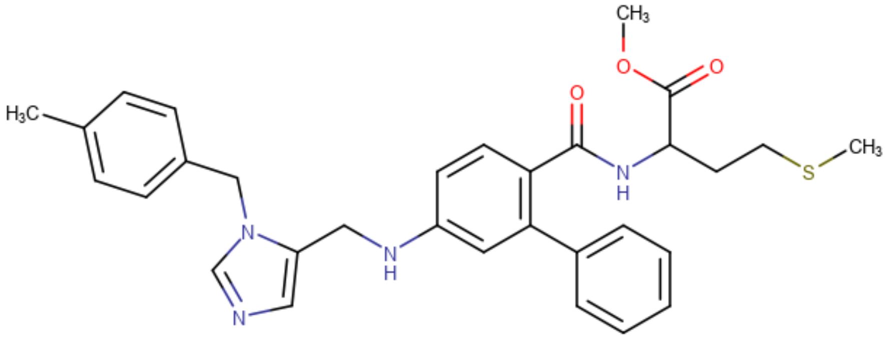
Encoder

Latent vector

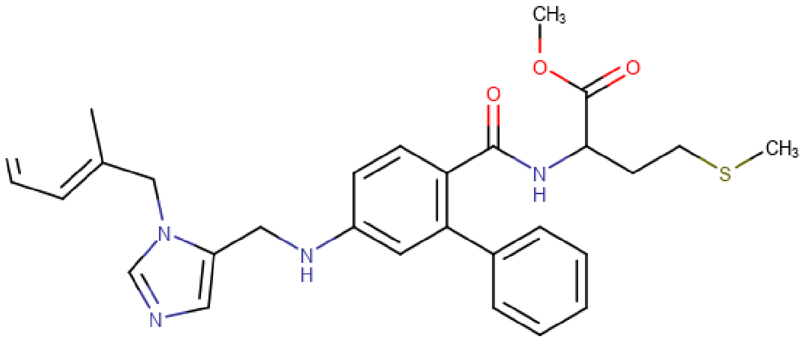
Decoder

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

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



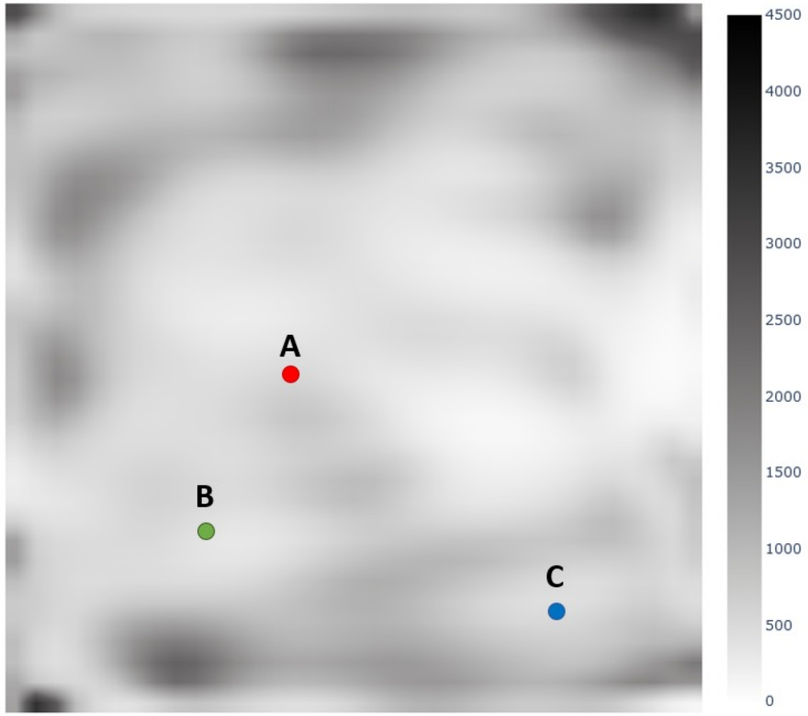
dependence is a problem!



