UGM Boston Oct. 4th 23

Chemaxon

Are reaction data FAIR ... and what can we do with that?

Gerd Blanke
Technical director,
StructurePendium
Technologies GmbH

Are reaction data FAIR?

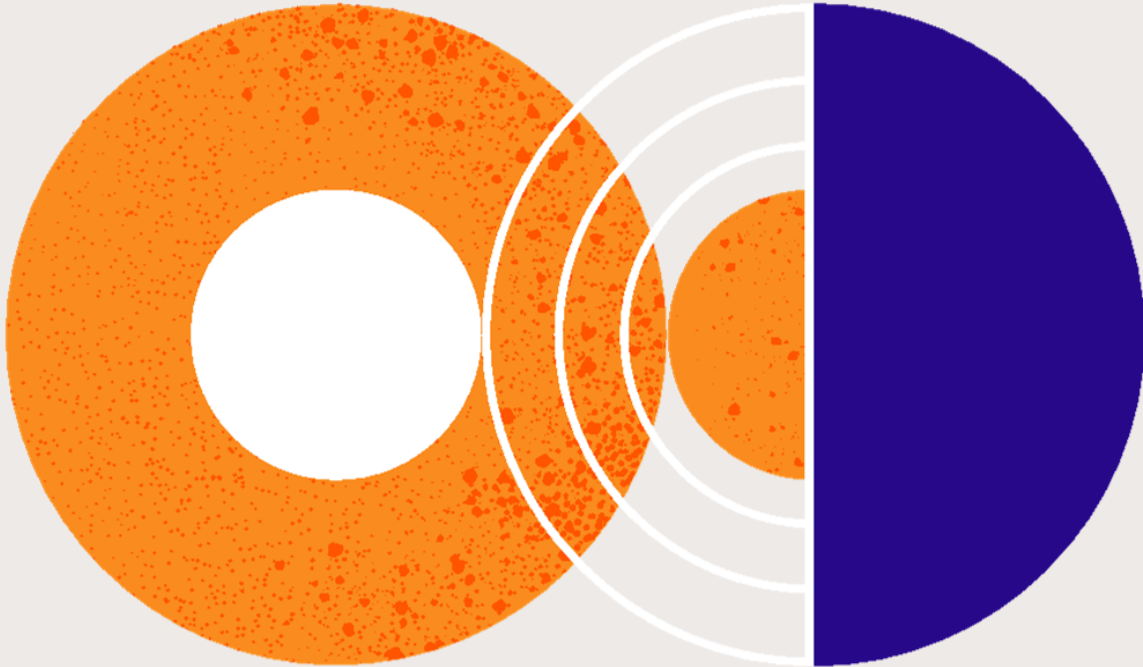
… and what can we do with it?

UGM Boston Oct. 4th 23

Chemaxon

Gerd Blanke

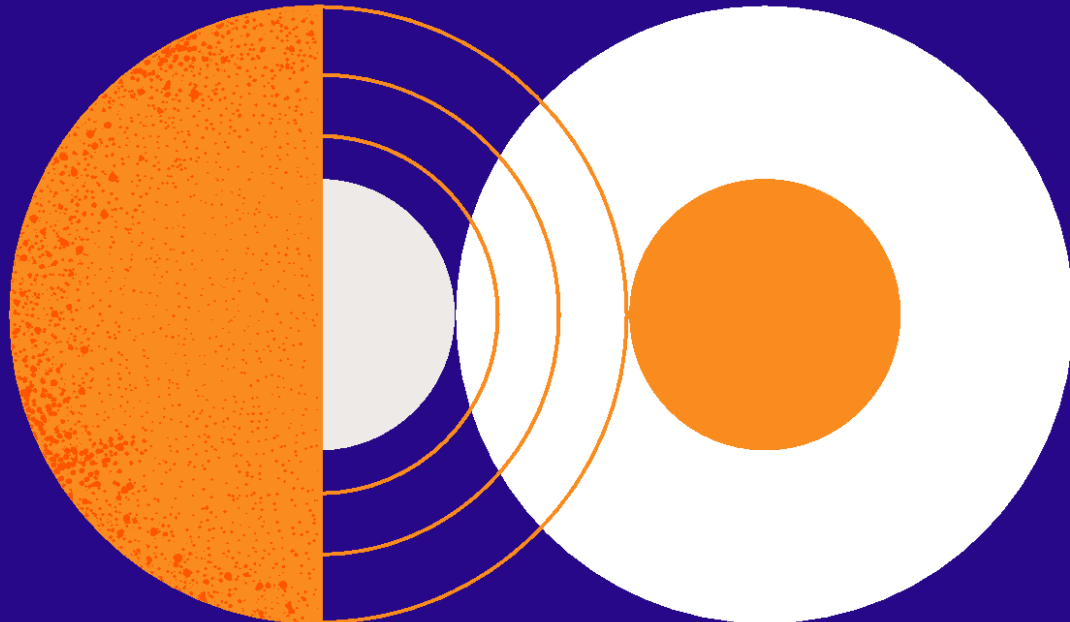
StructurePendium Technologies GmbH, Essen, Germany
Technical Director of the InChI Trust, Cambridge UK



