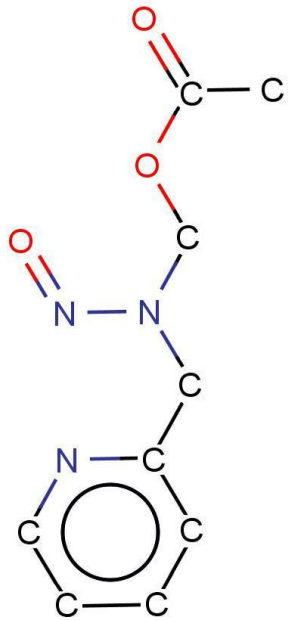
# Chemical space modeling and visualization with generative topographic maps

G. Marcou, A. Varnek, D. Horvath+

+Laboratoire de Chémoinformatique, UMR7140, Université de Strasbourg

# 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

.....

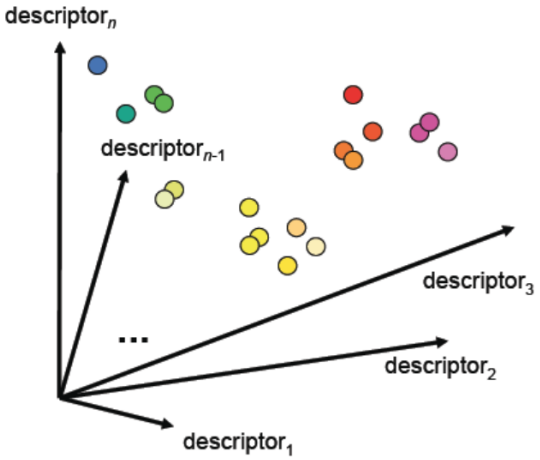
**Descriptor vector**

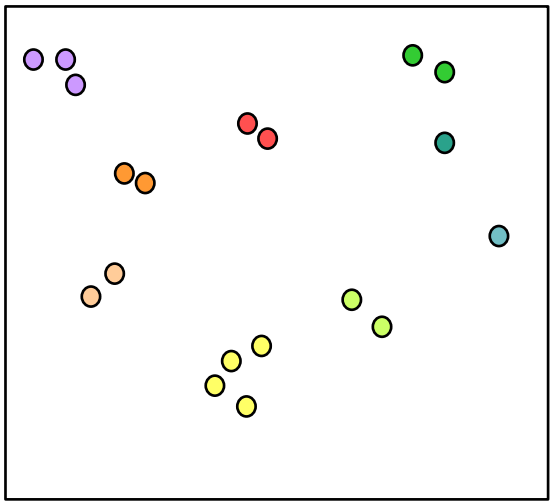
| Descriptor | Value |
|---|---|
| $D_1$ | $a_1$ |
| $D_2$ | $a_2$ |
| .... | ... |
| $D_i$ | $a_i$ |
| .... | ... |

**> 5000 types of descriptors are used**

# Chemography: cartography of chemical space

**Data visualization** → **dimensionality reduction problem**



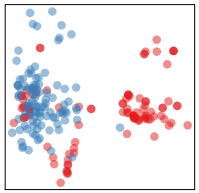
**Data space**
***(N-dimensional)***

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

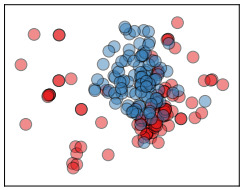
**Latent space**
***(2-dimensional)***

# Dimensionality reduction methods

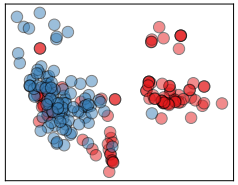
**Acetylcholinesterase dataset (DUD) : 100 actives and 100 inactives**



