UniStra activities within the *BigChem* project:

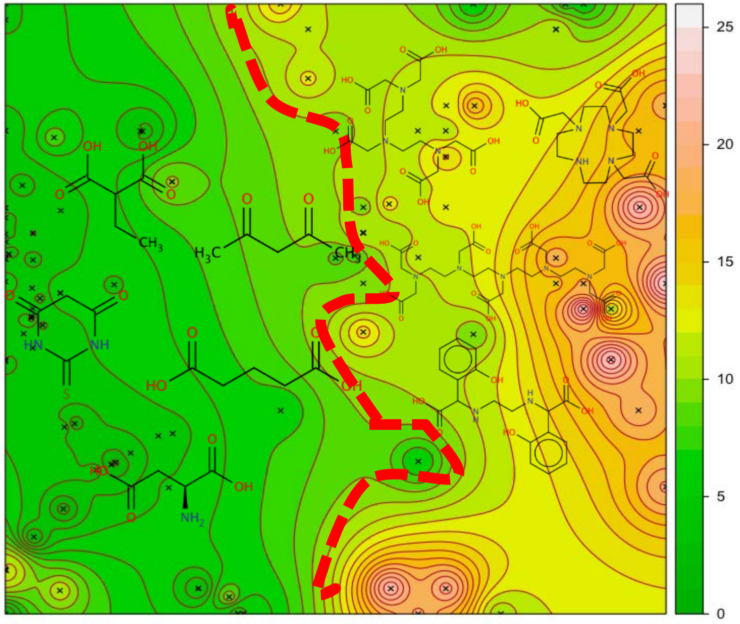
- **data visualization and modeling using GTM approach;**
- **chemical reactions mining with Condensed Graphs of Reactions**

