Reading and Writing Molecular File Formats for Data Exchange of Small Molecules, Biopolymers and Reactions

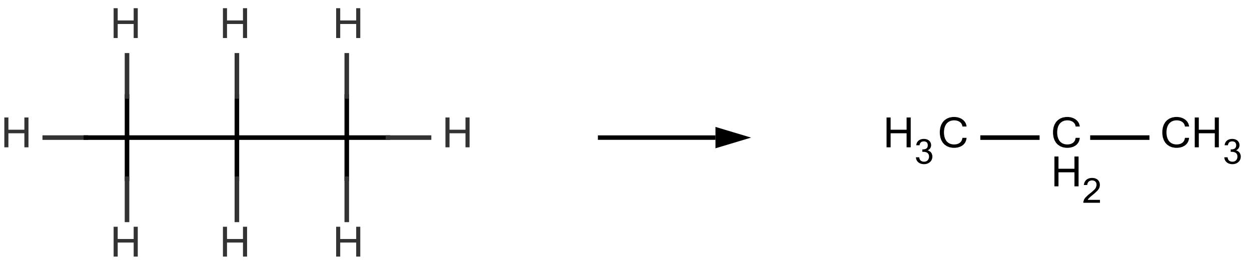
Roger Sayle

NextMove Software, Cambridge, UK

246th ACS National Meeting, Indianapolis, IN, Wednesday 11th September 2013

Connection Table Compression

- One distinguishing feature between computational chemistry (COMP) and cheminformatics (CINF) is the representation of hydrogens.
- Typically in organic chemistry, approximately half of the bonds in a "full" connection table are the bonds to terminal hydrogen atoms.



HYDROGENS IN FILE FORMATS

- Implicit vs. Explicit is also preserved with file I/O.
- SMILES: [H]C([H])([H])[H] vs C (or [CH4]).
- MDL connection tables:

```
[CH4]
     RDKit

  1  0  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
M  END
$$$$
```

```

 OpenBabel10021215582D

  5  4  0  0  0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
  1  5  1  0  0  0  0
M  END
$$$$
```

VALENCE MODELS

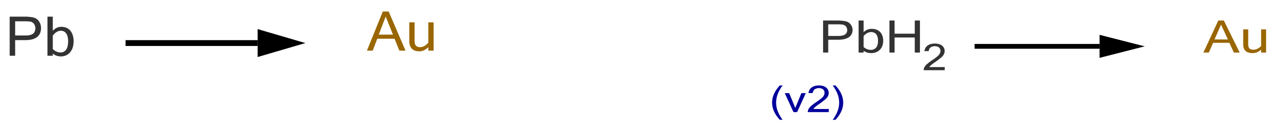
- To make life interesting, developers of molecule file formats that support implicit hydrogens (Daylight and MDL) allow omission of the number of implicit hydrogens on an atom, to save space, instead relying on the deriving the number of implicit hydrogens for the atom’s default valence in a given environment.
- Both formats can fully specify implicit hydrogen count/valence, but alas not all readers/writers support these conventions.

EXAMPLE VALENCE MODEL

- As an example, Daylight’s valence model for SMILES states that all aromatic nitrogens have no implicit hydrogens by default.
- If a hydrogen is present, it should be written as “[nH]”, and if it’s charged as “[nH+]”.
- The number of hydrogens never needs to be guessed, and the molecular formula is precisely specified.

THE PERIL OF VALENCE MODELS

A popular desktop sketching program is responsible for the poor reaction yields of traditional alchemists.



Saving the above reaction as an MDL RXN file and reloading it changes the hydrogen count due to bugs in the software. In practice, such errors lose the distinction between sodium metal and sodium hydride (etc.) in pharmaceutical ELNs.

The MDL/Accelrys Valence Model

- The correct valence is specified by MDL/ISIS (and as documented in Accelrys' "Chemical Representation")

- Neutral carbon should be four valent.
  - Everyone agrees on this (hopefully).
- +1 Nitrogen cation should be four valent.
- +1 Carbon cation should be three valent.
- -4 charge Boron should a have valence 1.
  - OEChem says 0, OpenBabel says 3, RDKit says 7...

