BigChem

Big data visualization and modelling using Generative Topographic Mapping (GTM)

Arkadii Lin

Supervisors: Prof. Alexandre Varnek

Dr. Bernd Beck

UNIVERSITÉ DE STRASBOURG

Boehringer Ingelheim

Outline

1. **Background & Education**
2. **Introduction to GTM approach**
3. **Software development**
4. **Chemical libraries comparison**
5. **Conclusions** and **Plans for the next 6 months**

1

Background & Education

Boehringer Ingelheim

Background & Education

Arkadii LIN

PhD student at the **University of Strasbourg**, now at **Boehringer Ingelheim Co.**

Age: 25 **Nationality:** Russian

Specialty: Chemoinformatics **Master:** Kazan Federal University, Russia (2015)

- 60 hours of University courses;
- 19 on-line lectures given by the partners of BigChem project;
- First BigChem School “Introduction to Chemoinformatics”, Munich, October 2016; *Second BigChem School "Chemical databases“ (Barcelona, April 2017);*
- 3rd Kazan Summer School on Chemoinformatics (Kazan, Russia, July 2017);
- 8th meeting of Chemoinformatics Society in France SFCi2017 (Orleans, France, October 2017).

2

Introduction to GTM

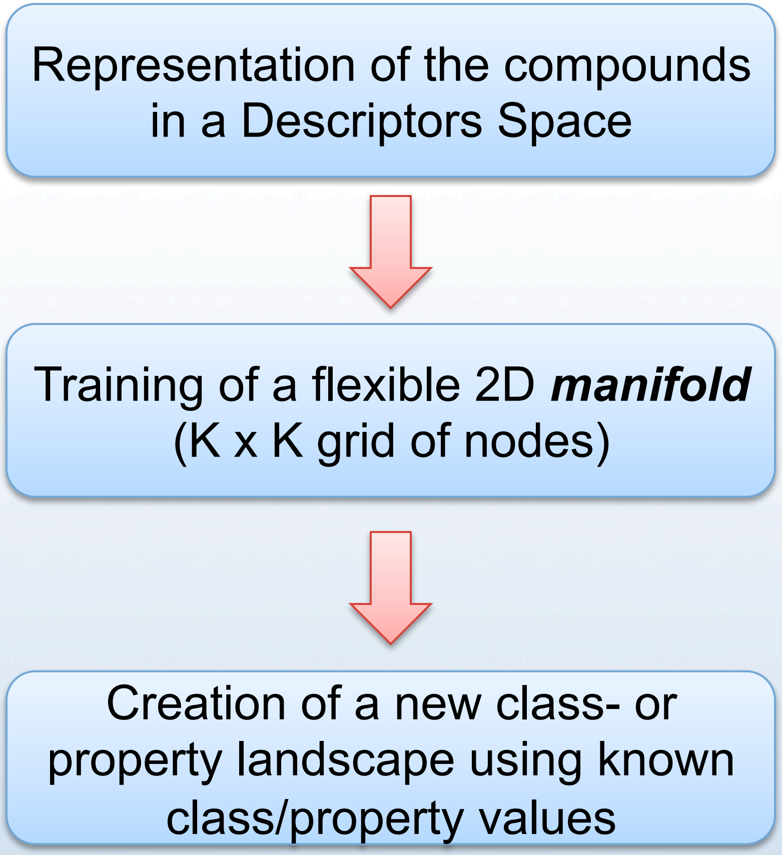
Generative Topographic Mapping (GTM) approach

Chemical space
(estimated size is 10^{63})

GTM already today allows to visualize and to analyze millions of compounds, projecting it onto 2D latent space.

Generative Topographic Mapping (GTM) approach

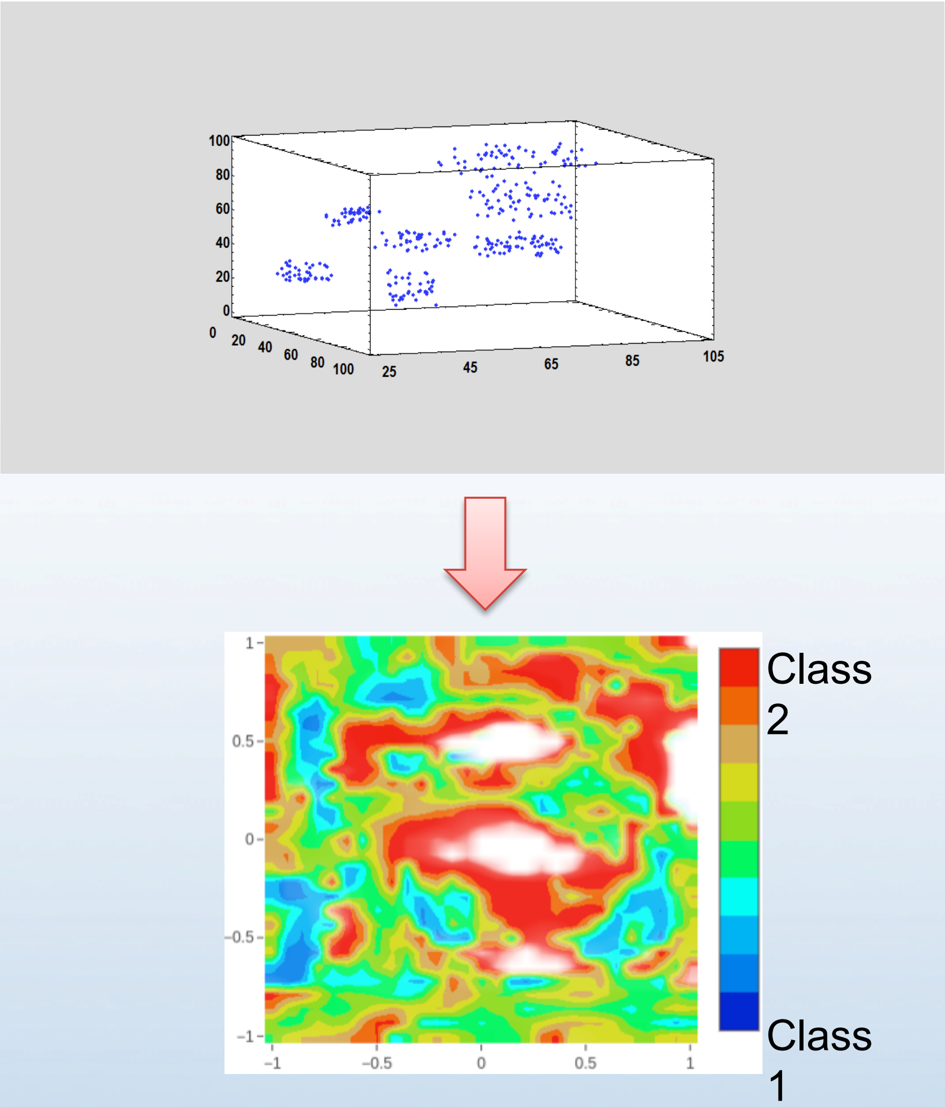
General workflow



Representation of the compounds in a Descriptors Space

Training of a flexible 2D ***manifold*** (K x K grid of nodes)

