Zoom Cloud Meetings | Zoom Meeting | 2022-04-05 16:06:28 UTC | norwid

Zoom Meeting

View

nbehrnd | Jonathan Goodman | Jonathan Goo... | Thomas Doern...

nbehrnd | Jonathan Goodman | Jonathan Goodman | Thomas Doerner

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | Ulrich Schatzs...

burt.leland (OpenEye) | Ingvar Lagerstedt | Marc Nicklaus | Ulrich Schatzschr

1/3

Bob Belford | Falko FRENZE... | Andrey Yerin -... | Aleksandr Beli...

Bob Belford | Falko FRENZEL (BfR, G... | Andrey Yerin - ACD/Lab... | Aleksandr Belinskiy

Alexander Awad | Daniel Lowe | Evan Bolton | Richard Kidd

Alexander Awad | Daniel Lowe | Evan Bolton | Richard Kidd

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Gudrun Gygli to Everyone 06:03 PM

Sorry, have to leave for another meeting. Thanks for all the intersting info!

Sunghwan Kim, NIH to Everyone 06:04 PM

Does the current version of Polymer-InChI distinguish (1) between LDPE and BDPE (that is, different polymers from the same monomer) and (2) polymers with different degree of polymerization (i.e., polymers with different numbers of monomers).

Sorry, LDPE vs HDPE (low- vs high-density PEs)

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | 2022-04-05 16:06:42 UTC | norwid

Zoom Meeting

View

Heather
Heather

Lars B
Lars B


ggrethe

Peter Linstrom
Peter Linstrom


Blanke


Steve Heller

Leah McEwen
Leah McEwen


Hinnerk Rey - Else

2/3


Nau, Heike (ELS-FRK)

Schwörer, Mar...
Schwörer, Markus (ELS...

Sunghwan Ki...
Sunghwan Kim, NIH

Nicki Davis
Nicki Davis (She/Her)

2/3

Roger Sayle
Roger Sayle

Ian B
Ian B


Alex Henderson, Manc...


Sonja Herres-Pawlis

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Gudrun Gygli to Everyone 06:03 PM

Sorry, have to leave for another meeting. Thanks for all the intersting info!

Sunghwan Kim, NIH to Everyone 06:04 PM

Does the current version of Polymer-InChI distinguish (1) between LDPE and BDPE (that is, different polymers from the same monomer) and (2) polymers with different degree of polymerization (i.e., polymers with different numbers of monomers).

Sorry, LDPE vs HDPE (low- vs high-density PEs)

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

View

Sunghwan Ki...
Sunghwan Kim, NIH

Nicki Davis
Nicki Davis (She/Her)

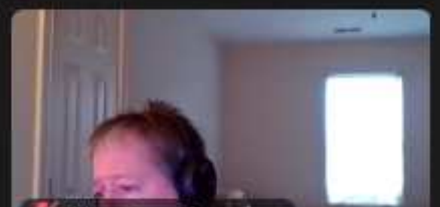
Roger Sayle
Roger Sayle

Ian B
Ian B


Henderson, Manc...


Sonja Herres-Pawlis

Anders
Anders


Hunter Moseley

3/3


rudy potenzone

Mick Kappler
Mick Kappler


Chris Jakober (CARB)

Yulia Borodina
Yulia Borodina

Barrie Walker
Barrie Walker


Gabe Weatherhead

Sandra Simon
Sandra Simon

Jane
Jane

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Gudrun Gygli to Everyone 06:03 PM

Sorry, have to leave for another meeting. Thanks for all the intersting info!

Sunghwan Kim, NIH to Everyone 06:04 PM

Does the current version of Polymer-InChI distinguish (1) between LDPE and BDPE (that is, different polymers from the same monomer) and (2) polymers with different degree of polymerization (i.e., polymers with different numbers of monomers).

Sorry, LDPE vs HDPE (low- vs high-density PEs)

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 16:31:04 UTC | norwid

Zoom Meeting

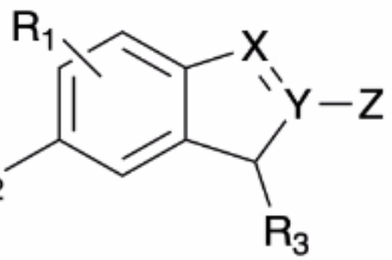
Sandra Simon | Jane | Markus Sitzm... | View

Sandra Simon | Jane | Markus Sitzmann

You are viewing Jonathan Goodman's screen | View Options

More Complex Variability

Use Te in place of Zz because RDKit can handle it



R_1 = H, Me
R_2 = Me, Et, Pr, Bu
R_3 = Ph, tolyl
X = N, CH
Y = C, N+
Z = Cl, CH_2Cl

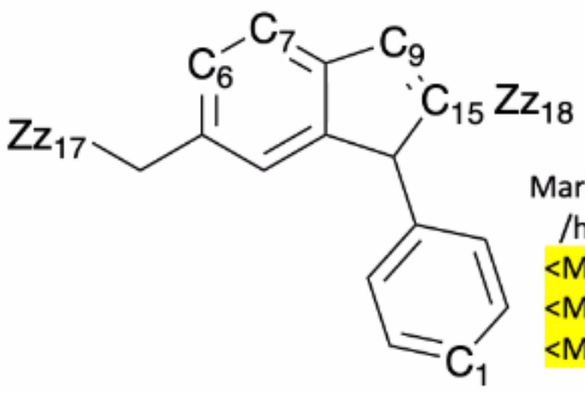
128 distinct structures

Enumerate structures

CANONICAL!
(based on template)

Software available on Github

Generate from molfile v3000



MarkInChI=1B/C16H12Zz2/c17-10-11-6-7-13-9-15(18)16(14(
/h1-9,16H,10H2
<M>H!C!C2H5Zz/c1-2-3/h2H2,1H3!C3H7Zz/c1-2-3-4/h2-3H
<M>CH2ClZz/c2-1-3/h1H2!Cl
<M>9-C!N<M>15-C!N+<M>1H-C!H<M>6H,7H-C!H

Abdullah Kattineh

Unmute | Start Video | 43 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Gudrun Gygli to Everyone 06:03 PM

Sorry, have to leave for another meeting. Thanks for all the intersting info!

Sunghwan Kim, NIH to Everyone 06:04 PM

Does the current version of Polymer-InChI distinguish (1) between LDPE and BDPE (that is, different polymers from the same monomer) and (2) polymers with different degree of polymerization (i.e., polymers with different numbers of monomers).

