International Chemical Identifier for Reactions (RInChI)

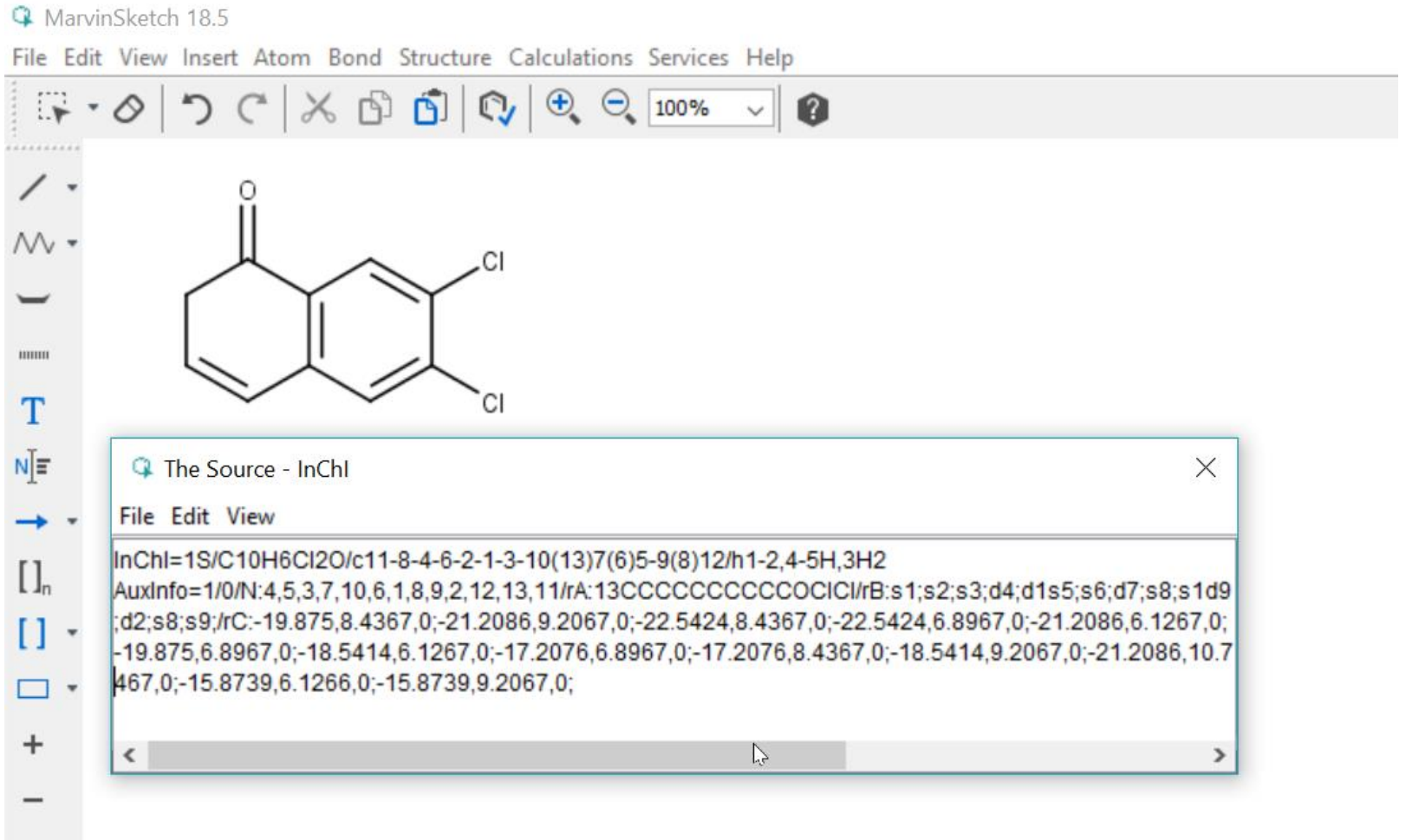
What is RInChI and how does it revolutionize the handling of reaction databases?