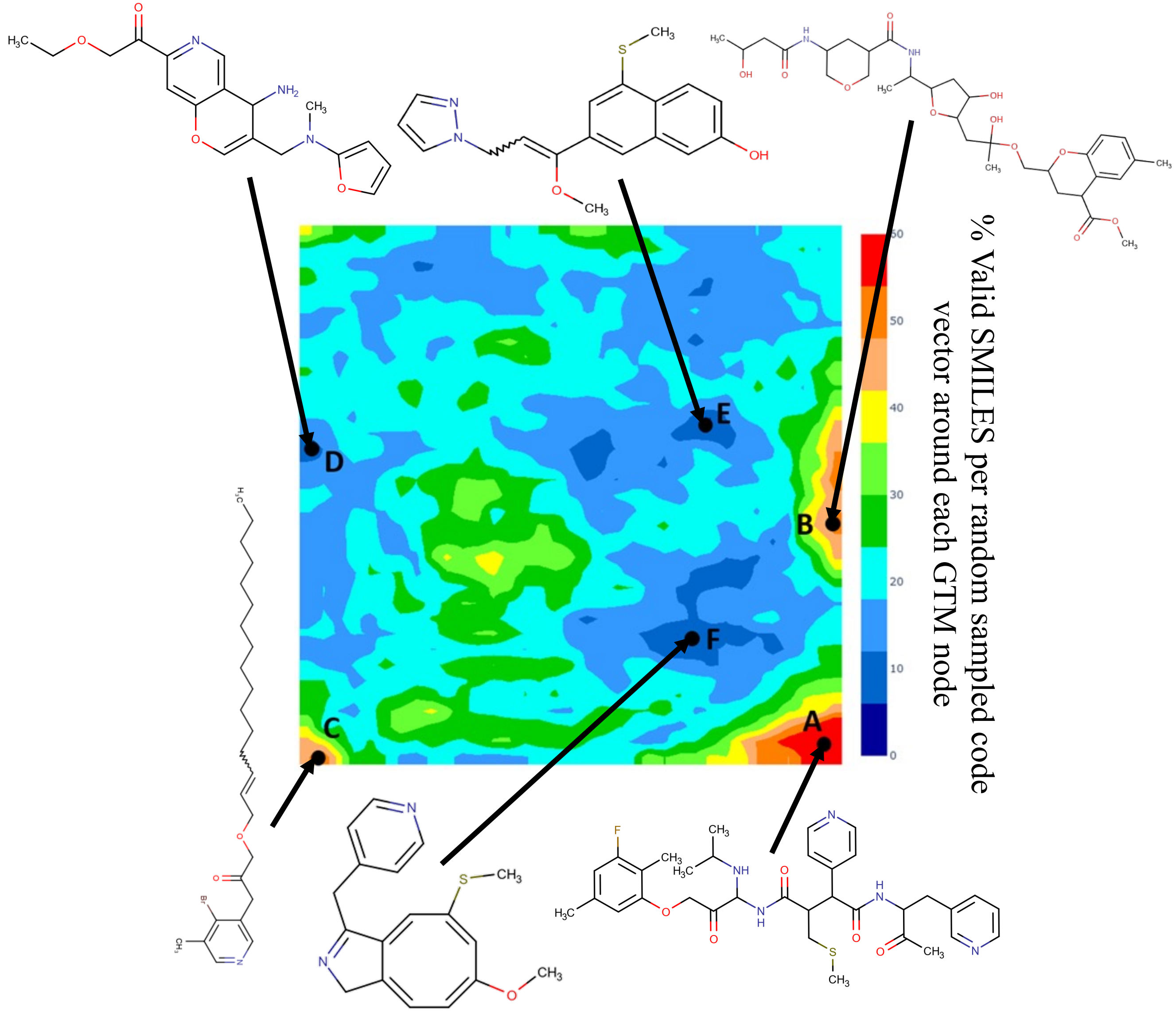
A: c1cc(C(NC(C(OC)=O)CCSC)=O)c(-c2ccccc2)cc1NCc1cncn1Cc1ccc(C)cc1

B: N(Cc1cncn1Cc1ccc(C)cc1)c1cc(-c2ccccc2)c(C(=O)NC(C(OC)=O)CCSC)cc1

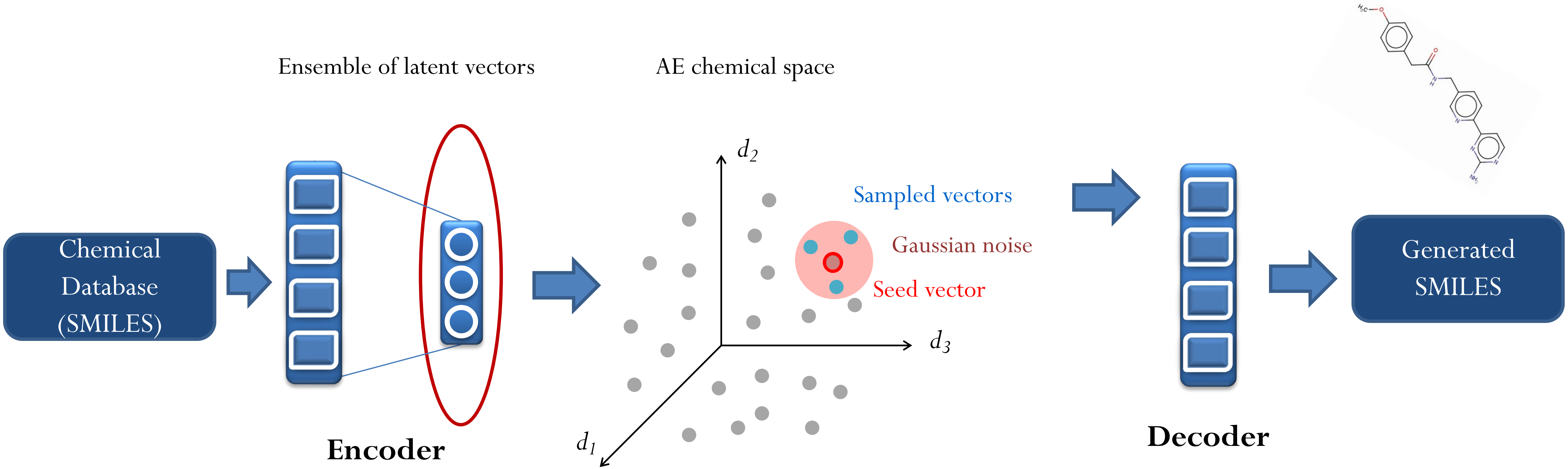
C: c1cc(C)ccc1Cn1cncc1CNc1ccc(C(=O)NC(CCSC)C(OC)=O)c(-c2ccccc2)c1



Latent Vector Reversibility is a Regional Property!

