BUILDING ON SAND

Standard InChIs on non-standard molfiles

John Mayfield, Roger Sayle

NextMove Software Ltd

State and Future of the IUPAC InChI, NIH, 16-18 Aug 2017

MDL VALENCE (MDLBENCH1)

	2012			2017		
	Version	Accuracy	Precission	Version	Accuracy	Precission
CDK	1.4.13	92.65%	95.11%	2.0	100.00%	100.00%
Open Babel	2.3.90	91.73%	93.34%	GitHub	100.00%	100.00%
MDL/BIOVIA Direct	8.0	90.30%	99.76%	2017	97.67%	97.73%
OEChem	1.9	97.20%	99.78%	20170613	97.20%	99.78%
ChemAxon	5.1	88.98%	92.99%	17.17	93.13%	97.33%
GGA/EPAM Indigo	1.1.4	70.80%	97.52%	1.3.0.r16	97.22%	97.22%
RDKit	2012.09	13.62%	22.74%	2017.03.03	67.30%	85.83%

Valence defined either **explicitly** (safe) or implicitly as a **default** value

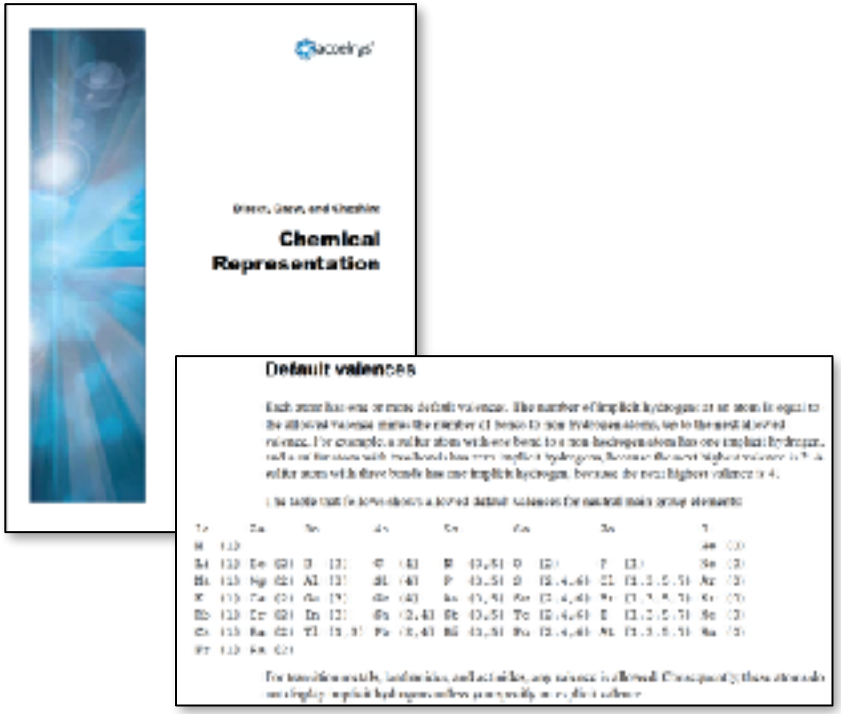
“The correct valence is specified by MDL/ISIS”

Roger Sayle, MDL Bench, Cheminformatics Toolkits: A Personal Perspective, ***RDKit UGM***, Oct 2012