Gerd Blanke[1], Günter Grethe, Hans Kraut[2], Jonathan Goodman[3]

ChemAxon User Meeting Europe 2018, Budapest, March 21st, 2018

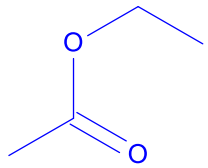
[1] CΞC StructurePendium Technologies GmbH, Reulsbergweg 5, D-45257 Essen, Germany
[2] InfoChem Gesellschaft für chemische Information mbH, Aschauer Str. 30, D-81549 Munich, Germany
[3] Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK

What about InChI



What about InChI

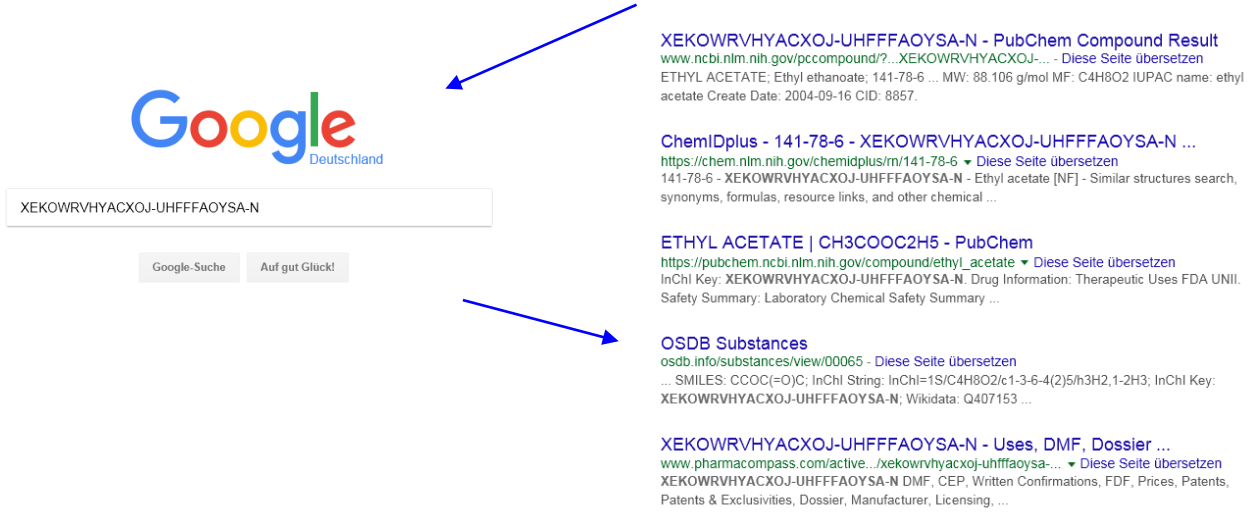
- InChI
  - IUPAC International Chemical identifier is a unique representation of a compound it describes.
  - Broadly used by e.g. PubChem, compound vendors, …
  - Current release 1.05 published in November 2016
  - The InChI format and algorithm are non-proprietary and the software is open source



InChI=1S/C4H8O2/c1-3-6-4(2)5/h3H2,1-2H3

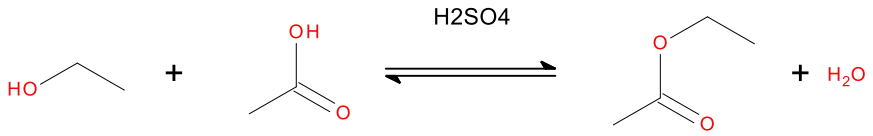
AuxInfo=1/0/N:5,1,4,2,6,3/rA:6CCOCCO/rB:s1;s2;s3;s4;d2;/rC:5.2468,-6.2392,0;6.2697,-5.6486,0;6.2697,-4.4671,0;7.2928,-3.8764,0;8.3159,-4.4671,0;7.2928,-6.2393,0;

