What's going on at the International Chemical Identifier for Reactions (RInChI)

Proposed developments to incorporate into the next release of Reaction InChI, including process information in a machine-readable format, and atom mapping

David Nicolaides[1], Gerd Blanke[2], Günter Grethe, Hans Kraut[3], István Öri[4], Jan Holst Jensen[5], Jonathan Goodman[6]

ChemAxon User Meeting, Budapest, May 22nd, 2019

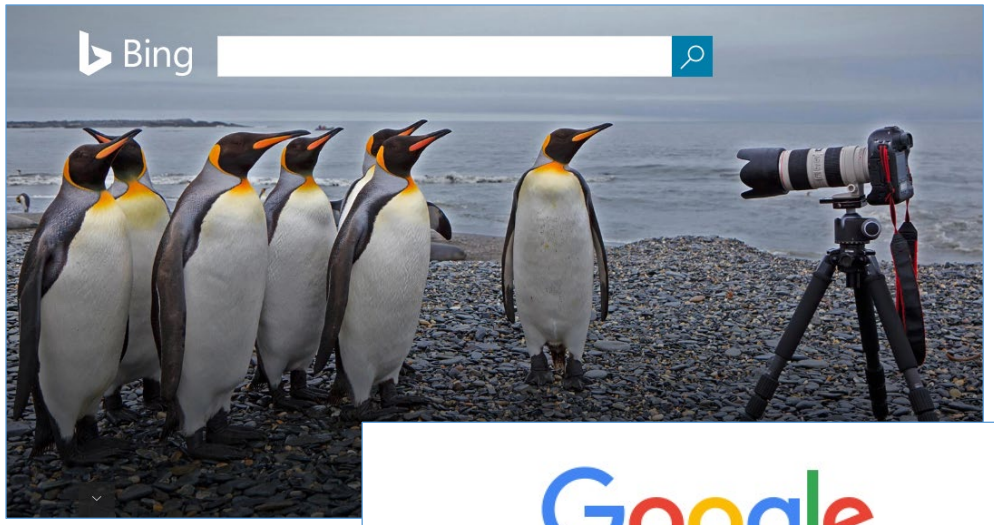
[1] Dassault Systèmes, 334 Cambridge Science Park, Cambridge, CB4 0WN, UK
[2] CΞC StructurePendium Technologies GmbH, Reulsbergweg 5, D-45257 Essen, Germany
[3] InfoChem Gesellschaft für chemische Information mbH, Aschauer Str. 30, D-81549 Munich, Germany
[4] ChemAxon Ltd., Graphisoft Park, Záhony utca 7, Budapest, 1031 Hungary
[5] BiochemFusion ApS, Charlottenlund, Denmark
[6] Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK

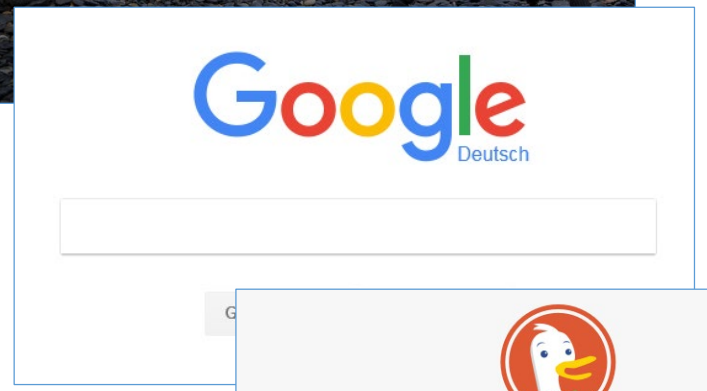
Last year we asked:
What do we need for reactions …

To run searches in the Web?

To provide easy ways to handle reactions in databases ?

To compare reactions between different sources?





IUPAC International Chemical Identifier for Reactions

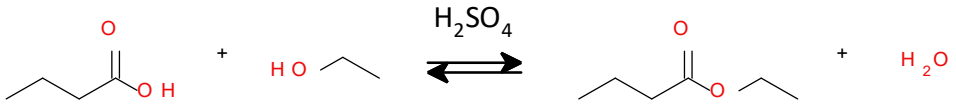
Reaction InChI

(RInChI)

RInChI 1.0 released March 2017

- RInChI
  - The RInChI is calculated from the InChIs of each reactant, product and agent
  - RAuxInfo consists of the AuxInfo of each reaction component
  - Recalculation of RXN/RD files from RInChI and RAuxInfo
- Long-RInChIKey
  - Calculated from InChIKeys of each reactant, product and agent.
- Short-RInChIKey
  - Fixed length hash over all reagents, products and agents
- Web-RInChIKey
  - Fixed length hash developed from the reaction components but ignoring the specific role within the reaction.