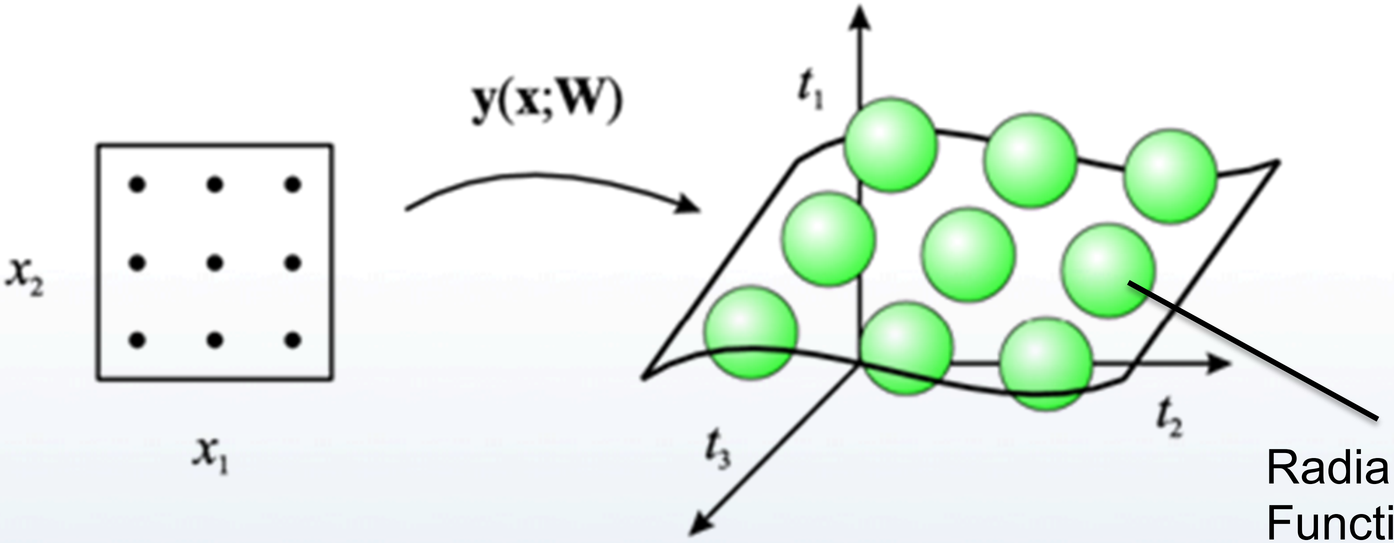
Creation of a new class- or property landscape using known class/property values



GTM Concept

2D Latent space

Initial data space



Radial Basis Function (**Gaussian function**)

GTM generates a data probability distribution in both initial and latent data spaces. As a result, each compound is associated to each node with it’s own probability.

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

3

Software development

Boehringer Ingelheim

Existing GTM Software Tools

ISIDA Fragmentor2017 – a tool for ISIDA Fragment descriptors generation.

GTMapTool2016 – a tool for GTM manifold training and new compounds projection.

GAConfig – a tool with the implemented Genetic Algorithm for the best descriptors and GTM meta-parameters selection.

All these tools are developed in the Laboratory of Chemoinformatics, University of Strasbourg.

GTM: Software Development

GTM classification models creation:

- **GTMClass tool** – creates a classification model using obtained GTM responsibilities of a training set to predict a class for a new compound.

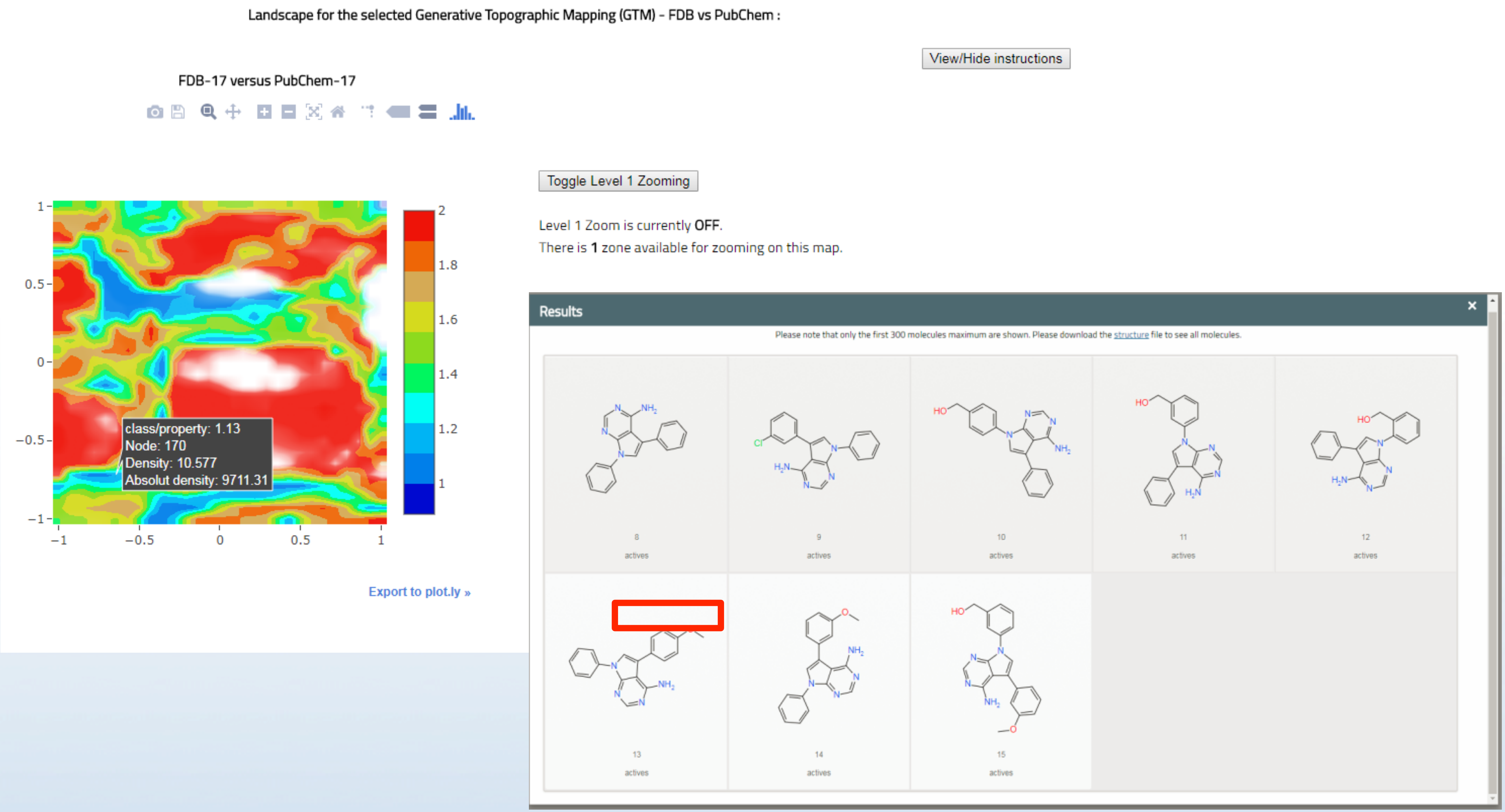
GTM regression models creation:

- **GTMReg tool** – creates a regression model using obtained GTM responsibilities of a training set to predict a property for a new compound.