InChIKey=XEKOWRVHYACXOJ-UHFFFAOYSA-N

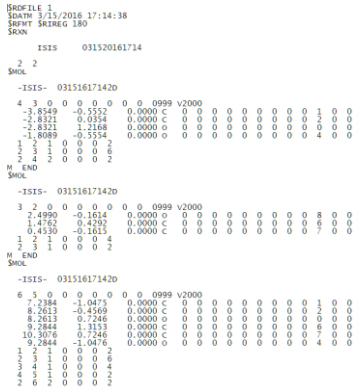


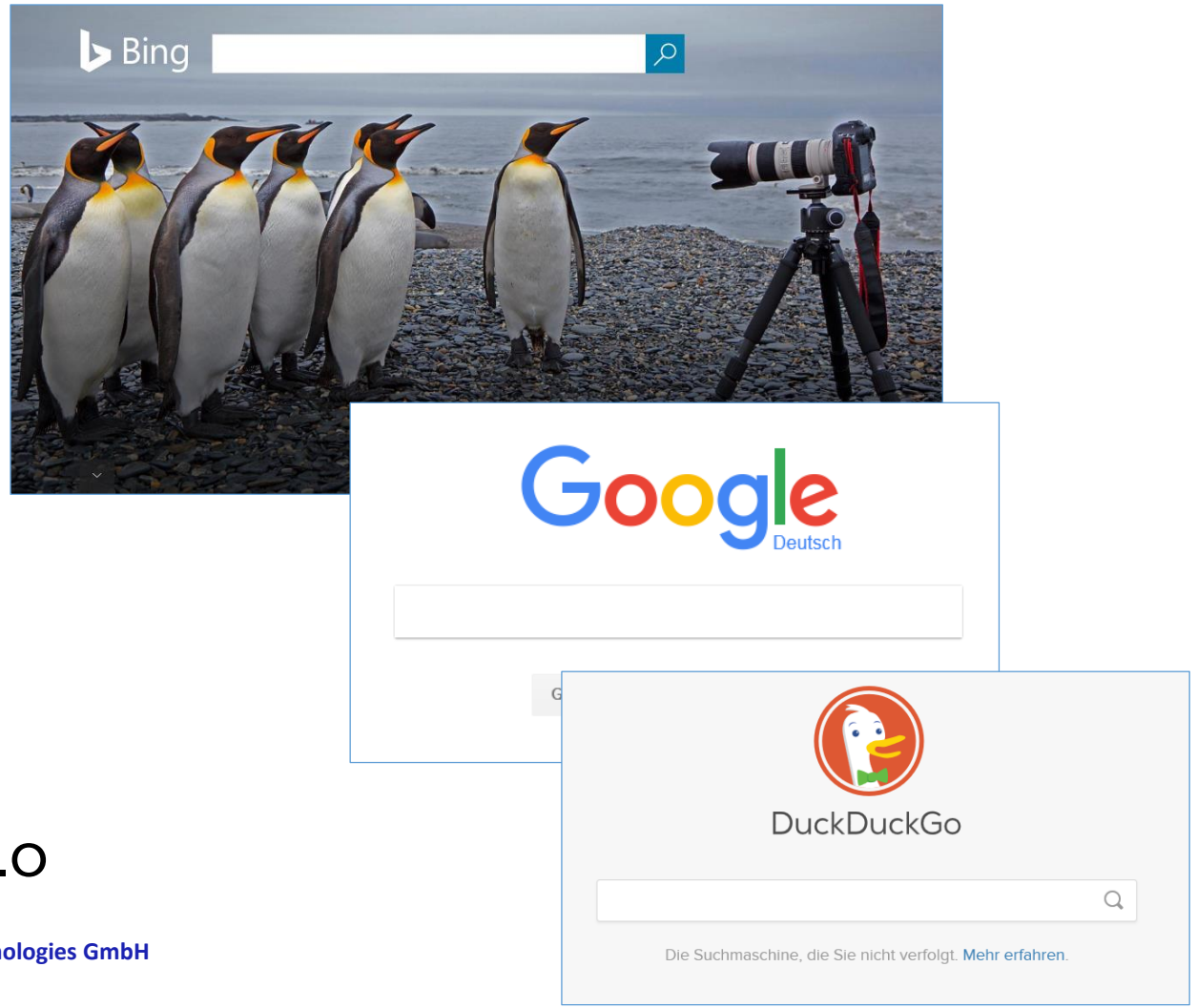
What do we need for reactions …

To run searches in the Web?



