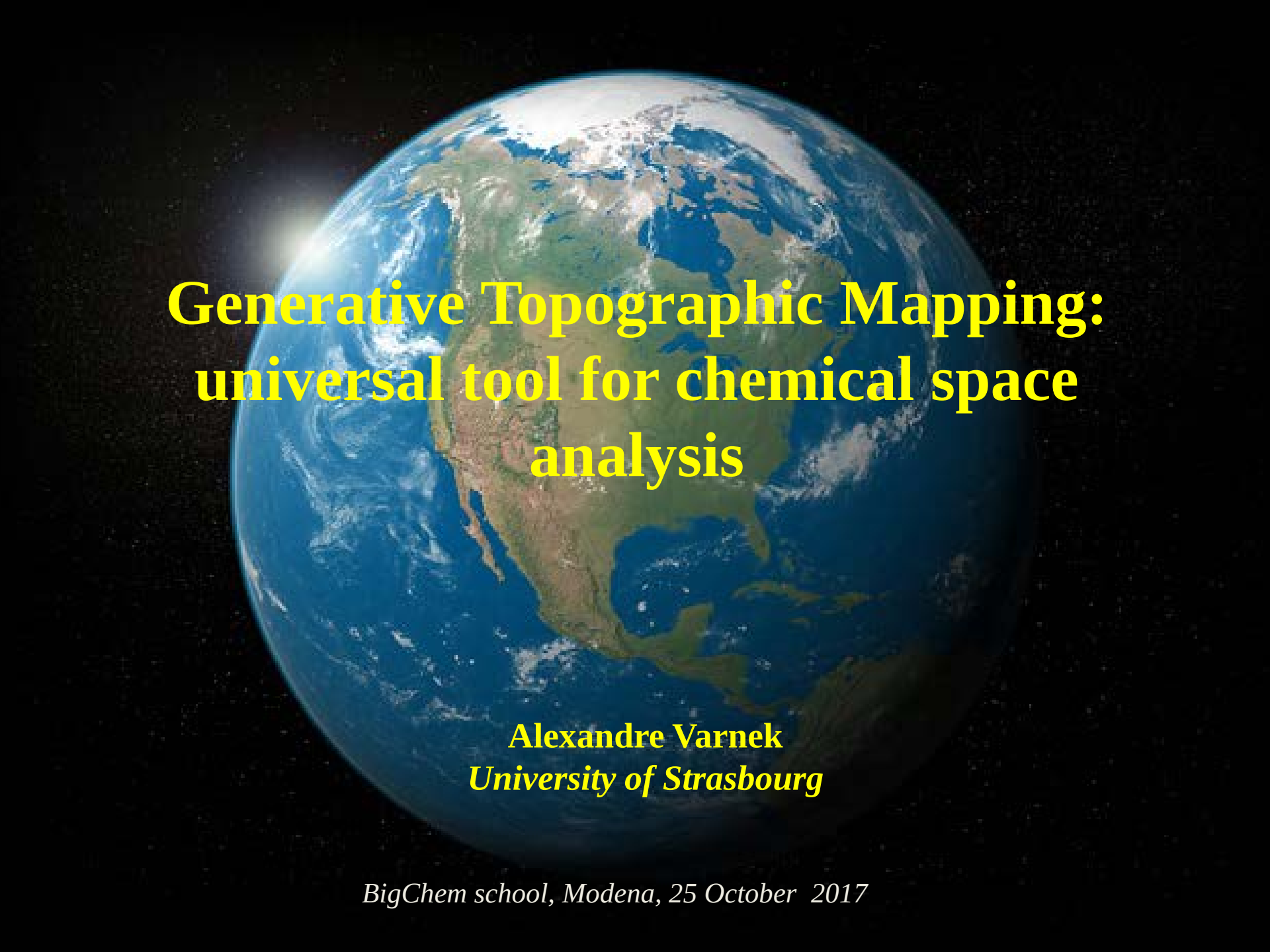


Generative Topographic Mapping: universal tool for chemical space analysis

Alexandre Varnek
University of Strasbourg

BigChem school, Modena, 25 October 2017

Chemical universe: Big Data problem

- 10^8 compounds are currently available

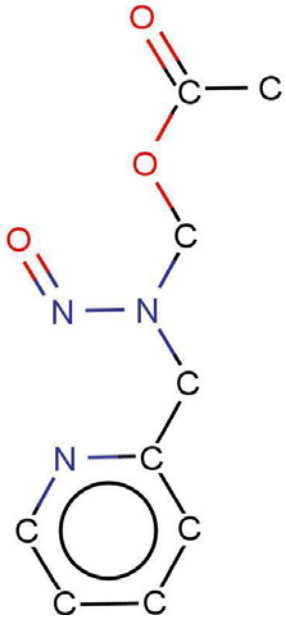
- 10^{33} drug-like molecules could be synthesized *

How to represent this huge chemical space and to navigate in this space ?

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

.....

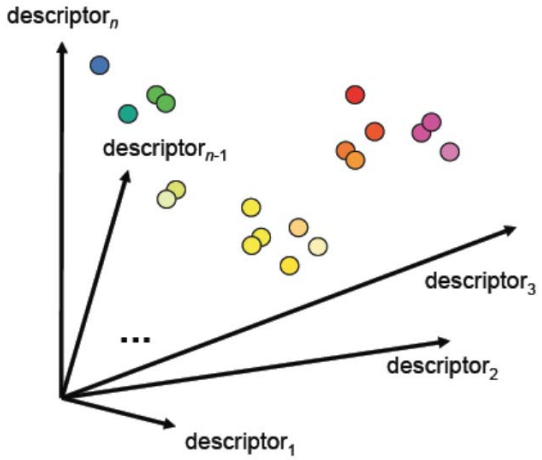
Descriptor vector

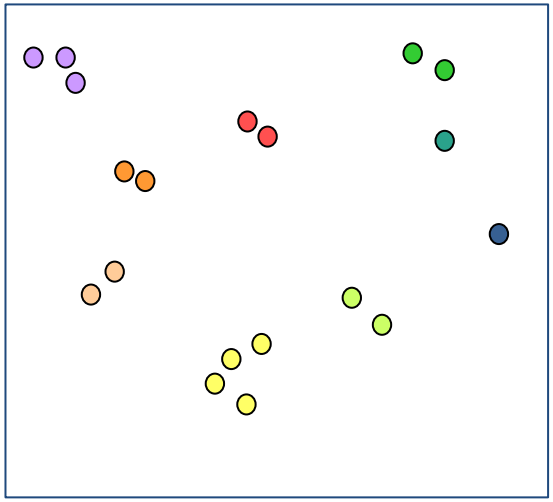
Descriptor	Value
D_1	a_1
D_2	a_2
....	...
D_i	a_i
....	...

> 5000 types of descriptors are used

Chemography: cartography of chemical space

Data visualization => dimensionality reduction problem

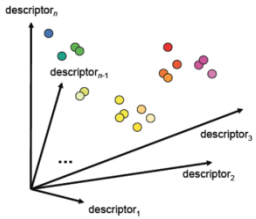


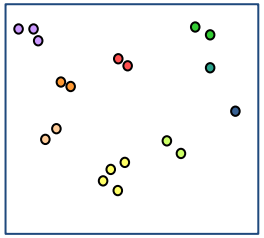


Data space
(N-dimensional)

Latent space
(2-dimensional)

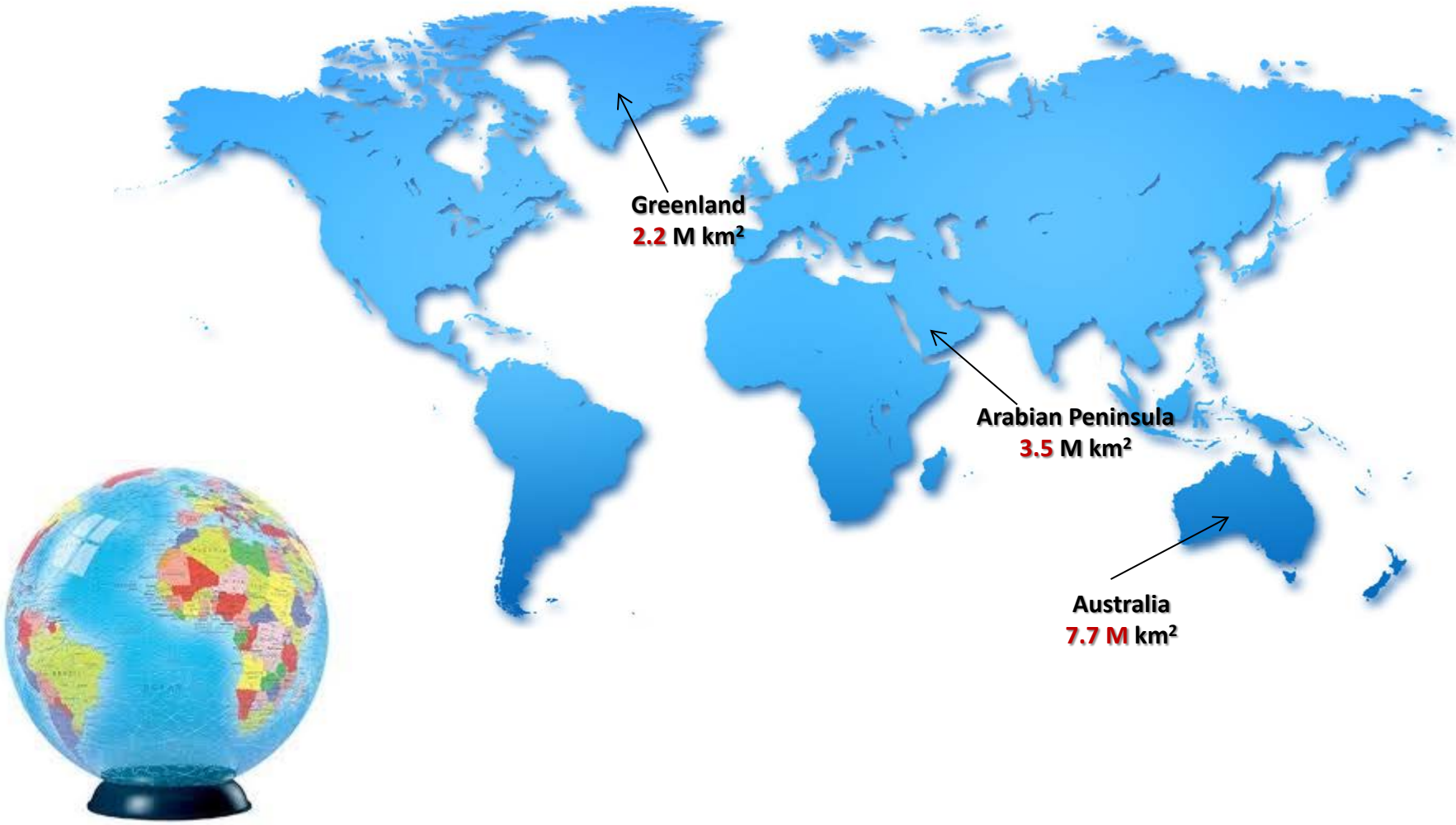
Dimensionality reduction requirements



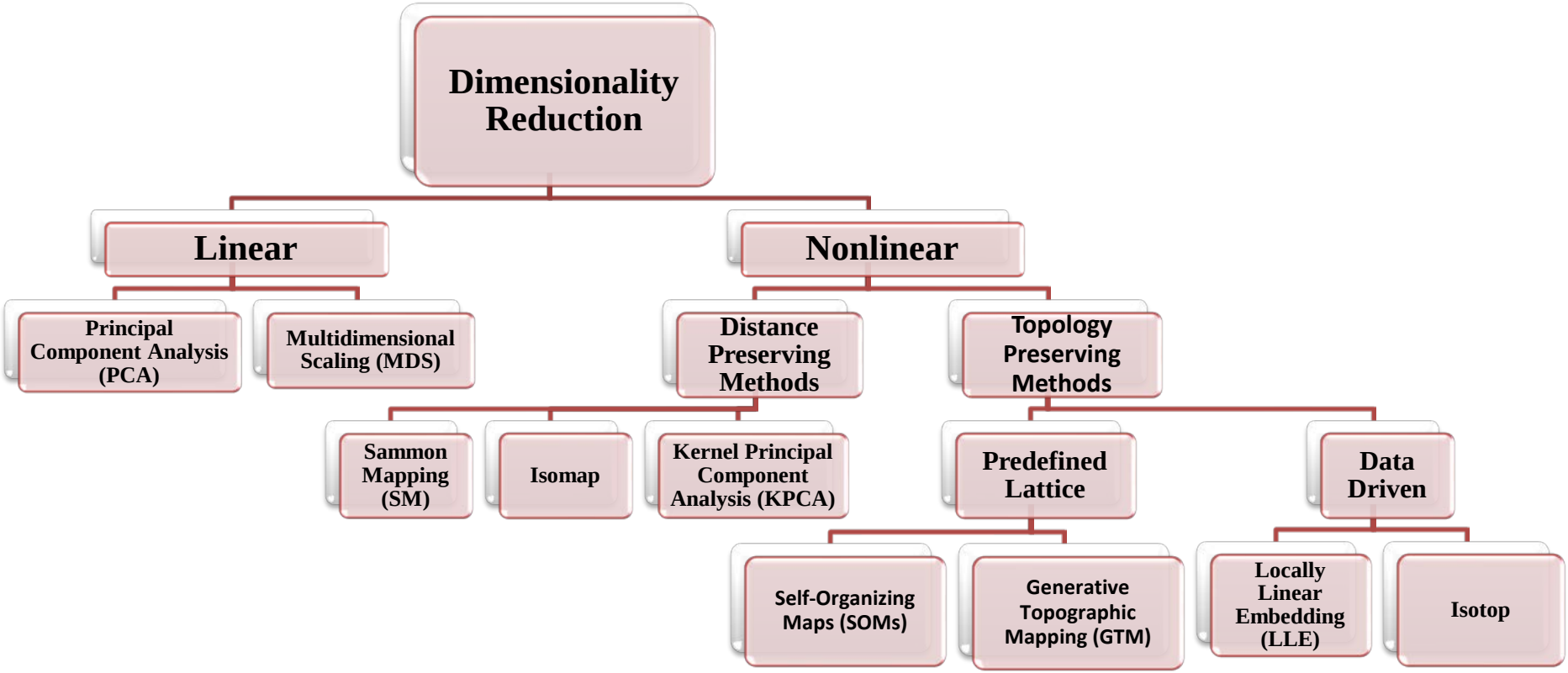


- **minimal information loss,**
- **topology preservation,**
- **distance preservation**

Dimensionality Reduction: information loss

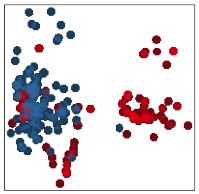


Taxonomy of Dimensionality Reduction Techniques

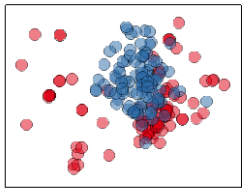


Dimensionality reduction methods

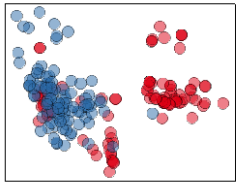
Acetylcholinesterase dataset (DUD) : 100 actives and 100 inactives



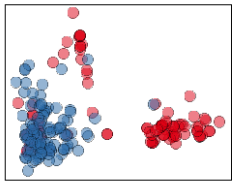
Multi-Dimensional Scaling



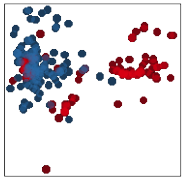
Canonical Correlation Analysis



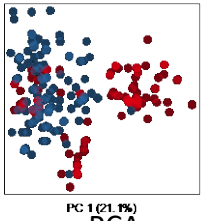
Independent Component Analysis



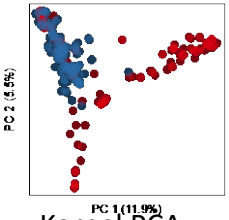
Exploratory Factor Analysis



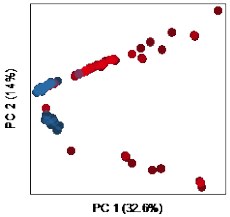
Sammon map



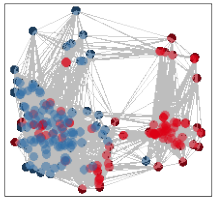
PCA



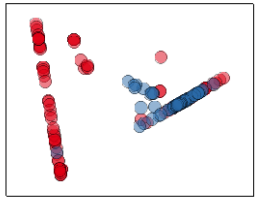
Kernel PCA (RBF kernel)



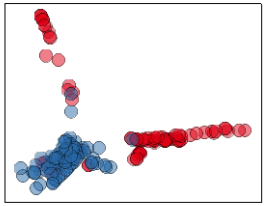
Kernel PCA (polynomial kernel)



