Tautomers in InChI

Marc C. Nicklaus

NIH InChI Workshop, March 22-24, 2021

IUPAC Project #2012-023-2-800 “Redesign of Handling of Tautomerism for InChI V2”
https://iupac.org/project/2012-023-2-800

CADD Group

Chemical Biology Laboratory
Center for Cancer Research
National Cancer Institute
National Institutes of Health

National Cancer Institute

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health

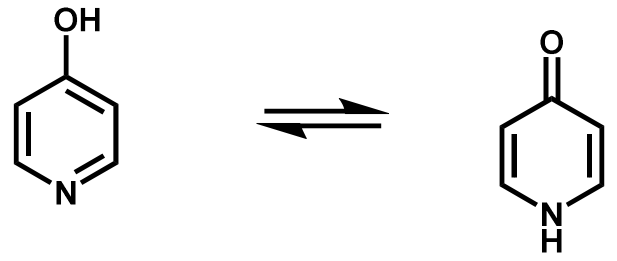
Tautomerism

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

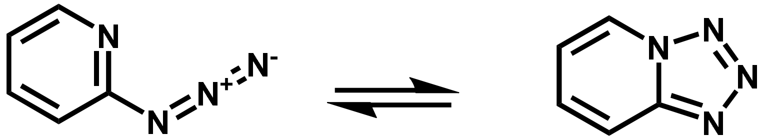


enol form

keto form

- Valence tautomerism:

rearrangement of bonds w/o migration of atoms

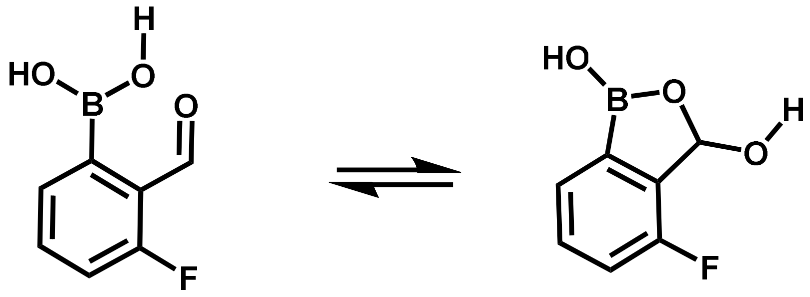


azide form

tetrazole form

- Ring-chain tautomerism:

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



acyclic form

cyclic form

How does current InChI handle tautomerism?

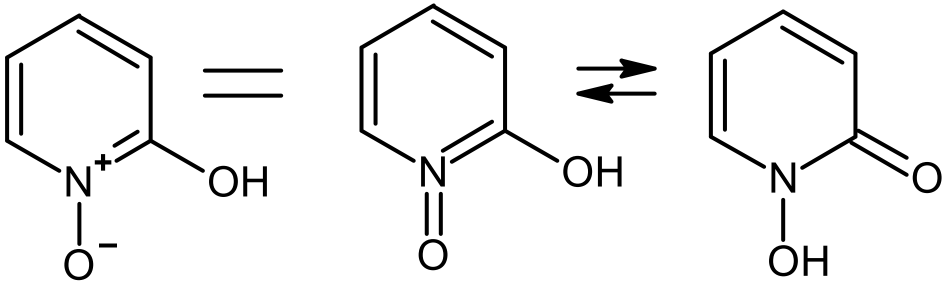
- InChI is in principle designed to be tautomer-invariant
- Standard InChI handles a limited range of tautomerism types
- One can turn on additional tautomeric types in non-standard InChI via options: KET, 15T
- It was recognized early on that important types of tautomerism are missing

Proposal by Dmitrii Tchekhovskoi in 2012

Why new version

- 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

InChI[Key] only Partially Recapitulates a More Complete Set of Rules

InChI Calculation Type >		Standard	{DONOTADDH W0}	
	Database	Tautomeric	InChI Success	Strict InChI
Database	Size	Part	Rate (%)	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