GTM maps visualization:

- **GTMVis tool** (desktop)– creates an HTML file with an interactive GTM map. With this tool the user is able to explore his map interactively;
- **Online GTM** – allows the user to explore the already created GTM maps online.

Online GTM



Landscape for the selected Generative Topographic Mapping (GTM) - FDB vs PubChem :

FDB-17 versus PubChem-17

Level 1 Zoom is currently **OFF**.

There is **1** zone available for zooming on this map.

class/property: 1.13
Node: 170
Density: 10.577
Absolut density: 9711.31

Results

Please note that only the first 300 molecules maximum are shown. Please download the structure file to see all molecules.

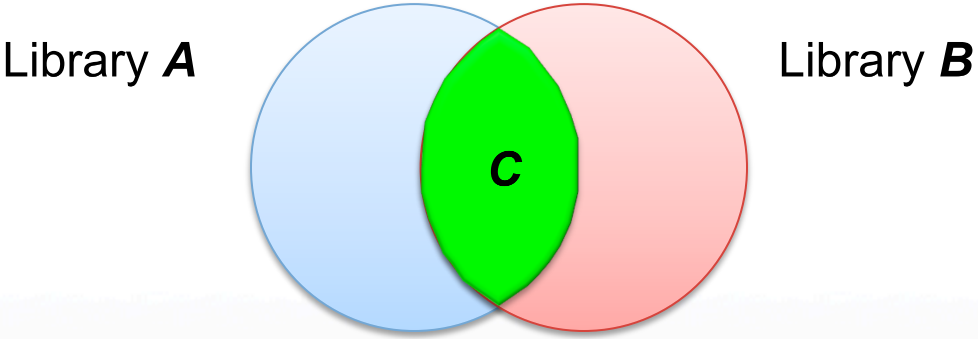
http://infochim.u-strasbg.fr/onlineGTM

4

Chemical Libraries Comparison

Boehringer Ingelheim

Chemical Libraries Comparison



- Does the library ***A*** contain any unique chemotypes?
- Is there any new chemotypes in the library ***B***?
- Can we increase the diversity of the library ***A*** in terms of poorly presented in ***A*** chemotypes (scaffolds) using the library ***B***?

Initial Data

ChEMBL-17 ≈ 102 000 mol.

PubChem-17 ≈ 11 000 000 mol.

FDB-17 ≈ 10 000 000 mol.

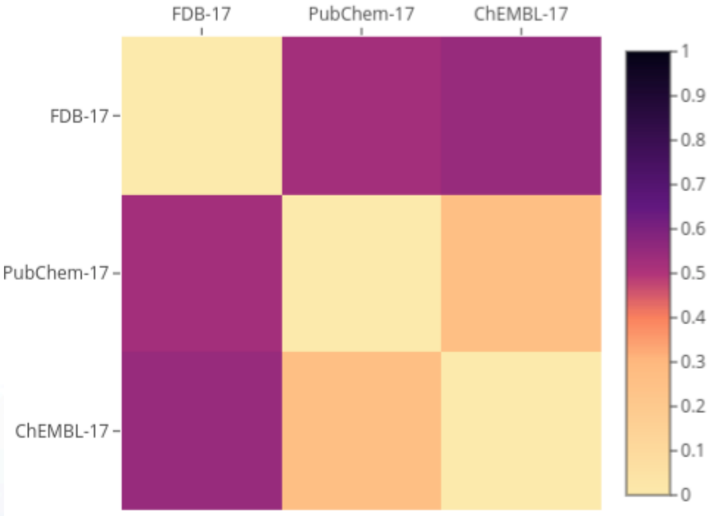
ChEMBL-17 + **PubChem-17** + **FDB-17** ≈ 21 100 000 compounds

Each compound contains less than **18** heavy atoms.

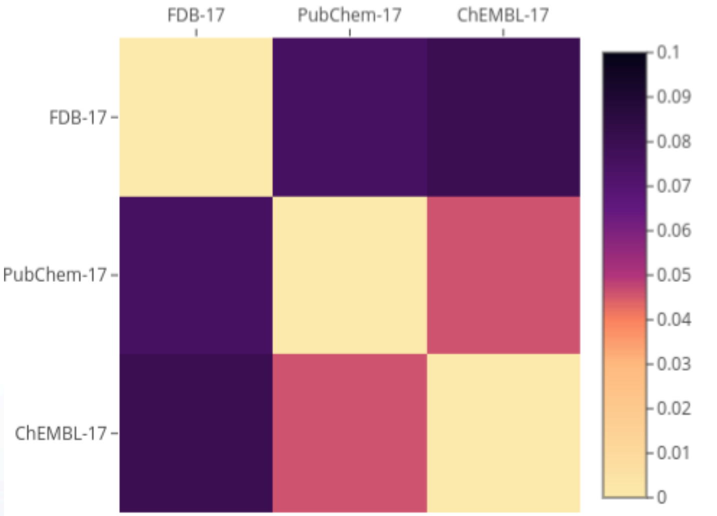
All the structures were standardized following some basic rules, such as aromatization, removing explicit hydrogens, transformation of the common groups (for instance, NO_2). ISIDA Fragment descriptors were used to describe the compounds.

Ricardo Visini, Mahendra Awale, and Jean-Louis Reymond ***Fragment Database FDB-17***, *Journal of Chemical Information and Modeling* **2017** *57* (4), 700-709
DOI: 10.1021/acs.jcim.7b00020

Quantitative Libraries Comparison



Tanimoto distance
(1-Tc)



Euclidean distance

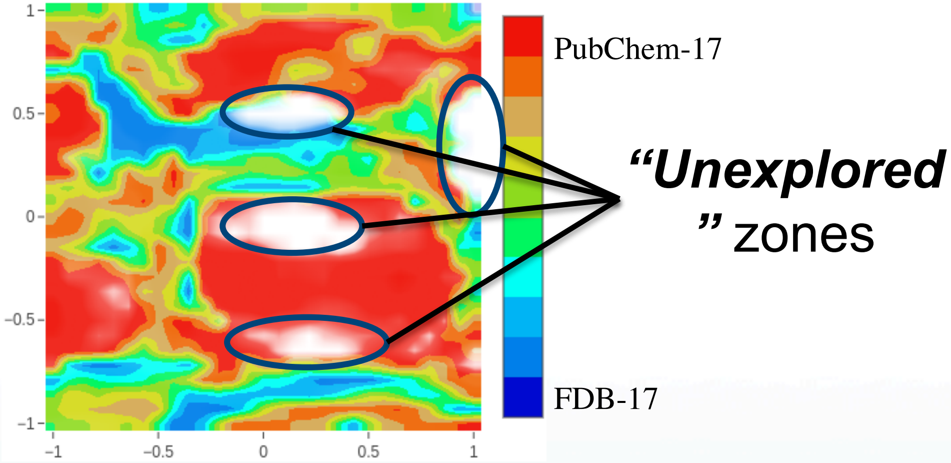
In terms of library size:

