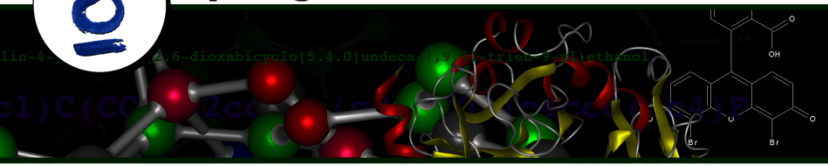
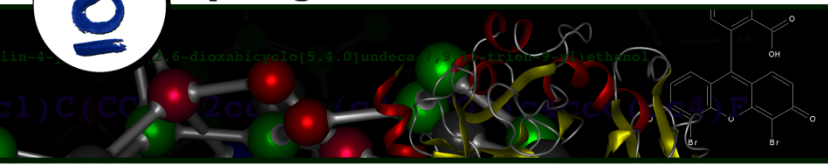


Canonical Line Notations

InChI vs SMILES

Krisztina Boda



Overview

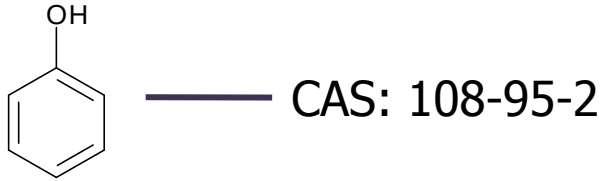
- Compound naming
- InChI
- SMILES
- Molecular equivalency
  - Isomorphism
  - Kekule
  - Tautomers
- Finding duplicates

What's Your Name?

1. Unique numbers

- CAS registry number
- BRN Beilstein Registry Number
- EC (European commission) number
- CID (PubChem compound id)

one-to-one

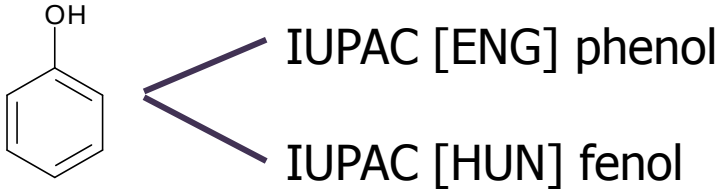


2. Chemical Nomenclatures

- IUPAC name(s)
- Traditional name(s)

  [carbolic acid, benzenol, phenylic acid, std.]

one-to-many

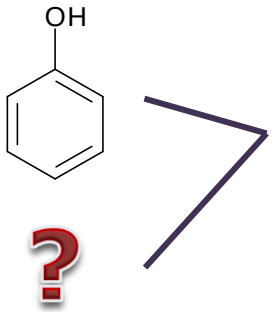


What's Your Name?

3. Hashed identifiers

- NCI/CADD structure identifiers
- InChIKey

many-to-one

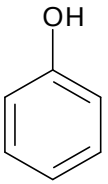


InChIKey= ISWSIDIOOBJBQZ-UHFFFAOYSA-N

4. Chemical line notations

- WLN (Wiswesser Line Notation)
- **SMILES**
- **InChI**

one-to-one ?



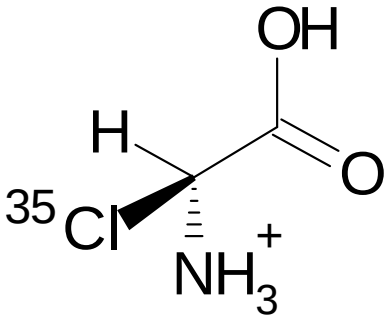
InChI=1S/C6H6O/c7-6-4-1-3-5-6/h1-5,7H

c1ccccc1O

InChI

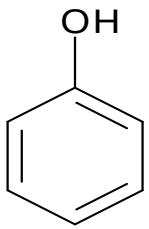
InChI Identifier (**In**ternational **Ch**emical **I**dentifier)

- Developed by IUPAC and NIST [first release 2005]
- One reference implementation
- Multiple hierarchical layers
  - No bond order
- Not a file format!



SMILES

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**pecification)

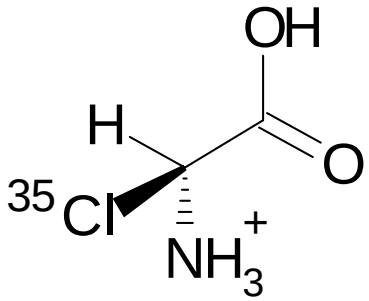


c1ccccc1O
Oc1ccccc1
C1=CC=CC=C1O
C1=CC=C(C=C1)O
[CH]1=[CH][CH]=[CH][CH]=C1[OH]

- Canonical SMILES [OEChem: c1ccc(cc1)O]
  - Unique name for each molecule in one system
  - Not a global identifier
- Canonical Isomeric SMILES
  - Encode isotope, double bond and chiral configuration

C(C(=O)O)([NH3+])Cl

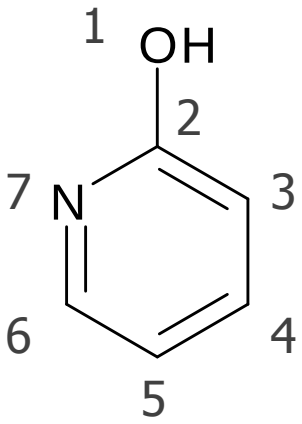
[C@H](C(=O)O)([NH3+])[35Cl]



[1] SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules
David Weininger, J. Chem. Inf. Comput. Sci., **1988** (28)

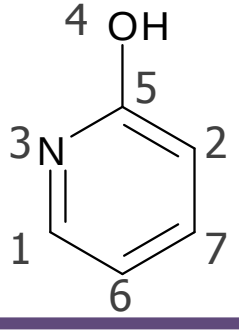
Molecule Registration

Is this already in my database?



