Improvements in InChI treatment of stereoconfiguration

Andrey Yerin, ACD/Labs

Status and Future of the IUPAC InChI: context and use cases
August 24, 2017, Bethesda, US
(updated in August 2019)

Advanced Chemistry Development, Inc. (ACD/Labs)

Possible improvements in InChI treatment of stereo

Main starting point of these presentation was discussions during and after the "Expanding IUPAC Standards for Chemical Information" EMBL-EBI Industry Programme workshop in Hinxton, UK, 20-21 March 2017.

Several participants agreed that the first and most important extension of InChI stereo should be a support of enhanced stereo in MOL V3000 that is the main subject of this presentation.

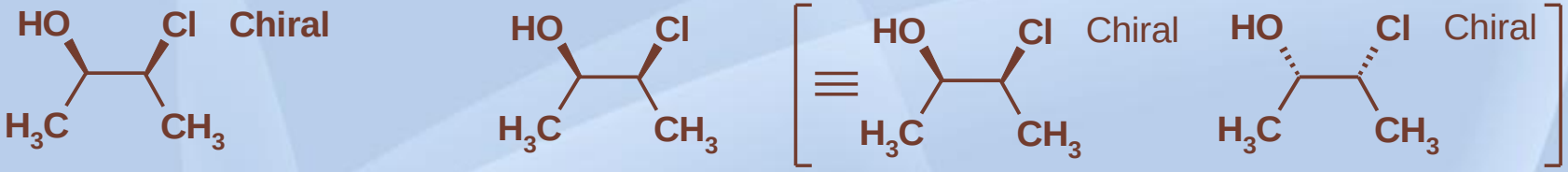
Other improvements may address the following issues and areas mentioned in InChI-discuss lists and encountered in ACD/Labs work with InChI:

- Retaining of 'stereo flag' in structure generation from InChI;
- Treatment of ambiguous representations;
- Recognition of atropisomerism, e.g. chiral biphenyls;
- Non-tetrahedral centers of chirality;
- Advanced stereo procedures

Enhanced Stereochemical Representation - reasons

MOL V2000 - 'chiral flag' definition and limited to:

- global stereoconfiguration – flag defines the whole structure, not specific centers;
- only two options available:
  - chiral flag – the specified stereoisomer only
  - no flag – racemic mixture – drawn structure AND its enantiomer



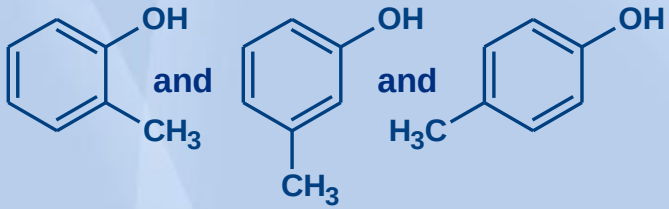
'Chiral flag' problems in MOL V2000:

- complex stereo relations cannot be represented with single structure;
- relative configuration – drawn OR its enantiomer – cannot be represented;
- absent flag default – fully valid representation treated as racemic mixture.

The last is still a very common reason of misinterpretation of stereo in MOL and SDFiles.

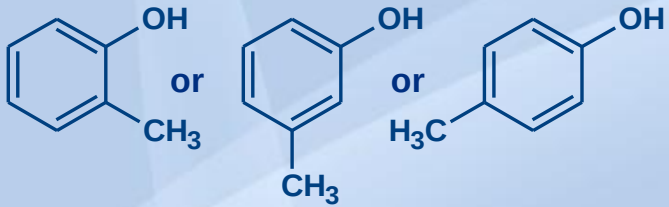
Digression – OR/AND relation in representation

A representation of isomers with explicit multiple structures is inconvenient.
Industrial cresol with CAS RN 1319-77-3 is a mixture of three isomers – "AND mixture"



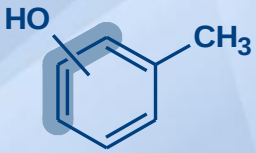
	Three structures	Single structure
Formula:	$C_{21}H_{24}O_3$	C_7H_8O
Mol.Mass:	324.4134	108.1378

Metabolic assay or other investigation may lead to definitely single product without knowledge of exact position of modification – "OR mixture"



	Three structures	Single structure
Formula:	$C_{21}H_{24}O_3$	C_7H_8O
Mol.Mass:	324.4134	108.1378

Markush structures are used to represent variable substitution in both cases



Formula: C_7H_8O
Mol.Mass: 108.1378

Still it lacks a differentiation between AND/OR mixtures.