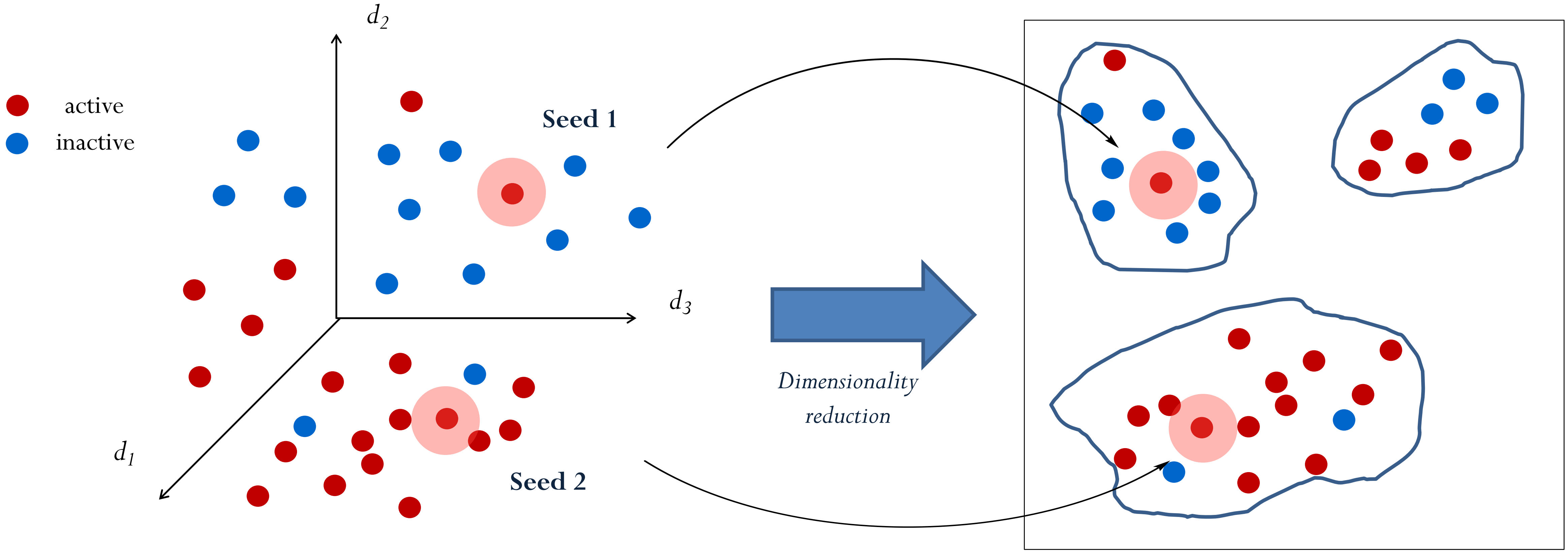


AutoEncoder: sampling using a seed vector

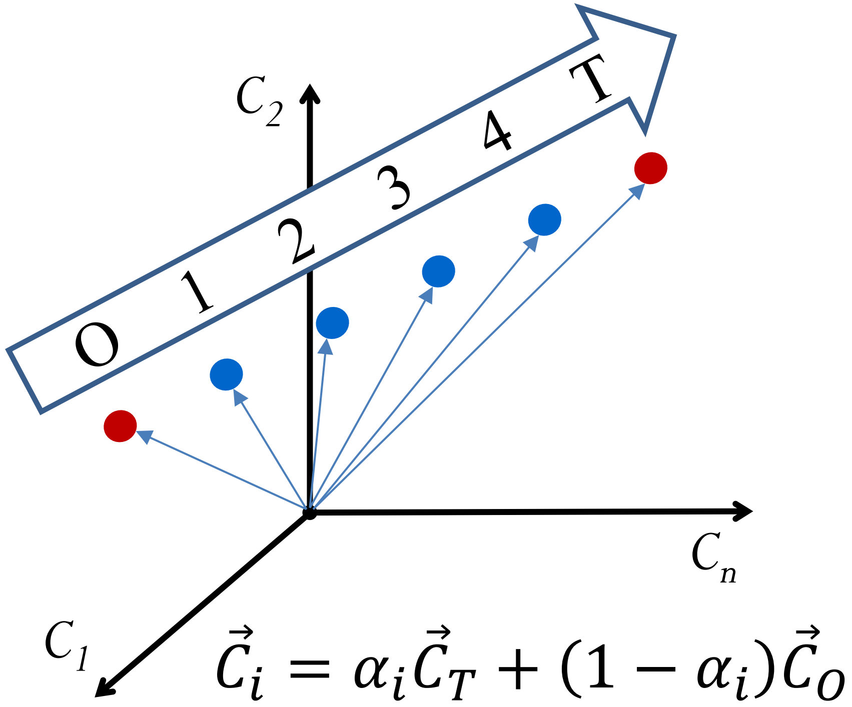


Goal: to identify a seed vector from which valid structures possessing a given activity can be generated