Sorry, LDPE vs HDPE (low- vs high-density PEs)

Andrey Yerin - ACD/Labs, IUPAC to ... 06:07 PM

LDPE and HDPE refer more to material/preparation procedures than to structure. Also not general for all polymers, so most probably should be specified in metadata, not in InChI

Sorry need to go from the office. Will join fr4om home

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

nbehrnd

Jonathan Goo...

nbehrnd

Jonathan Goodman

Jonathan Goodman

You are viewing Jonathan Goodman's screen

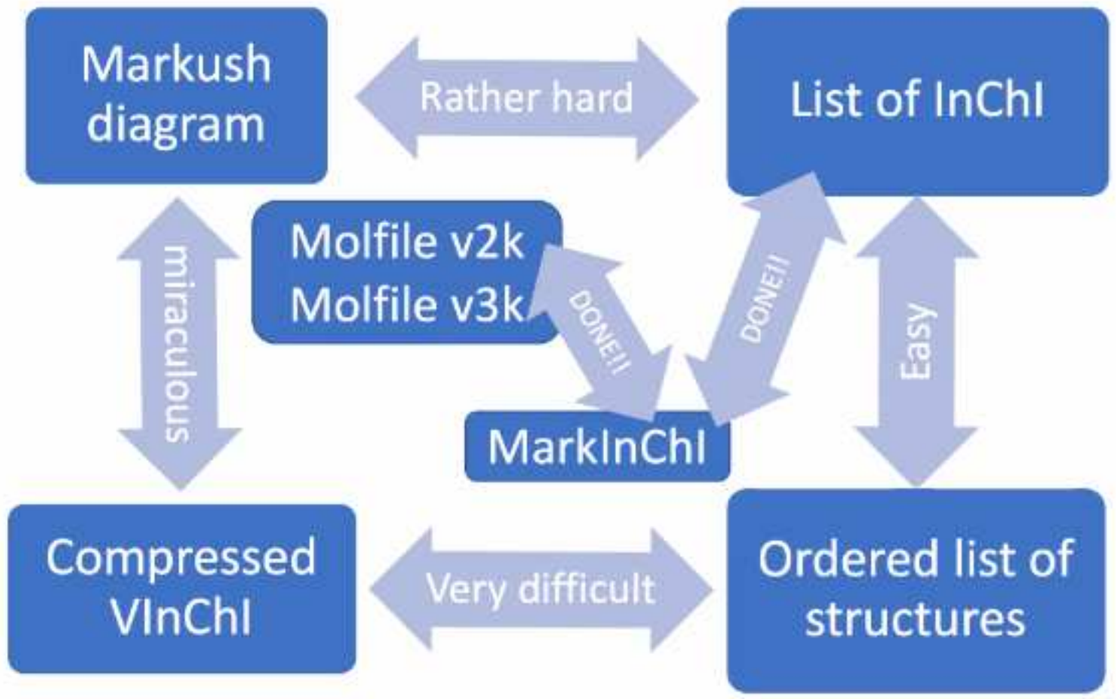
View Options

View

What is to be done?

a) Assume there are millions of structures in each Markush

b) Need to be able to calculate ∩ and ∪



Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Gudrun Gygli to Everyone 06:03 PM

Sorry, have to leave for another meeting. Thanks for all the intersting info!

Sunghwan Kim, NIH to Everyone 06:04 PM

Does the current version of Polymer-InChI distinguish (1) between LDPE and BDPE (that is, different polymers from the same monomer) and (2) polymers with different degree of polymerization (i.e., polymers with different numbers of monomers).

Sorry, LDPE vs HDPE (low- vs high-density PEs)

Andrey Yerin - ACD/Labs, IUPAC to ... 06:07 PM

LDPE and HDPE refer more to material/preparation procedures than to structure. Also not general for all polymers, so most probably should be specified in metadata, not in InChI

Sorry need to go from the office. Will join fr4om home

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:38:24 UTC | norwid

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Tautomerism

Tautomers are isomers that can readily transform into each other through chemical equilibrium reactions

Strongly environment-dependent (pH, solvent, temperature, time, ...)

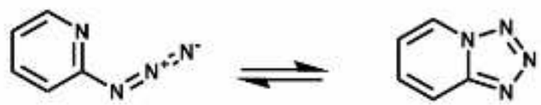
- Prototropic tautomerism:

intramolecular movement of a hydrogen atom

enol form *keto form*

- Valence tautomerism:

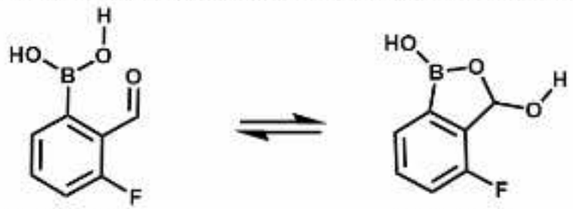
rearrangement of bonds w/o migration of atoms



azide form *tetrazole form*

- Ring-chain tautomerism:

movement of the proton accompanied by opening/closing of a ring



acyclic form *cyclic form*

2

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Aleksandr Belinskiy to Everyone 07:31 PM

Evan, WIPO is introducing from 07/2022 new standard ST. 26 for representation of biosequences which include in XML scheme description of non-standard aminoacids and chemical modifications, is not clear how the latter information would be used in search. Could your work influence next WIPO standard development for biosequence representation?

Yulia Borodina to Everyone 07:33 PM

SPL standard had introduced it in 2018... A lot of FDA data on proteins are available in this format on DailyMed. InChI is a part of the standard.

Ulrich Schatzschneider to Everyone 07:34 PM

If you want a real challenge: SARS-CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:38:51 UTC | norwid

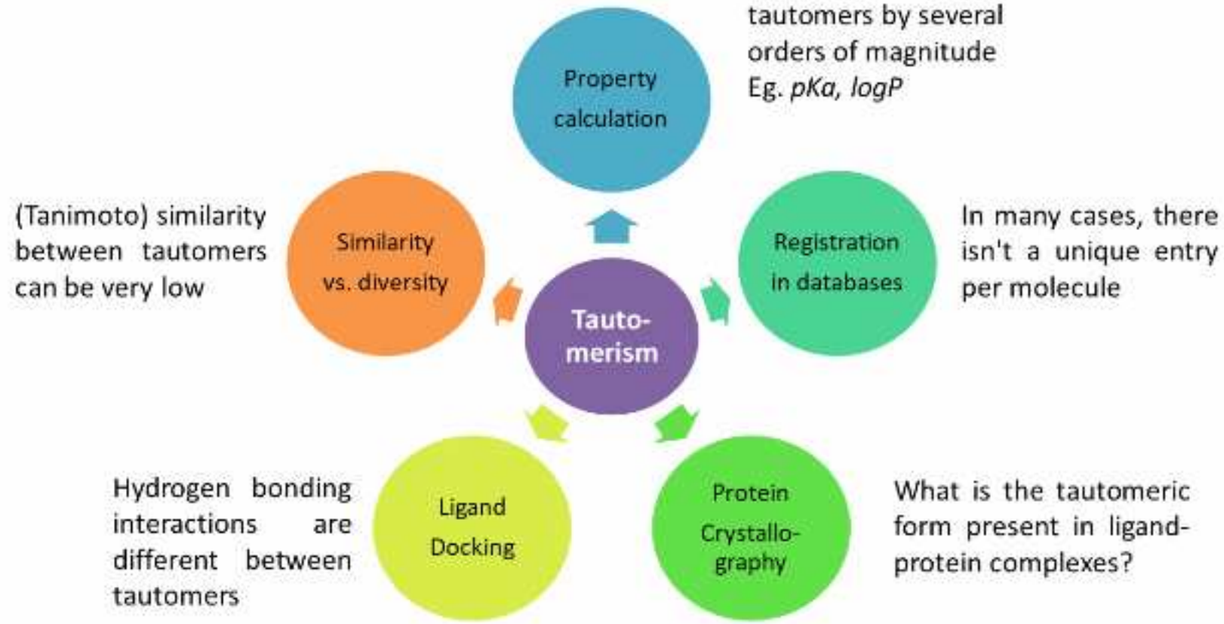
Zoom Meeting

burt.leland (O... | Ingvar Lagers... | View

burt.leland (OpenEye) | Ingvar Lagerstedt | Marc Nicklaus

You are viewing Marc Nicklaus's screen | View Options

Why worry about tautomers?



The existence of multiple tautomeric forms of the same molecule can create **problems**!!!

3

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Aleksandr Belinskiy to Everyone 07:31 PM

Evan, WIPO is introducing from 07/2022 new standard ST. 26 for representation of biosequences which include in XML scheme description of non-standard aminoacids and chemical modifications, is not clear how the latter information would be used in search. Could your work influence next WIPO standard development for biosequence representation?