Not for InChI yet, but probably it is time to distinguish AND/OR relations in InChI stereo.

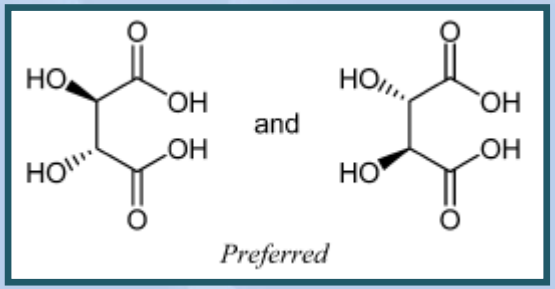
IUPAC recommendations

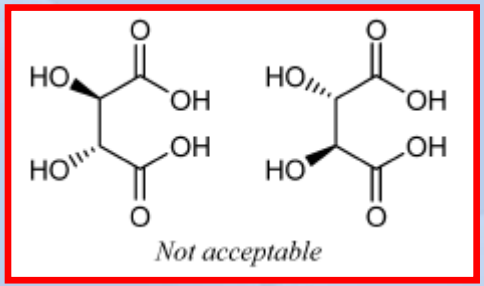
Graphical Representation of Stereochemical Configuration
(IUPAC Recommendations 2006)
***Pure Appl. Chem.*, Vol. 78, No. 10, pp. 1897–1970, 2006**

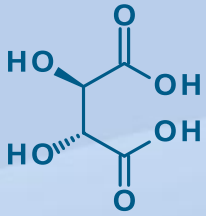
MOL V3000 enhanced stereo conventions are mentioned but it is stated that
"Given the lack of consensus among different systems and the poor general acceptance of any given one, the safest approach currently is to depict multiple structure diagrams explicitly, optionally accompanied by additional descriptive text"
and
"a comprehensive discussion of arbitrary mixtures of diastereoisomers (and mixtures in general) is outside the scope of this document."

Use of labels for configurations and isomers:
The use of the terms "racemate" and "rac" should be avoided as structural labels.
The terms "relative" and "rel" should similarly be avoided in structural diagrams.
So it is quite close to the old 'chiral flag' concept while with AND/OR distinction.

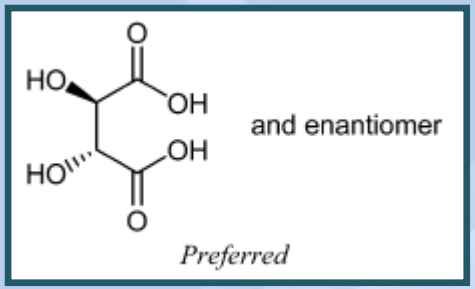
IUPAC recommendations

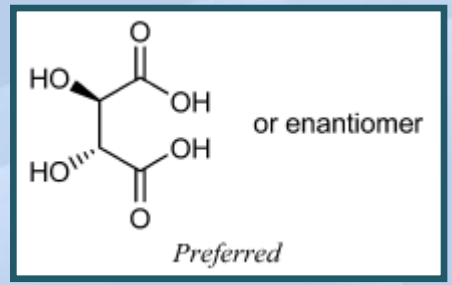


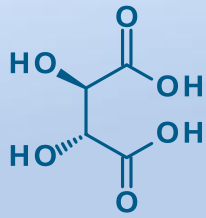




AND Enantiomer







OR Enantiomer

BIOVIA, ChemDraw

These statements were fully reasonable for 2006 and with main aim at printed structures.

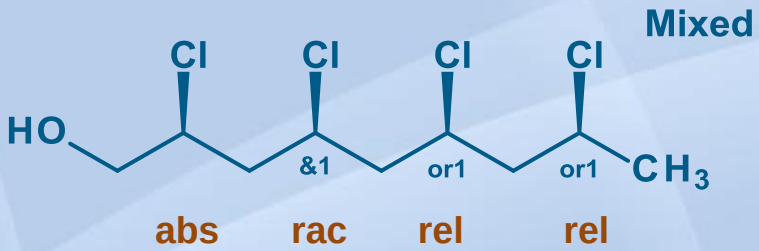
Still it seems after 15 years of existence of enhanced stereo concept and its growing acceptance the position can be reconsidered, especially for electronic representation.