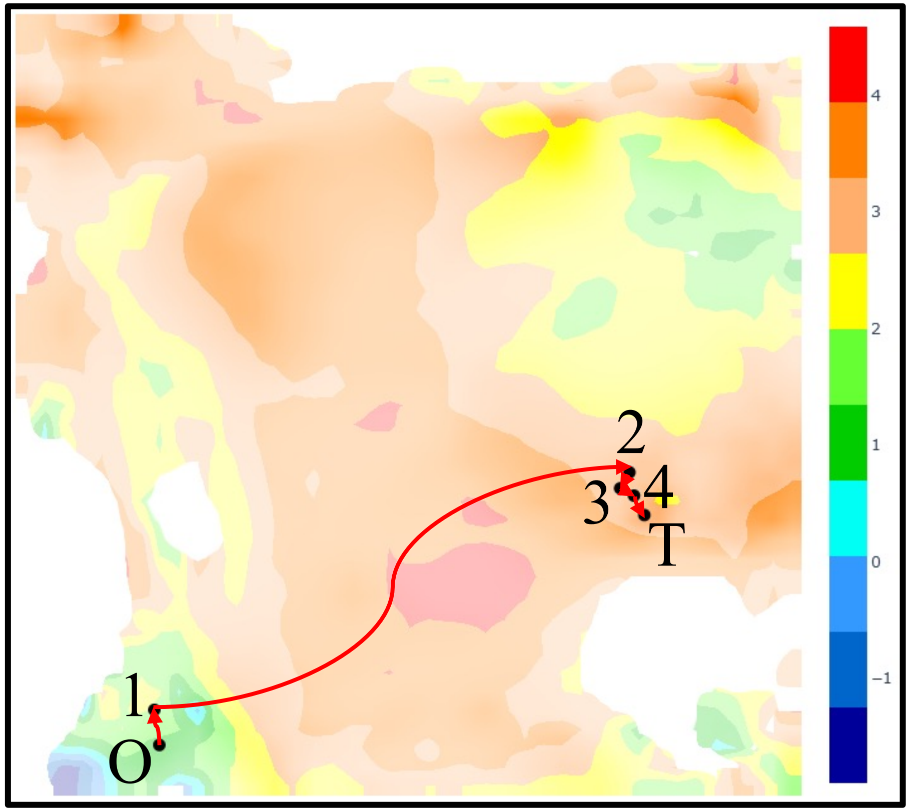
AutoEncoder chemical space: choice of a seed vector

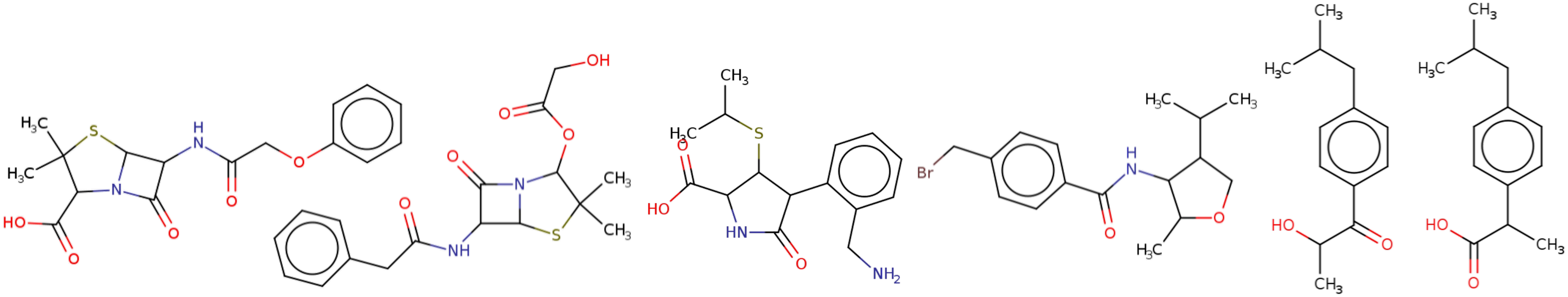


Sampling from the **Seed 2** (belonging to a cluster of actives) has more chance to generate active molecules than from the **Seed 1** (singleton)