Yulia Borodina to Everyone 07:33 PM

SPL standard had introduced it in 2018... A lot of FDA data on proteins are available in this format on DailyMed. InChI is a part of the standard.

Ulrich Schatzschneider to Everyone 07:34 PM

If you want a real challenge: SARS-CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Who can see your messages?

To: Everyone

Type message here...

burt.leland (O...

burt.leland (OpenEye)

Ingvar Lagers...

Ingvar Lagerstedt

View

You are viewing Marc Nicklaus's screen | View Options

Not Just an Academic Question

Tautomeric pairs (conflicts) – via NCI/CADD identifiers[1]

[1] Sitzmann *et al.* SAR QSAR Environ. Res. **2008**, *19*, 1–9

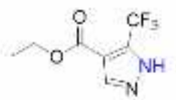
Aldrich Market Select (AMS) database :
5,755,574 molecules (2012-09 version)

31,155 conflicts ⟶ **62,869 molecules**

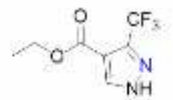
n-tuples	Conflicts
2	30,619
3	514
4	21
5	1

Guasch, L. *et al.* JCIM 56, **2016**, 2149–2161.

Examples (prices per 1 g):



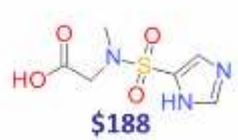
$31



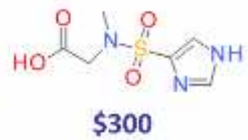
$251

$313

$350



$188



$300

Same original supplier!

4

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

SPL standard had introduced it in 2018... A lot of FDA data on proteins are available in this format on DailyMed. InChI is a part of the standard.

Ulrich Schatzschneider to Everyone 07:34 PM

If you want a real challenge: SARS-CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms 😉

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:41:10 UTC | norwid

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

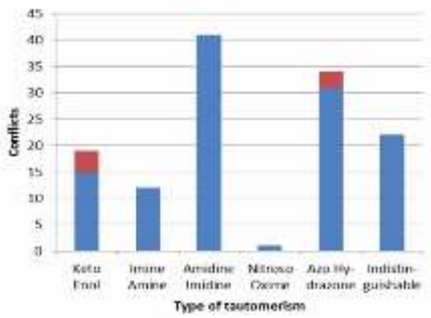
burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

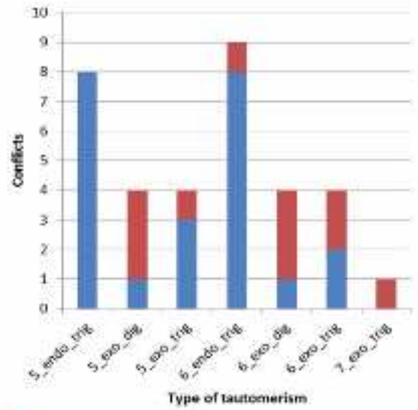
Experimental Verification

Analysis with both ^{1}H NMR and ^{13}C NMR experiments

Prototropic tautomerism



Ring-chain tautomerism



Blue: samples are the same substance. Red: samples are different substances.

Guasch, L. *et al.* JCIM *56*, **2016**, 2149–2161.

5

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

SPL standard had introduced it in 2018... A lot of FDA data on proteins are available in this format on DailyMed. InChI is a part of the standard.

Ulrich Schatzschneider to Everyone 07:34 PM

If you want a real challenge: SARS-CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Who can see your messages?

To: Everyone

Type message here...

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Tautomerism is Widespread

Tautomerism is not just interesting, important, and potentially costly – it is widespread:

- Tautomerism possible for an average of 71% structures tautomeric across ~401 million molecules

Dhaked, D. K. *et al.* J. Chem. Inf. Model. **2020**, 60, 3, 1253–1275

6

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

SPL standard had introduced it in 2018... A lot of FDA data on proteins are available in this format on DailyMed. InChI is a part of the standard.

Ulrich Schatzschneider to Everyone 07:34 PM

If you want a real challenge: SARS-CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

How does InChI up to version 1.05 handle tautomerism?

- InChI is in principle designed to be tautomer-invariant
- Standard InChI handles a limited range of tautomerism types
- One can turn on additional tautomeric types in non-standard InChI via options: KET, 15T
- It was recognized early on that important types of tautomerism are missing

7

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:42:55 UTC | norwid

Take a screenshot of the entire screen

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

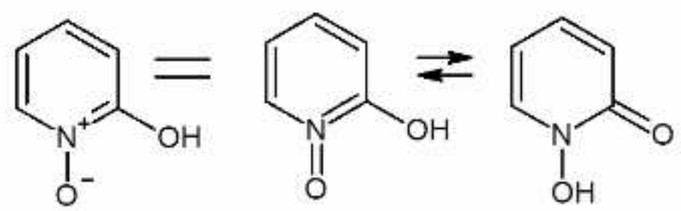
You are viewing Marc Nicklaus's screen | View Options

Why a new version of tautomerism handling needed?

Proposal by Dmitrii Tchekhovskoi in 2012:

- Another breaking change:

Add 1,4-oxime/nitroso tautomerism



InChI=1S/C5H5NO2/c7-5-3-1-2-4-6(5)8/h1-4,7H

InChI=1S/C5H5NO2/c7-5-3-1-2-4-6(5)8/h1-4,8H

8

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:43:28 UTC | norwid

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Rule Set

All rules expressed as SMIRKS, applied in CACTVS

		nn (need not be contiguous)		Number of rules
CACTVS Standard Rules	02–12	Common	New extensions: PT_11_*0m*, *m* = 1 - 4	24
	13–21	Rare		
Prototropic Rules PT_*nn*_*mm* — New Rules	22–49	(includes subrules PT_*nn*_0*m*, *m* > 0)		30
Ring-Chain Rules RC_*nn*_*mm* — New Rules	03–24	← Includes ... rules from ref. (1) (includes subrules RC_*nn*_0*m*, *m* > 0)		54
Valence Tautomerism Rules VT_*nn*_*mm* — New Rules	01–10	(includes subrules VT_*nn*_0*m*, *m* > 0)		11

Total number of rules: 119

(1) Guasch L. *et al.*, J. Chem. Inf. Model. **2014**, 54, 2423−2432
Dhaked D. *et al.*, J. Chem. Inf. Model. **2020**, 60, 3, 1090−1100
Dhaked D. *et al.*, J. Chem. Inf. Model. **2020**, 60, 3, 1253−1275
Dhaked D., Nicklaus M. ChemRxiv 10.26434/chemrxiv.14779254.v1

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:44:52 UTC | norwid

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Tautomer Enumeration Tool

https://cactus.nci.nih.gov/tautomerizer/

NCI/CADD Group

Tautomerizer - Predict tautomers based on 80+ rules

Introduction | Form | Individual Rule Pages | Rules Sources | Help

Enter the structure in SMILES format

1. Input Structure SMILES: Structure Editor

Submit

2. Single step or Multi step:

Single step ○ Multi step

3. Activate rules:

○ Activate all rules
○ Activate standard rules
○ Activate only new rules
○ Enter your own rule as SMIRKS:
● Activate custom rule set via following checkboxes:

Select rules

☐ PT_02_00 - **1,5 (thio)keto/(thio)enol** - [O,S,Se,Te;X1:1]=[Cz1H0:2][C:5]=[C:6][CX4z0,NX3:3][#1:4]>>[#1:4][O,S,Se,Te;X2:1][Cz1:2]=[C:5][C:6]=[Cz0,N:3]
○ Select example: C1=CC(C=C(C1=O)C)=O
Run Example

