Tautomerism in InChI V1.06

Marc C. Nicklaus
(NCI, NIH)

InChI Open Days, April 5, 2022

Working Group: Redesign of Handling of Tautomerism for InChI V2
Project No.: 2012-023-2-800

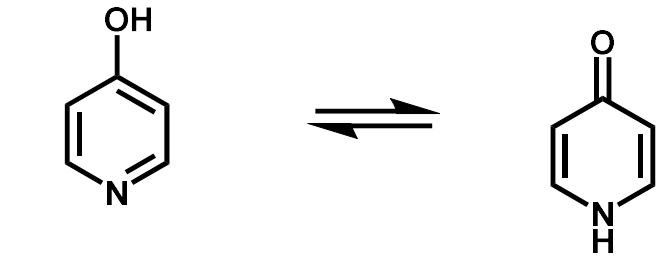
Tautomerism

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

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

- Prototropic tautomerism:

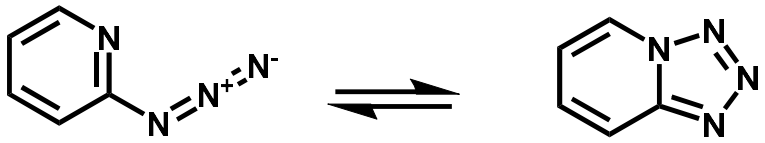
intramolecular movement of a hydrogen atom



enol form *keto form*

- Valence tautomerism:

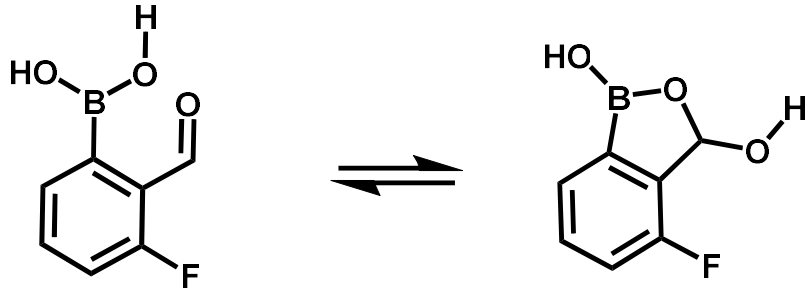
rearrangement of bonds w/o migration of atoms



azide form *tetrazole form*

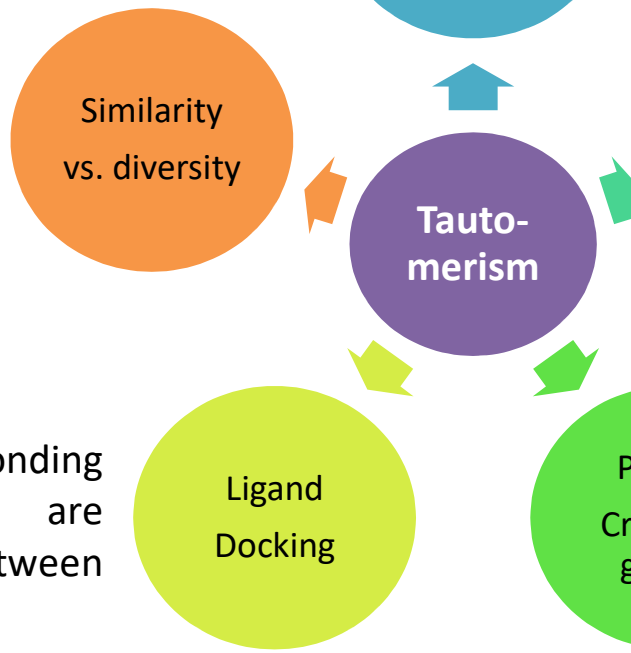
- Ring-chain tautomerism:

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



acyclic form *cyclic form*

Why worry about tautomers?