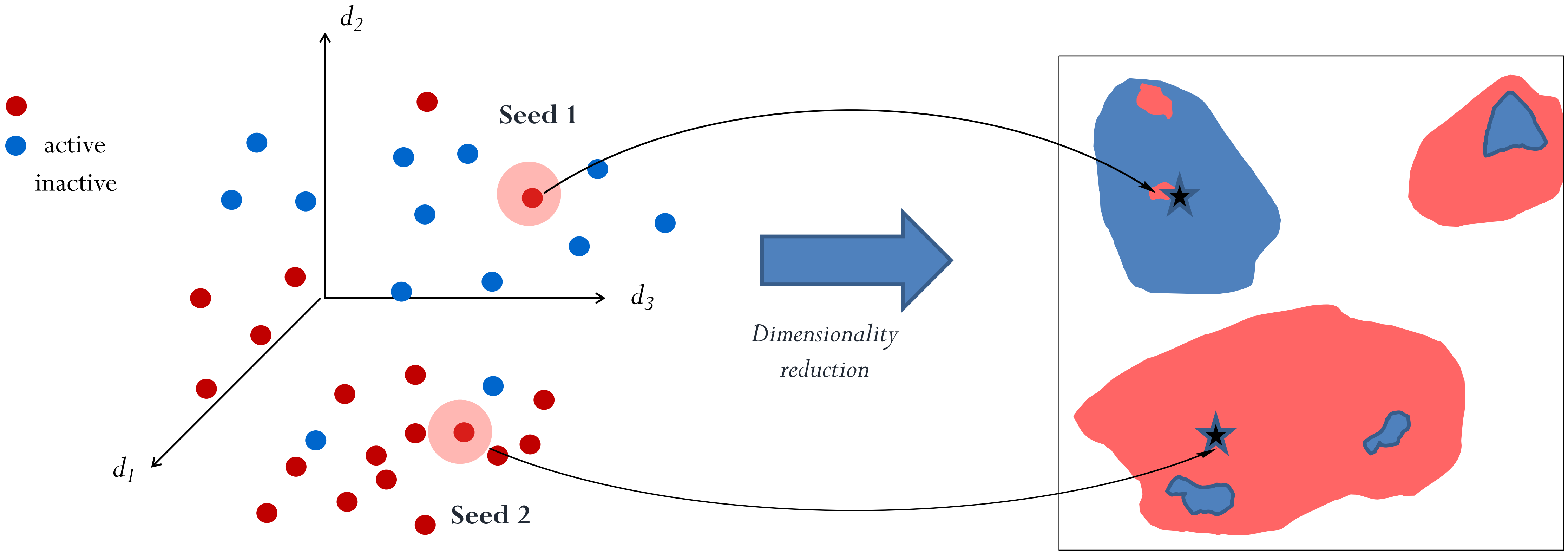
$$\vec{C}_i = \alpha_i \vec{C}_T + (1 - \alpha_i)\vec{C}_O$$





Origin: 1 2 3 4 Target:

AutoEncoder chemical space analysis with GTM

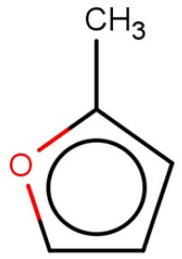


Generative Topographic Map showing data (activity) distribution

Generative Topographic Map (GTM) can be used for seed selection, chemical space exploration and activity prediction

