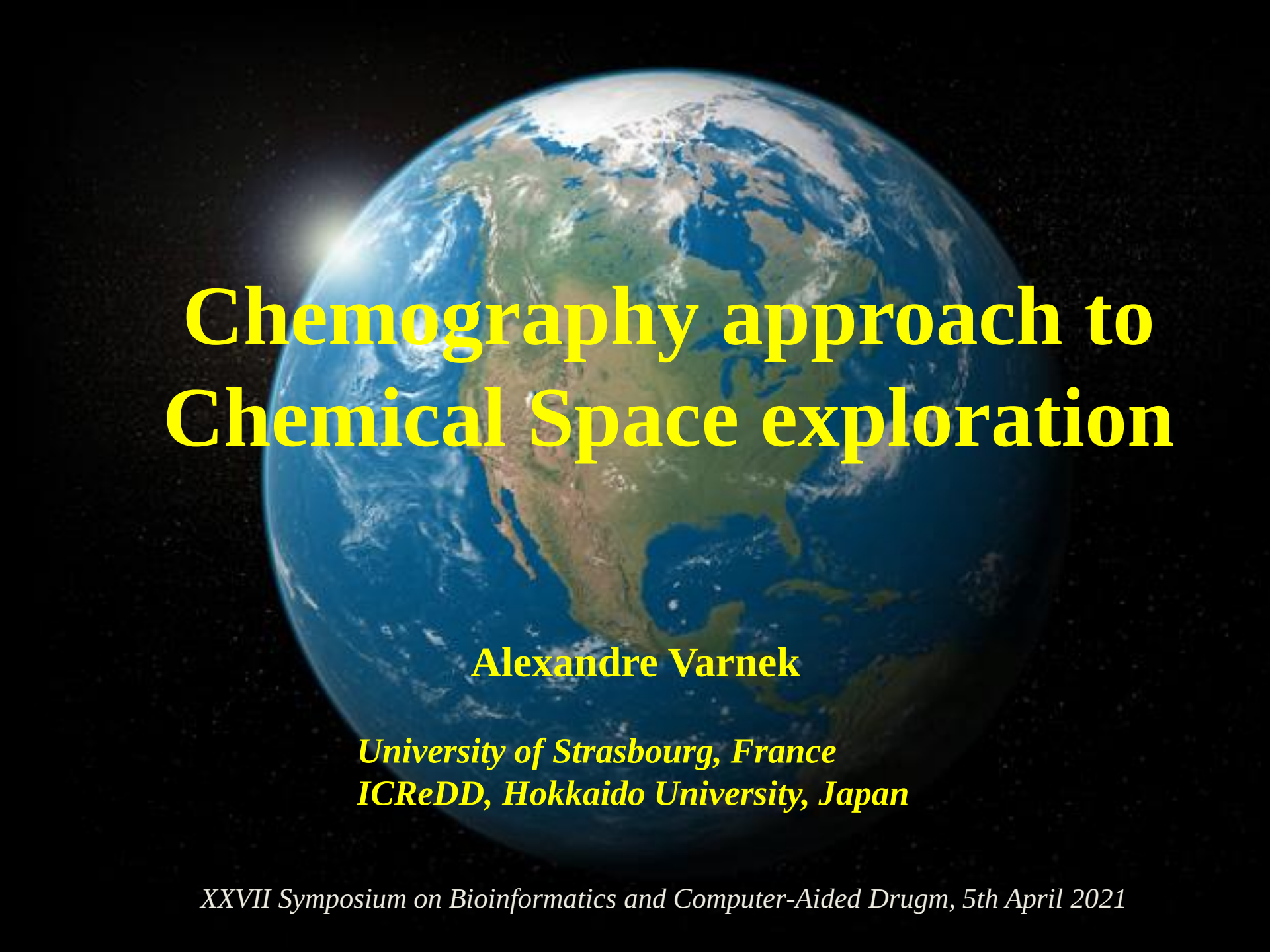


Chemography approach to Chemical Space exploration

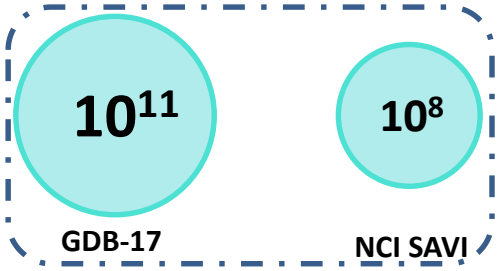
Alexandre Varnek

University of Strasbourg, France
ICReDD, Hokkaido University, Japan

XXVII Symposium on Bioinformatics and Computer-Aided Drugm, 5th April 2021

Sizes of selected chemical data collections

Public



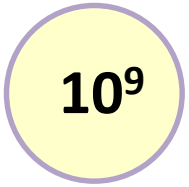
Virtual compounds

Real compounds

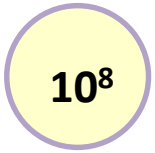
Approved drugs

~10^{4}

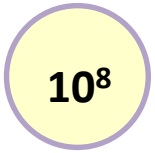
Commercial



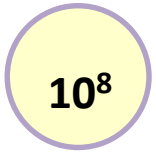
Enamine REAL Space



Enamine
REAL « Database »



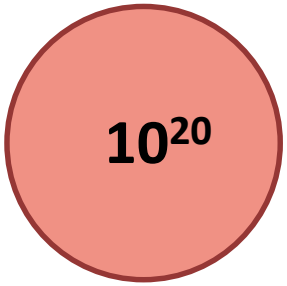
ZINC15



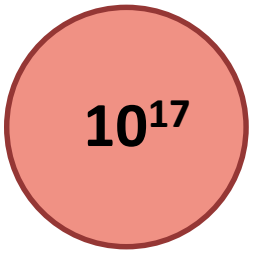
Sigma Aldrich
« in-stock »

eMolecules Plus

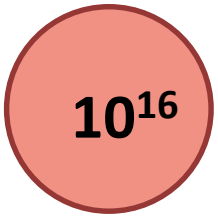
Proprietary



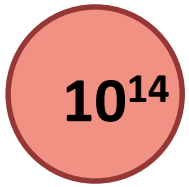
Merck MASSIV 2018



AstraZeneca 2018

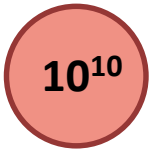


Evotec
EVOSpace 2016



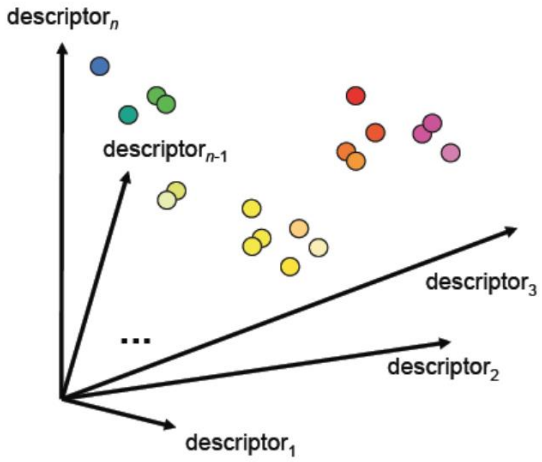
Pfizer
PGVL 2008

Boehr.-Ing.
BICLAIM 2012

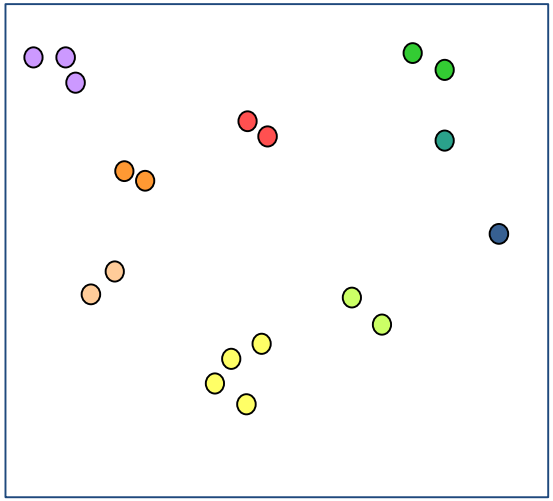


Lilly
LPC 2016

Data visualization: dimensionality reduction problem



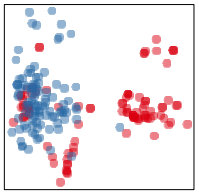
Data space
(N-dimensional)



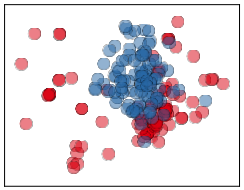
Latent space
(2-dimensional)

Dimensionality reduction methods

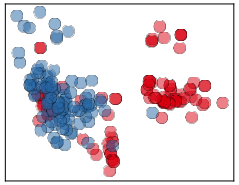
Acetylcholinesterase dataset (DUD) : 100 actives and 100 inactives
ISIDA descriptors



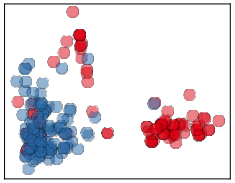
Multi-Dimensional Scaling



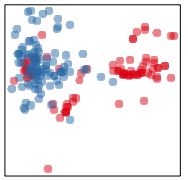
Canonical Correlation Analysis



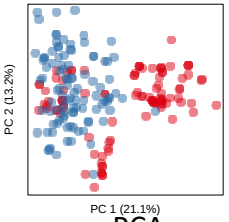
Independent Component Analysis



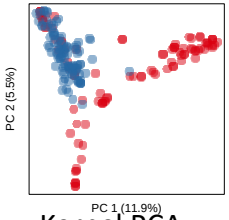
Exploratory Factor Analysis



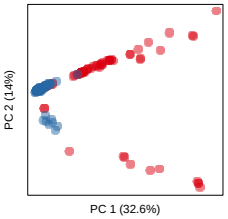
Sammon map



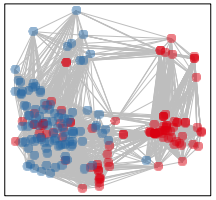
PCA



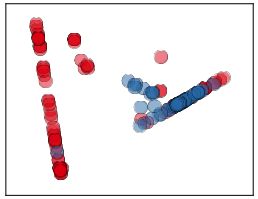
Kernel PCA (RBF kernel)



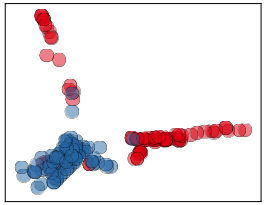
Kernel PCA (polynomial kernel)



