AGAINST SMILESPLUS: WHY DAVID WEININGER WAS OPPOSED TO (IUPAC) STANDARDIZATION OF SMILES

Roger Sayle

NextMove Software, Cambridge, UK

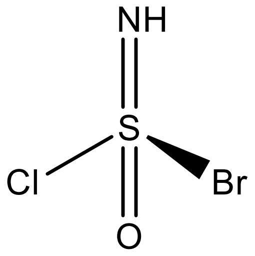
IN A NUTSHELL

- David Weininger, the creator of SMILES and founder of Daylight Chemical Information Systems (DCIS) was a product of 1960’s American culture.
- He believed that people should have the freedom to do what they want, rather than preventing or being prevented by other people.
- In 1999, when Dave’s friend Steve Heller send out an announcement to an meeting on standardizing canonicalization, his position was to just let IUPAC do whatever they want.

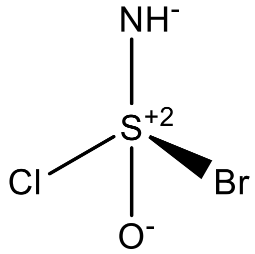
SCIENTIFIC HUBRIS

- Dave once explained that one of his scientific heros was Emil Fischer, 1902 Nobel prize in chemistry.
- Alas it would be almost impossible to have a conversation as so much has changed over the last 100 years.
- Chemists 100 years from now will look back on us like we were cavemen.
- The valence-bond model of chemistry has been obsoleted by quantum mechanics since the 1930s, but persists by it’s computationally tractable.

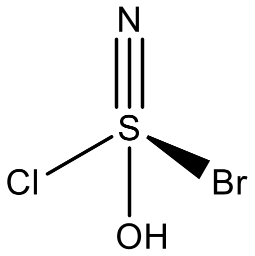
RECENT STEREOCHEMISTRY



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

I WAS ONCE LIKE YOU ARE NOW

- In May 1998, whilst at Daylight I composed a proposal document “StableSmiles” in Word 6.0 on MacOS 10, that analysed the portability issues at the time (Daylight v4.41) and proposed two changes to improve the stability of SMILES between releases.
  - Kekule SMILES: C1=CC=CC=C1
  - Heavy Elements: [#106]

ORIGINAL SMILES "TRIP TEST"

SMILES	CANSMI	CORINA	CORINA	CONCORD	COBRA
	4.41	1.6	WWW	3.2.1	3.21A
C1.C1	Y	Y	Y	N	N
C%00CC%00	Y	Y	Y	N	N
C(C.C)C	Y	Y	Y	N	N
C(C)1CC1	Y	N	N	N	Y
C(.C)	Y	Y	Y	N	N
C()	Y	Y	N	Y	Y
(CO)=O	N	N	N	N	N
(C)	N	N	N	N	N
.C	N	N	N	Y	Y
C..C	N	Y	N	Y	Y
C.	N	Y	Y	Y	Y
C=(O)C	N	Y	N	N	Y
C((C))	N	Y	N	Y	N
C.(C)	N	Y	N	Y	N
C1CC(=1)	N	Y	N	N	N
C1CC(1)	N	N	N	N	N
C(C.)	N	Y	N	N	N
C==C	N	Y	N	N	N

HISTORY REPEATING ITSELF

- Those who don’t study history are doomed to repeat it - Winston Churchill.
  - IUPAC/CAS naming
  - Wiswesser line notation
  - IUPAC/Dyson line notation
  - Standard Molecular Data (SMD) format
  - SMILES variants (e.g. CONCORD SMILES, Chortles)
  - V2000/V3000 Mol file
  - InChI and InChIKey
  - ISO11238

THE ROAD TO HELL…

- … is paved with good intentions.
- I believe that all standardization efforts are driven by a will to do something good for the community.
- Why so many of them fail (and sometimes harm) is driven by sociology, we're only human after all.
- Ultimately, the best form of standards are "de facto" standards, where (Darwinian) selection and utility drives adoption rather than edict.