AutoEncoder: Sequence-to-sequence Architecture

ChEMBL



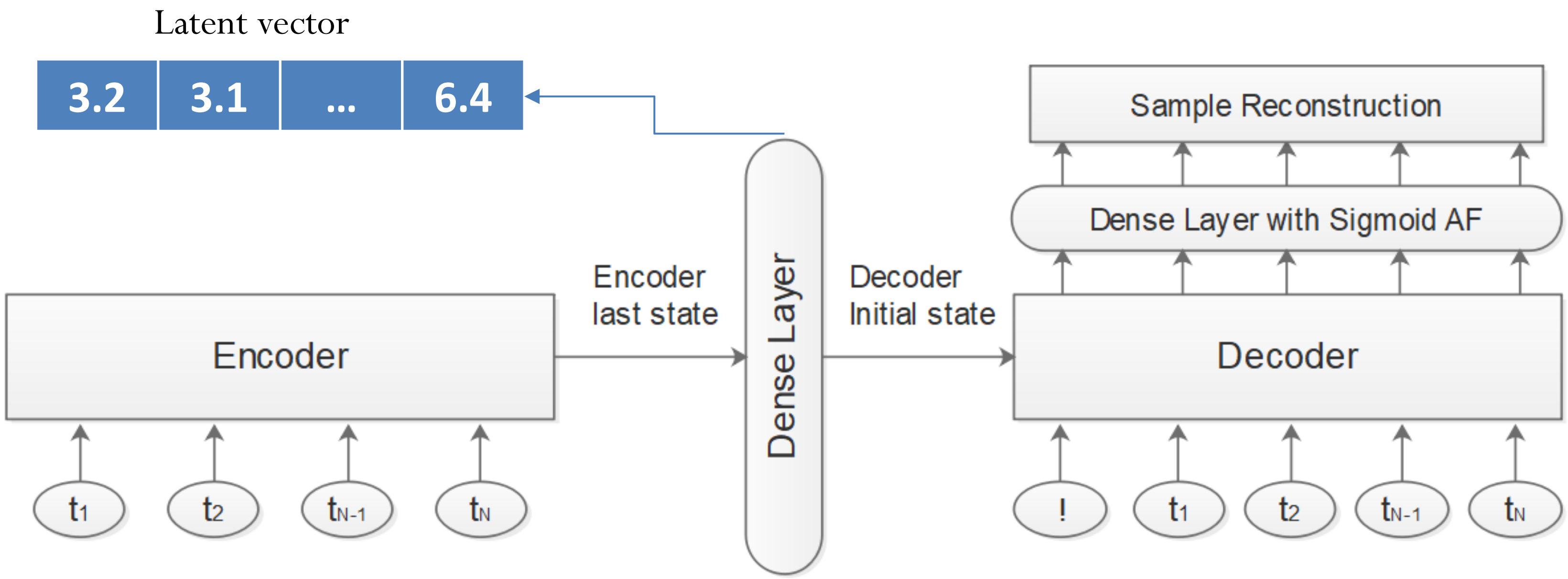
SMILES string

characters

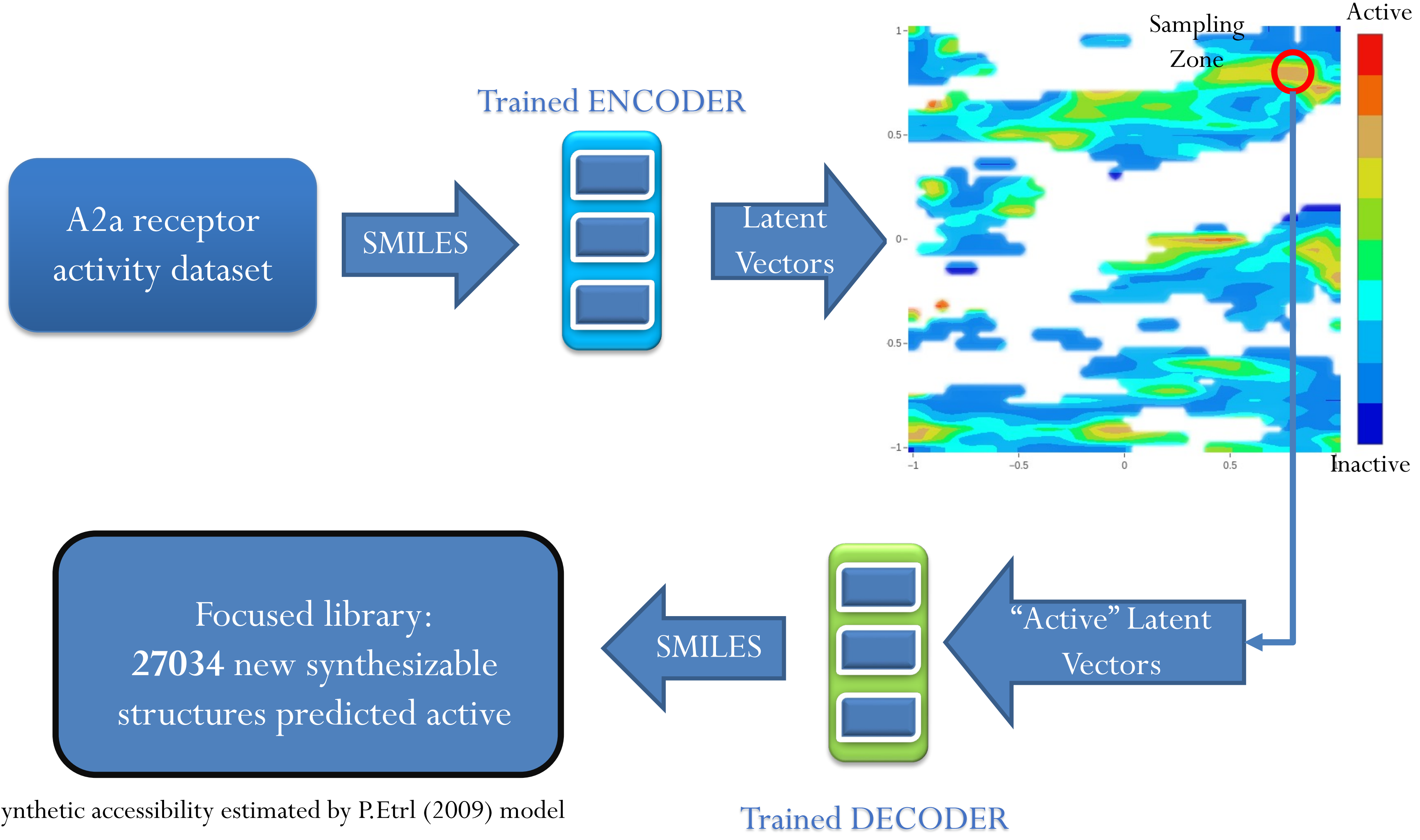
	o	1	c	(	c	c	c	1	)	C
c	0	0	1	0	1	1	1	0	0	0
1	0	1	0	0	0	0	0	1	0	0
(	0	0	0	1	0	0	0	0	0	0
)	0	0	0	0	0	0	0	0	1	0
o	1	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	1

One – hot encoding

- Sequence-to-sequence autoencoder
- One-hot representation of SMILES
- Bidirectional Long Short-Term Memory (LSTM) encoder
- Unidirectional LSTM decoder
- Latent vector of 128 components