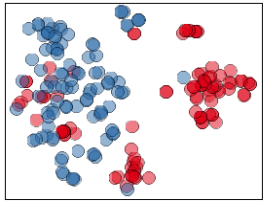
Isomap



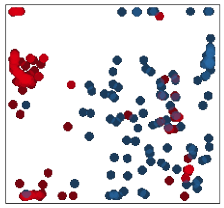
Locally Linear Embedding



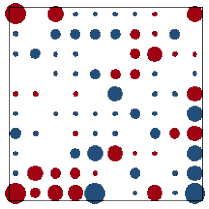
Laplacian Eigenmaps



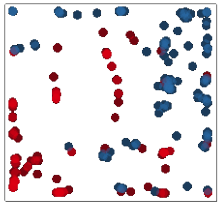
t-SNE



Autoencoder dimensionality reduction



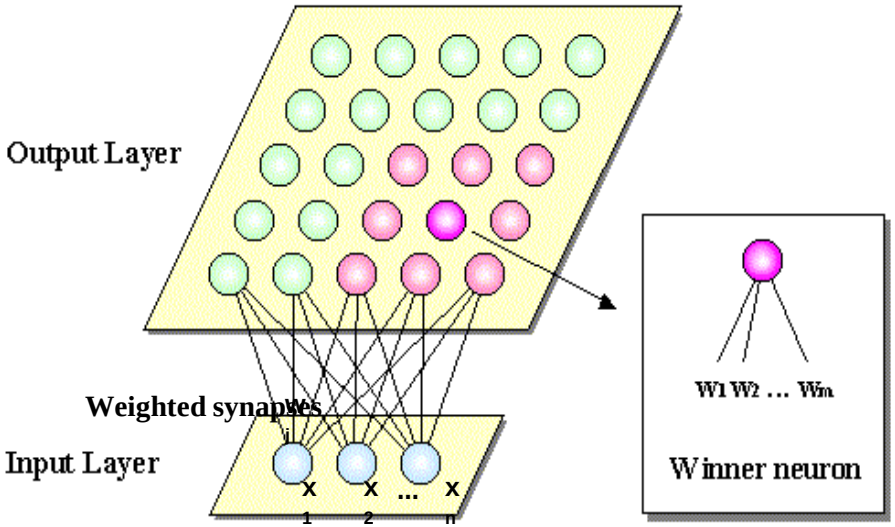
SOM



GTM

Generative Topographic Mapping approach

Teuvo Kohonen



- Nonlinear unsupervised approach
- Simple interpretation
- Topology preservation
- Can be used for classification purposes

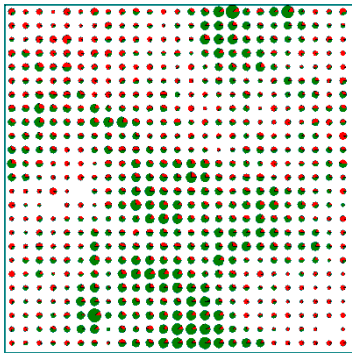
Limitations of SOMs

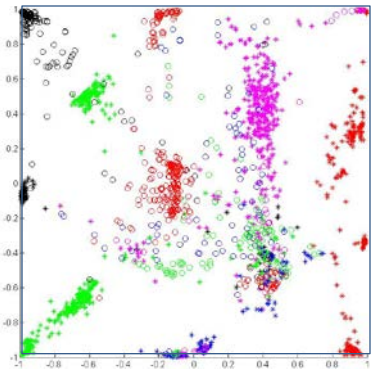
- the absence of a cost function to be optimized in training;
- the lack of a theoretical basis for choosing methods parameters;
- the absence of any general proofs of convergence;
- models do not define a probability density

C.M.Bishop, M.Svensen, C.K.I.Williams, « The Generative Topographic Mapping», *Neural Computation*, 10, No. 1, 215-234 (1998)

Self-Organizing Maps (SOM)

Generative Topographic Mapping (GTM)



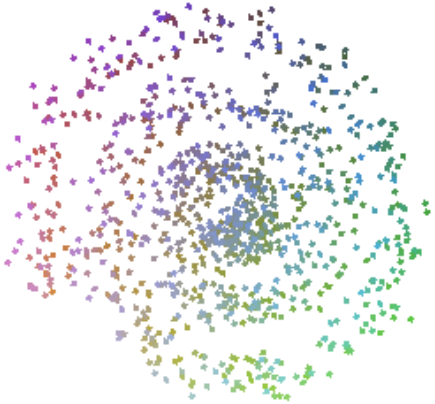


Teuvo Kohonen

Christopher Bishop

GTM overcomes most of limitations of SOMs without introducing disadvantages

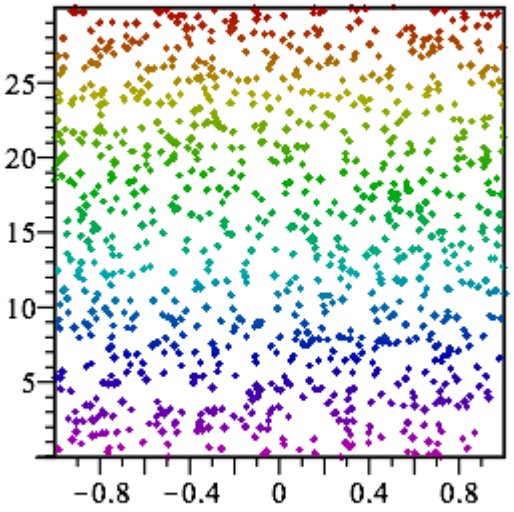
Generative Topographic Mapping : algorithm



Insertion a manifold

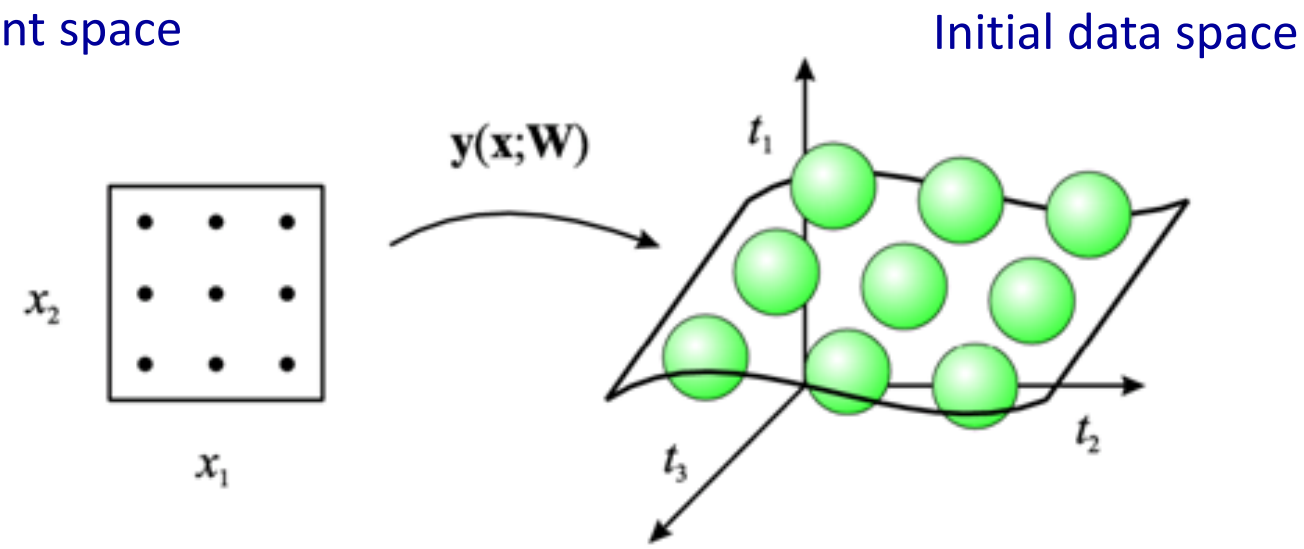
Projection of data points onto the manifold

Unbending manifold



2D latent space

Generative Topographic Mapping (GTM)



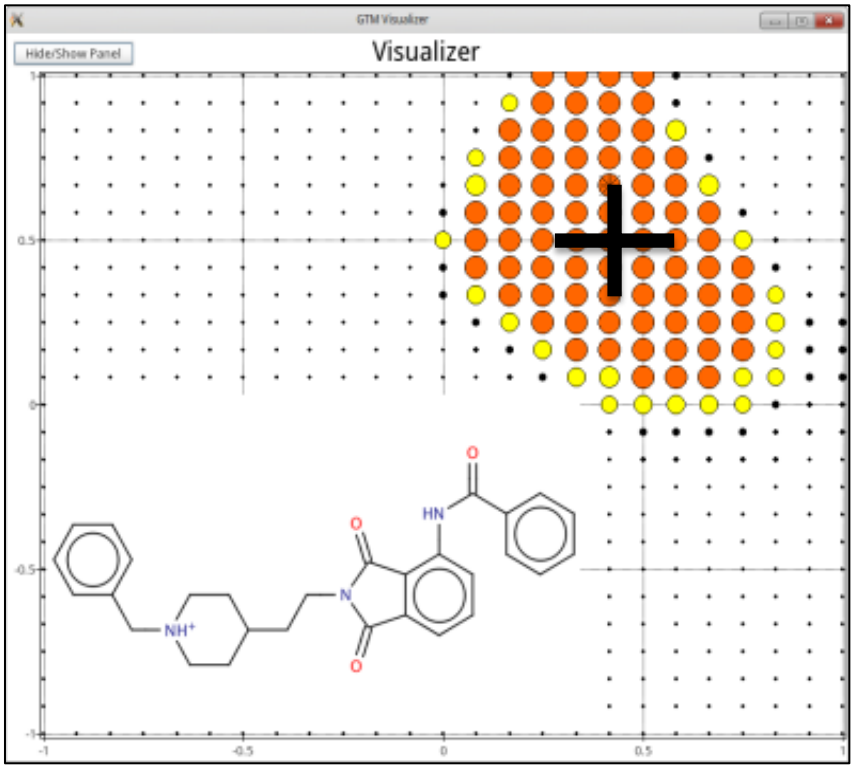
GTM generates a data probability distribution in ***both initial and latent data spaces*.**

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

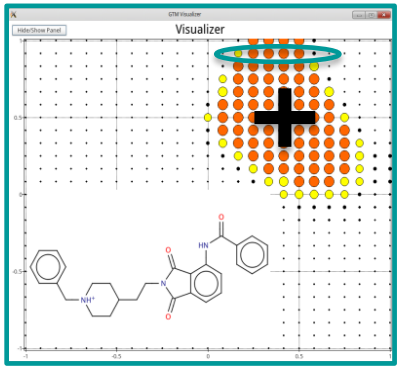
GTM: Probability Density distribution in the latent space

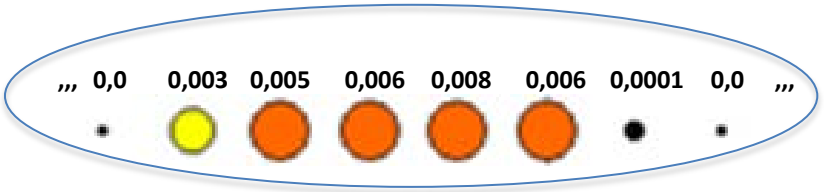
Projection of an object on GTM is described by the probability distribution (***responsibilities***) over the lattice nodes.



Probabilities (responsabilities) of acetylcholinesterase ligand projected into latent space

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 : areas of application

Conformational space Analysis

Data visualisation and analysis

Libraries comparison

Virtual screening

Structure-Activity modeling

Library design

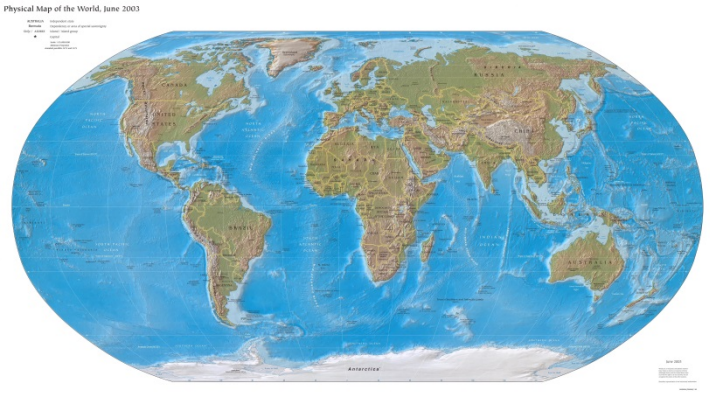
Inverse QSAR

Extraction of privilege patterns

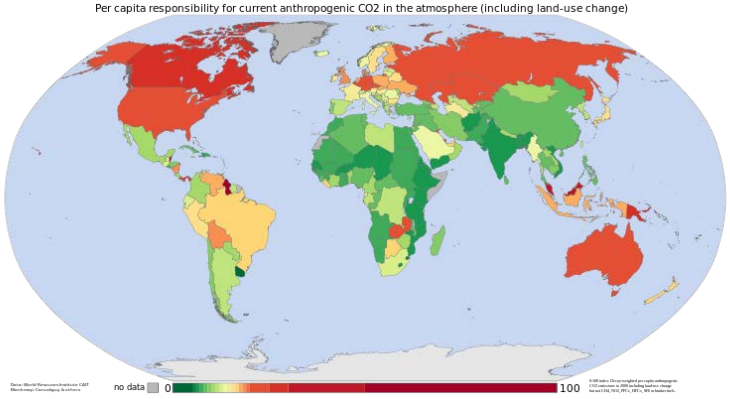
Chemical data analysis and activity prediction

Properties mapping

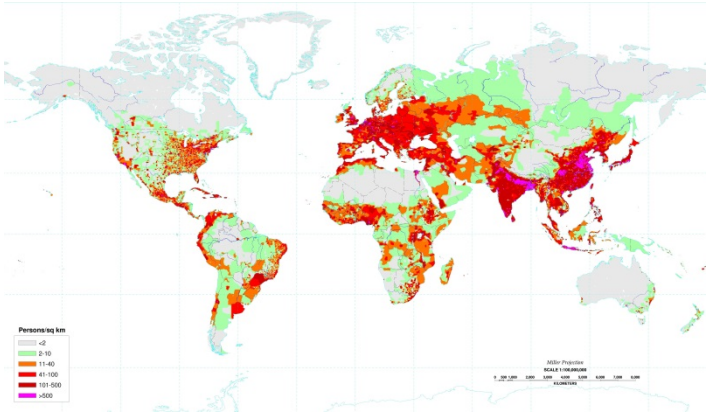
political map



physical map



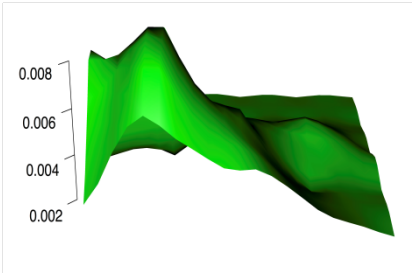
CO_2 emission



population density

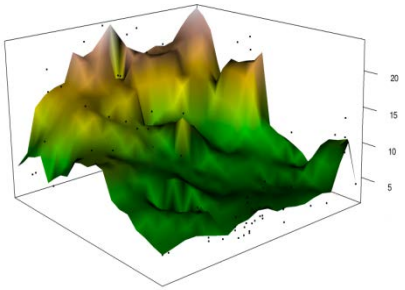
GTM property (activity) landscape

one molecule



$A_i R_{ik}$ - Activity value assigned to a node *k*

dataset



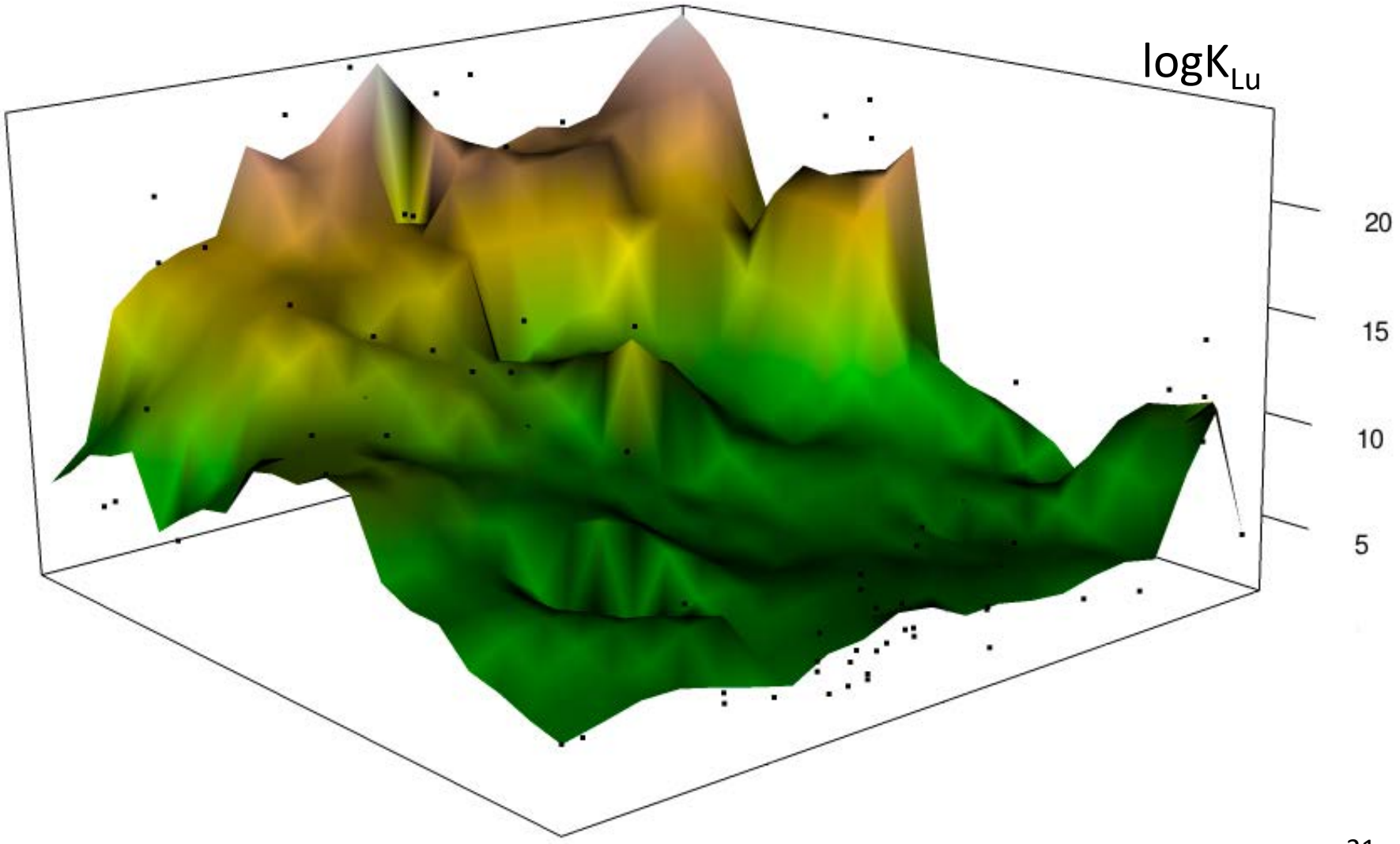
$$\overline{A}_k = \frac{\sum_i A_i R_{ik}}{\sum_i R_{ik}}$$