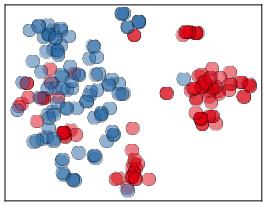
Isomap



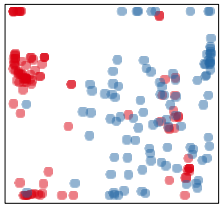
Locally Linear Embedding



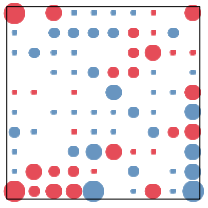
Laplacian Eigenmaps



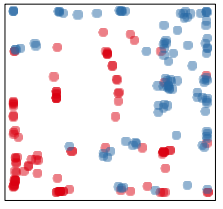
t-SNE



Autoencoder dimensionality reduction

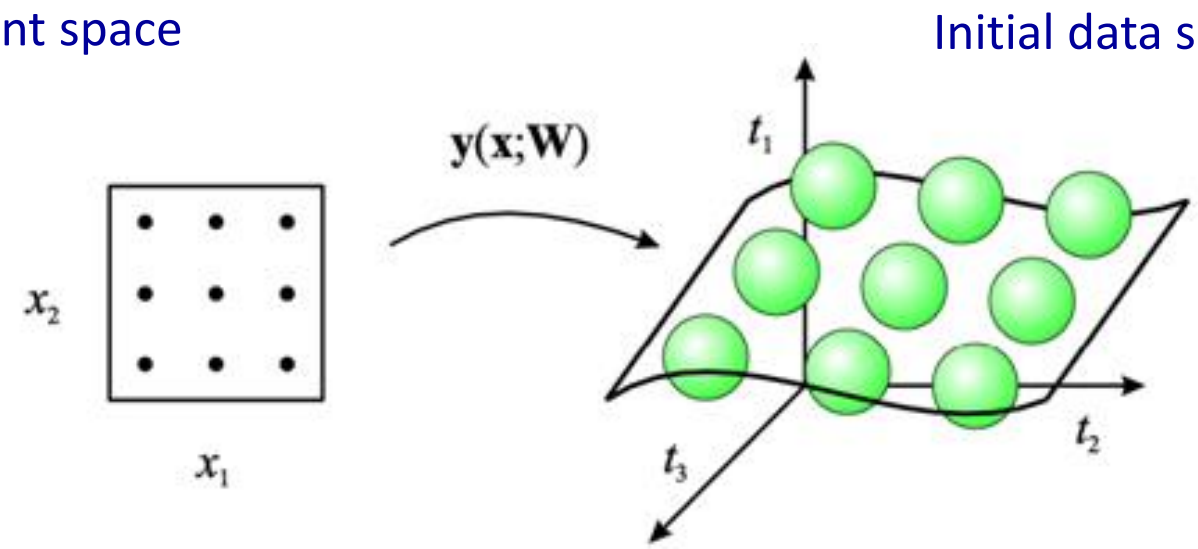


SOM



GTM

Generative Topographic Mapping



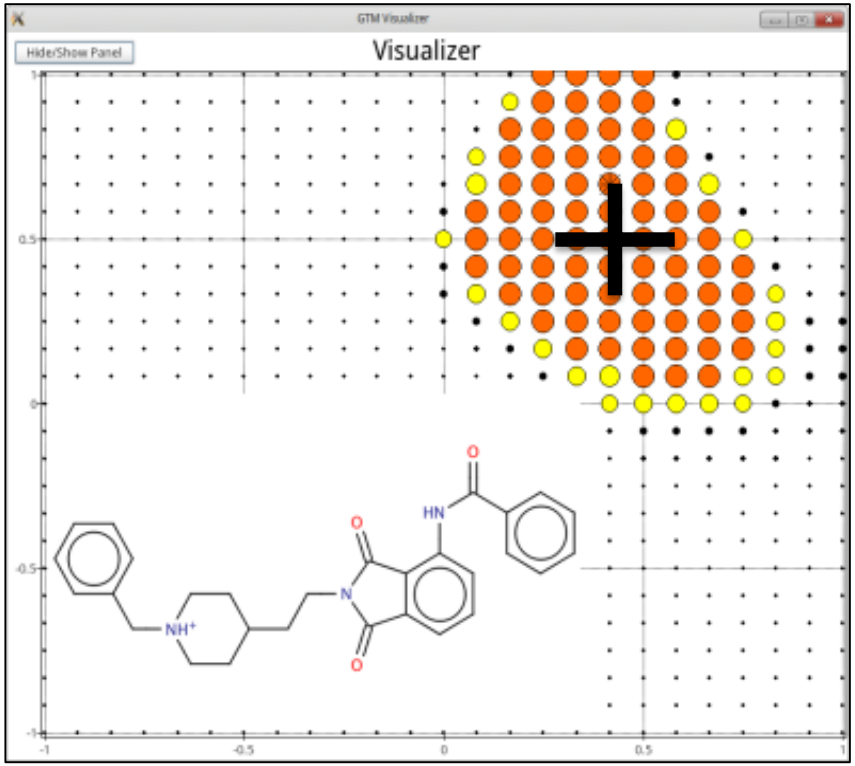
GTM generates a data probability distribution in ***both initial and latent data spaces*.**

This opens an opportunity to use GTM not only to visualize the data but also for structure-property modeling tasks

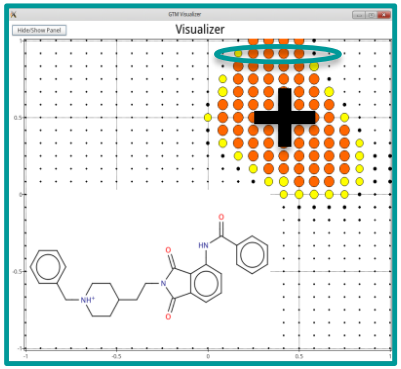
- C. M. Bishop *Pattern Recognition and Machine Learning*. 2006 Springer
- N. Kireeva. I.I. Baskin. H. A. Gaspar. D. Horvath. G. Marcou and A. Varnek. *Mol. Informatics. 2012.* 31. 201-312

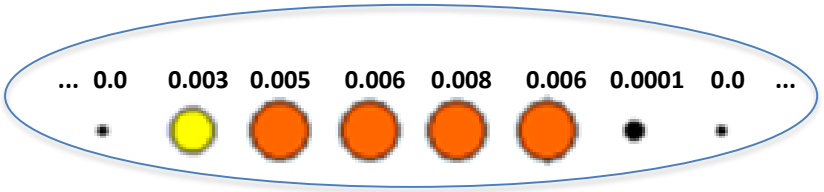
GTM: Probability Density distribution in the latent space

Projection of an object on GTM is described by the probability distribution (***responsibilities***) over the lattice nodes.



GTM descriptors for molecules and datasets





Map resolution: $N_{nodes} = K*K$

Standard setting: K =25, N_{grid} = 625

Molecule → responsibilities' vector $\{R_{tk}\}$ of N_{nodes} length

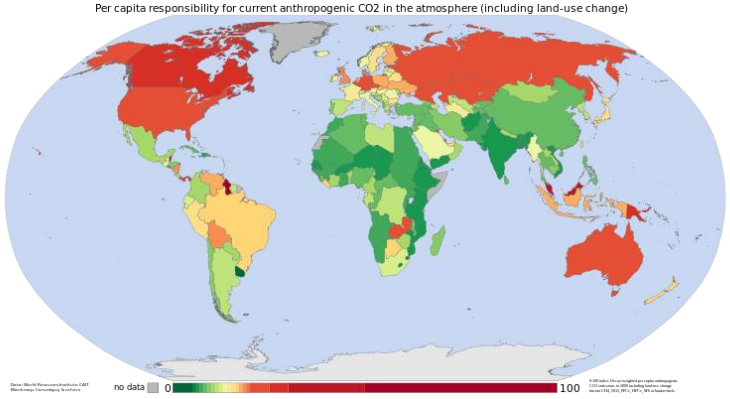
Dataset → normalized cumulated responsibilities' vector of N_{nodes} length

