DIGITAL CHEMICAL REPRESENTATIONS

Roger Sayle

NextMove Software, Cambridge, UK

TOO MUCH HISTORY!

- William J. Wiswesser, “107 Years of Line-Formula Notations (1861-1968)”, Journal of Chemical Documentation, Vol. 8, No. 3, pp. 146-150, 1968.
- Bonnie Lawlor, “Chemical Structure Association (CSA) Trust: Advancing Scientific Discovery for Fifty Years”, Chemical Information (CINF) Bulletin, Vol. 67, No. 4, Winter 2015.
- Andrew Dalke, “Weininger’s Realization”, blog 2016/12/02
  - http://www.dalkescientific.com/writings/diary/archive/2016/12/02/Weiningers_realization.html
- Committee on Modern Methods of Handling Chemical Information, National Academy of Science & National Research Council, “Survey of Chemical Notation Systems”, Publication 1150, Washington DC, 1964.

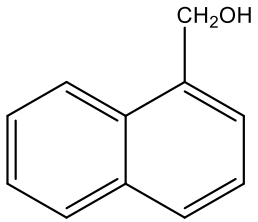
A LITTLE HISTORY (BONNIE LAWLOR)

The emergence of punch-card technologies during the middle of the last century renewed interest in these notations, and in 1949 the International Union of Pure and Applied Chemistry (IUPAC) invited the submission of simple notations that would be suitable for international adoption. They ultimately chose a notation submitted by G. Malcom Dyson, but it was one of the other seven notations that were submitted that caught the attention of those working in the field [1]

[1] It should be noted that the selection of the Dyson notation was criticized, and a petition was signed by about 1,000 chemists, including several who had submitted notations for consideration, stating that the Wiswesser Notation had not been given adequate consideration. The appeal was taken to the American Chemical Society and the National Academy of Sciences - National Research Council who requested that the National Science Foundation do a study, the results of which showed that more testing of both notations should be done before any decision was made. This was not done and the Dyson Notation was selected. A cloud hung over the decision because Dyson was the chair of the IUPAC Commission that called for the submission of notations