RInChI 1.0



- **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=
- **Long-RInChIKey**=SA-EUHFF-LFQSCWFLJHTTHZ-UHFFFAOYSA-N-FERIUCNNQQJTOY-UHFFFAOYSA-N--OBNCKNCVKJNDBV-UHFFFAOYSA-N-XLYOFNOQVPJJNP-UHFFFAOYSA-N--QAOWNCQODCNURD-UHFFFAOYSA-N
- **Short-RInChIKey**=SA-FUHFF-JEVIJXCZCL-UFTQDZUCXS-QAOWNCQODC-NUHFF-NUHFF-NUHFF-ZZZ
- **Web-RInChIKey**=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA

What‘s new: ChemAxon‘s implementation

- Demo

What's next

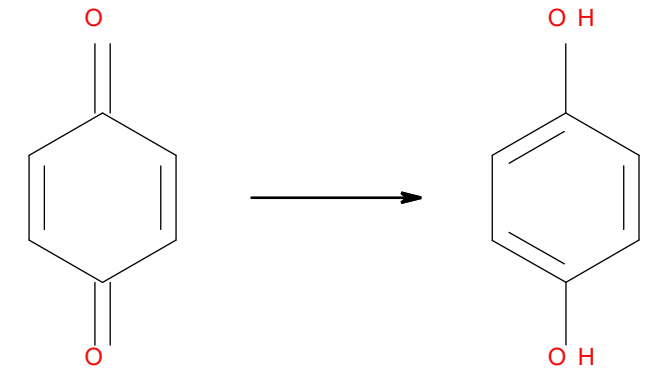
- Plans for the RInChI release 2.0
  - Technical issues
  - Additional import formats
  - Reaction representation
  - Make RInChI useable for AI

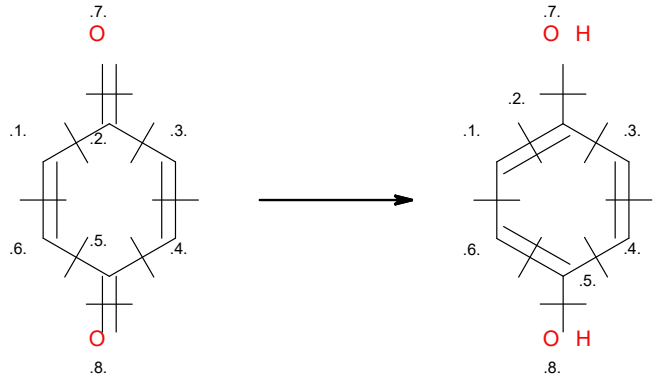
Additional import/export formats

- Reactions built out of the InChIs of each of the reaction components
- Reaction SMILES
- UDM format (Pistoia Alliance, new version released 29-Nov-2018)

Reaction representation

- Atom mapping
  - Atom mapping is used to mark the reaction centers
  - Atom mapping is handled as aux(iliary) layer
    - Proposed name: MapAuxInfo, version 1.00.1
      - MapAuxInfo=1.00.1/
  - We are not implementing a mapping algorithm into RInChI but use the information provided by the RXN file as delivered (by the author)
    - Notes:
      - RXN files only allow atom mapping between starting materials and products but not between starting materials/products and any agents as agents are not part of the RXN file definition



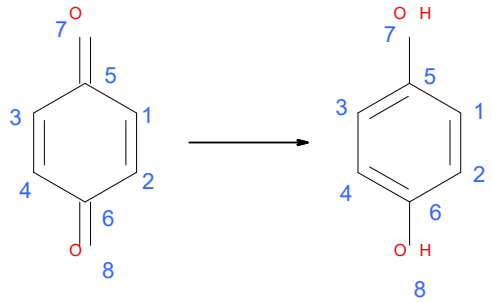


Reaction representation

- Atom mapping
  - Only identify those atoms that are kept during the reactions
  - Rules for the atom identification
    - Each atom is identified by
      - its relationship to the layer 2 or 3 (as educt/product layers in RInChI)
        - This relationship is defined by the position in front of or behind the separator "<>" and does not need to be written explicitly.
      - Its membership to its molecule (1 as first in the layer, 2 as second, ….)
      - Its RInChI number
    - Use a dash "-" as separator between each number
    - In case of equivalent mappings list the InChI numbers of all equivalent positions ordered ascending, separated by a comma "," and terminated by brackets "(" , ")" on the 3rd layer.
    - Use "<>" to indicate the mapping (analog the separator between layer 2 and 3)
    - Separate each string by a semicolon
    - Order the strings alphabetically ascending

Reaction representation

- Atom mapping
  - Quinone example



The numbers in blue represent the InChI numbering

Quinone reduction

```
RInChI=1.00.1S/C6H4O2/c7-5-1-2-
6(8)4-3-5/h1-4H<>C6H6O2/c7-5-1-2-
6(8)4-3-5/h1-4,7-8H/d+
```