THE "MDLBENCH.SDF" BENCHMARK

- To test the fidelity of MDL valence implementations in available SD file readers, we evaluate the SMILES generated from a standard reference test file.
- This SD file contains 10,208 connection tables:
  - 114 different elements plus "D" and "T"
  - 11 charge states (from -4 to +6 inclusive)
  - 8 environments (minimum valences from 0 to 7)
  - 116*11*8 = 10208

Recall/Precision and Subsets

- Rather than assess programs on the 10K theoretically possible atomic valences, a more realistic benchmark is to consider the environments encountered in practice (such as the 199 found in Accrelyrs' MDDR database, version 2011.2).
- An alternative restriction is to consider only the 10 "organic" elements (B,C,N,O,P,S,F,Cl,Br,I).

24 MDL MOL FILE READERS LATER...

						Complete Set (10208)				Organic Subset (880)				MDDR Observed (199)			
Toolkit	fails	invalid	dots	stars	bizarre	wrong	good	percent	fraction	wrong	good	percent	fraction	wrong	good	percent	fraction
OEChem v1.9	0	0	0	264	0	22	9922	97.20%	99.78%	5	875	99.43%	99.43%	0	199	100.00%	100.00%
MDL Direct v8	968	0	0	0	0	22	9218	90.30%	99.76%	5	875	99.43%	99.43%	0	199	100.00%	100.00%
Daylight v4.8.2	0	0	0	880	0	552	8776	85.97%	94.08%	196	684	77.73%	77.73%	1	198	99.50%	99.50%
Indigo 1.1.4	1917	0	0	880	0	184	7227	70.80%	97.52%	79	383	43.52%	82.90%	1	197	98.99%	99.49%
Pipeline Pilot v9.0	0	0	0	968	0	243	8997	88.14%	97.37%	27	853	96.93%	96.93%	3	196	98.49%	98.49%
Dotmatics PinPoint	0	0	0	968	0	650	8590	84.15%	92.97%	46	834	94.77%	94.77%	3	196	98.49%	98.49%
ChemDraw v12	0	704	0	0	0	548	8956	87.74%	94.23%	53	827	93.98%	93.98%	3	196	98.49%	98.49%
SDFilter	1276	0	0	847	0	462	7623	74.68%	94.29%	125	645	73.30%	83.77%	2	195	97.99%	98.98%
InfoChem [internal]	0	0	0	0	0	829	9379	91.88%	91.88%	285	595	67.61%	67.61%	4	195	97.99%	97.99%
Pipeline Pilot v8.5	0	0	0	968	0	2252	6988	68.46%	75.63%	371	509	57.84%	57.84%	5	194	97.49%	97.49%
ChemAxon 5.10	0	0	0	440	0	684	9084	88.99%	93.00%	248	632	71.82%	71.82%	6	193	96.98%	96.98%
CACTVS 3.407	33	0	6	352	2	576	9239	90.51%	94.13%	196	684	77.73%	77.73%	7	192	96.48%	96.48%
CACTVS 3.383	11	792	6	0	2	576	8821	86.41%	93.87%	196	684	77.73%	77.73%	7	192	96.48%	96.48%
Balloon v1.40	0	0	0	0	0	726	9482	92.89%	92.89%	252	628	71.36%	71.36%	8	191	95.98%	95.98%
AvalonTools v1.1b	0	0	0	0	0	732	9476	92.83%	92.83%	270	610	69.32%	69.32%	8	191	95.98%	95.98%
Schrodinger Canvas	616	0	0	0	44	872	8676	84.99%	90.87%	259	621	70.57%	70.57%	11	188	94.47%	94.47%
OpenBabel 2.3.90	0	0	0	176	66	659	9307	91.17%	93.39%	242	638	72.50%	72.50%	12	187	93.97%	93.97%
MOE 2011.10	0	0	0	0	4221	936	5051	49.48%	84.37%	94	340	38.64%	78.34%	5	186	93.47%	97.38%
CDK 1.4.13 [noh]	264	0	0	0	0	485	9459	92.66%	95.12%	165	715	81.25%	81.25%	14	185	92.96%	92.96%
RDKit +NextMove	2339	0	66	0	0	1143	6660	65.24%	85.35%	285	222	25.23%	43.79%	9	182	91.46%	95.29%
InChI=1S/...	880	0	6293	0	0	141	2894	28.35%	95.35%	85	795	90.34%	90.34%	11	113	56.78%	91.13%
CDK 1.4.13 [imph]	9876	0	0	0	0	80	252	2.47%	75.90%	30	63	7.16%	67.74%	12	98	49.25%	89.09%
RDKit 2012_09	4029	0	66	0	0	4722	1391	13.63%	22.75%	285	222	25.23%	43.79%	100	88	44.22%	46.81%
InfoChem +RDKit	3395	0	66	634	0	4735	1378	13.50%	22.54%	285	222	25.23%	43.79%	100	88	44.22%	46.81%