Expectation of activity $\bar{A}_k$ in node ***k***

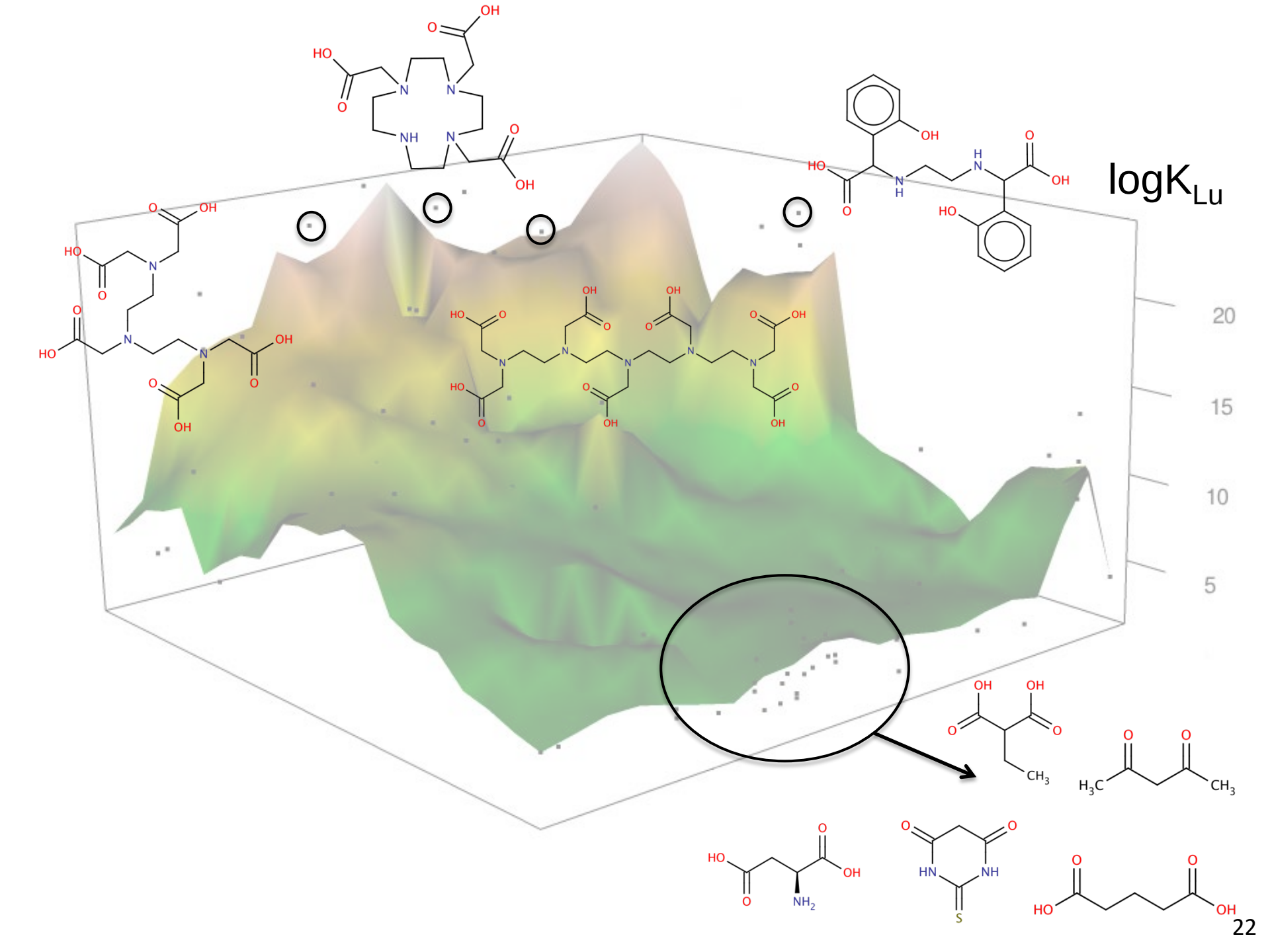
Dataset → property lanscape vector $\{\bar{A}_k\}$ of N_{nodes} length

GTM activity landscape

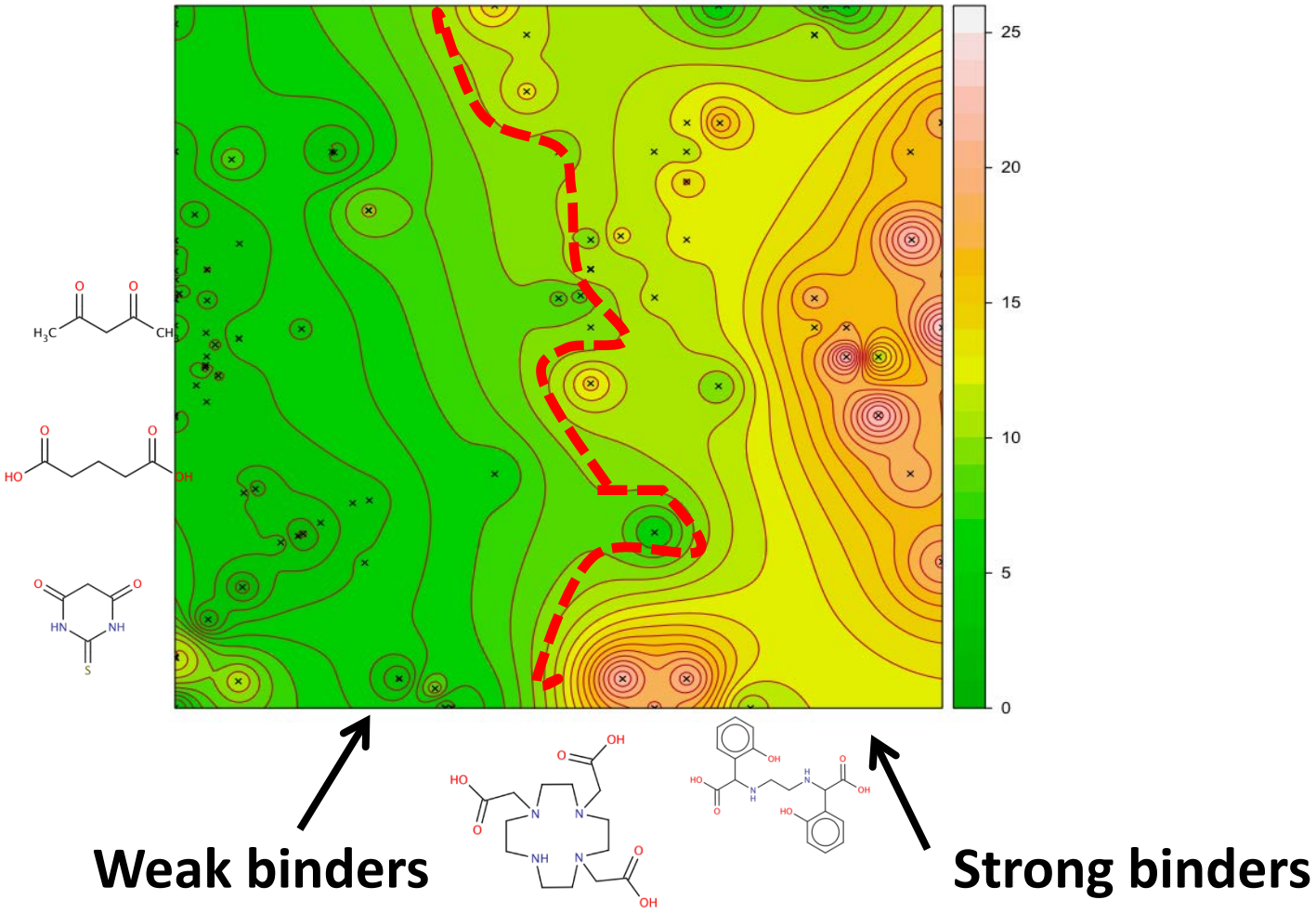
Stability of Lu^{3+} complexes with organic molecules



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

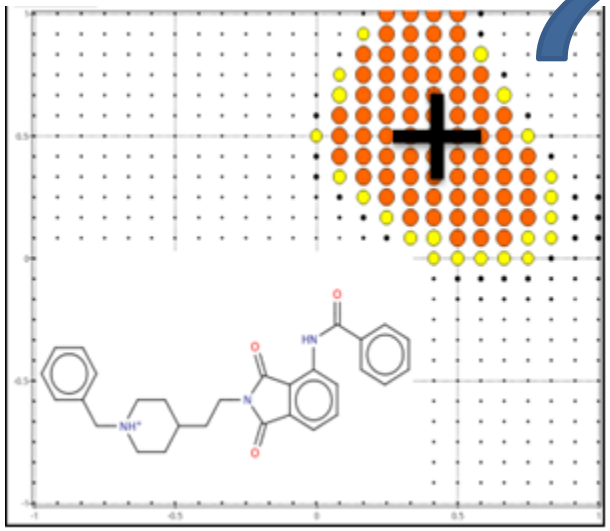

logK$_{Lu}$
20
15
10
5

Activity landscape for Lu^{3+} complexation

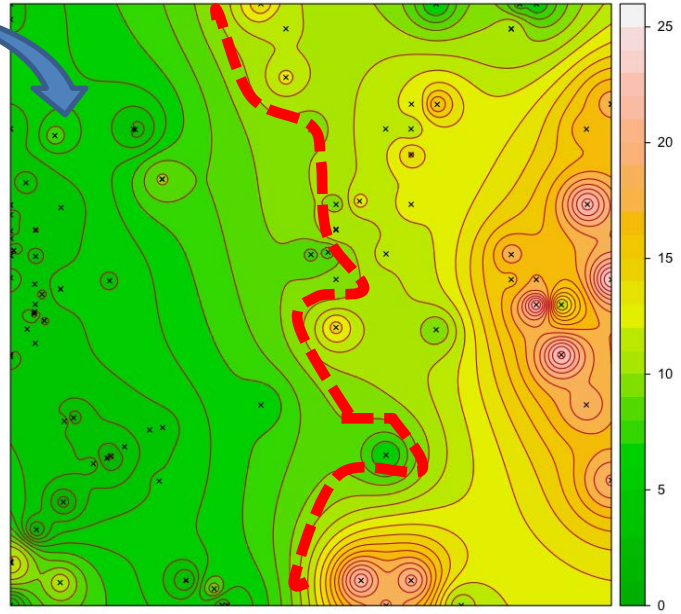


H. Gaspar, I. Baskin, D. Horvath, G. Marcou, A. Varnek *Mol. Informatics,* 2015, 34 (6-7), 348-356

Activity landscape: prediction on a test



Responsabilities' distribution of a test set cmpd



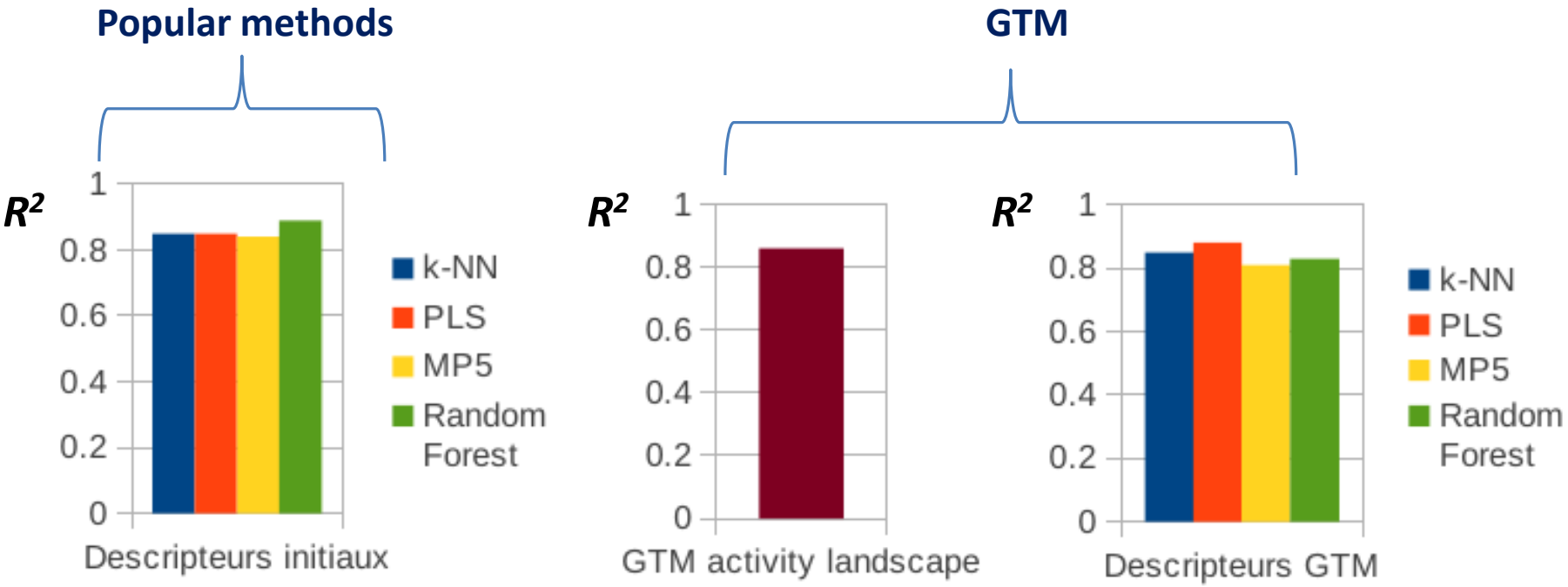
GTM for the training set

Predictions on a test set

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

Performance of GTM-based QSAR models

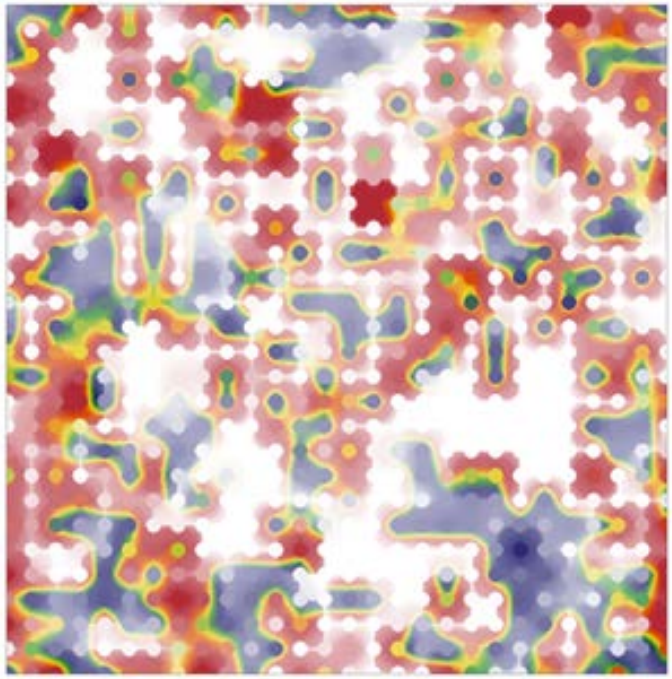
Regression models for LogK (Lu^{3+}) using ISIDA descriptors