MDL VALENCE-MAGEDDON

BIOVIA 2017 changes the interpretation of MDL files

Changes MF of **213,097** records in PubChem Compound



accelrys

Direct, Query, and Reaction

Chemical Representation

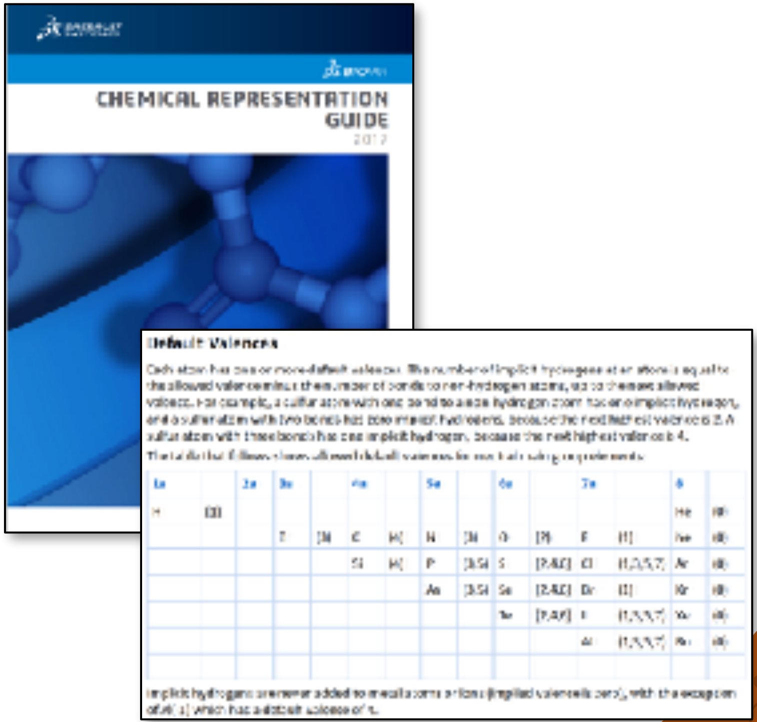
Default valences

Each atom has one or more default valences. The number of implicit hydrogens on 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		0	
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.



DASSAULT SYSTÈMES

BIOVIA

CHEMICAL REPRESENTATION GUIDE

2017

Default Valences

Each atom has one or more default valences. The number of implicit hydrogens on 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 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.

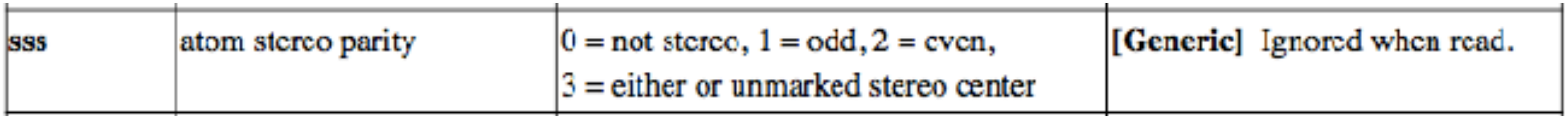
MDL MASS DELTA (MDLBENCH2)

BIOVIA Direct 2017	^{11}B	^{128}Te	^{266}Sg
CDK 2.0	^{11}B	^{130}Te	^{258}Sg
ChemAxon 17.17	^{11}B	^{130}Te	^{0}Sg
DataWarrior 4.6.0	^{11}B	^{130}Te	^{0}Sg
InChI 1.0.5	^{11}B	^{130}Te	^{269}Sg
Indigo 1.3.0b	^{11}B	^{128}Te	^{271}Sg
OEChem 20170613	^{11}B	^{130}Te	^{263}Sg
Open Babel 2.4.1	^{10}B	^{127}Te	^{271}Sg
RDKit 2017.03.03	^{11}B	^{130}Te	^{271}Sg

MDL files originally stored **atomic mass** delta

- InChI inherited this decision
- Resolved by **`M  ISO`** in molfile

STEREO PARITY (MDLBENCH3)

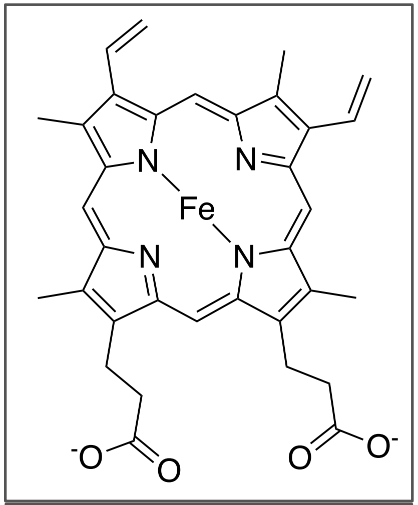


sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	[Generic] Ignored when read.