Alexandre Varnek

Laboratory of Chemoinformatics, University of Strasbourg

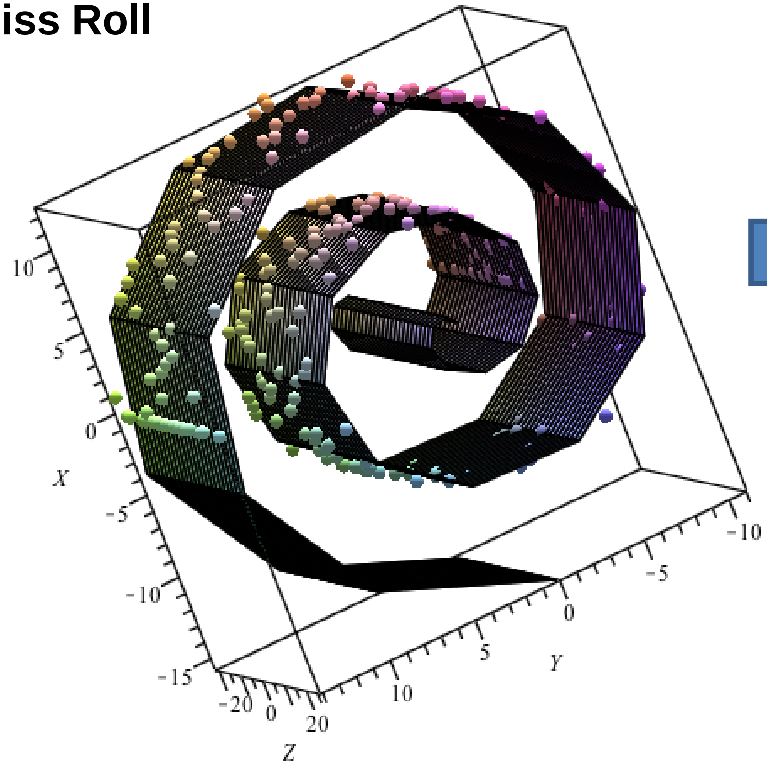
Munich, 18th January 2016

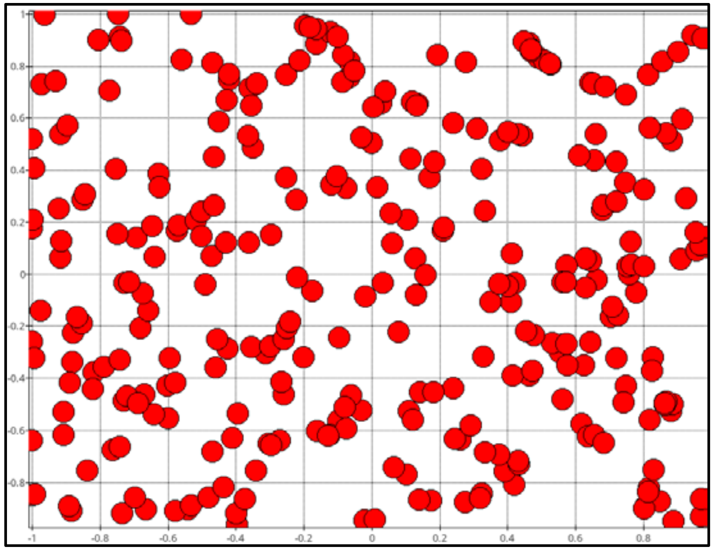
Generative Topographic Mapping (GTM)



Generative Topographic Mapping (GTM)

Swiss Roll





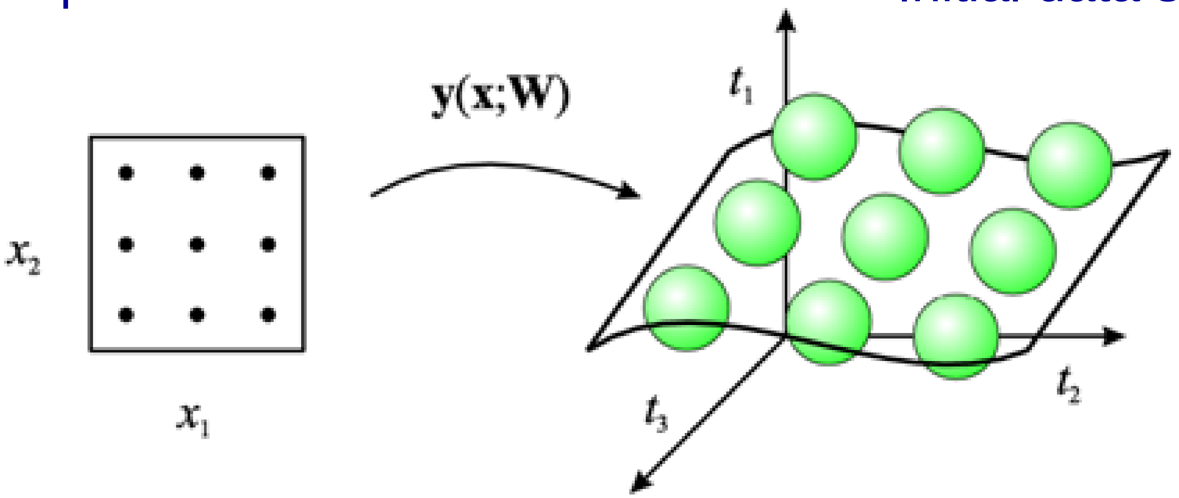
- GTM relates the latent space with a 2D “rubber sheet” (*manifold)* injected into the high-dimensional data space.

- The visualization plot is obtained by projecting the data points onto the manifold and then letting the “rubber sheet” relax to its original form.

Generative Topographic Mapping (GTM)

latent space

Initial data space

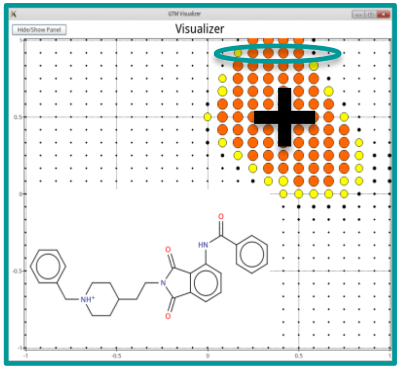


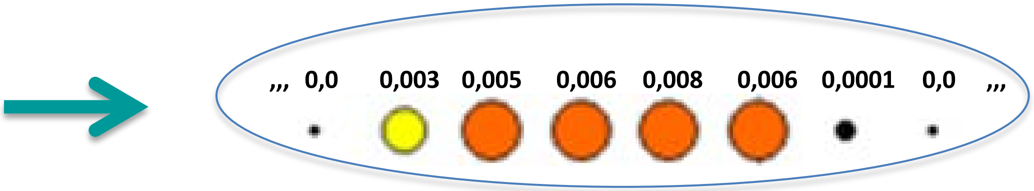
GTM generates a data probability distribution in ***both initial and latent data spaces*.**

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 a, D. Horvath, G. Marcou and A. Varnek, *Mol. Informatics, 2012,* **31***,* 201-312