GTM-based models preform similarly to those obtained with popular machine-learning methods

H. A. Gaspar , I. I. Baskin, G. Marcou, D. Horvath, A. Varnek *Mol. Informatics,* 2015

Class landscape for antiviral activity



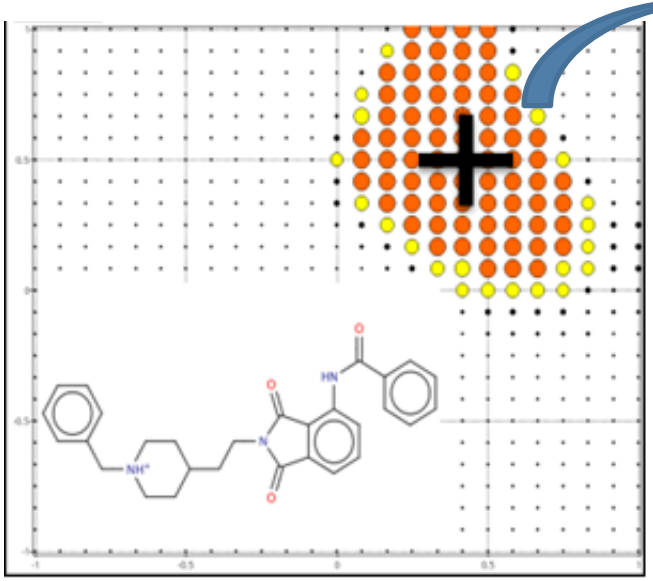
inactives

actives

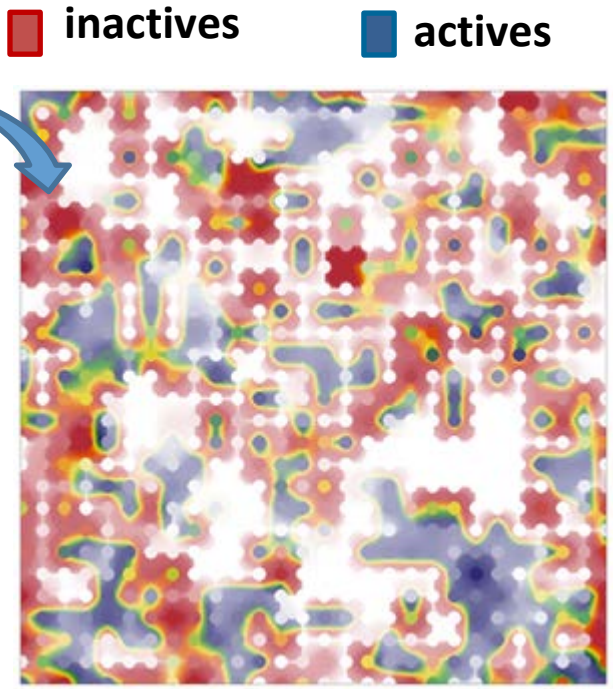
Class landscape can be used to predict a class (“*active*”/”*inactive*”) for any new compound

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

Class landscape for antiviral activity



Responsabilities' distribution of a test set cmpd



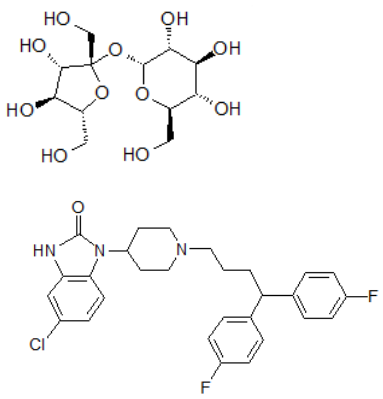
Probabilities to be "*active*" or "*inactive*") for a new compound is estimated using the Bayes equation

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

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

Toward « universal » map of chemical space

What do we expect from an “universal” map of the Chemical Space?



Map of the chemical space should:

- be representative with respect to the variety of known chemotypes;
- be able to distuinguish different activity classes and different chemotypes;
- be able to accomodate novel structures and activities in agreement with the neighborhood behavior principle

Universal map of chemical space

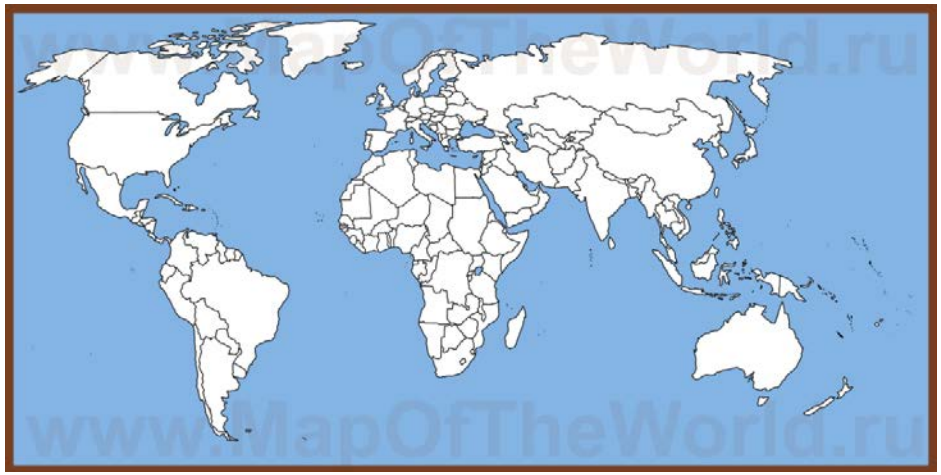
- As any machine-learning method, GTM is limited by the data size. For very large data sets, only part of molecules can be used for the manifold construction. Thus, a representative subset (a **"Frame" Set**) must be selected for this purpose.

Examples of Frame Sets

Non-representative

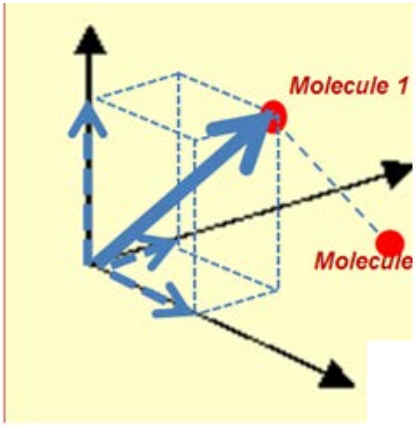
Hecataeus of Miletus (c. 550 – 476 BCE)

Representative



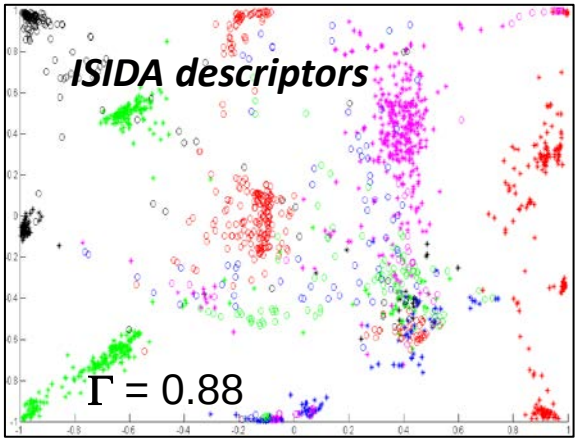
contour lines world map

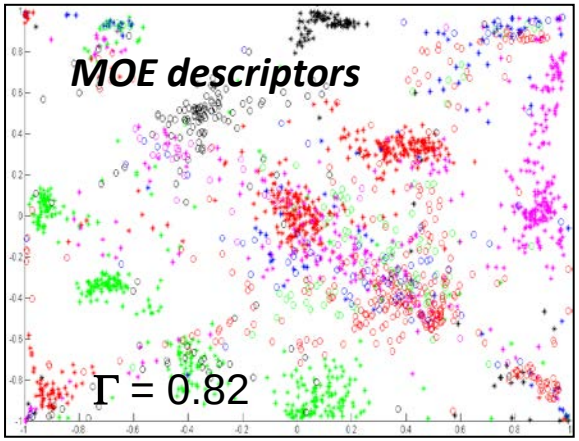
Choice of descriptors – a vital issue for chemical space construction

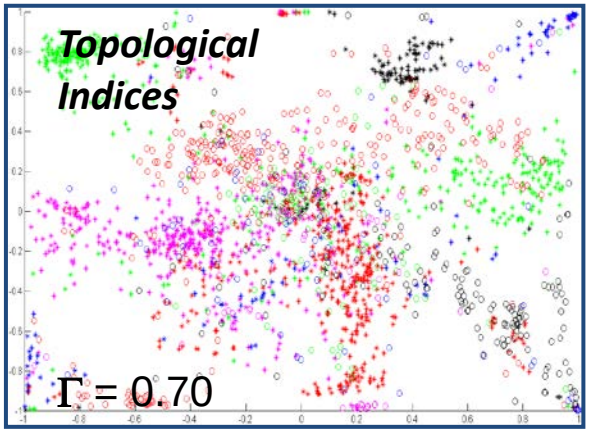


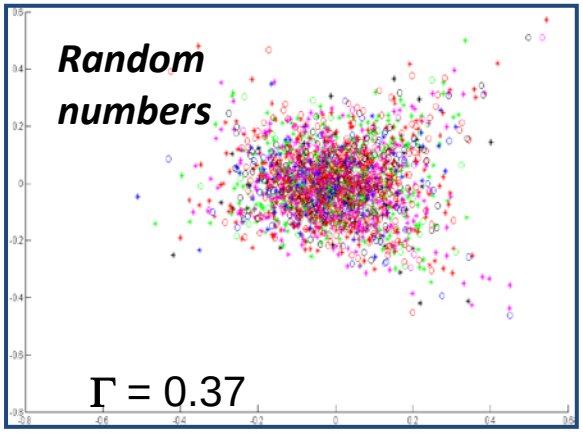
- Positioning of objects in the initial and latent spaces depends on the choice of molecular descriptors
- Is there a way to select some « optimal » descriptors for maps construction ?

DUD dataset: *GTM as a function of type of descriptors*



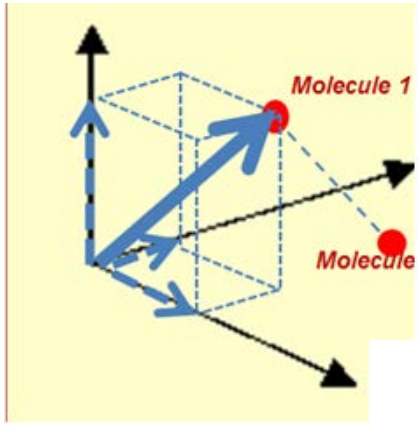






Γ-score – measures the ability of a model to produce clustering of similar structures in latent space. For ideal classes separation, **Γ** = 1.

Chemical space construction driven by (Q)SAR models



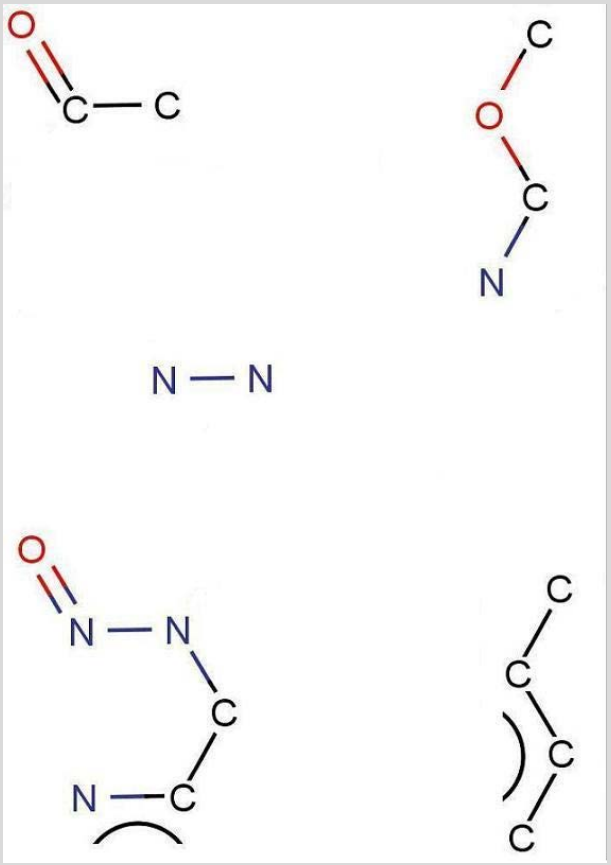
- Optimal descriptors are supposed to provide with the best GTM-based regression or classification models built on some « scoring » dataset(s).

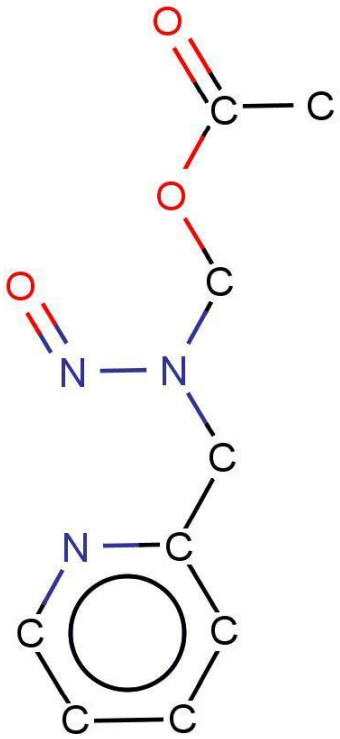
P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath J Comput Aided Mol Des, 2015, 29(12),1087-1108

ISIDA fragment descriptors

Sequences

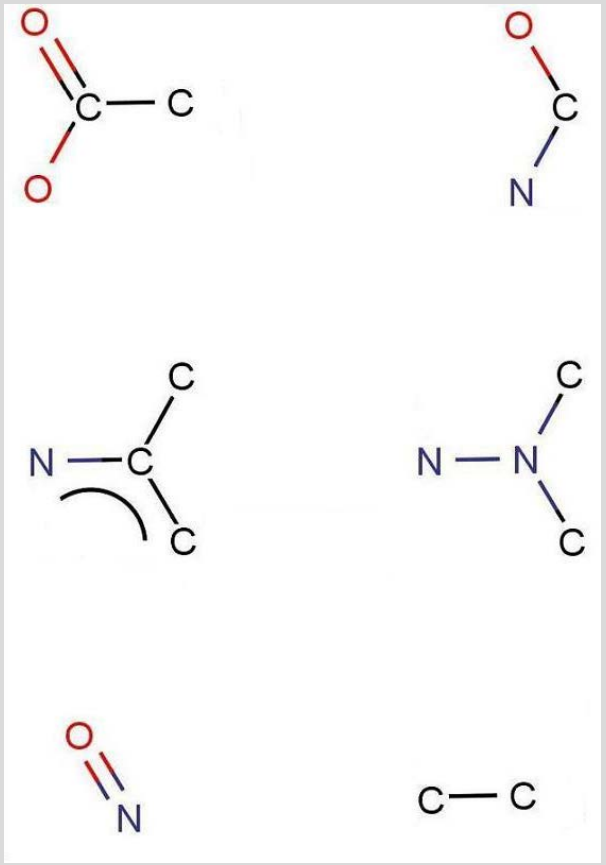
containing $2 < N < 15$ atoms



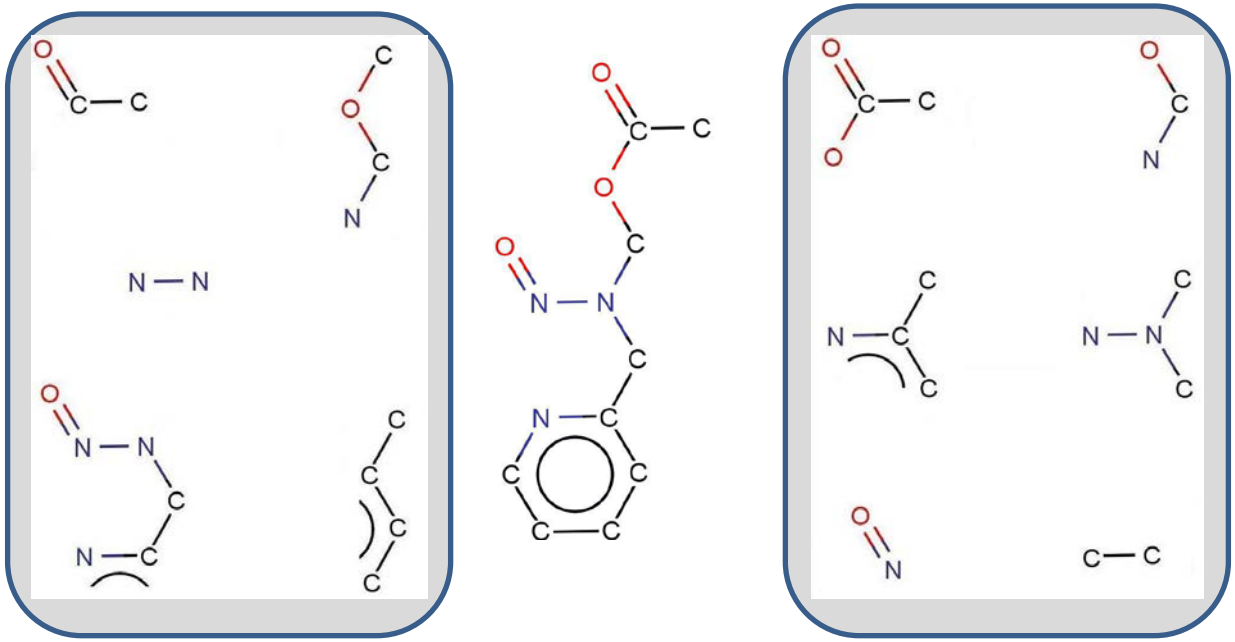


Augmented Atoms:

selected atoms with their closest neighbours

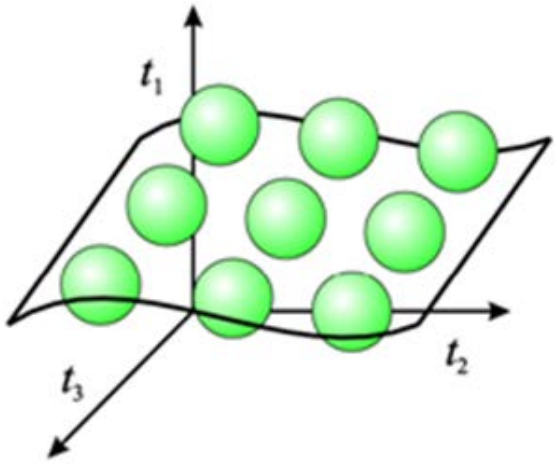


ISIDA fragment descriptors



Several hundreds types of fragment descriptors can be generated for one same data set