UGM Boston Oct. 4th 23

Chemaxon

About chemical reactions



Chemical reactions are back on stage

General introduction of Electronic Lab Notebooks (ELN) have made chemical reactions generally electronically available

- Primary storage reason: IP protection

Chemical reactions are back on stage

The work of authors like Marvin Segler and Mark Waller (University of Münster, Germany) or Alexei Lapkin (Cambridge university, UK) showed in the middle of the 2010 years that ML is the key to extend the usage of reaction data to the optimization and prediction of chemical reactions

Chemical reactions are back on stage

Software tools for reaction optimization and predictions are on the market and e.g. offered by CAS/ACS, Elsevier/Reaxys, Molecule one, SYNTHIA Merck, Darmstadt)

Chemical reactions are back on stage

- FAIR (Findable, Accessible, Interoperable, Reusable) data is another keyword dominating the current collaboration discussions.

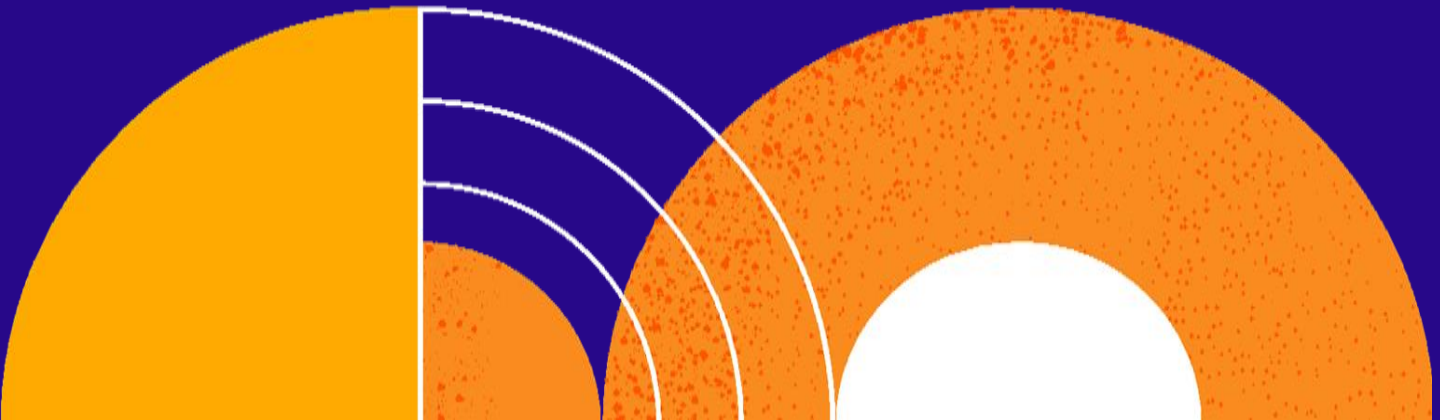
- FAIR data are seen to be the base of the collaboration of chemists within a company and between different organisations

UGM Boston Oct. 4th 23

Chemaxon

Can reaction data be FAIR?

The nitty gritty detail section.



Graphical representation for chemical reactions

Reaction schema

Agents
solvents, catalysts, reagents ...

Reactants $\xrightarrow{\text{Agents}}$ **Products**

- Reactant: starting material
- Product: resulting material
- Agent: material that is necessary to run the rection but does not materially participate in the creation of the products
  - Solvents
  - Catalysts
  - Reagents

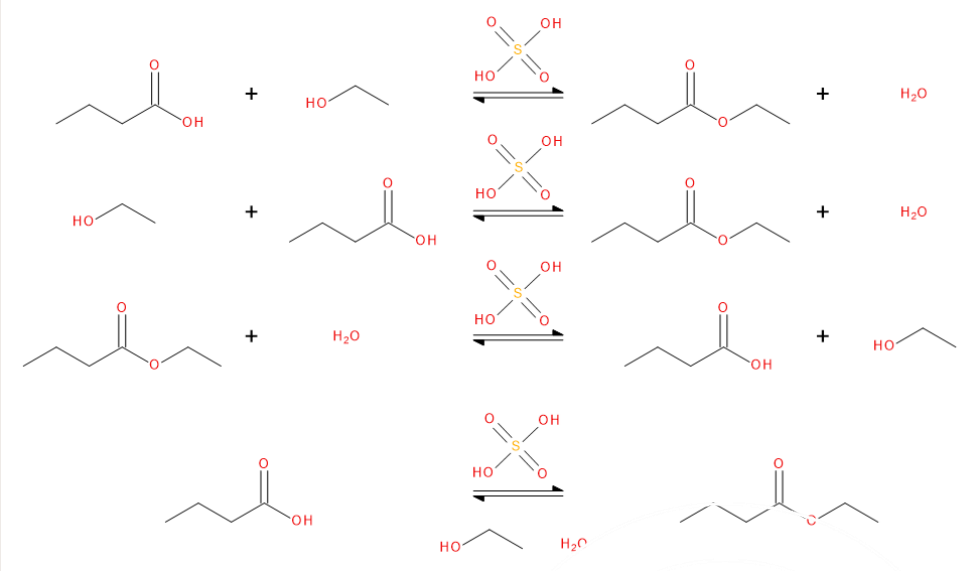
Graphical representation for chemical reactions