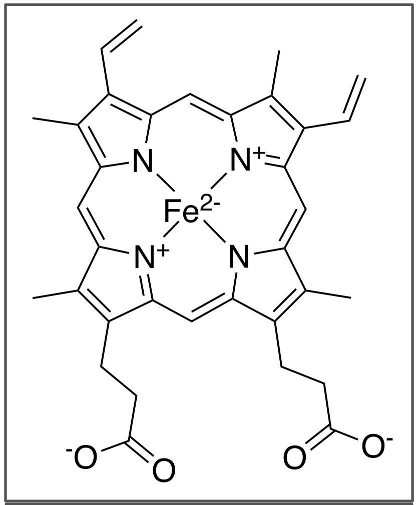
	0D				2D				3D			
	0	1	2	3	0	1	2	3	0	1	2	3
ChemAxon 17.17	-	*S*	*R*	-	-	-	-	-	*R*	*R*	*R*	*R*
CDK 2.0	-	*S*	*R*	-	-	-	-	-	*R*	*R*	*R*	-
Open Babel 2.4.1	-	*S*	*R*	-	-	-	-	-	*R*	*R*	*R*	*R*
OEChem 20170613	-	*S*	*R*	-	-	*S*	*R*	-	*R*	*R*	*R*	*R*
InChI 1.0.5	-	-	-	-	-	-	-	-	*R*	*R*	*R*	*R*
RDKit 2017.03.03	-	-	-	-	-	-	-	-	-	-	-	-
BIOVIA Direct 2017	-	-	-	-	-	-	-	-	-	*R*	*R*	*R*
Indigo 1.3.0b	-	-	-	-	-	-	-	-	*R*	*R*	*R*	*R*

Table shows default behaviour, often can be tweaked – Open Babel and CDK have options to use parity value for 2D input.