GTM landscapes

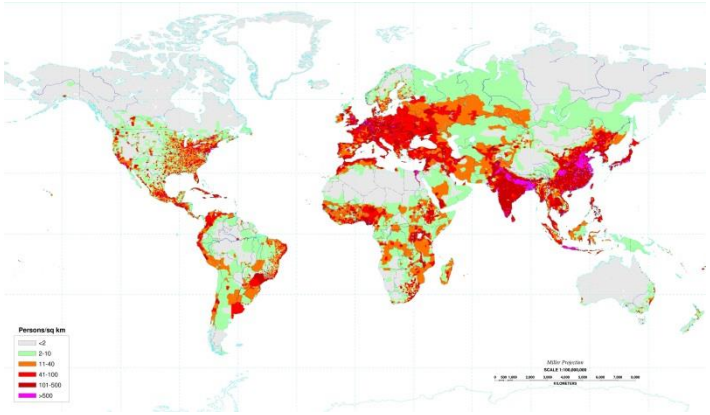
Properties mapping

political map

physical map



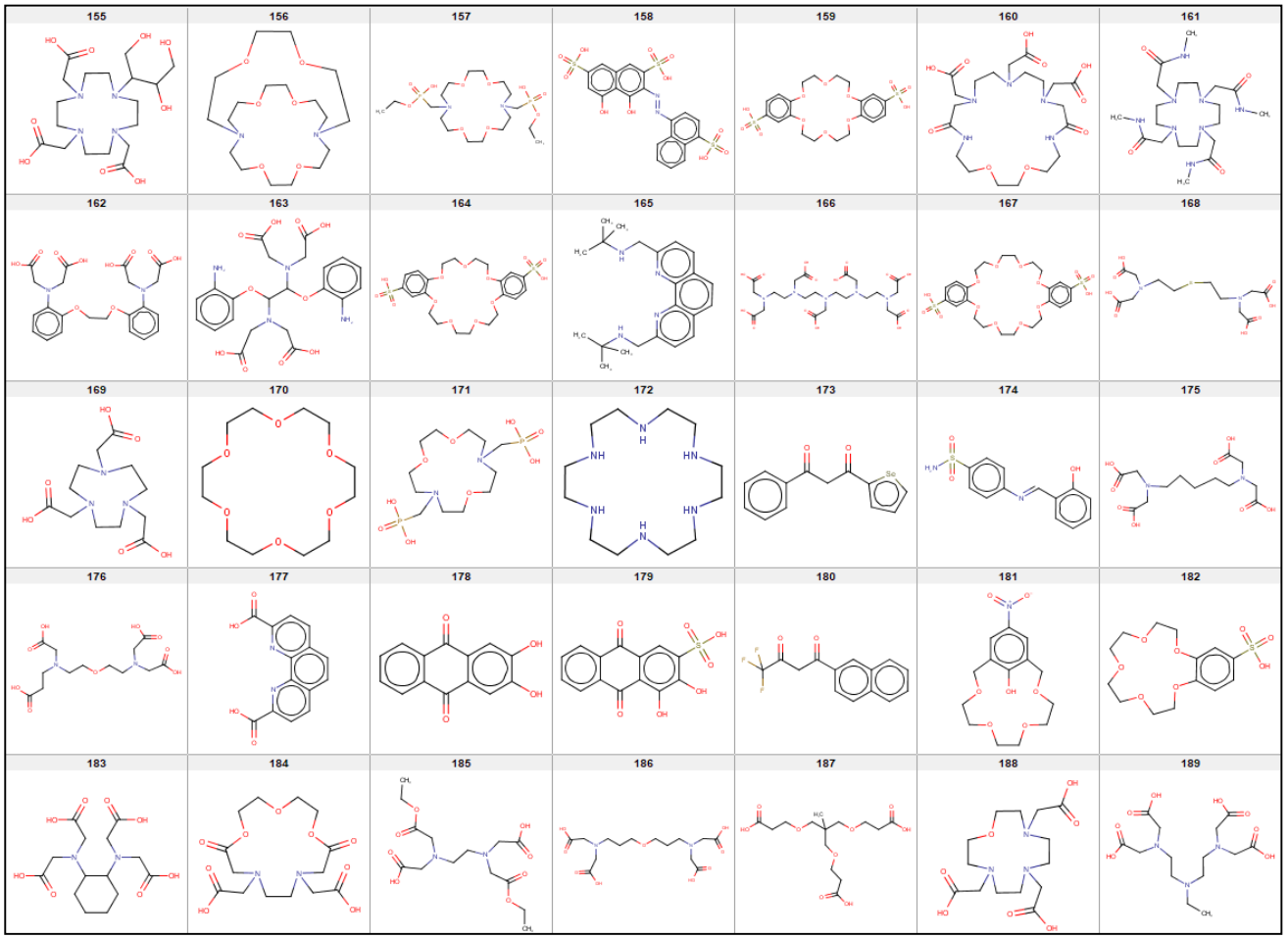
CO_2 emission



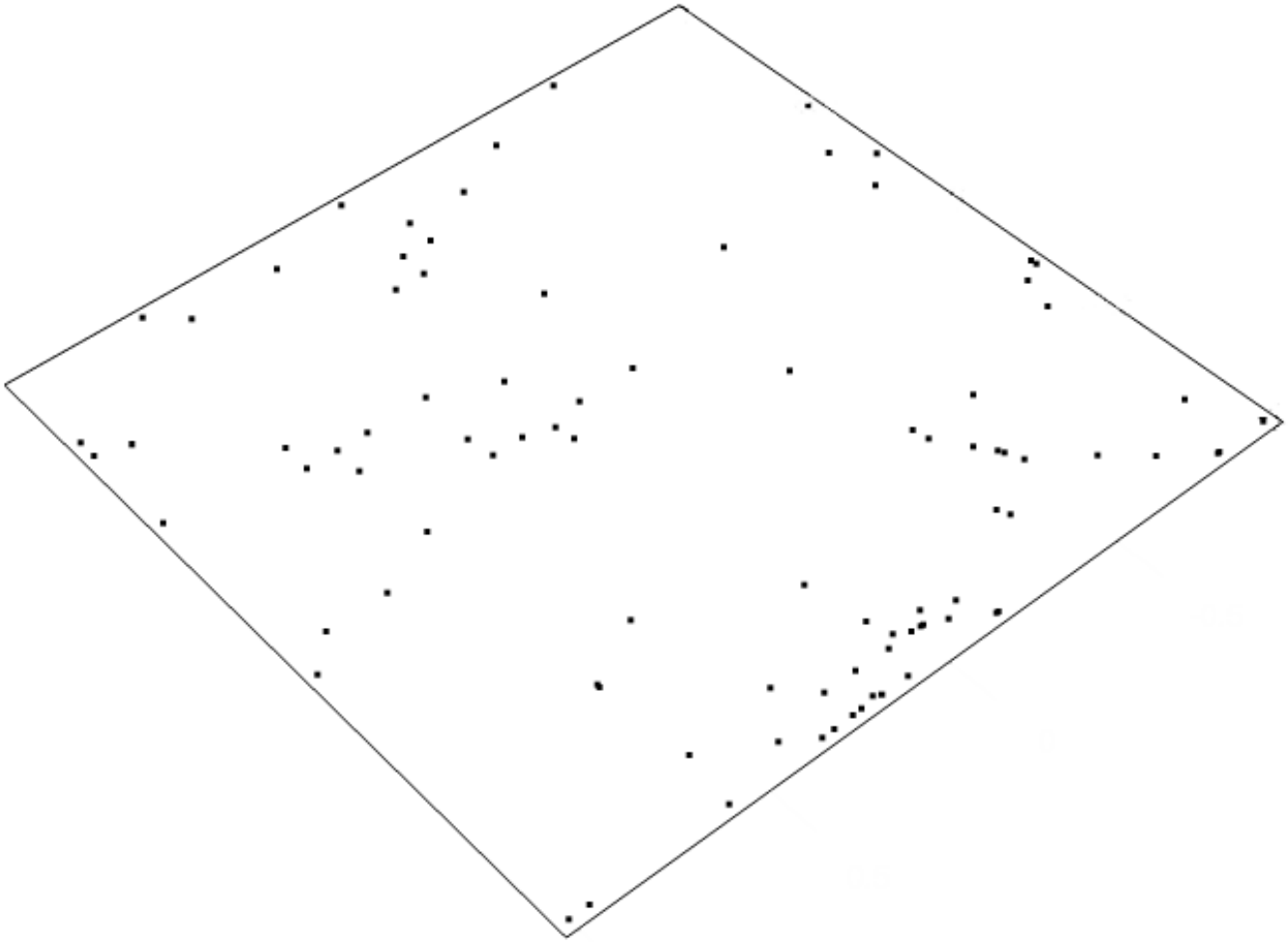
population density

Case study: chemical space of metal binders

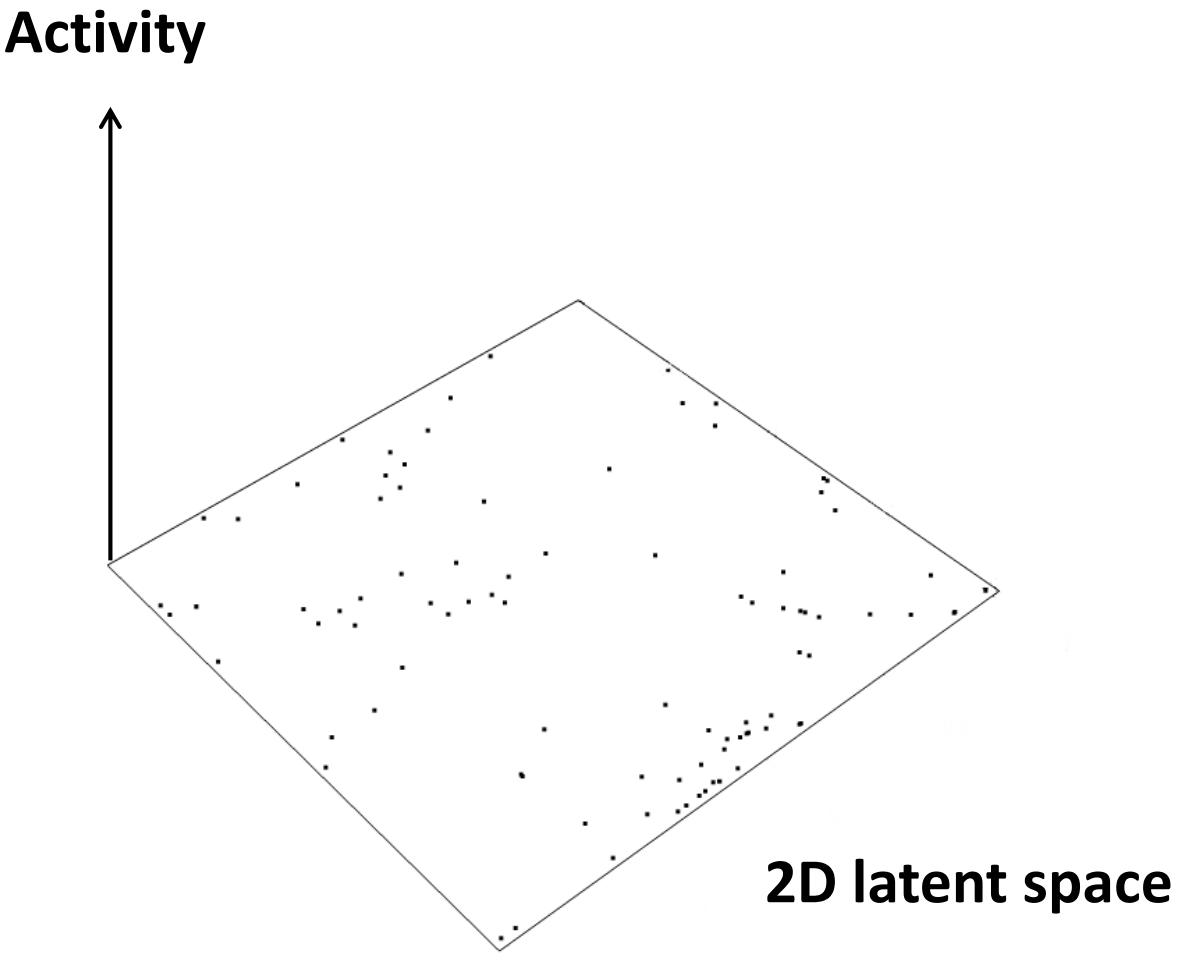
Data: 102 organic molecules which complex the Lu^{3+} cation in water