1964: DYSON/IUPAC VS. WISWESSER

$B6_2CQ3$



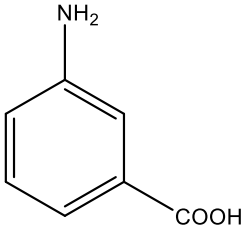
L66J B1Q

C_5C_23Q3

$(CH_3CH_2)_3COH$

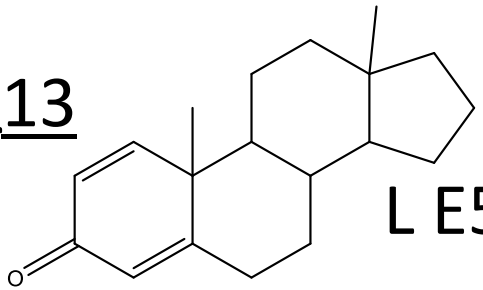
QZ2&2&2

B6CX1N3



ZR CVQ

$A6_3513b7C38EQ13$

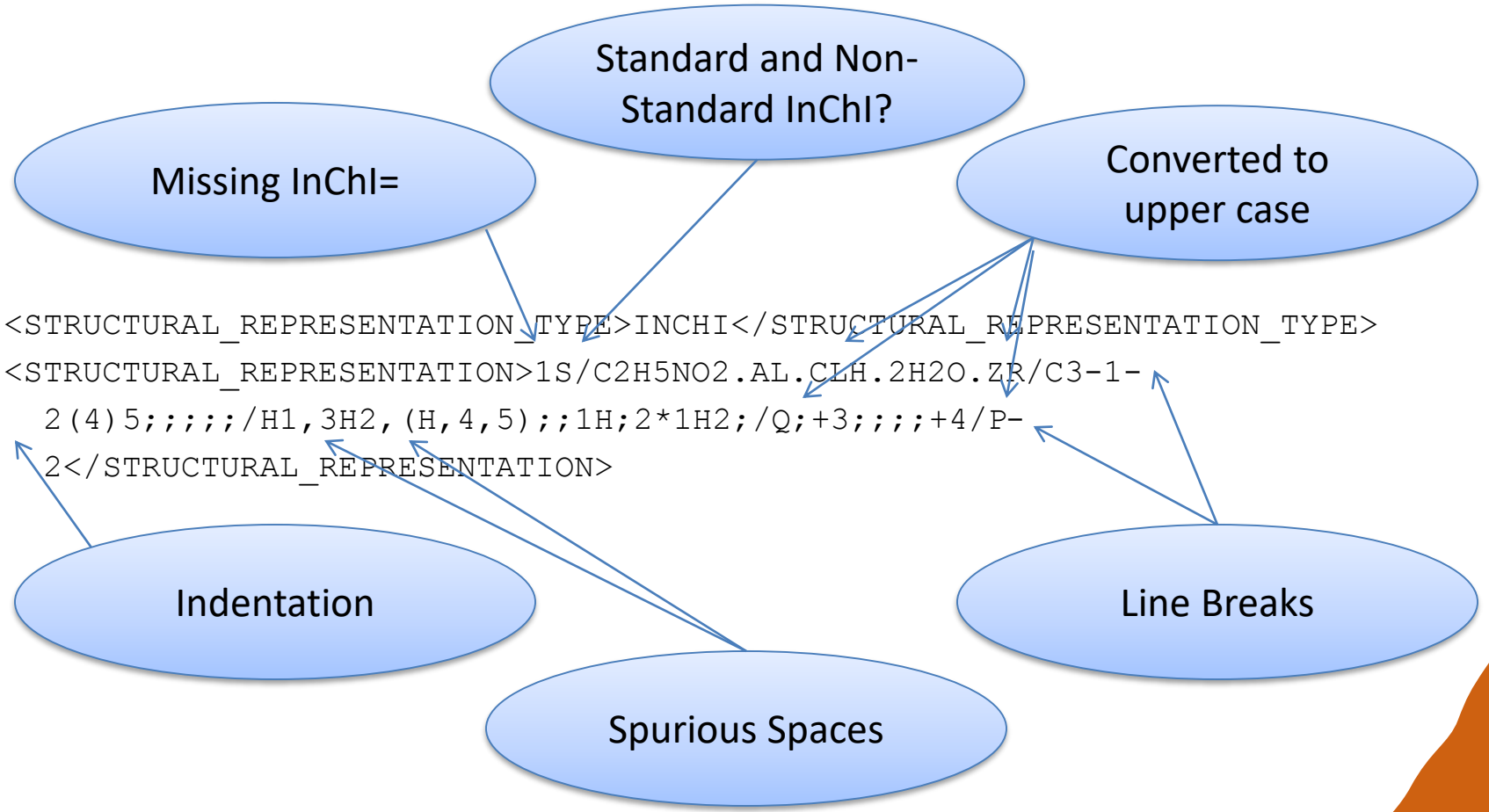


L E5 B666 OV AHTTT&J A E

WHY USE INCHI?

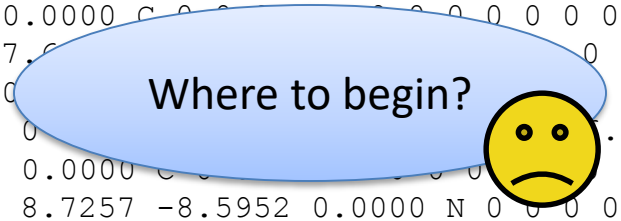
- “InChI is a unique representation/identifier for defined chemical structures. Probably marginally better than previous ones”

ISO11238 §B.2.2 INCHI IN XML EXAMPLE



§B.2.4 V2000 MOL FILE IN XML EXAMPLE

<STRUCTURAL_REPRESENTATION_TYPE>MOL</STRUCTURAL_REPRESENTATION_TYPE>
<STRUCTURAL_REPRESENTATION>30 29 0 0 0 0 0 0 0 0999 V2000 9.9563 -7.3055 0.0000 Y
1 1 0 0 0 0 0 0 0 0 0 0 0 15.0355 -4.8847 0.0000 * 0 0 0 0 0 0 0 0 0 0 0 0 13.3609 -
8.0134 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 13.8867 -9.9869 0.0000 O 0 5 0 0 0 0 0 0 0 0 0
0 6.4178 -6.8678 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 5.8872 -4.8955 0.0000 O 0 5 0 0 0 0
0 0 0 0 0 0 6.7218 -5.7285 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 13.0541 -9.1519 0.0000 C
0 0 0 0 0 0 0 0 0 0 0 0 13.3408 -6.8634 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0 13.8599 -
4.8881 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 13.0301 -5.7260 0.0000 C 0 0 0 0 0 0 0 0 0 0 0
0 5.9099 -9.9441 0.0000 O 0 5 0 0 0 0 0 0 0 0 0 0 6.4492 -7.9743 0.0000 O 0 0 0 0 0 0
0 0 0 0 0 0 6.7482 -9.1149 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 7.8605 -5.4221 0.0000 C 0
0 0 0 0 0 0 0 0 0 0 0 11.8897 -5.4263 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 11.9147 -9.4555
0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 7.8855 -9.4263 0.0000 C 0 0 0 0 0 0 0 0 0 0
7.6897 -8.0305 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 7.6 0 0 0 0
0 0 0 0 0 8.7018 -6.2618 0.0000 N 0 0 0 0 0 0 0 0 C 0 0
0 0 0 0 0 0 0 0 0 0 10.4700 -5.2524 0.0000 C 0 0 0 .2664
0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 12.0761 -6.8427 0.0000 0
12.0891 -8.0218 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 8.7257 -8.5952 0.0000 N 0 0 0 0 0 0
0 0 0 0 0 0 11.0839 -8.6223 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 10.4848 -9.6275 0.0000
C 0 0 0 0 0 0 0 0 0 0 0 0 9.3057 -9.6139 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0 10 2 1 0 0 0 0
8 3 2 0 0 0 0 25 24 1 0 0 0 0 8 4 1 0 0 0 0 27 18 1 0 0 0 0 7 5 2 0 0 0 0 26 28 1 0 0 0 0
7 6 1 0 0 0 0 19 27 1 0 0 0 0 15 7 1 0 0 0 0 20 21 1 0 0 0 0 17 8 1 0 0 0 0 30 27 1 0 0 0
0 11 9 2 0 0 0 0 30 29 1 0 0 0 0 11 10 1 0 0 0 0 20 19 1 0 0 0 0 16 11 1 0 0 0 0 22 21 1
0 0 0 0 14 12 1 0 0 0 0 23 24 1 0 0 0 0 14 13 2 0 0 0 0 18 14 1 0 0 0 0 26 25 1 0 0 0 0
21 15 1 0 0 0 0 29 28 1 0 0 0 0 24 16 1 0 0 0 0 23 22 1 0 0 0 0 28 17 1 0 0 0 0 M CHG 4
1 3 4 -1 6 -1 12 -1 M ISO 1 1 90 M END </STRUCTURAL_REPRESENTATION>