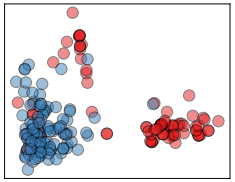
**Multi-Dimensional Scaling**



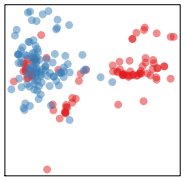
**Canonical Correlation Analysis**



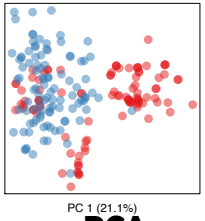
**Independent Component Analysis**



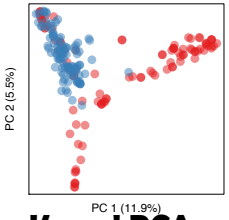
**Exploratory Factor Analysis**



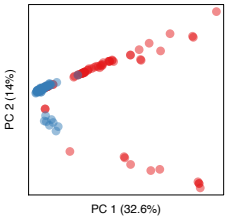
**Sammon map**



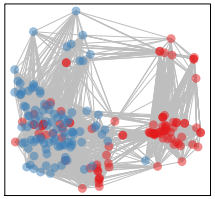

**PCA**



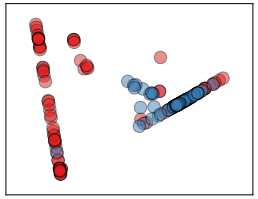

**Kernel PCA (RBF kernel)**



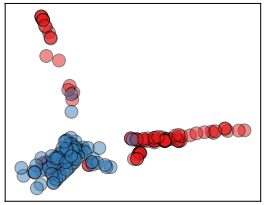

**Kernel PCA (polynomial kernel)**



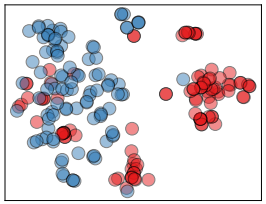
**Isomap**



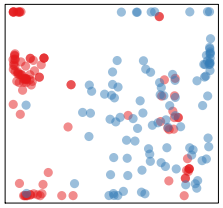
**Locally Linear Embedding**



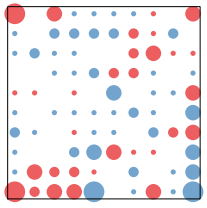
**Laplacian Eigenmaps**



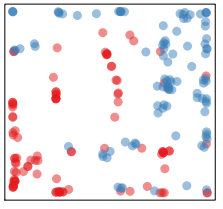
**t-SNE**



**Autoencoder dimensionality reduction**



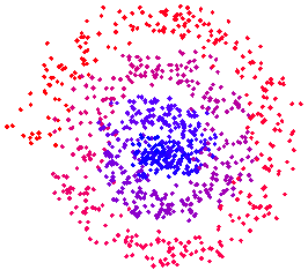
**SOM**

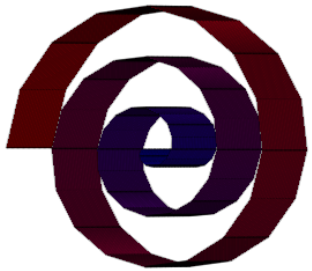


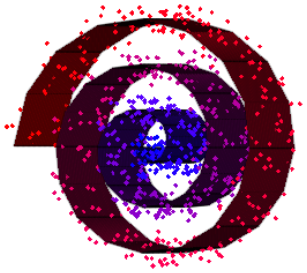
**GTM**

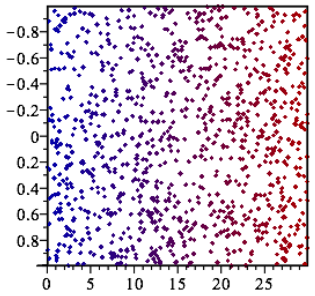
# THE GENERATIVE TOPOGRAPHIC MAPPING ALGORITHM

# Generative Topographic Map










Initial space

- A dataset is distributed in the initial space

Probability model

- Gaussian distribution centered in a 2D manifold

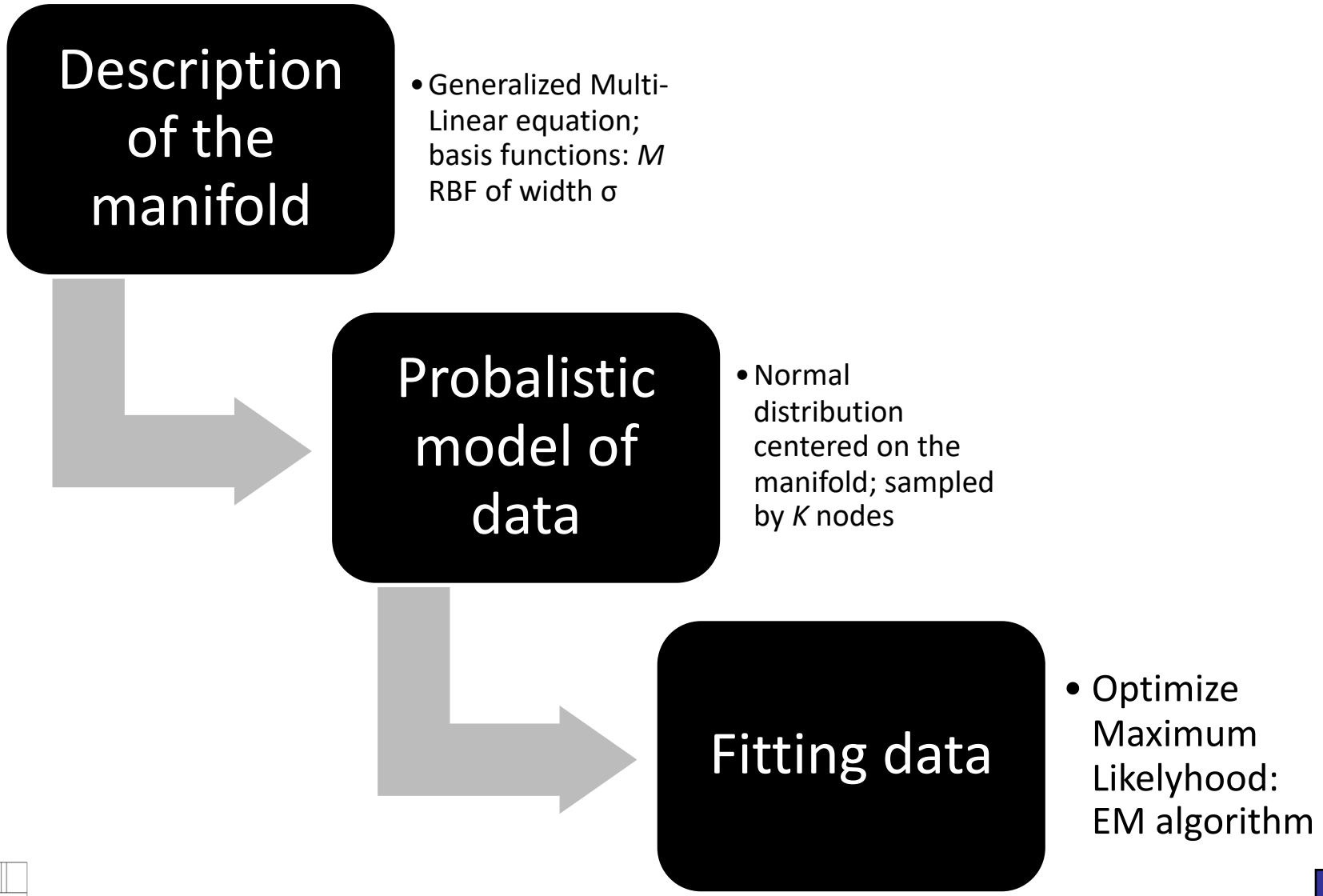
Fitting

- The fitted manifold maximizes the dataset likelihood

Unfolding

- Projected data to the manifold are unfolded appear as 2D map

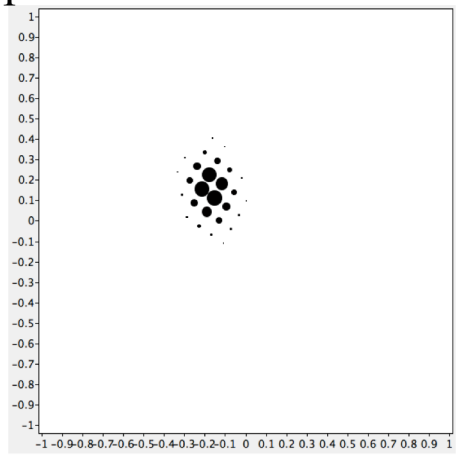
# GTM logic

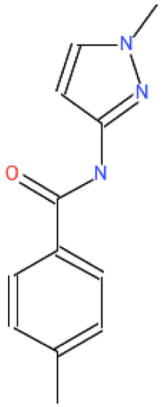

Description of the manifold
• Generalized Multi-Linear equation; basis functions: *M* RBF of width σ
Probalistic model of data
• Normal distribution centered on the manifold; sampled by *K* nodes
Fitting data
• Optimize Maximum Likelyhood: EM algorithm

# GTM responsibilities

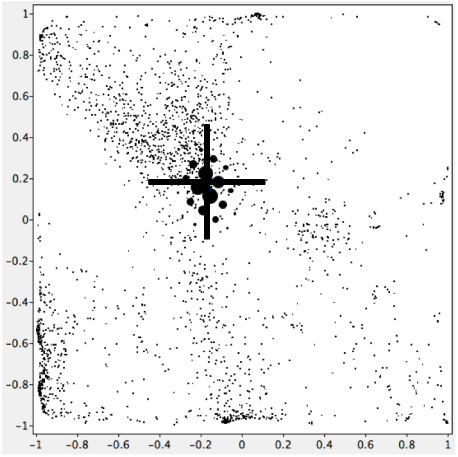
- **Responsibility:**
  - ✔ the probability that a node generated a data point.

A molecule appears on the manifold either as a responsibility pattern…

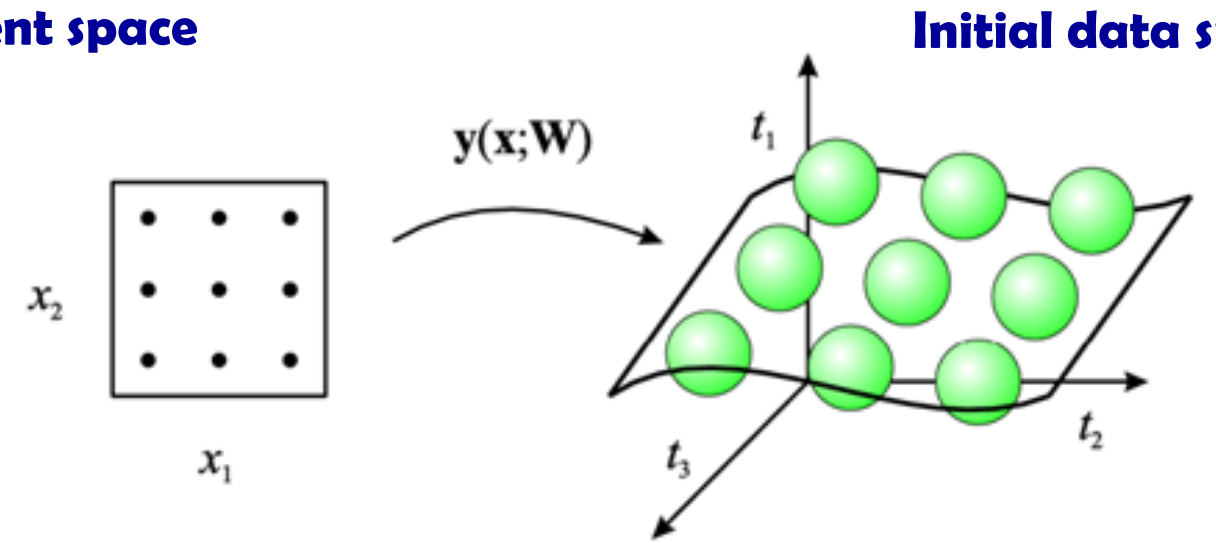





… or as an average coordinate on the manifold.