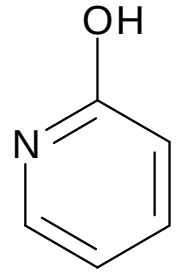
=

Isomorph



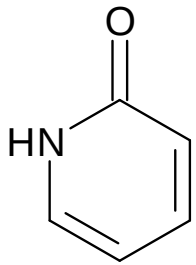
=

Kekule

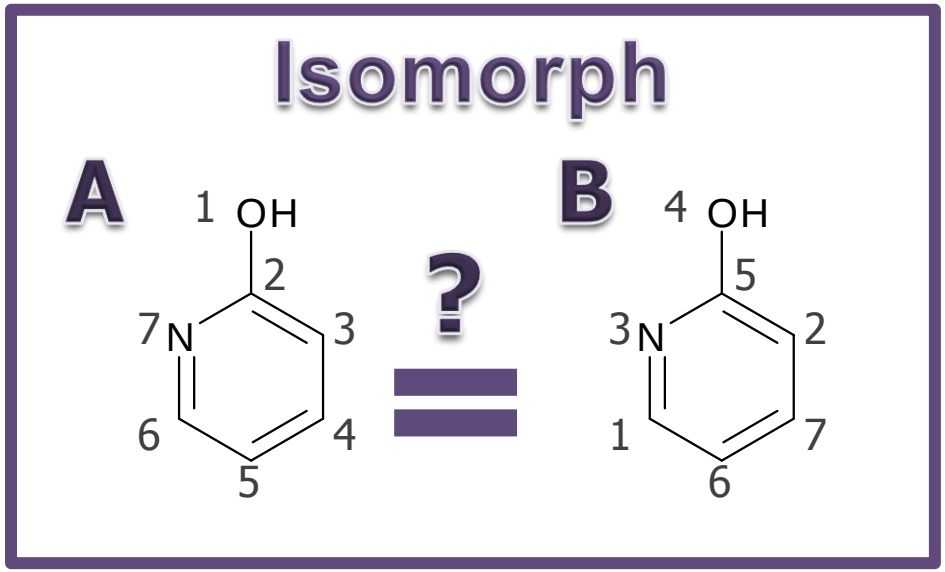


=

Tautomer

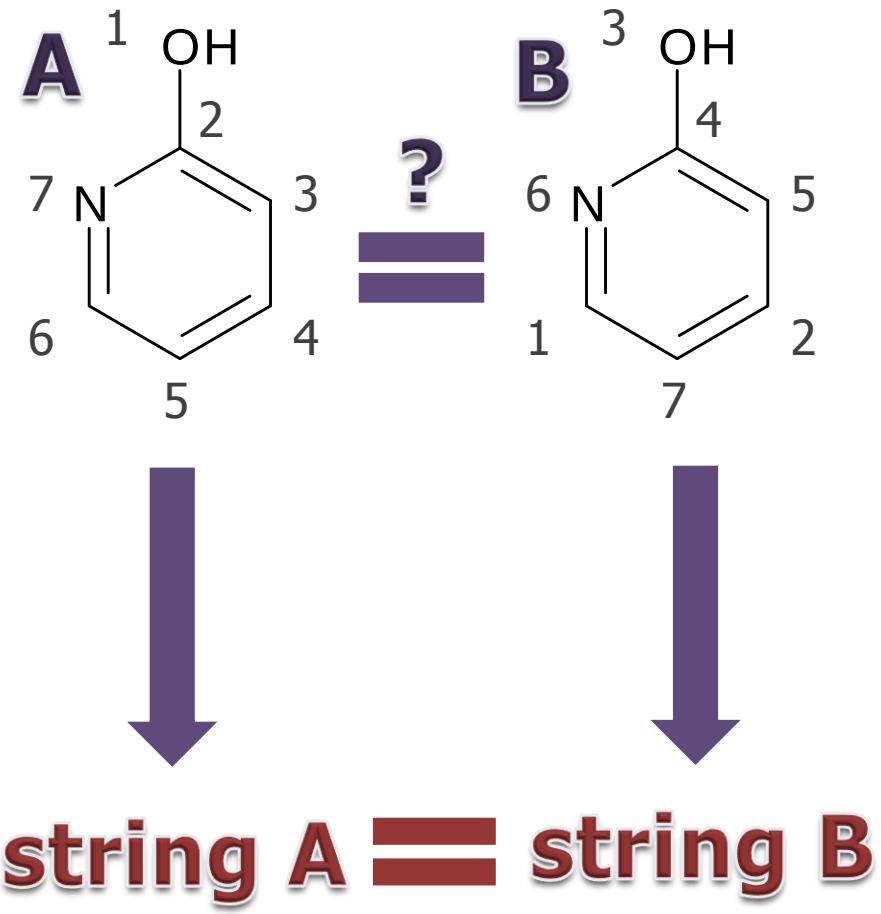


Molecule Equivalency - Isomorphism



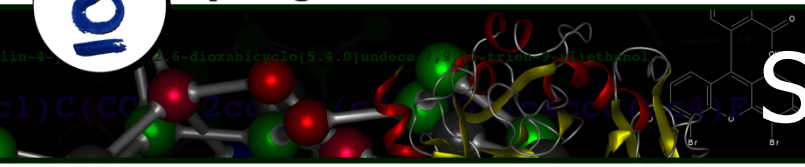
- NP problem
- Complexity = HvyAtm!
- Impractical for molecule registration!

Canonicalization



- Transforms any connection table into a unique canonical form (independent of atom indices)
- Comparing molecules ⇨ comparing unique, unambiguous string representations

SMILES/InChI Generation



Input structure

Canonical SMILES (OpenEye)

Pre-processing

- aromaticity perception

Canonicalization (partitioning algorithm)

SMILES generation

InChI

Pre-processing [Normalization]

- bond normalization
- disconnect salts/metal atoms
- eliminating radicals
- normalize mobile hydrogens, variable protonation and charge
- tautomer detection

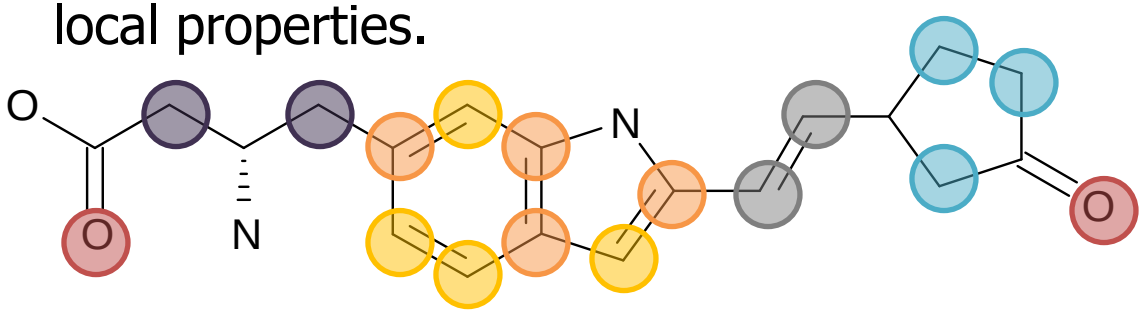
Canonicalization (Morgan algorithm)

InChI generation

Canonicalization

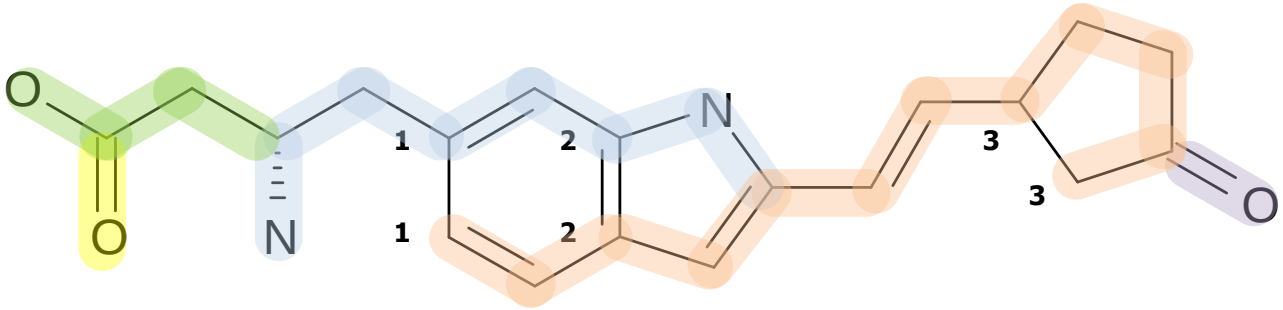
Finding unique atom ordering

1. Initial partitioning of atoms by distinguishing them by their local properties.



2. Recursively refining atom classes by distinguishing atoms by their neighbors.
3. When each atom belong to a unique partition ⇨ unique canonical atom order exists.

SMILES - Generation



c1cc2cc([nH]c2cc1C[C@@H](CC(=O)O)N)/C=C/C3CCC(=O)C3

Based on the canonical ordering of atoms, a unique SMILES is generated by depth-first search (DFS).