GTM of Lu^{3+} binders

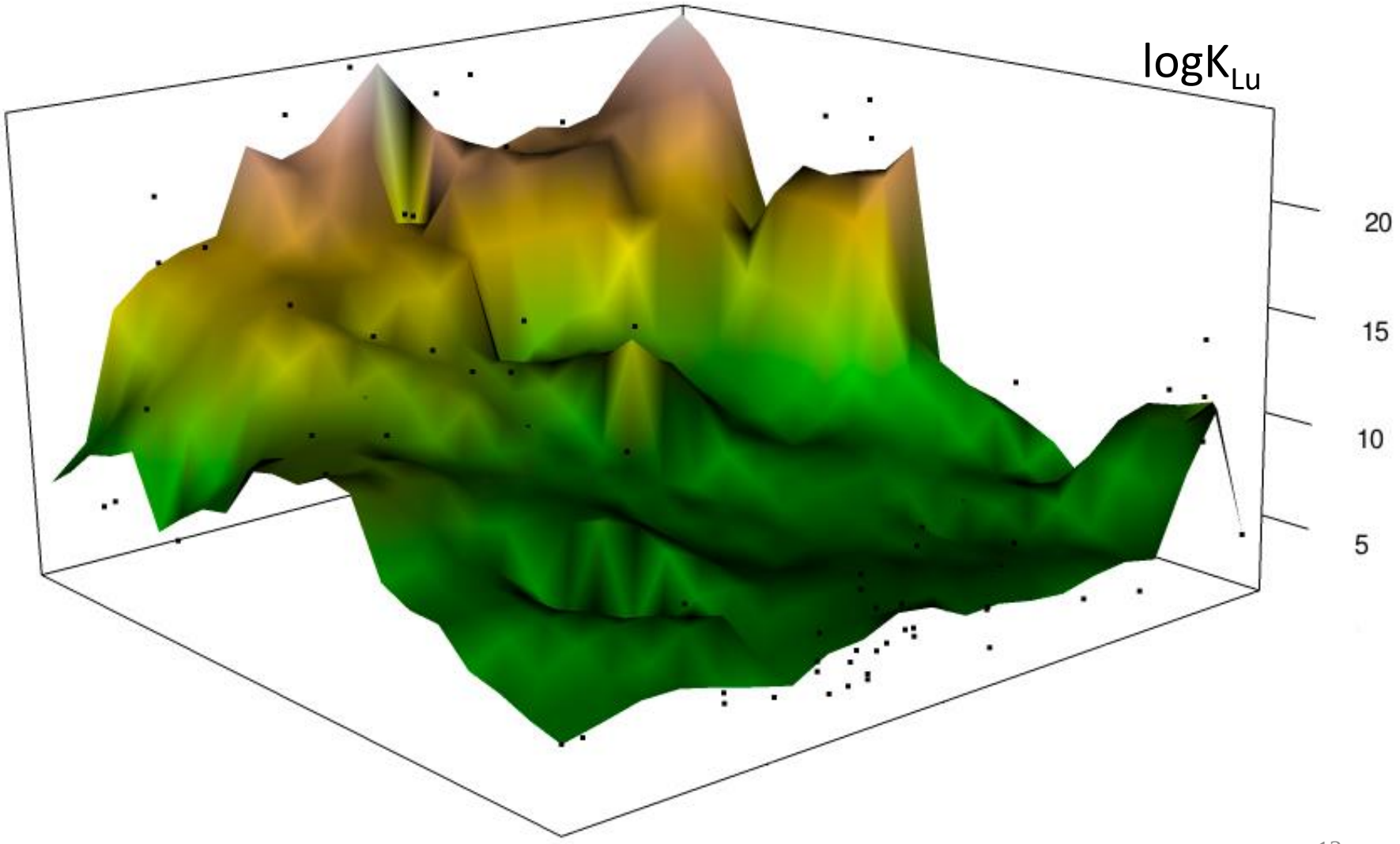


Activity landscape for Lu^{3+} binders

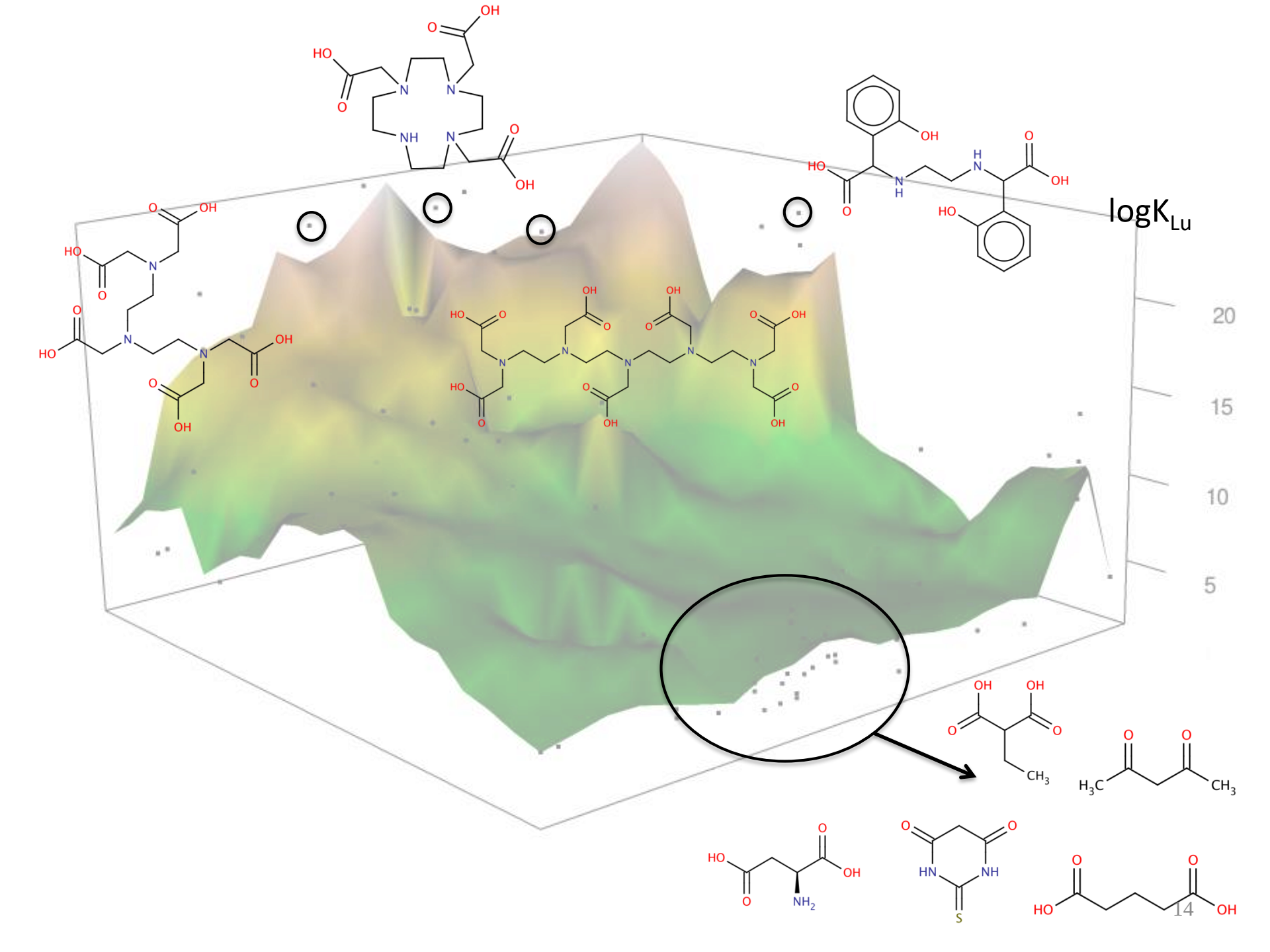


Activity landscape for Lu^{3+} binders

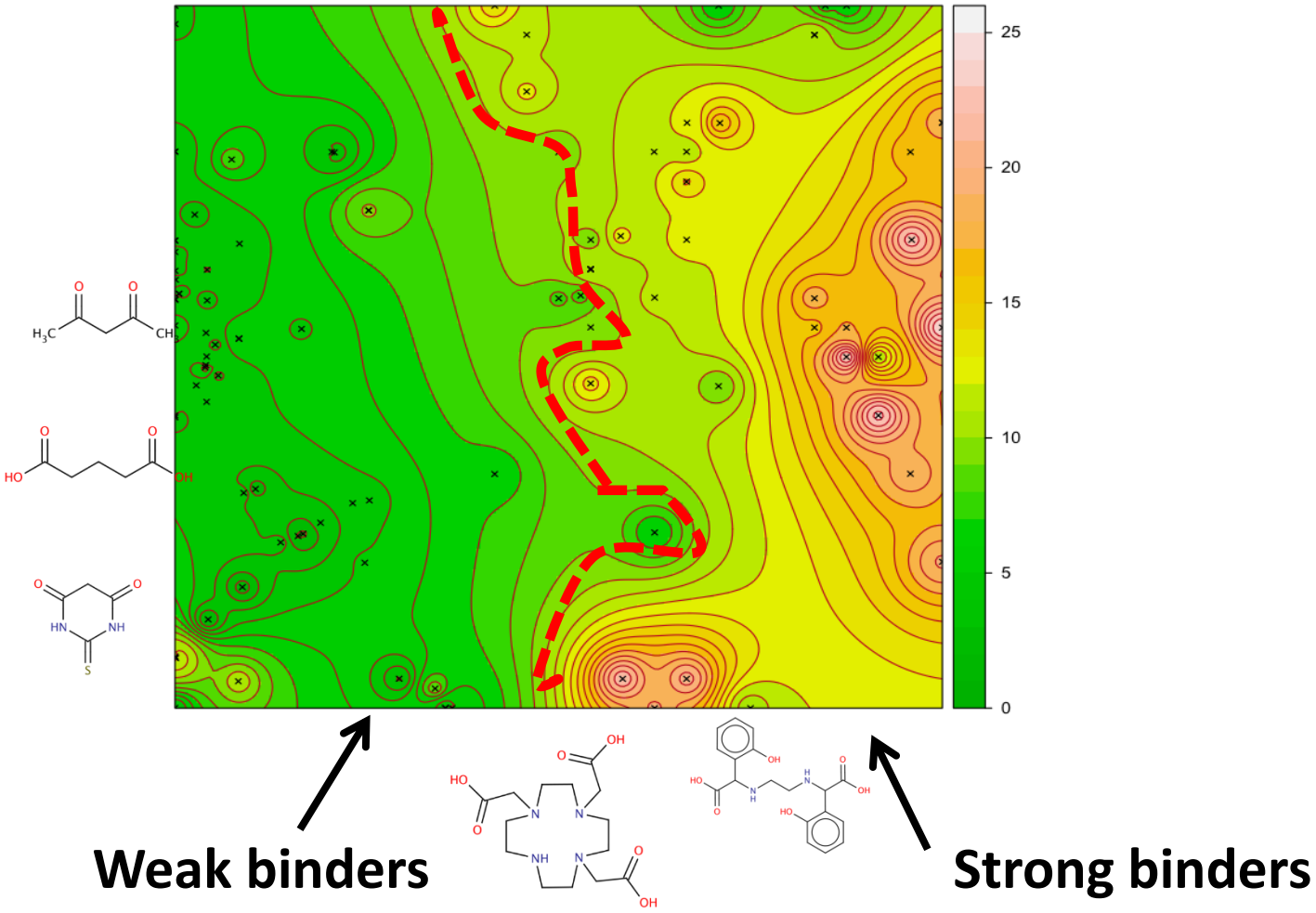
Stability of Lu^{3+} complexes with organic molecules



H. A. Gaspar . I. I. Baskin. G. Marcou. D. Horvath. A. Varnek *Mol. Informatics,* 2015, 34 (6-7), 348-356


logK$_{Lu}$
20
15
10
5

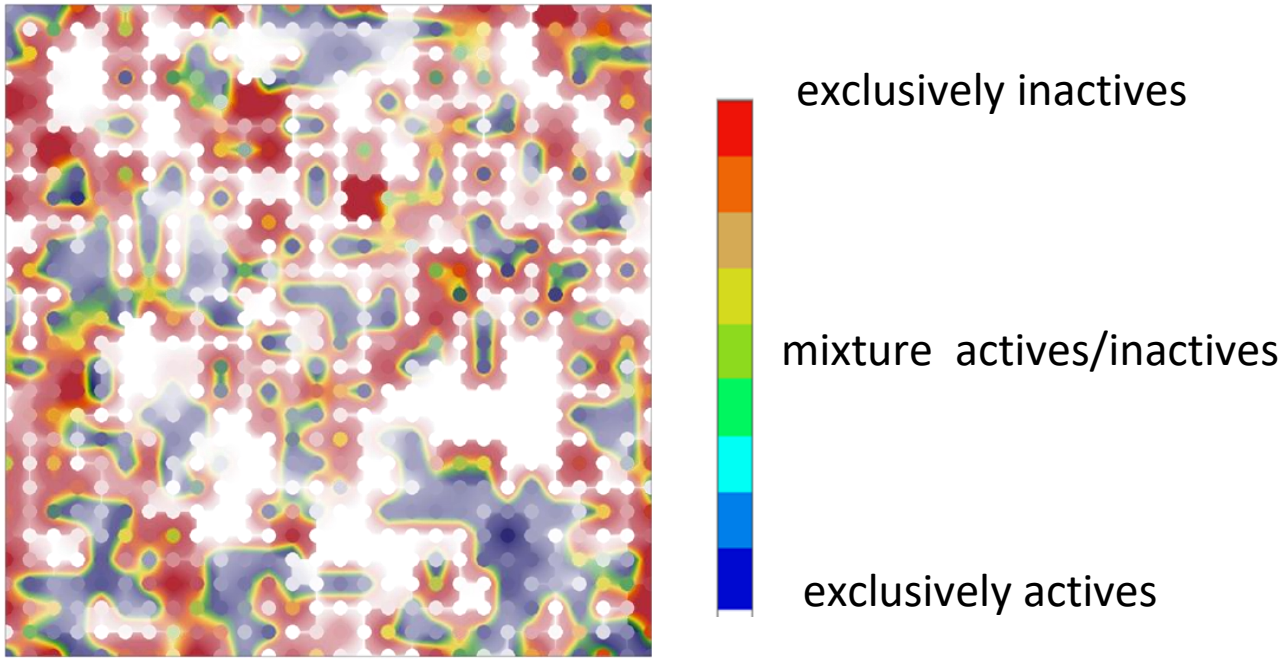
Activity landscape for Lu^{3+} complexation



H. Gaspar, I. Baskin, D. Horvath, G. Marcou, A. Varnek *Mol. Informatics,* 2015, 34 (6-7), 348-356

Class landscapes

ChEMBL (1.7 M cmds) : class landscape of antiviral compounds



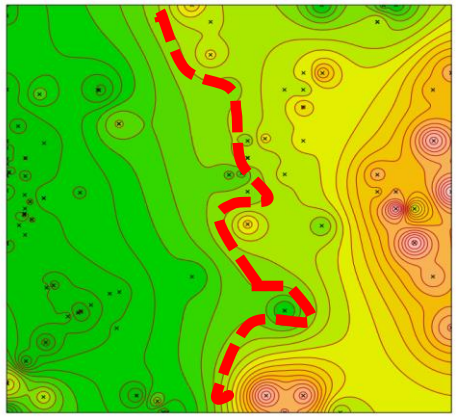
K. Klimenko, G. Marcou, D. Horvath, A. Varnek *J. Chem. Inf. Model,* 2016, 56, 1438−1454

GTM Landscape as predictive tool

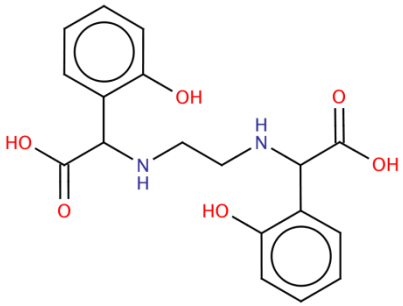
Activity landscape

Regression

$$\hat{A}_j(\text{test}) = \sum_k \bar{A}_k(\text{training}) R_{jk}(\text{test})$$

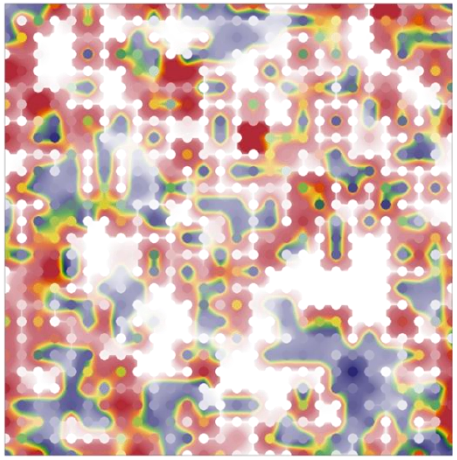


Predicted activity value



new compound

Projection



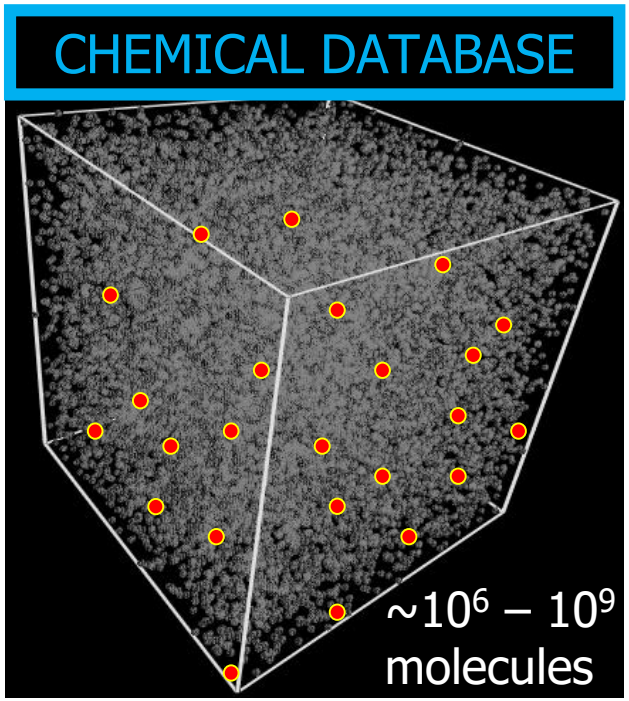
Predicted category "active" or "inactive"

Classification

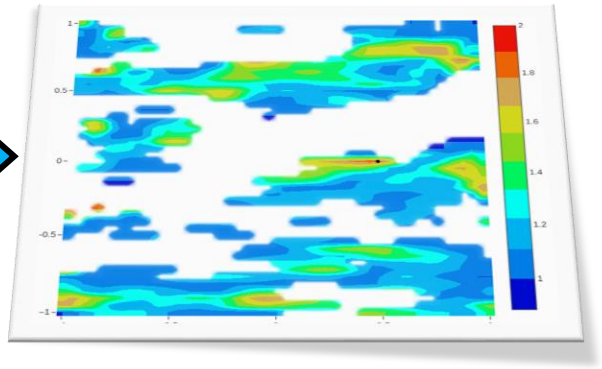
$$P(c_i|\mathbf{x}_k) = \frac{P(\mathbf{x}_k|c_i) \times P(c_i)}{\sum_i P(\mathbf{x}_k|c_i) \times P(c_i)}$$