# GTM outputs




**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**

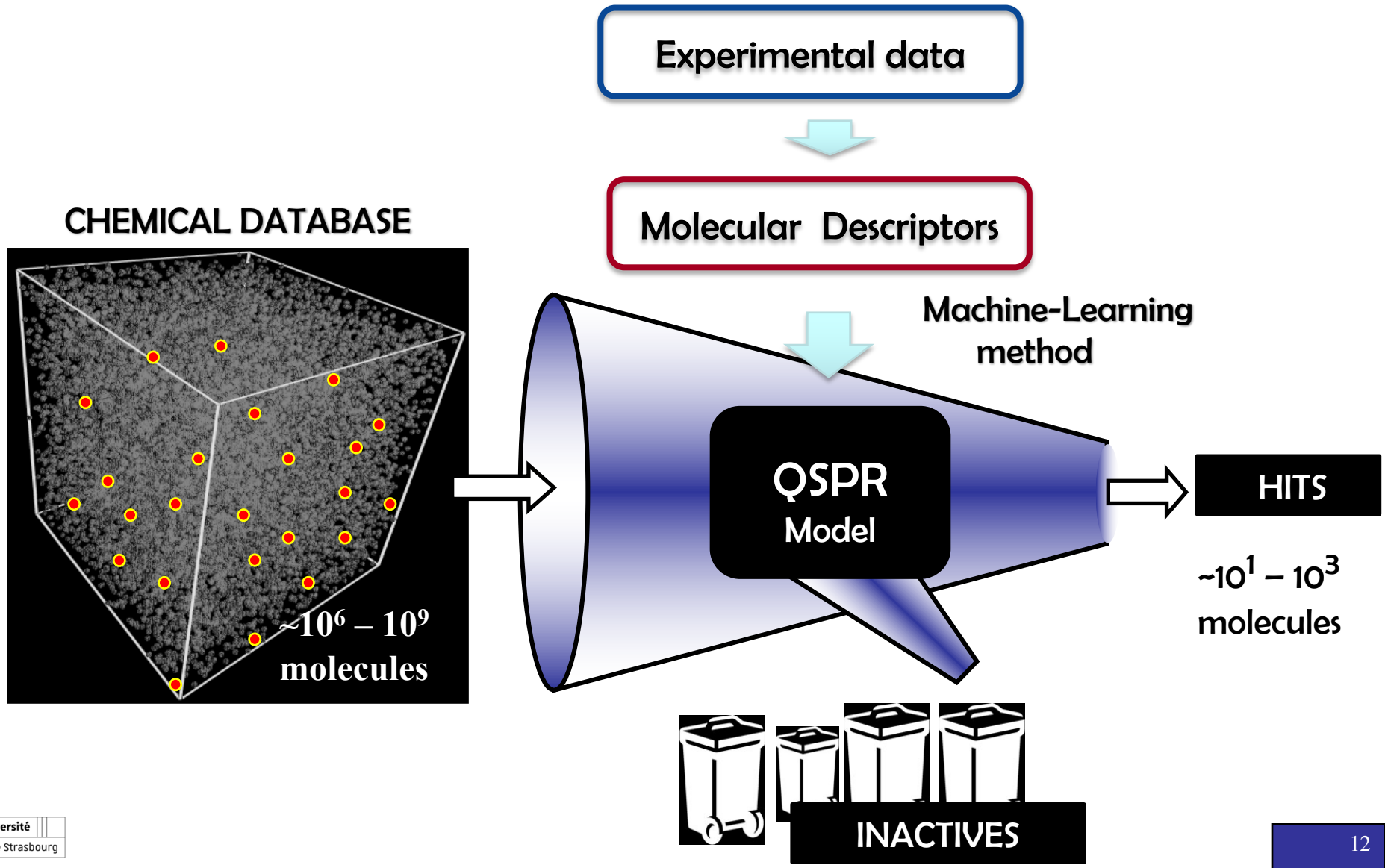
Solvent design embedding GTM

# QSPR AND GTM STUDY OF ELETROLYTIC SOLVENT INGREDIENT

# Objectives of electrolytic solvent design

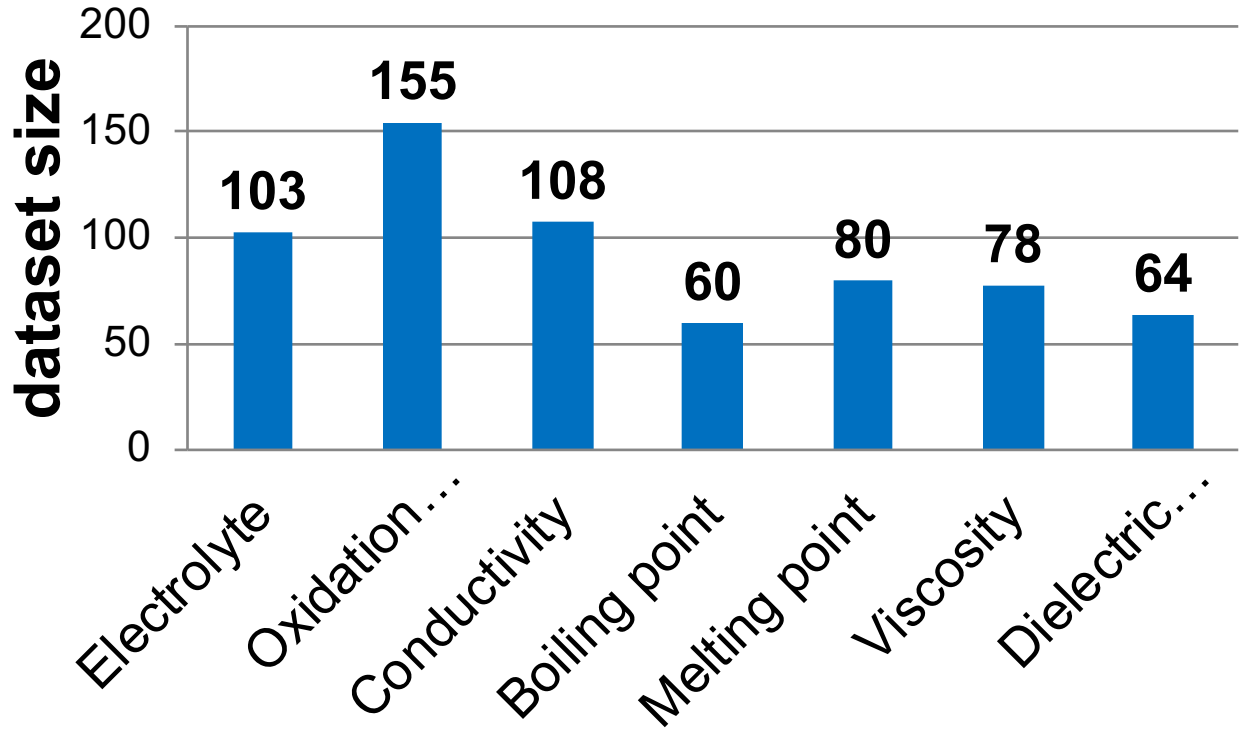
- Optimal properties for liquid electrolyte:
  - High Oxidation potential (> 4 V)
  - Low melting point (< 20°C)
  - High boiling point (> 50°C)
  - Low viscosity (< 3 mPa.s)
  - High conductivity (few mS/cm)
  - High dielectric constant (> 20)

# In silico design of new electrolyte ingredients



# Dataset

- **Data from partners at Chimie ParisTech**




- **Modeling of 6 experimental properties**

# Consensus Models Performance

| Property | RMSE 10-fold CV | RMSE on new structures | Largest error |
|---|---|---|---|
| Oxidation Potential | 0.45 | 0.18 | 0.4 |
| Conductivity ($mS.cm^{-1}$) | 2.13 | 0.49 | 1.2 |
| Boiling point (°C) | 37.0 | 14.0 | 26.0 |
| Melting point (°C) | 32.5 | 45.0 | 176.0 |
| Log Viscosity (Log mPa.s) | 0.13 | 0.04 | 0.06 |
| Dielectric Constant | 20.96 | Not determined | Not determined |

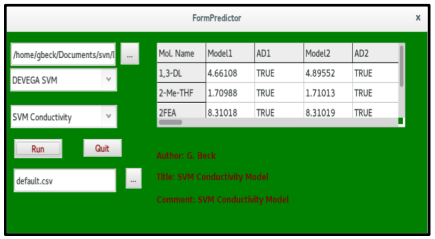
# Virtual Screening

Virtual Chemical Library Enumeration

- 8070 Chemical Structures

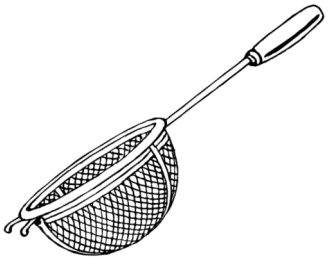
Physical Properties Estimations

- Oxidation Potential
- Melting Point
- Boiling Point
- Viscosity
- Conductivity
- Dielectric Constant




Filters

- **Oxidation potential > 4 V**
- **Boiling Point> 50 ° C**
- **Melting Point< 20 ° C**
- **3832 molecules remaining**

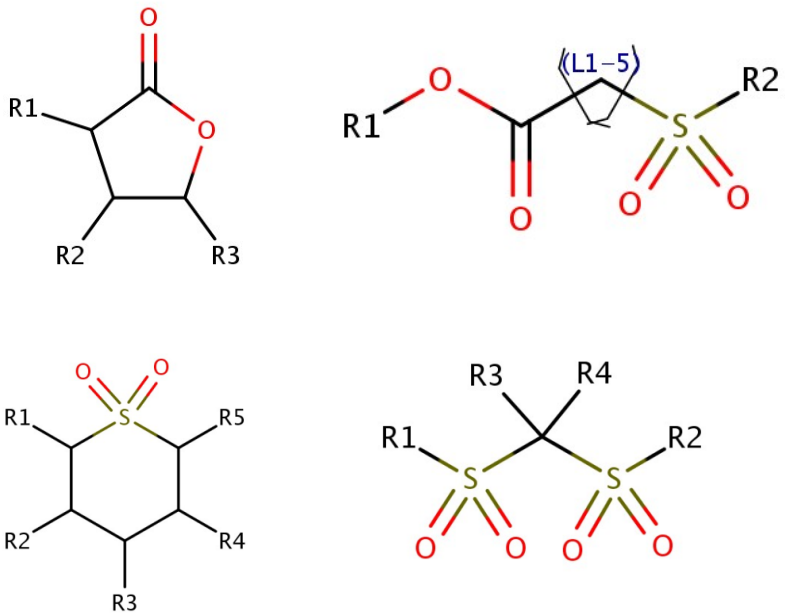


Ranking

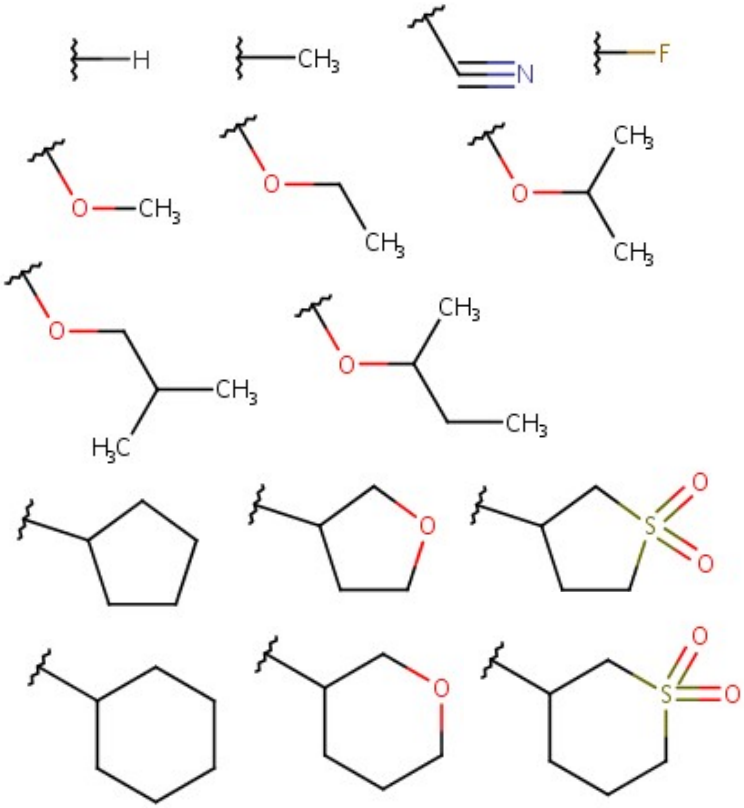
- Missing property values are replaced by an average
- Scoring based on rank product
- **Top 20 best candidates**