Generation of the focused library for Adenosine A2a receptor



A2a receptor activity dataset

SMILES

Trained ENCODER

Latent Vectors

"Active" Latent Vectors

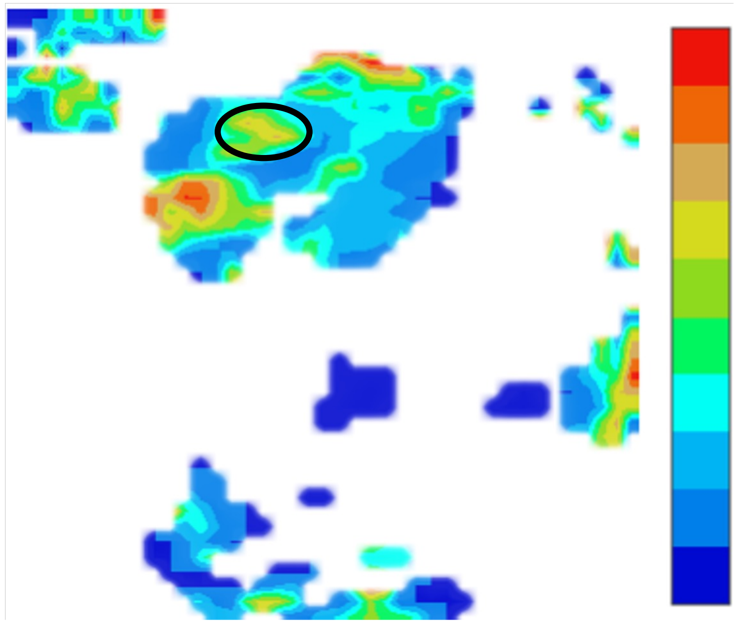
Trained DECODER

SMILES

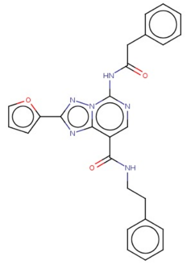
Focused library:
27034 new synthesizable structures predicted active

*synthetic accessibility estimated by P.Etrl (2009) model

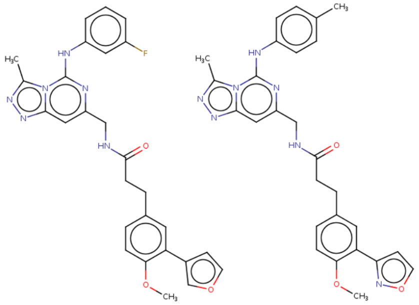
Case study: Generation of inhibitors of A2a receptor



Existing



de novo



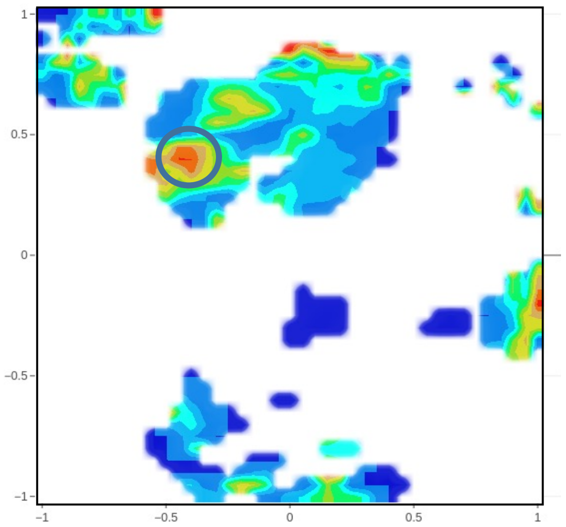
- *Generated structures are enriched with new scaffolds*
- *According to docking experiments they are efficiently able to bind A2a*

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

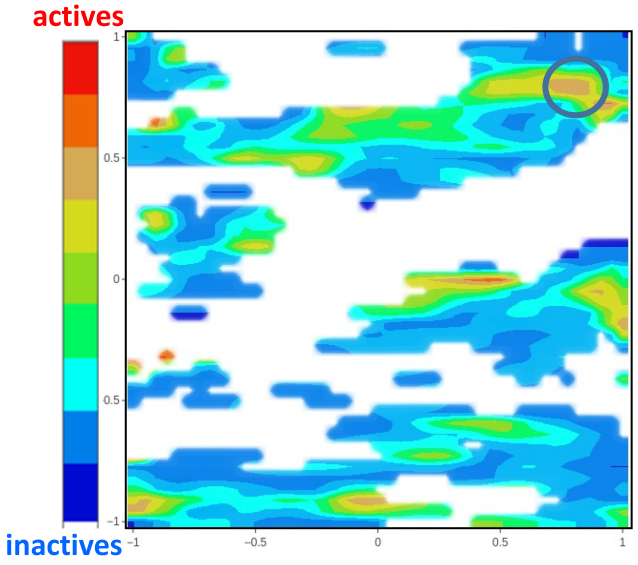
Soving the inverse-QSAR problem using a Conditional Variational Autoencoder