- RXN/RD files





- Reaction smiles:

CCO.CC(=O)O>OS(=O)(=O)O>CCOC(=O)C.O

What do we need for reactions …

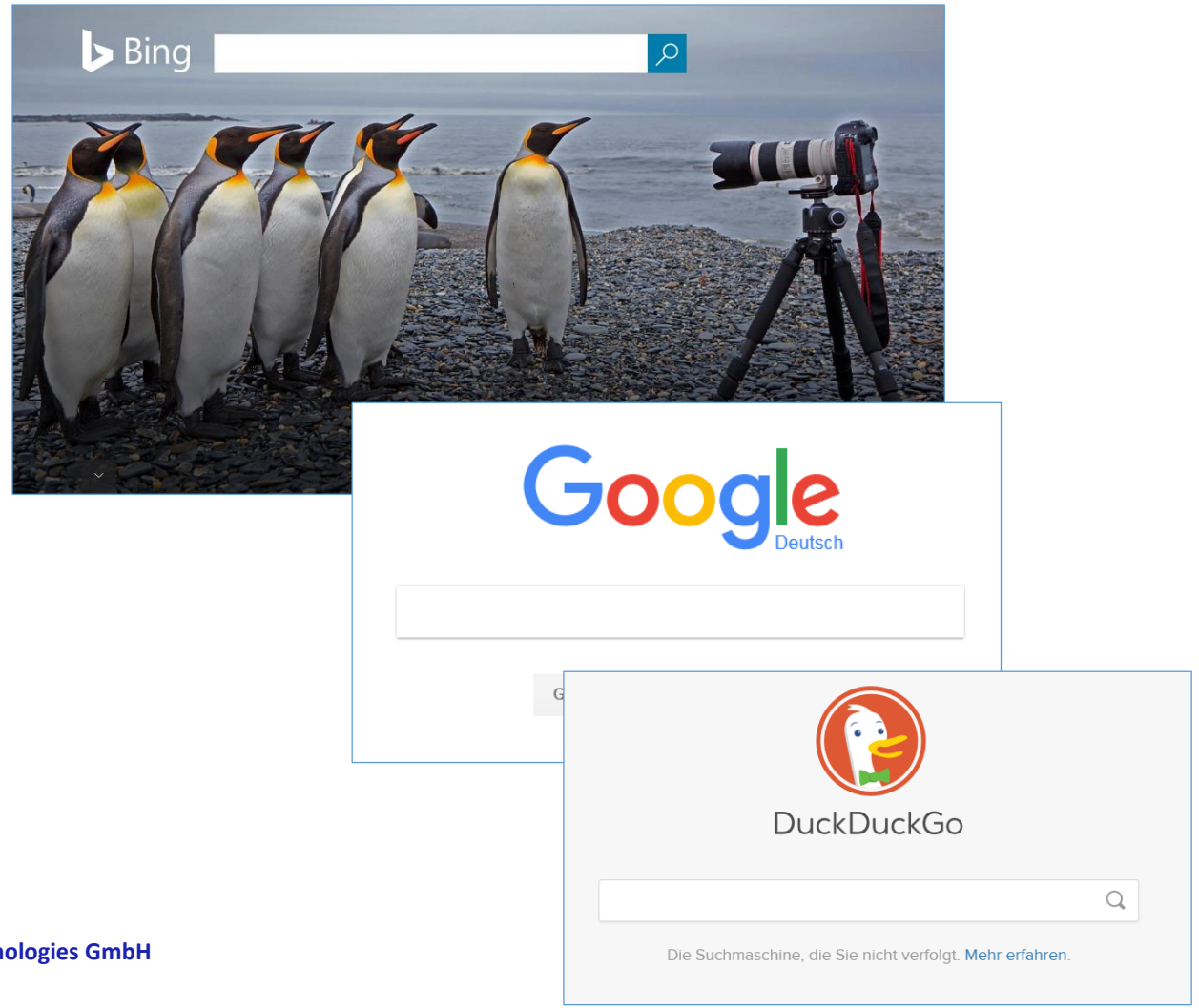
To run searches in the Web?

A single string providing a unique, i.e. canonicalized, representation of a reaction

IUPAC International Chemical Identifier for Reactions

Reaction InChI

(RInChI)