Pre and post reaction steps

- Pre-reaction steps / preparation work
  - E.g. drying solvents, activation of catalysts by heating
- Post-reaction steps / work up of products
  - E.g. separation of products from reaction materials, clean-up of product by “washing”, crystallization of product
- Each of these steps participate in the results of a reaction

Graphical representation for chemical reactions

- No guidelines how to order components
- Equilibrium reaction
  - Reactants and products are interchangeable
  - Note: most of the digital formats know A -> B only
- Solvents may be seen as reactants, and products, and/or agents

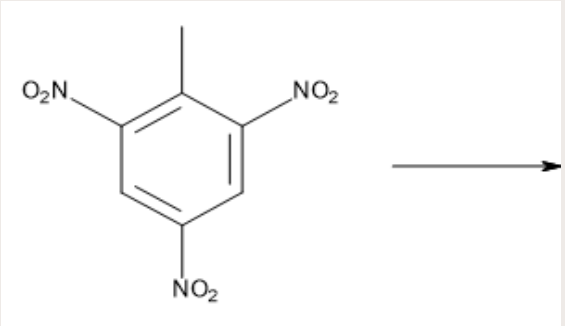


Esterification of butyric acid with ethanol

Graphical representation for chemical reactions

Special cases

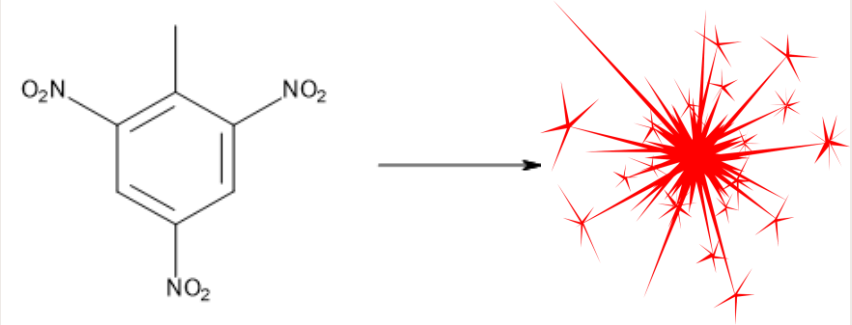
- **“Half reactions“:**
  Only the reactants or only the products are known



Graphical representation for chemical reactions

Special cases

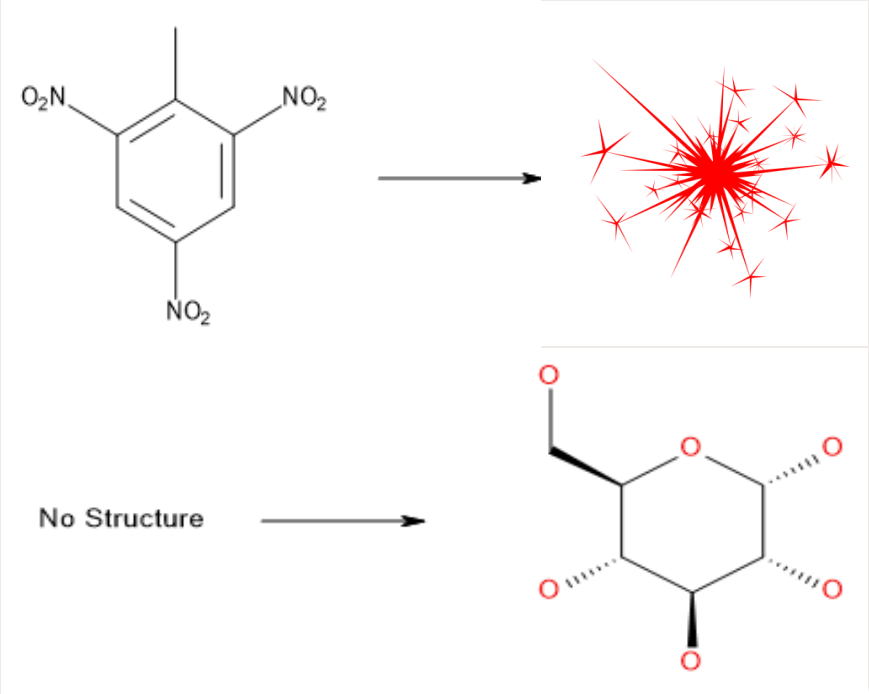
- **“Half reactions“:**
  Only the reactants or only the products are known



Graphical representation for chemical reactions

Special cases

- **“Half reactions“:** Only the reactants or only the products are known
- **“No structures”** are used in case a compound cannot be represented by a chemical structure
  - E.g. biopolymers, natural products



Degradation of starch to glucose

Graphical representation for chemical reactions

The graphical representation for chemical reactions is not uniquely defined and needs guidance for consistent storage!

- **Does IUPAC provide rules?**
  - Definitions for the terms used in chemical reactions including related data (e.g. yield, conversion, isomer ratio for products)
  - Guidelines for the graphical representations

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

CHEMICAL NOMENCLATURE AND STRUCTURE REPRESENTATION DIVISION*

GRAPHICAL REPRESENTATION STANDARDS FOR CHEMICAL REACTIONS**

(IUPAC Recommendations 2019)

Prepared for publication by

LINDA S. PRESS and JEFFERY B. PRESS
Press Consulting Partners, 22 Bearberry Lane, Brewster NY 10509, USA

KEITH T. TAYLOR
Ladera Consultancy, 4791 Mesa Meadows Drive, Sparks NV 89436

Graphical representation for chemical reactions

The graphical representation for chemical reactions is not uniquely defined and needs guidance for consistent storage!