Toward universal map(s) of chemical space

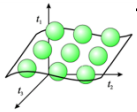
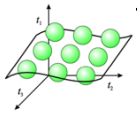
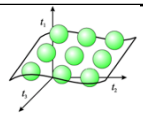


Activity	Model's performance
Act_1	BA_1
Act_2	BA_2
…….	
SCORE	**<BA>**

More activities are used, more universal is a map

P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath J Comput Aided Mol Des, 2015, 29(12),1087-1108

Toward universal map(s) of chemical space

Descriptors type	MAP	SCORE
 $Descr_1$	Map_1	$SCORE_1$
 $Descr_2$	Map_2	$SCORE_2$
…….		…….
 $Descr_N$	Map_N	$SCORE_N$

Decriptors leading to largest score are selected

P. Sidorov, H. Gaspar, G. Marcou, A. Varnek, D. Horvath J Comput Aided Mol Des, 2015, 29(12),1087-1108

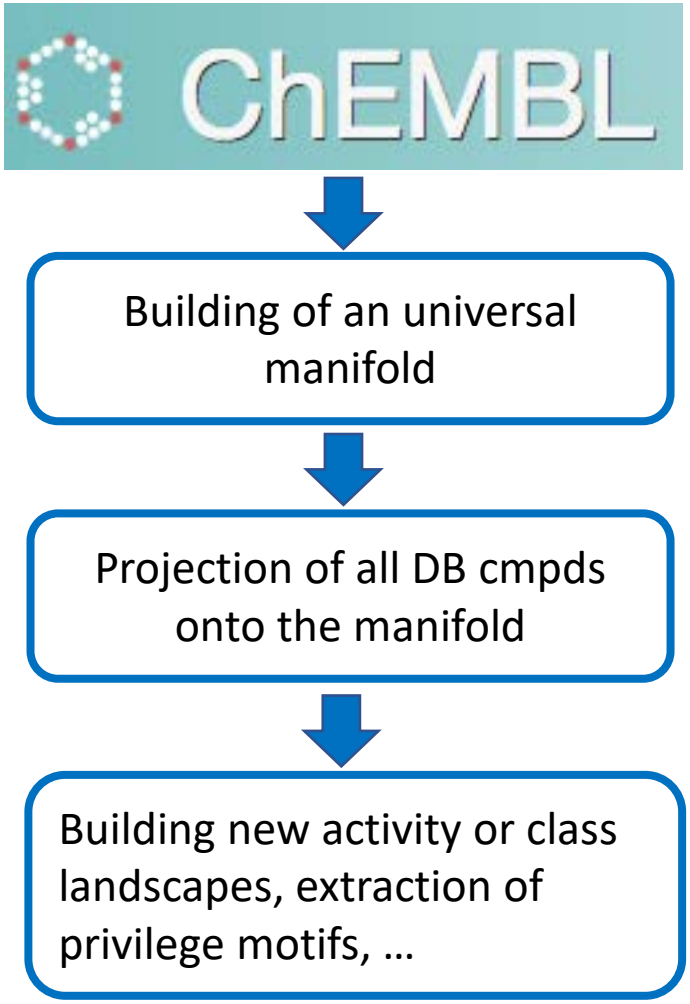
Chemical space of druglike compounds

2.2 M cmpds

Basic assumption:

If the manifold is trained on activity predictions of ligands for >100 different biological targets, it may also accommodate other biological activities and compounds.

Chemical space of druglike compounds



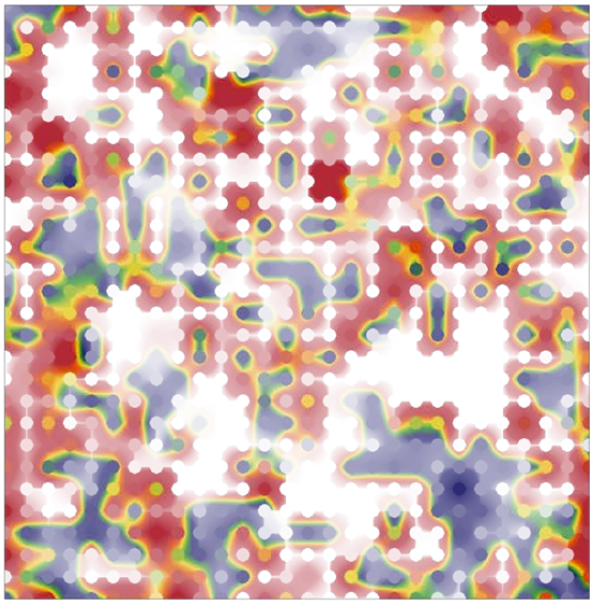
Universal manifold:

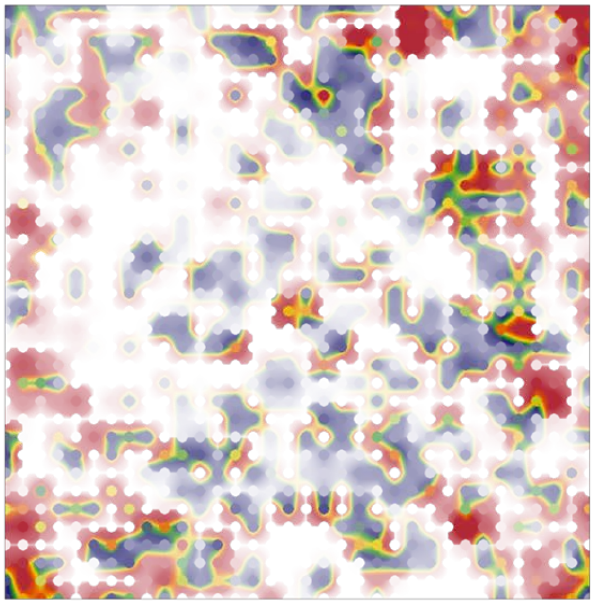
- Optimized for 144 sets (GPCR, kinases, …)
- Validated on > 450 sets

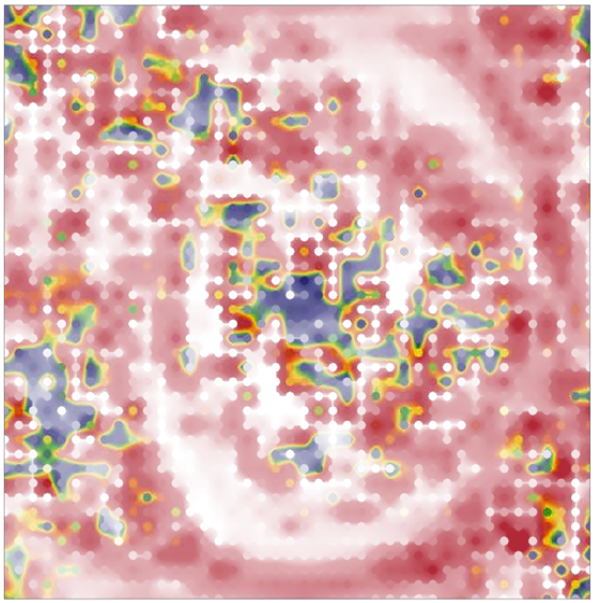
ChEMBL: chemical space of antiviral compounds

3 “universal” maps based on different types of ISIDA descriptors

inactives **actives**



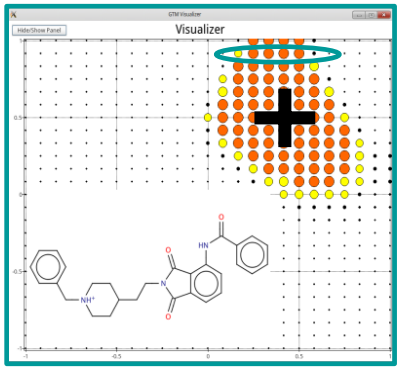


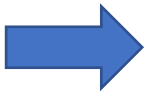


K. Klimenko, G. Marcou, D. Horvath, A. Varnek *J. Chem. Inf. Model.* 2016, 56, 1438−1454

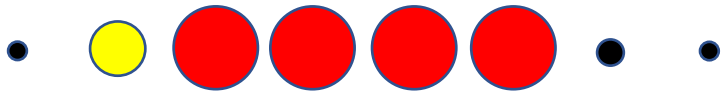
Privilege patterns extraction

From responsibility vector to binned pattern





Responsibility Vector

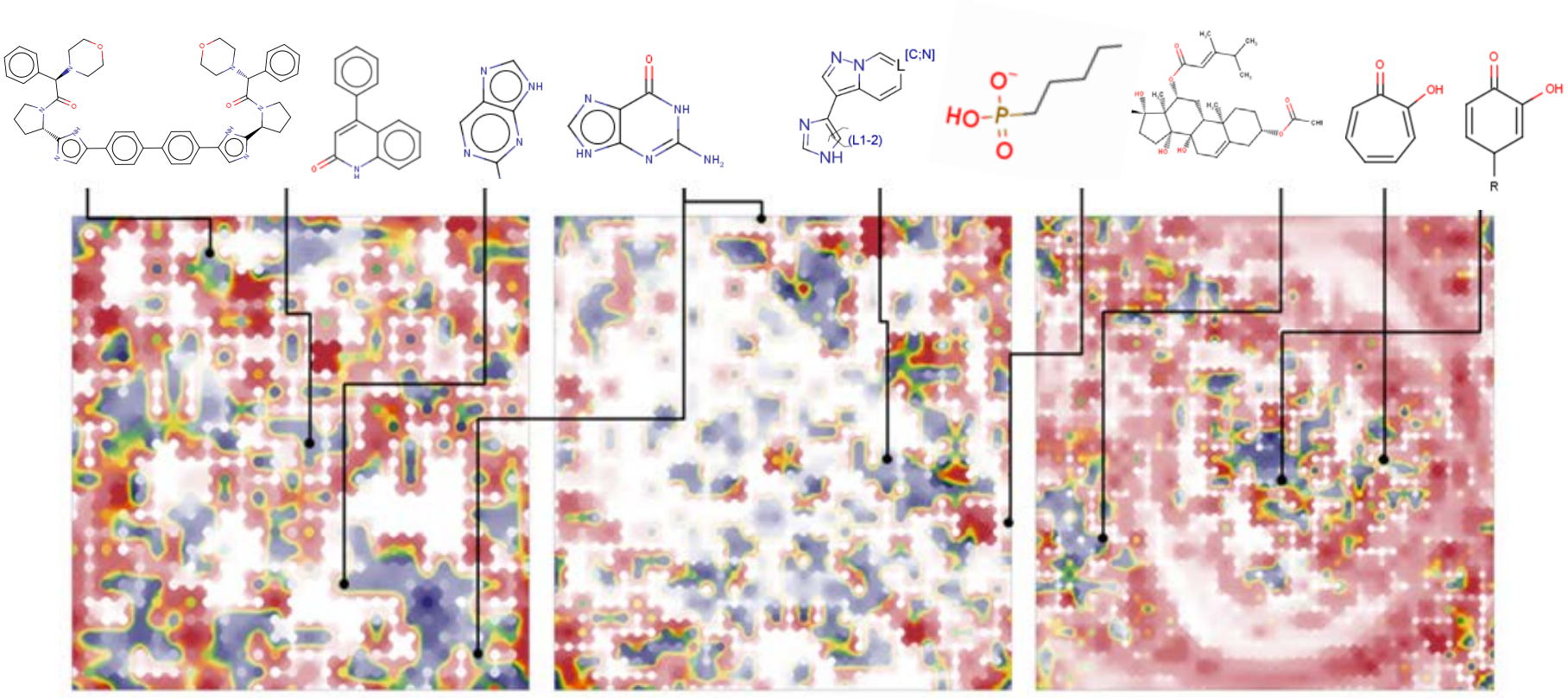


Node Number n	...	183	184	185	186	187	188	....
Probability of Residence R_n	0.0	0.005	0.020	0.022	0.031	0.021	0.002	0.0

Node Number n	...	183	184	185	186	187	188	....
Binned "Pattern" vector	0	1	2	2	3	2	0	0

ChEMBL: chemical space of antiviral compounds

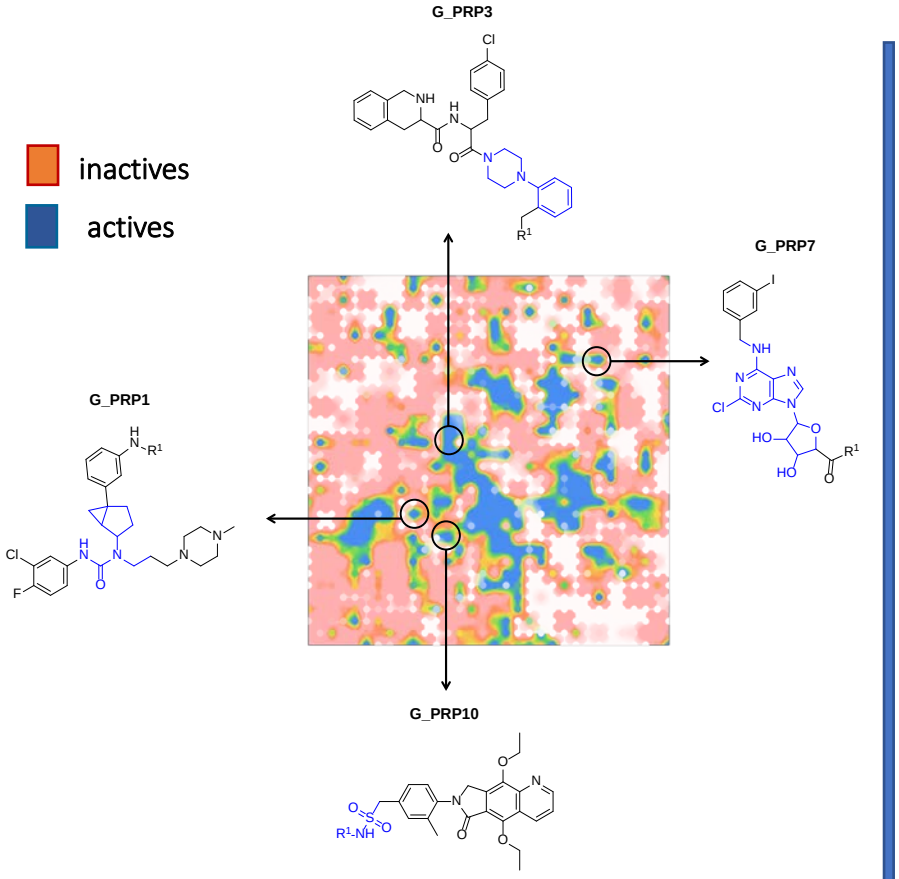
3 different maps are needed in order to extract the privilege patterns for main classes of antivirals