ChEMBL-17 << PubChem-17 ~ FDB-17

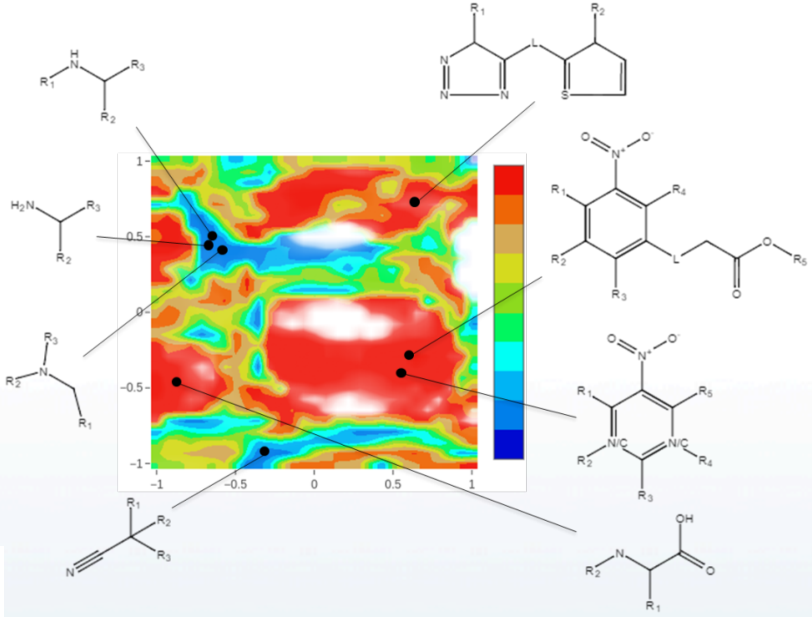
In terms of chemical space coverage:

- ChEMBL-17 ~ PubChem-17
- PubChem-17 ≠ FDB-17
- ChEMBL-17 ≠ FDB-17

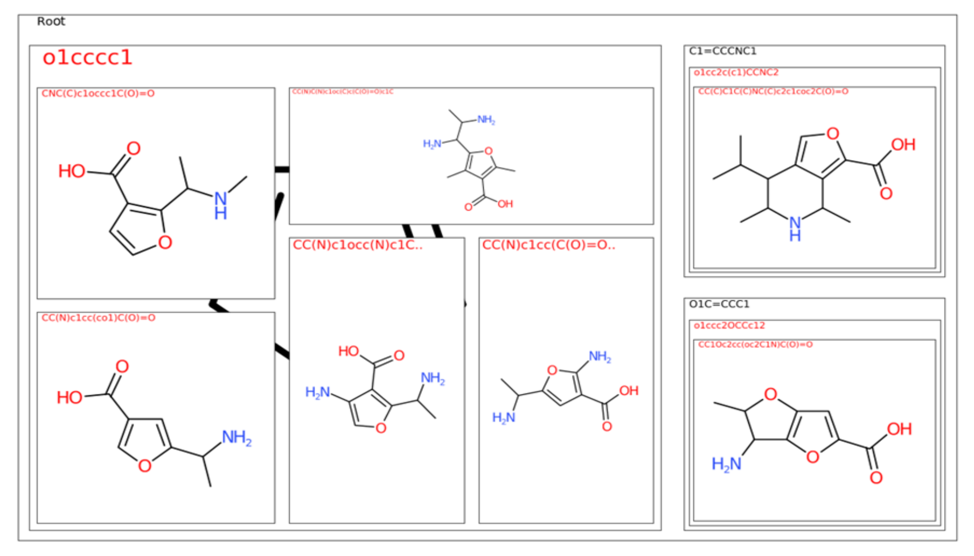
Libraries Comparison using GTM maps



“Unexplored” zones

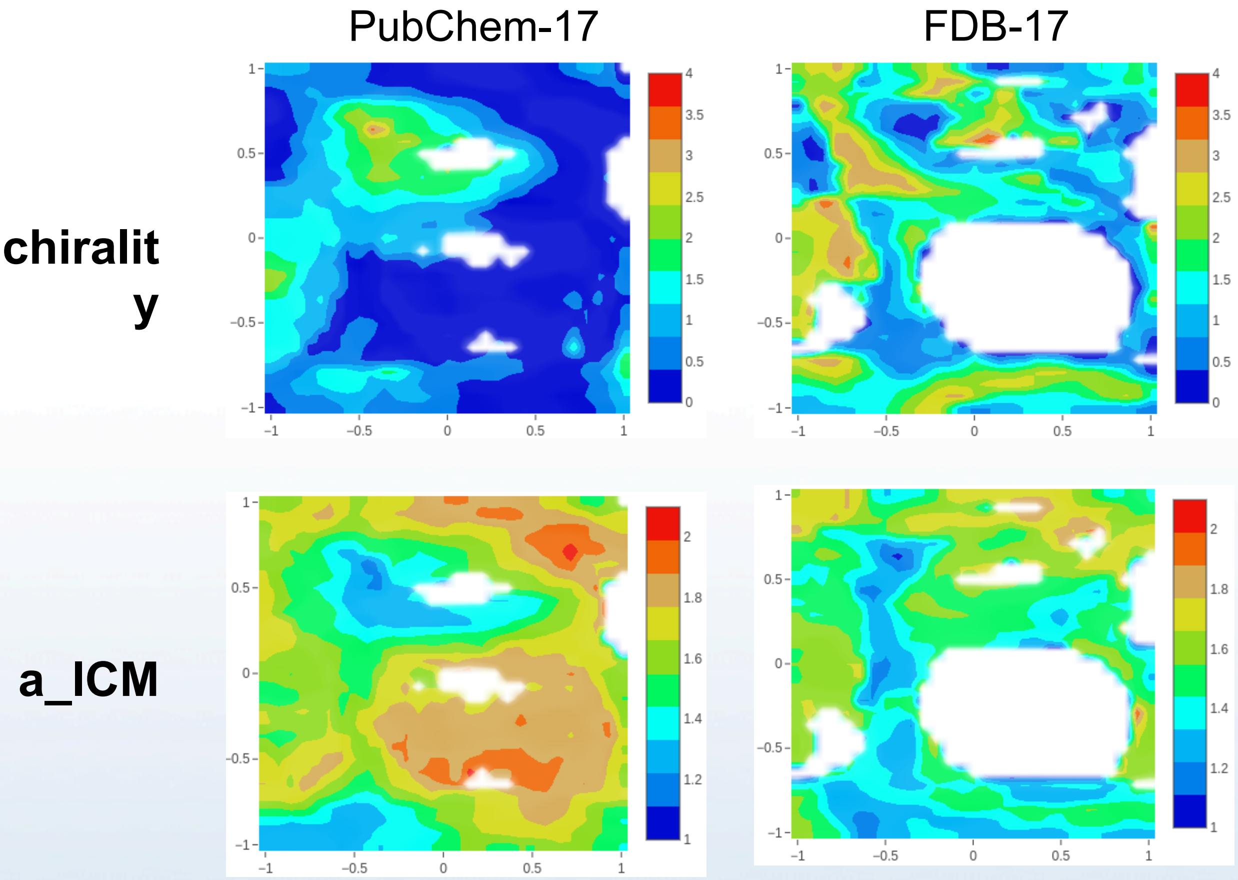


FDB-17 **PubChem-17**



Deep „*mining*“ in one of the areas of the ***PubChem-17/FDB-17*** map retrieved some unique structures, which are presented only in FDB-17.