- **Does IUPAC provide rules?**
  - Status: still not officially released in 2023!
  - Note: the 2019 version does not provide guidelines how to order reactants or products

INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

CHEMICAL NOMENCLATURE AND STRUCTURE REPRESENTATION DIVISION*

GRAPHICAL REPRESENTATION STANDARDS FOR CHEMICAL REACTIONS**

(IUPAC Recommendations 2019)

Prepared for publication by

LINDA S. PRESS and JEFFERY B. PRESS
Press Consulting Partners, 22 Bearberry Lane, Brewster NY 10509, USA

KEITH T. TAYLOR
Ladera Consultancy, 4791 Mesa Meadows Drive, Sparks NV 89436

Reaction conditions

Example: Esterification of butyric acid with ethanol

- Component data

	Time	Component	EtOH	Acid	Ester	H_2O	Acid
Summary	1800		0.5	1	.6	0,6	60
Unit			l	mol	mol	mol	ml
Timepoint	0		0.5	1	0	0	0
	300		0.5	.9	.1	.1	10
	600		0.5	.8	.2	.2	20
	900		0.5	.7	.3	.3	30
	1200		0.5	.6	.4	.4	40
	1500		0.5	.5	.5	.5	40
	1800		0.5	.4	.6	.6	40

- Reaction data

	Time	Temperature		pH		Stirring	
Summary	1800.0	20.0	100.0	7.0	5.0	1000.0	1500.0
Timepoint		°C				rpm	
0.0		20.0		7.0		1000.0	
300.0		40.0		6.5		1000.0	
600.0		60.0		6.0		1000.0	
900.0		90.0		5.5		1000.0	
1200.0		100.0		5.0		1000.0	
1500.0		100.0		5.0		1500.0	
1800.0		100.0		5.0		1500.0	

Reaction conditions

Example: Esterification of butyric acid with ethanol

- Component data

	Time	Component	EtOH	Acid	Ester	H_2O	Acid
Summary	1800		0.5	1	.6	0,6	60
Unit			l	mol	mol	mol	ml
Timepoint	0		0.5	1	0	0	0
	300		0.5	.9	.1	.1	10
	600		0.5	.8	.2	.2	20
	900		0.5	.7	.3	.3	30
	1200		0.5	.6	.4	.4	40
	1500		0.5	.5	.5	.5	40
	1800		0.5	.4	.6	.6	40

- Reaction data

	Time	Temperature		pH		Stirring	
Summary	1800.0	20.0	100.0	7.0	5.0	1000.0	1500.0
Timepoint		°C				rpm	
0.0		20.0		7.0		1000.0	
300.0		40.0		6.5		1000.0	
600.0		60.0		6.0		1000.0	
900.0		90.0		5.5		1000.0	
1200.0		100.0		5.0		1000.0	
1500.0		100.0		5.0		1500.0	
1800.0		100.0		5.0		1500.0	

Typical for research and literature databases

Reaction conditions

Example: Esterification of butyric acid with ethanol

- Component data

	Time	Component	EtOH	Acid	Ester	H_2O	Acid
Summary	1800		0.5	1	.6	0,6	60
Unit			l	mol	mol	mol	ml
Timepoint	0		0.5	1	0	0	0
	300		0.5	.9	.1	.1	10
	600		0.5	.8	.2	.2	20
	900		0.5	.7	.3	.3	30
	1200		0.5	.6	.4	.4	40
	1500		0.5	.5	.5	.5	40
	1800		0.5	.4	.6	.6	40

- Reaction data

	Time	Temperature		pH		Stirring	
Summary	1800.0	20.0	100.0	7.0	5.0	1000.0	1500.0
Timepoint		°C				rpm	
0.0		20.0		7.0		1000.0	
300.0		40.0		6.5		1000.0	
600.0		60.0		6.0		1000.0	
900.0		90.0		5.5		1000.0	
1200.0		100.0		5.0		1000.0	
1500.0		100.0		5.0		1500.0	
1800.0		100.0		5.0		1500.0	

Typical for development

- **Reaction optimization**
- **Automated data capture**

What is needed to make reaction FAIR?

Findability, Interoperability

- Common understanding of the component roles and their order within the groups of reactants, products and agents
  - What are reactants, what are products, what goes over/under the reaction arrow (agents)
  - How do you place the components in the reactant, product and agent block (what comes first)
  - Clear rules for the reaction direction
- Consistent drawing rules for the chemical structures
  - Structure checker and normalizer

What is needed to make reaction FAIR?