SMILES Validation

Test Set
10M compounds
(ACD, Aldrich, Maybridge, MDDR, NCI *etc.*)

repeat 100 ×

Randomly Renumbering Atoms

Canonical IsoSMILES Generation

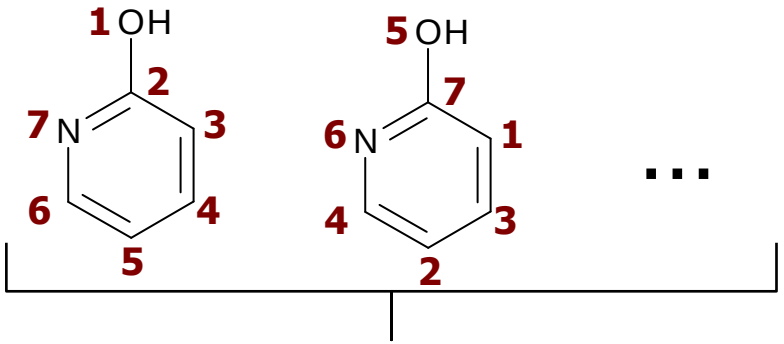
Identical String

no

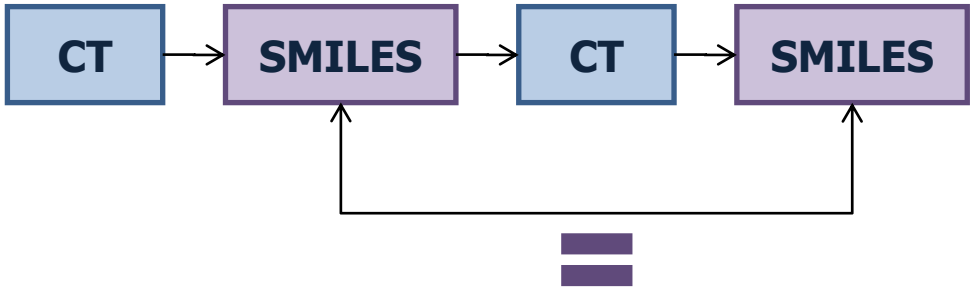
yes

FAILURES

Name has to be independent of atom numbering!

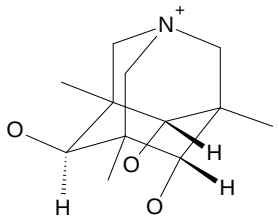


c1ccnc(c1)O



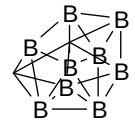
SMILES Validation

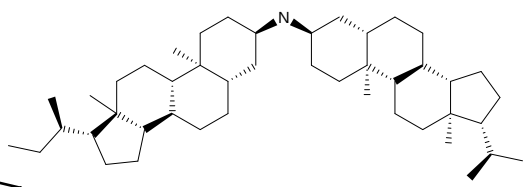
7

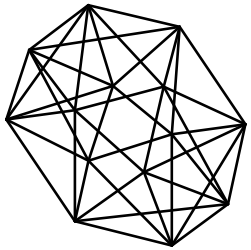


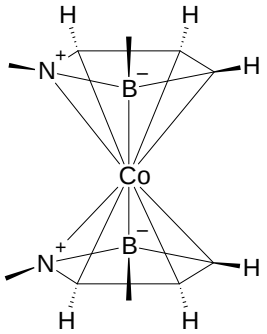
adamantane-like ring systems

63



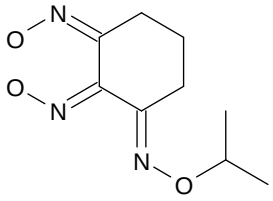






highly-complex ring systems

8



SMILES bond stereo representation problem

Number of total failures: 78 (0.0008%)

InChI Validation

Test Set
10M compounds
(ACD, Aldrich, Maybridge, MDDR, NCI *etc.*)

repeat 100 ×

Randomly Renumbering Atoms

InChI Generation

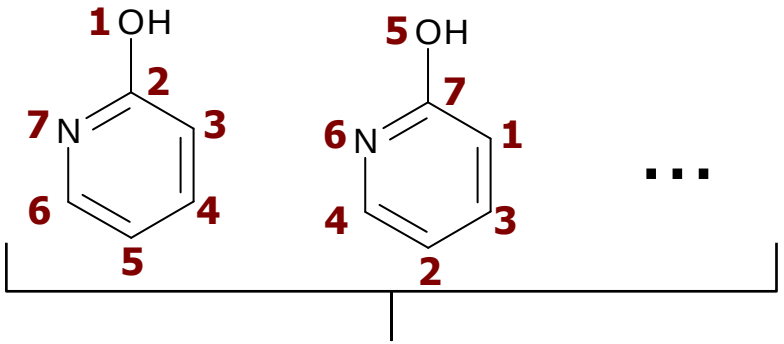
Identical String

no

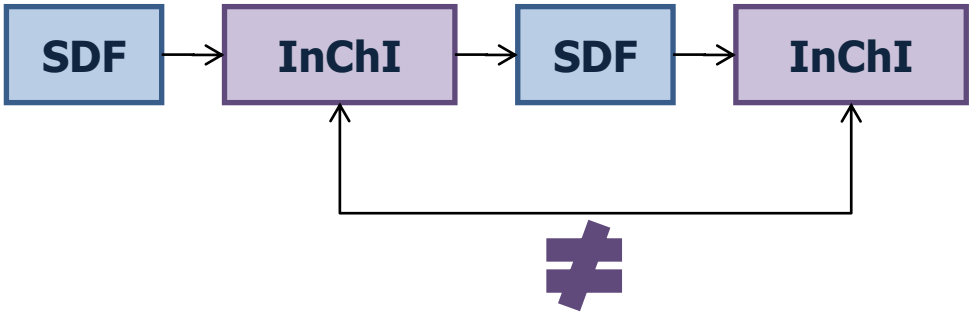
yes

FAILURES

Name has to be independent of atom numbering!

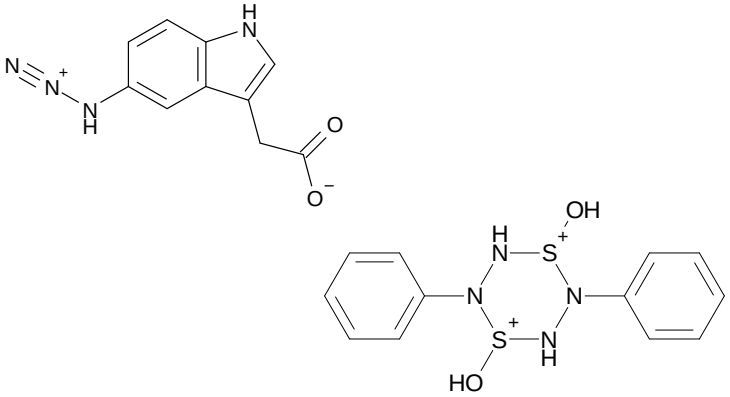


InChI=1S/C5H5NO/c7-5-3-1-2-4-5/h1-4H,(H,6,7)