# Virtual Chemical Library

- Enumeration of a virtual dataset of sulfones and esters
- Up to 2 subtitutions

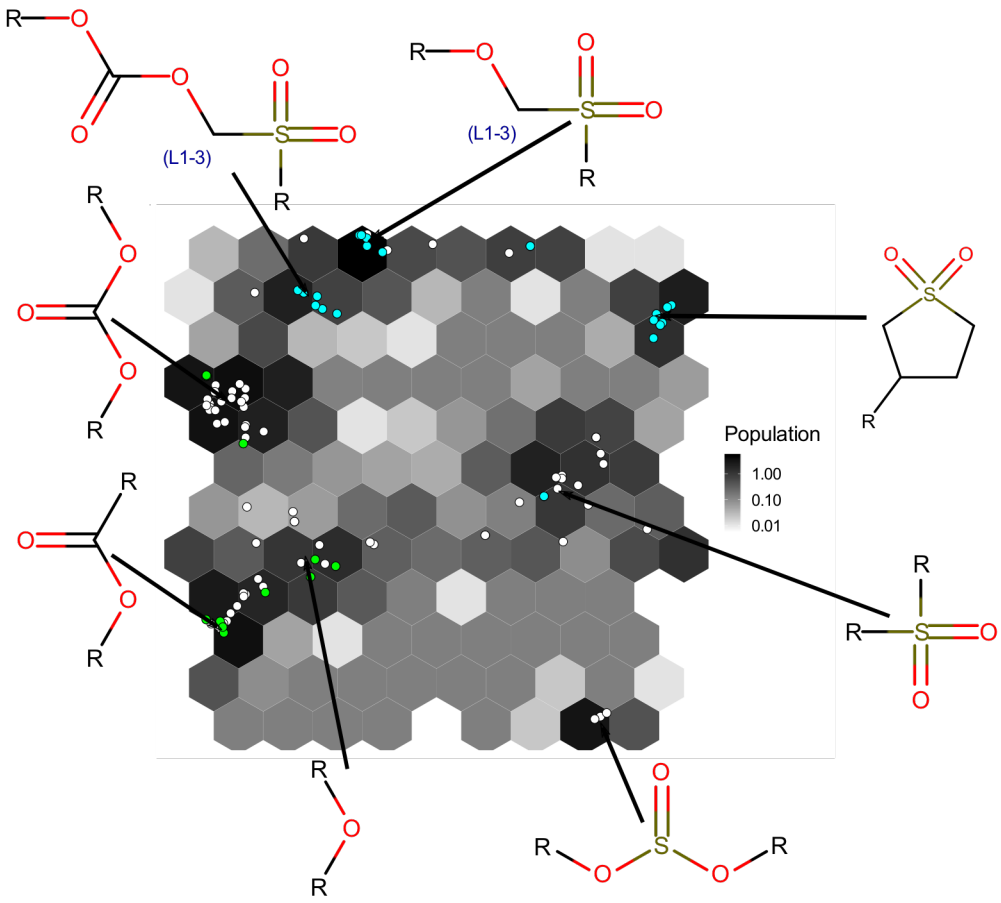



Example of Scaffolds




R-groups

# Map of the Electrolyte Chemical Space




High populated region

Low populated region

Initial data

New esters

New sulfones

# Property maps: monitoring the Chemical Space exploration

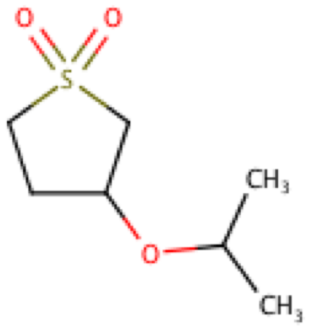
Oxidation Potential

Initial dataset

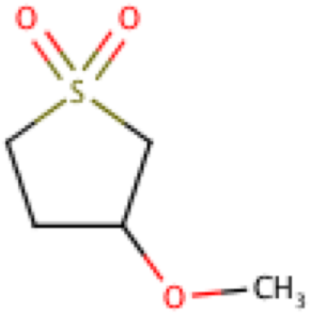
Ionic Conductivity

With new esters and sulfones

# Best sulfone candidates




Melting Point (°C): < -70
Boiling Point (°C): 299.5
Viscosity (mPa.s): 15.82
Ionic Conductivity ($mS.s^{-1}$): 0.7
Oxydation Potential (V): 5.30




Melting Point (°C): < -70
Boiling Point (°C): 281
Viscosity (mPa.s): Not determined
Ionic Conductivity ($mS.s^{-1}$): 1.1
Oxydation Potential (V): 5.10

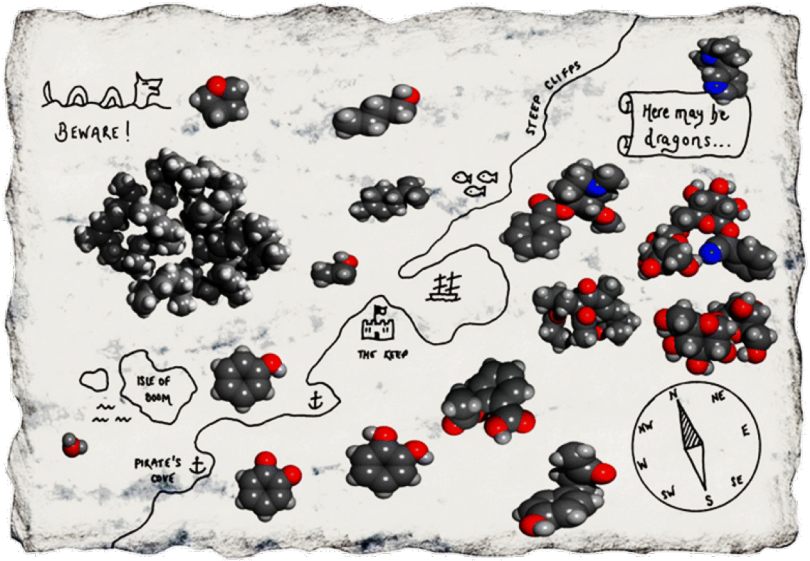
# Summary

- **QSPR Modeling**
  - ✔ 6 models
    - **Oxidation potential, conductivity, boiling point, melting point, viscosity dielectric constant**
  - ✔ Implemented on DEVEGA Predictor software
- **Virtual Screening**
  - ✔ Virtual library of 8070 esters and sulfones
  - ✔ Identification of two new sulfones ongoing formulation
- **Electrolytes data visualization**
  - ✔ 2D map, 6 different activity landscapes
  - ✔ Identification of promising areas in the chemical space

GTM and profiling applications

# TOWARD « UNIVERSAL » MAP OF CHEMICAL SPACE

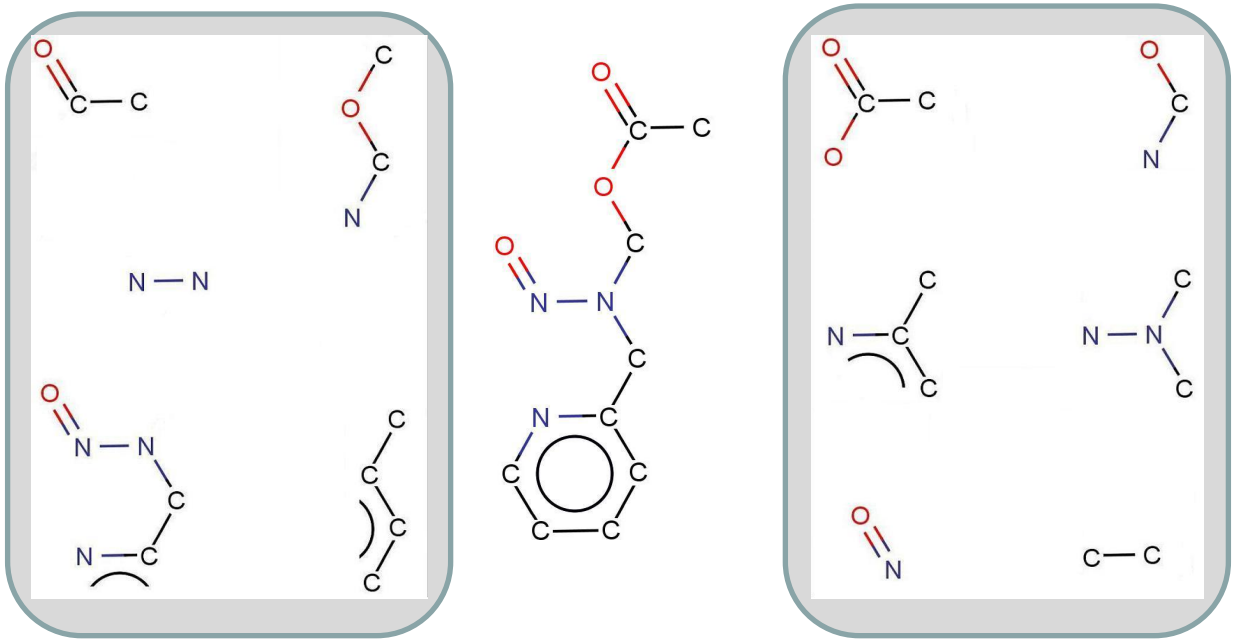
# What do we expect from an "universal" map of the Chemical Space?