Mapping for each atom

```
2-1-1 <> 3-1-1
2-1-2 <> 3-1-2
2-1-3 <> 3-1-3
2-1-4 <> 3-1-4
2-1-5 <> 3-1-5
2-1-6 <> 3-1-6
2-1-7 <> 3-1-7
```

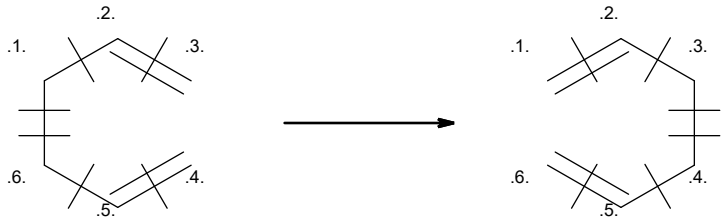
Skip the trailing 2 and 3 as they are defined by the reaction separator “<>”

```
1-1 <> 1-1
1-2 <> 1-2
1-3 <> 1-3
1-4 <> 1-4
1-5 <> 1-5
1-6 <> 1-6
1-7 <> 1-7
```

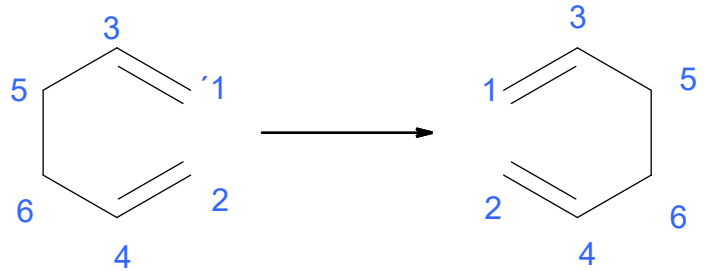
```
MapAuxInfo=1.00.1/1-1<>1-1;1-2<>1-2;1-
3<>1-3;1-4<> 1-4;1-5<>1-5;1-6<>1-6;1-
7<>1-7
```

Reaction representation

- Atom mapping
  - Cope rearrangement



(original atom and bond mapping)



Based on InChI atom numbering

Cope rearrangement

RInChI=1.00.1S/C6H10/c1-3-5-6-4-2/h3-4H,1-2,5-6H2<>C6H10/c1-3-5-6-4-2/h3-4H,1-2,5-6H2/d+

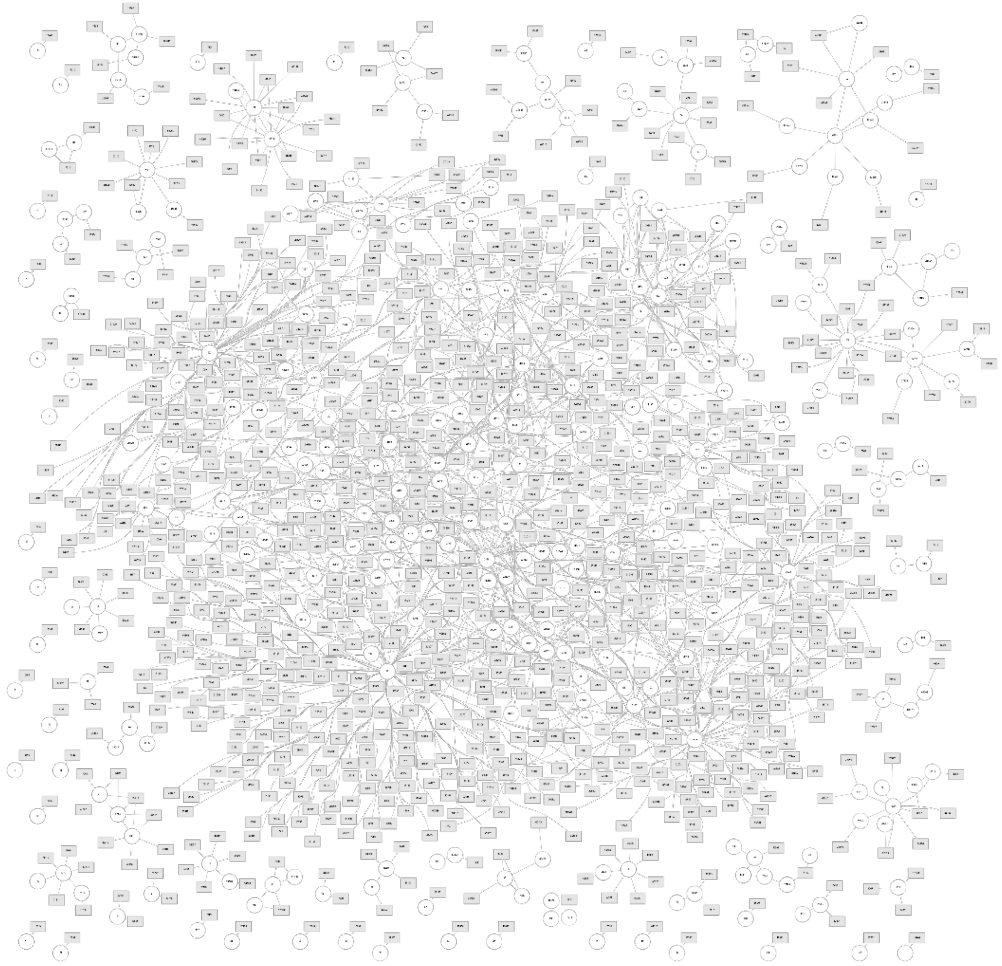
RInChI=1.00.1S/<><>C6H10/c1-3-5-6-4-2/h3-4H,1-2,5-6H2