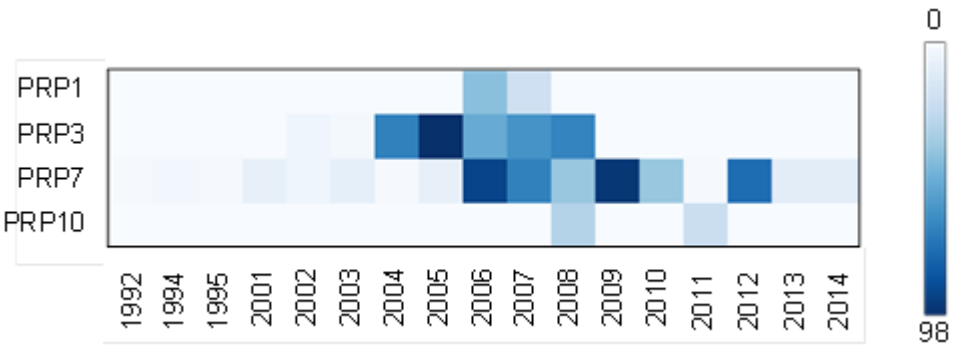


3 “universal” maps based on different types of ISIDA descriptors

GTM: extraction of privilege structural patterns

ChEMBL: class landscape for GPCR ligands





Evolution of privileged structural motifs of GPCR ligands in ChEMBL

S. Kayastha, D. Horvath, E. Gilberg, M. Gütschow, J. Bajorath, A. Varnek J. Chem. Inf. Model., 2017, 57, 1218−1232

Big Data challenge

Comparison of databases: GDB vs PubChem

GDB-17 – computer-generated virtual molecules containing up to 17 heavy atoms *

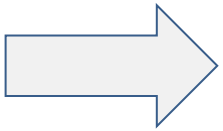
- whole set: 1.66 *10^{11} *virtual* molecules
- « lead-like » subset used in this study: **10 M** molecules

PubChem-17 – subset of **10.8 M** *real* molecules containing up to 17 heavy atoms exracted from the PubChem database

* L. Ruddigkeit et al. J Chem Inf Model 2012, 52, 2864–2875

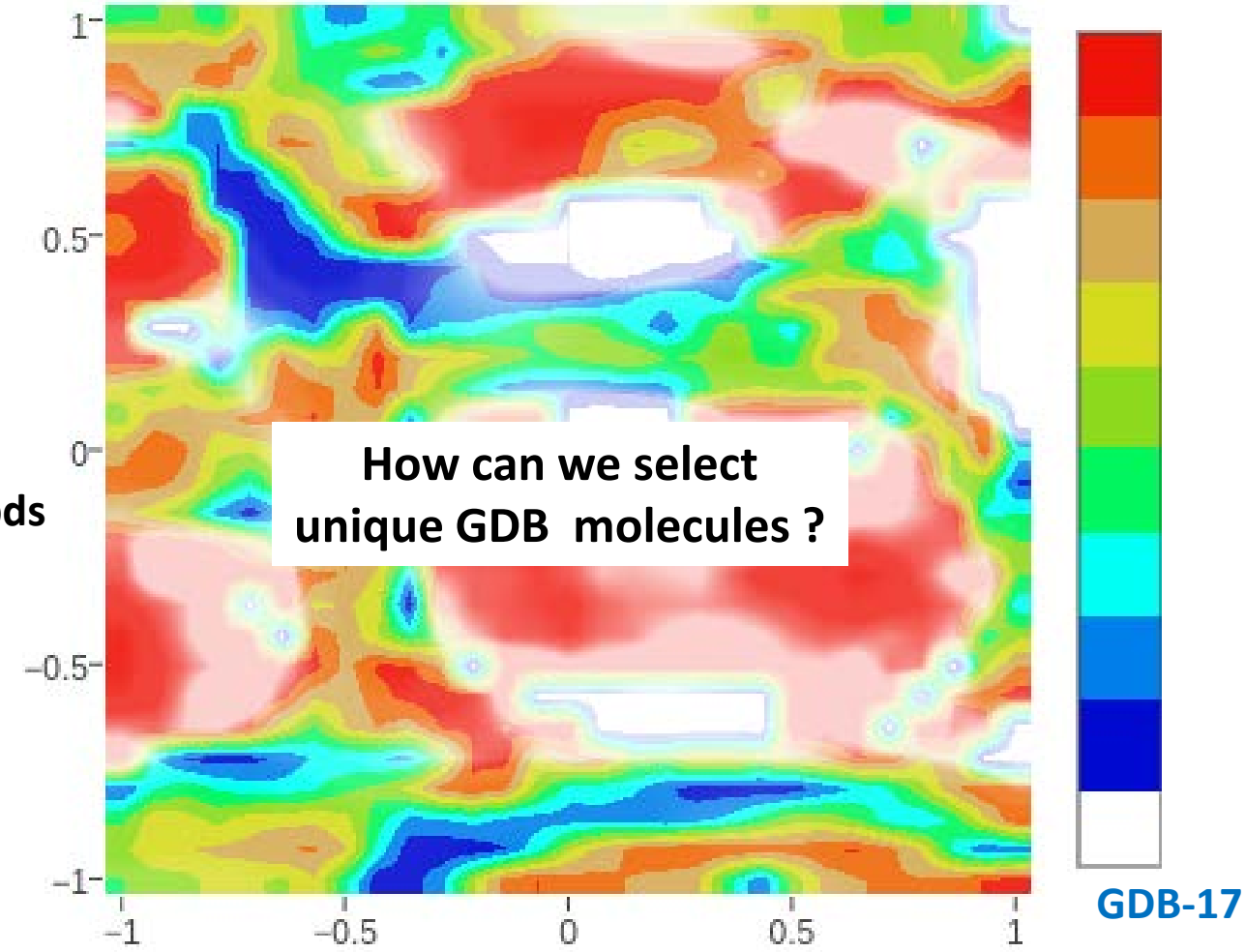
Comparison of databases: GDB vs PubChem

PubChem-17



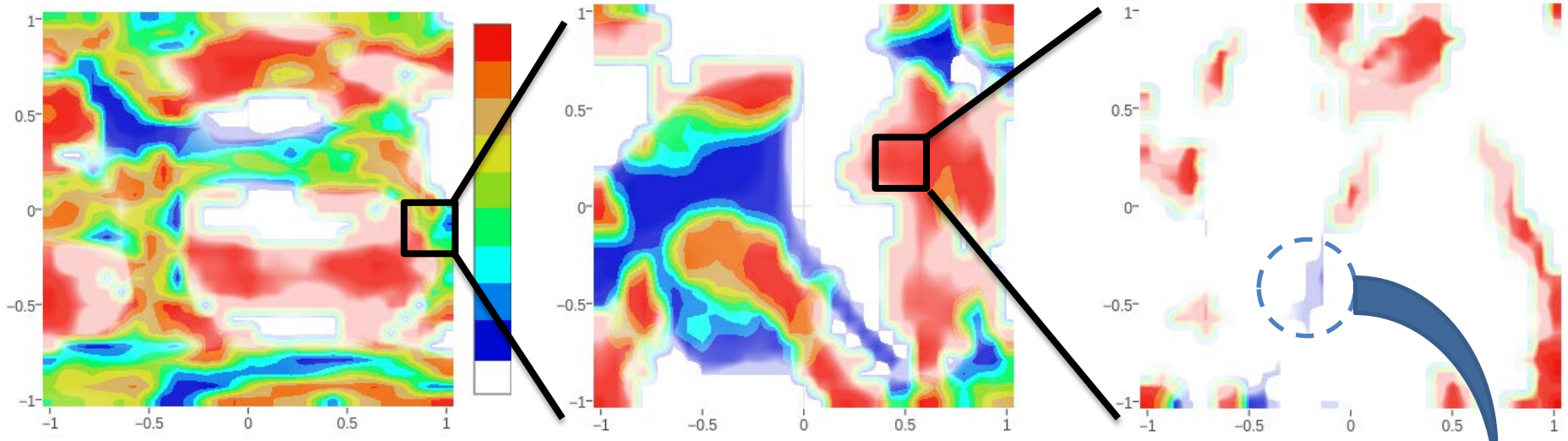
21*10^6 cmpds

GDB-17



Chemography: hierarchical GTM

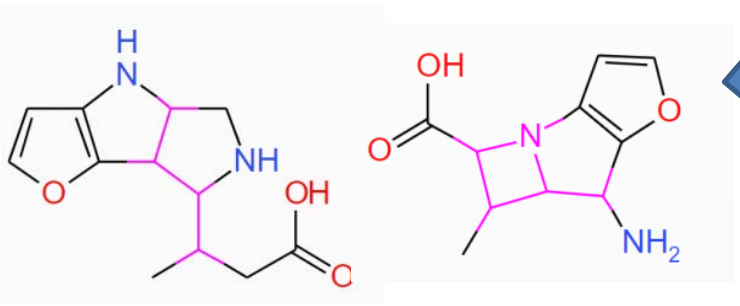
PubChem-17 ***vs*** **GDB-17**



$21*10^6$ cmpds

$0.65*10^6$ cmpds

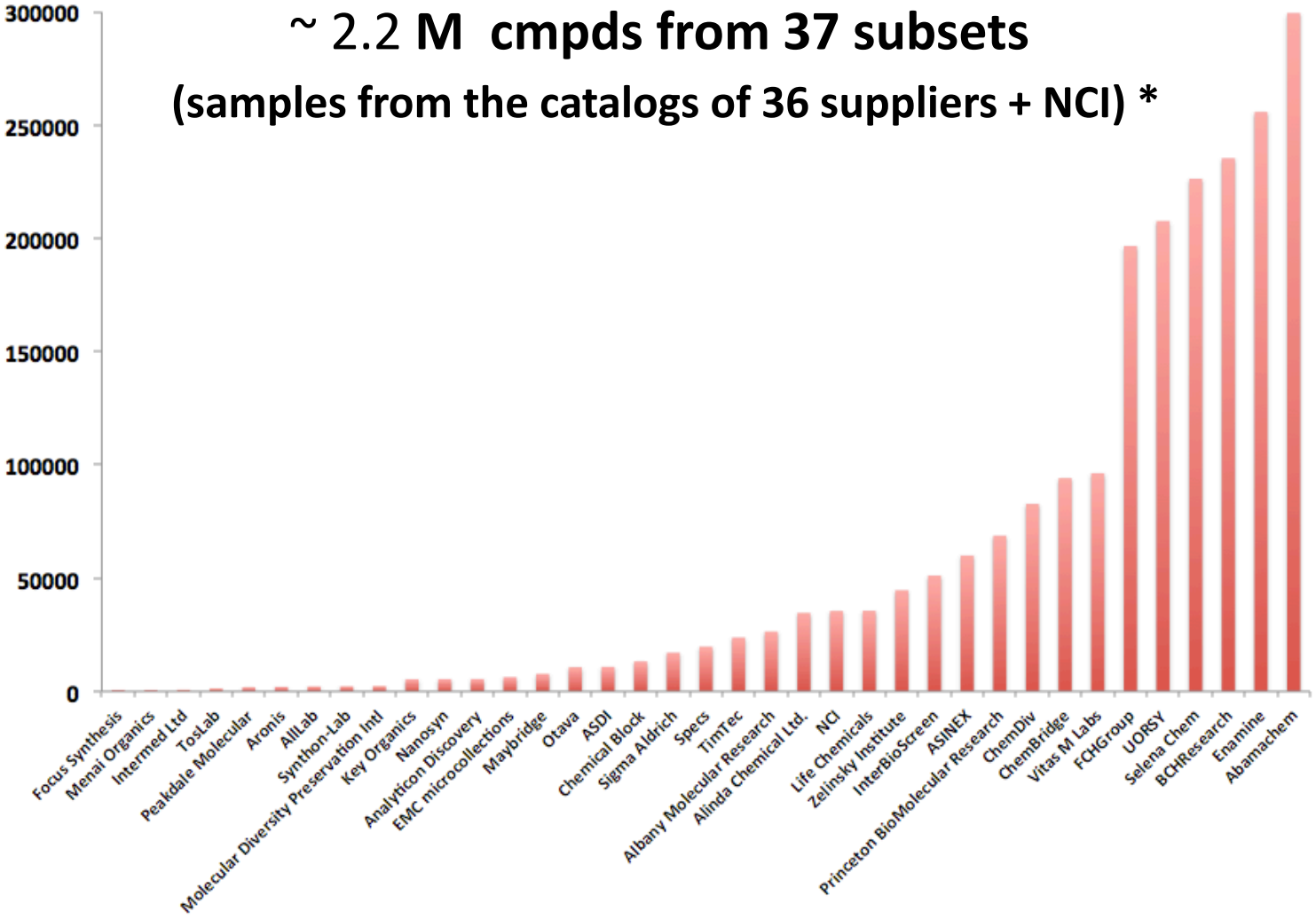
$2.5*10^3$ cmpds



No analogues in PubChem

Databases comparison and properties profiling

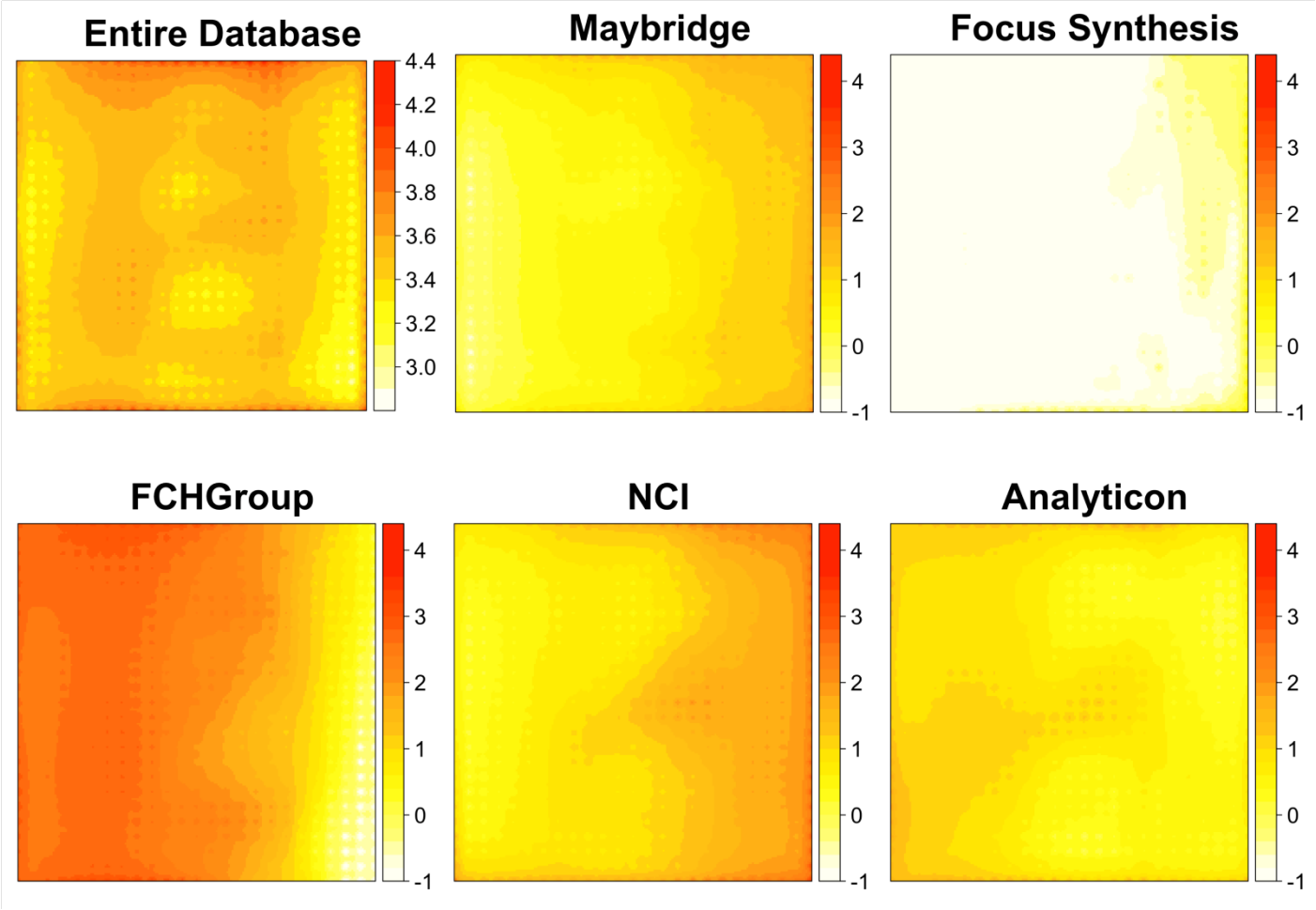
Comparison of suppliers databases



* **Data from:** T. Petrova et al. *Med. Chem. Commun.*, 2012, 3, 571-579

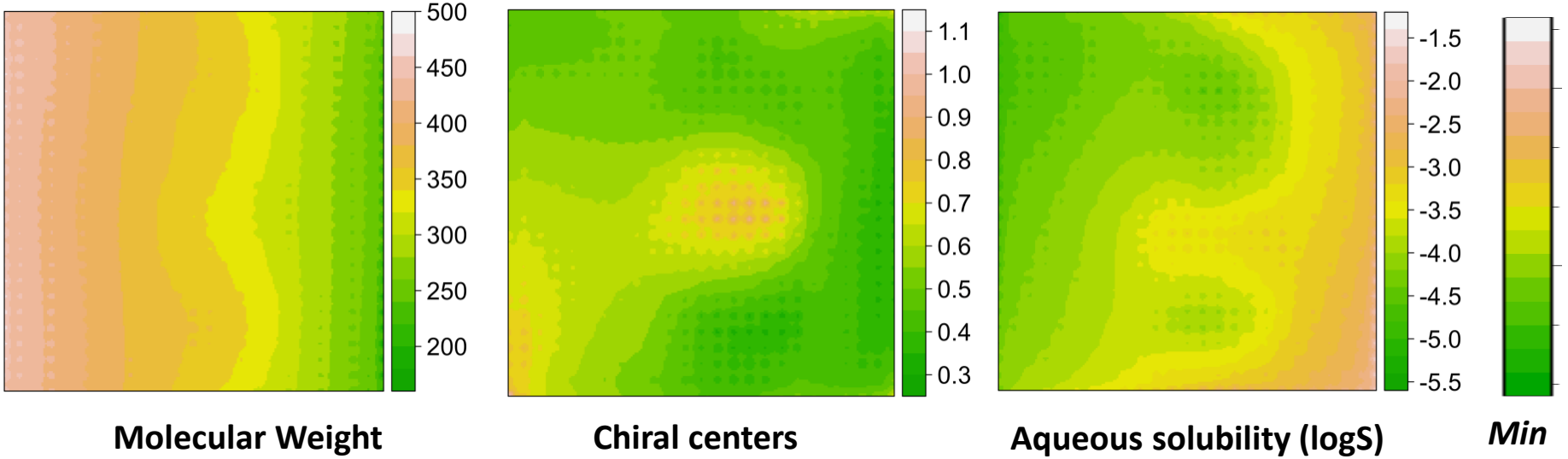
GTM for the Suppliers DB (> 2 M cmpds)