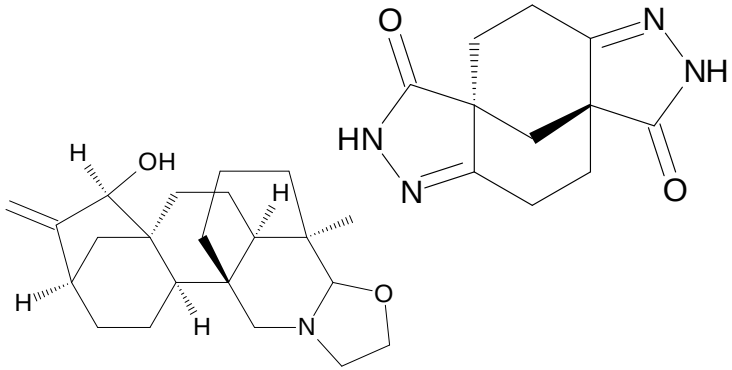
InChI Validation

28



charge layer

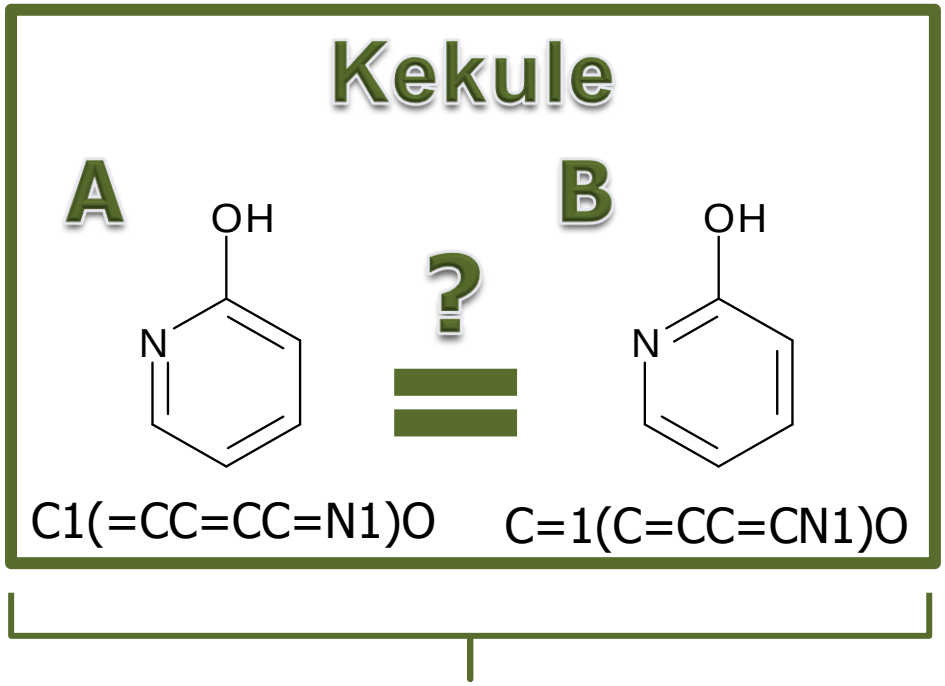
27



stereo layer

Number of total failures: 55 (0.0006%)

Molecule Equivalency (Kekule)



Canonical SMILES

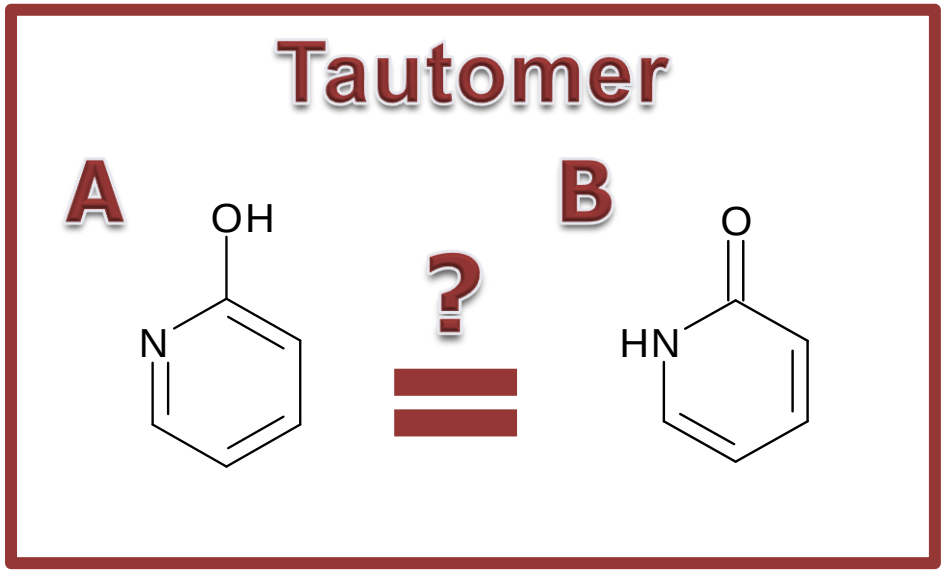
- aromatic form (no bond between aromatic atoms)

InChI

- no bond information

InChI=1S/c5H5NO/c7-5-3-1-2-4-6-5/h1-4H,(H,6,7)

Molecule Equivalency - Tautomer



- J. Comput.-Aided Mol. Des. (June 2010)

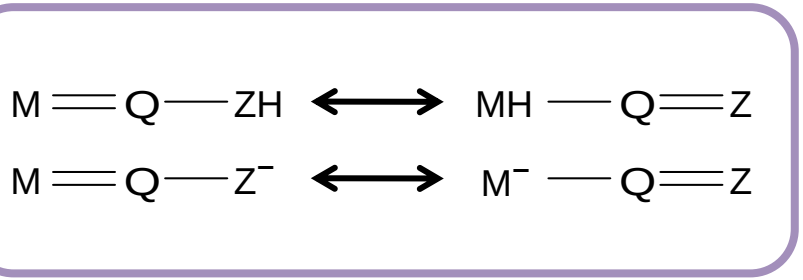
For registration:

Canonical tautomer does **not have** to be

- the chemically most reasonable
- the lowest in energy

Canonical tautomer **has to be** identical for every tautomer

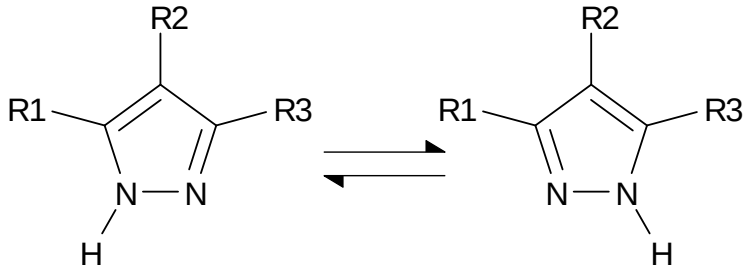
InChI- Tautomer Detection

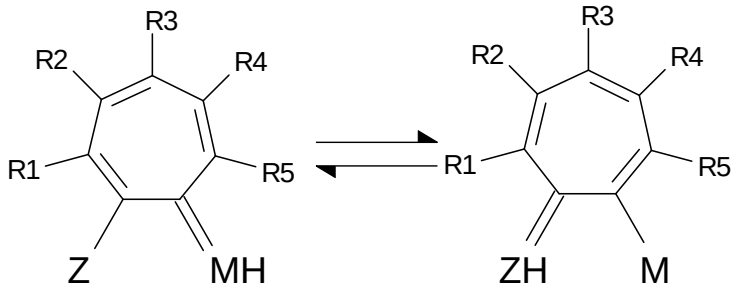


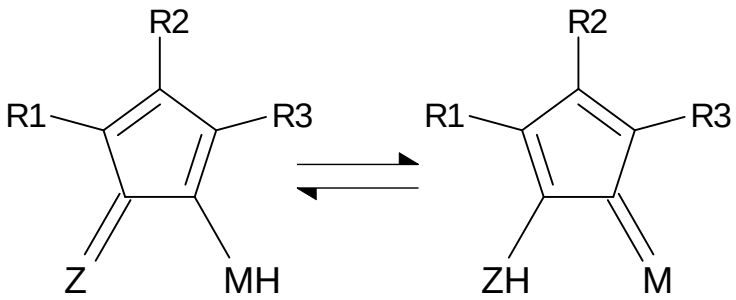
M,Z = trivalent N, bivalent O, S, Se, Te