GTM descriptors for molecules and datasets





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

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

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

Dataset → normalized cumulated responsibilities' vector of $\boldsymbol{N_{nodes}}$ 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

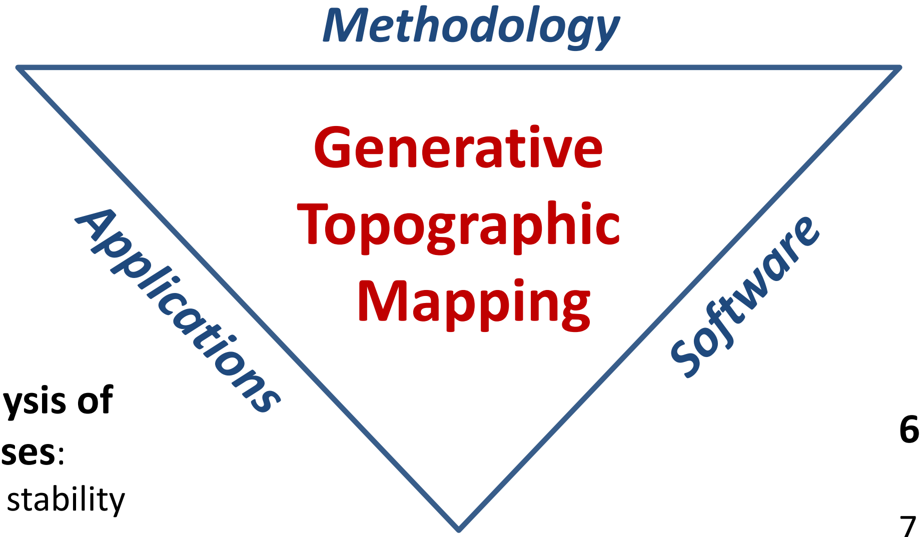
2.2 Profiling models

2.3 Applicability Domain of Models

3. In silico design

3.1 GTM Activity landscapes

3.2 Chemical structures generation (« inverse » QSAR)



Methodology

Generative Topographic Mapping

Applications

Software

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

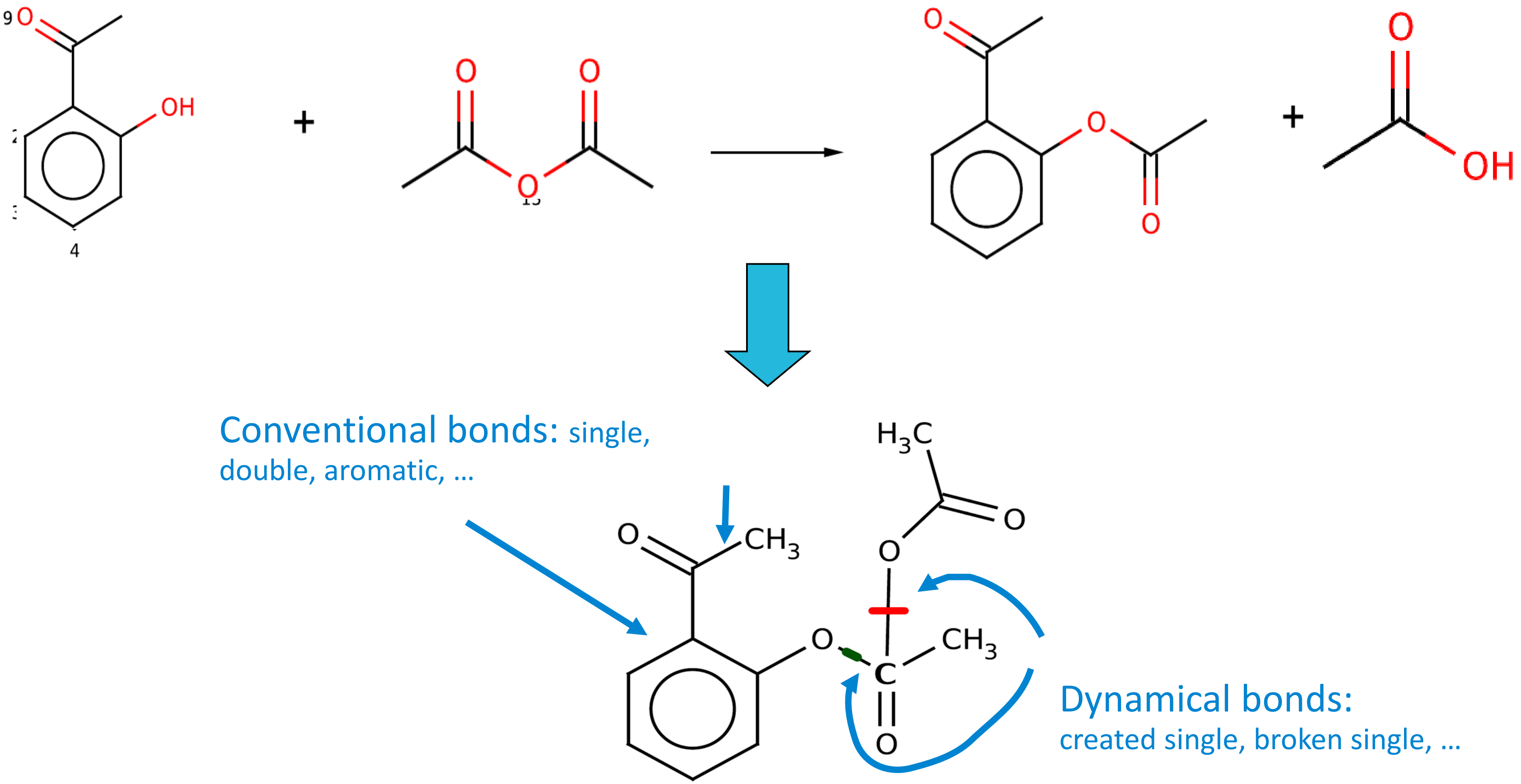
7.3 On-line GTM tools

References

- N. Kireeva, I. Baskin, H. Gaspar, D. Horvath, G. Marcou, A. Varnek, Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison, *Mol. Informatics* 2012, 31 (3-4), 301-312
- H Gaspar, G Marcou, D Horvath, A Arault, S Lozano, P Vayer, A Varnek, Generative Topographic Mapping-Based Classification Models and Their Applicability Domain: Application to the Biopharmaceutics Drug Disposition Classification System (BDDCS), *J. Chem. Inf. Model.,* 2013, 53 (12), 3318-3325