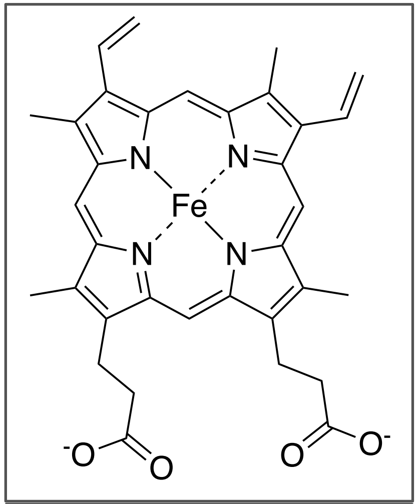
ZERO-ORDER BONDS



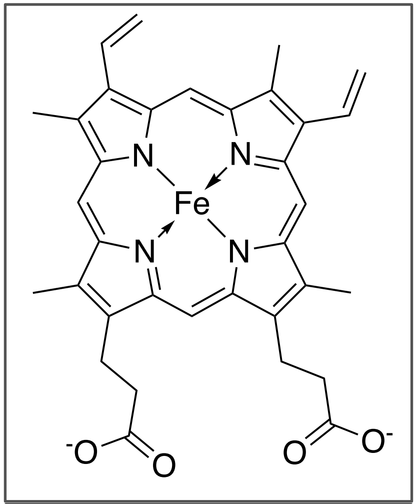
Omitted



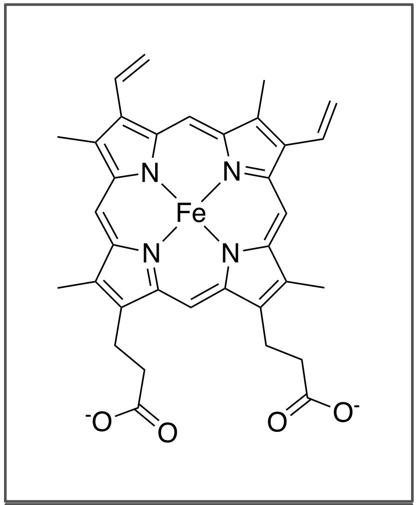
Charge Separated



Dashed



Coordination



Plain

Bonding required to describe **configuration**

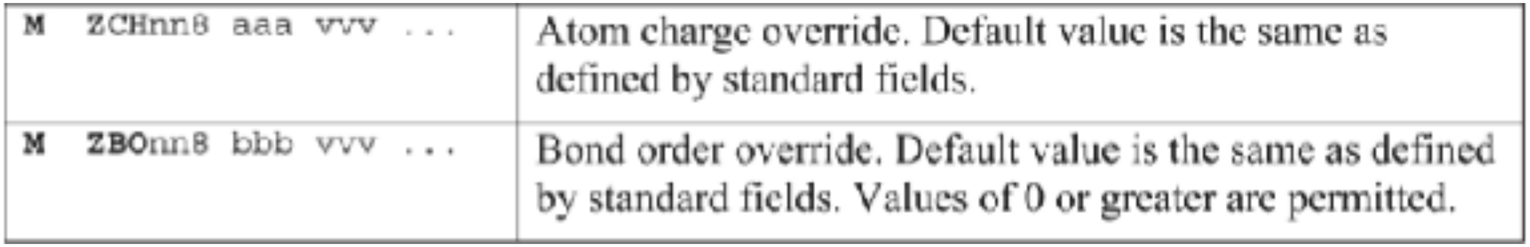
Representation part of the solution (and sometimes part of the problem), normalisation still required

How can they be represented in a **molfile**?

CTAB REPRESENTATION

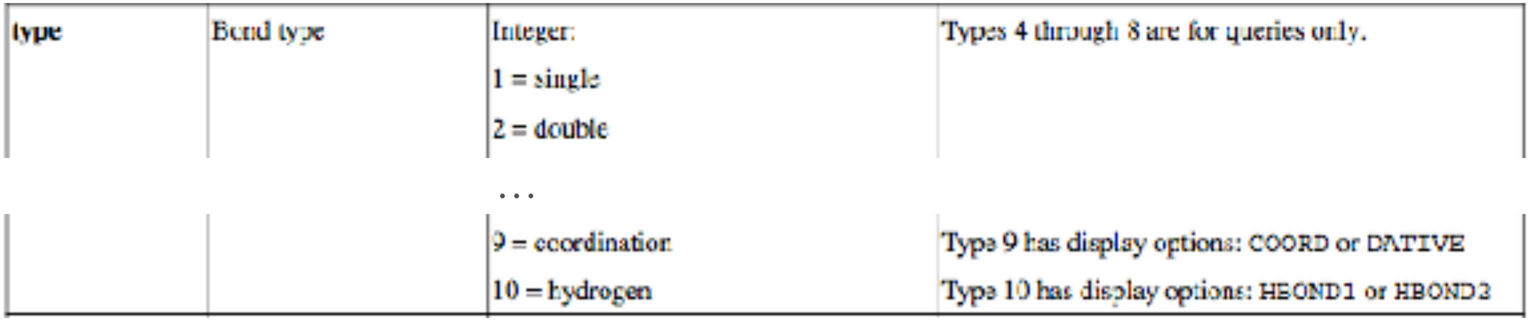
(Syntax Extensions)

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



M ZCHnn8 aaa vvv ...	Atom charge override. Default value is the same as defined by standard fields.
M ZBOnn8 bbb vvv ...	Bond order override. Default value is the same as defined by standard fields. Values of 0 or greater are permitted.

CTfile Formats “Nov 2011 onwards” V3000 only, many tools allow it in V2000



type	Bond type	Integer: 1 = single 2 = double	Types 4 through 8 are for queries only.
		…	
		9 = coordination 10 = hydrogen	Type 9 has display options: COORD or DATIVE Type 10 has display options: HBOND1 or HBOND2

CTAB REPRESENTATION

(Semantic Extensions)

PubChem SD File Formatted Data V2.0.1
ftp://ftp.ncbi.nih.gov/pubchem/specifications

```
BondTypeID        Meaning
----------        -----------------
     5            Dative Bond
     6            Complex Bond
     7            Ionic Bond
   255            Unspecified or Unknown Connectivity
```

ChemAxon specific information in MDL MOL files,
http://docs.chemaxon.com

```
M  STY  1   1 DAT
M  SAL   1  2  12  29
M  SDT   1 MRV_COORDINATE_BOND_TYPE
M  SDD   1     0.0000    0.0000    DR    ALL  0       0
M  SED   1 31
```

SUMMARY

Systematic benchmarks highlight differences in **interpretations**

- Often simple to change, but can need agreement
- Chemistry is a moving target

Existing different ways the format has been **enhanced** to handle zero-order bonds