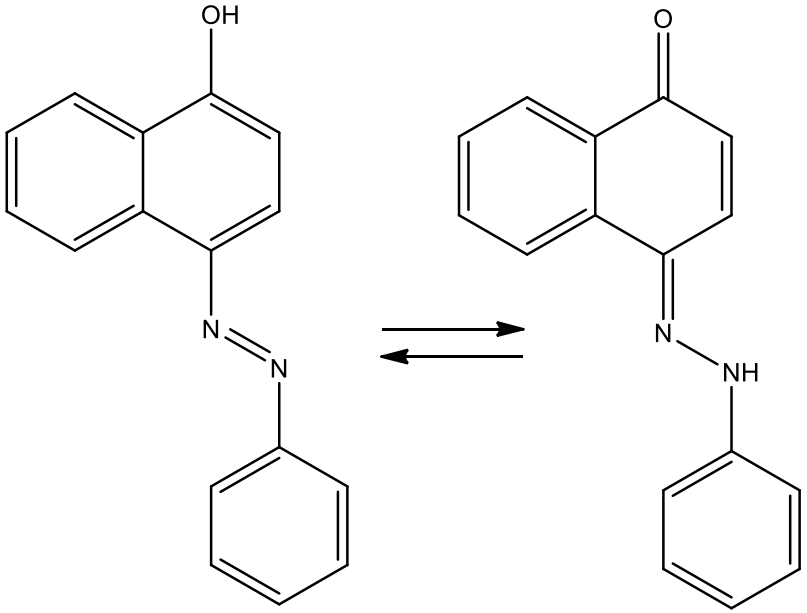


CHEMINFORMATICS' CUTTING EDGE

1. Mesomers and Valence Representations
2. Protonation States
3. Tautomers (prototropic, C-type, ring-chain, ring-ring)
4. Stereochemistry, Configuration and Conformation
5. Biomolecules (Peptides, Nucleic Acids, Sugars and Lipids)
6. Reactions
7. Inorganics, organometallics and intermetallic alloys.
8. Polymers
9. Mixtures (Salts, Solvents, Alloys, Formulations, etc.)
10. Patterns and Transformations.
11. Part, Position, Count and Class Variation (Markush)
12. Physical States and Forms.
13. Radicals, Excited and Metastable nuclear states.

TAUTOMERS

- Tautomers are molecular isomers that easily interconvert by migration of hydrogen atoms[1].

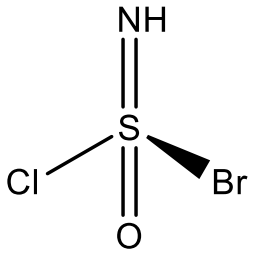


InChI=1S/C16H12N20/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,19H
InChI=1S/C16H12N20/c19-16-11-10-15(13-8-4-5-9-14(13)16)18-17-12-6-2-1-3-7-12/h1-11,17H

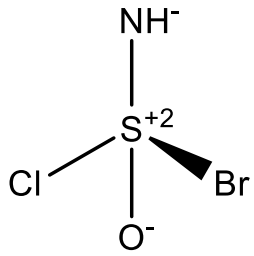
1. R. Sayle, "So you think you understand tautomerism?", JCAMD, 24(6-7):485-496, June 2010.