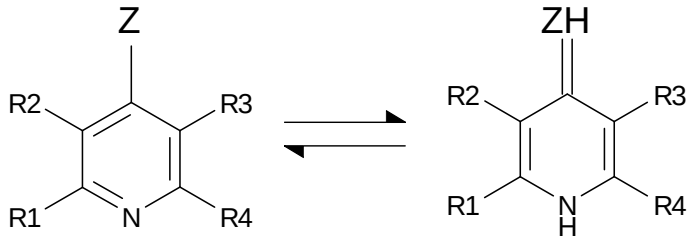
Q = C, N, S, P, Sb, As, Se, Te, Br, Cl, I

H = H, D, T

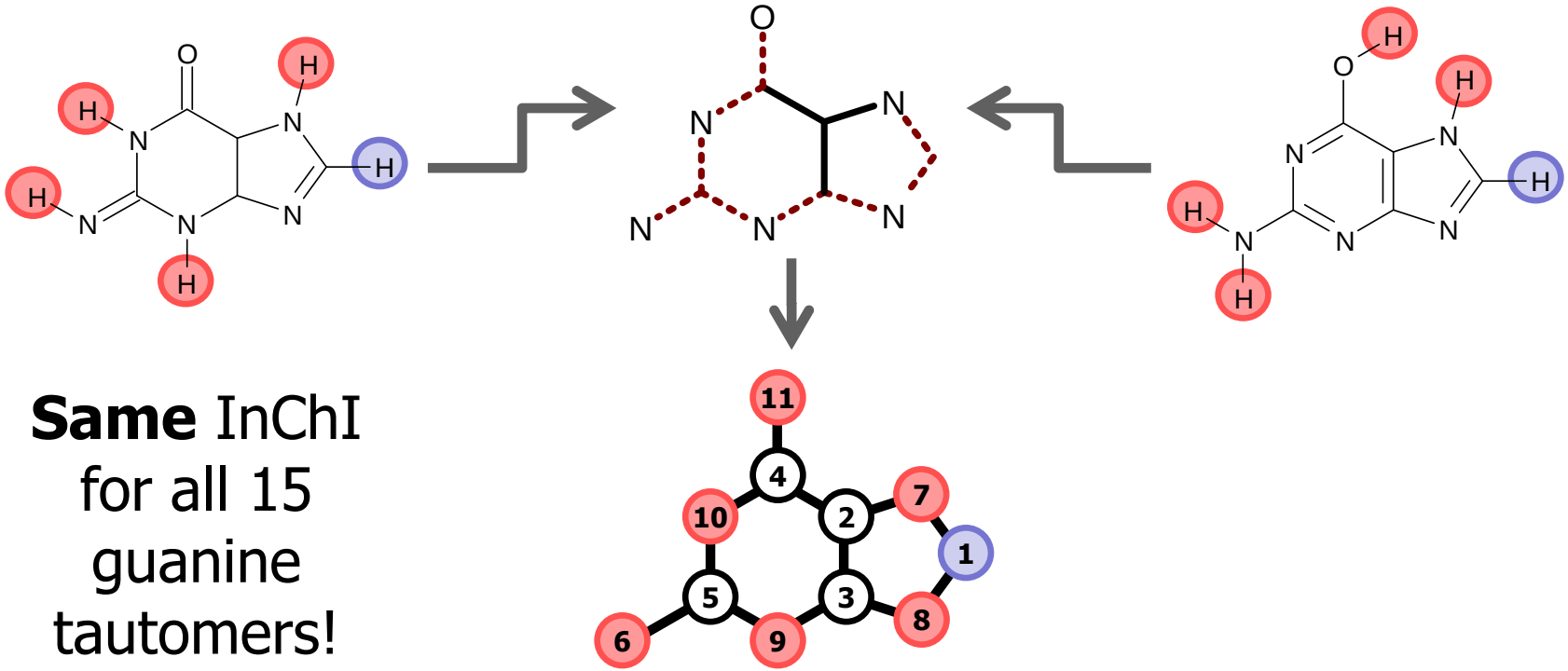








InChI - Handling Tautomers



Same InChI for all 15 guanine tautomers!

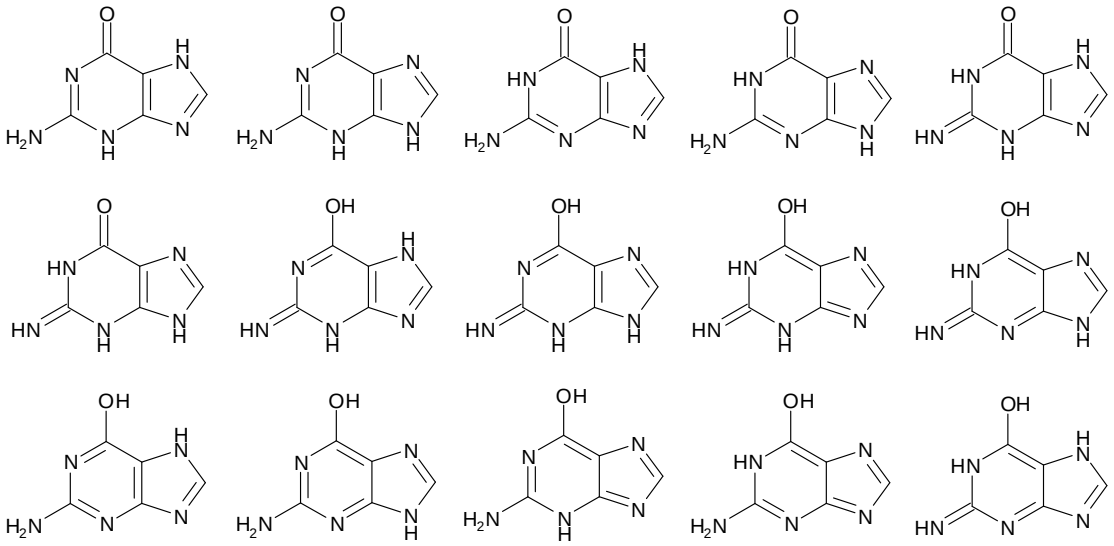
InChI = 1S/C5H5N5O/c6-5-93-2(4(11)10-5)7-1-8-3/h1H,(H4,6,7,8,9,10,11)

hydrogen layer

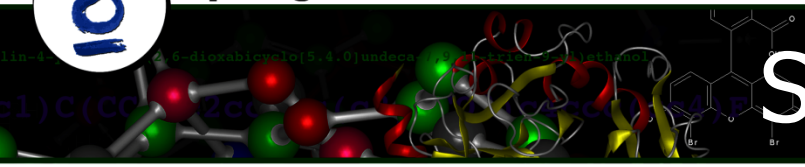
SMILES – Handling tautomers

- Identifying and handling tautomers is not part of the language
- Tautomeric structures are explicitly specified
  - No tautomeric bond
  - No mobile hydrogens

c1[nH]c2c(n1)[nH]c(nc2=O)N
c1[nH]c2c(=O)[nH]c(nc2n1)N
c1[nH]c2c(n1)c(=O)nc([nH]2)N
c1[nH]c2c(n1)c(=O)[nH]c(n2)N
c1[nH]c2c(n1)[nH]c(=N)[nH]c2=O
c1[nH]c2c(n1)c(=O)[nH]c(=N)[nH]2
c1[nH]c2c(n1)nc(nc2O)N
c1nc-2c(nc([nH]c2n1)N)O
c1[nH]c2c(n1)c(nc(n2)N)O
c1nc-2c([nH]c(nc2n1)N)O
c1[nH]c2c(n1)[nH]c(=N)nc2O
c1[nH]c2c([nH]c(=N)nc2n1)O
c1[nH]c2c(n1)c(nc(=N)[nH]2)O
c1nc-2c([nH]c(=N)[nH]c2n1)O
c1[nH]c2c(n1)c([nH]c(=N)n2)O