- Can cause unexpected behaviour elsewhere
- Normalisation still difficult

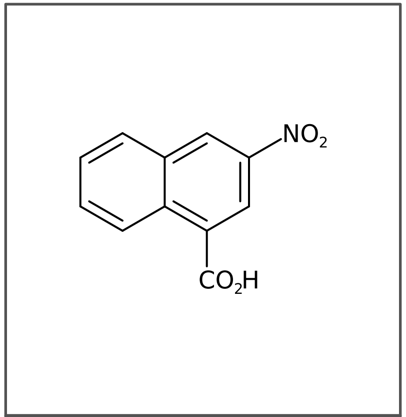
Acknowledgements

Noel O’Boyle and Shuzhe Wang

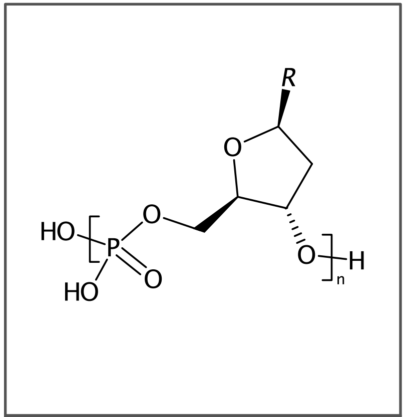
ENDS

SGROUPS

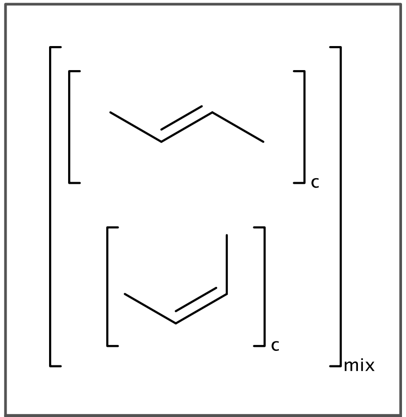
Annotation layer over **part** of a **structure**



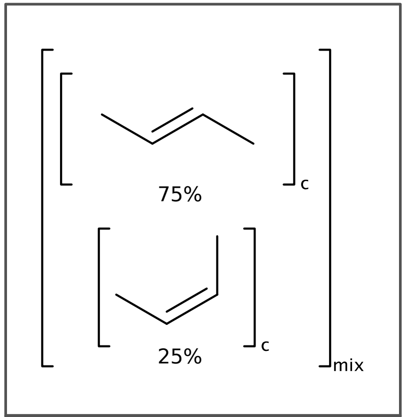
Display Shortcut



Polymer



Mixture



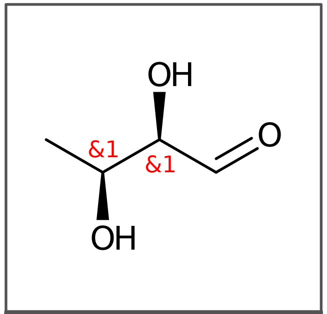
Data

Gushurst *et al*. The substance module: the representation, storage, and searching of complex structures. ***J. Chem. Inf. Comput. Sci.*** (1991)

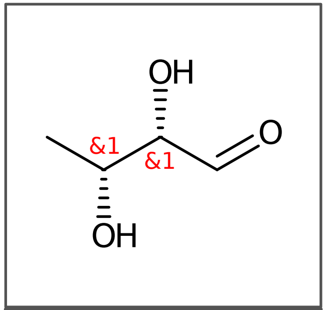
Blanke G. Sgroups – Abbreviations, Mixtures, Formulations, Polymers, Structures with Statistical Distribution and Other Special Cases. ***Online - StructurePendium Technologies GmbH***

ENHANCED STEREO 1

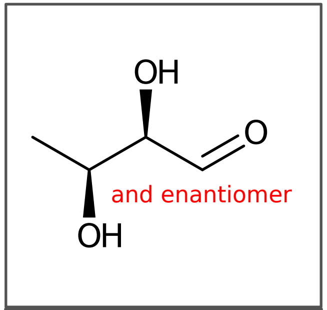
Enhanced stereo is for handling **racemic mixtures** and **relative stereochemistry**



