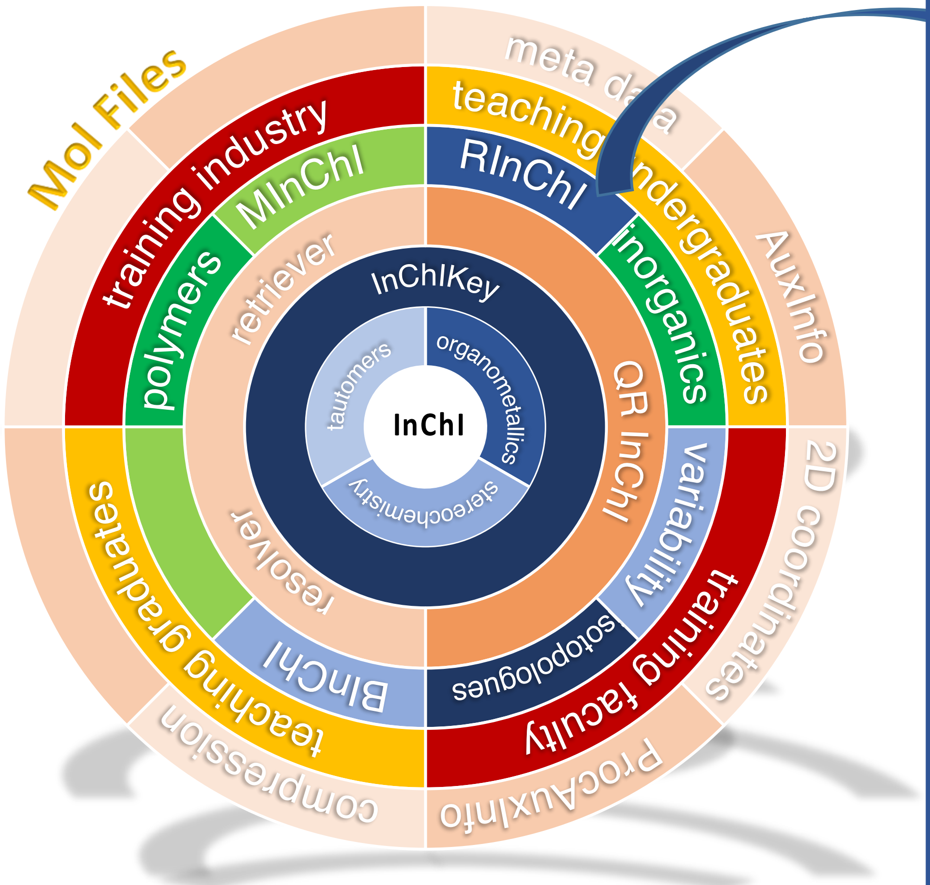
RInChI – What is next?

NIH InChI workshop March 22-24, 2021

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

[1] StructurePendium Technologies GmbH, Essen, Germany; [2] San Diego, CA, US; [3] InfoChem GmbH, Munich, Germany; [4] ChemAxon, Budapest Hungary; [5] Biochemfusion AsP, Copenhagen, Denmark; [6] University of Cambridge, Department of Chemistry, Cambridge, UK

The International Chemical Identifier for Reactions (RInChI) in the InChI Ecosystem



Canonical Identifier for Reactions

with minimal data

RInChI

auxiliary Information

ProcAuxInfo

Reactions with unexpected results

atom mapping

MapAuxInfo

Stereochemistry

layer

InChI upgrades will drop in (eg extended stereochemistry)