SMILES/InChI Generation



Canonical SMILES (OpenEye)

Input structure

InChI

Pre-processing

- aromaticity perception

QuacPac

Canonicalization (partitioning algorithm)

SMILES generation

Pre-processing [Normalization]

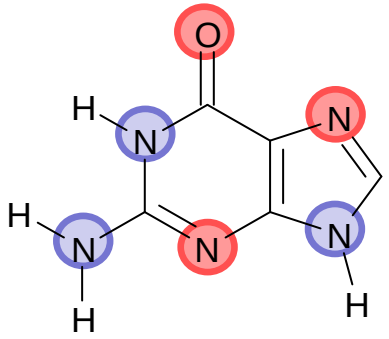
- bond normalization
- disconnect salts/metal atoms
- eliminating radicals
- normalize mobile hydrogens, variable protonation and charge
- tautomer detection

Canonicalization (Morgan algorithm)

InChI generation

QuacPac – Canonical Tautomer

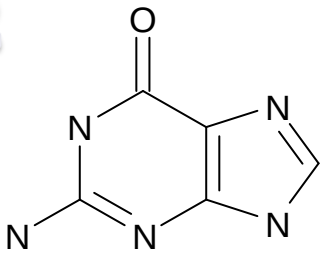
1



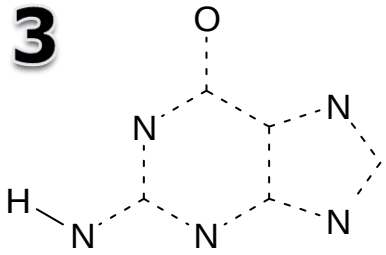
donor acceptor

Global algorithm that systematically generates **all** tautomers!

2

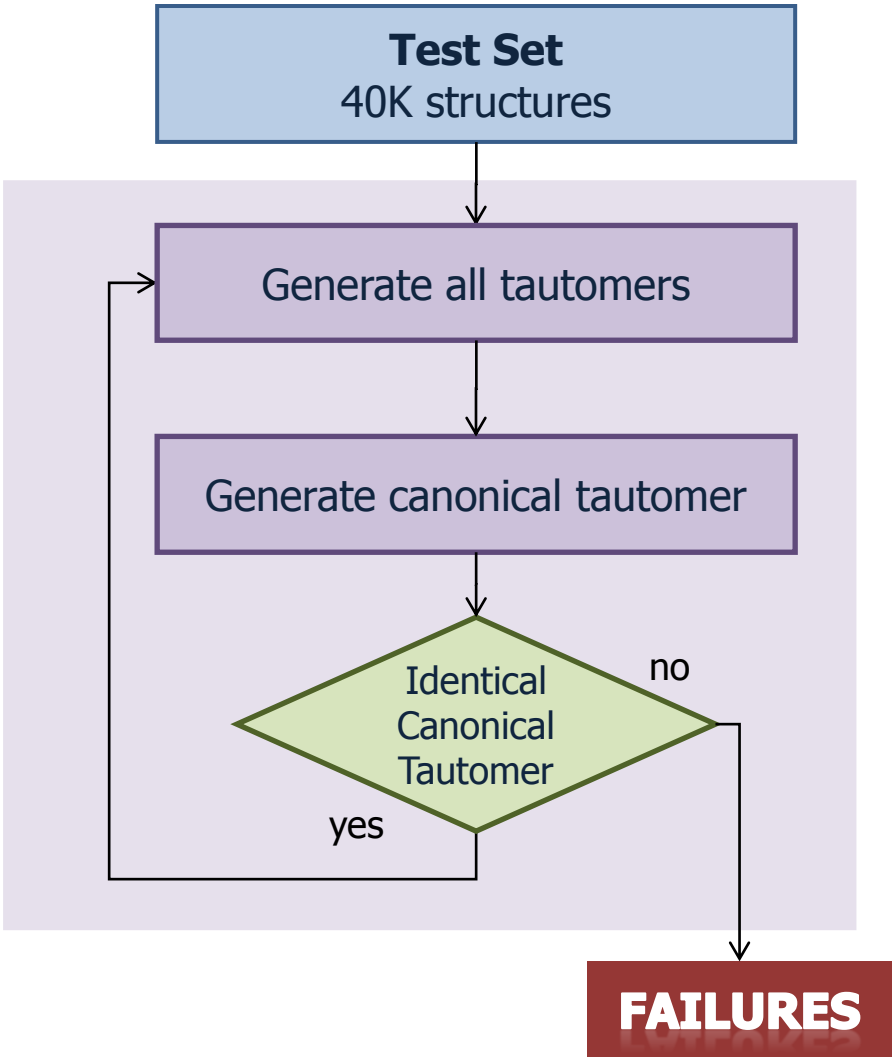


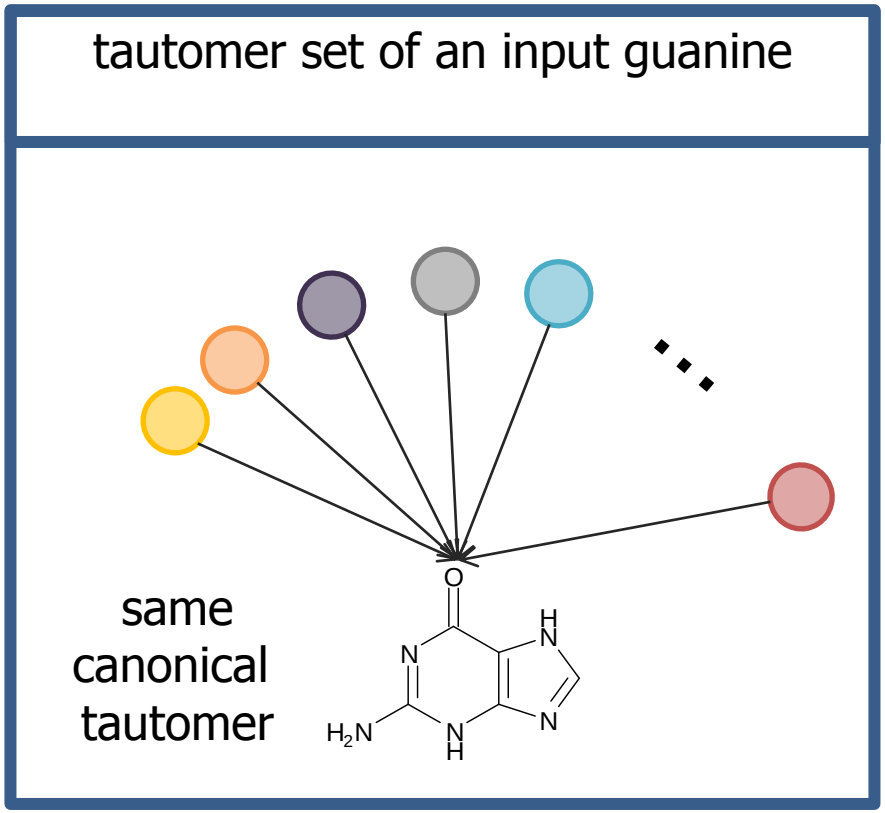
3



1. atom typing
2. removing (+counting) hydrogens
3. unspecifying bond types
4. enumerating tautomers (DFS)
   - protonating/deprotonating atom
5. Empirical, rule-based scoring to find the canonical tautomer