INCHI ISSUES (PWN2OWN AT INCHI CON)

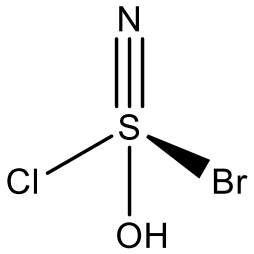
- Some recent limitations with InChI stereochemistry



InChI=1S/BrClHNOS/c1-5(2,3)4/h3H/t5-/m1/s1



InChI=1S/BrClHNOS/c1-5(2,3)4/h3H

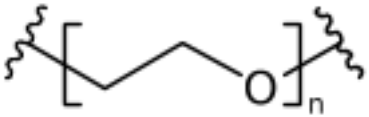


InChI=1S/BrClHNOS/c1-5(2,3)4/h4H

POLYMERS

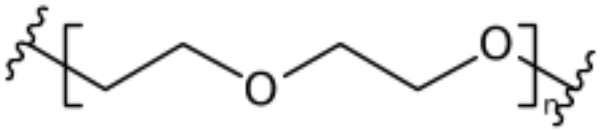
- The InChI Trust has added polymer support to InChI.
- Or has it?

Experimental Beta Option **-Polymers**



PEG_n

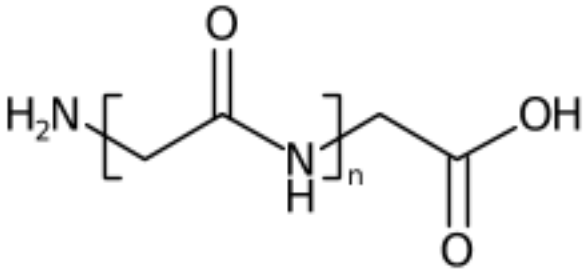
InChI=1B/C2H4O/c1-2-3-1/h1-2H2/z101-1-3(1,2,1,3,2,3)
InChIKey=IAYPIBMASNFSPL-GCGQHNKHBA-N