Enhanced Stereochemical Representation - details

Enhanced stereo was introduced by MDL about 2002 with MOL file V3000

Main principles

- **stereo mark is a property of specific center, not the whole structure;**
- **three variants – absolute, relative and racemic;**
- **relative and racemic centers can be joined in groups**



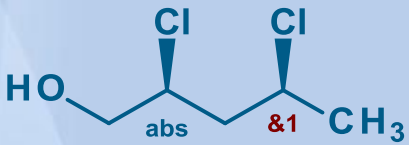
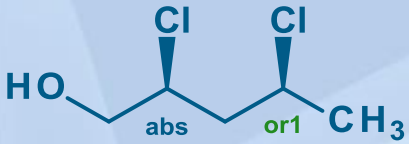
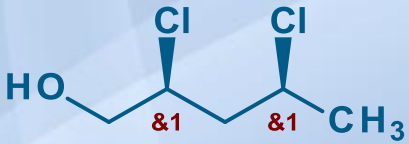
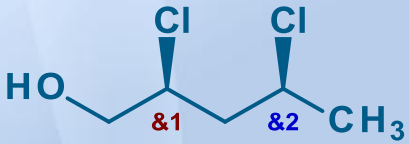
Designations:

- **absolute – no label by default or abs;**
- **racemic (AND) – & with number for group;**
- **relative (OR) – or with number for groups;**
- **complex stereo relation is indicated with "Mixed" label**

This label is hardly necessary and can be hidden

Enhanced Stereochemical Representation - details

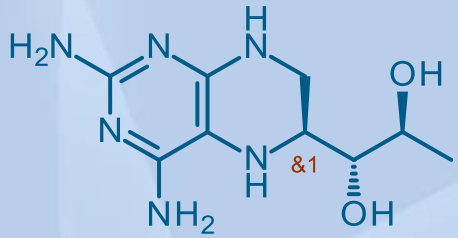
Enhanced stereo conventions allow representation of all possible stereo relations using single structure with corresponding marks

Examples	Defines mixture of
	2R,4S **AND** 2R,4R
	2R,4S **OR** 2R,4R
	2R,4S **AND** 2S,4R
	2R,4S **AND** 2R,4R **AND** 2S,4R **AND** 2S,4S **(full racemate)**

Enhanced Stereochemical Representation - examples

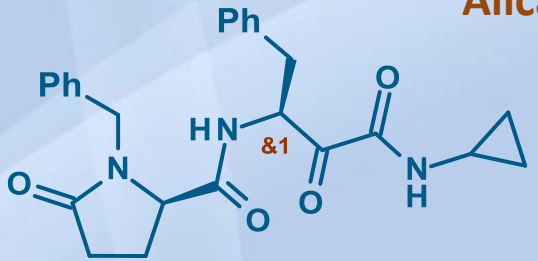
Industrial chemicals are obviously (mostly?) “AND mixtures”
In relation to stereo it means presence of absolute and racemic centers together.

Ronopterin (INN)



Defines mixture of S,R,S and R,R,S
(*RS,R,S*)

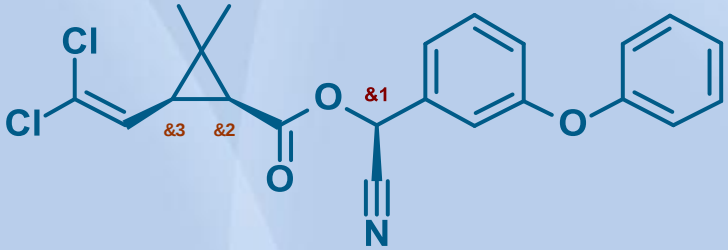
Alicapistat (INN)



Defines mixture of R,R and R,S
(*R,RS*)

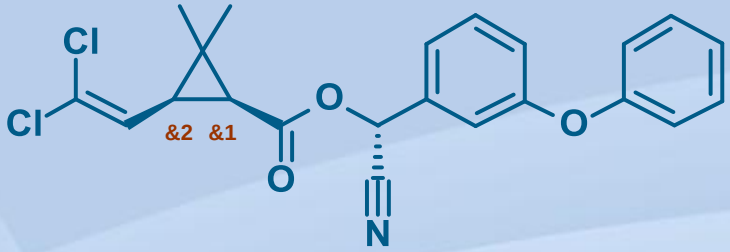
Enhanced Stereochemical Representation - examples