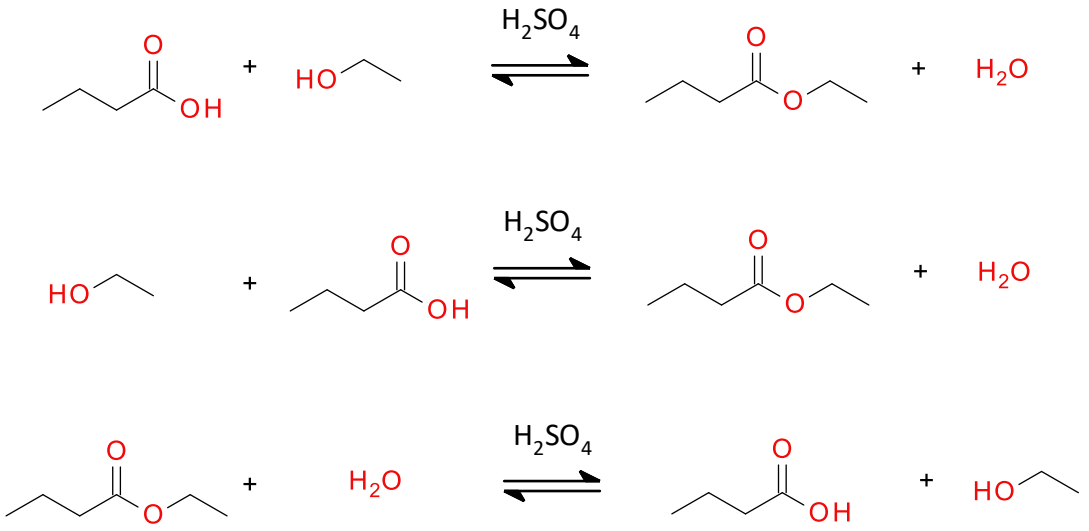
Introduction to RInChI

- 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.

Introduction to RInChI

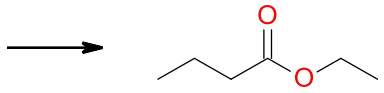
- Format:

RInChI=1.00.1S/*layer2*<>*layer3*<>*layer4*/*d*(+,-,=)/*u*#2-#3-#4



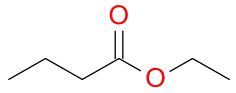
RInChI=1.00.1S/C2H6O/c1-2-3/h3H,2H2,1H3!C4H8O2/c1-2-3-4(5)6/h2-3H2,1H3,(H,5,6)<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3!H2O/h1H2<>H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)/d=

Half-reactions and reactions with No-Structures



- Half reactions like the one above where only reactant(s) or products(s) are given follow the rules above with the assumption that the missing group is represented by a null string and the rule that empty strings are not added to the RInChI:

RInChI=1.00.1S/<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3/d+

No Structure 

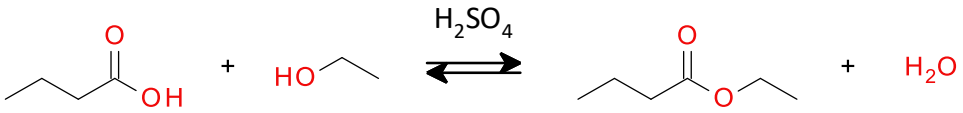
- No-structures are ignored within their group but are added as an additional flag at the right end of the RInChI string using the format u#1-#2-#A with # as integer representing the total number of No-Structures in group1, group2 and the agent group as 3rd group.

RInChI=1.00.1S/<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3/d+/u1-0-0

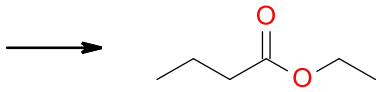
Introduction to Long-RInChIKey

- Format

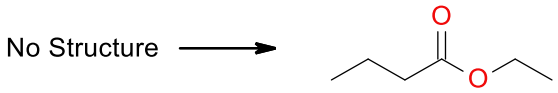
Long-RInChIKey=SA-(F,B,E,U)UHFF-layer3--layer4--layer5 and (F,B,E,U) = F, B, E, or U



Long-RInChIKey=SA-EUHFF-LFQSCWFLJHTTHZ-UHFFFAOYSA-N-FERIUCNNQQJTOY-UHFFFAOYSA-N--OBNCKNCVKJNDBV-UHFFFAOYSA-N-XLYOFNOQVPJJNP-UHFFFAOYSA-N--QAOWNCQODCNURD-UHFFFAOYSA-N