Variations across tautomers by several orders of magnitude
Eg. *pKa, logP*

(Tanimoto) similarity between tautomers can be very low

In many cases, there isn't a unique entry per molecule

Hydrogen bonding interactions are different between tautomers

What is the tautomeric form present in ligand-protein complexes?

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

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

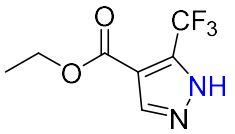
Aldrich Market Select (AMS) database :
5,755,574 molecules (2012-09 version)

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

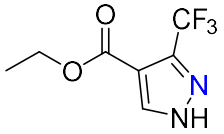
n-tuples	Conflicts
2	30,619
3	514
4	21
5	1

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

Examples (prices per 1 g):



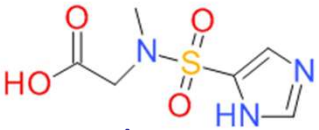
$31



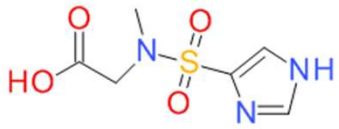
$251

$313

$350



$188



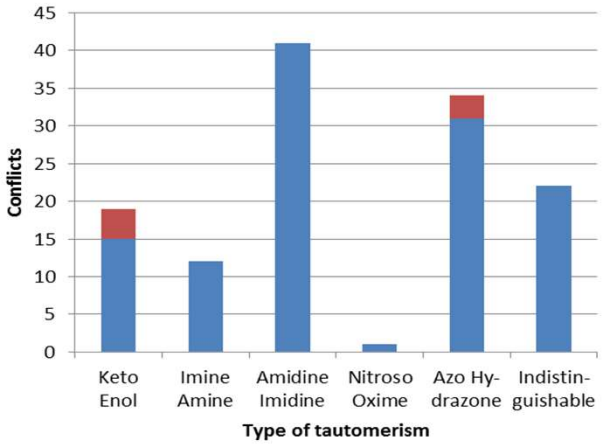
$300

Same original supplier!

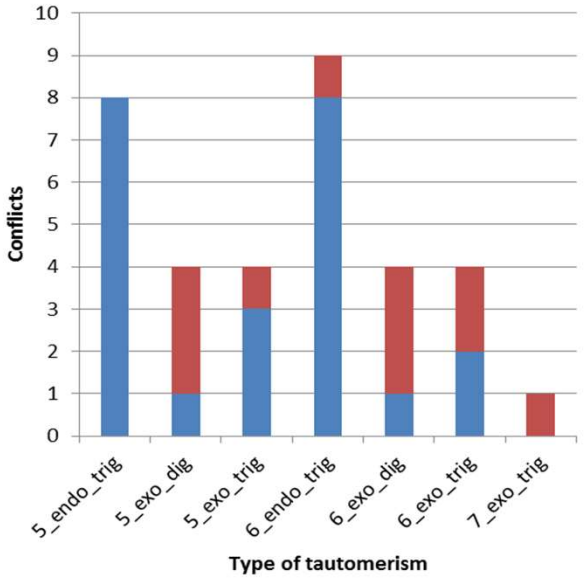
Experimental Verification

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.

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

How does InChI up to version 1.05 handle tautomerism?

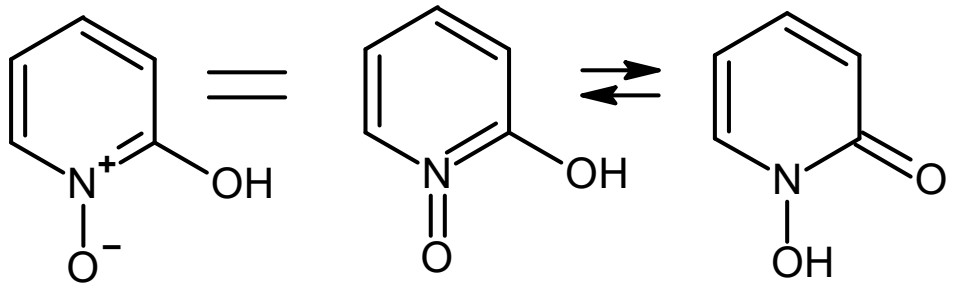
- InChI is in principle designed to be tautomer-invariant
- Standard InChI handles a limited range of tautomerism types
- One can turn on additional tautomeric types in non-standard InChI via options: KET, 15T
- It was recognized early on that important types of tautomerism are missing

Why a new version of tautomerism handling needed?