GTM Property Maps



Theoretically generated ***FDB-17*** is significantly richer in chiral compounds than ***ChEMBL-17*** and ***PubChem-17***. However, it has smaller entropy of the element distribution in a molecule, described here by **a_ICM** (MOE).

5

Conclusions and Outlook

Boehringer
Ingelheim

Conclusions

- New software tools for GTM based models creation and visualization were developed.
- An exhaustive map of the fragment-like chemical space represented by 21.1M molecules has been created and visualized.
- Three chemical databases (ChEMBL-17, PubChem-17, FDB-17) were compared using class- and property landscapes.
- Regions with highly imbalanced population of FDB-17 versus PubChem-17 compounds were successfully mined for original structures.

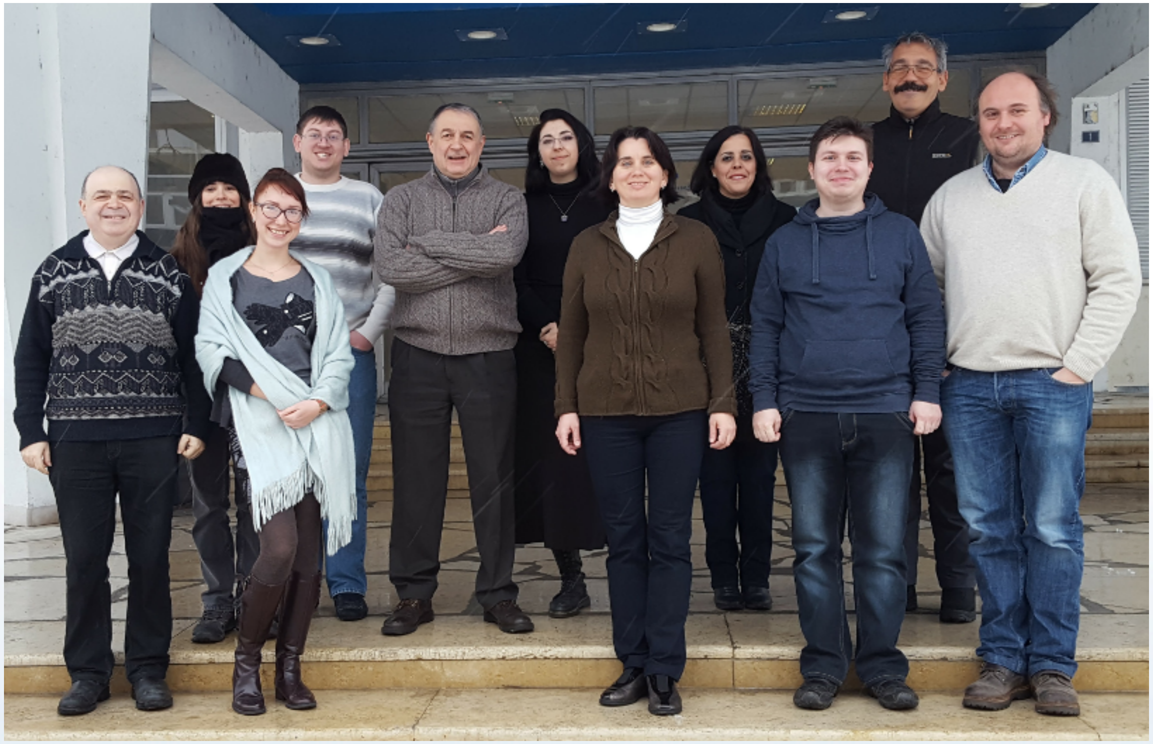
A paper with the results of this project was submitted to *ChemMedChem* journal.

Plans for the next 6 months

There are 2 projects for the next 6 months:

- To try the GTM approach as a tool for Virtual Screening and compare it with Similarity approach.
- To compare the Boehringer Ingelheim compound pool with some public collections.

Thank you!



Prof. Alexandre Varnek
Dr. Gilles Marcou
Dr. Dragos Horvath

Dr. Fanny Bonachera
Dr. Olga Klimchuk

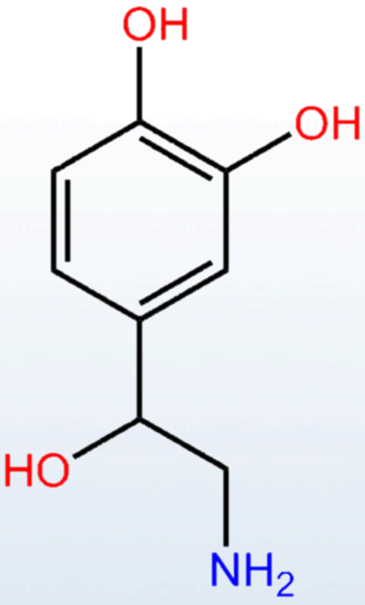
BigChem, Modena, Italy, 2017

Supporting materials

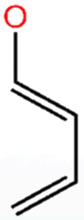
Boehringer Ingelheim

ISIDA Fragment descriptors

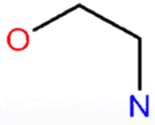
Counts of fragments of different nature in the molecule.