MDLBENCH.SDF RESULTS @ RDKIT UGM

ToolKit	Failures	Incorrect	Correct	Recall	Precision
OEChem v1.9	264	22	9922	97.20%	99.78%
CDK v1.4.13	264	486	9458	95.11%	95.11%
OpenBabel v2.3.90	176	668	9364	91.73%	93.34%
CACTVS v3.407	352	511	9339	91.49%	94.81%
MDL Direct v8.0	968	22	9218	90.30%	99.76%
ChemAxon v5.10	440	685	9083	88.98%	92.99%
Pipeline Pilot v9.0	968	243	8997	88.14%	97.37%
ChemDraw v12.0	704	548	8956	87.74%	94.23%
GGA Indigo v1.1.4	2797	184	7227	70.80%	97.52%
MOE v2011.10	4221	936	5051	49.48%	84.37%
RDKit v2012_09	4095	4723	1390	13.62%	22.74%

MDLBENCH.SDF SUMMARY

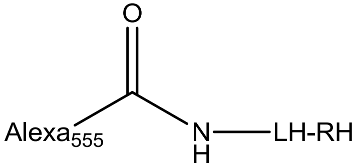
- Although many toolkits can read MDL SD files, they don’t all perfectly agree on the semantics.
- Different readers, sometimes written by the same company, can interpret the exact same MOL file as different molecules.
- Fortunately, most compounds encountered in the pharmaceutical industry fall into the widely understood “well-behaved” subset.

MDLBENCH.SDF POSTSCRIPT

- Following this experiment, feedback was passed back to several vendors/developers and the benchmark including the expected results were published online.
- Patches were submitted to RDKit and OpenBabel to achieve 100% conformance on this benchmark.
- Many vendors, including ChemAxon, Dotmatics, Optibrium and Xemistry, and open source projects, including MayaChem and Balloon, have announced improvements based on this work.
- As a result the community converges on a standard.

Part 2: Further Support

LABEL PRESERVATION (MDL ALIAS)



Can be converted to a molfile as

```
NextMove07171309292D

  5  4  0     0  0  0  0  0  0999 V2000
    0.0000    0.0000    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
    1.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000   -0.8660    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.5000    0.8660    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.5000    0.8660    0.0000 *   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  2  0  0  0  0
  2  4  1  0  0  0  0
  4  5  1  0  0  0  0
A    1
Alexa555
A    5
LH-RH
M  END
```

V3000 MOL FILES

- Accelrys' introduction of the version 3000 molfile introduces a chicken and egg problem for the cheminformatics community.
- Adoption of the format is hampered by the limited support for v3000 molfiles amongst third-party tools, but equally the v3000 files in circulation create problems for developers.

REACTIONS AS 1ST CLASS CITIZENS

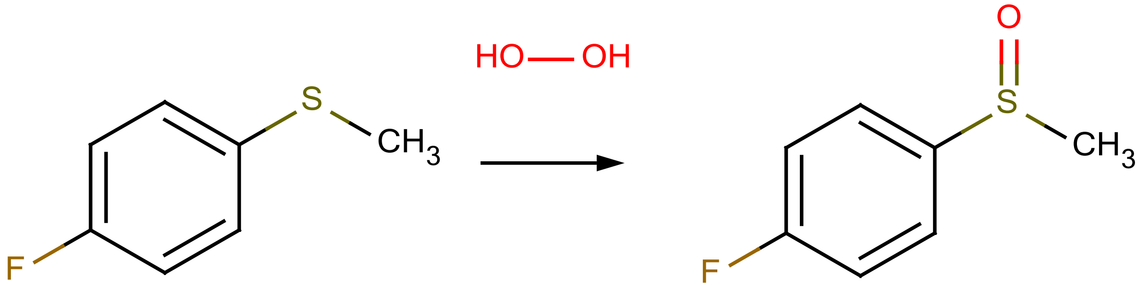
- Although all toolkits support small molecule exchange, support for reactions is generally poorer.
- This situation is compounded by treating reactions as different objects/types than regular molecules, and sometimes introducing format distinctions based upon the content of a file.
- Many formats, including MDL mol file (via CPSS), ISIS Sketch, SMILES, ChemDraw CDX and CDXML merrily encode both molecules and reactions.