Proposal by Dmitrii Tchekhovskoi in 2012:

- Another breaking change:

Add 1,4-oxime/nitroso tautomerism

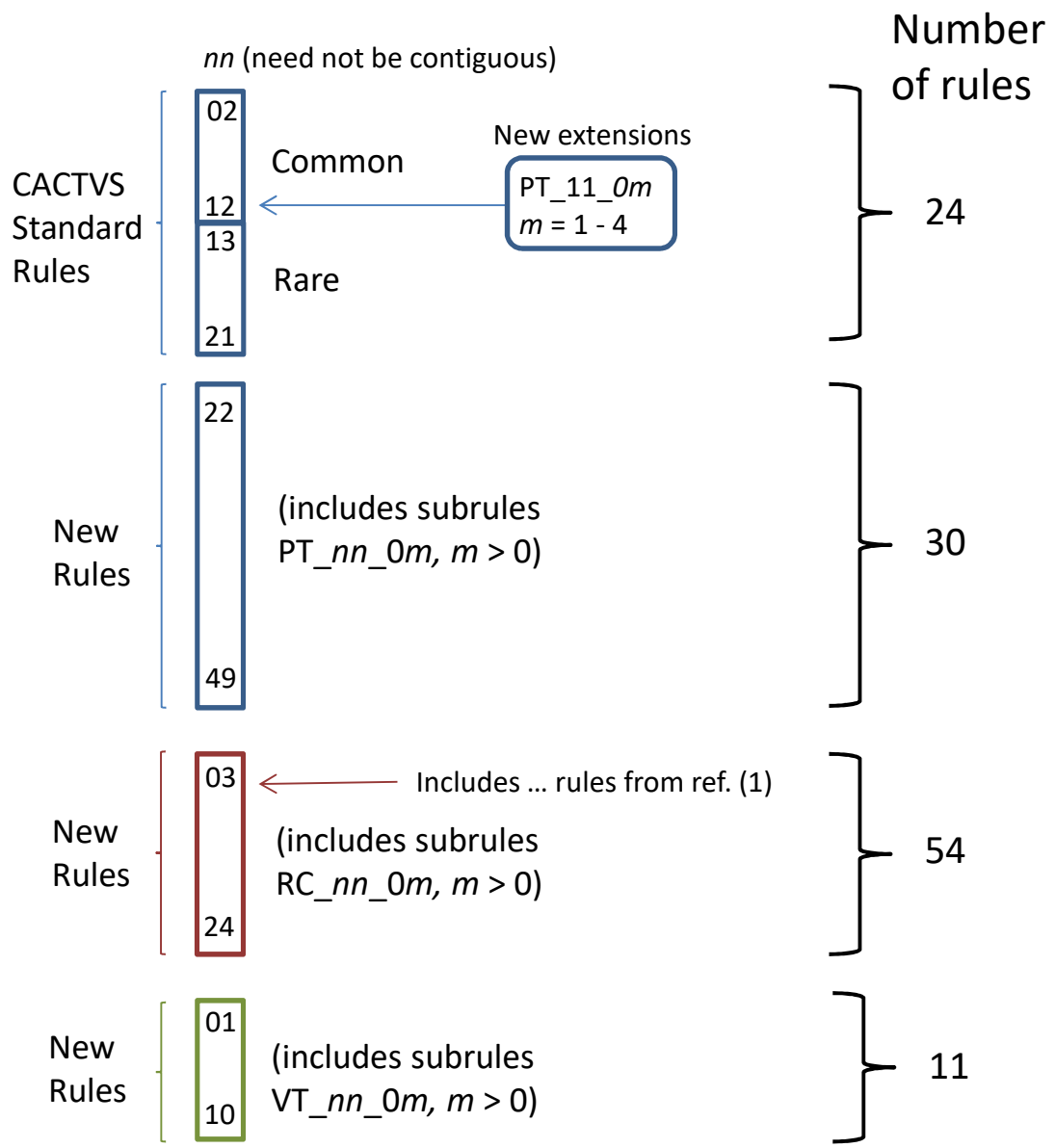


InChI=1S/C5H5NO2/c7-5-3-1-2-4-6(5)8/h1-4,7H

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

Rule Set

All rules expressed as SMIRKS, applied in CACTVS



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

Tautomer Enumeration Tool

https://cactus.nci.ih.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

Hitesh Patel

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

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

Rules applied in chemoinformatics toolkit CACTVS

Devendra Dhaked

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

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**

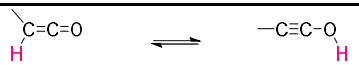
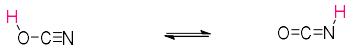
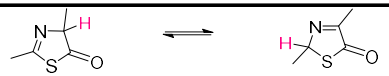
Note: These were all the rules that could be added; for others, the effort was unsuccessful!


Igor Filippov

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


Wolf-D. Ihlenfeldt

New Rules Implemented

PT_06_00	 1,3 heteroatom H-shift	[CX{2-3}z{0-1},N,n,S,s,O,o,Se,Te:1]=[NX2,nX2,CX3,c,P,p:2][N,n,S,O,Se,Te:3][#1:4] >>[#1:4][CX4z{0-1},N,n,S,O,Se,Te:1][NX2,nX2,CX3z{0-1},c,P,p:2]=[N,n,S,s,O,o,Se,Te:3]