☐ PT_03_00 - **simple (aliphatic) imine** - [#1,a,O:5][NX2:1]=[Cz{1-2}:2][CX4R{0-2}:3][#1:4]>>[#1,a,O:5][NX3:1]([#1:4])[Cz:2]=[C:3]
○ Select example: [C]1(CC[C]CC1)=[N]
Run Example

☐ PT_04_00 - **special imine** - [Cz0R0X3:1]([C:5])=[C:2][Nz0:3][#1:4]>>[#1:4][Cz0R0X4:1]([C:5])[c:2]=[nz0:3]
○ Select example: C(CC1=NC=C[NH]1)(C)C
Run Example

10

Hitesh Patel

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

CoV-2 RNA genome is approx. 30.000 bp = around 1.200.000 atoms ☺

Yulia Borodina to Everyone 07:35 PM

@Ulrich

Evan Bolton to Everyone 07:39 PM

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

InChI[Key] (V. 105) only Partially Recapitulates a More Complete Set of Rules

InChI Calculation Type >		Standard	(DONOTADDH W0)	
Database	Database Size	Tautomeric Part	InChI Success Rate (%)	Strict InChI Success Rate (%)
CSD	319,201	203,108	26.25	13.46
ChEMBL	1,820,035	1,578,290	62.15	28.55
AMS	8,409,644	7,204,965	64.77	29.85
PUBCHEM	96,502,282	78,807,315	56.64	29.47
CSDB	141,743,903	127,543,398	71.27	31.90

Rules applied in chemoinformatics toolkit CACTVS

Devendra Dhaked

InChI Calculation Type >		Non-standard	(DONOTADDH W0 RECMET NEWPS SPXYZ SAsXYZ Fb Fnud KET 15T)	
Database	Database Size	Tautomeric Part	InChI Success Rate (%)	Strict InChI Success Rate (%)
CSD	319,201	203,108	48.83	30.90
ChEMBL	1,820,035	1,578,290	73.91	37.46
AMS	8,409,644	7,204,965	71.99	36.32
PUBCHEM	96,502,282	78,807,315	66.52	38.26
CSDB	141,743,903	127,543,398	78.70	38.97

Dhaked D. *et al.*, J. Chem. Inf. Model. **2020**, 60, 3, 1253–1275

InChI Success Rate: At least two rule-enumerated tautomers have same InChIKey

Strict InChI Success Rate: **All** rule-enumerated tautomers have same InChIKey

11

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Ulrich Schatzschneider to Everyone 07:46 PM

How fixed are these rules? Are they "hard-coded" into the main code and need a lot of effort to add/modify or is there another more low-key mechanism to append in case further exceptions are identified?

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

New Rules: Integrated in experimental version of InChI 1.06

- New rules, as implemented in CACTVS, expressed as SMIRKS
- InChI doesn't have a SMIRKS parser
- Adding new tautomeric rules requires **code changes in the core of InChI**

- We picked ~20 prototropic rules as candidates for implementation in InChI
- No ring-chain or valence tautomerism rules – impossible to add to current InChI

- Igor Filippov was able to add **six new rules**

Note: These were all the rules that could be added; for others, the effort was unsuccessful!

Igor Filippov

Six new rules implemented in InChI library (based on V. 1.06 code) integrated in CACTVS.

Wolf-D. Ihlenfeldt

12

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Thanks Aleksandr! Yes, one could imagine being able to work with other formatted data. The use cases of InChI for search tend to be focused around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Ulrich Schatzschneider to Everyone 07:46 PM

How fixed are these rules? Are they "hard-coded" into the main code and need a lot of effort to add/modify or is there another more low-key mechanism to append in case further exceptions are identified?

Who can see your messages?

To: Everyone

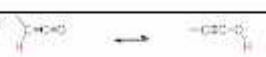
Type message here...

burt.leland (O... | Ingvar Lagers... | View

burt.leland (OpenEye) | Ingvar Lagerstedt | Marc Nicklaus

You are viewing Marc Nicklaus's screen | View Options

New Rules Implemented

PT_06_00	 1,3 heteroatom H shift	[CX{2-3}r{0-1}:N,n,S,s,O,o,Se,Te:1]-[NX2,nX2,CX1,c,P,p:2][N,n,S,O,Se,Te:3][#1:4] >>[#1:4][CX3r{0-1}:N,n,S,O,Se,Te:1][NX2,nX2,CX3r{0-1},c,P,p:2]-[N,n,S,s,O,o,Se,Te:3]
PT_13_00	 ketone-enol exchange	[O,S,Se,Te;X1:1]-[C:2]-[C:3][#1:4]>>[#1:4][O,S,Se,Te;X2:1][C:2]=[C:3]
PT_16_00	 nitroso/oxime	[#1,1][O,R:2][N+0,r3]=[CX3,4]>>[O,R:2]-[N+0,r3][CX4,4][#1,1]
PT_18_00	 cyanic/iso cyanic acids	[#1:1][O:2][C:3]-[N:4]>>[O:2]=[C:3]=[N:4][#1:1]
PT_22_00	 imine/imine	[#1,1][CX4,2][NX2,3]-[CX3,4]>>[CX3,2]-[NX2,3][CX4,4][#1,1]
PT_39_00	 nitrone/oxaziridine Beckmann rearrangement	[CX3,NX2:1]=[NX3+:2][O-:3].[CX4:4][#1:5]>>[#1:5][CX4,NX3:1][NX3+:2][O-:3]-[CX3:4]

Note that example structures are just that: examples. Similar for the names. The SMIRKS are really defining the rule!

In InChI, new code has to be written!

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Ulrich Schatzschneider to Everyone 07:46 PM

How fixed are these rules? Are they "hard-coded" into the main code and need a lot of effort to add/modify or is there another more low-key mechanism to append in case further exceptions are identified?

The problem at least in inorganic compounds for almost everything is "the rule is there is no general rule"

Who can see your messages?

To: Everyone

Type message here...

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

What have we gained with the six new rules?

Total of 8 rules: KET, 15T, PT_06_00, PT_13_00, PT_16_00, PT_18_00, PT_22_00, PT_39_00

Counting various identifiers for recent version of PubChem (2020-02 Compound database):

90,600,000 compounds analyzed for the PubChem file(s)

Number of cases where S, N, T same:	26,339,099
Number of cases where S, N same, T diff:	28,104,489
Number of cases where N, T same, S diff:	10,892,602
Number of cases where S, N, T diff:	25,252,306

Note: Most analyses were done with InChKeys but could as well have been done with InChIs

90,390,917 unique Standard InChIKeys found

87,322,470 unique non-standard InChIKeys (with KET and 15T turned on)
87,322,472 unique Tauto InChIKeys with KET and 15T turned on

Differences vs. the non-standard InChIKey count:

84,249,356 (-3.519%, -3,073,114) unique Tauto InChIKeys (with all 8 rules turned on)

87,448,002 (0.144%, 125,532) unique Tauto InChIKeys with just KET turned on
90,184,001 (3.277%, 2,861,531) unique Tauto InChIKeys with just 15T turned on
86,004,390 (-1.509%, -1,318,080) unique Tauto InChIKeys with just PT_06_00 turned on
90,310,034 (3.421%, 2,987,564) unique Tauto InChIKeys with just PT_13_00 turned on
90,284,657 (3.392%, 2,962,187) unique Tauto InChIKeys with just PT_16_00 turned on
90,326,333 (3.440%, 3,003,863) unique Tauto InChIKeys with just PT_18_00 turned on
88,785,161 (1.675%, 1,462,691) unique Tauto InChIKeys with just PT_22_00 turned on
90,324,463 (3.438%, 3,001,993) unique Tauto InChIKeys with just PT_39_00 turned on

14

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

around near-field 'sameness' .. so the search scheme might partially fill this use case.

