UniStra activities within the *BigChem* project:

- data visualization and modeling using GTM approach;
- chemical reactions mining with Condensed Graphs of Reactions

*Alexandre Varnek*

*Laboratory of Chemoinformatics, University of Strasbourg*

Munich, 18th January 2016
Generative Topographic Mapping (GTM)
Generative Topographic Mapping (GTM)

Swiss Roll

- GTM relates the latent space with a 2D “rubber sheet” (manifold) injected into the high-dimensional data space.
- The visualization plot is obtained by projecting the data points onto the manifold and then letting the “rubber sheet” relax to its original form.
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
GTM descriptors for molecules and datasets

Map resolution: $N_{nodes} = K^2$
Standard setting: $K = 25$, $N_{grid} = 625$

Molecule $\longrightarrow$ responsibilities’ vector $\{R_{tk}\}$ of $N_{nodes}$ length

Dataset $\longrightarrow$ normalized cumulated responsibilities’ vector of $N_{nodes}$ length
1. Chemical Space analysis
   1.1 *Big Data* problem: visualization and analysis of large databases
   1.2 Concept of « universal » chemical spaces

2. Structure-property modeling
   2.1 Individual classification and regression models
   2.2 Profiling models
   2.3 Applicability Domain of Models

3. In silico design
   3.1 GTM Activity landscapes
   3.2 Chemical structures generation (« inverse » QSAR)

4. Visualization and analysis of popular chemical databases:
   ChEMBL, SuppliersDB, IUPAC stability constants DB ...

5. Chemical Reactions Data visualization and analysis using the Condensed Graph of Reaction method
   (see Figure 3)

6. New modules in the ISIDA package:
   7.1 ISIDA/GTM
   7.2 *Stargate* GTM
   7.3 On-line GTM tools
References


Condensed Graph of Reaction (CGR)
Conventional bonds: single, double, aromatic, ...

Dynamical bonds: created single, broken single, ...

CGR: a pseudo-molecule representing a given reaction
Reaction can be encoded by a descriptor vector which can be used in structure-reactivity modeling, similarity searching, clustering, etc.
1. **Automatized processing of raw reaction data**
   1.1 Reaction data curation
   1.2 Atom-to-Atom Mapping

2. **Structure-Reactivity modeling**
   2.1 Classification and regression models
   2.2 Similarity-based approach

3. **Automatized reactions classification**
   3.1 Data visualization and clustering
   3.2 Extraction of reaction signatures

4. **Expert system for protecting groups reactivity**

5. **Predictive models for:**
   5.1 Reaction rate (substitution, elimination, cycloaddition and bio-orthogonal reactions)
   5.2 Tautomeric equilibrium constants
   5.3 Regioselectivity of enzymatic reactions
   5.4 Activity cliffs and bioisosters

6. **Visualization and analysis of reaction databases using GTM:**
   Reaxys, FlowReact DB

7. **New modules in the ISIDA package:**
   7.1 ISIDA/GGR designer
   7.2 Mapper
   7.3 On-line reactivity predictor
References


• F. Hoonakker, N. Lachiche, A. Varnek, A. Wagner Condensed Graph of Reaction: considering a chemical reaction as one single pseudo molecule *Intern. J. Artificial Intelligence Tools*, 2011, **20**, (2), 253-270