**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;**

# ISIDA fragment descriptors

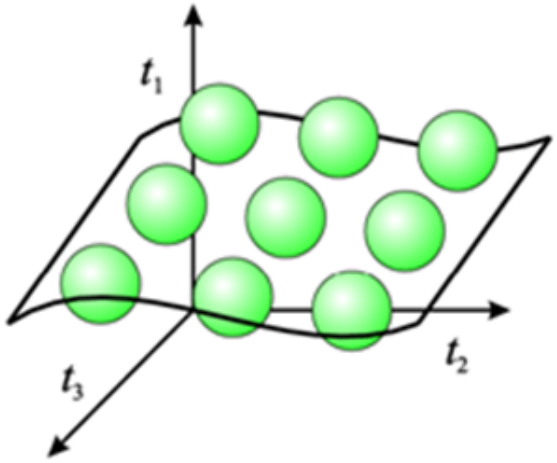


**Several hundreds types of fragment descriptors can be generated for one same data set**

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

# Toward universal map(s) of chemical space

**Selected initial space ($Descr_i$)**

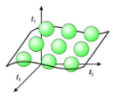
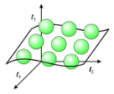
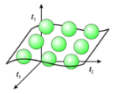



| Activity | Model's performance |
|---|---|
| $Act_1$ | $BA_1$ |
| $Act_2$ | $BA_2$ |
| ..... | |
| SCORE | <BA> |

**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

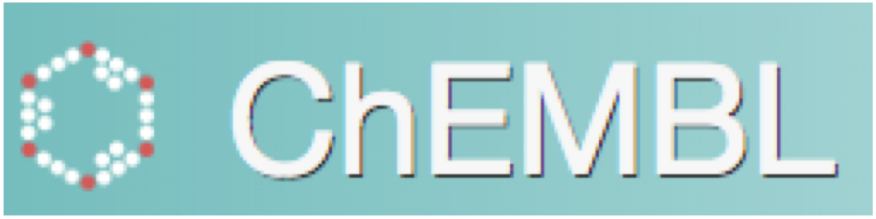
# Toward universal map(s) of chemical space

| 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

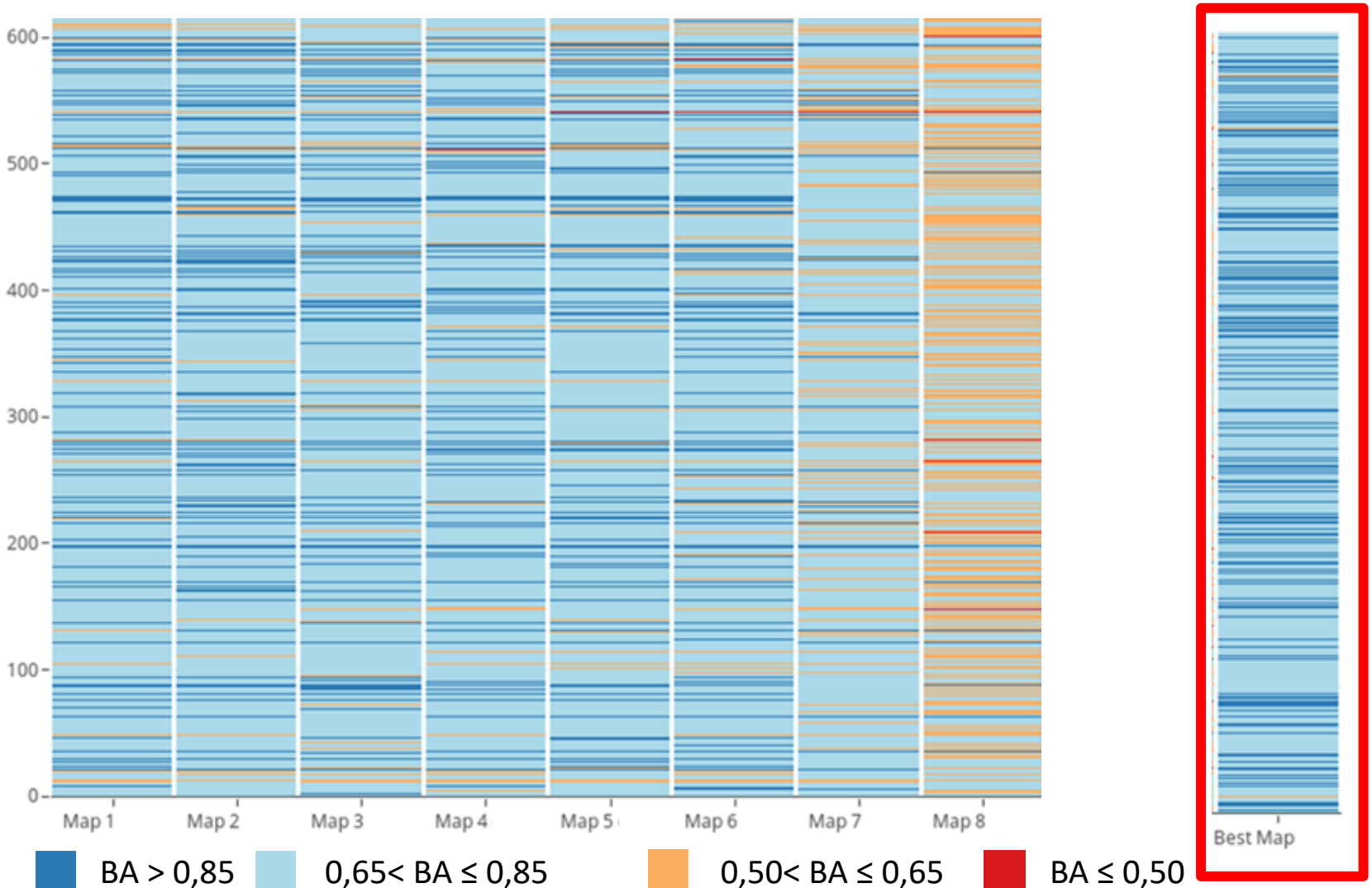



> 1.7 M structures
11538 targets

- **8 maps corresponding to different types of descriptors have been constructed,**

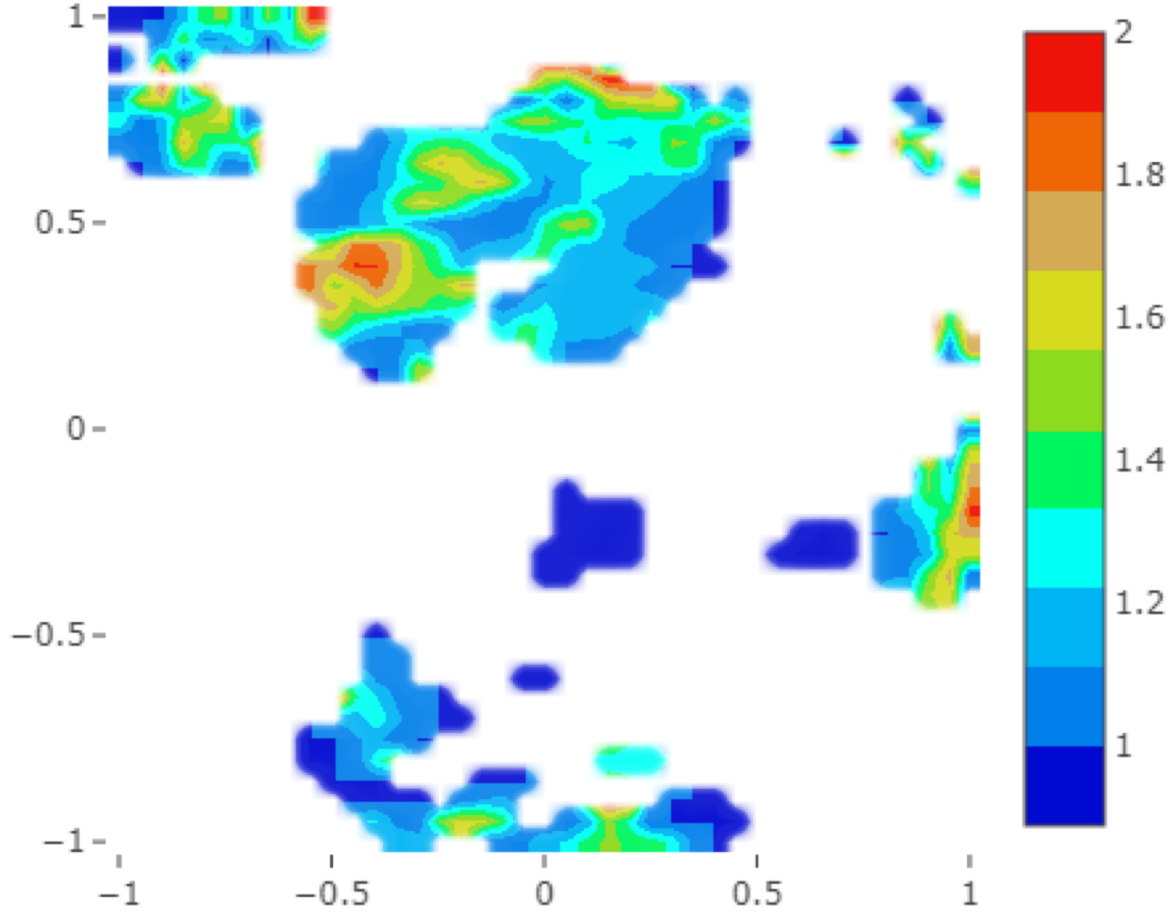
- **they correctly predict classes (active / inactive) for the ligands against > 600 targets and for various phenotypic activities (antiviral, antimalarial, ...)**