Cypermethrin (ISO), pesticide



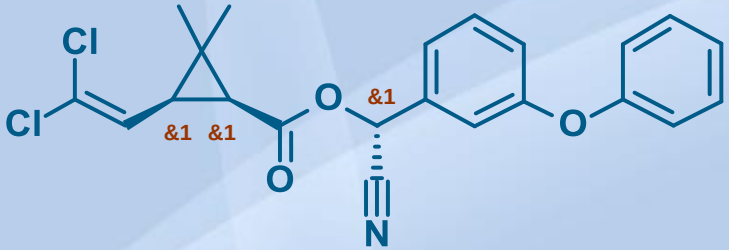
All 8 stereoisomers

ζ-Cypermethrin



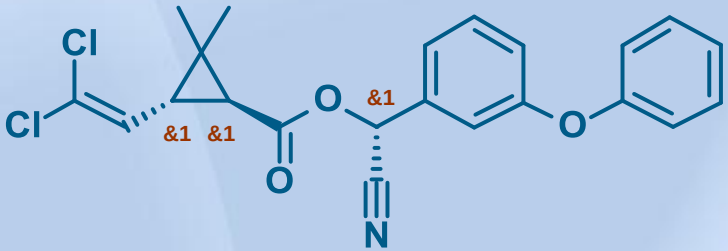
Four isomers - S,1R,3S, S,1R,3R, S,1S,3R, S,1S,3S

α-Cypermethrin



Two isomers - R,1S,3S and S,1R,3R

θ-Cypermethrin



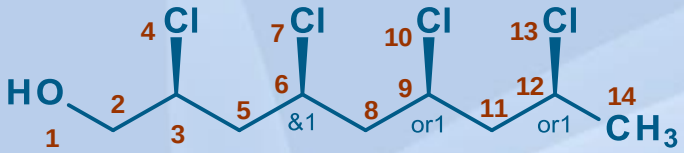
Two isomers - R,1S,3R and S,1R,3S

Enhanced stereo allows to represent selectively complex mixtures of stereoisomers

Enhanced Stereochemical Representation – InChI support

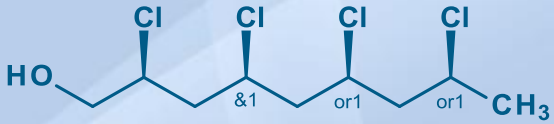
What should be done in InChI to support enhanced stereo?
Just few simple procedures are necessary:

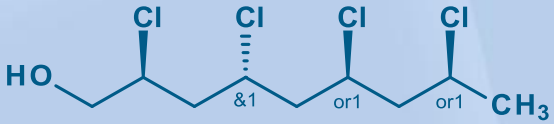
- Read MOL V3000 stereo marks;
- Add new enhanced stereo layer – like marks itself;
- Recognize equivalence of reversed stereo within each group;
- Return stereo marks while conversion of InChI to structure.

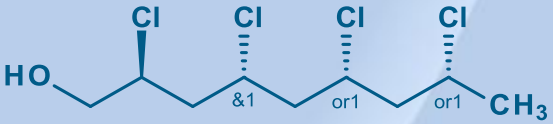


Something like …/e3a,6&1,9o1,12o1/…
a – abs, &1 – AND group 1, o1 – OR group 1, together with atoms numbers – 3,6,9,12 here

These three structures should be recognized as equivalent:



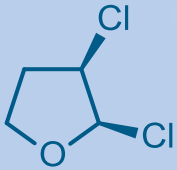




Can be introduced even in InChI version 1.x being just an addition of new layer.

Additional stereo support – lost relative or racemic stereo

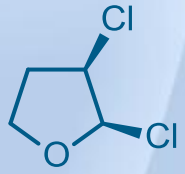
The problem was mentioned in InChI-discuss lists.
Currently InChI with global relative or racemic configurations can be generated either via options or taking into account V2000 'chiral flag' (racemic)



Standard (absolute)
InChI=1S/C4H6Cl2O/c5-3-1-2-7-4(3)6/h3-4H,1-2H2/t3-,4+/m1/s1

Racemic
InChI=1/C4H6Cl2O/c5-3-1-2-7-4(3)6/h3-4H,1-2H2/t3-,4+/s2

Racemic
InChI=1/C4H6Cl2O/c5-3-1-2-7-4(3)6/h3-4H,1-2H2/t3-,4+/s3

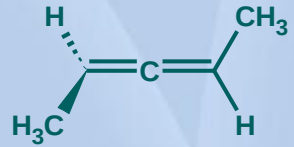


These InChIs return the same structure while structure generation.
So the information about relative or racemic relations is lost.