RInChI – What it is

Agents

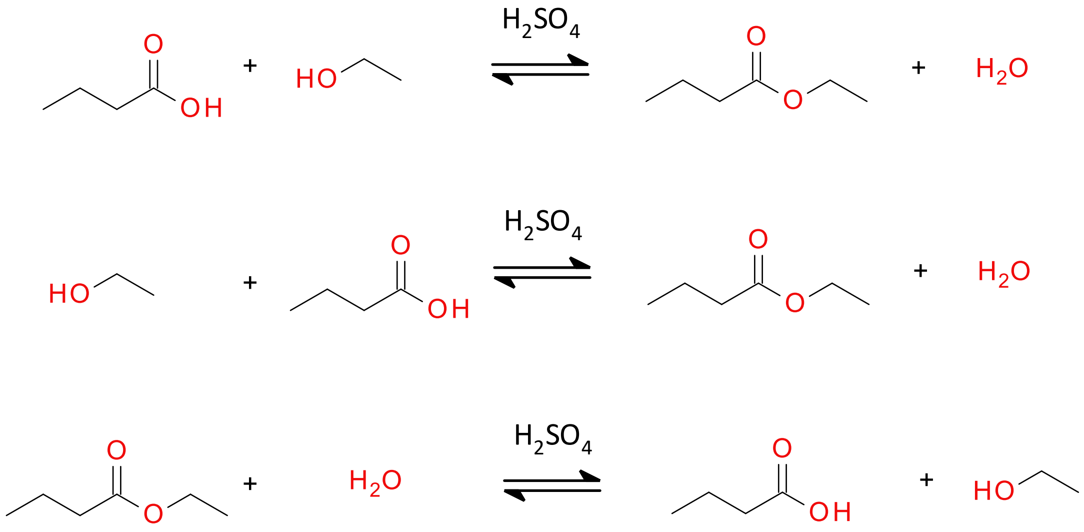
solvents, catalysts ...

Reactants → Products

- The RInChI format is a hierarchical, layered, unique description of a reaction with different levels based on the Standard InChI representation (version 1.04) of each structural component participating in a reaction.
  - RInChI
    - The RInChI is calculated from the InChIs of each reactant, product and agent
  - 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 – What it is

- Example: Esterification, equilibrium reaction



- 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=

- RInChI is the unique identifier for chemical reactions

RInChI – What it is

- Example: Esterification, equilibrium reaction
  - Long-RInChIKey

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

    - Long-RInChIKeys are a valuable tool for the database storage of reactions. Beside uniqueness checks, they allow the identification of each reaction component by simple text searches based on Standard InChIKeys
      - Any molecule within a reaction can be identified by its InChIKey.
        - That allows exact molecule searches based on text searches for InChIs within any reaction databases containing the long RInChIKey.
      - The InChIs from the long RInChIKey can be easily used to build synthesis trees
    - The long RInChIKey does not provide a fixed length.

RInChI – What it is

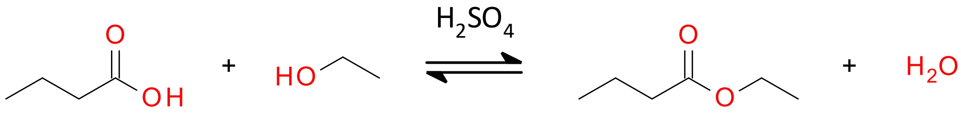
- Example: Esterification, equilibrium reaction
  - Short-RInChIKey

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

    - The Short-RInChIKey has a fixed length of 55 letters plus 8 hyphens as separators resulting in a total of 63 characters.
    - The fixed length of Short-RInChIKey makes it suitable for
      - exact searches of reactions in databases (and in the WEB)
      - indexing reactions in databases
      - linking identical reactions in different databases.

RInChI – What it is

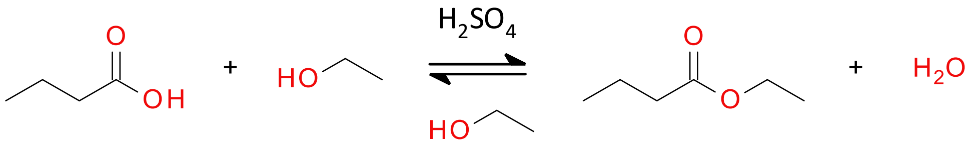
- Example: Esterification, equilibrium reaction
  - Web-RInChIKey



Reaction 1

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

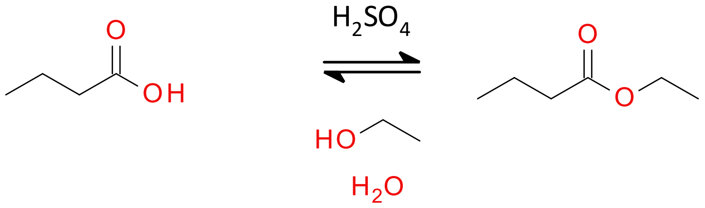
Web-RInChIKey=UTLWRJSGXVLTKYLGZ-NUHFFFADPSCTJSA