SUPPORT FOR EMPTY REACTIONS

- An interesting class of informatics problems concerns molecules with no atoms, and reactions with no reactants and/or products.
- In SDF and RDF file formats, these records can have associated data, but it is not uncommon for this to be skipped/lost by many tools.
- A SMILES example might be ">> title".
- It is not uncommon for chemists to draw nothing after an arrow to indicate a reaction failure, or nothing before one to indicate compound purchase.

AGENTS, SOLVENTS AND CATALYSTS

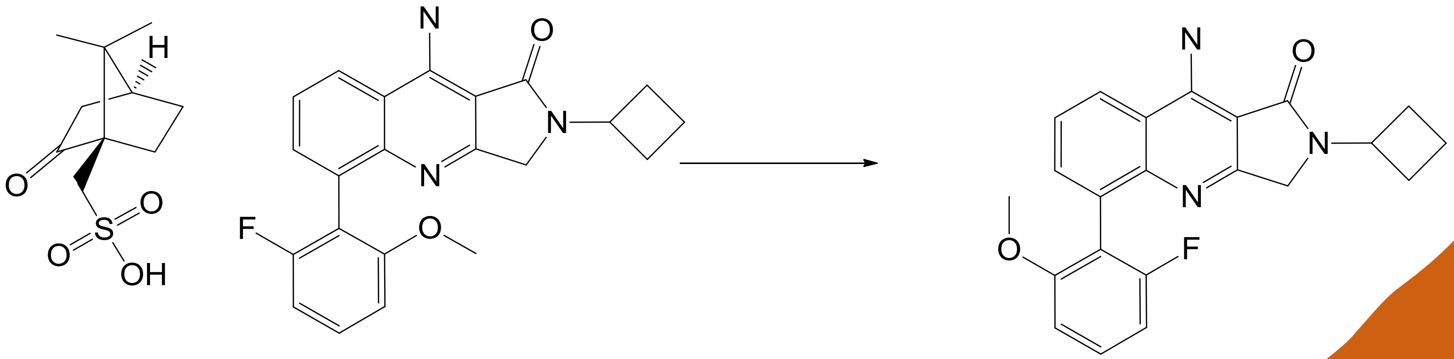
An extremely useful extension to the MDL RXN file format introduced by ChemAxon is support for agents. After the reactant count and product count fields, allowing an agent count avoid information loss, capturing items drawn above/below a reaction arrow.



```
CSc1ccccc1(F)>OO>CS(=O)c1ccc(cc1)F
```

Sketch File Support

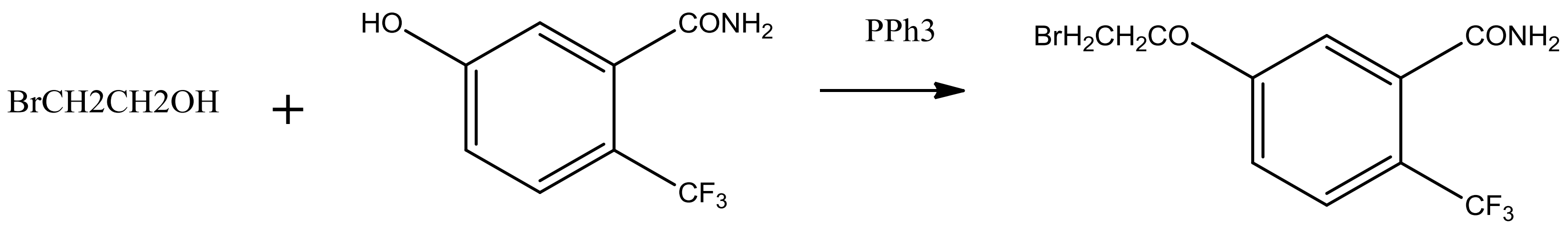
- The community should tackle better handling of sketch file formats, such as ISIS Sketch, CDX, CDXML, Marvin and ACD/Labs sk2.
- The challenge is that these formats are intended to be consumed by human beings not computers, but they do capture details otherwise lost in translation.