# Perfomance of universal maps to classify actives/inactives for 618 ChEMBL targets




Université de Strasbourg

# Case study: virtual screening of DUD-E using activity landscape for *Adenozine receptor A2a*

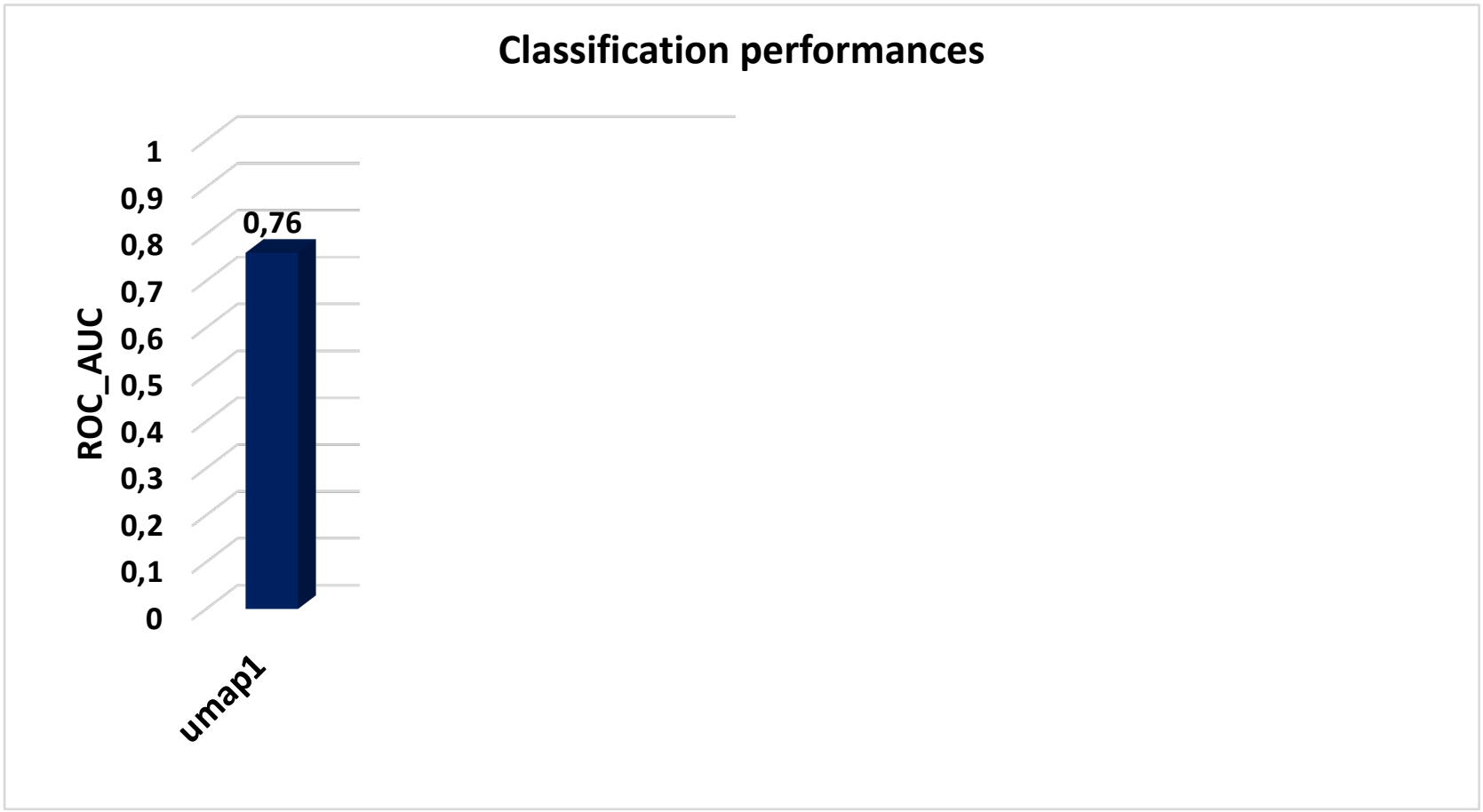



***Active***

***Inactive***

**ChEMBL data:**
**1303 actives**
**3618 inactives:**

# Virtual screening of A2a inhibitors



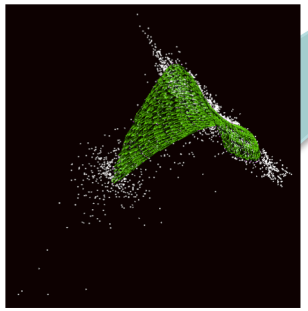
GTM-based information on QSAR models

# INTERPRETATION OF CLASSIFICATION MODELS

# Workflow

GTM model of the data

- Extract coordinates of each node

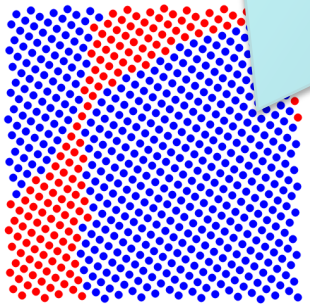


Building QSAR model on the data

- SVM, RF

Apply QSAR model on nodes

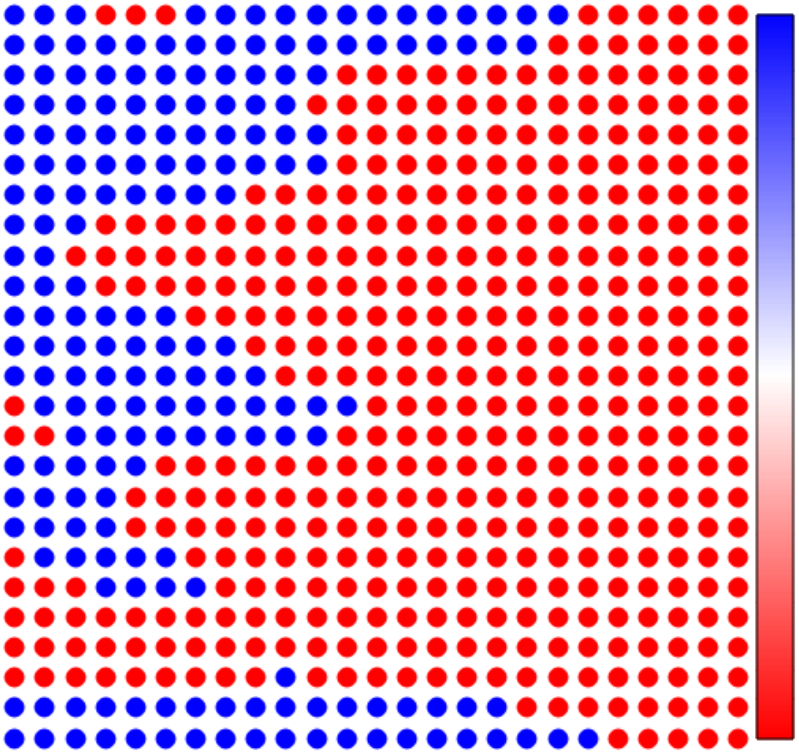
- Assign predictions to nodes



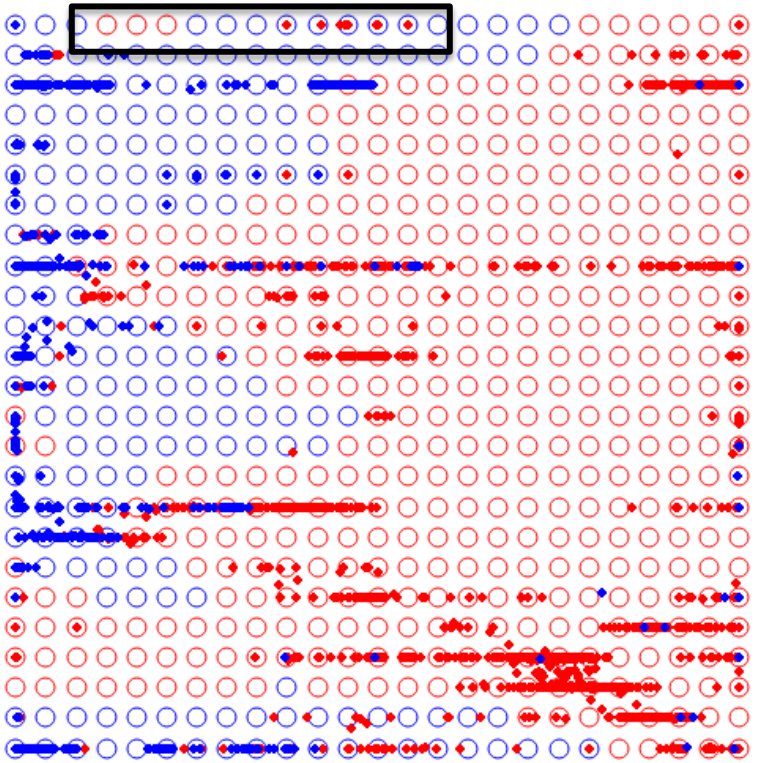
Visualize decision boundary

- Color nodes according to QSAR predictions

# A2A Linear SVM model




A2A C-SVM model, linear kernel