Reproducibility

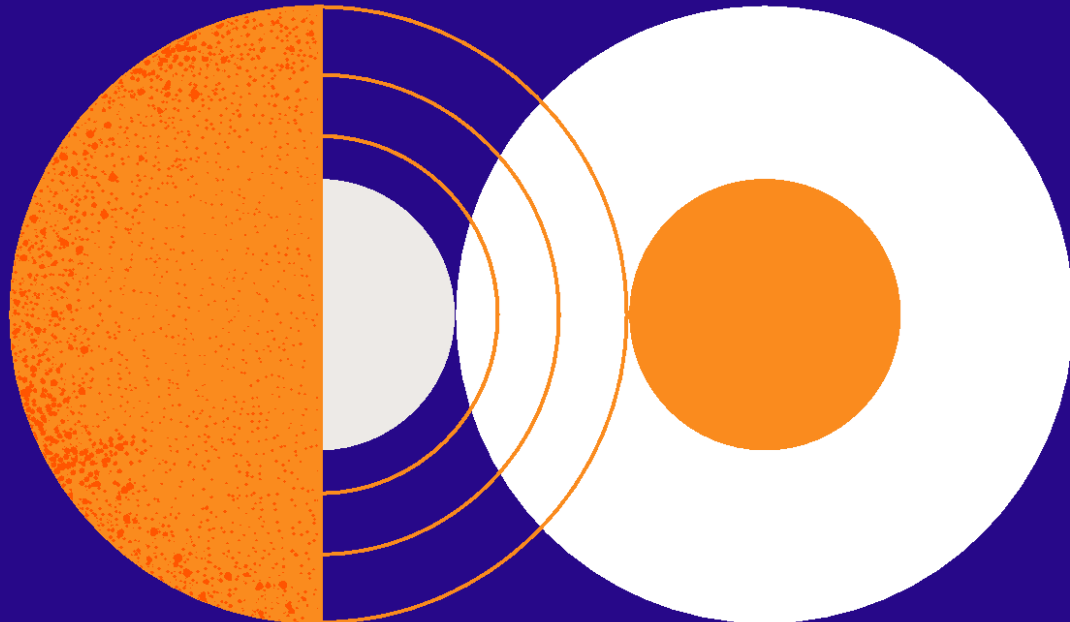
- Sufficient information of the “cooking recipe”
  - Consistent, transparent, and reproducible condition handling
    - Clear criteria for conditions (e.g. yield classification)
      - Use ontologies to make conditions consistent
    - Failed reactions (i.e. reactions under conditions leading to unexpected results) must be captured as well
      - Generally not found in public databases

- Note: Reaction automation and enhanced measurement systems are going to simplify the data capturing.

UGM Boston Oct. 4th 23

Chemaxon

… and what can we do with FAIR reactions



ELN data beyond IP protection

ELN reaction data present your in-house synthesis knowledge

- Make R&D data driven
  - Re-use your in-house knowledge in the structure design workflow
    - Integrate it into your in silico research: develop target molecule, check the synthesizability, run virtual screenings, identify lead compounds for physical realization
- Make the data available for synthesis optimization and prediction

ELN data beyond IP protection

Examples

- Eli Lilly runs an automated synthesis center in San Diego
  - Reported at the Noordwijkerhout conference in 2018 (already)
- Janssen/J&J provides virtual High Throughput Experiments (HTE) to support the optimization of conditions for reactions that failed
  - Reported at the ACS Fall Meeting 2023
- Connor Coley (MIT) is working on self-learning reactors

ELN data beyond IP protection

Short introductions into ML and AI methods for reactions

- Graph databases
  - Ideal storage place for reaction pathways
  - Based on Graph AI methods tool for pathway optimizations and predictions
  - Ask Chemaxon consulting for a Neo4j based reaction graph database

- Finger print methods
  - Fingerprints as representation of the entire reaction
  - Fingerprints of reaction components
  - Various attempts reported

ELN data beyond IP protection

Short introductions into ML and AI methods for reactions

- Google Translate
  - Developed at ETH Zürich, supported by IBM and Google
  - Based on Reaction Smiles
    Example esterification CCCC(=O)O.OCC>>CCOC(CCC)=O.O
  - Reactants as words of language 1 and products as words of language 2
  - Train Google Translate to “translate” language 1 into language 2
  - Reactants -> encode -> (chemical) language model -> decode -> products

ELN data beyond IP protection

Short introductions into ML and AI methods for reactions

- Google Translate
  - BERT (Bidirectional Encoder Representations from Transformers)
    - Currently most used model
    - Atom mapping of reactions (alternative: may common subgraph methods)
    - Identification or reaction mechanisms (alternative: RSS search sets)
    - Optimization and prediction of reaction pathways

ELN data beyond IP protection

Short introductions into ML and AI methods for reactions

- The enumerated reaction pathway approach
  - SAVI (Synthetically Accessible Virtual Inventory, NIH), ENAMINE
    - Simple chemical reactions with generally high success rate are used to continuously enumerate the chemical space
    - Over 100 million virtual compounds (SAVI), more than a billion at Enamine
    - Use transformation pathway for the compound of interest as synthesis route

ELN data beyond IP protection

Integrate with existing tools

- Reaction prediction and optimization tools are on the market
  - CAS, Elsevier (based on work by Mark Waller), Molecule one, Synthia (Merck), etc.
    - Synthia offers fully automated synthesis robot
    - Lee Cronin (University of Glasgow, Lee Cronin Group) has developed automated synthesis workflows based on protocols
      - Chemputation: combines chemistry with computation approaches

ELN data beyond IP protection

How do I make my reaction data ML ready?