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

RInChI – What it is

- Example: Esterification, equilibrium reaction
  - Web-RInChIKey (continued)
    - The Web-RInChIKey has a fixed length of 47 characters with 17 letters in the major and 15 letters in the minor layer.
    - The Web-RInChIKey is a component-based representation of reactions that is agnostic about the role of any component within a reaction and the direction of the reaction. Reactions with identical components but different roles are represented by identical Web-RInChIKeys.
      - The results of the searches may have to be refined in a second step to filter out the inappropriate hits.
    - Intended usage
      - Searches over reaction databases with unknown drawing model
        - Especially web seraches
      - Comparison of reaction databases with different drawing models

RInChI – What it is

- RInChI 1.0 released in March 2017
- Publication: International chemical identifier for reactions (RInChI), Grethe *et al.* J Cheminform (2018) 10:22 (May 2018)
  - https://doi.org/10.1186/s13321-018-0277-8
- First publisher introducing RInChI
  - Beilstein Institute, Frankfurt (Main), Germany
- RInChI implementations by software vendors
  - The Biovia software packages Draw, Direct and Pipeline Pilot (version 2019, released in December 2018) include RInChI
  - ChemAxon Marvin editor, released 2019
  - Knime nodes by Lhasa Ltd, UK, 2019

RInChI – What it is

- Known projects
  - EMBL reaction database, 2019
  - SAVI, NIH
  - Ontochem
  - Elsevier Entellect Reaction workbench, released August 2020
  - Tested in industry projects

RInChI – What' next

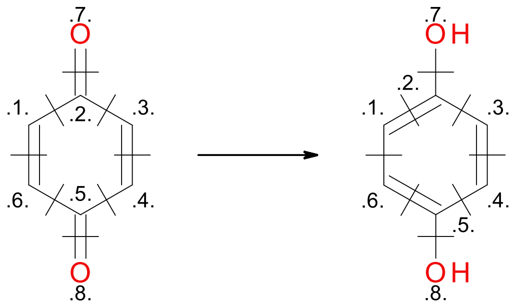
- In summer 2018 the RInChI group has started to work on the next release (RInChI 2.0).
- Addressed issues
  - Technical enhancements
    - Especially upgrade to the latest InChI version 1.06
  - Additional input and output formats
  - Reaction mapping (MapAuxInfo)
  - Workarounds for stereochemistry and tautomer restrictions
  - Failing reactions
  - Address needs for big data analysis methods (under investigation)
  - Move to GitHub

RInChI – What' next

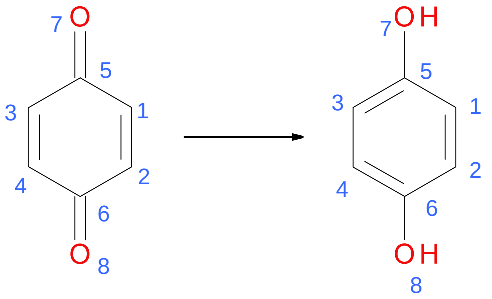
- Atom mapping
    - Atom mapping is used to mark the reaction centers
    - Atom mapping is handled as aux(iliary) layer
        - Layer name: MapAuxInfo, version 1.00.1
            - MapAuxInfo=1.00.1/
    - We are not implementing a mapping algorithm into RInChI but use the information provided by 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.
    - The RInChI representation of the atom mapping is canonical!

RInChI – What’ next

- Atom mapping: Example Quinone reduction



Mapping as defined in RXN file



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 (including layer and molecule)

```
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
```

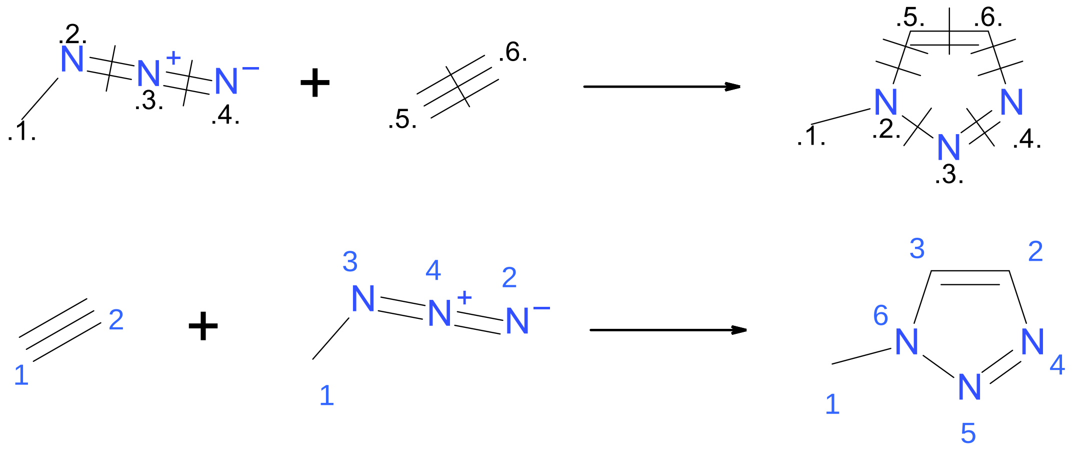
Skip the trailing 2 and 3 as they are defined by the reaction separator “<>” and the 4th layer is not available

```
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
```