- H Gaspar, II Baskin, G Marcou, D Horvath, A Varnek, GTM-Based QSAR Models and Their Applicability Domains, *Mol. Informatics,* 2015, *DOI: 10.1002/minf.201400153*
- H. Gaspar , I. I. Baskin, G. Marcou, D. Horvath, A. Varnek Stargate GTM: bridging descriptor and activity spaces. *J. Chem. Inf. Model., 2015,* DOI: 10.1021/acs.jcim.5b00398
- H Gaspar, II Baskin, G Marcou, D Horvath, A Varnek, Chemical Data Visualization and Analysis with Incremental GTM: Big Data Challenge, *J. Chem. Inf. Model.*, 2015, 55 (1), 84–94
- P. Sidorov, H. A. Gaspar, Helena; A. Varnek, G. Marcou, D. Horvath Mappability of drug-like space: towards a polypharmacologically competent map of drug-relevant compounds, *Accepted in J. Comp. Aided Mol. Design 2015*

Condensed Graph of Reaction (CGR)

Condensed Graph of Reaction



Conventional bonds: single, double, aromatic, …

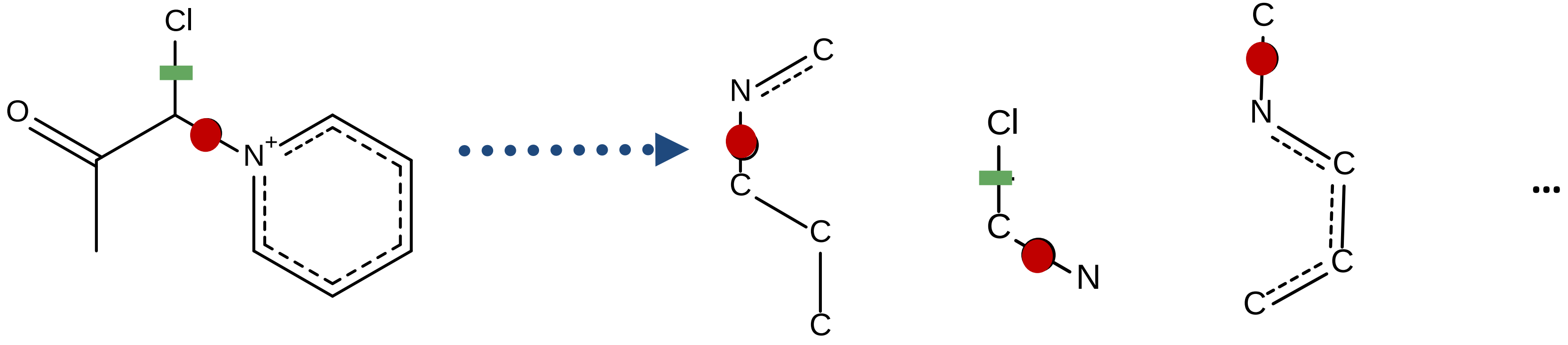
Dynamical bonds: created single, broken single, …