Thanks Ulrich! Yes .. there are some really good ones out there .. the PDB has some real large ones these days (ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Ulrich Schatzschneider to Everyone 07:46 PM

How fixed are these rules? Are they "hard-coded" into the main code and need a lot of effort to add/modify or is there another more low-key mechanism to append in case further exceptions are identified?

The problem at least in inorganic compounds for almost everything is "the rule is there is no general rule"

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Tautomer Structures Extracted from Experimental Literature: "Tautomer Database"

https://cactus.nci.nih.gov/download/tautomer/

Release 3 - November 2019

2,819 Tautomeric Tuples Comprising 5,977 Structures

Structurally different tuples: 1,776 (comprising **3,884 different structures**)
since some tuples are differentiated from each other only by experimental conditions such as solvent, spectroscopy method, etc.

See https://doi.org/10.1021/acs.jcim.9b01156
and https://doi.org/10.26434/chemrxiv.10790369.v1 for literature about this database.

Dhaked D. *et al.*, J. Chem. Inf. Model. **2020**, 60, 3, 1090–1100

15

Unmute | Start Video | 42 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

(ribosomes, etc.).

Thanks Jonathan! Yes, some can be described in purely atomic terms .. however, usually some piece is or pieces are missing .. for many reasons (e.g., attached to a cell wall) .. it makes for an interesting problem to describe what is meaningful based on the use case.

Ulrich Schatzschneider to Everyone 07:46 PM

How fixed are these rules? Are they "hard-coded" into the main code and need a lot of effort to add/modify or is there another more low-key mechanism to append in case further exceptions are identified?

The problem at least in inorganic compounds for almost everything is "the rule is there is no general rule"

Hunter Moseley to Everyone 07:50 PM

Does CACTUS parse SMIRKS rules every time or are they pre-parsed and saved in compilable or actionable form?

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:54:05 UTC | norwid

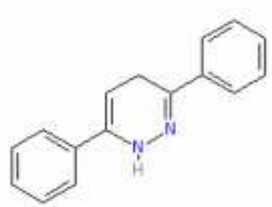
Zoom Meeting

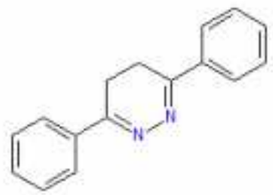
burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Only Tauto InChI Identifies Tautomers





Non-standard InChIKey:	UHFJDRXFXQDUHB-UHFFFAOYNA-N	GEHJYGJMMCSVQL-UHFFFAOYNA-N
Tauto InChIKey:	YAMMWMQCRXUUCQ-UHFFFAOYNA-N	YAMMWMQCRXUUCQ-UHFFFAOYNA-N

Case: 115_01 and _02

19

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

The problem at least in inorganic compounds for almost everything is "the rule is there is no general rule"

Hunter Moseley to Everyone 07:50 PM

Does CACTUS parse SMIRKS rules every time or are they pre-parsed and saved in compilable or actionable form?

Ulrich Schatzschneider to Everyone 07:53 PM

On PubChem: keep in mind that a good part of the inorganic structures is more or less nonsense due to "disconnection" - for example has cisplatin and transplatin as "synonyms" although cisplatin is a multibillion dollar anticancer drug while transplatin is inactive. Same with ferrocenes important in catalysis

Jonathan Goodman to Everyone 07:53 PM

Why does the tautomer code have to be in the core InChI? It seems an ideal post-processing step: does this Standard InChI have another tautomeric form and, if so, which one is preferred?

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Summary

Compared with a comprehensive set of tautomeric rules:

- Current Standard InChI recapitulates ~30% of amenable compounds
- Current Non-Standard InChI (KET, 15T) recapitulates ~37% of compounds
- Relative to Standard InChI, Non-Standard InChI (KET, 15T) equates 3.5% more compounds as tautomers of other compounds in a typical large database (e.g. PubChem)

Working group achievements:

- Six new prototropic rules were added to InChI code
- Relative to Standard InChI, "Tauto InChI" (KET, 15T, 6 new rules) equates 7% more compounds as tautomers of other compounds, i.e. yet 3.5% more than Non-Standard InChI

Availability:

- Experimental version of InChI 1.06 released with 6 rules added

21

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

The problem at least in inorganic compounds for almost everything is "the rule is there is no general rule"

Hunter Moseley to Everyone 07:50 PM

Does CACTUS parse SMIRKS rules every time or are they pre-parsed and saved in compilable or actionable form?

Ulrich Schatzschneider to Everyone 07:53 PM

On PubChem: keep in mind that a good part of the inorganic structures is more or less nonsense due to "disconnection" - for example has cisplatin and transplatin as "synonyms" although cisplatin is a multibillion dollar anticancer drug while transplatin is inactive. Same with ferrocenes important in catalysis

Jonathan Goodman to Everyone 07:53 PM

Why does the tautomer code have to be in the core InChI? It seems an ideal post-processing step: does this Standard InChI have another tautomeric form and, if so, which one is preferred?

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Conclusion

Notes and Questions:

- **Tauto InChI is different InChI: many InChIKeys are different. Do not mix with non-standard InChI!**
- Maybe should have kind indicator "T" instead of "N": **WCHQBIYPPGCACF-UHFFFAOYTA-N** ?

- How to test: Which rules are realistic, which ones may be too strict?

Future outlook:
- Prototropic transforms: doubtful whether more can be added
- Ring-chain, valence tautomerism: likely incompatible with current InChI chemical structure model

- **To be able to add more rules, InChI code likely needs to be re-written**

22

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

and saved in compilable or actionable form?

Ulrich Schatzschneider to Everyone 07:53 PM

On PubChem: keep in mind that a good part of the inorganic structures is more or less nonsense due to "disconnection" - for example has cisplatin and transplatin as "synonyms" although cisplatin is a multibillion dollar anticancer drug while transplatin is inactive. Same with ferrocenes important in catalysis

Jonathan Goodman to Everyone 07:53 PM

Why does the tautomer code have to be in the core InChI? It seems an ideal post-processing step: does this Standard InChI have another tautomeric form and, if so, which one is preferred?

Ulrich Schatzschneider to Everyone 07:55 PM

What if different tautomers can be stabilized under different conditions - wouldn't selection of a preferred one then case problems as they have to be distinguished?

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 17:57:21 UTC | norwid

Zoom Meeting

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus | View

burt.leland (OpenEye) | Ingvar Lagerstedt

You are viewing Marc Nicklaus's screen | View Options

Acknowledgements

Markus Sitzmann
Waruna Yapamudiyansel
Megan L. Peach
James A. Kelley
Joseph J. Barchi
Jeff Saxe
Igor Filippov

Members of the IUPAC Working Group:
Gerd Blanke
Evan Bolton
Alex M. Clark
Bret Daniel
Devendra Dhaked
Laura Guasch
Wolf-Dietrich Ihlenfeldt
Gregory Landrum
John W. Mayfield
Hitesh Patel
Roger Sayle
Dmitrii Tchekhovskoi

Igor Pletnev (†)

23

Unmute | Start Video | 41 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

and saved in compilable or actionable form?