- **Sequence Fragments**

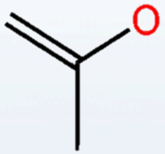


2

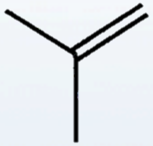


1

- **Atom-centered Fragments**

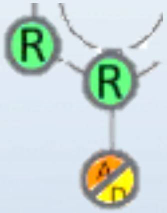


2

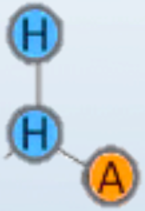


1

- **Property labels**

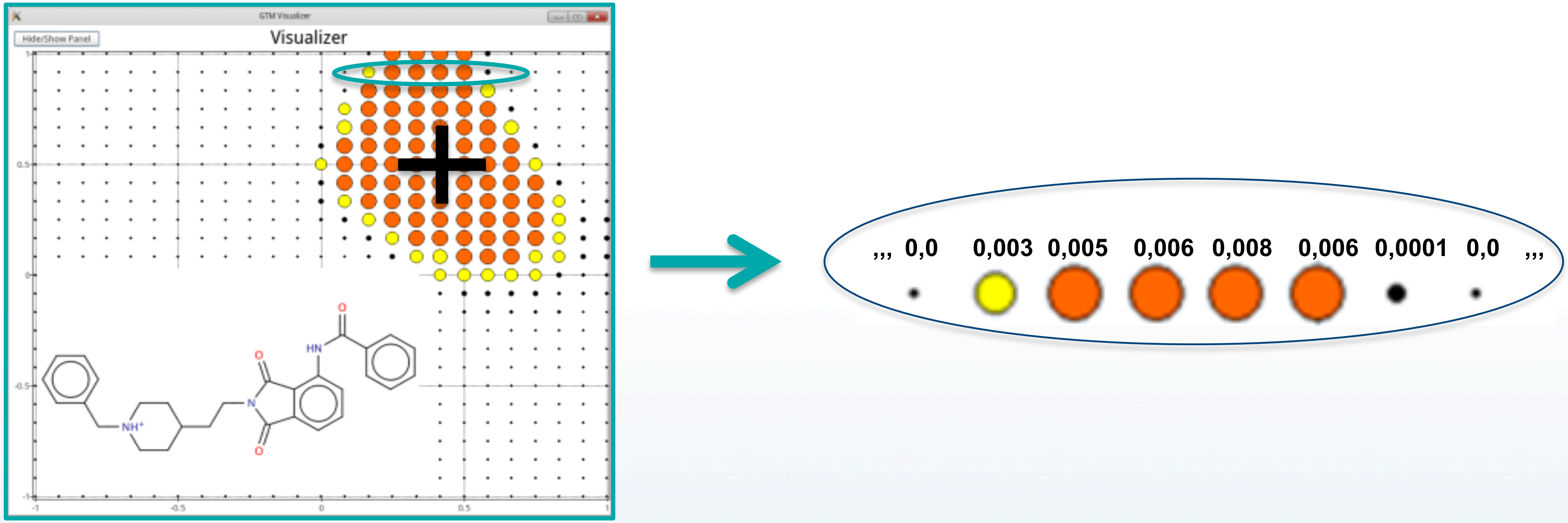


4



0

Initial compounds in GTM Latent space

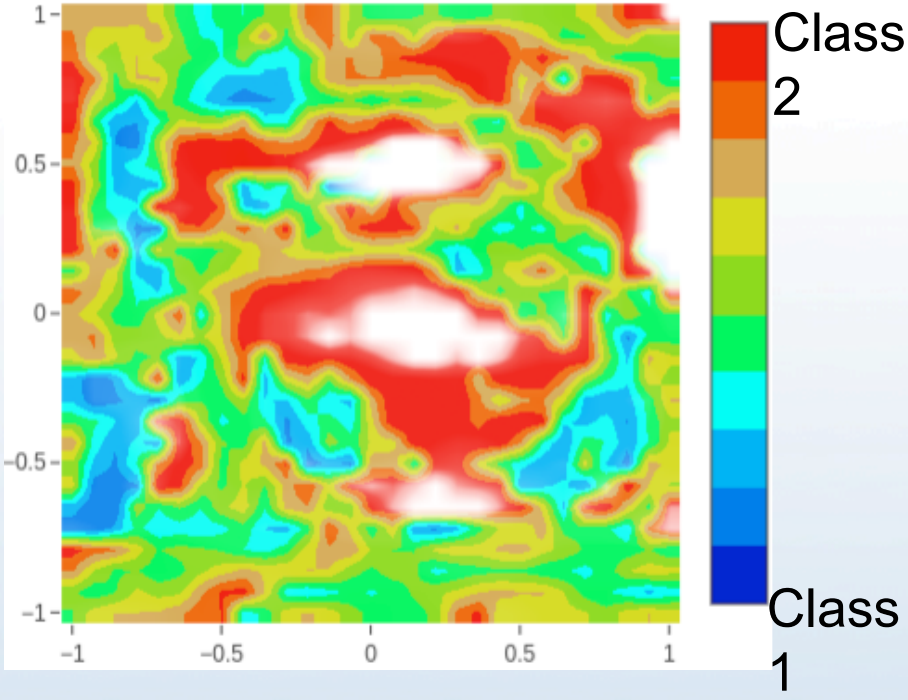


Each molecule in 2D latent space is described by a responsibilities' vector $\{\boldsymbol{R}\downarrow\boldsymbol{tk}\}$ of $\boldsymbol{N_{nodes}}$ length (a vector of normalized probabilities).

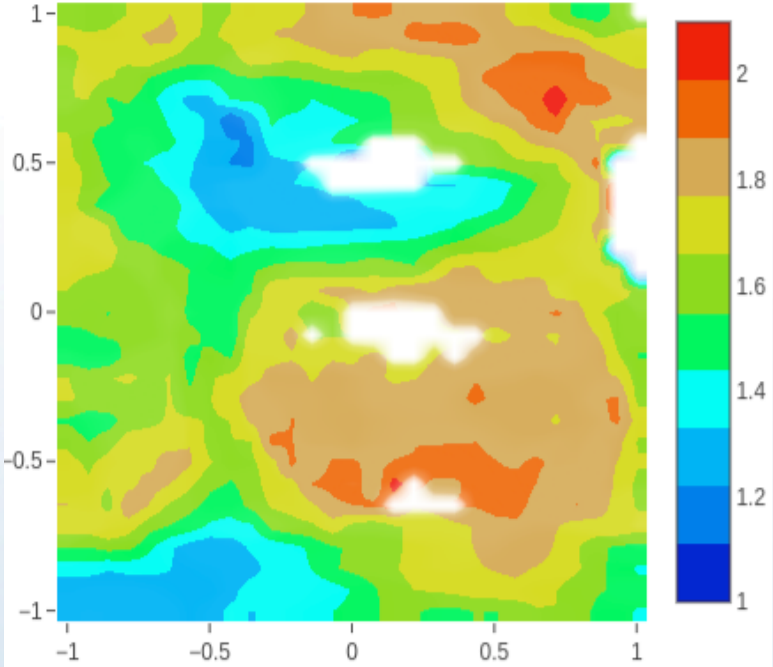
The entire data set is described by a cumulated vector of responsibilities.

Class/Property GTM maps

Cumulated vector of responsibilities × Known class/ property values = Class/Property map