RInChI – What' next

- Atom mapping: Example Click reaction



Click 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)

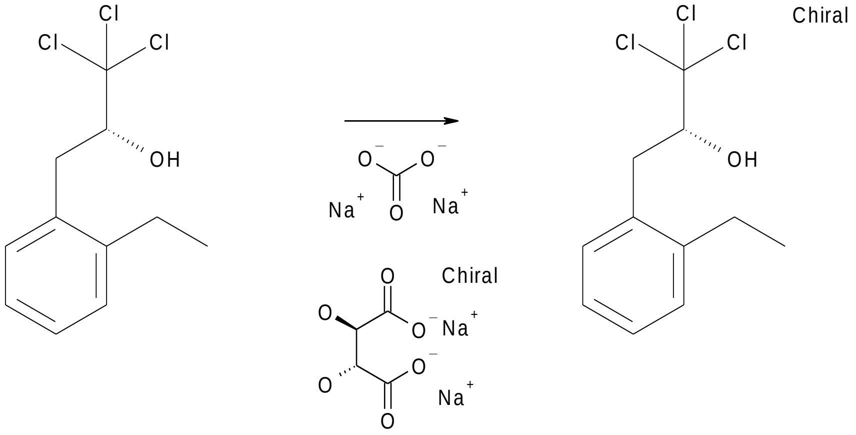
2-1 <> 1-1
2-2 <> 1-4
2-3 <> 1-6
2-4 <> 1-5

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

Note: The equivalence of the 2 acetylene atoms cannot be taken from the atom mapping in the RXN file.

RInChI – What' next

- Stereochemistry
  - The stereochemistry settings for the standard InChI are not sufficient to represent the stereochemistry of reactions
    - Standard InChI uses the parameter /SAbs (all stereo centers are absolute)
      - How can we represent (racemic) mixtures of stereoisomers that are the starting materials or products of a reaction with the absolutely known isomer?
        - Currently RInChI delivers A -> A

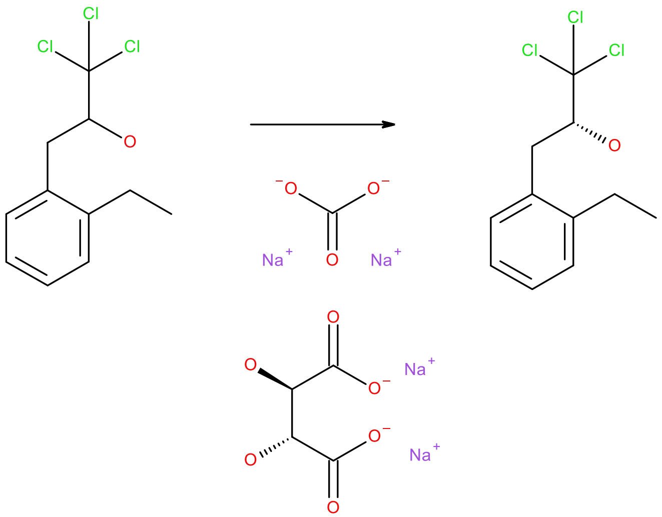


RInChI=1.00.1S/C11H13Cl3O/c1-2-8-5-3-4-6-9(8)7-10(15)11(12,13)14/h3-6,10,15H,2,7H2,1H3/t10-/m1/s1<>C11H13Cl3O/c1-2-8-5-3-4-6-9(8)7-10(15)11(12,13)14/h3-6,10,15H,2,7H2,1H3/t10-/m1/s1<>CH2O3.2Na/c2-1(3)4;;/h(H2,2,3,4);;/q;2*+1/p-2!C4H6O6.2Na/c5-1(3(7)8)2(6)4(9)10;;/h1-2,5-6H,(H,7,8)(H,9,10);;/q;2*+1/p-2/t1-,2-;;/m1../s1/d+