AutoEncoder *vs* Molecular descriptors space

ISIDA molecular descriptors space



Autoencoder latent space



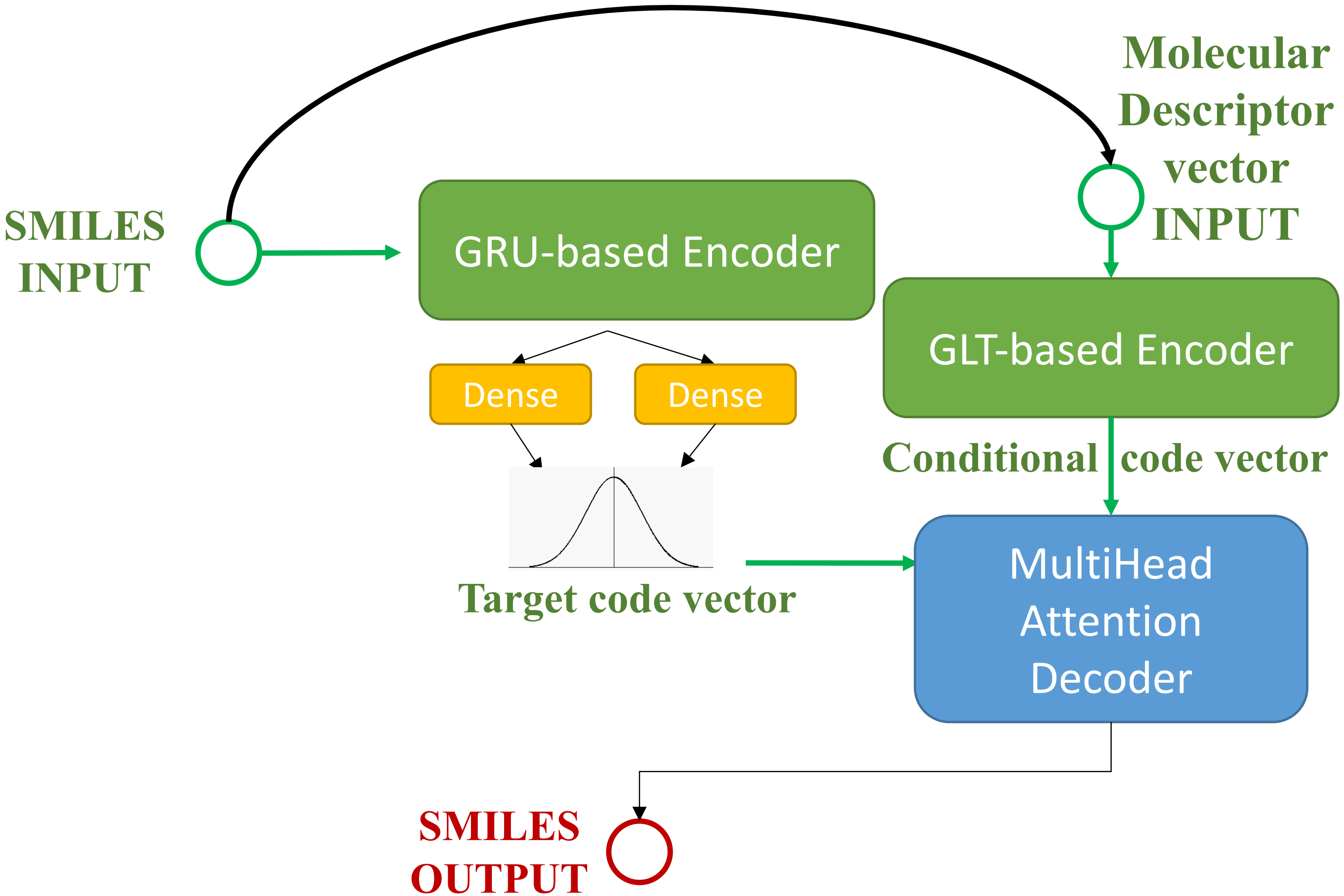
GTM Class landscapes for A2a-receptors binders (1303 actives and 3618 inactives)

- activity prediction
- no structure generation

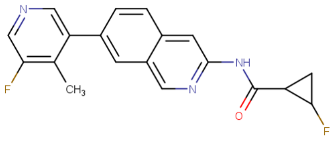
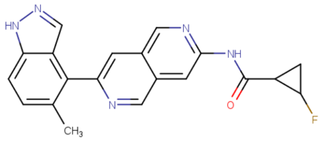
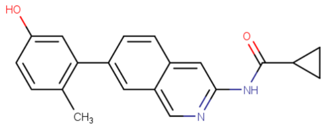
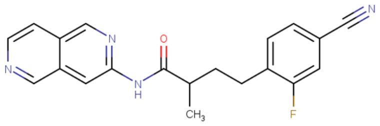
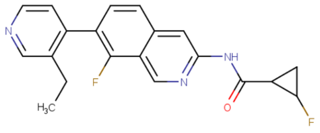
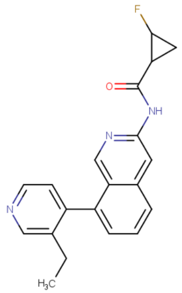
- activity prediction
- generation of new structures

<u>Goal</u>*:* development of deep-learning architecture able to generate structures with desired activities using any descriptor space (*inverse-QSAR problem*)

Attention-based Conditional Variational Autoencoder



Inverse-QSAR with ACoVAE

ChEMBL compounds		
 10.73	 10.70	 10.70
inverse-QSAR compounds		
 10.08	 9.45	 9.35

Structures and related pK_i values of the most potent *ABL Tyrosine kinase 1* ligands from ChEMBL and their counterparts generated with the ACoVAE tool

Collaboration

- ITN Marie-Curie BigChem
- ITN Marie Curie TubInTrain
- Institute of Organic Chemistry, Kiev, Ukraine
- Chumakov Center, Moscow, Russia

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

Chémoinformatique
STRASBOURG