CONDENSED GRAPH OF REACTION - SWISS-KNIFE TOOL FOR REACTION INFORMATICS

Dr. Timur I. Madzhidov
Kazan Federal University, Department of Organic and Medicinal Chemistry

tmadzhidov@gmail.com
A dream…

QSAR  SBDD  Similarity  Molecular dynamics  Quantum chemistry

Generative Neural Nets
... and reality

1) EtOH, SOCl₂, reflux, 3h
2) Me₂C(OMe)₂, cat. TsOH, EtOAc 150-200 mbar, <35 °C, 4h
3) MsCl, Et₃N, EtOAc

cryst. MeOH

3-pentanone (exc.)
cat. TfOH, EtOAc 40 °C, 100 mbar

TiCl₄, Et₃SiH
CH₂Cl₂, -34 °C, 2-6h

NaHCO₃
EtOH/H₂O, 60 °C, 1.5h
extr. and cryst. hexane

Ph₃P, Et₃N, MsOH
DMSO, 50 °C, 1h

NaN₃, NH₄Cl
EtOH/H₂O, 60-65 °C

NaN₃, H₂SO₄
DMSO, 35 °C, 4h

Ac₂O
Bu₂O, 0-25 °C

Bu₃P, cat. AcOH
EtOH/H₂O, 5-20 °C

H₃PO₄, EtOH
50-20 °C, cryst.

Chemoinformatics and molecular modeling lab.
Reaction is complex

two types of species: reactants and products;

unbalanced reactions: missing molecules

dependence on conditions

multi-step reactions
Condensed Graph of Reaction: why?

- Reaction space visualization
- Novel reaction prediction
- Data curation
- Reaction search
- Rate modeling
- Equilibrium constant prediction
- Reaction classification
- Condition prediction
CGR: history

- Yuri KIHO (1972)
- George VLADUTZ (1974) - Superimposed Reaction Skeleton Graph
- Shinsaku FUJITA (1986) - Imaginary Transition Structures
- Gérard KAUFFMAN (1990) - Condensed Graph of Reaction

Reactions in graph-based chemical space

Reactions classification
Reactions rules
Synthesis design

- Alexandre VARNEK (2005) - Condensed Graph of Reaction

Reactions in descriptors-based chemical space

Machine-learning models
Conventional bonds: single, double, aromatic, …

Dynamical bonds: created single, broken single, …

Nugmanov, R.I. et al. (2019) JCIM 59: 2516

https://github.com/cimm-kzn/CGRtools
Reaction can be encoded by a *descriptors vector* which can be used in data analysis or in structure-reactivity modeling.

Condensed Graph of Reaction: why?

**CGR as graph object**
- Reaction balancing
- Data curation
- Reaction search
- Novel reaction prediction
- Reaction classification

**CGR represented by descriptors**
- Condition prediction
- Reaction space visualization
- Equilibrium constant prediction
- Rate modeling
Tools

CGRtools
Data processing

CGRdb2.0
Data storage and search

CIMtools
QSPR in scikit-learn pipelines

https://github.com/cimm-kzn/CGRtools
https://github.com/icredd-cheminfo/CGRdb2
https://github.com/cimm-kzn/CIMtools
CGR as graph object
Reaction centers as reaction type markers

Signatures for reaction classification

AAM Fixing

Retrosynthetic rule extraction

Applicability domain - RTC
Different levels of reaction centers

ICClassify (*InfoChem*)

**0-Sphere (BROAD)**
Reaction centers only

**1-Sphere (MEDIUM)**
Reaction centers plus alpha atoms, excluding hydrogens

**2-Sphere (NARROW)**
Reaction centers plus beta atoms, excluding hydrogens and consecutive $sp^3$-atoms

---


Reaction types in USPTO database

- 219K “RC+1” motifs were found in 1,36M reactions
- 1063 motifs occur in ≥100 reactions
Reaction centers as applicability domains

$S_N^2$ reaction rate dataset

Model for $S_N^2$ reaction rate prediction

Reaction type control AD

Model = RF
AD = RF variance + RTC

CGR can be used for...