Class landscape

Universal maps: application to virtual screening



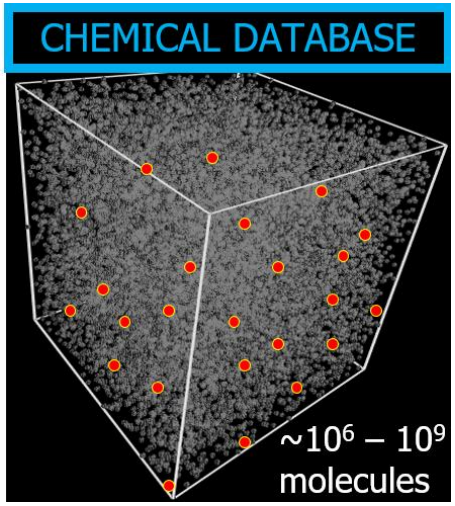
GTM activity or class landscape



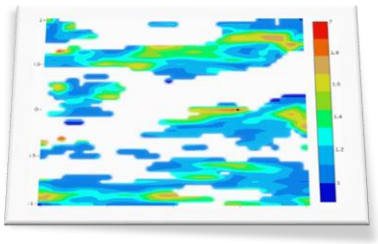
HITS

TRASH

Universal maps: application to virtual screening



GTM



In silico designed with GTM and experimentally validated compounds

- *Antiviral compounds*
- *Antimalarial compounds*
- *Solvents for Li-batteries*
- *Bromodomain (BRD4) inhibitors*

GTM : areas of application

Conformational space analysis

Data visualisation and analysis

Ligand to Protein docking

Libraries comparison

Sequences space analysis

Drugs repurposing

Structure-Activity modeling

Library design

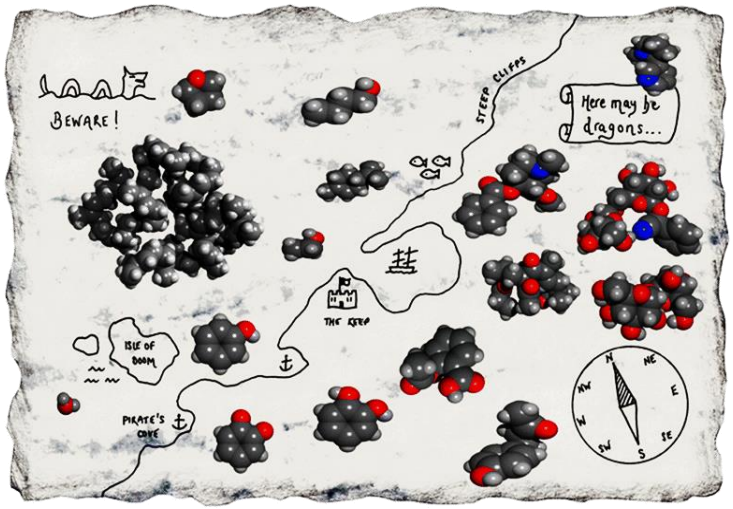
Virtual screening

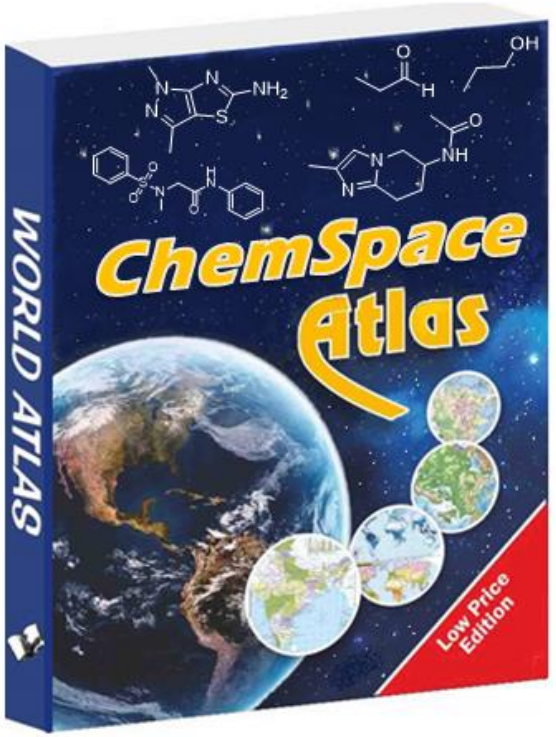
***de novo* design**

GTM : case studies

- **Visualisation and analysis of ultra-large data**
- **Artificial Intelligence driven design of novel molecular structures and reactions**

ChemSpace Atlas





Universal Map of chemical space

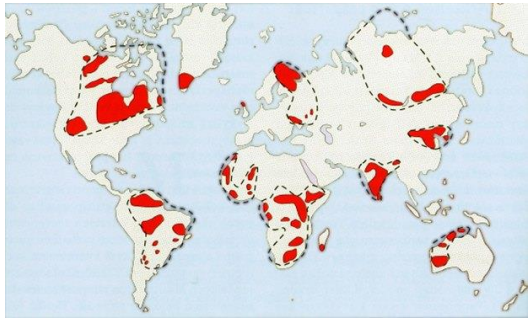
What do we expect from an “universal” map ?

Map of a chemical space is expected:

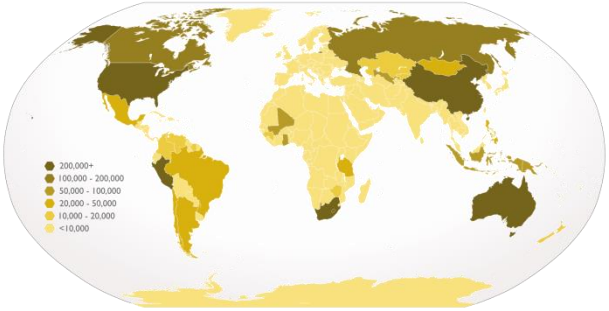
- to accommodate variety of known chemotypes;
- to be able to distinguish different activity classes;
- to separate actives and inactives within a given activity class;
- to be *neighbourhood behaviour (NB)* compliant, e.g., molecules grouped together are expected to display similar activities

«Universal » map

- Defines a frame of a relevant space
- Accommodates different landscapes



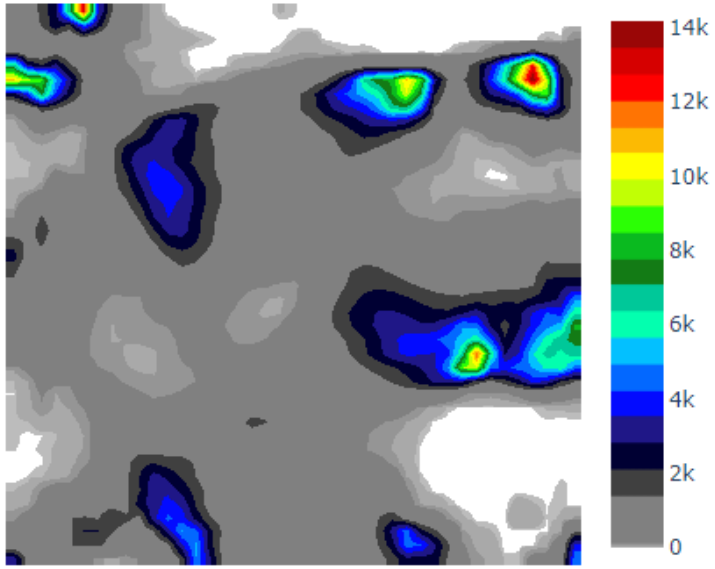
Highly prospective mineral regions



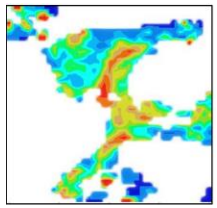
Gold production by country

«Universal » map

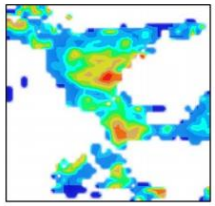
- Defines a frame of biological relevant chemical space
- ISIDA fragment descriptors are used
- Constructed on the basis of ChEMBL database compounds
- Predicts of > 700 biological activities



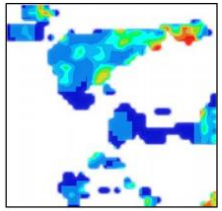
Density landscape of the ChEMBL database (1.7 M cmds)



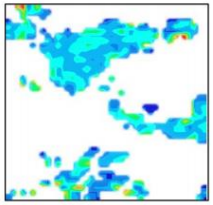
MAP kinase p38 alpha



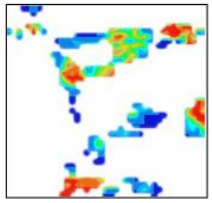
Vascular endothelial growth factor receptor 2



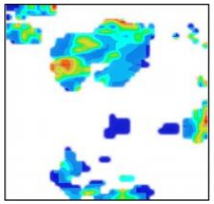
Cyclin-dependent kinase 2



Serine/threonine-protein kinase AKT



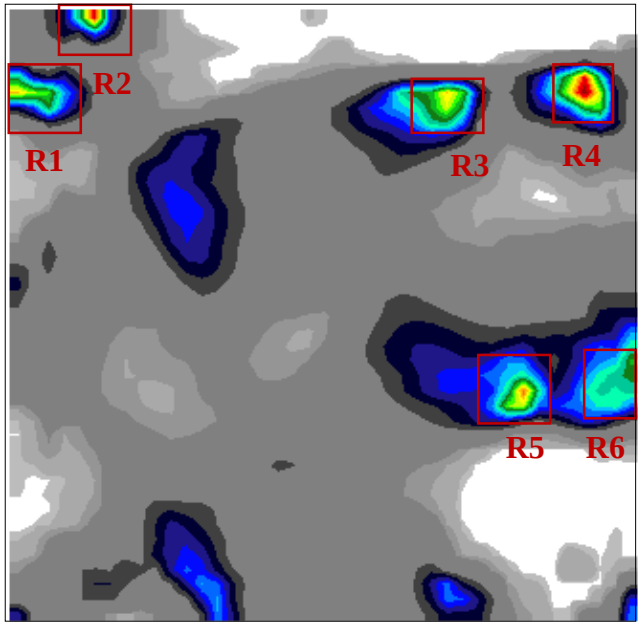
Phosphodiesterase 5A



Adenosine A2a receptor

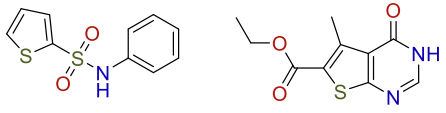
«Universal » map

chemotypes distribution

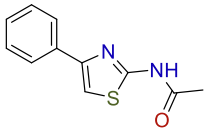


ChEMBL density landscape

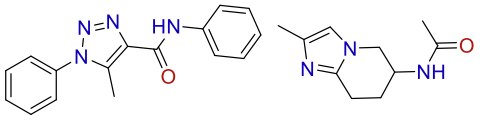
R1: Thiophenes

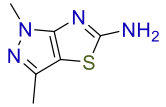


R2: Thiazoles

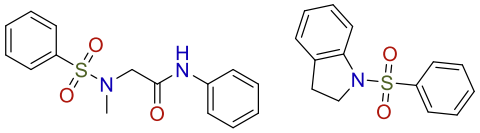


R3: Heterocyclic amides

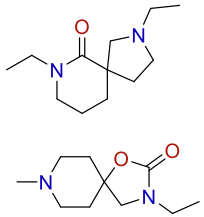




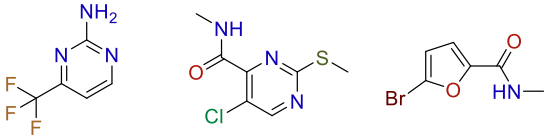
R4: Benzensulphonamides



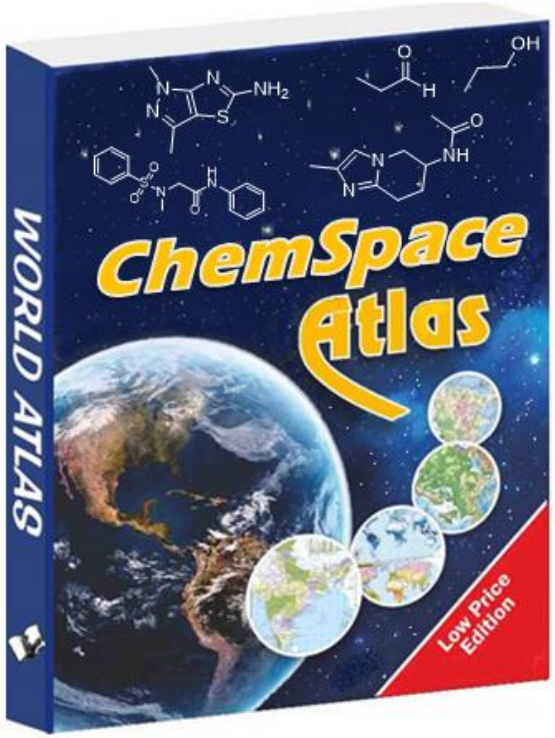
R5: Spiro-heterocyclic amides and carbamites