Data density distributions built on responsibility vectors



H.Gaspar, Igor I. Baskin, G.Marcou, D.Horvath and A.Varnek *J. Chem. Inf. Model.,* 2015, *55* (1), 84–94

Suppliers DB: GTM property landscapes

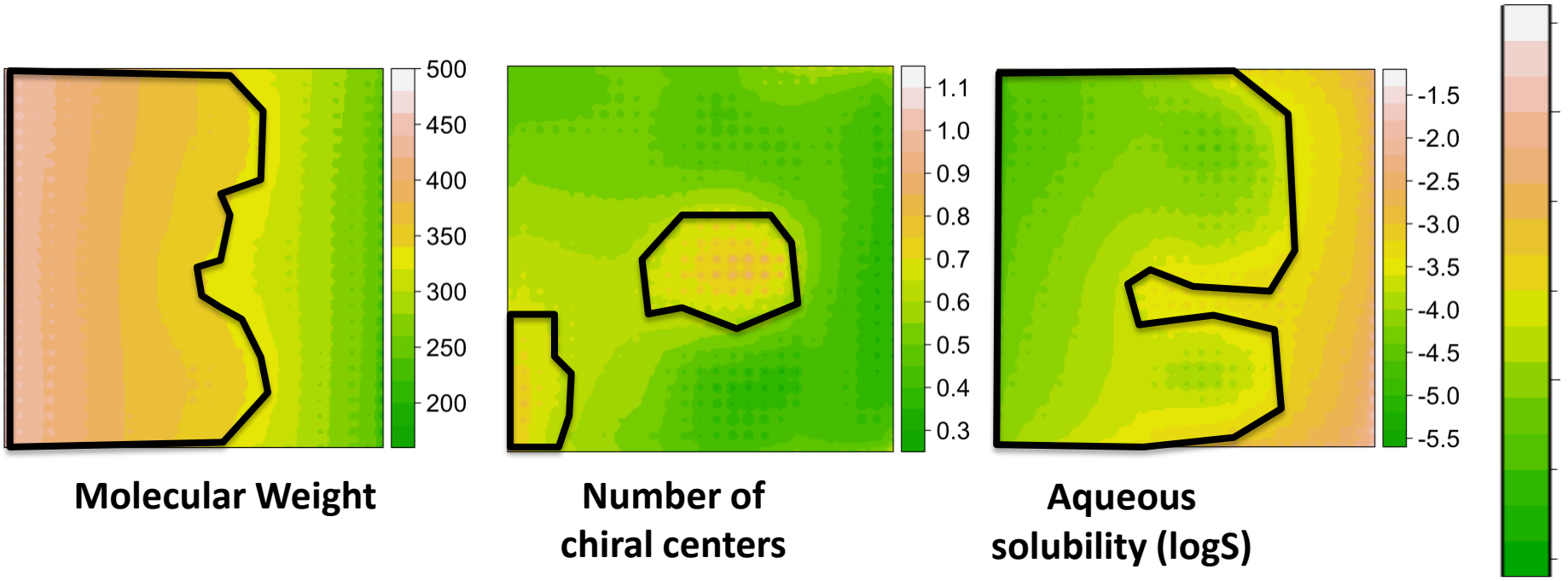


Molecular Weight **Chiral centers** **Aqueous solubility (logS)**

Suppliers DB: GTM property landscapes

High molecular weight **&** High chirality **&** Low solubility

Max



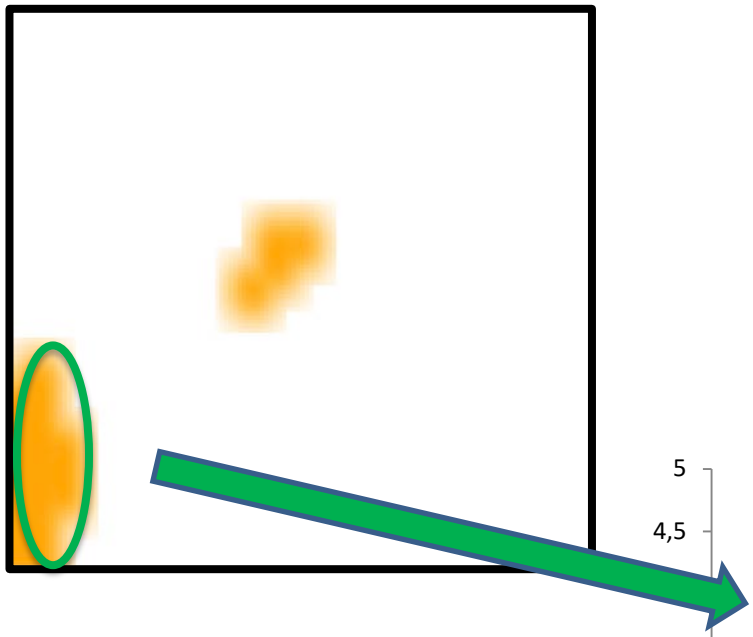
Molecular Weight

Number of chiral centers

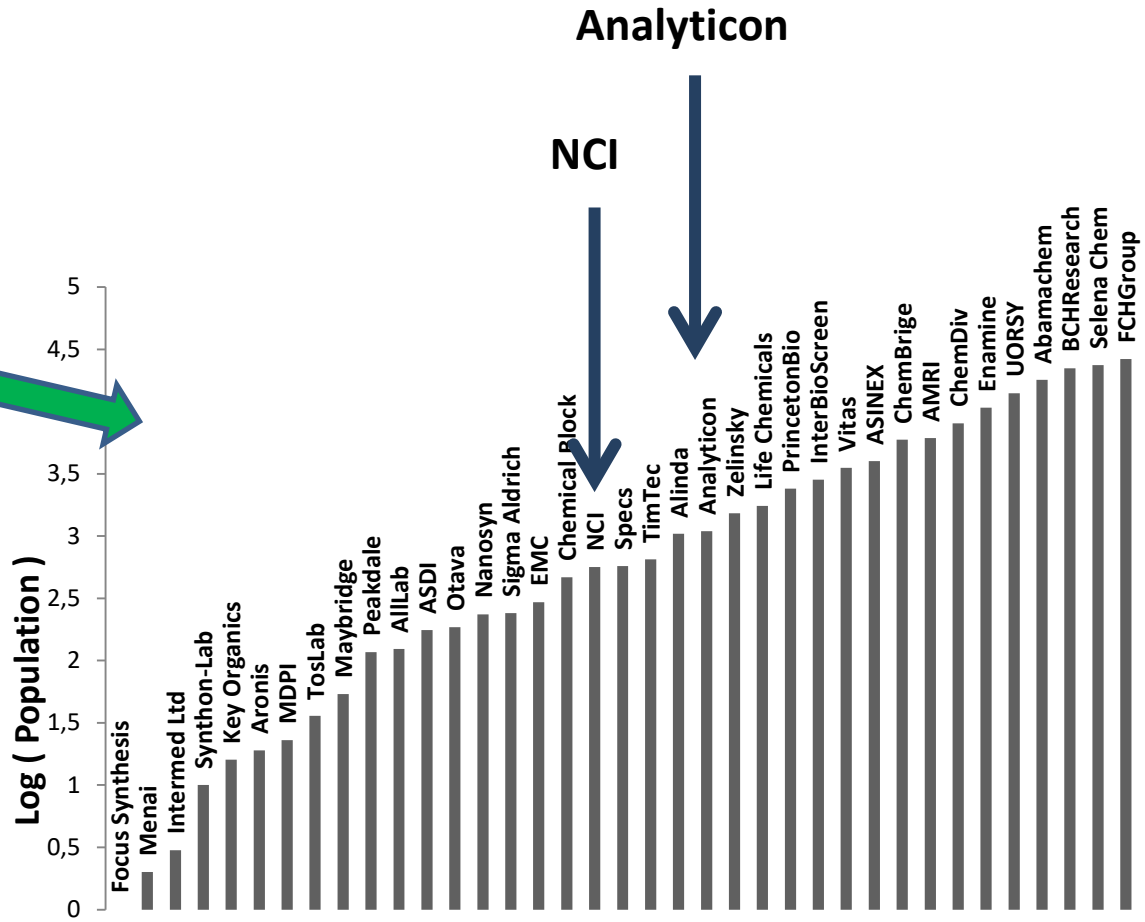
Aqueous solubility (logS)

Min

Suppliers DB map: regions of interest

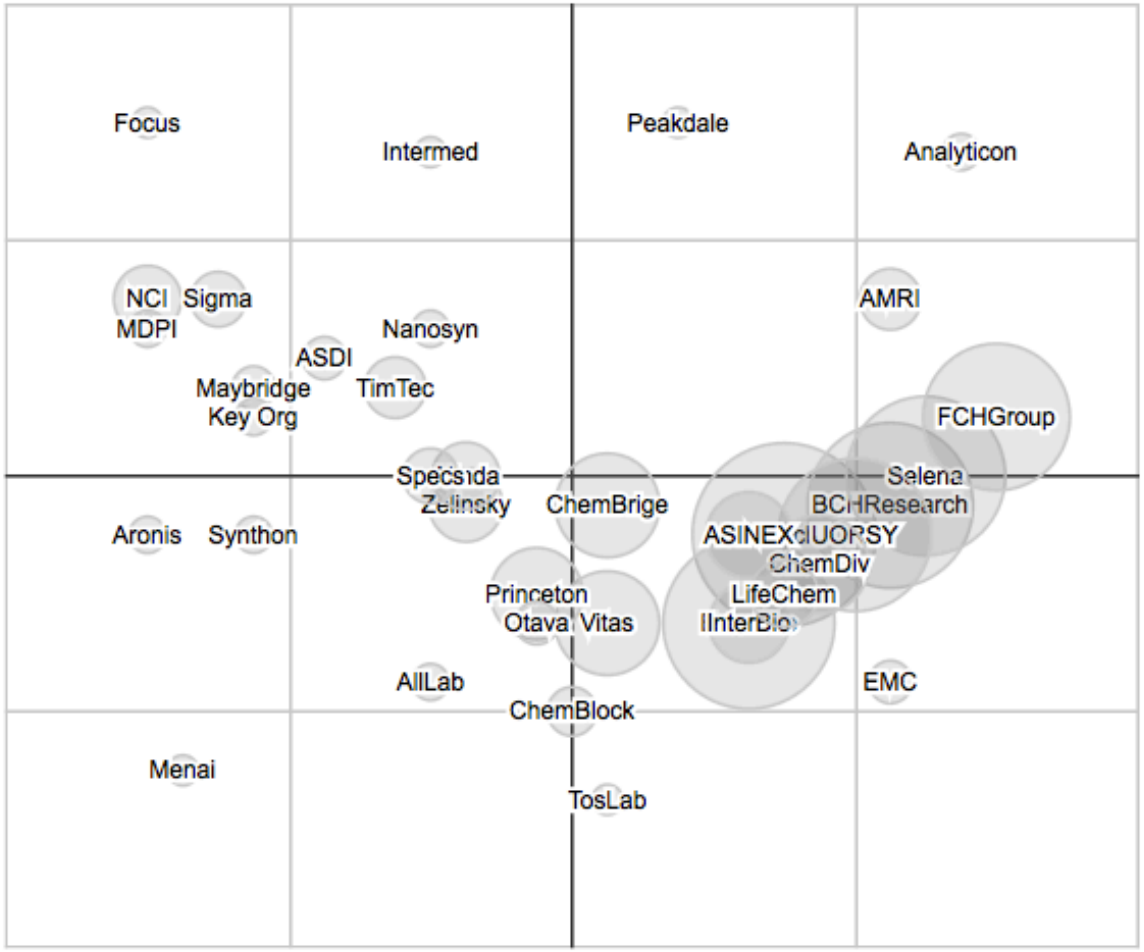


High molecular weight
&
High chirality
&
Low solubility



H.Gaspar, Igor I. Baskin, G.Marcou, D.Horvath and A.Varnek *J. Chem. Inf. Model.,* 2015, *55* (1), 84–94

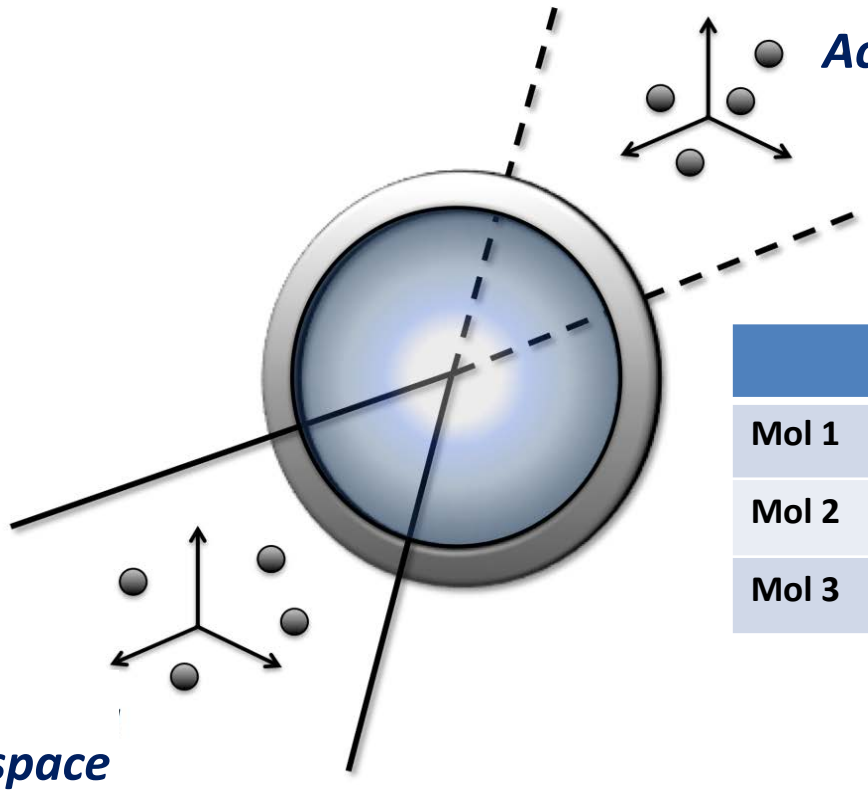
Meta-GTM: individual libraries as datapoints