FINAL WORDS

- Conflicting goals: Standardization vs. Innovation.
- Conflicting goals: Return on Investment.
- Serenity Prayer (c.f. Shepard’s Prayer)

  Grant me the serenity to accept the things I cannot change,
  Courage to change the things I can,
  And wisdom to know the difference.

DAVE'S NOT HERE

TOO MUCH HISTORY!

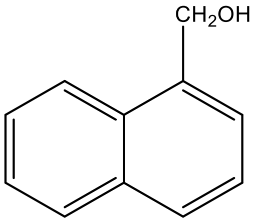
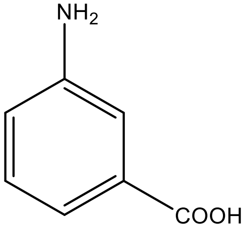
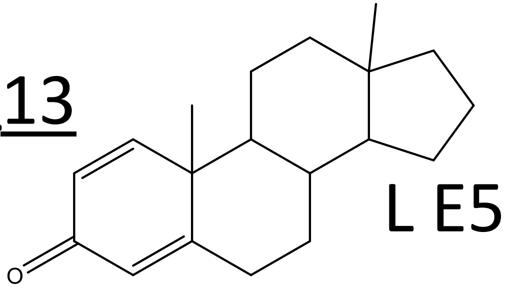
- William J. Wiswesser, “107 Years of Line-Formula Notations (1861-1968)”, Journal of Chemical Documentation, Vol. 8, No. 3, pp. 146-150, 1968.
- Bonnie Lawlor, “Chemical Structure Association (CSA) Trust: Advancing Scientific Discovery for Fifty Years”, Chemical Information (CINF) Bulletin, Vol. 67, 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)

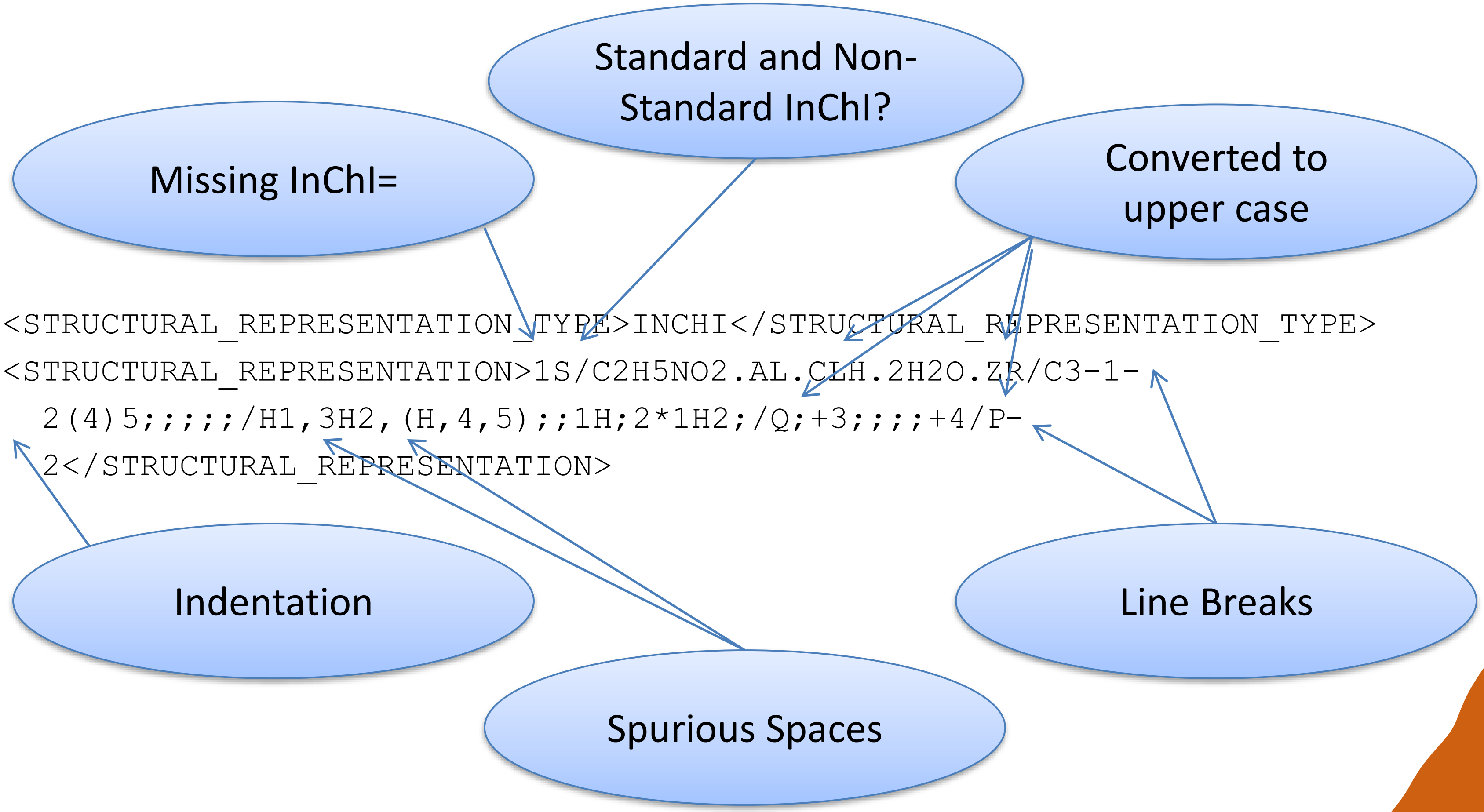
The emergence of punch-card technologies during the middle of the last century renewed interest in these notations, and in 1949 the International Union of Pure and Applied Chemistry (IUPAC) invited the submission of simple notations that would be suitable for international adoption. They ultimately chose 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$_2$CQ3		L66J B1Q
C$_5$C$_2$3Q3	$(CH_3CH_2)_3COH$	QZ2&2&2
B6CX1N3		ZR CVQ
A6$_3$513b7C38EQ$\underline{13}$		L E5 B666 OV AHTTT&J A E