Mapping for each atom

1-1 <> 1-5	1-4 <> 1-4
1-2 <> 1-6	1-5 <> 1-1
1-3 <> 1-3	1-6 <> 1-2

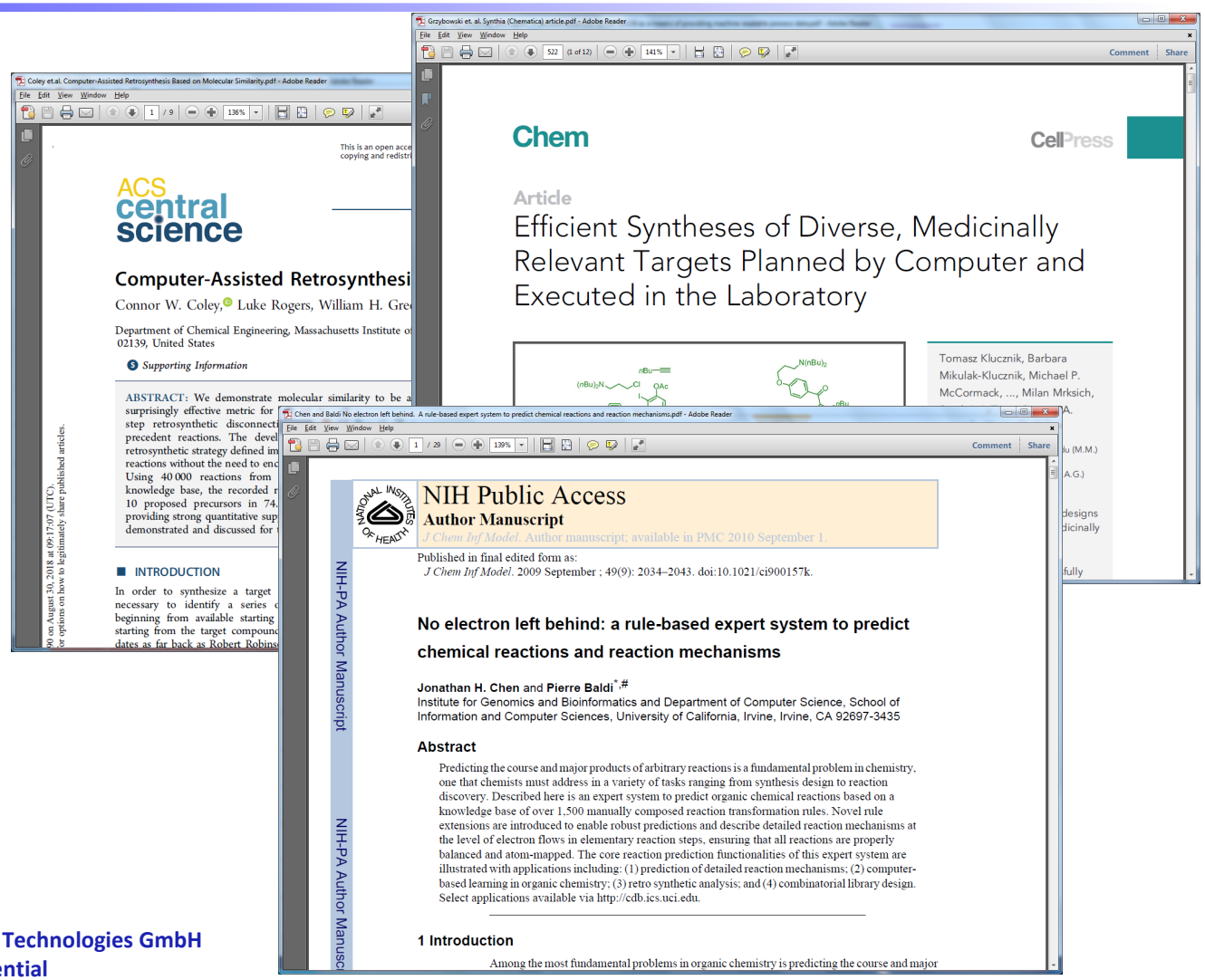
MapAuxInfo=1.00.1/1-1<>1-5;1-2<>1-6;1-3<>1-3;1-4 <>1-4;1-5<>1-1;1-6<>1-2