R6: Halloginated heterocycles



ChemSpace Atlas



Main features

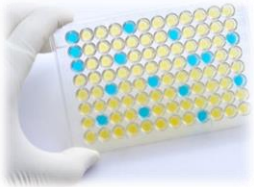
- polyvalent tool based on the GTM Universal Map
- accomodates > 3 billion cmpds
- assembles > 40000 hierarchically related maps of different scale and > 1.5 million activity landscapes

Main options

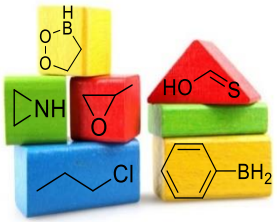
- Data visualization, search, subsets selection
- Automatized extraction of Maximal Common Structures
- Scaffold analysis
- Projection of new compounds
- Pharmacological profiling with respect to >700 biological targets

ChemSpace Atlas

The tool consists of 4 main parts:

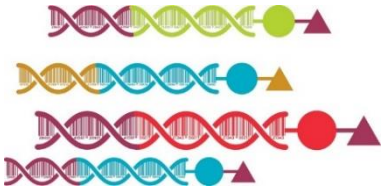


Screening Compounds



Building Blocks

Natural Products

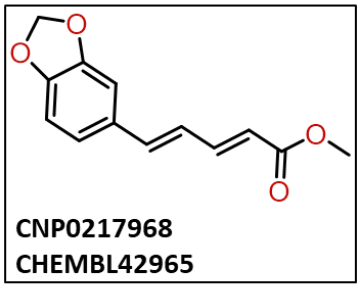


DNA-Encoded Libraries

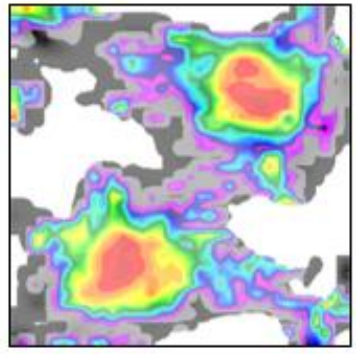
ChemSpace Atlas
discovery of synthetic analogs of natural products

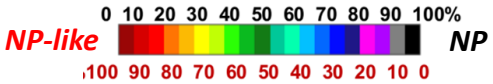
ChemAtlas NP database:
253 893 Natural Products + 586 235 synthetic analogs from ZINC

Query



Active against **Monoamine oxidase B**
(Rattus norvegicus)





Structures and activity profiles of synthetic analogues

NATURAL PRODUCTS NAVIGATOR

Dashboard

START HERE

Welcome

ANALYSIS

Chemspace tracker

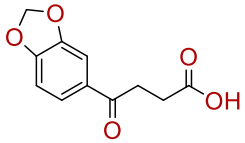
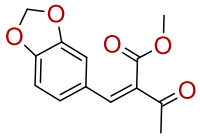
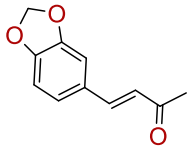
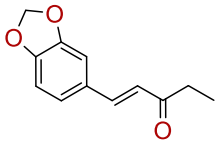
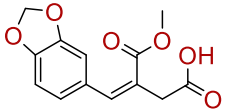
Activity prediction

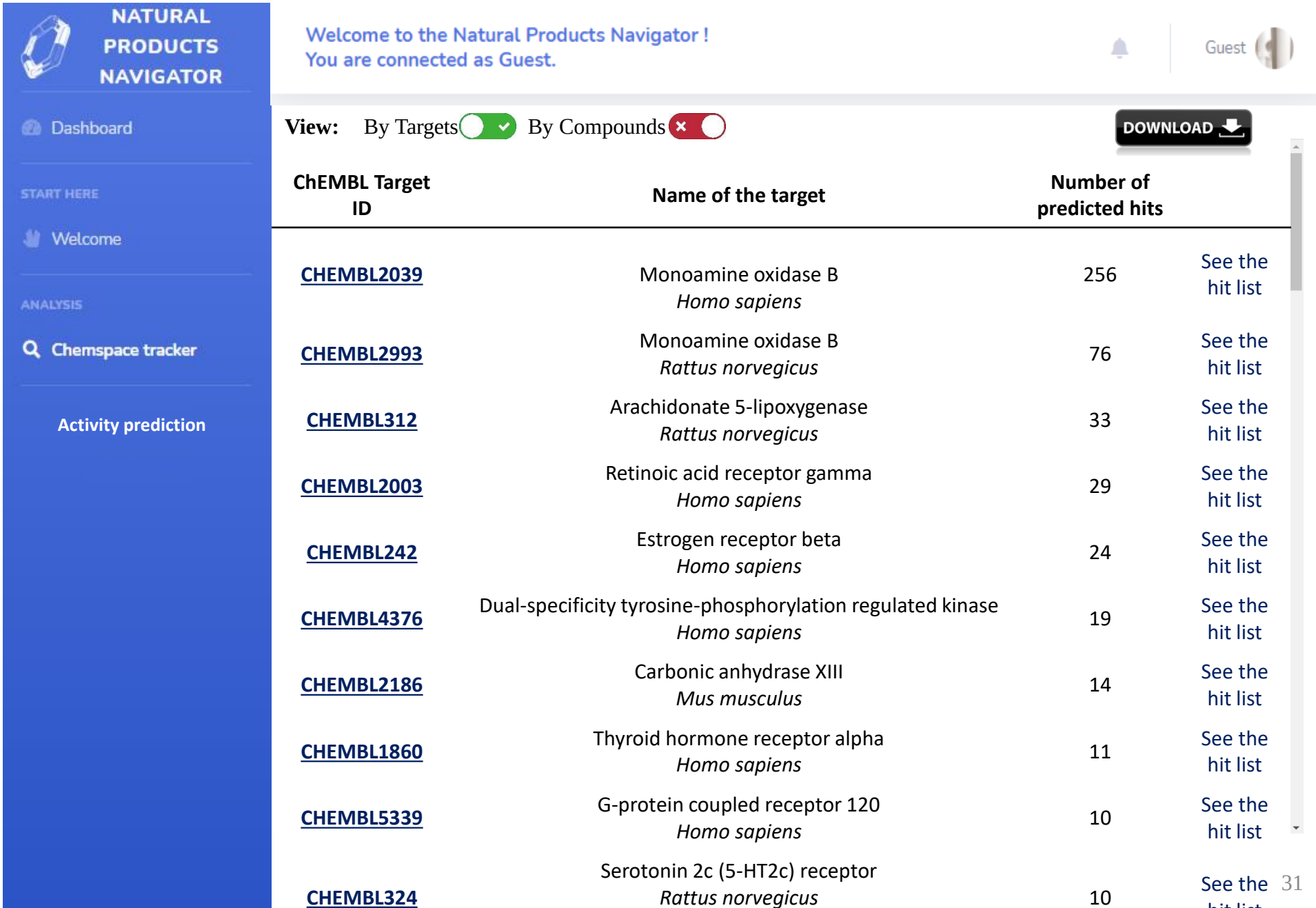
Welcome to the Natural Products Navigator !
You are connected as Guest.

Guest

View: By Targets By Compounds

DOWNLOAD

Compounds	Predicted activity (Target ID)
 ZINC000000040327	**CHEMBL2039 : Monoamine oxidase B** *(Homo sapiens)*
 ZINC000016138715	**CHEMBL2993: Monoamine oxidase B** *(Rattus norvegicus)*
 ZINC000001754404 ZINC000012417143	**CHEMBL2039 : Monoamine oxidase B** *(Homo sapiens)*
 ZINC000002015852 ZINC000020232188	**CHEMBL2039 : Monoamine oxidase B** *(Homo sapiens)* **CHEMBL4376 : Dual-specificity tyrosine-phosphorylation regulated kinase 2** *(Homo sapiens)*
 ZINC000005218035	**CHEMBL2993: Monoamine oxidase B** *(Rattus norvegicus)*



ChEMBL Target ID	Name of the target	Number of predicted hits	
CHEMBL2039	Monoamine oxidase B *Homo sapiens*	256	See the hit list
CHEMBL2993	Monoamine oxidase B *Rattus norvegicus*	76	See the hit list
CHEMBL312	Arachidonate 5-lipoxygenase *Rattus norvegicus*	33	See the hit list
CHEMBL2003	Retinoic acid receptor gamma *Homo sapiens*	29	See the hit list
CHEMBL242	Estrogen receptor beta *Homo sapiens*	24	See the hit list
CHEMBL4376	Dual-specificity tyrosine-phosphorylation regulated kinase *Homo sapiens*	19	See the hit list
CHEMBL2186	Carbonic anhydrase XIII *Mus musculus*	14	See the hit list
CHEMBL1860	Thyroid hormone receptor alpha *Homo sapiens*	11	See the hit list
CHEMBL5339	G-protein coupled receptor 120 *Homo sapiens*	10	See the hit list
CHEMBL324	Serotonin 2c (5-HT2c) receptor *Rattus norvegicus*	10	See the

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim

Chemography: Searching for Hidden Treasures

Yuliana Zabolotna, Arkadii Lin, Dragos Horvath, Gilles Marcou, Dmitriy M. Volochnyuk, and Alexandre Varnek*

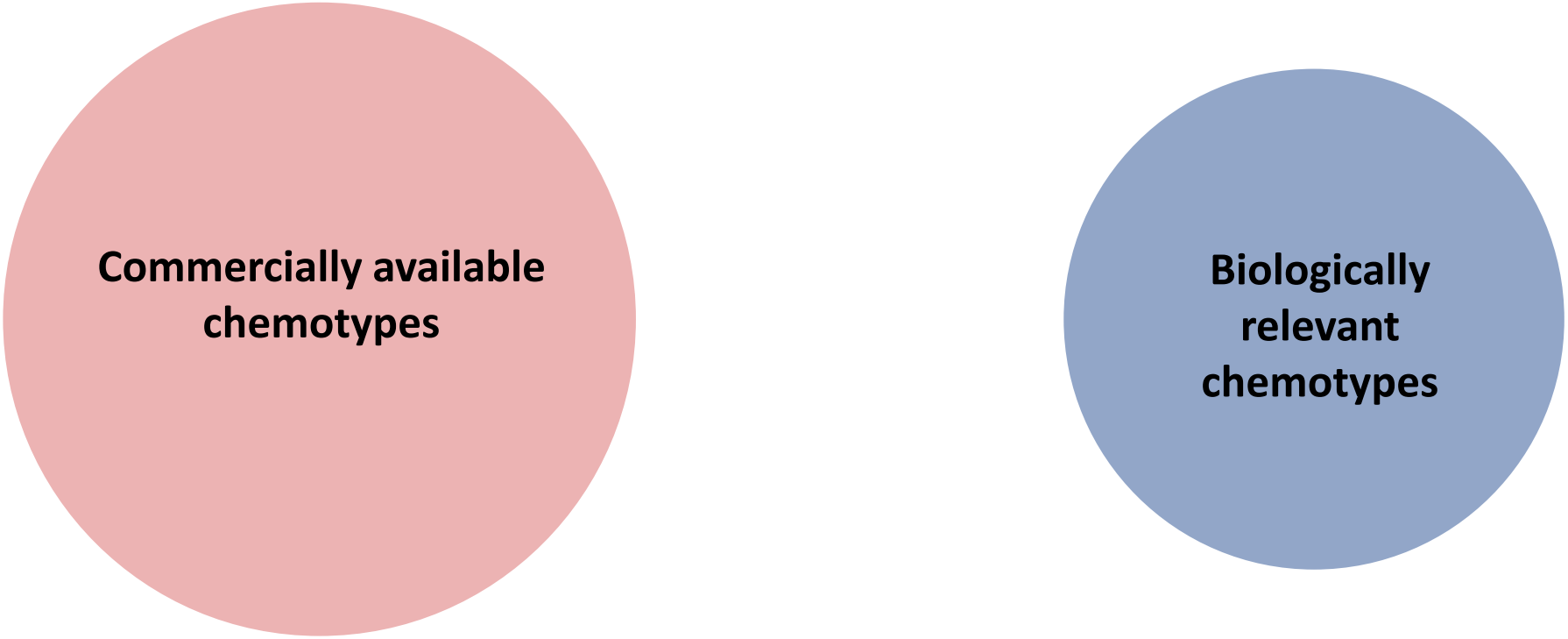
J. Chem. Inf. Model. 2021, 61, 1, 179–188