ISO11238 §B.2.2 INCHI IN XML EXAMPLE



Standard and Non-Standard InChI?

Missing InChI=

Converted to upper case

```
<STRUCTURAL_REPRESENTATION_TYPE>INCHI</STRUCTURAL_REPRESENTATION_TYPE>
<STRUCTURAL_REPRESENTATION>1S/C2H5NO2.AL.CLH.2H2O.ZR/C3-1-
  2(4)5;;;;;/H1,3H2,(H,4,5);;1H;2*1H2;/Q;+3;;;;+4/P-
  2</STRUCTURAL_REPRESENTATION>
```

Indentation

Line Breaks

Spurious Spaces

§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 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 [illegible] 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.[illegible] 0 0 0 0
0 0 0 0 0 8.7018 -6.2618 0.0000 N 0 0 0 0 0 0 0 [illegible] C 0 0
0 0 0 0 0 0 0 0 0 0 10.4700 -5.2524 0.0000 C 0 0 0 [illegible].2664
0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0 12.0761 -6.8427 0.0000 [illegible] 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>
```

Where to begin?

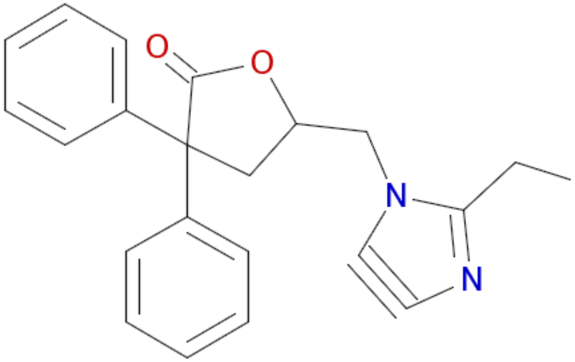
MOL FILE STANDARDIZATION

- 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).

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