Make RInChI useable for AI



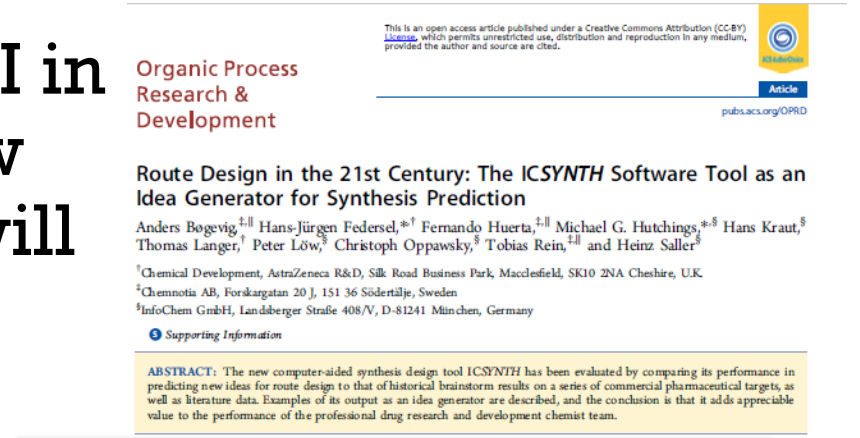
Can ML/AI be Applied to Predict Reactions?

- There has been a long history of software solutions aimed at "in-silico" route-finding.
- These have until now been regarded as predicting routes which are not as good as the practicing chemist would wish.
  - Though the "quality" has been steadily increasing.
- They typically embed tens of thousands of "chemistry rules", and apply these in a retrosynthetic search.



Can ML/AI be Applied to Predict Reactions?

- With the (over-hyped?) success of AI in many difficult problems, there is now the expectation that such methods will succeed where traditional methods (explicit chemistry rules) have not.
- Some initial results have been reported, including methods which equal or surpass chemists routes (in AB tests).
- In addition, the ML/AI community expects this to work!



Organic Process Research & Development

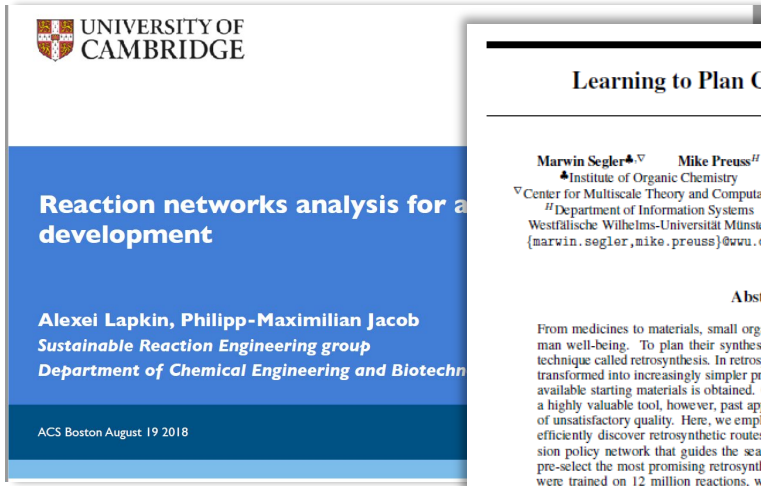
Article

pubs.acs.org/OPRD

Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction

Anders Bøgevig, Hans-Jürgen Federsel, Fernando Huerta, Michael G. Hutchings, Hans Kraut, Thomas Langer, Peter Löw, Christoph Oppawsky, Tobias Rein, and Heinz Saller

ABSTRACT: The new computer-aided synthesis design tool ICSYNTH has been evaluated by comparing its performance in predicting new ideas for route design to that of historical brainstorm results on a series of commercial pharmaceutical targets, as well as literature data. Examples of its output as an idea generator are described, and the conclusion is that it adds appreciable value to the performance of the professional drug research and development chemist team.



UNIVERSITY OF CAMBRIDGE

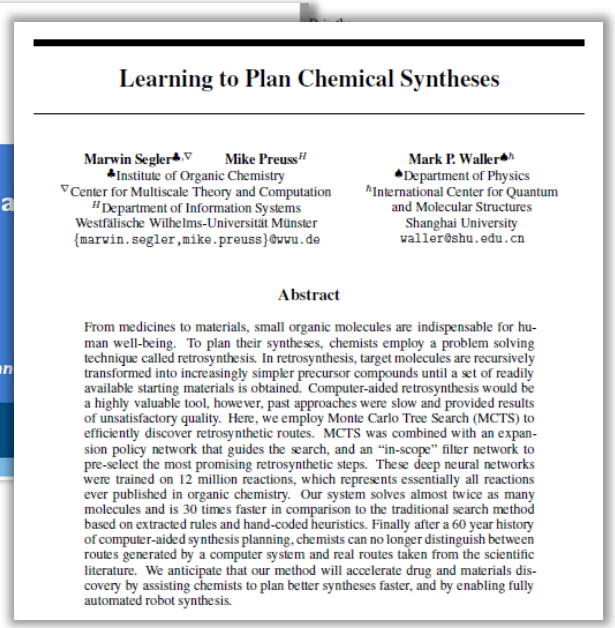
Reaction networks analysis for a development