Long-RInChIKey=SA-FUHFF---OBNCKNCVKJNDBV-UHFFFAOYSA-N

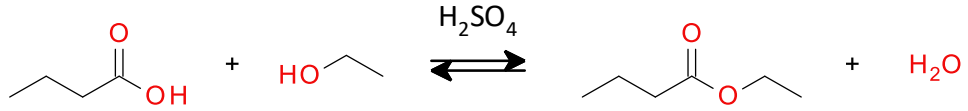


Long-RInChIKey=SA-FUHFF-MOSFIJXAXDLOML-UHFFFAOYSA-N--OBNCKNCVKJNDBV-UHFFFAOYSA-N

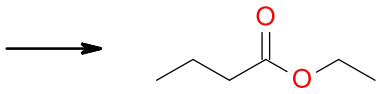
Introduction to Short-RInChIKeys

- Format

Short-RInChIKey=SA-(F,B,E,U)UHFF - hash over all major layers in RInChI 2 – hash over all major layers in RInChI 3 – hash over all major layer in RInChI 4 – hash over all minor layers plus sum of protonation states in RInChI 2 – hash over all minor layers plus sum of protonation states in RInChI 3 – hash over all minor layers plus sum of protonation states in RInChI 4 – ### with # = Z, A, B…

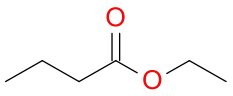


Short-RInChIKey=SA-FUHFF-JEVIJXCZCL-UFTQDZUCXS-QAOWNCQODC-NUHFF-NUHFF-NUHFF-ZZZ



Short-RInChIKey=SA-FUHFF-UHFFFADPSC-OBNCKNCVKJ-UHFFFADPSC-NUHFF-NUHFF-NUHFF-ZZZ

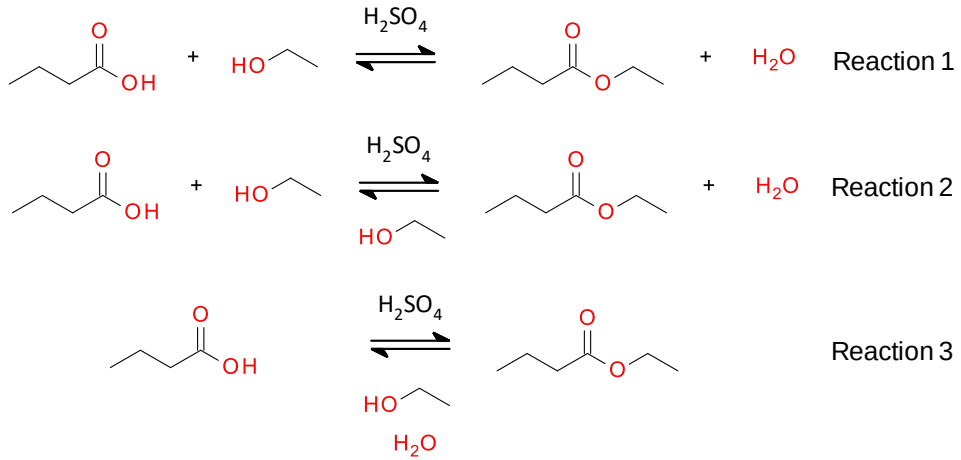
No Structure →



Short-RInChIKey=SA-FUHFF-UHFFFADPSC-OBNCKNCVKJ-UHFFFADPSC-NUHFF-NUHFF-NUHFF-AZZ

Introduction to Web-RInChIKey

- The depiction of a chemical reaction is not uniquely defined

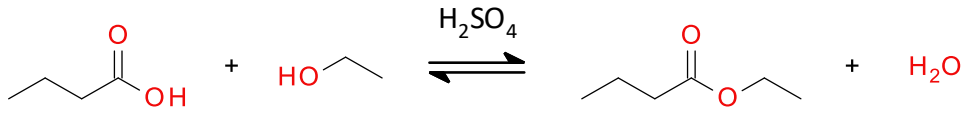


  - Especially the role of the components may differ depending on the rules used by the authors
  - To compare reactions from different sources the more flexible Web-RInChIKey has been defined