Total Synthesis of Cinchona Aalkaloids. 3. Synhesis of Quinuclidine Intermediates, G. Grethe, H.L.Lee, T. Mitt, M.R.Uskokovic, JACS, 100, 581, 1978

RInChI – What' next

- Stereochemistry rules
  - The structure representation is based on standard InChI
    - All marked tetrahedral stereocenters are represented as absolute centres.
  - Racemic centres without any relationship to other centres are represented by single bonds.



RInChI=1.00.1S/C11H13Cl3O/c1-2-8-5-3-4-6-9(8)7-
10(15)11(12,13)14/h3-6,10,15H,2,7H2,1H3<>
C11H13Cl3O/c1-2-8-5-3-4-6-9(8)7-10(15)11(12,13)14
/h3-6,10,15H,2,7H2,1H3/t10-/m1/s1<>CH2O3.2Na/c2-
1(3)4;;/h(H2,2,3,4);;/q;2*+1/p-2!C4H6O6.2Na/c5-
1(3(7)8)2(6)4(9)10;;/h1-2,5-6H,(H,7,8)(H,9,10);;/q;2*+1
/p-2/t1-,2-;;/m1../s1/d+

RInChI – What' next

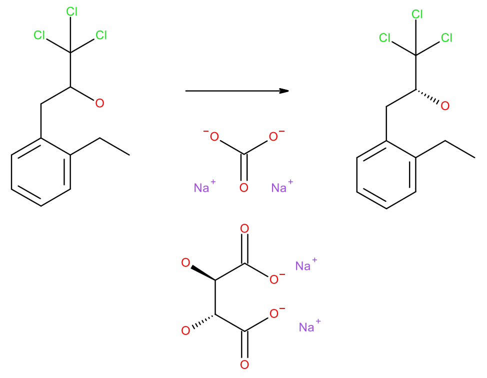
- Stereochemistry
  - To compensate the missing stereochemistry information the additional stereo layer "/st" is added at the end of the RInChI string
  - Racemic centres without any relationship to other centres are represented by single bonds.
  - Syntax of the additional stereo layer /st
    - Identification of the stereocenters (if needed)
      - The stereocenters are identified by the layer the related molecule belongs to, the order number of the molecule within the layer and the InChI number of the atom within the molecule (see rules for atom mapping as well)
    - Following attributes are used to characterize the stereocenters in the stereo layer /st
      - "mix" to indicate mixtures (analog AND# case)
      - "rac" to indicate racemates (as special mixture)
      - "pu" for pure but unknown (analog OR# case)
      - Note: Absolutely known centers are not marked in the stereo layer as they are identified by the standard InChI already

RInChI – What’ next

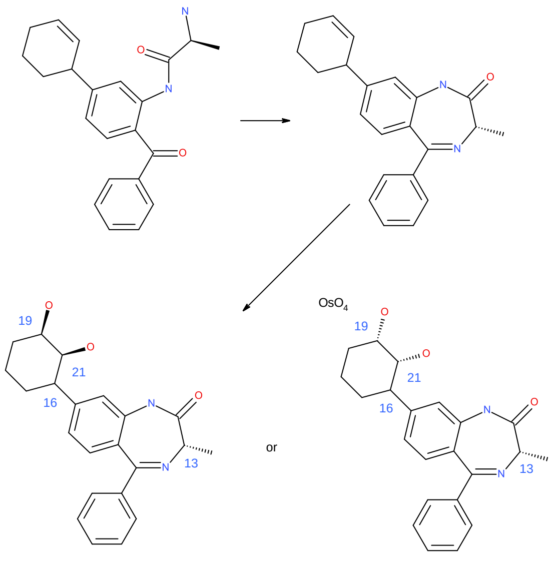
- Stereochemistry
  - Syntax of the additional stereo layer /st (continued)
    - If necessary, brackets are used to group related stereocenters, each group element is separated by commas
      - In case of brackets, set the stereo attribute to the closing bracket
      - If necessary, multiple brackets must be used to fully group the stereochemistry representation
      - Note: the grouping by brackets corresponds with the numbering schema for AND# and OR# in the representation of the enhanced stereochemistry
  - Note:
    - The RInChI stereo layer may change if the enhanced stereochemistry is introduced into InChI.