Alexei Lapkin, Philipp-Maximilian Jacob

Sustainable Reaction Engineering group

Department of Chemical Engineering and Biotechn

ACS Boston August 19 2018



Learning to Plan Chemical Syntheses

Marwin Segler, Mike Preuss, Mark P. Waller

Institute of Organic Chemistry

Center for Multiscale Theory and Computation

Department of Information Systems

Westfälische Wilhelms-Universität Münster

{marwin.segler,mike.preuss}@wwu.de

Department of Physics

International Center for Quantum and Molecular Structures

Shanghai University

waller@shu.edu.cn

Abstract

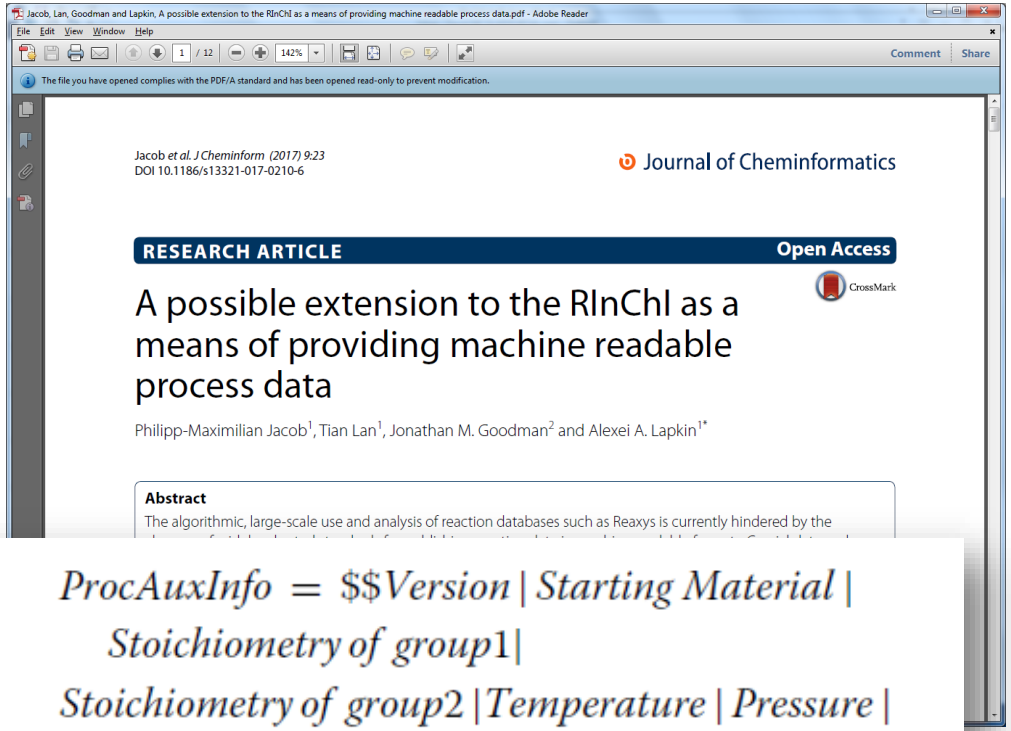
From medicines to materials, small organic molecules are indispensable for human well-being. To plan their syntheses, chemists employ a problem solving technique called retrosynthesis. In retrosynthesis, target molecules are recursively transformed into increasingly simpler precursor compounds until a set of readily available starting materials is obtained. Computer-aided retrosynthesis would be a highly valuable tool, however, past approaches were slow and provided results of unsatisfactory quality. Here, we employ Monte Carlo Tree Search (MCTS) to efficiently discover retrosynthetic routes. MCTS was combined with an expansion policy network that guides the search, and an "in-scope" filter network to pre-select the most promising retrosynthetic steps. These deep neural networks were trained on 12 million reactions, which represents essentially all reactions ever published in organic chemistry. Our system solves almost twice as many molecules and is 30 times faster in comparison to the traditional search method based on extracted rules and hand-coded heuristics. Finally after a 60 year history of computer-aided synthesis planning, chemists can no longer distinguish between routes generated by a computer system and real routes taken from the scientific literature. We anticipate that our method will accelerate drug and materials discovery by assisting chemists to plan better syntheses faster, and by enabling fully automated robot synthesis.