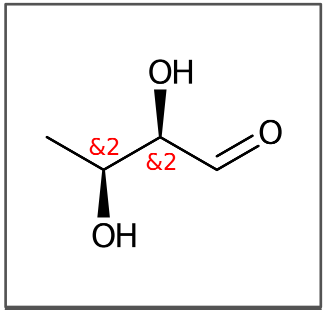
A



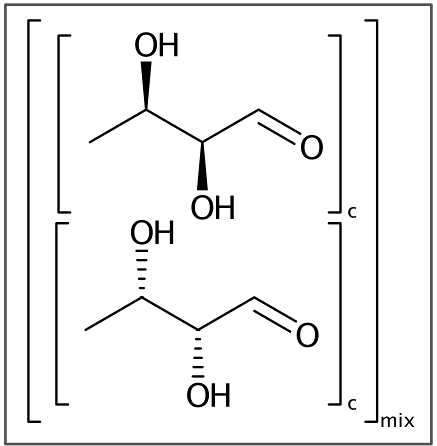
B



C



D



E

BIOVIA (NEMA-KEY)

A,B,C,D 47CZTH5YZKMZ9K3MVCCVHSUF2378UH

E NULL

ChemAxon (CXSMILES)

A,D C[C@H](O)[C@@H](O)C=O |&1:1,3,r|

B C[C@@H](O)[C@H](O)C=O |&1:1,3,r|

C C[C@H](O)[C@@H](O)C=O |r|

D C[C@@H](O)[C@H](O)C=O.C[C@H](O)[C@@H](O)C=O |…|

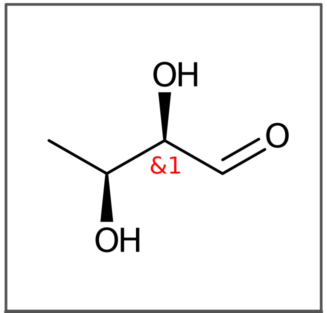
DataWarrior

A,B,C,D gNq`AjdmsURQAh@

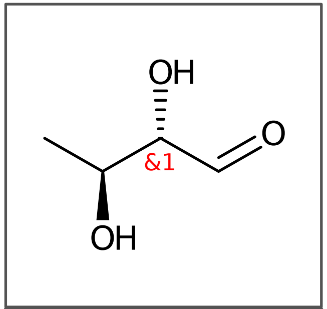
E dgLF@@rnT|bTtARfcUSUQHPUDtZP@

ENHANCED STEREO 1

Enhanced stereo is a shortcut for racemic **mixtures** and relative stereochemistry



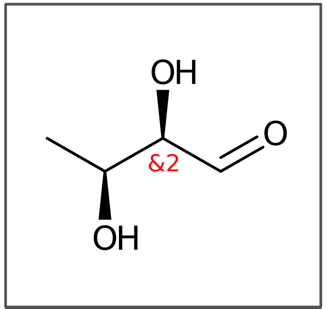
A



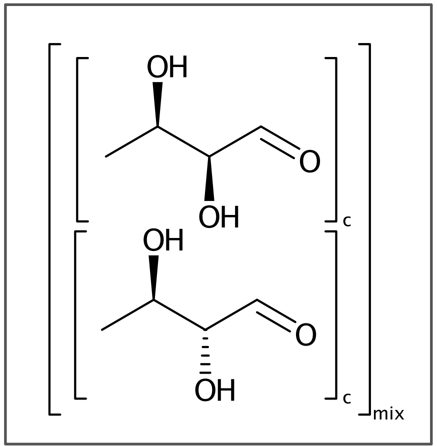
B

n/a

C



D



E

BIOVIA (NEMA-KEY)

A,B,D,E NULL

DataWarrior

```
A,B,D gNq`AjdmsURQA`@
E dgLF@@rnT|bTtARfcUSUQHPUDdZP@
```

ChemAxon (CXSMILES)

```
A,D C[C@H](O)[C@@H](O)C=O |&1:3,r|
B C[C@@H](O)[C@H](O)C=O |&1:1,r|
D C[C@@H](O)[C@@H](O)C=O.C[C@H](O)[C@@H](O)C=O |…|
```