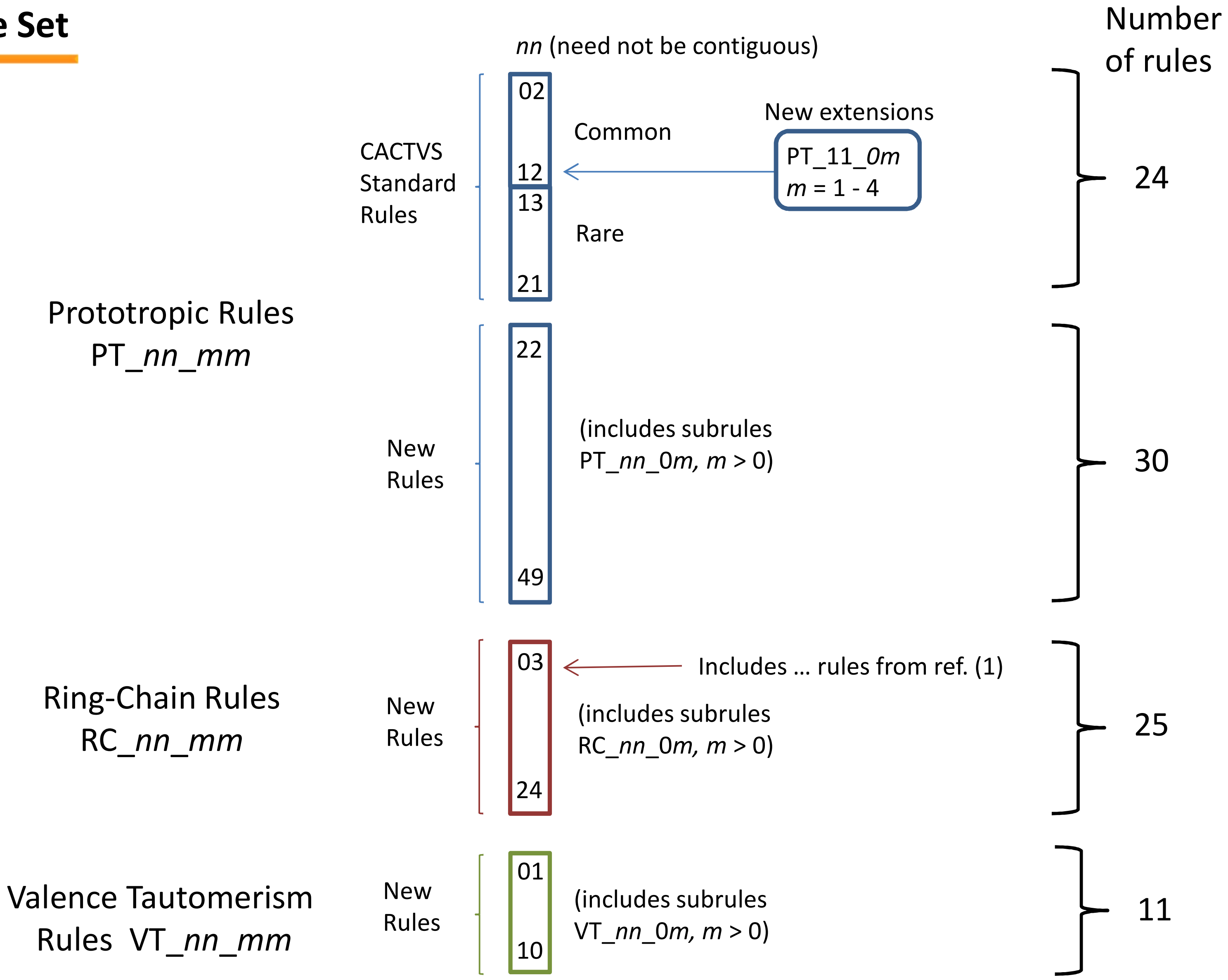
InChI Calculation Type >		Non-standard	{DONOTADDH W0 RECMET NEWPS SPXYZ SAsXYZ Fb Fnud KET 15T}	
	Database	Tautomeric	InChI Success	Strict InChI
Database	Size	Part	Rate (%)	Success Rate (%)
CSD	319,201	203,108	48.83	30.90
ChEMBL	1,820,035	1,578,290	73.91	37.46
AMS	8,409,644	7,204,965	71.99	36.32
PUBCHEM	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