Newly planned additional information

- Mark failing reactions
  - Currently discussed: direction layer /d! or an additional identifier
- Additional AUXInfo layers for statistical tools provided by vendors/publishers
  - Class codes by Infochem
    - AuxClassCode=1.0/narrow code; medium code, broad code
  - Under discussion: Transfom by Reaxys
  - These AuxInfos can only be used if the vendor's product is licensed

Newly planned additional information Processing Information in "ProcAuxInfo"

- A proposal has been made for a "ProcAuxInfo" layer in a paper by authors from Cambridge University
  - Available online (open access) at https://jcheminf.springeropen.com/articles/10.1186/s13321-017-0210-6
- This is a rich format supporting (amongst other things):
  - Composition
  - Processing Conditions
  - Yields over time
- Together with sections of the MInChI format this is the base for the newly planned "ProcAuxInfo"


Jacob, Lan, Goodman and Lapkin, A possible extension to the RInChI as a means of providing machine readable process data.pdf - Adobe Reader

Jacob *et al. J Cheminform* (2017) 9:23
DOI 10.1186/s13321-017-0210-6

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

A possible extension to the RInChI as a means of providing machine readable process data

Philipp-Maximilian Jacob[1], Tian Lan[1], Jonathan M. Goodman[2] and Alexei A. Lapkin[1*]

Abstract

The algorithmic, large-scale use and analysis of reaction databases such as Reaxys is currently hindered by the

ProcAuxInfo = $$*Version* | *Starting Material* | *Stoichiometry of group*1| *Stoichiometry of group*2 | *Temperature* | *Pressure* | *Time* : *Conversion* | *Yield* | *Amount of group*1 *fed* | *Amount of group*2 *fed* | *Amount of group*3 *fed* | *Volume of reactor*

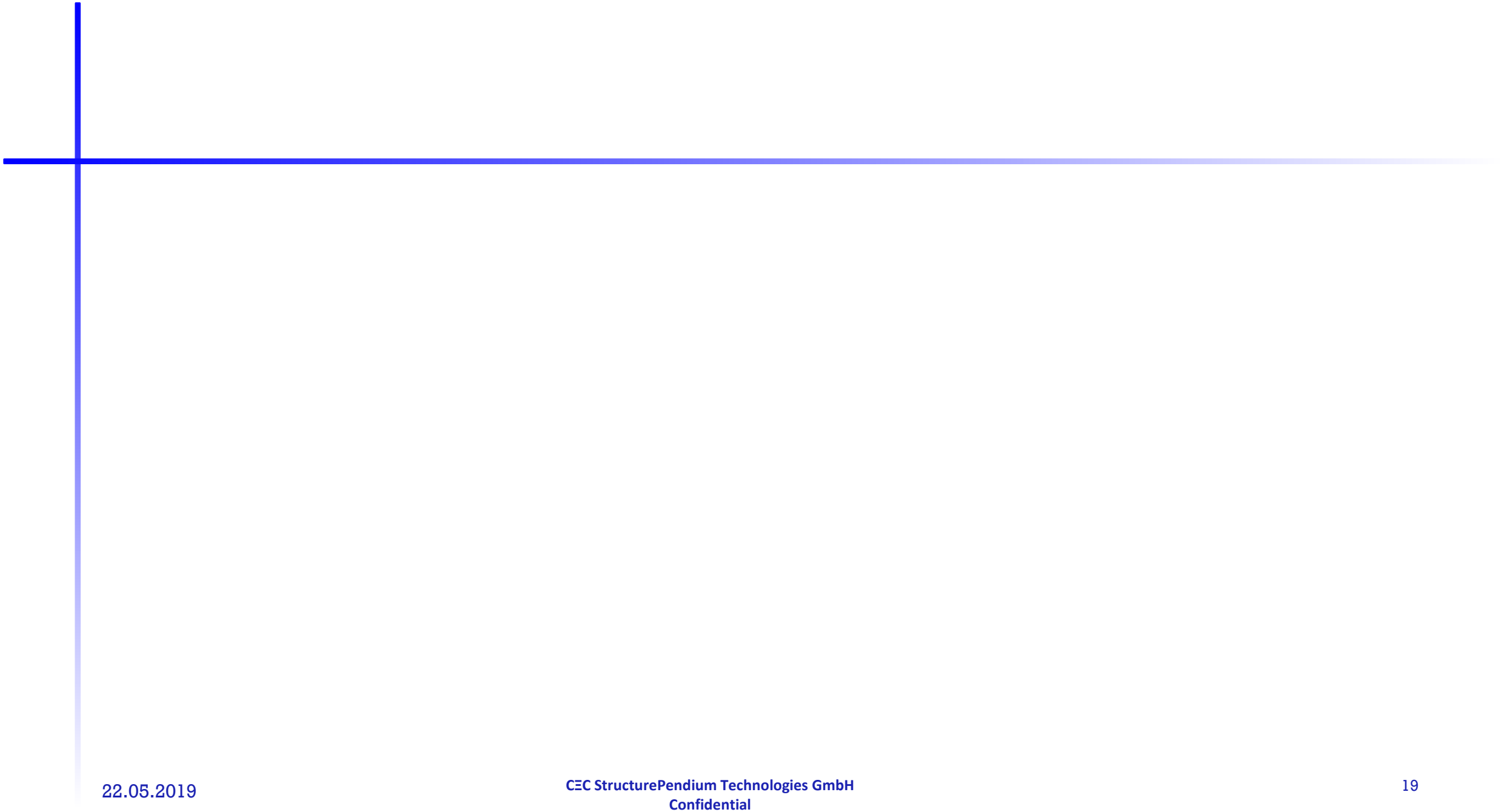
Deliverables, questions and remarks?