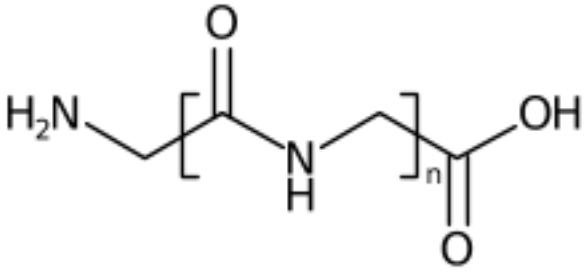
PEG_2n

InChI=1B/C4H8O2/c1-2-6-4-3-5-1/h1-4H2/z101-1-6(1,2,1,5,2,6,3,4,3,5,4,6)
InChIKey=RYHBNJHYFVUHQT-UEXHMUNTBA-N

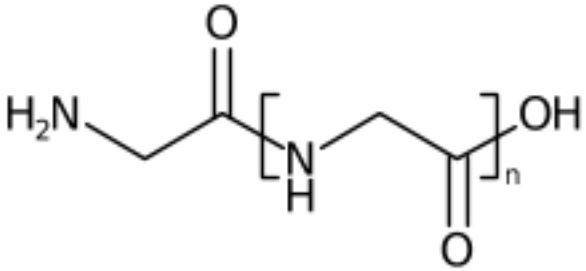
INCHI POLYMERS
CAPPING GROUP FRAME SHIFT



InChI=1B/C4H8N2O3/c5-1-3(7)6-2-4(8)9/h1-2,5H2,(H,6,7)(H,8,9)/z101-1,3,6-7(2-6,5-1)
InChIKey=YMAWOPBAYDPSLA-NPFRMHEKBA-N



InChI=1B/C4H8N2O3/c5-1-3(7)6-2-4(8)9/h1-2,5H2,(H,6,7)(H,8,9)/z101-2-3,6-7(1-3,4-2)
InChIKey=YMAWOPBAYDPSLA-LCMLVUGIBA-N



InChI=1B/C4H8N2O3/c5-1-3(7)6-2-4(8)9/h1-2,5H2,(H,6,7)(H,8,9)/z101-2,4,6,8(3-6,9-4)
InChIKey=YMAWOPBAYDPSLA-YYSJEQPSBA-N

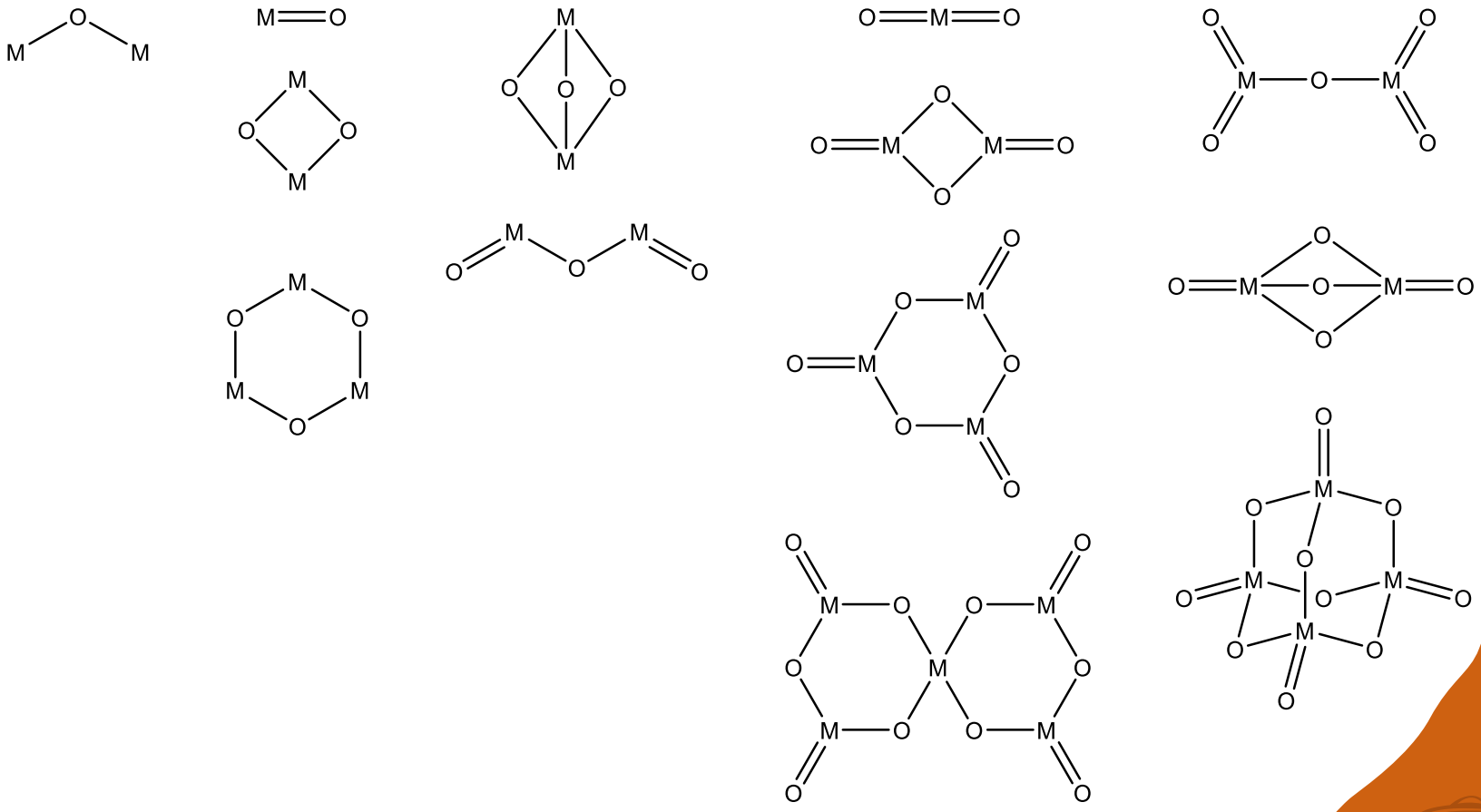
NEUTRAL COMPONENT DUPLICATION

- Duplicated components lead to different InChI
  - Water (O)
    - InChI=1S/H2O/h1H2
    - XLYOFNOQVPJJNP-UHFFFAOYSA-N
  - Wet water (O.O)
    - InChI=1S/2H2O/h2*1H2
    - JEGUKCSWCFPDGT-UHFFFAOYSA-N
  - Dilute water (O.O.O)
    - InChI=1S/3H2O/h3*1H2
    - JLFVIEQMRKMAIT-UHFFFAOYSA-N
- Goodman's Hypothesis: How many InChI-Keys?

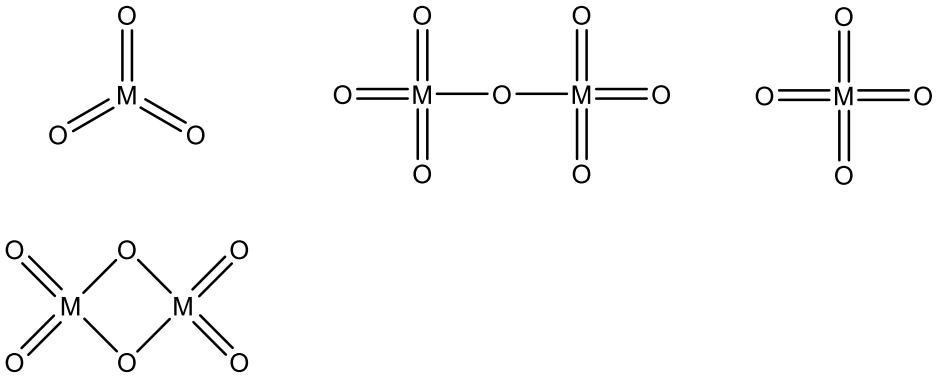
METALLIC OXIDES (AND FRIENDS)

- One problem area for chemical representation are compounds that have no discrete chemical structure but are defined by the ratios of their elements.
- One of John Dalton's case studies (in his 1808 "A New System of Chemical Philosophy") was on tin oxides, SnO and SnO_2, often represented as O=[Sn], O=[Sn]=O

REPRESENTATIONS OF RATIOS



REPRESENTATIONS OF RATIOS



- And these are just the neutral binary metal oxides, there are even more permutations for ions (permanganates, perchlorate) and halides (aluminium chloride) and so on.
- Fortunately, a defining feature of a substance is that it has zero net charge.