Decision boundary projected on the GTM manifold

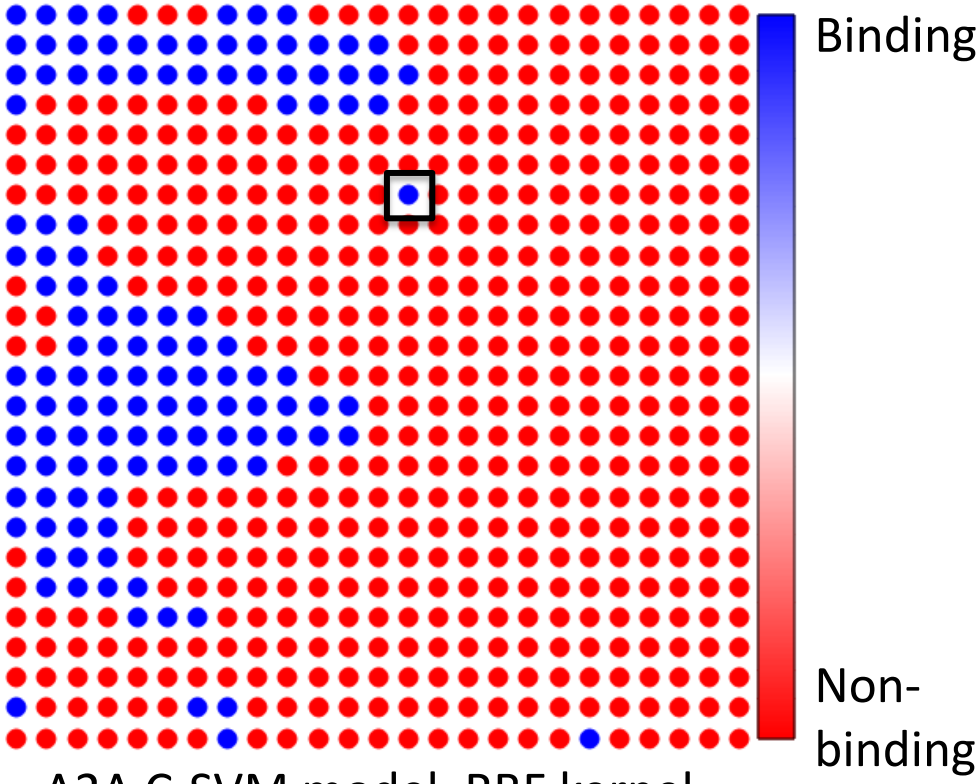


Projection of new molecules on the manifold

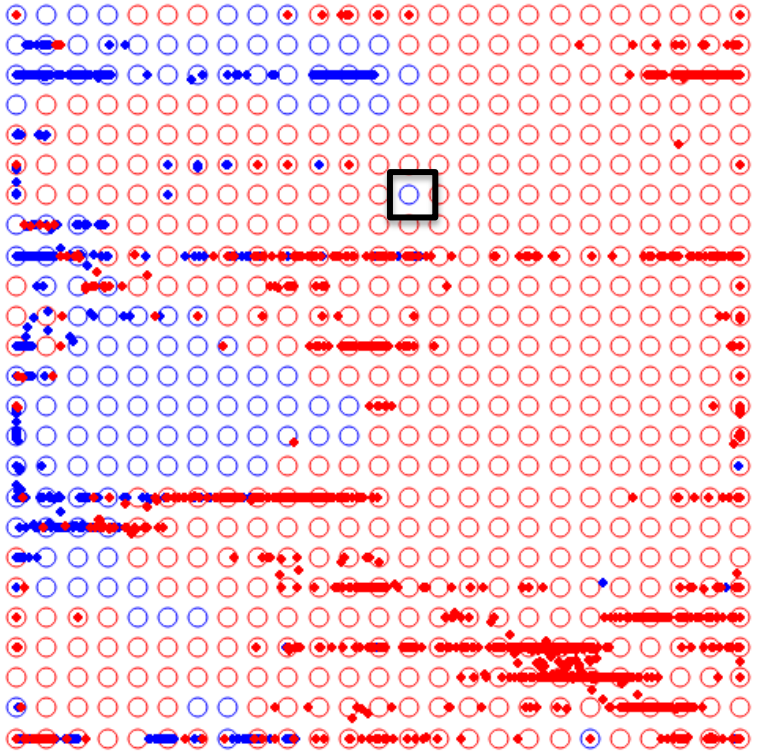
It reveals inconsistencies:

- Class inconsistencies
- Decision supported by no molecules

# A2A RBF SVM model




A2A C-SVM model, RBF kernel
Decision boundary projected on the GTM manifold



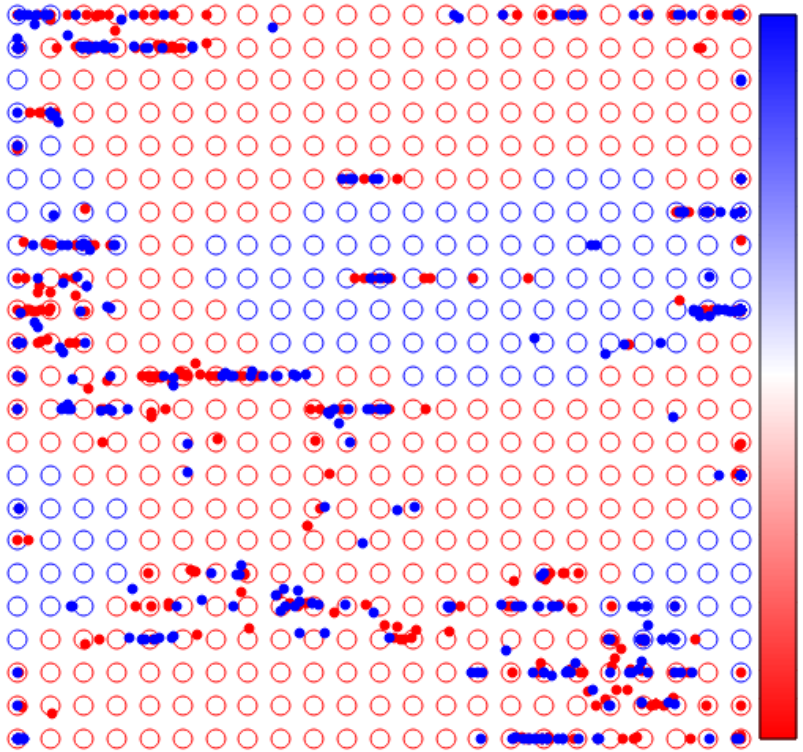
Projection of new molecules on the manifold

It reveals some singular decisions boundaries

# Support Vector

- **SVM models are based on similarities to a limited set of instances, the Support Vectors**
  - ✔ Support Vectors can be located on the GTM
- **There are two kinds of Support Vectors**
  - ✔ Bound Support Vectors:
    - **Support Vectors actually misclassified**
    - **Support Vectors inside the margin**
  - ✔ Unbound Support Vectors:
    - **Support Vectors that are located on the margin**