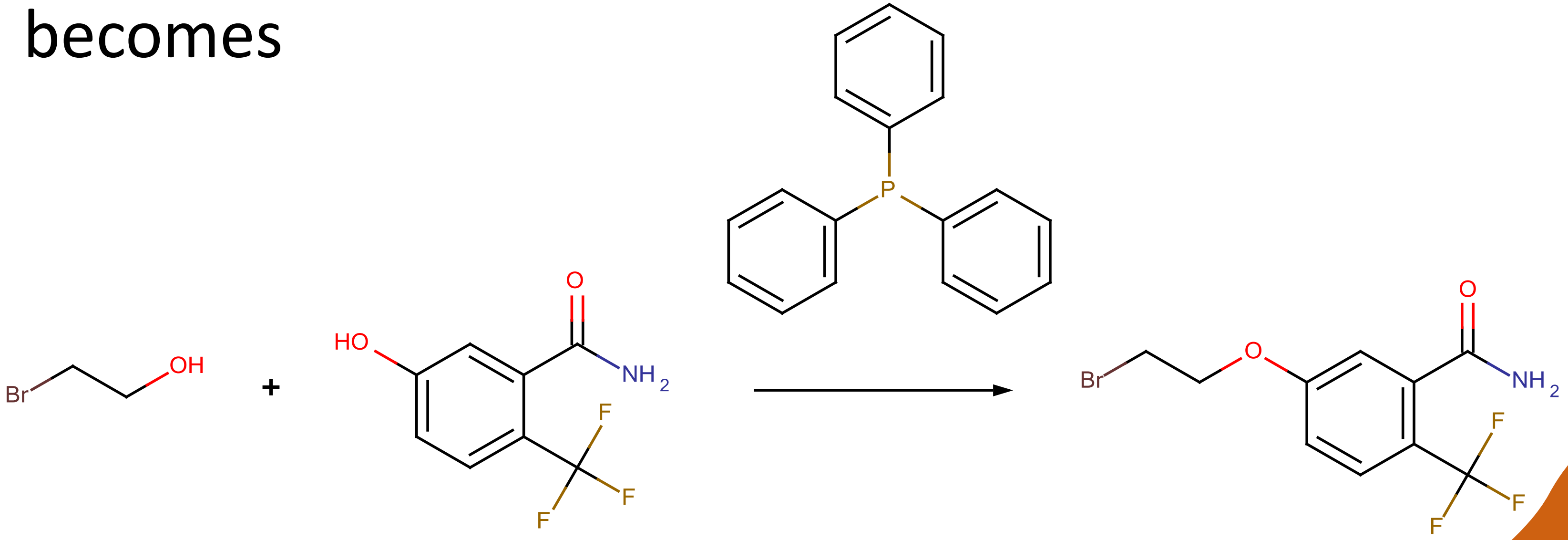
Decrypting CDX & CDXML Files

- CambridgeSoft are to be congratulated for publicly documenting their CDX and CDXML file formats.
- Unfortunately this online ChemDraw developer resource is no longer being kept up to date.
- Mistakes: “arrow” is encoded by object 0x8021.
- New tags: object 0x802b encodes “annotation”.
- Proprietary property tags: USPTO’s “PageDefinition”.
- Support for reading and writing isotopic information in CDXML files has been contributed to Open Babel.

SUPERATOM EXPANSION



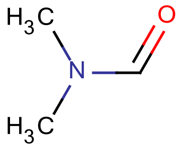
becomes

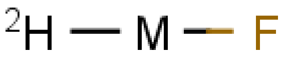


CHEMDRAW SUPERATOM CORRECTION

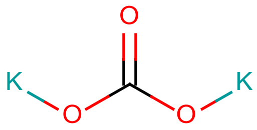
Chemist drew **Chemist Intended** **Chemist got**