- Reaction storage
  - Provide multiple reaction formats
    - RXN (to document the original depiction)
    - Reaction Smiles (use canonical form of one provider)
    - RInChI (Unique reaction identifier)
  - The components of the reactions must be normalized
  - Implement atom mapping for each reaction
    - Some reactions like the Cope rearrangement only become unique with mapping

ELN data beyond IP protection

How do I make my reaction data ML ready?

- Reaction storage
  - Keep reaction relationships
    - Multistep reactions are single reactions that are related to each other
  - Keep component roles (reactants, products, agents …)
    - Clear guidance for the input of component rules
  - Keep reaction types for classifications
    - E.g. Diels-Alder reaction
    - If not available during input identify it by RSS or ML based approaches

ELN data beyond IP protection

How do I make my reaction data ML ready?

- Conditions
  - Capture data properly on ELN level (already)
    - Identify the properties you need for ML
      - Temperature, pressure, atmosphere, time, yield, …
    - ELN needs mandatory fields for requested properties
  - Where text cannot be avoided use NLP for extraction and ontologies to unify the data output

ELN data beyond IP protection

How do I make my reaction data ML ready?

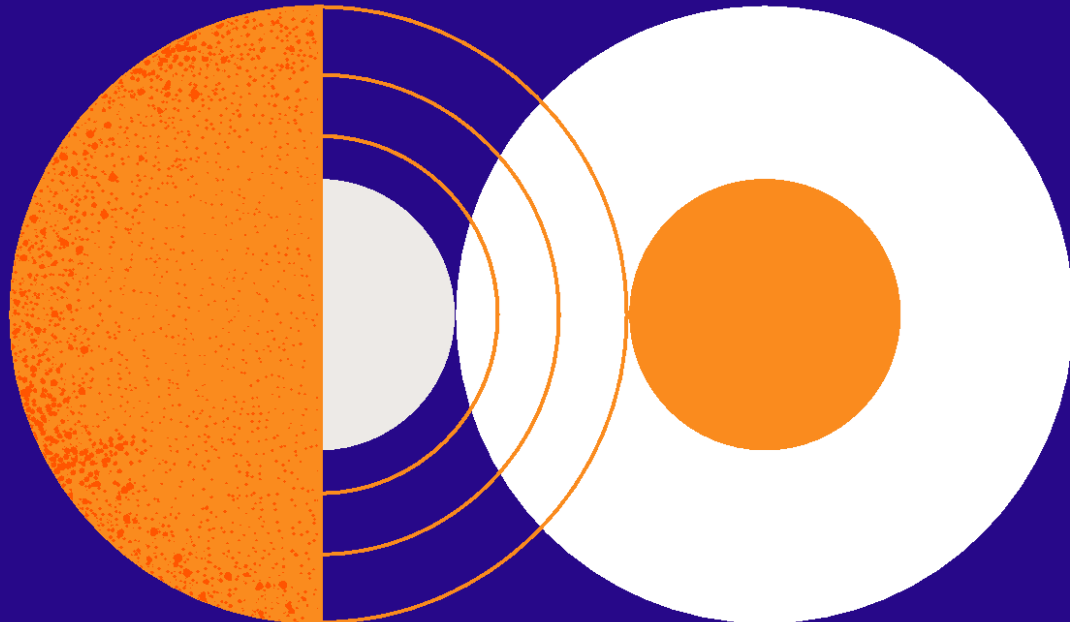
- Conditions
  - For data in Development, an automated data capturing should be developed to implement the handling of an continuing data flow
    - Keep all data versus keep summary data
      - For summary data you need classifications to let data be compared with other data of other reaction types

UGM Boston Oct. 4th 23

Chemaxon

Addendum

Utilities for the handling of reaction data



Utilities for reaction data

Reaction InChI (RInChI)

The RInChI is a unique identifier for chemical reactions based on InChIs as the representation of the reaction components for reactants, products, and reagents
Esterification of butyric acid:

```
RInChI=1.00.1S/C2H6O/c1-2-3/h3H,2H2,1H3!C4H8O2/c1-2-3-
4(5)6/h2-3H2,1H3,(H,5,6)<>C6H12O2/c1-3-5-6(7)8-4-2/h3-
5H2,1-2H3!H2O/h1H2<>H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)/d=
```