RInChI – What’ next

- Examples



RInChI=1.00.1S/C11H13Cl3O/c1-2-8-5-3-4-6-9(8)7-10(15)11(12,13)14
/h3-6,10,15H,2,7H2,1H3<> C11H13Cl3O/c1-2-8-5-3-4-6-9(8)7-10(15)
11(12,13)14/h3-6,10,15H,2,7H2,1H3<> CH2O3.2Na/c2-1(3)4;;
/h(H2,2,3,4);;/q;2*+1/p-2!C4H6O6.2Na/c5-1(3(7)8)2(6)4(9)10;;/h1-
2,5-6H,(H,7,8)(H,9,10);;/q;2*+1 /p-2/t1-,2-;;/m1../s1/d+



InChI=
1S/C22H24N2O3/c1-13-22(27)24-18-12-15(16-8-5-9-19
(25)21(16)26)10-11-17(18)20(23-13)14-6-3-2-4-7-14/h2-
4,6-7,10-13,16,19,21,25-26H,5,8-9H2,1H3,(H,24,27)/t13-
,16?,19+,21-/m0/s1

InChI=
1S/C22H24N2O3/c1-13-22(27)24-18-12-15(16-8-5-9-19
(25)21(16)26)10-11-17(18)20(23-13)14-6-3-2-4-7-14/h2-
4,6-7,10-13,16,19,21,25-26H,5,8-9H2,1H3,(H,24,27)/t13-
,16?,19-,21+/m0/s1

RInChI=1.00.1S/C22H22N2O/c1-15-22(25)24-20-14-18(16-8-4-2-5-9-
16)12-13-19(20)21(23-15)17-10-6-3-7-11-17/h3-4,6-8,10-
16H,2,5,9H2,1H3,(H,24,25)/t15-,16?/m0/s1<> C22H24N2O3/c1-13-
22(27)24-18-12-15(16-8-5-9-19(25)21(16)26)10-11-17(18)20(23-
13)14-6-3-2-4-7-14/h2-4,6-7,10-13,16,19,21,25-26H,5,8-
9H2,1H3,(H,24,27)/t13-,16?,**19+,21-**/m0/s1/d+ **/st(3-1-19,3-1-21)pu**

RInChI – What' next

- Reaction data / ProcAuxInfo, failing reaction
  - Original introduction of ProcAuxInfo
    - Purpose: handle the optimization of flow chemistry experiments
    - DOI: 10.26434/chemrxiv.6954908_Philipp-Maximilian_Jacob (03-Feb-2019)
  - Goal for RInChI
    - Transport the non-structure related reaction data in a format that is easily machine readable/writeable.
      - Format reaction data in a way that they can be immediately consumed by AI/ML methods.
      - Format reaction data in a way that they can feed synthesis robots
        - Taken from a discussion with AstraZeneca in August 2020
  - Instead of using the original (M)InChI like notation format as originally proposed for the ProcAuxInfo a JSON format is developed
    - JSON reader are available in most of the programming languages and data processing tools

RInChI – What' next

- Reaction data / ProcAuxInfo, failing reaction
  - Classification of reaction data
    - Reaction related data
      - Data related to the entire reaction like protecting atmosphere, total yield, reaction vessel
        - Literature reaction databases provide reaction data mostly on the "entire reaction level" only
      - Timepoint depending data like temperature, pressure, etc.
        - Transpose time depending data as a summary to the "literature level"
    - Component related data
      - Data related to each of the components (e.g. melting point, color)
      - Timepoint depending data like concentration, relationship to phases of the reaction
        - Note: literature data are mainly summarized over the timepoint values
    - Work-up procedures for products

RInChI – What’ next

- Reaction data / ProcAuxInfo, failing reaction
  - The ProcAuxInfo layer begins with
    - **/ProcAuxInfo=1.00/{.....}**
      - 1.00 is related to the ProcAuxInfo version (first version = 1.00)
      - {.....} to define the JSON string
      - The assembly of the ProcAuxInfo is not relying on InChI
        - No reference to InChI version
    - Failing reactions
      - A reaction fails not becaue of the RInChI as such but because of the conditions the reaction is run under
        - The failing Reaction flag must be reaction condition related
      - Use an “X” to mark a reaction as failing
      - **ProcAuxInfo=1.00(X)/{.....}**
    - Notes:
      - It may become necessary to reference any related ontology system and the version used in the ProcAuxInfo layer information