each library is described by GTM descriptors


STARGATE
GTM
Metro Goldwyn Mayer
TRADE
MARK

Stargate GTM (S-GTM) setup



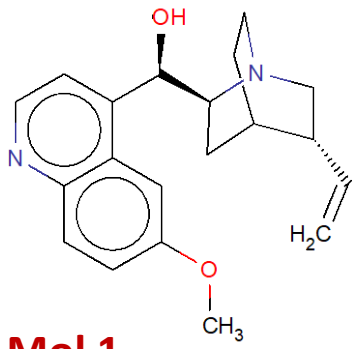
Activities space

	Act1	Act2	Act3
Mol 1	6.24	9.35	7.11
Mol 2	7.32	10.22	8.53
Mol 3	6.03	6.78	7.11

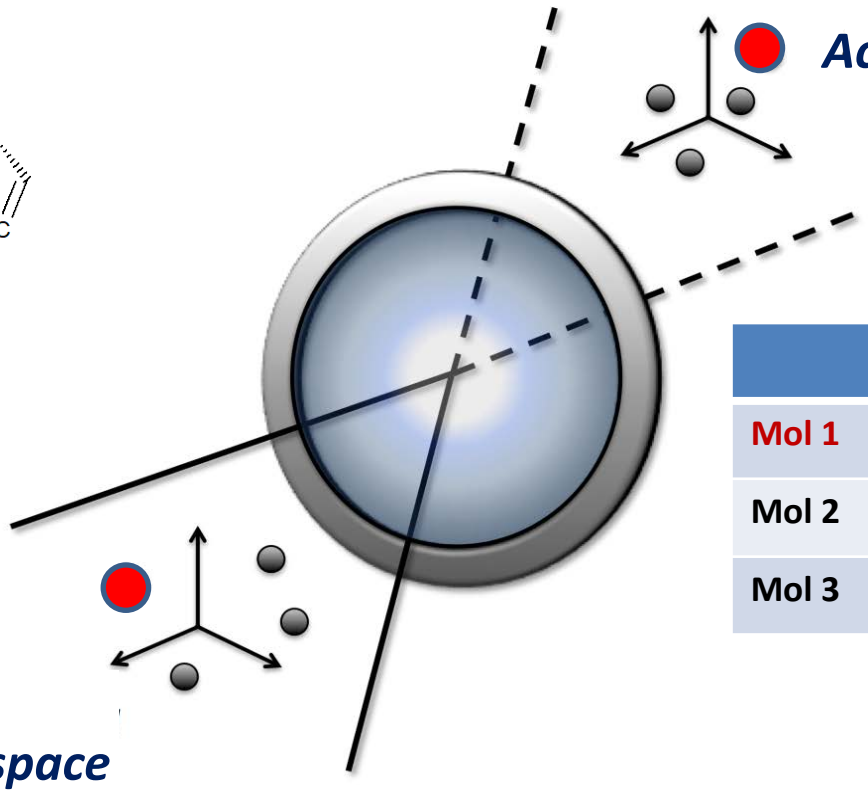
Descriptors space

	Descr1	Descr2	Descr3
Mol 1	0.22	5.43	1.12
Mol 2	7.18	0.96	-5.42
Mol 3	-1.01	7.41	10.63

Stargate GTM (S-GTM) setup



Mol 1



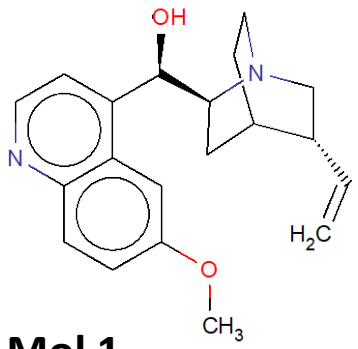
Activities space

	Act1	Act2	Act3
Mol 1	**6.24**	**9.35**	**7.11**
Mol 2	7,32	10,22	8,53
Mol 3	6.03	6.78	7.11

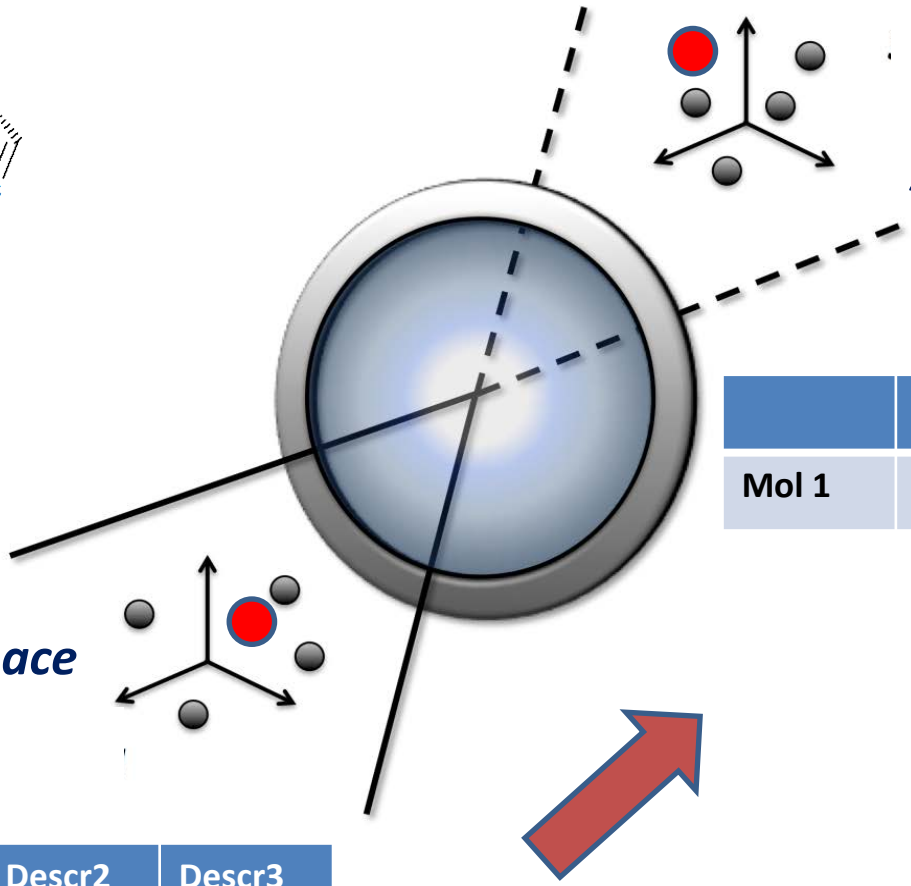
Descriptors space

	Descr1	Descr2	Descr3
Mol 1	**0.22**	**5.43**	**1.12**
Mol 2	7.18	0.96	-5.42
Mol 3	-1.01	7.41	10.63

S-GTM: prediction of pharmacological profile



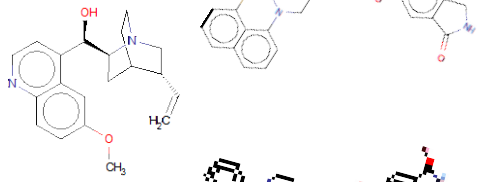
Mol 1

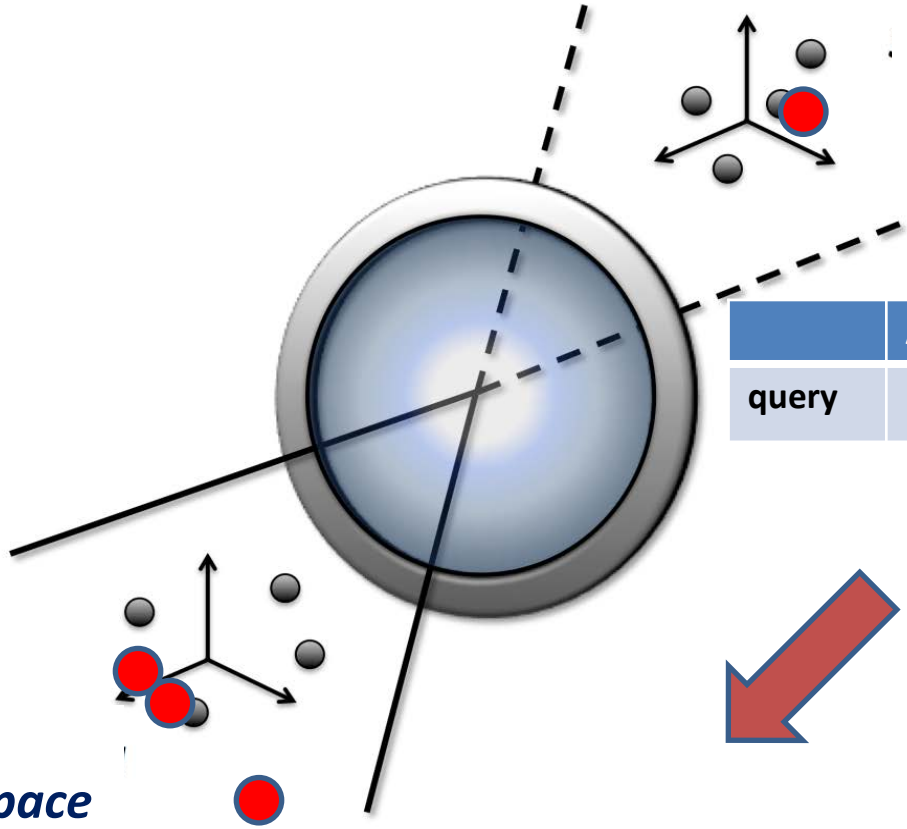


	Act1	Act2	Act3
Mol 1	??	??	??

	Descr1	Descr2	Descr3
Mol 1	0.22	5.43	1.12

S-GTM: discovery of structures corresponding to a given pharmacological profile





Activities space

	Act1	Act2	Act3
query	6	9	6

Descriptors space

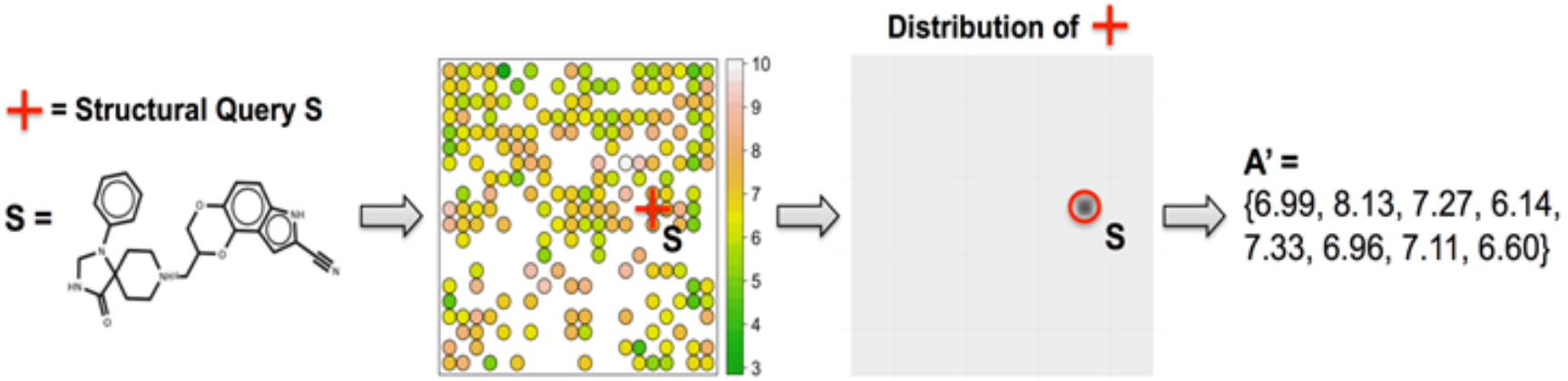
	Descr1	Descr2	Descr3
Mol 1	??	??	??
Mol 2	??	??	??
Mol 3	??	??	??

Case study: S-GTM models for a set of 8 GPCR activities

ID	Target name	Mol 1	Mol 2	...	Mol 1325
A2a	Alpha-2a adrenergic receptor	AFFINITY		VALUES	
D2	Dopamine D2 receptor	AFFINITY		VALUES	
D3	Dopamine D3 receptor	AFFINITY		VALUES	
D4	Dopamine D4 receptor	AFFINITY		VALUES	
S1a	Serotonin 1a (5-H1a) receptor	AFFINITY		VALUES	
S2a	Serotonin 2a (5-H2a) receptor	AFFINITY		VALUES	
S7	Serotonin 7 (5-HT7) receptor	AFFINITY		VALUES	
ST	Serotonin transporter	AFFINITY		VALUES	

S-GTM: retrieval of structures corresponding to a given pharmacological profile

Descriptors space → ***Activities space***



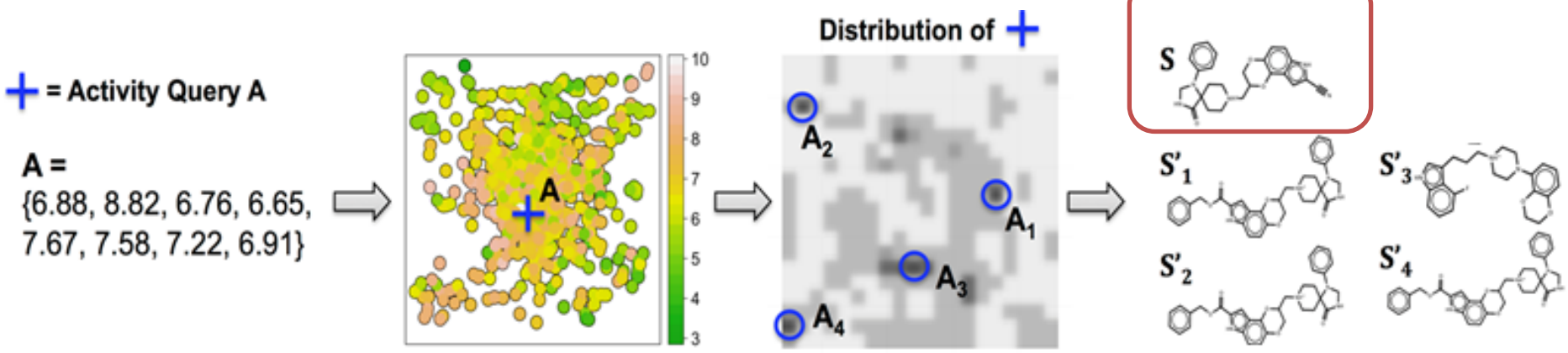
EXP

A =
{6.88, 8.82, 6.76, 6.65,
7.67, 7.58, 7.22, 6.91}

H. A. Gaspar , I. I. Baskin, G. Marcou, D. Horvath, A. Varnek *J. Chem. Inf. Model.,* 2015, 55 (11), 2403–2410

S-GTM: retrieval of structures corresponding to a given pharmacological profile

Activities space → *Descriptors space*



H. A. Gaspar , I. I. Baskin, G. Marcou, D. Horvath, A. Varnek *J. Chem. Inf. Model.,* 2015, 55 (11), 2403–2410

ISIDA-GTM Software

- **GTM Algorithms:**

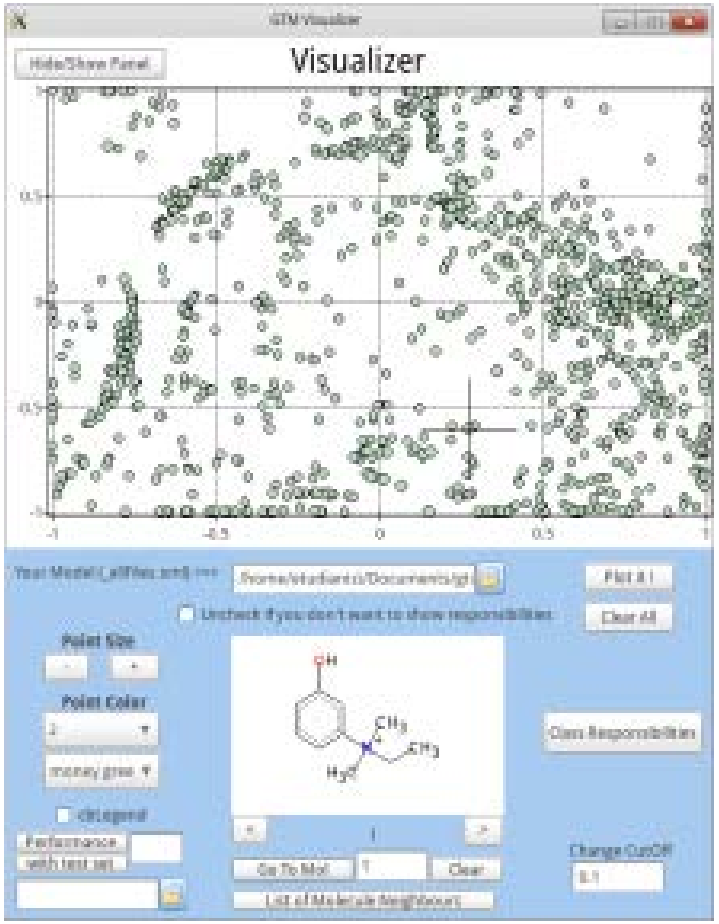
  « classical », incremental and kernel,

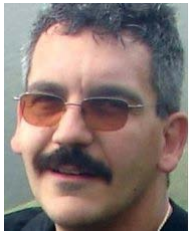
- **Modeling**:

  regression and classification models

- **Visualization**:

  data points, data probability distribution, activity landscapes, chemical structures





Chém Informatique
STRASBOURG