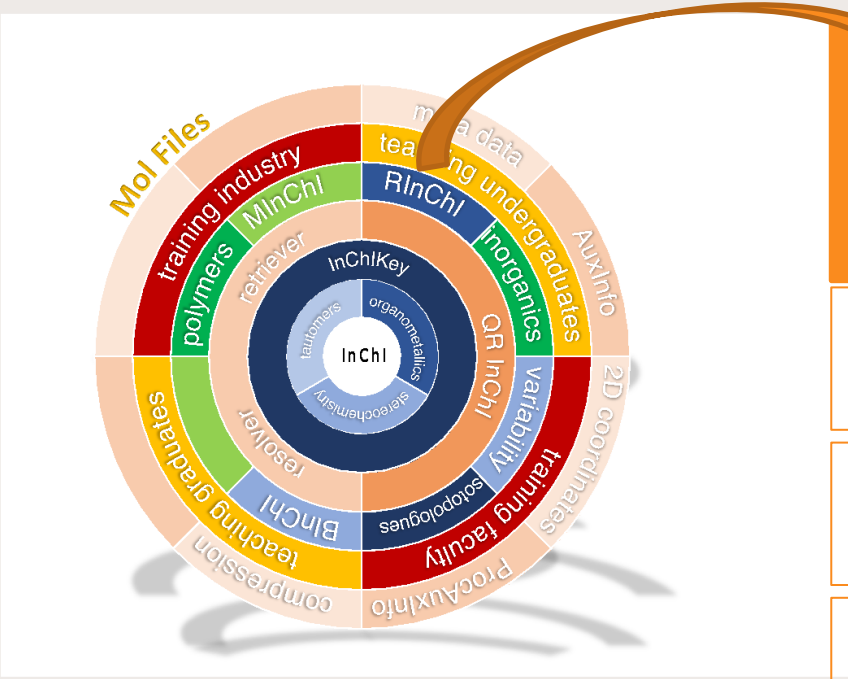
Utilities for reaction data

Reaction InChI Keys (RInChIKeys)

- Long-RInChIKey
  - Each component is represented by Standard InChIs
- Short-RInChIKey
  - Fixed length key for exact match searches and reaction comparisons
- Web-RInChIKey
  - Fixed length key that contains all components of a reaction but without role assignment for Web searches or searches in databases with unknown or different data model

Utilities for reaction data

RInChI auxiliary information layers



RInChI

Auxiliary Information
ProcAuxInfo
"Failed" reactions

Atom mapping
MapAuxInfo

Stereochemistry
layer

Agent roles
AgentInfo

No-structures
NoStructInfo

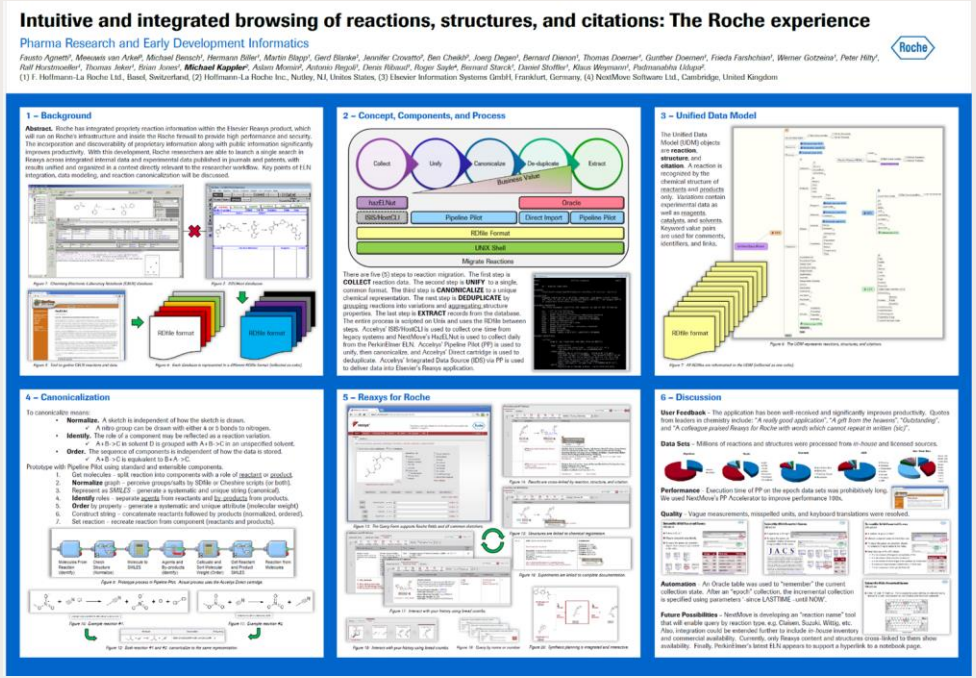
InChI upgrades will drop in

https://github.com/IUPAC-InChI/RInChI

Utilities for reaction data

The Unified Data Model (UDM) for reactions

The Roche UDM project was started in 2012 to export reaction data from different data sources into Reaxys as a common database server for in-house and external reaction data.