DALTON SMILES AND DALTON INCHI

- A molecular representation that correctly captures the ratio of elements, but not necessarily connectivity.
  - O=[Si]=O — Silicon Dioxide (c.f. Wikipedia history)
  - O=P(=O)OP(=O)=O — Phosphorus pentoxide
  - [C] — Diamond, Graphene, Fullerenes
  - [C]=[C] — Graphene, Fullerene
- Extension to mixtures, where each component is listed, but not necessarily the relative amounts of each.
  - Cl.O — Hydrochloric acid
  - Cu1OS(=O)(=O)O1.O — Copper(II) sulfate hydrate
  - [Fe].[Cl] — Iron chloride

EXAMPLE CSD NAMES FROM CCDC

- catena(Tetra-aqua-tetrakis(mu!2$-formato-O,O')-bis(formato-O)-di-barium-copper)
- (mu!2$-2,5-bis((Phenylimino)methyl)benzene-1,4-diyl-C,C',N,N')-bis(eta$5!-pentamethylcyclopentadienyl)-dichloro-di-iridium

IDENTIFIERS VS. REPRESENTATIONS

- Compounds are composed of atoms in defined whole-number ratios, where all atoms of an element are identical.
- It is this statement that allows us to claim that two compounds (or crystals) are identical, and can be captured by a canonical form or universal identifier.
- Without it, substances or mixtures of arbitrary composition are each unique, and one can only talk of similarity and equivalence, not of equality.

METAL ALLOYS

- **AdmiraltyBrass**
  - Cu 69 %
  - Zn 30 %
  - Sn 1 %
- **RollsRoyceTurbineAlloy1**
  - Ni 29.2-37 %mass
  - Co 29.2-37 %mass
  - Cr 10-16 %mass
  - Al 4-6 %mass
  - Zr 0.04-0.07 %mass

SEA WATER COMPOSITION

- **SeaWater**
  - Water 1 liter
  - Salts 41.953 g
    - NaCl 58.490 %
    - MgCl2.6H20 26.460 %
    - Na2SO4 9.750 %
    - CaCl2 2.765 %
    - KCl 1.645 %
    - NaHCO3 0.477 %
    - KBr 0.238 %
    - H3BO3 0.071 %
    - SrCl2.6H20 0.095 %
    - NaF 0.007 %

ATMOSPHERIC COMPOSITION

- **Air**
  - Nitrogen 78.084 %v
  - Oxygen 20.964 %v
  - Argon 0.9340 %v
  - Carbon dioxide 0.04 %v
  - Neon 0.001818 %v
  - Helium 0.000524 %v
  - Methane 0.00018 %v
  - Krypton 0.000114 %v
  - Hydrogen 0.000055 %v
- **Martian Atmosphere**
  - Carbon dioxide 95.97 %v
  - Argon 1.93 %v
  - Nitrogen 1.89 %v
  - Oxygen 0.146 %v
  - Carbon monoxide 0.0557 %v

IBM'S BAD SIMILARITY

- Vidal05 and Grant06 described a fast chemical similarity measure based upon SMILES strings.
- Typically, similar molecules have similar SMILES.
- In US20080002810, IBM claim a patent for "System and Method for Identifying Similar Molecules" that also uses n-gram similarity but on InChI strings.
- Unfortunately, similar molecules don't have similar (similar runs of characters in) InChI strings!
- Benchmarking shows the IBM method to be one of the worst chemical similarity measures to date.

CONCLUSIONS

- It's fantastic to be here and be part of the on-going process (which has a long legacy).
- Even when flawed, chemical line notations help frame our understanding of chemistry.
- Extending discrete compounds to continuous compounds is the current state-of-the-art.
- But for mixtures and beyond, the future is in both "similarity" and canonical forms.

ACKNOWLEDGEMENTS

- The team at NextMove Software
  - John Mayfield
  - Noel O'Boyle
- And the Cheminformatics Community (including)
  - Daniel Lowe
  - Leah McEwen
  - Philip Skinner
  - Evan Bolton
  - Greg Landrum
  - Ian Bruno
  - Jonathan Goodman
  - Andrew Dalke
- And many thanks for your time!

DEFINITIONS

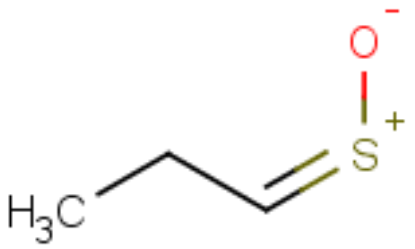
- A substance has constant chemical composition.
- A mixture is two or more different substances.
- A chemical compounds has two or more atoms, with a fixed proportion/ratio of it's elements.
- Non-stoichiometric compounds = mixtures.
- Intermetallic alloys = compounds.