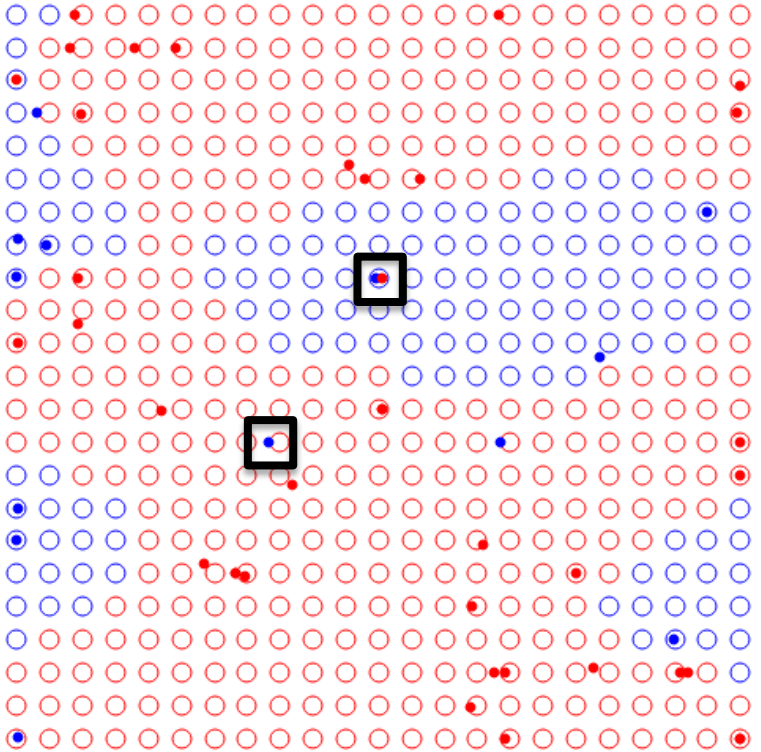
# hERG: Support Vectors
# Linear Kernel



Blocker

Non-blocker

Bounded Support Vectors

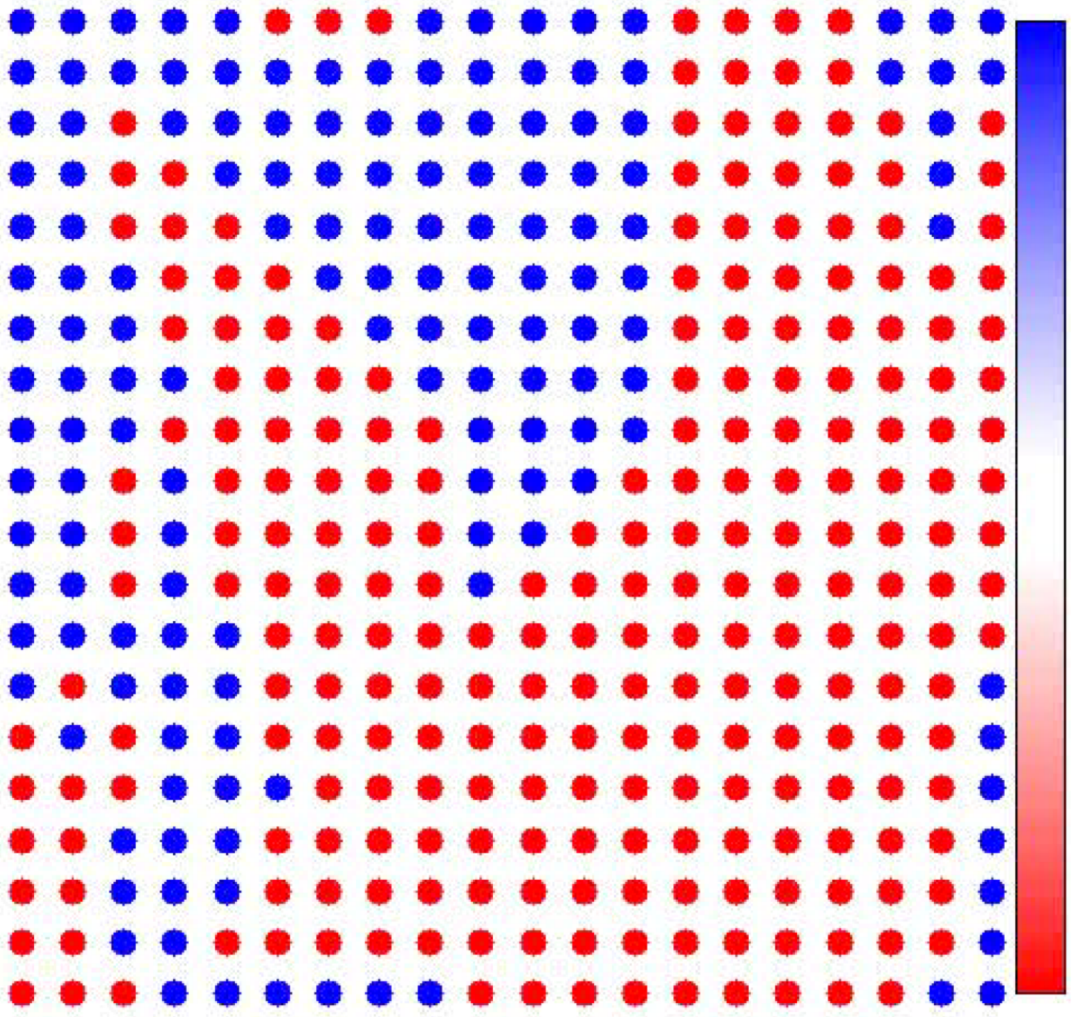


Unbounded Support Vectors

Bonded Support Vectors can be anywhere

Parts of the decision boundary can be away of the manifold as illustrated by some support vectors.

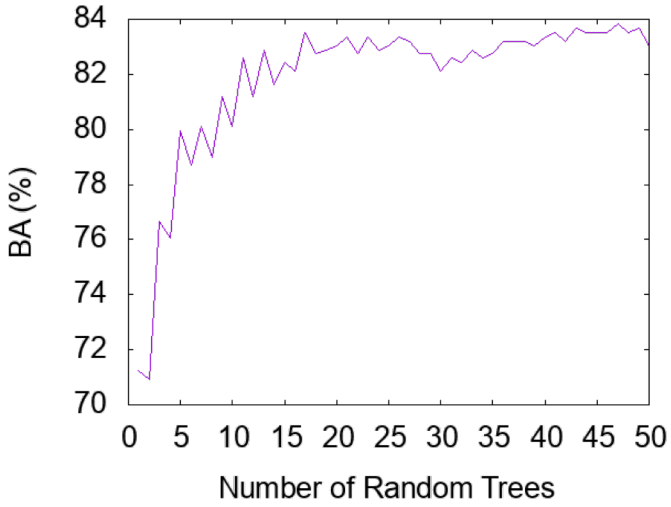
# THC: Random Forest



Decision boundary of Random Forest

Number of Random Trees: 1 to 50

BA (5-CV) from 0.71 to 0.84



# Summary

- **GTM allows to visualize decision boundaries of classifiers**
  - ✔ General agreement between classes and the boundary
  - ✔ Visualize the score associated to a decision: probability, confidence, etc
- **Visualization allows better understanding of intricate concepts**
  - ✔ Differences between linear / non-linear models
  - ✔ Support Vectors
  - ✔ Model changes with meta-parameters (SVM C or Γ; RF number of trees)
- **Visualization illustrates uncertain decision**
  - ✔ Applicability domain

# Conclusion

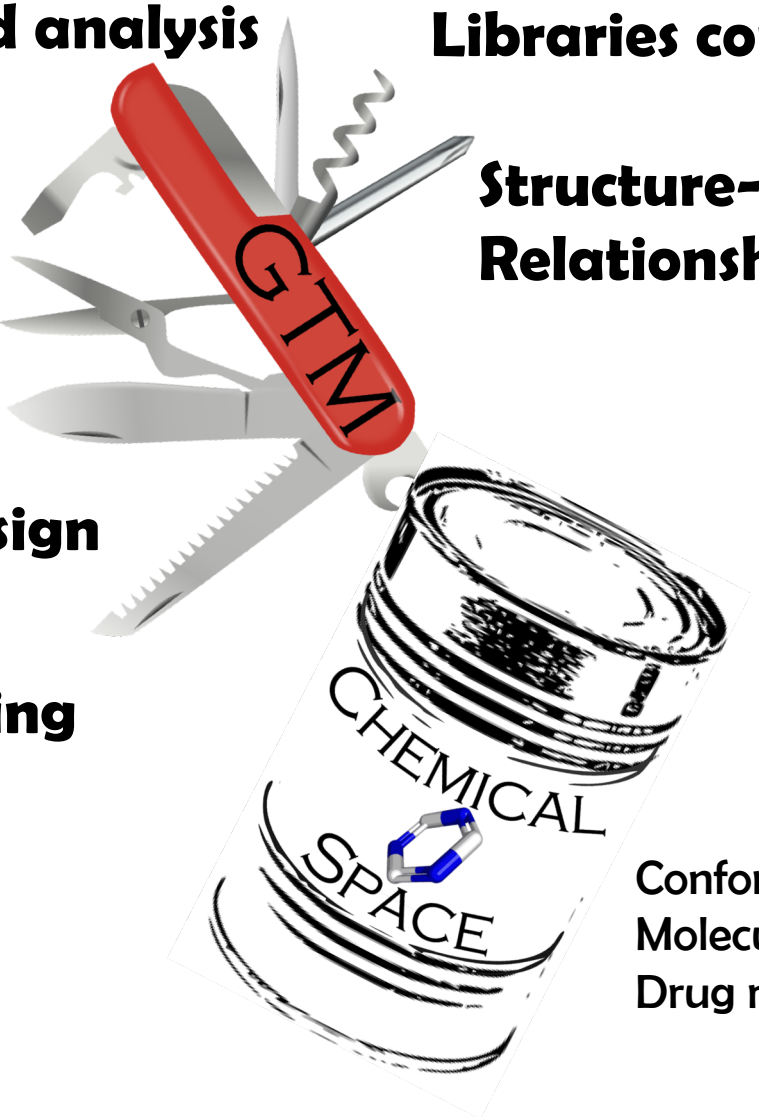
**Data visualisation and analysis**

**Libraries comparison**

**Virtual screening**

**Structure-Activity Relationship**

**Library design**




***de novo* design**

**Profiling**

Conformational space analysis
Molecular surfaces modelling
Drug repurposing

# Software

**ISIDA Software,**
**including Molecular Descriptors and**
**Generative Topographic Mapping,**
**are available for free:**

**http://infochim.u-strasbg.fr/spip.php?rubrique41**

**Tutorial on GTM is available here:**

**http://infochim.u-strasbg.fr/spip.php?rubrique246**

Chém Informatique
STRASBOURG

# Community

## Société Française de Chémoinformatique (SFCi)

Adhesion: 20 euros
Benefits:

- Workshop:
  **Journée Nationale de la Chémoinformatique**
- Directory including both academic and industrials
- PhD price
- Grants for participating to conference
- Grants for traveling

## New GDR currently in creation:
## **Big Data en Chimie**

Adhesion: academic teams only
Benefits:

- Workshops
- Coordination on research
- Grant for traveling between groups