Initial gold-bearing ore

Gold-enriched ore

Pure gold

Commercial vs Biologically relevant data



ZINC DATABASE

>1.3 billion cmpds

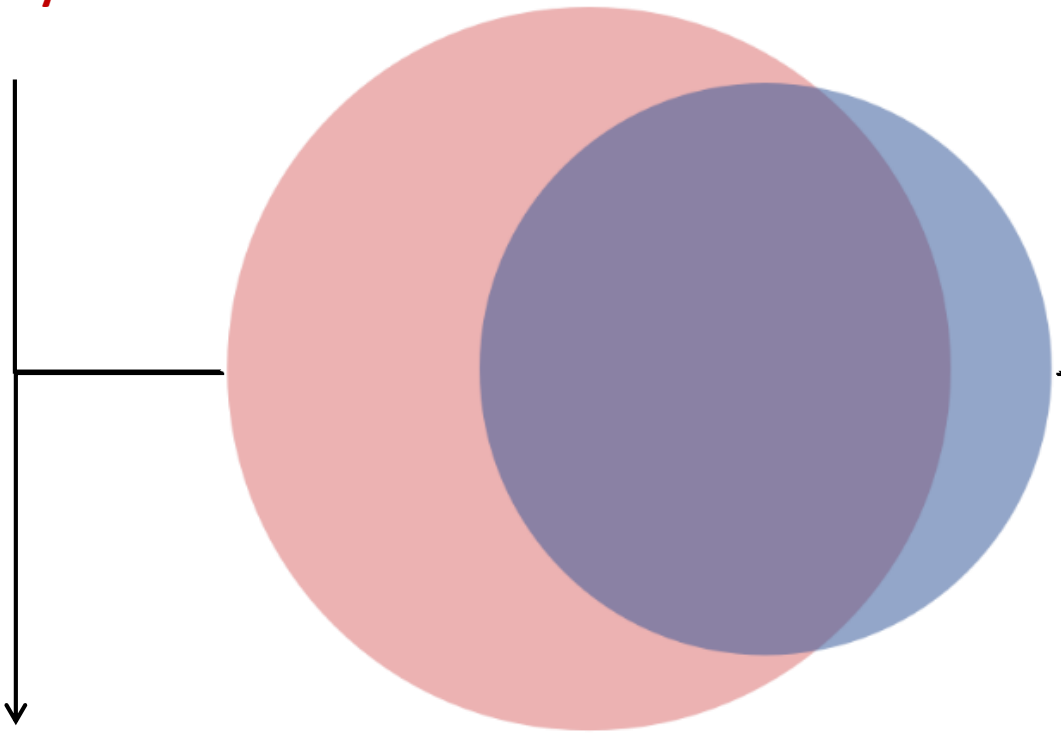
ChEMBL

>1.8 M cmpds

Commercial vs Biologically relevant data

Chemotypes never been biologically tested

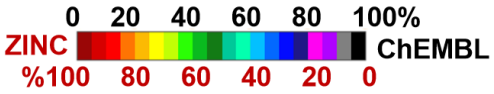
Chemotypes missing in the commercial chemical space

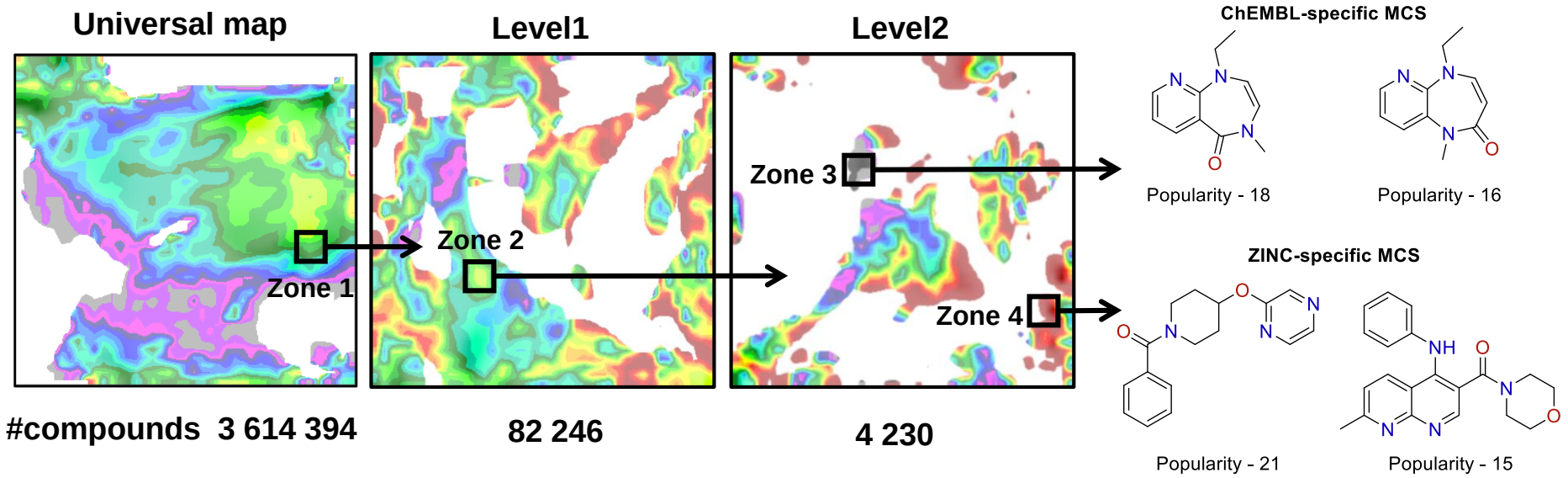


Enhancement of screening libraries

Biologicaly biased commercial libraries enhencement

Hierarchical GTM navigation of the chemical space



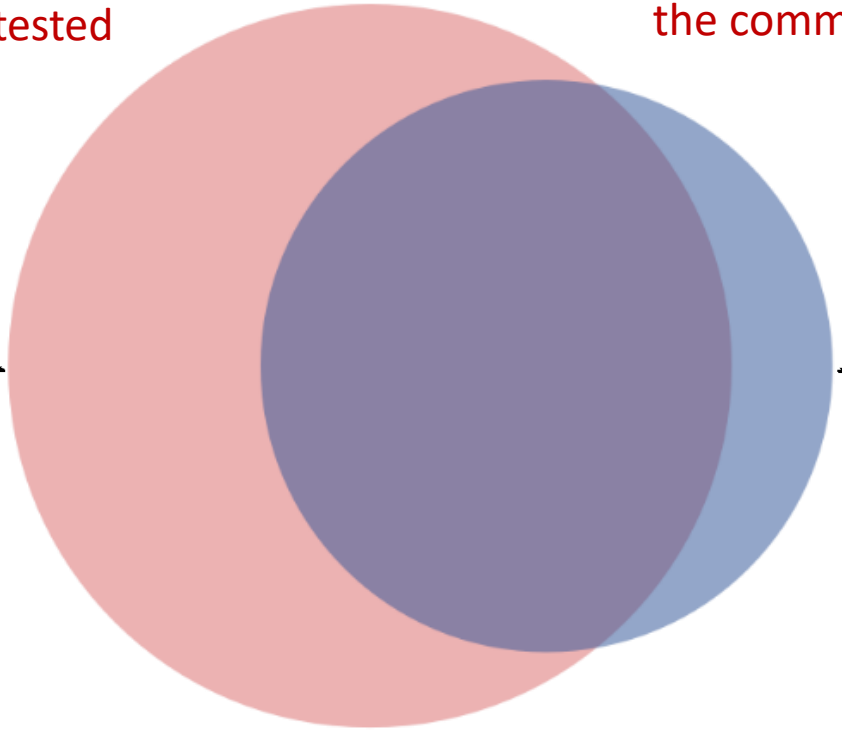


Yu. Zabolotna *et al., J. Chem. Inf. Model.* 2021, 61, 1, 179–188

Commercial vs Biologically relevant data

> 100K chemotypes never been biologically tested

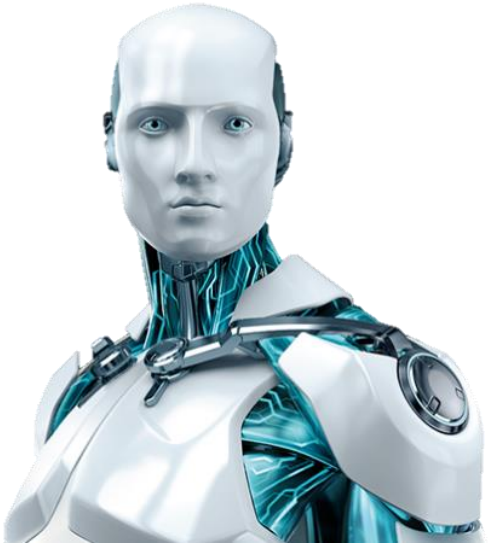
>20 K chemotypes missing in the commercial chemical space



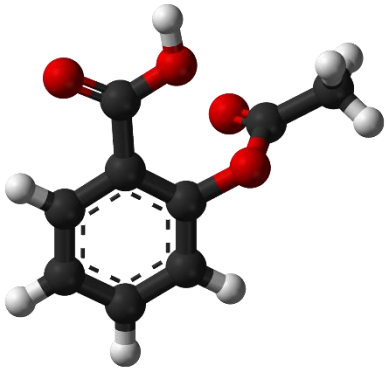
Enhancement of screening libraries

Biologicaly biased commercial libraries enhencement

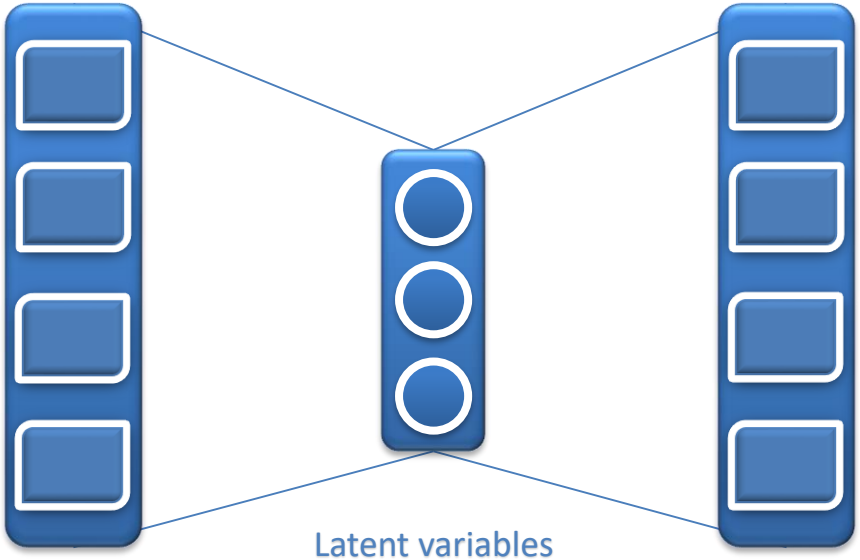
De novo design of molecules possessing desirable biological activity