Rule Set



nn (need not be contiguous)

Number of rules

Prototropic Rules PT_*nn*_*mm*

CACTVS Standard Rules

02

Common

12

13

Rare

21

New extensions

PT_11_*0m*
m = 1 - 4

24

New Rules

22

(includes subrules PT_*nn*_0*m*, *m* > 0)

49

30

Ring-Chain Rules RC_*nn*_*mm*

New Rules

03

Includes … rules from ref. (1)

(includes subrules RC_*nn*_0*m*, *m* > 0)

24

25

Valence Tautomerism Rules VT_*nn*_*mm*

New Rules

01

(includes subrules VT_*nn*_0*m*, *m* > 0)

10

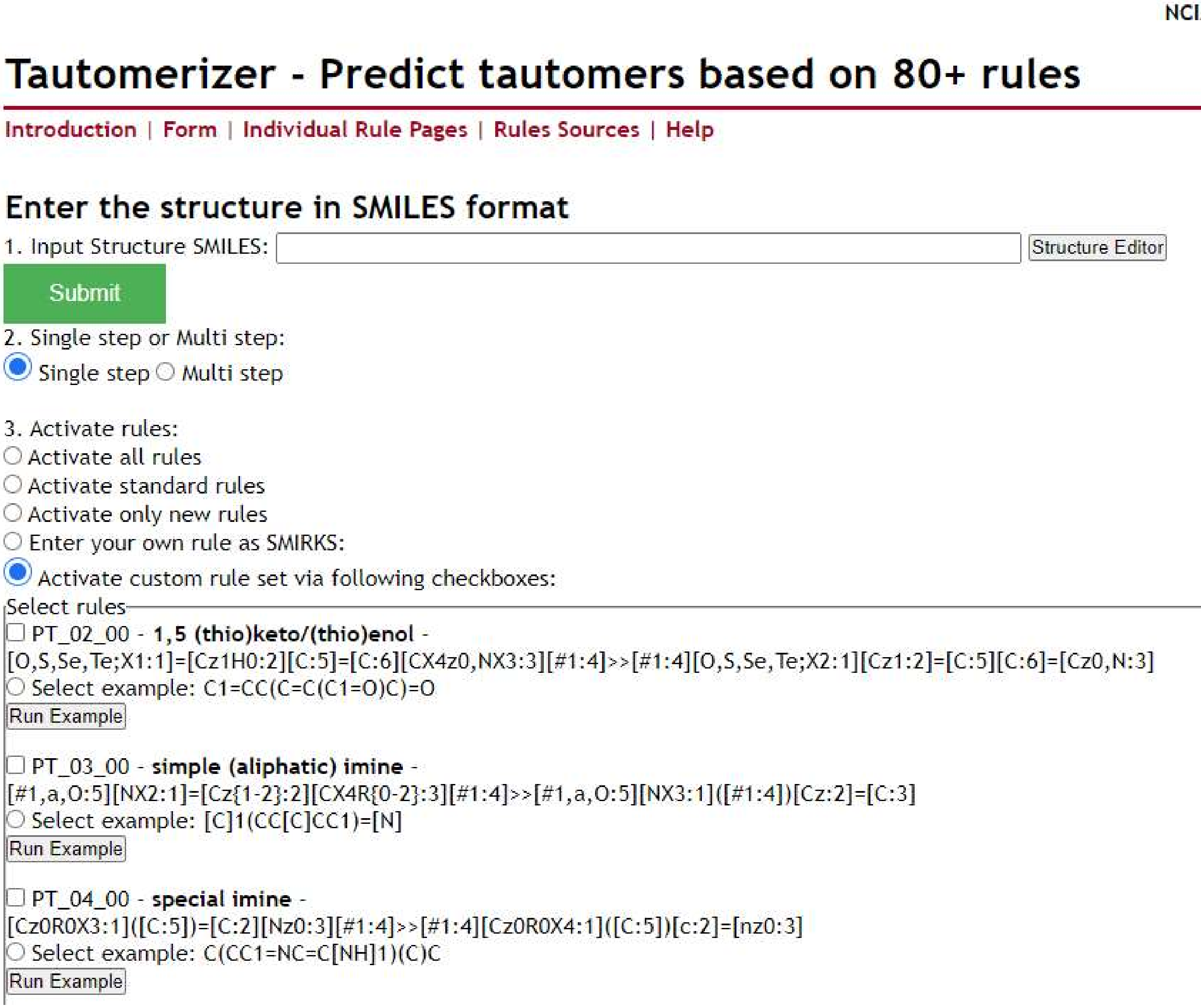
11

(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

Total number of rules: 90

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]

Hitesh Patel

New Rules: How, and which ones, to integrate in InChI

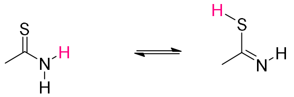
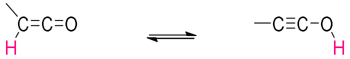
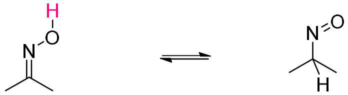
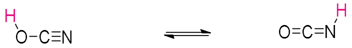
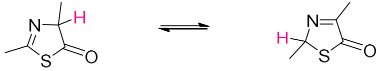
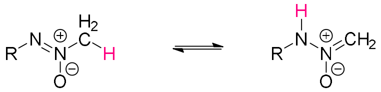
- New rules, as implemented in CACTVS, expressed as SMIRKS
- InChI doesn’t have a SMIRKS parser
- Adding new tautomeric rules requires code changes in the core of InChI

- 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

Igor Filippov

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?

Six new rules implemented in InChI library (based on V. 1.06 code) integrated in CACTVS.
== This is currently the only available implementation of these rules in InChI ==

Wolf-D. Ihlenfeldt

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

71,600,000 compounds analyzed

71,409,375 (100%) unique Standard InChIKeys
68,893,074 (-3.52%) unique Non-standard InChIKeys (with KET and 15T options turned on)