- Format

Web-RInChIKey=major layer-minor layer

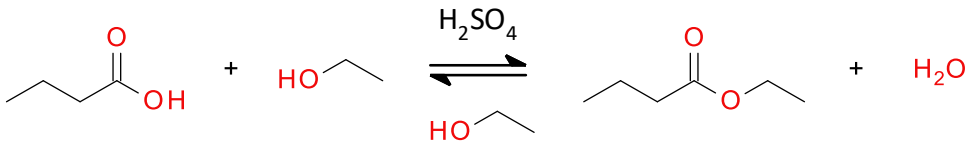
Web-RInChIKeys
Examples



Reaction 1

Short-RInChIKey=SA-EUHFF-JEVIJXCZCL-UFTQDZUCXS-QAOWNCQODC-NUHFF-NUHFF-NUHFF-ZZZ

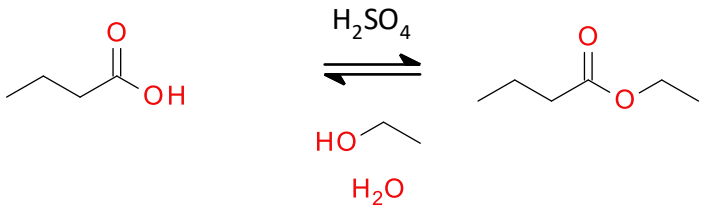
Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA



Reaction 2

Short-RInChIKey=SA-EUHFF-JEVIJXCZCL-UFTQDZUCXS-UAUFKIWNBD-NUHFF-NUHFF-NUHFF-ZZZ

Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA



Reaction 3

Short-RInChIKey=SA-EUHFF-FERIUCNNQQ-OBNCKNCVKJ-DNBJJWMYJT-NUHFF-NUHFF-NUHFF-ZZZ

Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA

CΞC StructurePendium Technologies GmbH
Confidential

RInChI (usage)

- RInChI is the only index allowing the comparison of reactions including the related agents (catalysts, solvents, etc.)
  - Uniqueness identifications (by Short-RInCIKeys)
  - Identifications of duplicates based on different reaction representation rules or reaction database models (by Web-RInChIKeys)
- Identification of equilibrium reactions within reaction databases
- Identification of reaction components based on InChIKey searches (based on Long-RInChIKeys) or on RInChIs based on InChIs
- Reaction Web-searches

Input formats

- RXN files
- RD files
  - Hierarchical
    - traditional MDL REACCS format
  - Flat
    - Mainly used in the context of chemical electronic labjournals
    - Simplifies the work with PP component
- Molfiles with component roles

CEC StructurePendium Technologies GmbH
Confidential

Status of RInChI development

- First RInChI release in March 2017
- RInChI team internal Pipeline Pilot RInChI test environment
- Used within "Synthetically Accessible Virtual Inventory" (SAVI), NCI, Bethesda MD
  - Fully integrated into CACTVS tool kit by Xemistry
- RInChIs for 4.5 million reaction of SPRESI generated by InfoChem (Hans Kraut and colleagues)
  - SPRESI is a chemical structure and reaction database that includes 5.52 million structures, 4.26 million reactions and 32 million factual data entries extracted from 675,000 references and 164,000 patents with the major focus to organic chemistry built since 1970.
  - Results
    - 239 reactions could not be converted.
    - 4,564,718 reactions with RInChIs and RInChIKeys

Deliverables

- Deliverables since March 2017
  - RInChI libraries for Linux and Windows
    - MAC version needs work up
  - Application example
    - RInChI-cmdline tool
    - Example application to use molfiles with reaction roles.
- Download (under open-source agreement)
  - http://www.inchi-trust.org/downloads/

Deliverables

- Test Web-site provided by University of Cambridge, UK
  - http://www-rinchi.ch.cam.ac.uk/
- Development Web-Site
  - http://proteax.dk/rinchi/rinchi.py/demo
- Additional tools built for RInChI testing
  - Oracle cartridge for RInChIs based on Oracle 11
  - Pipeline Pilot components
  - The tools can be made available on consulting base by StructurePendium Technologies GmbH

Thanks

- InChI Trust, Cambridge
  - IUPAC Division VIII and IUPAC's Committee on Publications and Cheminformatics Data Standards (CPCDS)
- RInChI working group
  - Günter Grethe
  - Jonathan Goodman (University of Cambridge, UK)
  - Hans Kraut (InfoChem GmbH, Munich, Germany)
  - Gerd Blanke (StructurePendium Technologies GmbH, Essen, Germany)
  - Jan Holst Jensen, Colin Bachelor, Keith Taylor, Martin Schmidt, Sandy Lawson

Questions and remarks?

- Send to
  - RInChI@StructurePendium.com