RInChI – What' next

- ProcAuxInfo – Concatenation rules
    - Each of the component is identified by its position in the RInChI layer (similar to the atom mapping and stereochemistry handling)
    - The JSON is set up out of the components described above using the following concatenation rules
        1. ProcAuxInfo Version and failure information
        2. General component properties
        3. General reaction properties
        4. Components' summary
        5. Reaction's summary
        6. Components' workup (to be detailed)
        7. Time point data
            1. Component data (per timepoint)
            2. Reaction data (per timepoint)
    - Only add sections that are (explicitly) filled with data

RInChI – What’ next

- ProcAuxInfo – Example: Esterification with 50% yield in a flask
    - Component data

	Time	Component	2-1	2-2	3-1	3-2	4-1
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	

RInChI – What’ next

- **ProcAuxInfo=1.00**/{"**components**": [{"2-1": {"form": "fluid","color": "colorless", "density": 0.783, "purity": 99.9}},{"2-2": {"form": "fluid","color": "colorless", "density": 1.06, "purity": 99.5}},{"3-1": {"form": "fluid","color": "colorless", "density": 0.783, "purity": 99.9}},{"3-2": {"form": "fluid","color": "colorless", "density": 0.894}},{"4-1": {"form": "fluid","color": "colorless", "density": 1.84, "purity": 99.9}}],"reaction": [{"properties": {"yield": 50,"vessel": "flask"}}], "**compsummary**": [{"time": 1800,"components": [{"2-1": {"l": 0.5}},{"2-2": {"mol": 1}},{"3-1": {"mol": 0.6}},{"3-2": {"mol": 0.6}},{"4-1": {"ml": 60}}]}], "**reaccompany**": [{"time": 1800,"temperature": [20,100],"ph": [7,5],"stirring": [1000,1500]}],"**timeprop**": [{"timepoint": 0,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 1},"3-1": {"mol": 0},"3-2": {"mol": 0},"4-1": {"ml": 0}}],"reaction": [{"temperature": 20,"ph": 7,"stirring": 1000}]},{"timepoint": 300,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 0.9},"3-1": {"mol": 0.1},"3-2": {"mol": 0.1},"4-1": {"ml": 10}}],"reaction": [{"temperature": 40,"ph": 6.5,"stirring": 1000}]},{"timepoint": 600,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 0.8},"3-1": {"mol": 0.2},"3-2": {"mol": 0.2},"4-1": {"ml": 20}}],"reaction": [{"temperature": 60,"ph": 6,"stirring": 1000}]},{"timepoint": 900,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 0.7},"3-1": {"mol": 0.3},"3-2": {"mol": 0.3},"4-1": {"ml": 30}}],"reaction": [{"temperature": 90,"ph": 5.5,"stirring": 1000}]},{"timepoint": 1200,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 0.6},"3-1": {"mol": 0.4},"3-2": {"mol": 0.4},"4-1": {"ml": 40}}],"reaction": [{"temperature": 100,"ph": 5,"stirring": 1000}]},{"timepoint": 1500,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 0.5},"3-1": {"mol": 0.5},"3-2": {"mol": 0.5},"4-1": {"ml": 40}}],"reaction": [{"temperature": 100,"ph": 5,"stirring": 1500}]},{"timepoint": 1800,"components": [{"2-1": {"l": 0.5},"2-2": {"mol": 0.4},"3-1": {"mol": 0.6},"3-2": {"mol": 0.6},"4-1": {"ml": 0.6}}],"reaction": [{"temperature": 100,"ph": 5,"stirring": 1500}]}]}

RInChI – What' next

- Towards Open Source
  - Evaluate GitHub for the RInChI Open Source development
    - Develop release governance
    - Include automated testing
  - Next release is supposed to be developed "classically" but based on a publicly available GitHub repository
  - Further development of RInChI on GitHub has to be governed by the RInChI group

Thanks

- All organizers of the NIH for this meeting
- InChI Trust, Cambridge
  - IUPAC Division VIII and IUPAC's Committee on Publications and Cheminformatics Data Standards (CPCDS)
- RInChI working group
  - 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)