- Download (under open-source agreement)
  - http://www.inchi-trust.org/downloads/
- Answers may be found in
  - International chemical identifier for reactions (RInChI)

    G. Grethe, G. Blanke, H. Kraut and J. M. Goodman *J. Cheminformatics* 2018, **10**, 22. DOI: 10.1186/s13321-018-0277-8
- Else send your questions and remarks to
  - RInChI@StructurePendium.com

Thanks

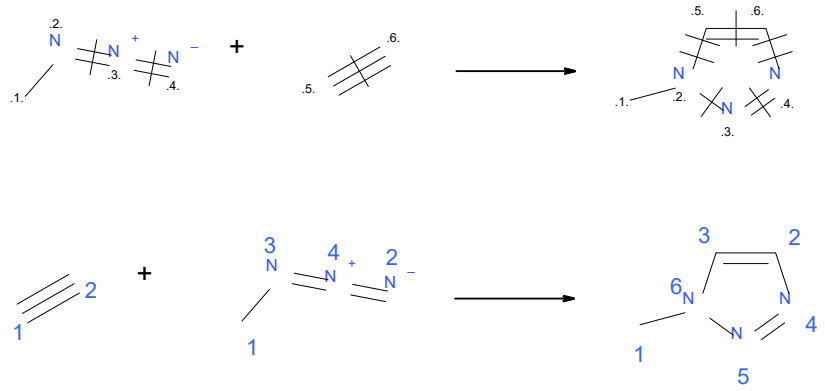
- InChI Trust, Cambridge
  - IUPAC Division VIII and IUPAC's Committee on Publications and Cheminformatics Data Standards (CPCDS)
- RInChI working group
  - David Nicolaides (Dassault Systèmes, Cambridge, UK)
  - Gerd Blanke (StructurePendium Technologies GmbH, Essen, Germany)
  - Günter Grethe,
  - Hans Kraut (InfoChem GmbH, Munich, Germany)
  - István Öri (ChemAxon Ltd, Budapest, Hungary)
  - Jan Holst Jensen (BioChemFusion AsP, Denmark)
  - Jonathan Goodman (University of Cambridge, UK)
- ChemAxon for the technical partnership and for this talk



Backups

Reaction representation

- Atom mapping
  - Example Klick reaction



Klick reaction

```
RInChI=1.00.1S/C2H2/c1-2/h1-2H!CH3N3/c1-3-
4-2/h1H3<>C3H5N3/c1-6-3-2-4-5-6/h2-
3H,1H3/d+
```

Mapping for each atom

- Note that both atoms of Acetylene are equivalent for the mapping process and must be represented by using the related bracket notation

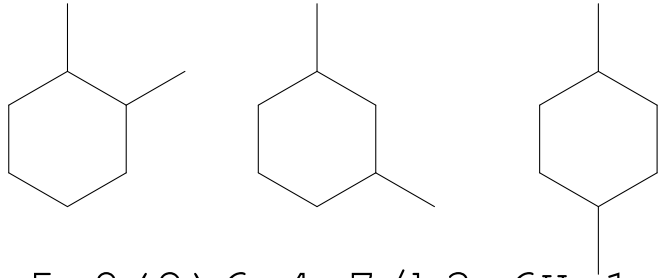
```
1-1 <> 1-(3,2)
1-2 <> 1-(2,3)
```

```
2-1 <> 1-1
2-2 <> 1-4
2-3 <> 1-6
2-4 <> 1-5
```

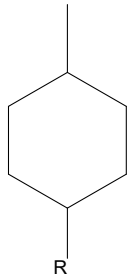
```
MapAuxInfo=1.00.1/1-1<>1-(3,2);1-2<>1-
(2,3);2-1<>1-1;2-2<>1-4;2-3<>1-6;2-
4<>1-5
```

Reaction representation

- Discussed potential enhancements based on discussed InChI improvements
  - Positional isomers



    - InChI=1S/C8H10/c1-7-3-5-8(2)6-4-7/h3-6H,1-2H3/pi-1,7(3,5)
  - Markush structures (Variable structures)



R = H, Me, Ph, SPh, OSiMe3

    - InChI=1S/C6H6R/c7-6-4-2-1-3-5-6/h1-5,7H/vs-f3,7H