Ulrich Schatzschneider to Everyone 07:53 PM

On PubChem: keep in mind that a good part of the inorganic structures is more or less nonsense due to "disconnection" - for example has cisplatin and transplatin as "synonyms" although cisplatin is a multibillion dollar anticancer drug while transplatin is inactive. Same with ferrocenes important in catalysis

Jonathan Goodman to Everyone 07:53 PM

Why does the tautomer code have to be in the core InChI? It seems an ideal post-processing step: does this Standard InChI have another tautomeric form and, if so, which one is preferred?

Ulrich Schatzschneider to Everyone 07:55 PM

What if different tautomers can be stabilized under different conditions - wouldn't selection of a preferred one then case problems as they have to be distinguished?

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

View

burt.leland (O... | Ingvar Lagers... | Marc Nicklaus

burt.leland (OpenEye) | Ingvar Lagerstedt | Marc Nicklaus

You are viewing Yulia Borodina's screen | View Options

FDA | OFFICE OF DATA, ANALYTICS, AND RESEARCH

Challenge announcement

FDA plans to open a public challenge in collaboration with the InChI Trust and IUPAC Working Group on tautomers

Challenge title (tentative): Crowdsourced evaluation of InChI-based tautomer identification

Timeframe (tentative): September 2022 – January 2023

Purpose: Determine how well InChI algorithm agrees with experimental determination of tautomers in chemical databases

Why it Matters: Tautomers are molecules that interconvert at normal conditions and should therefore be considered the same substance for the purpose of regulation of their therapeutic use and/or tracing their presence as impurities in products

Incentives for participants:

- Recognition by FDA as participant in helping to enable improvements in health outcomes
- Opportunity to co-author a paper

4

Unmute | Start Video | 40 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

There is also the issue of knowing the energy barrier in the solvent of interest

If you can put the tautomer in the bottle (and it does not scramble quickly, say seconds, then it may be important .not. to collapse?)

Me to Everyone 08:02 PM

@Ulrich Check the different Smiles/ non-standard InChI about 2-hydroxy pyridine/2-pyridone e.g. on https://en.wikipedia.org/wiki/2-Pyridone

Ulrich Schatzschneider to Everyone 08:02 PM

Does the barrier really matter? It's again "outside Information". Isn't it enough that a person preparing a structure thinks it is sufficiently different from something else that it should be differentated?

Me to Everyone 08:02 PM

The interchange may be dynamic (solution), or frozen (e.g., solid state/ crystal).

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

Challenge Background:

- Since 2010, the InChI Trust has supported the continual development of a standard way to encode molecular information and to facilitate the search for such information in databases and on the web
- The InChI identifier is computed from chemical structure and is not sensitive to certain variations of structure representations.
- However, it doesn't recognize some cases of structures being tautomers of each other.
- Tautomer Working Group suggested a few modifications to the InChI algorithm that would recognize more tautomer types

Challenge Objectives:

- Members of Industry, Government, and Academia will be invited to test the InChI algorithm modifications on their real chemical samples by:
  1. Running their database of structures through specified computational analyses
  2. Comparing identified tautomeric sets with their internal analytical data, most likely NMR spectroscopy
  3. Determining whether members of these sets can interconvert at normal conditions
  4. Reporting the summaries of their analyses for tautomeric rule sets
- As a result, participants will:
  - Help with the creation of better tautomer-invariant identifiers
  - Find previously unknown tautomeric duplicates in their databases

5

Unmute | Start Video | 40 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Ulrich Schatzschneider to Everyone 08:02 PM

Does the barrier really matter? It's again "outside Information". Isn't it enough that a person preparing a structure thinks it is sufficiently different from something else that it should be differentated?

Me to Everyone 08:02 PM

The interchange may be dynamic (solution), or frozen (e.g., solid state/ crystal).

Evan Bolton to Everyone 08:03 PM

If you can separate it in the bottle and do separate chemistry on it .. yes, it matters very much to be able to associated data to it separately

Marc Nicklaus to Everyone 08:03 PM

@Jonathan: Yes, it would make things much easier if (s0me part of) the tautomerism part of InChI V2 would not be in the very core of the code. But in V1 it is.

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

rudy potenzone | Mick Kappler | Mick Kappler | Yulia Borodina

View

You are viewing Yulia Borodina's screen | View Options

FDA | OFFICE OF DATA, ANALYTICS, AND RESEARCH

What is precisionFDA?

https://precision.fda.gov

7

Unmute | Start Video | 40 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

again "outside Information". Isn't it enough that a person preparing a structure thinks it is sufficiently different from something else that it should be differentated?

Me to Everyone 08:02 PM

The interchange may be dynamic (solution), or frozen (e.g., solid state/ crystal).

Evan Bolton to Everyone 08:03 PM

If you can separate it in the bottle and do separate chemistry on it .. yes, it matters very much to be able to associated data to it separately

Marc Nicklaus to Everyone 08:03 PM

@Jonathan: Yes, it would make things much easier if (s0me part of) the tautomerism part of InChI V2 would not be in the very core of the code. But in V1 it is.

@Mick: The supplemental tables should be on ChemRxiv. If not, contact me.

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

Unmute | Start Video | 40 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

The interchange may be dynamic (solution), or frozen (e.g., solid state/ crystal).

Evan Bolton to Everyone 08:03 PM

If you can separate it in the bottle and do separate chemistry on it .. yes, it matters very much to be able to associated data to it separately

Marc Nicklaus to Everyone 08:03 PM

@Jonathan: Yes, it would make things much easier if (s0me part of) the tautomerism part of InChI V2 would not be in the very core of the code. But in V1 it is.

@Mick: The supplemental tables should be on ChemRxiv. If not, contact me.

Ulrich Schatzschneider to Everyone 08:06 PM

In a "hydrogen-pruned" version of a structure tautomers will get the same identifier (prune = non-bridging H completely removed/ignored) ...

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

rudy potenzone

Mick Kappler

Mick Kappler

Yulia Borodina

View

You are viewing Yulia Borodina's screen

View Options

Overview Challenges News Experts About UNII Search Request Access Log In

precisionFDA

A secure, collaborative, high-performance computing platform that builds a community of experts around the analysis of biological datasets in order to advance precision medicine.

EXPERT HIGHLIGHT

New Tools: Processing Electronic H ... and Increasing Potential Adverse Ev...

Roselie A. Bright, ScD, MS and Summer R...

Most clinical safety systems, including ... and patients, despite well-documentr... condition is reportable, and burden... true authorship of literature writt... notice the approaching army, des...

Expert Q&A ☆ About Th...

6,000+ Users
107 Terabytes of Data
136 Unique Publicly Available Apps
31 Challenges
Cited in a Total of 78 Scientific Publications
11 Contributed/Written Scientific Publications
21 Training Videos
5 Training Workshops

FISMA FedRAMP

RECENT EXPERT BLOGS

Mark Stewart and Grace Collins
Harmonizing Tumor Mutational Burden (TMB) Assessment to Inform Cancer Treatment Decisions
Nov 16, 2021

Andrew Kennedy
FDA's Traceability Challenge Opens the Door to Conversations on Innovation
Jun 30, 2021

View All Expert Blogs

precisionFDA Challenges

UPCOMING CHALLENGES

NCTR Indel Calling from ...

STARTS 05/02/2022 → ENDS 07/26/2022

Genetic variation involving Indels (inser... protein coding and often lead to changes i... variants, it is important that the tools for ind... FDA's National Center for Toxicological Resear... pipelines to identify indels in oncopanel sequencing...

View Details →

LATEST NEWS

Upcoming FDA Public Meeting on Biomarker Terminology
Mar 23, 2022 View →

Expert Blog: New Tools: Processing Electronic Healthcare Notes and Using Them to Find New and Increasing Potential Adverse Events, Even if Unattributed
Feb 15, 2022 View →

Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Ulrich Schatzschneider to Everyone 08:06 PM

In a "hydrogen-pruned" version of a structure tautomers will get the same identifier (prune = non-bridging H completely removed/ignored) ...