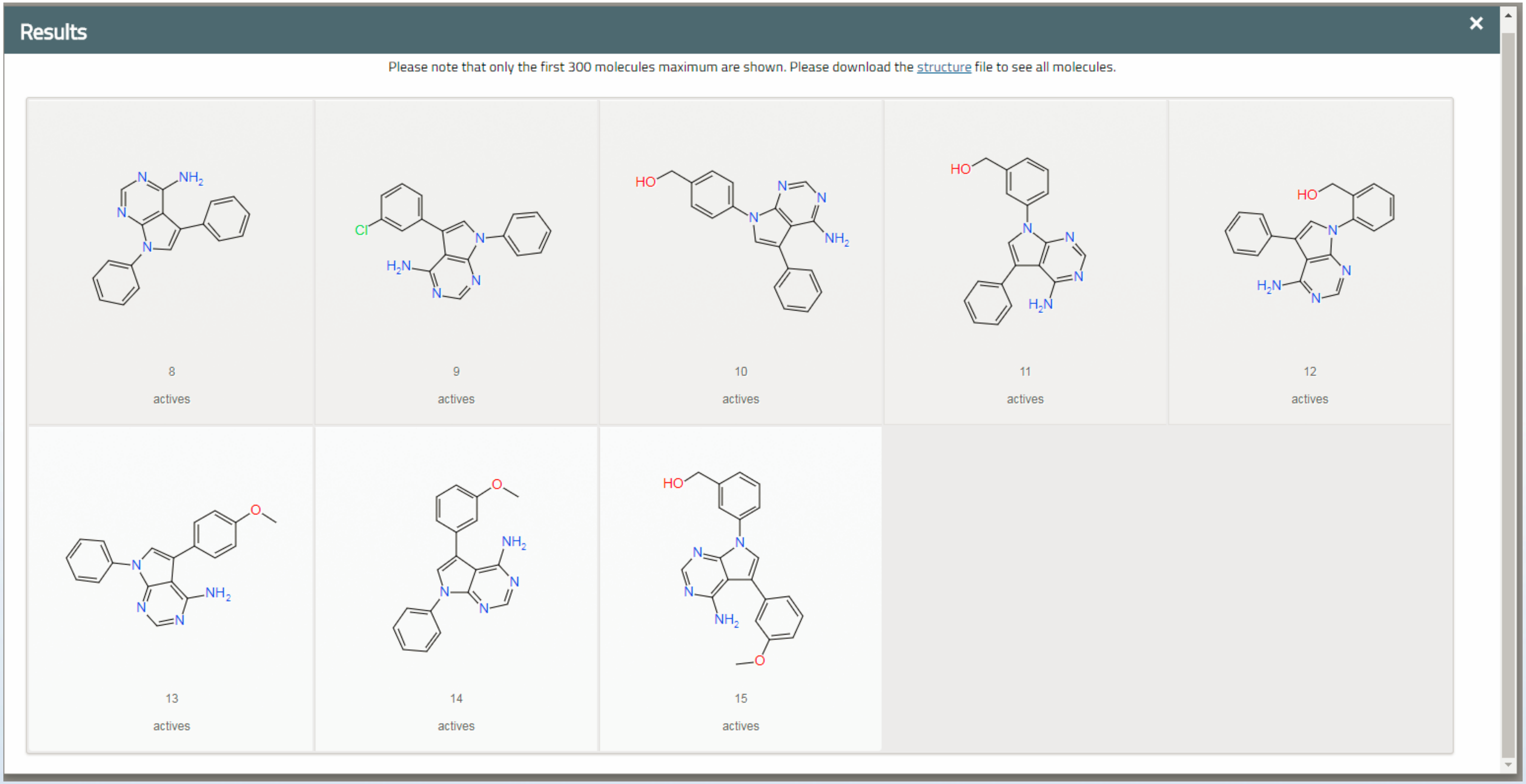


Class map



Property map

Online GTM



Results

Please note that only the first 300 molecules maximum are shown. Please download the structure file to see all molecules.