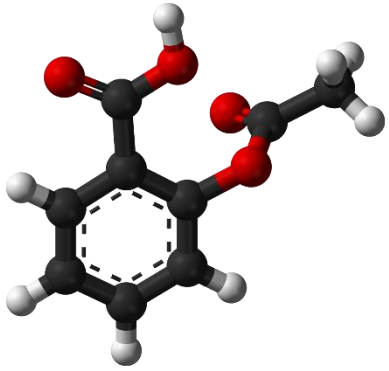


Autoencoder performing SMILES reconstruction



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





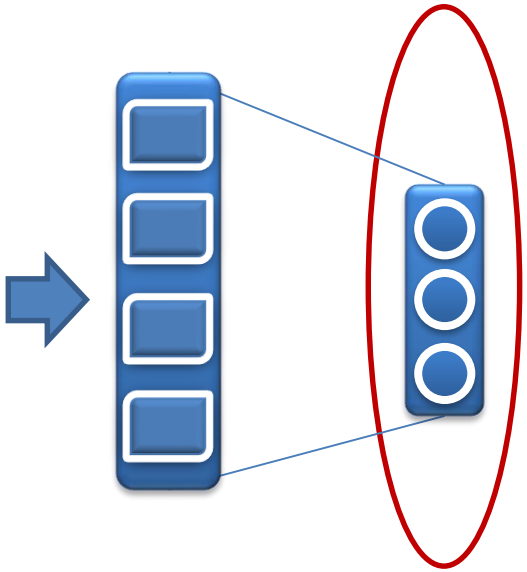
O=C(C)Oc1ccccc1C(=O)O

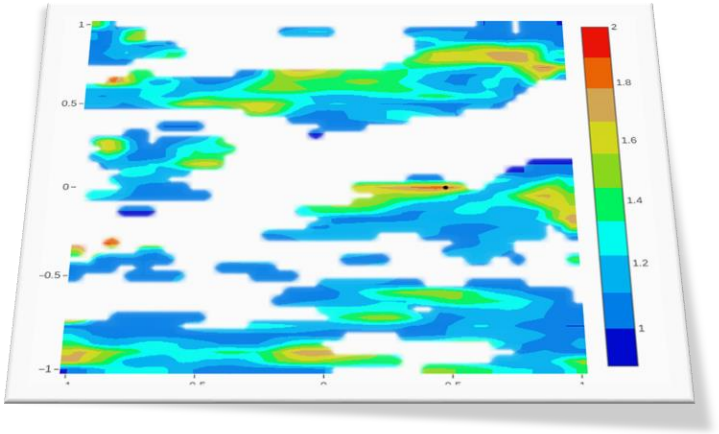
Chemical structure → **Real numbers encoding** → **Chemical structure**

Building GTM on latent variables of autoencoder

Latent variables
(*vector on real numbers)*

Chemical Database (SMILES)

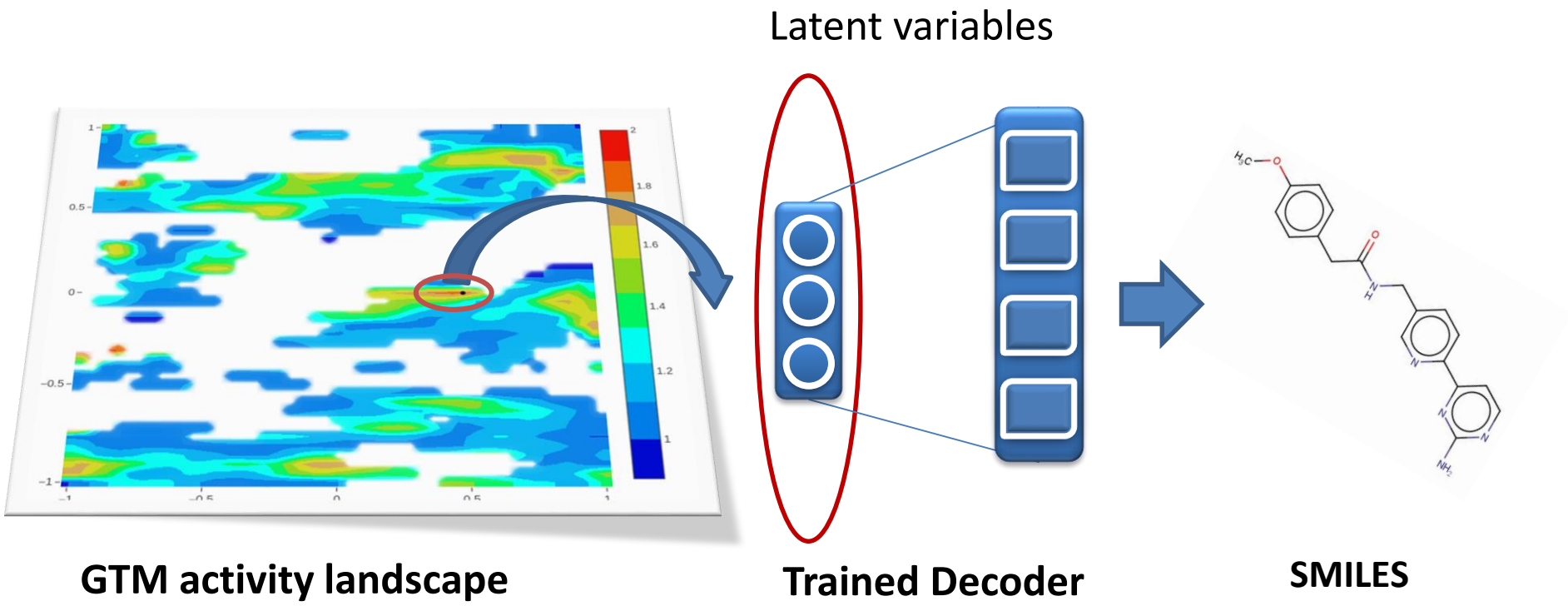




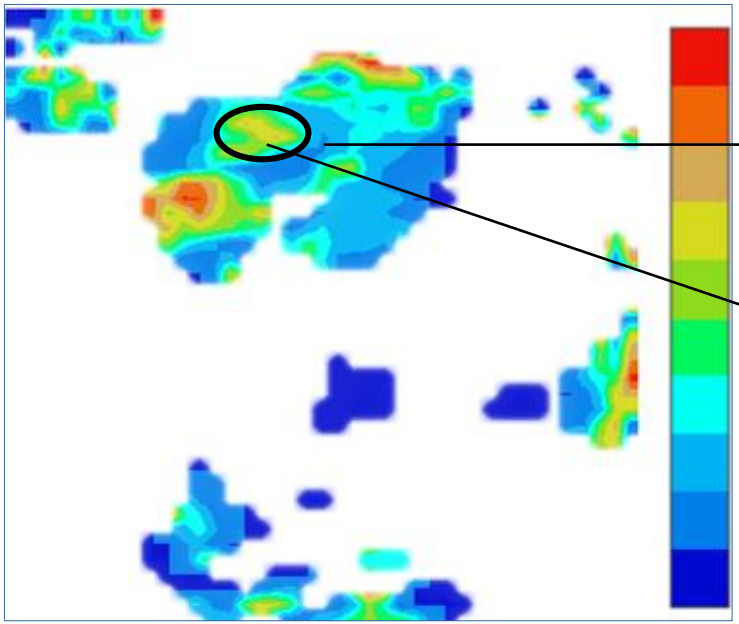
Trained Encoder

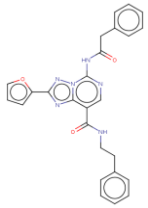
GTM

Generation of novel structures from specific areas of the map

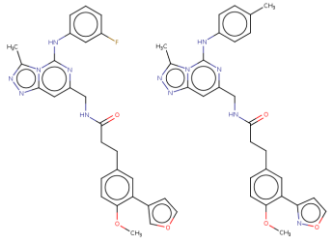


Case study: Generation of inhibitors of A2a receptor





Existing structure

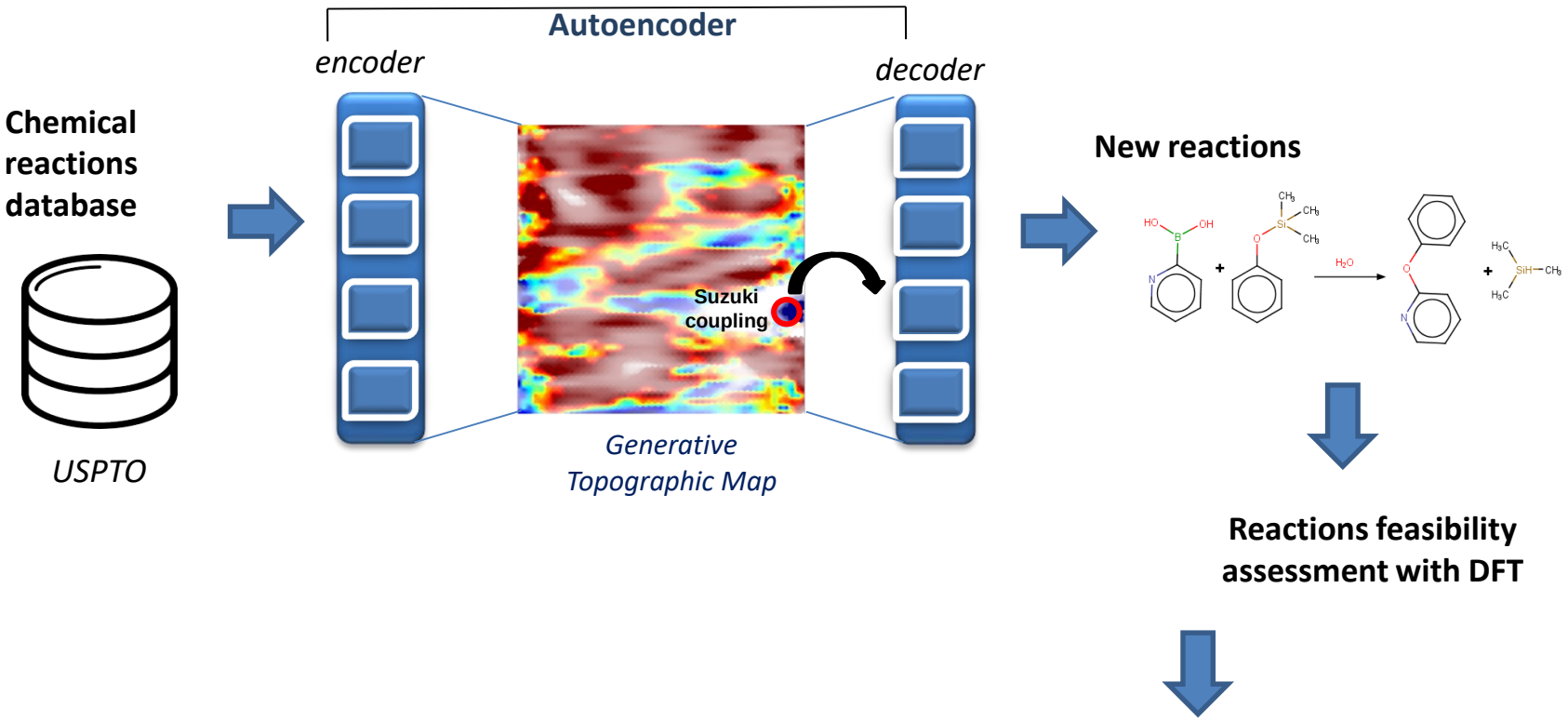


***de novo* structures**

- ***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 types of chemical reactions

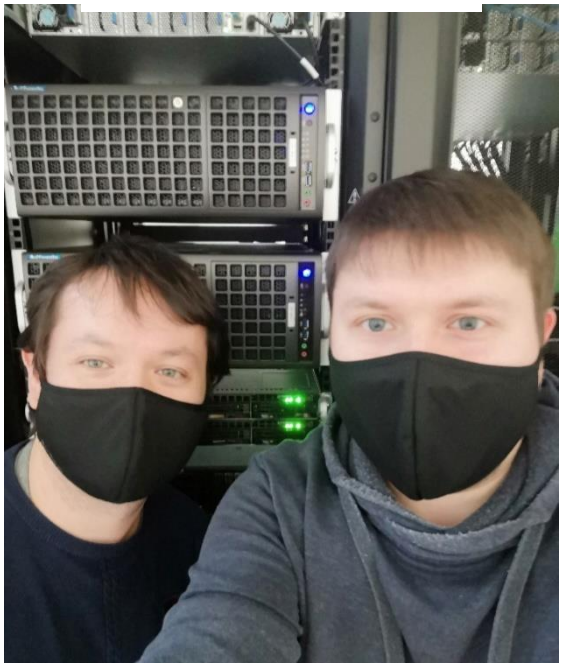


- **13 new (with respect to the training data) Suzuki-like reactions have been detected**
- **5 of them have been found in recent publications**

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

Chém Informatique
STRASBOURG

ICReDD
Institute for Chemical Reaction Design and Discovery
HOKKAIDO UNIVERSITY



Collaboration

- ITN Marie-Curie BigChem
- Federal University of Kazan
- Chumakov Research Center RAS

- Enamine
- eMolecules
- Janssen Pharmaceutical
- TOTAL
- SOLVAY