Thomas Doerner to Everyone 08:07 PM

Re. interconverting and not interconverting tautomers: Would the combination of a very fine-grained identifier part (no tautomer merging) plus an aggregation part which aggregates different tautomers (to aid finding tautomers) as an overall InChI address the various needs ? (the members of the organometallics InChI team may remember this kind of proposal... :-)

Marc Nicklaus to Everyone 08:07 PM

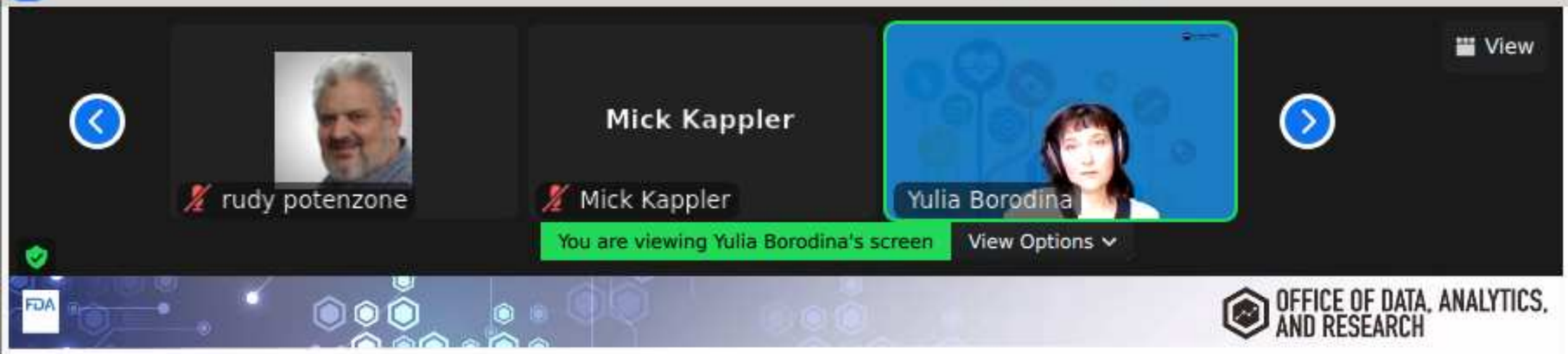
@Daniel: Since I did this analysis in CACTVS. I don't think I saw a difference in speed when the 6 rules were added. But running the InChI library within CACTVS is generally significantly slower than the stand-alone InChI executable.

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting



Data on pFDA are secure and private unless shared

- All data brought to precisionFDA are private by default, cannot be viewed by anyone else
- One can use public apps with private data
- Adding files to precisionFDA:
  - Upload through web UI under My Home->Files
  - Upload from command line with pFDA command line uploader https://precisionfda-staging.dnanexus.com/docs/files#files-uploading
  - Import from a public URL via the URL Fetch app

10

Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

Ulrich Schatzschneider to Everyone 08:06 PM

In a "hydrogen-pruned" version of a structure tautomers will get the same identifier (prune = non-bridging H completely removed/ignored) ...

Thomas Doerner to Everyone 08:07 PM

Re. interconverting and not interconverting tautomers: Would the combination of a very fine-grained identifier part (no tautomer merging) plus an aggregation part which aggregates different tautomers (to aid finding tautomers) as an overall InChI address the various needs ? (the members of the organometallics InChI team may remember this kind of proposal... :-)

Marc Nicklaus to Everyone 08:07 PM

@Daniel: Since I did this analysis in CACTVS. I don't think I saw a difference in speed when the 6 rules were added. But running the InChI library within CACTVS is generally significantly slower than the stand-alone InChI executable.

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting

rudy potenzone | Mick Kappler | Yulia Borodina

You are viewing Yulia Borodina's screen | View Options | View

FDA | OFFICE OF DATA, ANALYTICS, AND RESEARCH

Tautomeric rule sets for testing

1-rule-sets:

- KET: keto-enol tautomerism recognized
- 15T: 1,5 H-transfer recognized
- PT_06_00: "1,3 heteroatom H shift" recognized
- PT_13_00: "keten/ynol exchange" recognized
- PT_16_00: "nitroso/oxime" tautomerism recognized
- PT_18_00: "cyanic/iso-cyanic acids" tautomerism recognized
- PT_22_00: "imine via imine" tautomerism recognized
- PT_39_00: "nitrone/azoxy or Behrend rearrangement" recognized

2-rule-set:

- KET + 15T

5-rule-set:

- PT_13_00+PT_16_00+PT_18_00+PT_22_00+PT_39_00

6-rule-set:

- PT_06_00+PT_13_00+PT_16_00+PT_18_00+PT_22_00+PT_39_00

7-rule-set:

- 15T+PT_06_00+PT_13_00+PT_16_00+PT_18_00+PT_22_00+PT_39_00

8-rule-set:

- KET+15T+PT_06_00+PT_13_00+PT_16_00+PT_18_00+PT_22_00+PT_39_00

13

Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

InChI address the various needs ? (the members of the organometallics InChI team may remember this kind of proposal... :-)

Marc Nicklaus to Everyone 08:07 PM

@Daniel: Since I did this analysis in CACTVS. I don't think I saw a difference in speed when the 6 rules were added. But running the InChI library within CACTVS is generally significantly slower than the stand-alone InChI executable.

Ulrich Schatzschneider to Everyone 08:08 PM

@Thomas: Good to bring this up again - I think the issue of how finely "grained" an identifier should be still seems to be unclear ...

Andrey Yerin to Everyone 08:09 PM

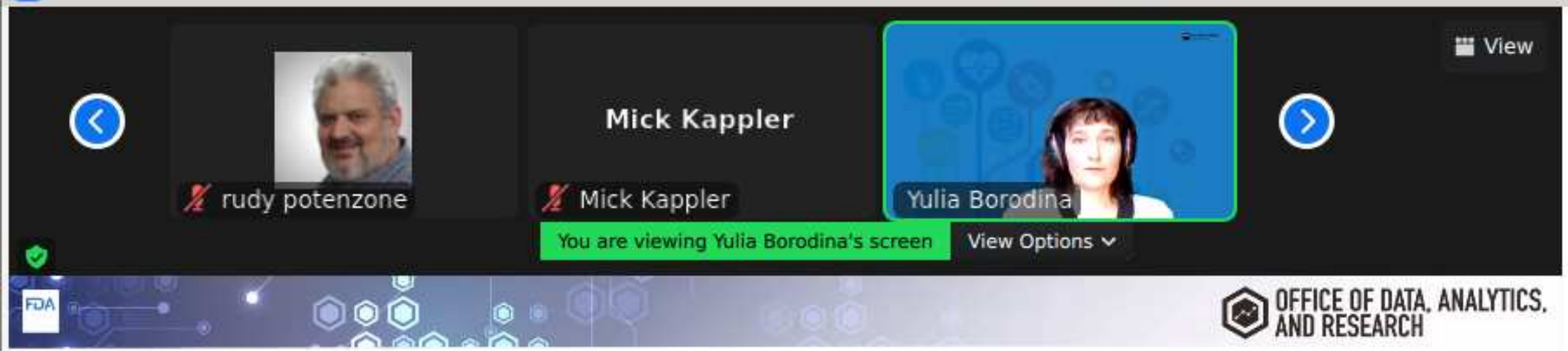
I would say that tautomer module can be in 'preparation' module. There one 'best taut' can be created and in this way tauts will have the same InChI. While may need format changes - no need to idicate distributed location of H

Who can see your messages?