CGR: a pseudo-molecule representing a given reaction

ISIDA/CGR fragment descriptors

Condensed graph of reaction

ISIDA fragment descriptors



2	**1**	**2**	**...**

Reaction can be encoded by a descriptor vector which can be used in structure-reactivity modeling, similarity searching, clustering, etc

A. Varnek In: "*Chemoinformatics and Computational Chemical Biology*", J. Bajorath, Ed., Springer, 2010

1. Automatized processing of raw reaction data

1.1 Reaction data curation

1.2 Atom-to-Atom Mapping

2. Structure-Reactivity modeling

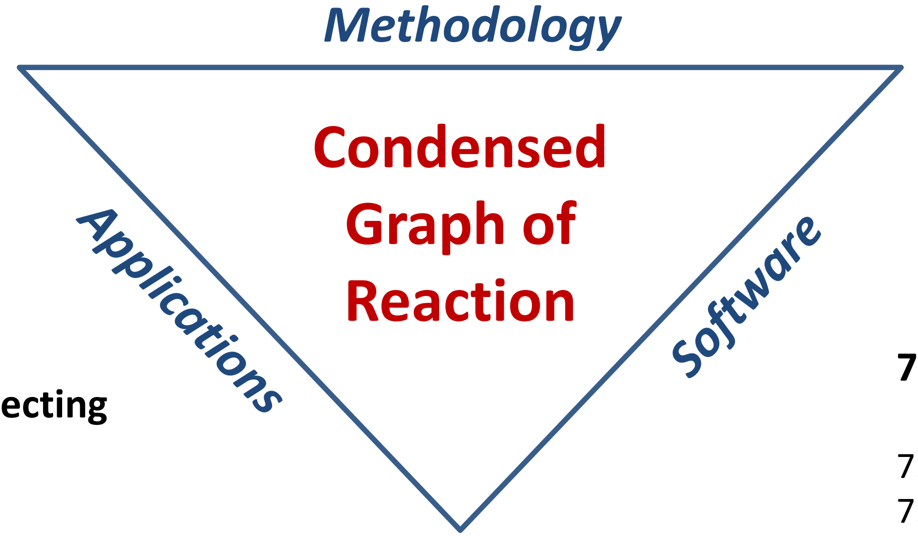
2.1 Classification and regression models

2.2 Similarity-based approach

3. Automatized reactions classification

3.1 Data visualization and clustering

3.2 Extraction of reaction signatures



Methodology

Applications

Software

Condensed Graph of Reaction

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

- A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov’ev *Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures.* J. Computer-Aided Molecular Design, *2005, **19**, 693-703*
- C. Muller, G. Marcou, D. Horvath, J. Aires-de-Sousa, A. Varnek *Models for identification of erroneous atom-to-atom mapping of reactions performed by automated algorithms*. *J. Chem. Inf. Model.* 2012, **52** (12), 3116–3122
- F. Hoonakker, A. Varnek and A. Wagner *A computer based method for calculate similarity between two reactions based on the concept of Condensed Graph of Reactions.* US Patent 11/779 255, PCT/IB2008/052851 from 17.07.2008
- A. de Luca, D. Horvath, G. Marcou, V. P. Solov'ev and A. Varnek *Mining Chemical Reactions Using Neighborhood Behavior and Condensed Graphs of Reactions Approaches* *J. Chem. Inf. Model.* 2012, **52** (9), 2325–2338
- 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
- T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A.V. Bodrov, A.I. Lin, I.I. Baskin, I.S. Antipin, A.A. Varnek *Structure-reactivity relationships in terms of the condensed graphs of reactions* *Russ. J. Org. Chem.*, 2014, **50** (4), 459-463
- G. Marcou, J. Aires de Sousa, D. Latino, A. Deluca, D. Horvath, V. Rietsch, and A. Varnek Towards an expert system for predicting reaction conditions: the Michael reaction case. *J. Chem. Inf. Model.*, 2015, **55**, 239−250