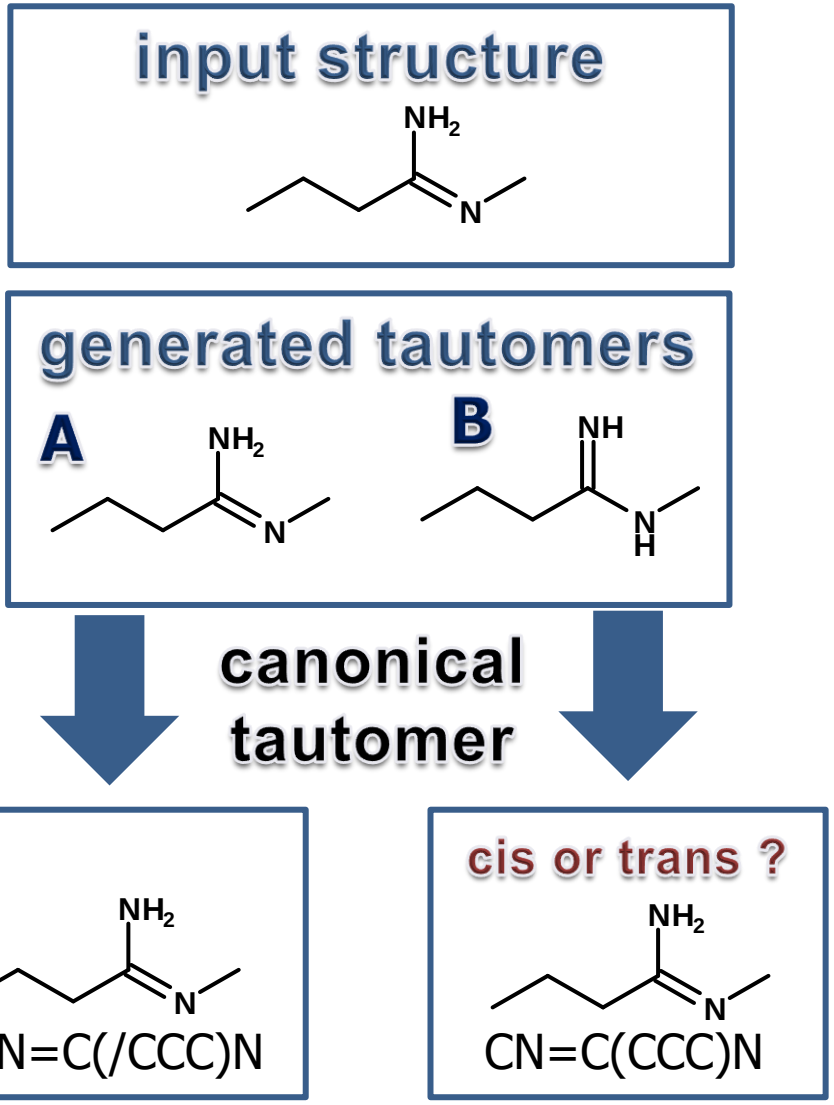
Validation – Canonical Tautomer



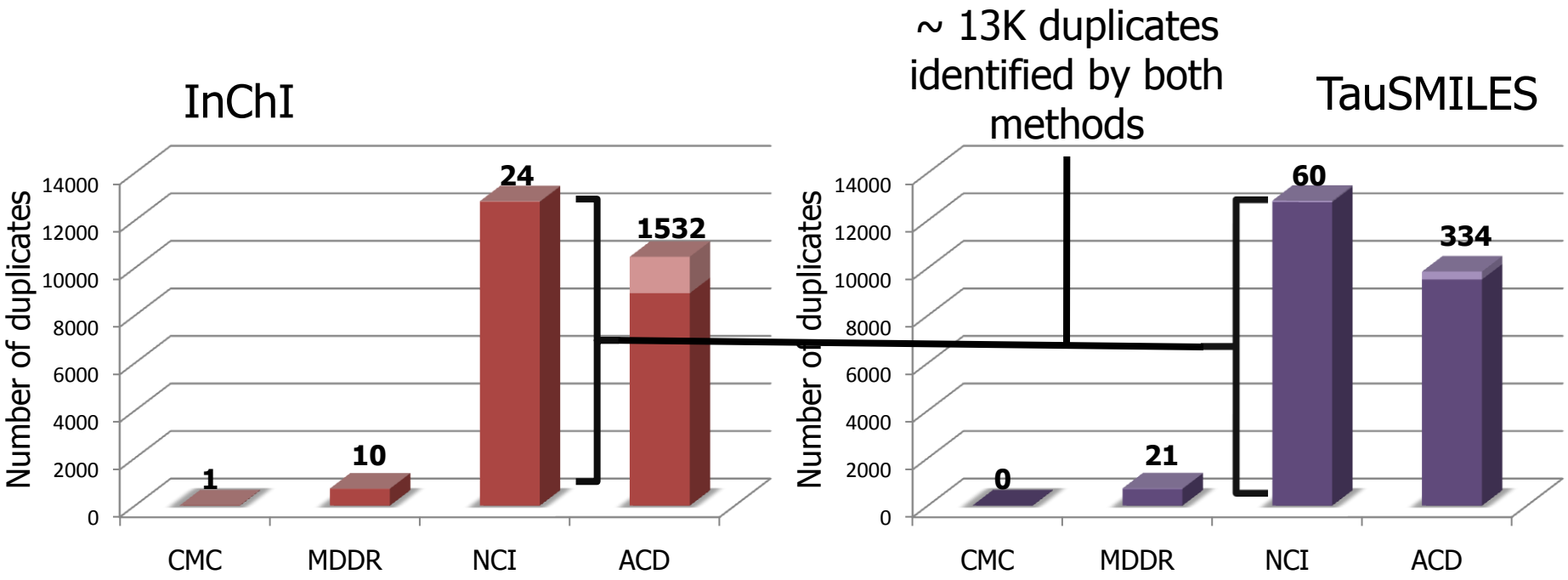


Tautomer Generation for SMILES

- Input: 40K
- Generated tautomers: 230K
- Failures: 272 (0.007%)
- Failures due to changing E/Z geometry 269