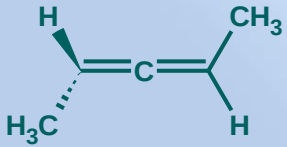
This problem can be solved in the same way as for enhanced stereo – centers should have and keep their *abs, rel, rac* attributes.

Additional stereo support - allenes

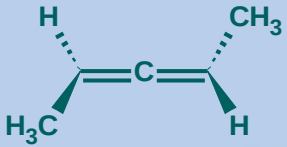
InChI works correctly with short allenes. Except some questionable presentations



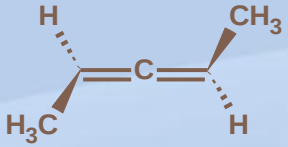
-YFKPBYRVSA-N



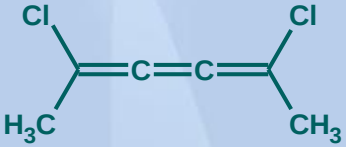
-RXMQYKEDSA-N



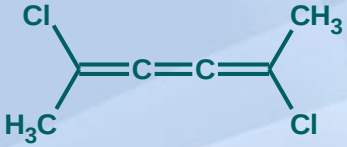
-UHFFFAOYSA-N



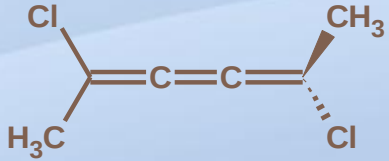
-YFKPBYRVSA-N



-WAYWQWQTSA-N



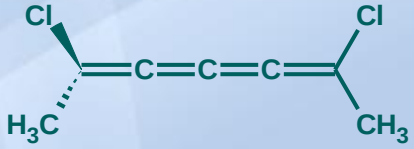
-AATRIKPKSA-N



-AATRIKPKSA-N

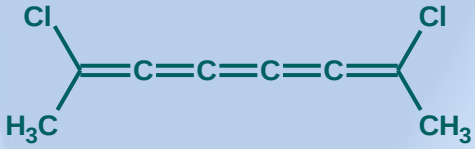
But configuration in longer allenes are not recognized.

M-isomer



-UHFFFAOYSA-N

Z-isomer



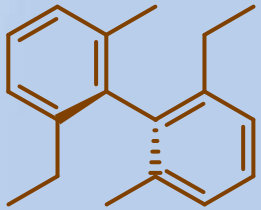
-UHFFFAOYSA-N

Hardly critical but formally incorrect. Should be easy to implement.

Additional stereo support - atropisomers

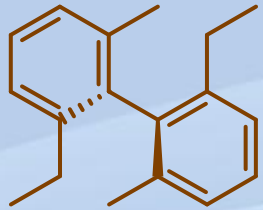
Designation of configuration in chiral biphenyls is not supported

***M*-isomer**



-UHFFFAOYSA-N

***P*-isomer**



-UHFFFAOYSA-N

Should be relatively easy to implement – the procedure is very close to M/P-allenes. The corresponding stereo can be put in 't' layer where both tetrahedral and allenes' stereo is encoded with +/- designation.

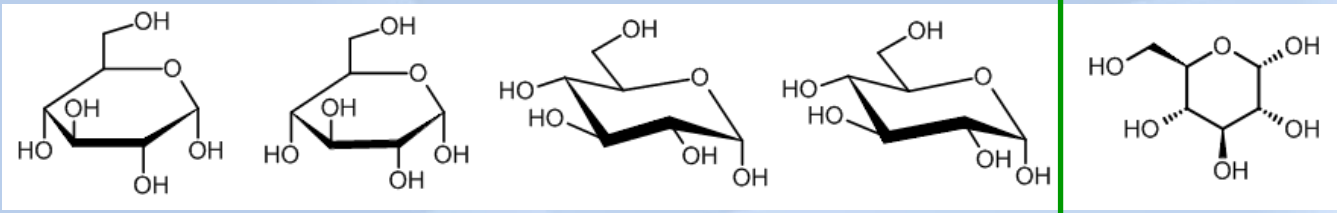
Again hardly critical but really desirable as soon as:

- **such stereoisomers are quite abundant;**
- **does not need much work;**
- **can be implemented in InChI version 1.x**

Additional stereo support – problem of Haworth presentation

Slide from my presentation at the EMBL-EBI Workshop in March

InChI Key second block for Haworth, conformational and Mills representations of α-D-glucopyranose