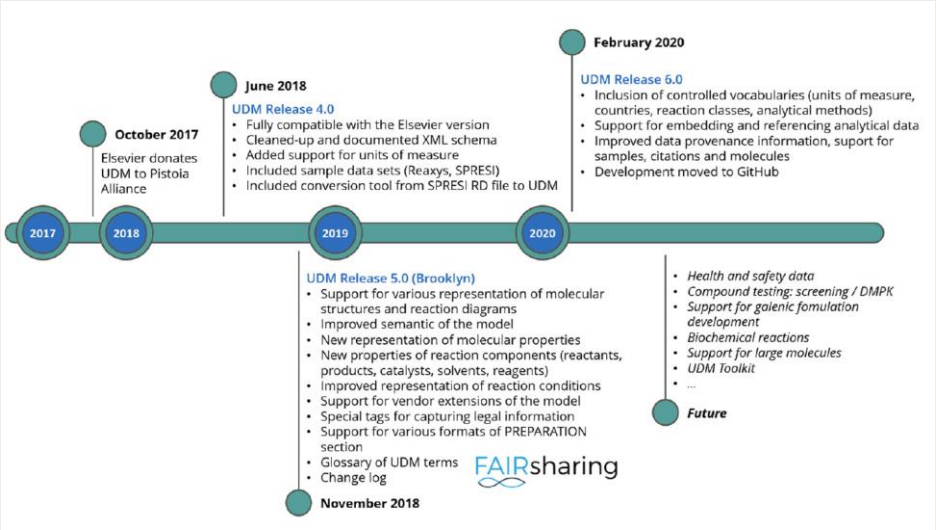


ACS Spring Meeting 2012

Utilities for reaction data

The Unified Data Model (UDM) for reactions

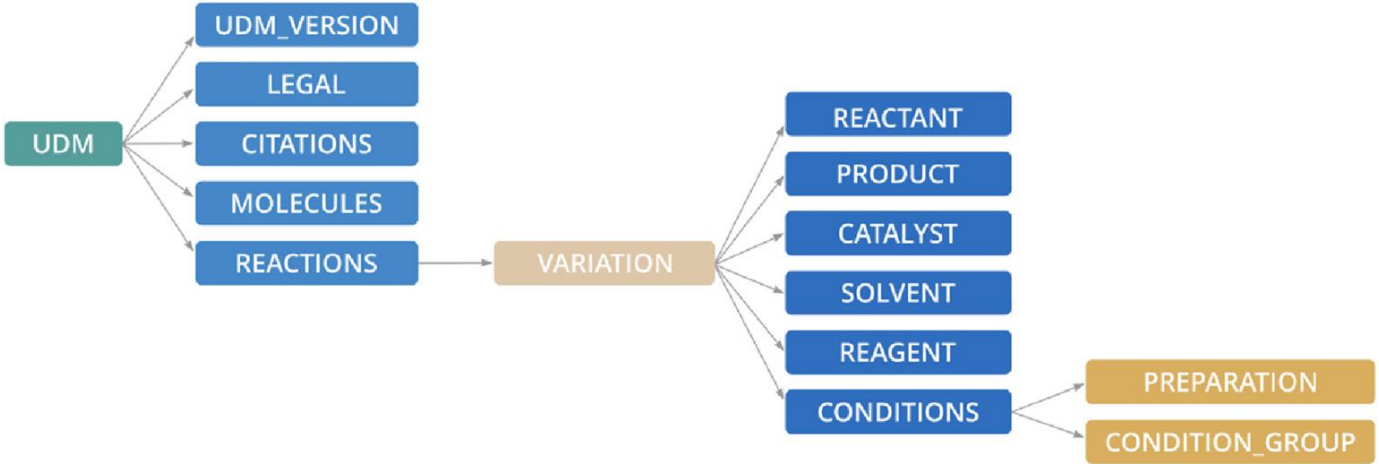
- UDM moved to Elsevier in 2013
- UDM with Elsevier up to 2017.
- UDM under Pistoia until 2020
- Open source since 2020
  - Planned extension
    - Simplified multi step reaction handling by referencing the IDs of ancestors and successor



Utilities for reaction data

The Unified Data Model (UDM) for reactions

Top level elements of the UDM



Utilities for reaction data

The Unified Data Model (UDM) for reactions

- Supported molecule formats: molfiles, InChI, Smiles, CDXML, WLN-Wiswesser line notation
- Supported reaction formats: RXN, RInChI, Reaction SMILES, CDXML
- Deduplication of molecules and reactions based on structures
- Simple alphanumerical controls, e.g. for DOIs

Utilities for reaction data

The Unified Data Model (UDM) for reactions

- Controlled vocabularies
  - County codes / names
  - Reaction classes from the RXNO name reaction ontology
    - https://github.com/rsc-ontologies/rxno
  - Analytical methods and result types taken from Allotrope Foundation Taxonomies (AFT)
  - Vocabulary of measurement units aligned with Allotrope Foundation Ontologies (AFO)
- The data model can be individually extended within the given hierarchy

Utilities for reaction data

The Unified Data Model (UDM) for reactions

- UDM
  - https://www.pistoiaalliance.org/projects/current-projects/unified-data-model/
  - https://github.com/PistoiaAlliance/UDM
  - UDM (Unified Data Model) for chemical reactions – past, present and future, Jarosław Tomczak, Elena Herzog, Markus Fischer, Juergen Swienty-Busch, Frederik van den Broek, Gabrielle Whittick, Michael Kappler, Brian Jones and Gerd Blanke https://doi.org/10.1515/pac-2021-3013

Utilities for reaction data

The Unified Data Model (UDM) for reactions

- Format comparison
  - Roger Sales, nextMove: “Data Formats for Reaction Databases: Lessons Learned from Pistoia UDM and Google/MIT’s Open Reaction Database (ORD)”, Talk at ACS Fall Meeting 2022
    - https://www.nextmovesoftware.com/talks/Sayle_DataFormatsForReactionDatabasesLessonsLearnedFromPistoiaUDMandORD_ACS_202208.pdf

UGM Boston Oct. 4th 23

Chemaxon

Thank you

Gerd Blanke

Gerd.Blanke@
StructurePendium.com