Identifying Duplicate Compounds

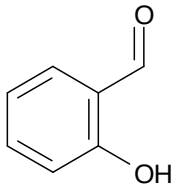


Both InChi & TauSMILES
Only InChI
Only TauSMILES

CMC	MDDR	NCI	ACD
14	704	12784	8940
1	10	24	1532
0	21	60	334

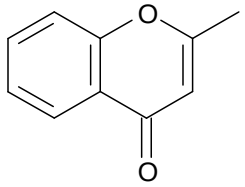
Duplicates

Isomorphs

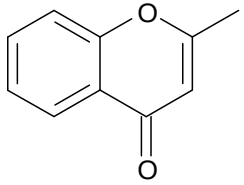


NSC numbers:
49178,51705,83559 ,83560,83561,
83562,97202,112278,187662

Kekules

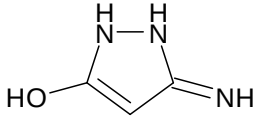


NSC numbers:
74869

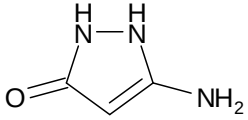


NSC numbers:
660048

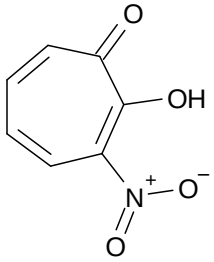
Tautomers



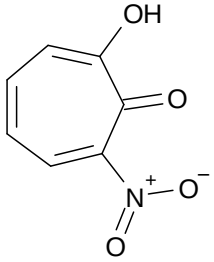
NSC numbers:
60188



NSC numbers:
408479

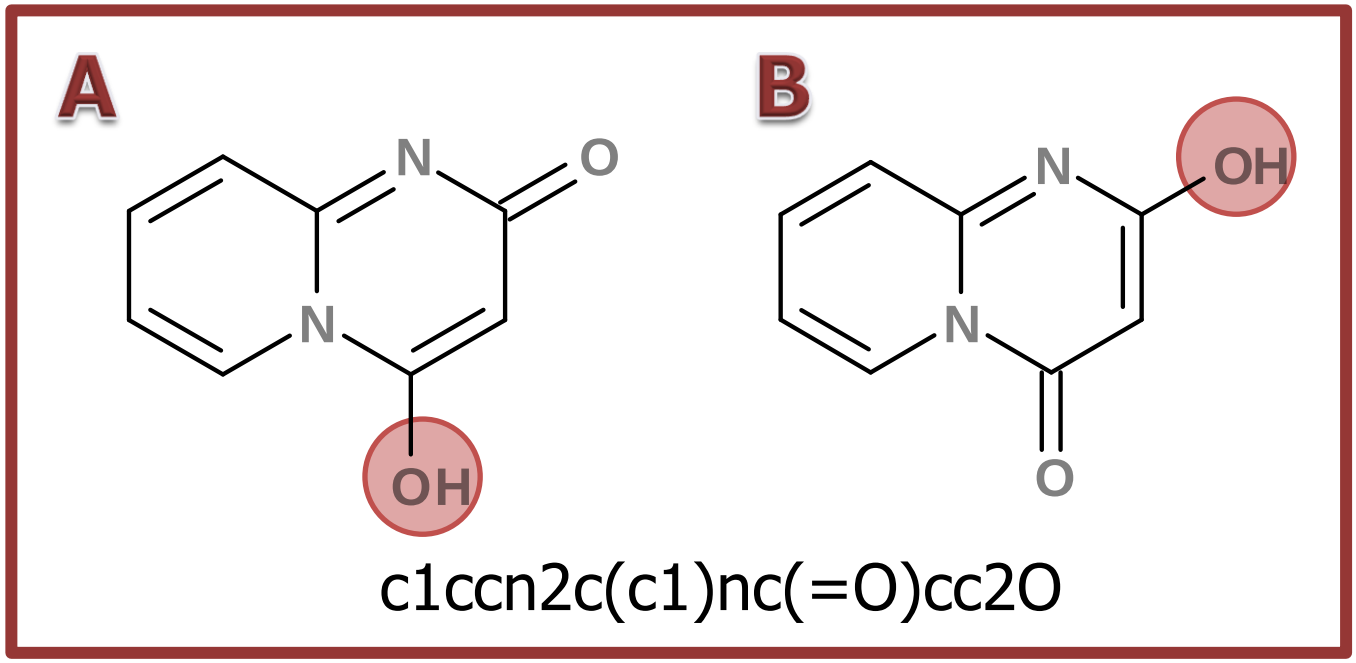


NSC numbers:
98689



NSC numbers:
98690

Differences – Tautomer recognition

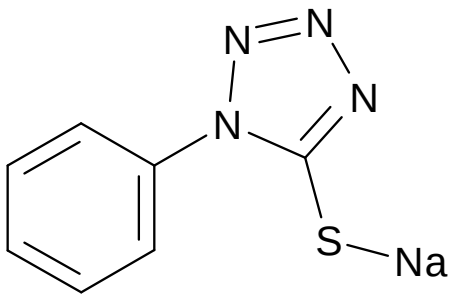


A C8H6N2O2/c11-7-5-8(12)10-4-2-1-3-6(10)9-7/h1-5,**12H**

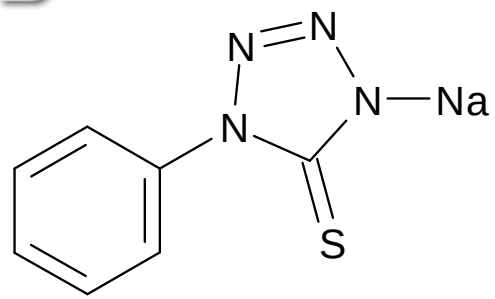
B C8H6N2O2/c11-7-5-8(12)10-4-2-1-3-6(10)9-7/h1-5,**11H**

InChI Removes Metal Bonds

A



B



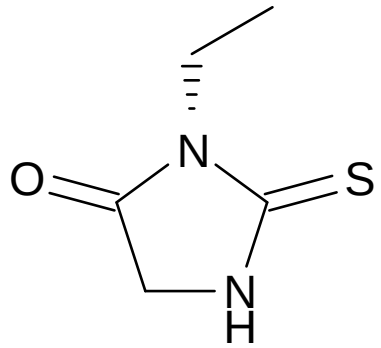
C7H6N4S.Na/c12-7-8-9-10-11(7)6-4-2-1-3-5-6;/h1-5H,(H,8,10,12);/q;+1/p-1

A c1ccc(cc1)[n@]2c(nnn2)S[Na]

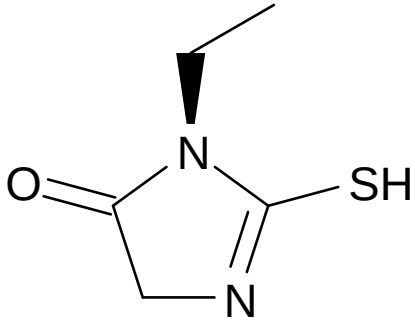
B c1ccc(cc1)[n@]2c(=S)[n@](nn2)[Na]

InChI Automatically Cleans Stereo

A



B



C5H8N2OS/c1-2-7-4(8)3-6-5(7)9/h2-3H2,1H3,(H,6,9)

A CC[N@]1C(=O)CNC1=S

B CC[N@@]1C(=O)CNC1=S

Conclusion

InChI

- Consistent unique identifier (because reference implementation)
- Normalization/tautomer identification part of the language

More consistent!

SMILES

- Algorithm dependent unique identifier
- Tautomer handling can be performed by separate application

More flexible!

	INCHI	SMILES
Human readable/writable	NO	YES
Compact	NO	YES
Expresses similarity (LINGOs)	NO	YES

Acknowledgments

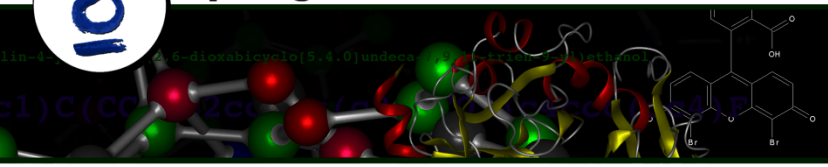
InChI TRUST

NextMove Software

Roger Sayle

OpenEye Scientific Software

Brian Cole
Ben Ellingson
Paul Hawkins



Thank you for your attention!

OpenEye Scientific Software

For more information, please contact us.

business@eyesopen.com

support@eyesopen.com

www.eyesopen.com

505-473-7385