PT_13_00	 keten-inol 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+0z1:3]=[CX3:4]>>[O;!R:2]=[N+0z1:3][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/azoxy or Behrend 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!

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

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

How Does InChI Perform on "Tautomer Database"?

3380 unique Standard InChIKeys found

2416 unique non-standard InChIKeys (with KET and 15T turned on) found

2210 unique Tauto InChIKeys (with all 8 rules turned on) found

2416 unique Tauto InChIKeys with KET and 15T turned on found for the PubChem file(s)

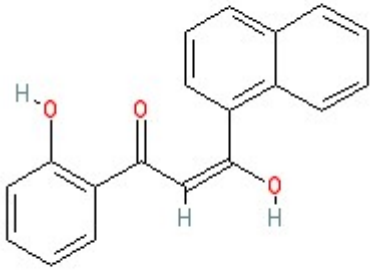
2210 (-8.526%, -206) unique Tauto InChIKeys (with KET, 15T and all 6 new rules by Igor F. turned on) found
2961 (22.558%, 545) unique Tauto InChIKeys with just KET turned on found
2809 (16.267%, 393) unique Tauto InChIKeys with just 15T turned on found
2240 (-7.285%, -176) unique Tauto InChIKeys with just PT_06_00 turned on found
3379 (39.859%, 963) unique Tauto InChIKeys with just PT_13_00 turned on found
3380 (39.901%, 964) unique Tauto InChIKeys with just PT_16_00 turned on found
3380 (39.901%, 964) unique Tauto InChIKeys with just PT_18_00 turned on found
3353 (38.783%, 937) unique Tauto InChIKeys with just PT_22_00 turned on found
3364 (39.238%, 948) unique Tauto InChIKeys with just PT_39_00 turned on found

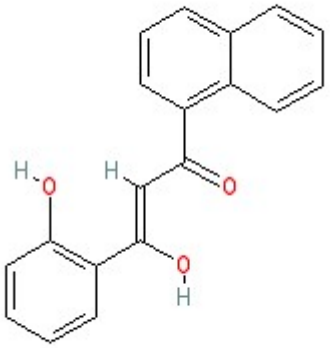
"Ideal" InChI algorithm would generate 1776 unique InChI[Key]s