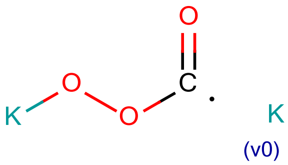
DMF



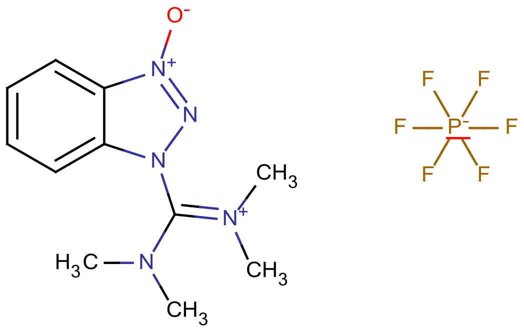


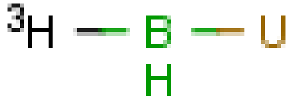
K2CO3



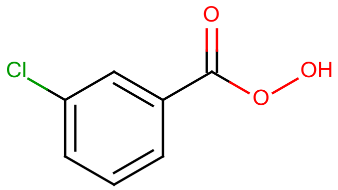


HBTU





mCPBA



Salt/Component Grouping

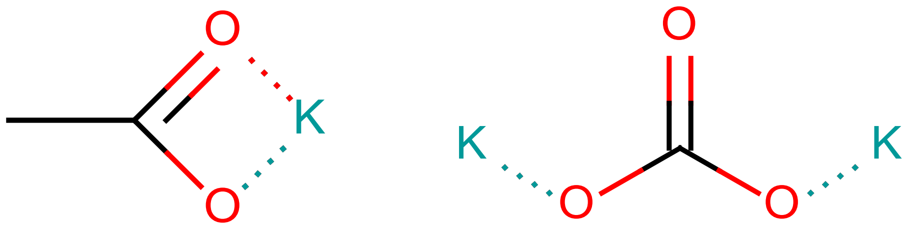
- Keeping track of the intended number and formulae of reactants, products and agents in a reaction, requires preserving salt form associations.
- This can be implemented by honoring the “group” information from the sketch as single disconnected components in MDL RXN and RD file output.
- These associations are traditionally lost in SMILES...

  …>CC(=O)[O-].[Cl-].[Cl-].[Fe].[K+].[Pd+2]…>…

  but can be retained via ChemAxon/GGA extensions to SMILES that appear as “[Na+].[Cl-] |f:0.1| salt”.

Zero Order Bonds

- Clark[1] has proposed a useful extension to MDL file format for representing bonds of order zero, “M ZBO” which solve a number representation issues in organometallics.

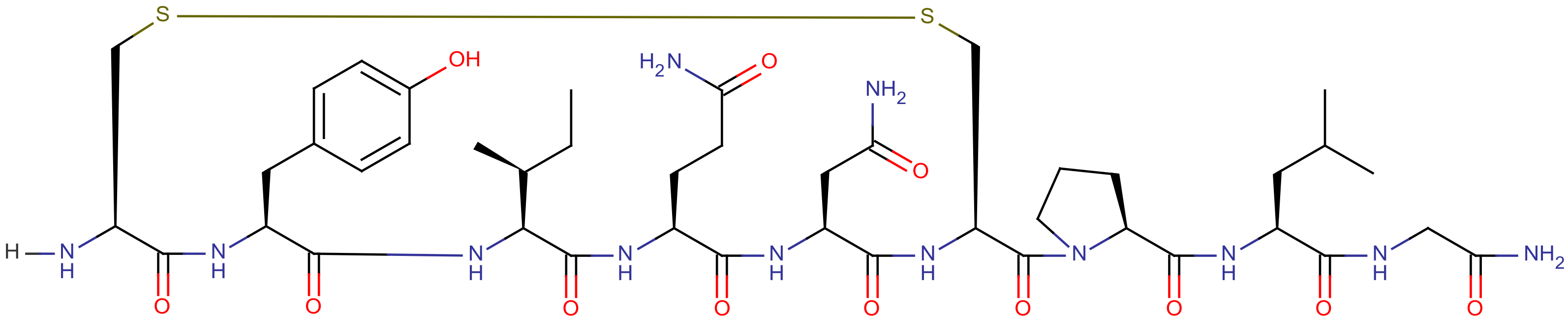


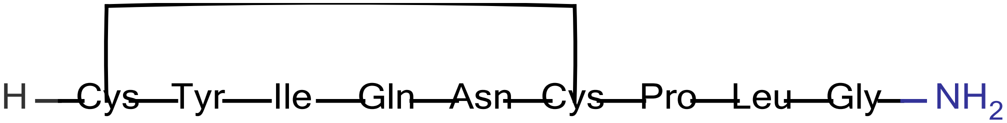
- In this presentation, we propose a novel extension to SMILES to encode/preserve these: [K]..OC(=O)O..[K]

1. Alex M. Clark, “Accurate Specification of Molecular Structures: The Case for Zero-Order Bonds and Explicit Hydrogen Counting”, J. Chem. Inf. Model. 51:3149-3157, 2011.