3D CGR proposed for Transition State storage and visualization


CGR SMILES as reaction representation

Autoencoder performing SMILES reconstruction

Chemical structure → Real numbers encoding → Chemical structure
Building GTM on latent variables of autoencoder

Latent variables (vector on real numbers)

Chemical Database (SMILES) → Trained Encoder → GTM

Generation of novel structures from specific areas of the map

GTM activity landscape

Latent variables

Trained Decoder

SMILES

AI-driven design of new Suzuki-like reactions

- 13 new (with respect to the training data) Suzuki-like reactions have been detected
- 5 of them have been found in recent publications

W. Bort et al., *Nature Scientific Reports*, 2021, 11, 3178
CGR encoded by descriptors
General concept

General workflow of “reaction QSAR”

Quantitative Structure-Reactivity Relationships

\[
\log K_T = f (\text{structure}, \text{solvent}, \text{temperature})
\]
Reaction rate and equilibrium constant prediction

$S_N2$ reaction rate constant

$E2$ reaction rate constant

Cycloaddition rate constant

Tautomerisation equilibrium constants


Benchmark of reaction descriptors

Models for
- reaction rate constant of Diels-Alder, $S_N2$ and E2 reactions
- tautomeric equilibrium constant

CGR descriptors were used in top ranked models
Reaction condition prediction complexity

\[ Y = f(X) \]

Optimal conditions

Descriptors of reaction structure

The same reaction could go at several conditions!

No knowledge which conditions are not good for particular reaction!
Condition modelling as ranking

1) (PtO₂, acid, 20 °C, 1 atm)
2) (Pd/C, 20 °C, 1 atm)
3) (Pt, 20 °C, 1 atm)

https://doi.org/10.3390/ijms23010248
Model performance

Dataset: ~90,000 hydrogenation reactions

- Null Model
- k Nearest Neighbors Model
- Likelihood Ranking Model

https://doi.org/10.3390/ijms23010248
Conclusions

- CGR is universal approach for reaction representation solving most of their complexity
- CGR can be manipulated as graphs or can be used for descriptor calculations
- CGRs as graph can be utilized for AAM check or correction, missing molecules identification, data curation, and effective applicability domain for reaction characteristics prediction
- CGRs can be encoded by SMILES and be coupled with generative neural networks for novel reaction generation
- CGR-based fragment descriptors can be applied for reaction characteristics modeling, condition recommendation, reaction space visualization
Acknowledgements

Kazan Federal University
Prof. Igor Antipin
Dr. Ramil Nugmanov (Janssen)
Dr. Timur Gimadiev (BIOCAD)
Dr. Marta Glavatskikh (UNC)
Valentina Afonina
Adelia Fatykhova
Asima Rakhimbekova
Dmitrii Zankov (UniStra)
Aigul Khakimova (BIOCAD)
Artem Kokorin (UniLux)
Ravil Mukhametgaliev
Tagir Akhmetshin (UniStra)
Etc...

University of Strasbourg
Prof. Alexandre Varnek
Dr. Gilles Marcou
Dr. Dragos Horvath
Dr. Olga Klimchuk
Dr. Fanny Bonachera
Dr. Arkadii Lin (InSilico)
William Bort
Iuri Casciuc (Syngenta)
Boris Sattarov (Qubit Pharma)

Hokkaido University
Dr. Pavel Sidorov

University of Olomouc
Dr. Pavel Polishchuk
Mariia Matveyeva
Alexandra Nikonenko

Technion, Israel
Dr. Igor Baskin

Janssen Pharmaceutica
Dr. Natalia Dyubankova
Dr. Jonas Verhoeven
Dr. Joerg Wegner

Elsevier (Reaxys)
Dr. Elena Herzog
Dr. Marcus Fischer

Министерство науки и высшего образования РФ, соглашение 14.587.21.0049 (проект RFMEFI58718X0049)