How do the variants of InChI perform against each other:

3884 - 3380 = **504 cases of tautomeric overlap** found for Standard InChIKey
3884 - 2416 = 1468 for non-standard InChIKey (with KET and 15T turned on) – **1381 tuples found**
3884 - 2210 = 1674 for Tauto InChIKey (with KET, 15T, 6 new rules) – **additional 191 tuples found**
638 structures not identified as tautomers of any other structure = **319 tautomeric pairs missed**

Non-standard InChI and Tauto InChI Identify Tautomers With Same Identifier

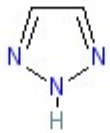


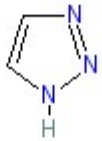


Non-standard InChIKey:	NIMVZAHKIIUSFT-UHFFFAOYNA-N	NIMVZAHKIIUSFT-UHFFFAOYNA-N
Tauto InChIKey:	NIMVZAHKIIUSFT-UHFFFAOYNA-N	NIMVZAHKIIUSFT-UHFFFAOYNA-N
[Standard InChIKey:	NIMVZAHKIIUSFT-UHFFFAOYNA-N	VABFYUCCCQCHOM-UNOMPAQXSA-N]

Case: 6_01 and _02

Non-standard InChI and Tauto InChI Identify Tautomers With Different Identifiers

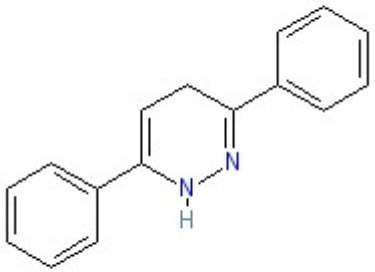


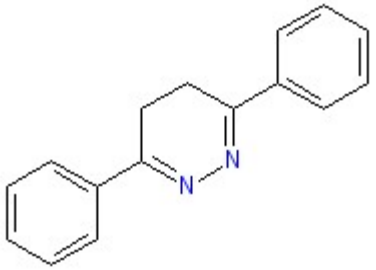


Non-standard InChIKey:	QWENRTYMTSOGBR-UHFFFAOYNA-N	QWENRTYMTSOGBR-UHFFFAOYNA-N
Tauto InChIKey:	WCHQBIYPPGCACF-UHFFFAOYNA-N	WCHQBIYPPGCACF-UHFFFAOYNA-N

Case: 2_01 and _02

Only Tauto InChI Identifies Tautomers

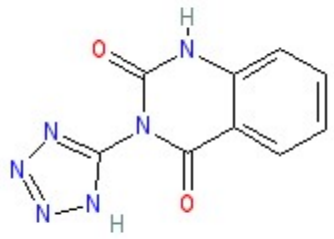


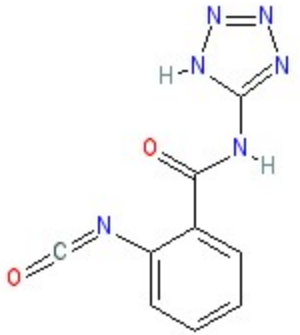


Non-standard InChIKey:	UHFJDRXFXQDUHB-UHFFFAOYNA-N	GEHJYGJMMCSVQL-UHFFFAOYNA-N
Tauto InChIKey:	YAMMWMQCRXUUCQ-UHFFFAOYNA-N	YAMMWMQCRXUUCQ-UHFFFAOYNA-N

Case: 115_01 and _02

No InChI Identifies Tautomers





Non-standard InChIKey:	BBLOKLTVQNRBCZ-UHFFFAOYNA-N	NGZXTQLEPBMFOH-UHFFFAOYNA-N
Tauto InChIKey:	FQQIRUPVTOGMNW-UHFFFAOYNA-N	DMNYUUUGNPYNGA-UHFFFAOYNA-N

Case: 205_01 and _02

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

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**

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 (†)