BIOMOLECULES AS MDL SGROUPS

- Peptides, nucleic acids and sugars can be annotated in a number of ways in molfiles, typically as abbreviated Sgroups that define leaving groups.

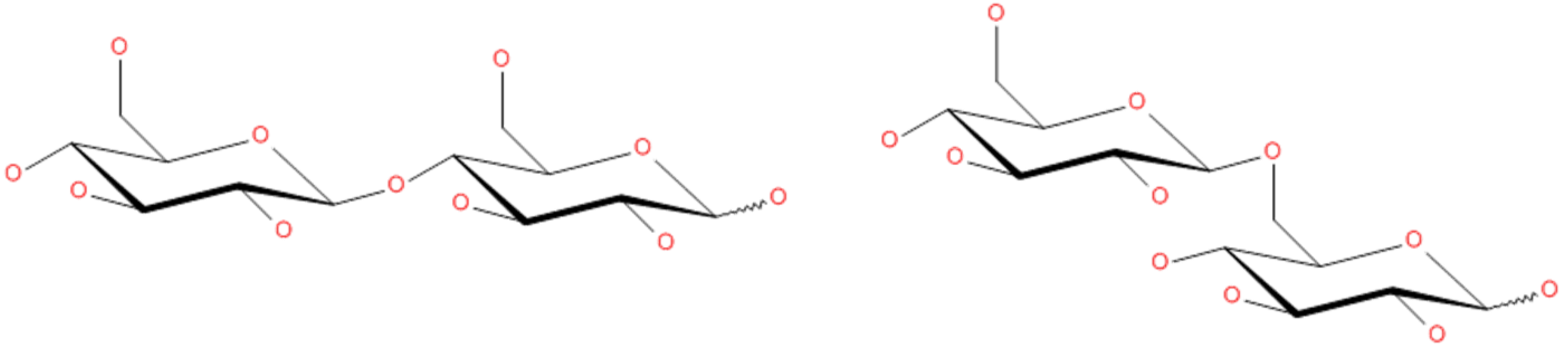




PEPTIDE1{C.Y.I.Q.N.C.P.L.G.[am]}$PEPTIDE1,PEPTIDE1,1:R3-6:R3$$$

VARIABLE ATTACHMENT POINTS

- Reported oligosaccharide structures often contain linkages whose exact attachment point is unknown.
- Glc(β1→?)Glc has four possible structures, including:



- Such structures can be round-tripped between IUPAC line notations, SMILES and MDL v2000 and v3000 molfiles (sometimes using common extensions).

VARIABLE ATTACHMENT POINTS

- Conveniently, these formats use the same “dummy atom” semantics to represent a set of atoms.
- **SMILES** does not support positional variation, but ChemAxon use an extension: `CCCl.Cl* |m:4:1.0|`
- **V3000 molfiles** from both Accelrys Draw and Marvin store variable attachments in the bond block

```
M  V30 7 1 8 7 ENDPTS=(3 3 2 1) ATTACH=ANY
```

- **V2000 molfiles** from ChemAxon Marvin and ACDLabs ChemSketch use their own incompatible extensions.

ACKNOWLEDGEMENTS

- Plamen Petrov, AstraZeneca, Molndal, Sweden.
- Sorel Muresan, AkzoNobel, Stenungsund, Sweden.
- Richard Hall, Astex Pharmaceuticals, Cambridge, UK.
- Markus Sitzmann, NCI, Washington DC, USA.
- Wolf-Dietrich Ihlenfeldt, Xemistry GmbH, Germany.
- Evan Bolton, PubChem, NCBI, Washington DC, USA.
- Daniel Lowe and Noel O’Boyle, NextMove Software, Cambridge, UK.

- Thank you for your time. Questions?