8

actives

9

actives

10

actives

11

actives

12

actives

13

actives

14

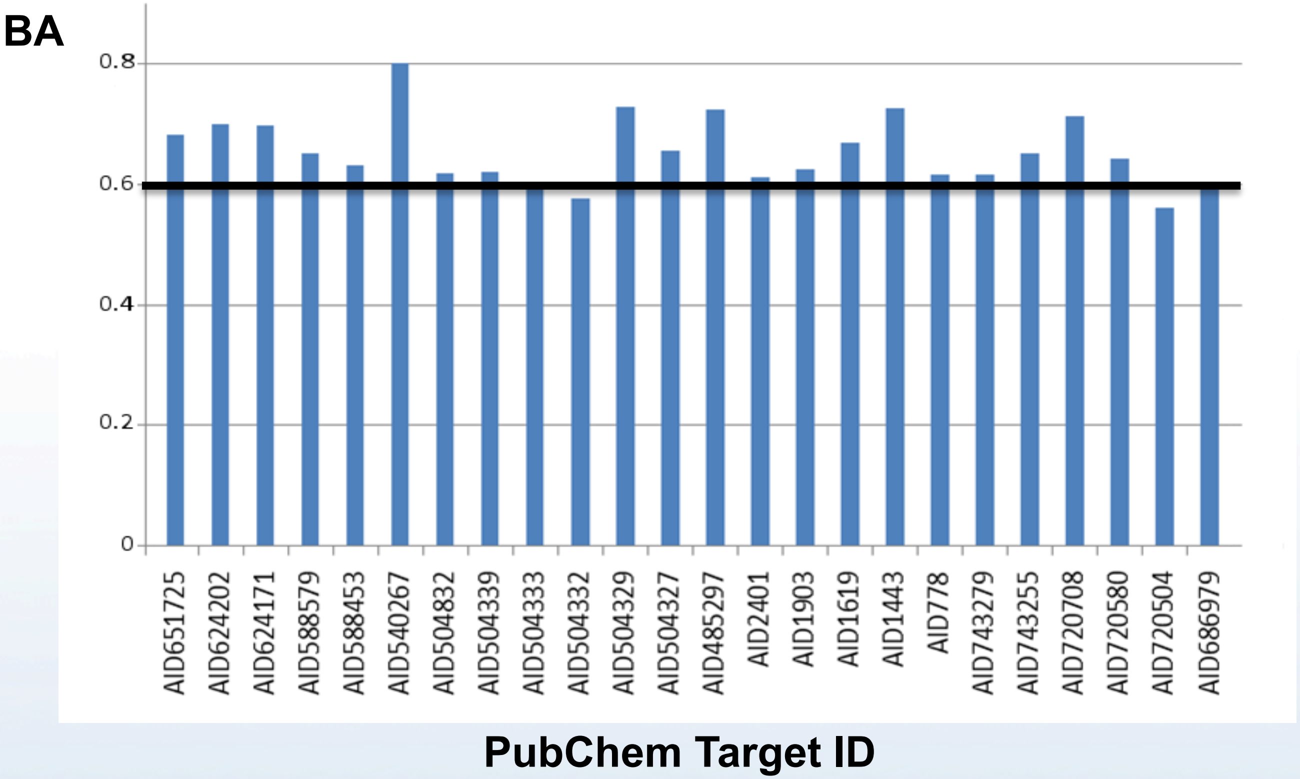
actives

15

actives

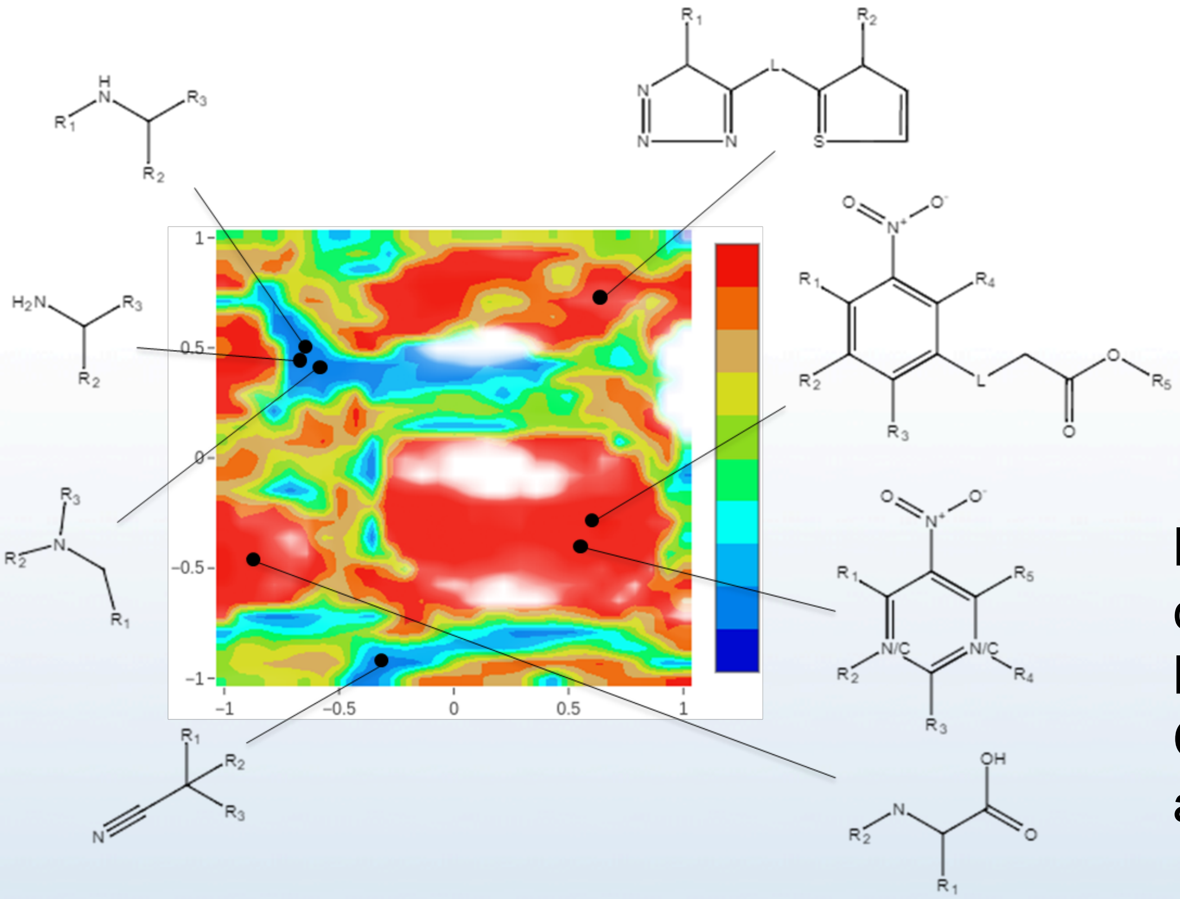
http://infochim.u-strasbg.fr/onlineGTM

GTM Map‘s Quality Control



3-fold Cross-Validated Balanced Accuracy (BA) of classification models for *active* vs *inactive* separation for **24** selected targets shows high predictive performance of the built GTM map.

FDB-17 versus PubChem-17

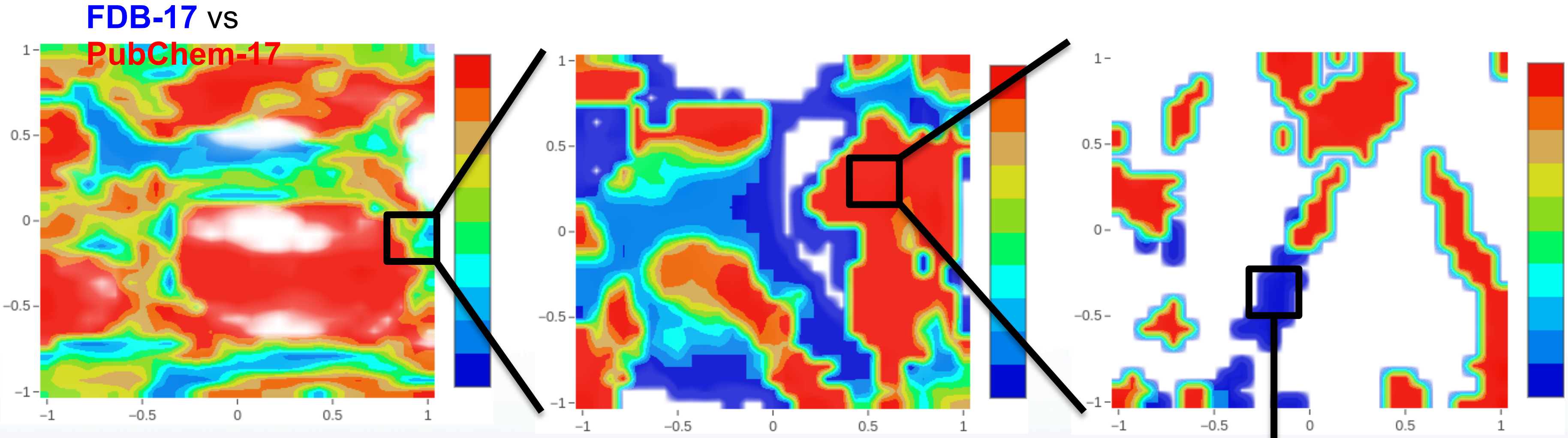


Large red area containes compounds with **halogens, NO_2 groups, C#C bonds** which are absent in **FDB-17**.

FDB-17

PubChem-17

GTM *Zooming*

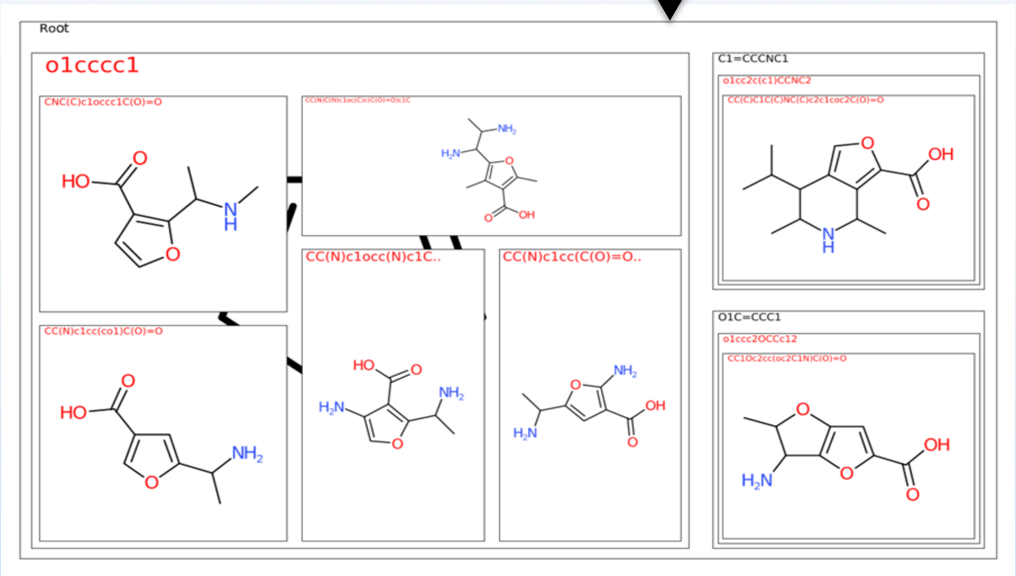


21 million compounds

650 000 compounds

2 500 compounds

Simple Scaffold analysis retrieved some unique for PubChem-17 structures.


Root
o1cccc1
CNC(C)c1occc1C(O)=O
CC(N)C(N)c1oc(C)c(C(O)=O)c1C
CC(N)c1cc(co1)C(O)=O
CC(N)c1occ(N)c1C..
CC(N)c1cc(C(O)=O..
C1=CCCNC1
o1cc2c(c1)CCNC2
CC(C)C1C(C)NC(C)c2c1coc2C(O)=O
O1C=CCC1
o1ccc2OCCc12
CC1Oc2cc(oc2C1N)C(O)=O