To: Everyone

Type message here...

Zoom Meeting



Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

InChI address the various needs ? (the members of the organometallics InChI team may remember this kind of proposal... :-)

Marc Nicklaus to Everyone 08:07 PM

@Daniel: Since I did this analysis in CACTVS. I don't think I saw a difference in speed when the 6 rules were added. But running the InChI library within CACTVS is generally significantly slower than the stand-alone InChI executable.

Ulrich Schatzschneider to Everyone 08:08 PM

@Thomas: Good to bring this up again - I think the issue of how finely "grained" an identifier should be still seems to be unclear ...

Andrey Yerin to Everyone 08:09 PM

I would say that tautomer module can be in 'preparation' module. There one 'best taut' can be created and in this way tauts will have the same InChI. While may need format changes - no need to idicate distributed location of H

Who can see your messages?

To: Everyone

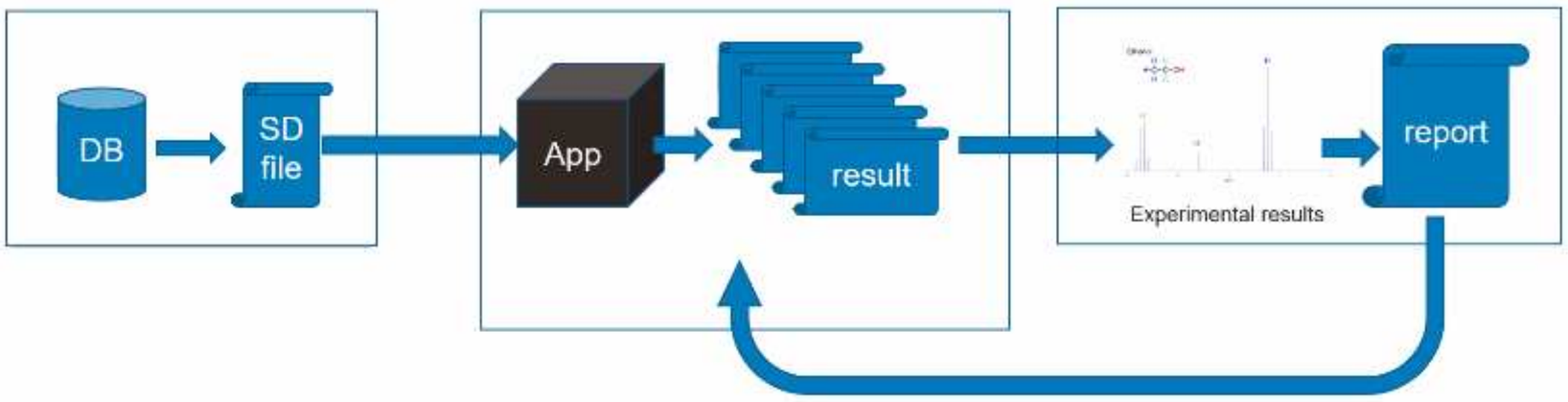
Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 18:11:02 UTC | norwid

Zoom Meeting

Workflow 1 – For participants who can upload files with structures

Participant site | precisionFDA platform | Participant site



14

Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

I would say that tautomer module can be in 'preparation' module. There one 'best taut' can be created and in this way tauts will have the same InChI. While may need format changes - no need to idicate distributed location of H

Ulrich Schatzschneider to Everyone 08:10 PM

For a regulatory person or sombody who wants to order a commercial compound "can be put in a bottle" is enough - if one needs to identify something in an ELN, it might be different depending on your type of chemistry. Not everybody is a synthetic chemist - think low-temperature spectroscopists or gas phase mass spectrometry people - all of them also ned identifiers!

Marc Nicklaus to Everyone 08:10 PM

@Ulrich and Evan: I think repository managers would be interested in knowing if molecules can tautomerically interconvert within, say, months. But this is part of the larger discussion of what the foremost use case of InChI should be.

Who can see your messages?

To: Everyone

Type message here...

Zoom Cloud Meetings | Zoom Meeting | Terminal | 2022-04-05 18:17:35 UTC | norwid

Zoom Meeting

Interested in participating?

Contact yulia.borodina@fda.hhs.gov

Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

attached reference I already thought so: Evolution of Novartis' Small Molecule Screening Deck Design https://doi.org/10.1021/acs.jmedchem.0c01332

Me to Everyone 08:15 PM

@Marc But what often happens is that you one hasn't one .or. the other , but one preferentially over the other (i.e., a blend). Cf. Tautobase, https://github.com/WahlOya/Tautobase

Ulrich Schatzschneider to Everyone 08:17 PM

@Hinnerk Problem is how many applications actually use "non-standard InChI" or at least allow you to switch between "standard" and "non-standard"? If non-standard is there but not use beyond small developer community it is almost as good as not there at all (= not of help)

Marc Nicklaus to Everyone 08:17 PM

@Ulrich: Absolutely, in out Tautomer Database we see "blends" very frequently, just like in Tautobase.

Who can see your messages?

To: Everyone

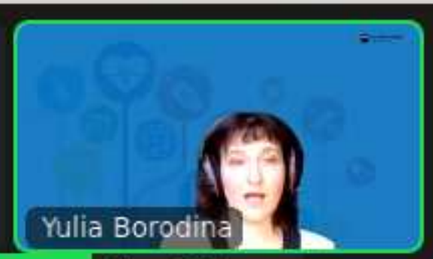
Type message here...

Zoom Meeting

rudy potenzone

Mick Kappler

Mick Kappler



Yulia Borodina

You are viewing Yulia Borodina's screen

View Options

View

FDA

OFFICE OF DATA, ANALYTICS, AND RESEARCH

Acknowledgments

Igor Pletnev (InChI Trust)
Steve Heller (InChI Trust)
Rudy Potenzone (InChI Trust)
Gerd Blanke (InChI Trust)
Marc Nicklaus (IUPAC WG on tautomers)
Wolf-Deitrich Ihlendfeldt (IUPAC WG on tautomers)
Andrey Yerin (ACDlabs)
Antony Williams (EPA)
Charles Lowe (EPA)
Lutz Weber (Ontochem)
Mitch Miller (Scientific Thinking LLC)
Yurii Moroz (Enamine)
Yuri Pevzner (Abbvie)

Elaine Johanson (FDA)
Annette Vernon (FDA)
Letria Hall (FDA)
Sarah Prezek (precisionFDA)
Vishal Thovarai (precisionFDA)
Sam Westreich (precisionFDA)
Omar Serang (precisionFDA)

24

Unmute | Start Video | 39 Participants | Chat | Share Screen | Record | Reactions | Leave

Chat

attached reference I already thought so: Evolution of Novartis' Small Molecule Screening Deck Design https://doi.org/10.1021/acs.jmedchem.0c01332

Me to Everyone 08:15 PM

@Marc But what often happens is that you one hasn't one .or. the other , but one preferentially over the other (i.e., a blend). Cf. Tautobase, https://github.com/WahlOya/Tautobase

Ulrich Schatzschneider to Everyone 08:17 PM

@Hinnerk Problem is how many applications actually use "non-standard InChI" or at least allow you to switch between "standard" and "non-standard"? If non-standard is there but not use beyond small developer community it is almost as good as not there at all (= not of help)

Marc Nicklaus to Everyone 08:17 PM

@Ulrich: Absolutely, in out Tautomer Database we see "blends" very frequently, just like in Tautobase.

Who can see your messages?

To: Everyone

Type message here...