ChemDraw	DVKNGEFBSA	DVKNGEFBSA	DVKNGEFBSA	DVKNGEFBSA	DVKNGEFBSA
ChemSketch	Refused	Refused	Refused	Refused	DVKNGEFBSA
Accelrys	SADXIQTQSA	PRPIVIPDSA	KYKSVUQTSA	SADXIQTQSA	DVKNGEFBSA
Marvin	SADXIQTQSA	PRPIVIPDSA	KYKSVUQTSA	SADXIQTQSA	DVKNGEFBSA

Note different Keys for visually the same Haworth and conformational pairs that differ only in drawing order (= internal numbering of atoms)

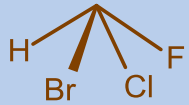
This problem should be resolved!

Improvement of InChI stereo – ambiguous presentations

InChI and InChIKey are generated for some obviously wrong arrangements

InChI=1S/C4H6Cl2/c1-3(5)4(2)6/h1-2H3/**b4-3-**

YRCBNVMRRBIDNF-ARJAWSKDSA-N



InChI=1S/CHBrClF/c2-1(3)4/h1H/**t1-**/m1/s1

YNKZSBSRKWVMEZ-PVQJCKRUSA-N

Such arrangements should lead to at least unknown stereo or even to refusal.
Every procedure should avoid generation of knowledge from garbage!

Additional changes can be made for ambiguous or one step from wrong presentations

-UPHRSURJSA-N

-UHFFFAOYSA-N

-SFOWXEAESA-N

Some valid arrangements are not recognized

-UHFFFAOYSA-N

Improvement of InChI stereo – non-tetrahedral centers

Encoding of non-tetrahedral stereo needs several procedures and solutions:

- **Unambiguous representation standards;**
- **Special elements list for specific polyhedral arrangements;**
- **Procedures to determine configurations;**
- **Encoding – more than two isomers are possible for each.**

Again considerations of representation for polyhedral arrangements can be found in

Graphical Representation of Stereochemical Configuration
(IUPAC Recommendations 2006)
***Pure Appl. Chem.*, Vol. 78, No. 10, pp. 1897–1970, 2006**

Every arrangement with up to six ligands is considered and all possible representations are classified as recommended, acceptable and not acceptable.
Really excellent source of examples to consider for non-tetrahedral stereo.

Still these recommendations need some reconsideration for electronic representation needs.

Improvement of InChI stereo – non-tetrahedral centers

Square planar configuration – *SP*-4 – 11 arrangements provided:
1 preferred, 3 acceptable and 7 not acceptable

- **1st is perfect and unambiguous;**
- **2nd – nothing bad except a bit uncommon perspective, can be accepted;**
- **3rd and 4th – look OK but missing some 'stereo indication' that specific configuration is really intended and not just occasional arrangement;**
- **5th – formally nothing bad and can be accepted;**
- **6th – questionable but still can lead to detection of flat square.**

In such way all designations for every configuration can be analyzed.

The nature of central atom should be taken into account. It can be done via lists of allowed configurations for elements.

Improvements in InChI treatment of stereo

Proposed tasks for improvement of InChi stereo:

1. Support of enhanced stereo in MOL V3000;
2. Retaining of ‘stereo flag’ in structure generation from InChI;
3. Recognition of atropisomerism, e.g. chiral biphenyls;
4. Recognition of sugars in Haworth and chair representations;
5. Better treatment of ambiguous and wrong representations;
6. Non-tetrahedral centers of chirality – mostly for organometallics, but not only;
7. Additional stereo procedures – new classes of stereo and specific cases mentioned during this meeting, e.g. with chiral center at sulfur.

How to proceed?

‘Enhanced stereo’ is clear and almost ready to implement – layer format to choose;

Other areas need further discussion and may benefit from arranging of special project team.

The considerations will be useful both for InChI and general recommendations on definition of stereochemical configurations.

Improvements in InChI treatment of stereo

The project was submitted in April 2019
Project team

Andrey Yerin, ACD/Labs, Russia;
Gerd Blanke, StructurePendium Technologies GmbH, Germany;
Igor Pletnev, Moscow State University, Russia;
Burt Leland, OpenEye Scientific Software Inc.
Jürgen Kammerer, Sanofi-Aventis Deutschland GmbH

Few days ago we received reviews and they are positive.
We will respond to the comments and hope that the project can formally start.

For questions and proposals:
Andrey Yerin, ACD/Labs,
IUPAC Chemical Nomenclature and Structure Representation (Division VIII)
erin@acdlasb.ru