66,353,137 (-7.08%) unique Tauto InChIKeys (with KET, 15T and all 6 new rules by Igor F. turned on)

Difference between Standard and Non-standard counts:	2,561,301
Difference between Standard and Tauto InChIKey counts:	5,056,238

Note: Numerous (non-standard) InChIKey values change when 6 new rules are turned on

Summary, Conclusions, and Questions for the Community

- Typically >80% of compounds in databases are amenable to one or more of 90 tautomeric rules
- Number of affected compounds per rule varies widely

- Current Standard InChI recapitulates ~30% of amenable compounds
- Current Non-Standard InChI (KET, 15T) recapitulates ~37% of compounds
- Only 3 out of 90 rules have Non-Standard InChI Success rates > 90%
- Only 7 rules have Non-Standard InChI Success rates > 50%
- 57 rules have Non-Standard InChI Success rates = 0%
- Question: Which ones are realistic, which ones may be too strict?

- Six new prototropic rules could be added to InChI code (and no, not 1,4-oxime/nitroso)

- Relative to Standard InChI, Non-Standard InChI (KET, 15T) equates 3.5% more compounds as tautomers of other compounds
- 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

- When to release InChI with the 6 new rules? In version 1.06x? Or 1.07? Or wait for InChI V.2?

- Prototropic transforms: doubtful whether more can be added to InChI
- 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

- Devendra Dhaked
- Laura Guasch
- Hitesh Patel
- Igor Filippov
- Wolf-Dietrich Ihlenfeldt
- Jeff Saxe

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
Igor Pletnev
Roger Sayle
Dmitrii Tchekhovskoi