INCHI VS. IUPAC

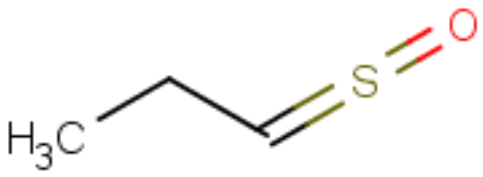
- **Physically impossible charge**
  - [C+7] InChI=1S/C/q+7
- **Physically impossible isotope**
  - [5C] InChI=1S/C/i1-7
- **Mononuclidic elements**
  - [23Na][19F] InChI=1S/FH.Na/h1H;/q;+1/p-1/i2*1+0
  - [Na]F InChI=1S/FH.Na/h1H;/q;+1/p-1
- **Formula Unit**
  - [Na]Cl InChI=1S/ClH.Na/h1H;/q;+1/p-1
  - [Na]Cl.[Na]Cl InChI=1S/2ClH.2Na/h2*1H;;/q;;2*+1/p-2
  - [Na]1Cl[Na]Cl1 InChI=1S/2ClH.2Na/h2*1H;;

INCHI ISSUES

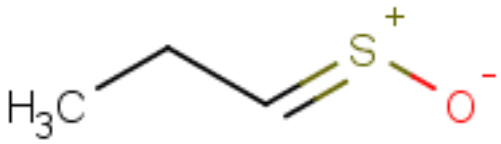
- Some important molecules are indistinguishable



CC\C=[S+]/[O-]



CCC=S=O

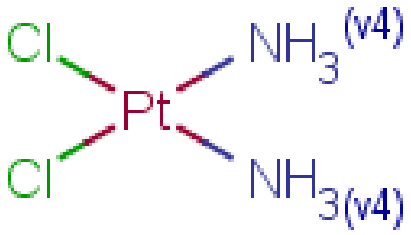


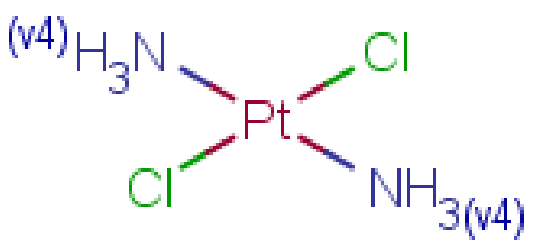
CC\C=[S+]\[O-]

InChI=1S/C3H6OS/c1-2-3-5-4/h3H,2H2,1H3

INCHI ISSUES

- Some important molecules are indistinguishable





[NH3][Pt]([NH3])(Cl)Cl [NH3][Pt]([NH3])(Cl)Cl

[NH3+][Pt-2]([NH3+])(Cl)Cl [NH3+][Pt-2]([NH3+])(Cl)Cl

[NH3][Pt@SP1]([NH3])(Cl)Cl [NH3][Pt@SP2]([NH3])(Cl)Cl

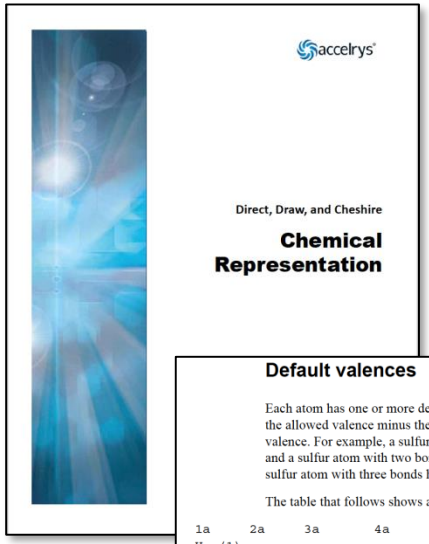
InChI=1S/2ClH.2H3N.Pt/h2*1H;2*1H3;/q;;;;+2/p-2

MOL FILE STANDARDIZATION

- RDKit's rdkit/Docs/Book/data/actives_5ht3.sdf
  - Contains 180 connection tables
    - RDKit outputs 180 molecules, with no warnings.
    - OEChem outputs 38 molecules, with 142 warnings.
    - ChemAxon outputs 10 molecules, with 1 warnings.
    - OpenBabel outputs 180 molecules, with 142 warnings.
    - InChI outputs 180 molecules, with 23 warnings.
  - Counts line of an offending records " 21 24999 V2000"
  - Hence, aaa is " 21", bbb is " 24", lll is "999", ccc (chiral flag) is "000" (i.e. not chiral).

MDL MOLFILE-AGEDDON

- Biovia 2017 changes the interpretation of MDL files.
- This affects over 213097 CIDs in PubChem!



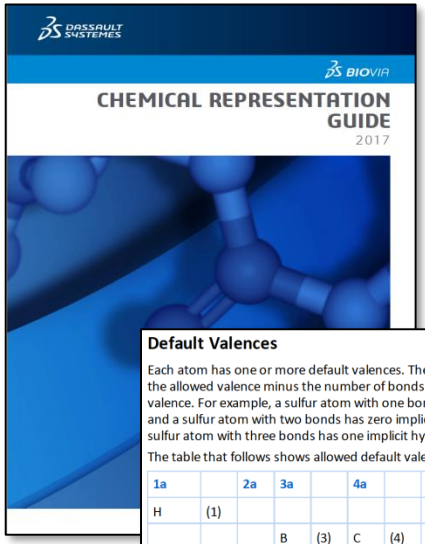
Default valences

Each atom has one or more default valences. The number of implicit hydrogens at an atom is equal to the allowed valence minus the number of bonds to non-hydrogen atoms, up to the next allowed valence. For example, a sulfur atom with one bond to a non-hydrogen atom has one implicit hydrogen, and a sulfur atom with two bonds has zero implicit hydrogens, because the next highest valence is 2. A sulfur atom with three bonds has one implicit hydrogen, because the next highest valence is 4.

The table that follows shows allowed default valences for neutral main group elements:

1a	2a	3a	4a	5a	6a	7a	8
H (1)							He (0)
Li (1)	Be (2)	B (3)	C (4)	N (3,5)	O (2)	F (1)	Ne (0)
Na (1)	Mg (2)	Al (3)	Si (4)	P (3,5)	S (2,4,6)	Cl (1,3,5,7)	Ar (0)
K (1)	Ca (2)	Ga (3)	Ge (4)	As (3,5)	Se (2,4,6)	Br (1,3,5,7)	Kr (0)
Rb (1)	Sr (2)	In (3)	Sn (2,4)	Sb (3,5)	Te (2,4,6)	I (1,3,5,7)	Xe (0)
Cs (1)	Ba (2)	Tl (1,3)	Pb (2,4)	Bi (3,5)	Po (2,4,6)	At (1,3,5,7)	Rn (0)
Fr (1)	Ra (2)						

For transition metals, lanthanides, and actinides, any valence is allowed. Consequently, these atoms do not display implicit hydrogens unless you specify an explicit valence.



Default Valences

Each atom has one or more default valences. The number of implicit hydrogens at an atom is equal to the allowed valence minus the number of bonds to non-hydrogen atoms, up to the next allowed valence. For example, a sulfur atom with one bond to a non-hydrogen atom has one implicit hydrogen, and a sulfur atom with two bonds has zero implicit hydrogens, because the next highest valence is 2. A sulfur atom with three bonds has one implicit hydrogen, because the next highest valence is 4.

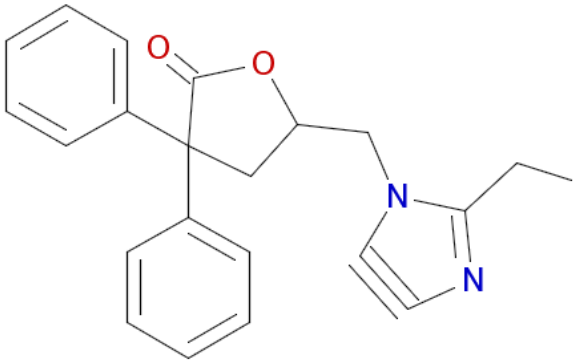
The table that follows shows allowed default valences for neutral main group elements:

1a		2a	3a		4a		5a		6a		7a		8	
H	(1)												He	(0)
			B	(3)	C	(4)	N	(3)	O	(2)	F	(1)	Ne	(0)
					Si	(4)	P	(3,5)	S	(2,4,6)	Cl	(1,3,5,7)	Ar	(0)
							As	(3,5)	Se	(2,4,6)	Br	(1)	Kr	(0)
									Te	(2,4,6)	I	(1,3,5,7)	Xe	(0)
											At	(1,3,5,7)	Rn	(0)

Implicit hydrogens are never added to metal atoms or ions (implied valence is zero), with the exception of Al(-1) which has a default valence of 4.

SMILES STANDARDIZATION

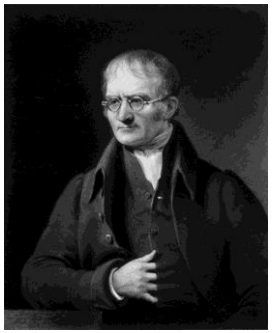
- CHEMBL 423544
  - ChEMBL uses Biovia Pipeline Pilot for its SMILES



- CCc1n[c]#[c]n1CC2CC(C(=O)O2)(c3ccccc3)c4ccccc4

- CCc1nc#cn1CC2CC(C(=O)O2)(c3ccccc3)c4ccccc4
- CCC1=NC#CN1CC2CC(C(=O)O2)(c3ccccc3)c4ccccc4

JOHN DALTON'S LEGACY



- In 1808, John Dalton published "A New System of Chemical Philosophy", in which he described his atomic theory, based upon the law of multiple proportion that revolutionized/defined molecular chemistry.
- Compounds are composed of atoms in defined whole-number ratios, where all atoms of an element are identical.
- Interestingly, 209 years later, boundary cases of this rule, define the frontiers of cheminformatics.