MOSFECCS v6 SMILES-Generator Technical Manual

1. Canonicalization: Generation of a unique SMILES code for a single molecule

Definition:
A canonical SMILES (unique SMILES) code is independent of the order in which atoms and bonds have been drawn or deleted and is independent of the 2D-coordinates or orientation of the 2D structural drawing.
Any way to draw a given structure leads to the same and only version of the SMILES. This requires an algorithm for *unique numbering* of the atoms or *canonicalization*.

The order of the atoms as stored in the computer's memory is dependent on the order of generation/deletion. Canonicalization requires a unique numbering (classification) of the atoms that depends only on the invariants stored with each atom and the molecular graph (directed and weighted to include the bond order and, for stereo-up/down bonds, the direction of bonds).

Procedure for molecules without topological symmetry
H-atoms can be either explicitly drawn or implicitly assumed on the basis of normal valency (for so called "organic" elements or any element with a defined standard valency ->*see* "chemical" constants section in appendix 2). To treat all H-atoms identically, the first step is to strip off all explicitly drawn hydrogen atoms and transform them into implicit H. To keep record of explicitly drawn hydrogen atoms, especially important for hydrides of elements without standard valency, they are registered in the properties of the non-H atom to which they are attached in the drawing.

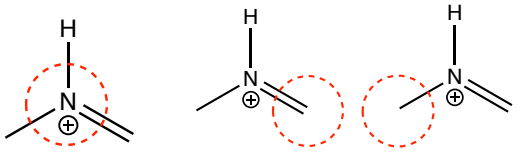
Next, a so-called invariant for each non-H atom is generated: it includes all information on the atom that does not depend on the properties of other atoms (same procedure as in Weininger et al. *J. Chem. Inf. Comput. Sci*. **1989**, 29,97-101).

The invariant is a 9-digit integer number composed of the following sections:

1. number of connections 1 digit
2. sum of bond orders to non-H atoms 2 digits
3. atomic number 2 digits
4. sign of charge 1 digit
5. absolute charge 1 digit
6. number of attached H 2 digits (0-4) or residue(10-99).

Because classification is based on numerical sorting of the invariants, section 1 has the highest priority and section 6 the lowest priority as a distinguishing criterium.

Examples:



invariant:	203071101	102060002	101060003
class:	3	2	1

The atoms are then sorted according to increasing invariants and assigned a class, which is the ranking in the ordered list of invariants.

In the next step, the classes of the atoms one connection (bond) away are used to generate a unique prime product: not the classes but the i-th prime numbers (if the bonding partner has class i) are multiplied and this prime product is added to the original class of the atom•10^{12} to give a new invariant. Sorting and classification (rank in the sorted list of the new invariants) gives the new class. This procedure is recursively iterated until the number of classes remains constant. In the absence of topological symmetry, it will generate as many classes as there are non-H atoms. With topological symmetry, symmetry equivalent atoms will still have the same class. See below on how these "ties" or degeneracies can be resolved.

Let us assume, that we have achieved such a unique classification with the number of classes equal to the number of atoms (one atom per class, see below).

The SMILES code is then generated by a DFS (depth-first search) of the molecular graph.

The first decision is where to start. Since the dead ends (atoms with only one connection) often belong to the lowest class (e.g. methyl groups), the DFS is started with the atom of lowest class.
The second type of decision is required at each branching point (any atom with more than two connections) and concerns the order in which the connected atoms (branches) are visited in the search (so called preordering). If the branches are sorted according to the class number of the next atom either in descending or ascending order, the resulting SMILES will be unique. The chosen rule is to visit the branches in the order of increasing class number of the directly bound atom at each branching point.

Rings:
Whenever a ring is followed in the DFS, the ring closure can be recognized because a previously visited atom is encountered again. Since this previously visited atom has already been inserted in the SMILES string at this point, the ring closures (from-to) must be stored separately for later insertion into the SMILES string (the original SMILES convention uses integer numbers as designators of ring closures).
Because the SMILES syntax does not allow ring closures by multiple bonds, a mechanism must ensure that rings are always closed with single bonds. This requires

sorting the branches with a different priority rule than “lower class number” upon entry into a ring.

In the original paper by Weininger et al. 1989, the rule is stated as follows:
When entering a ring:

a) Follow a multiple bond, if present, before a single bond.

Otherwise (not entering a ring or no multipe bond in ring):

b) Follow the branch with the lower class of the first atom.

Since, during a one-pass DFS, it is not known that a ring has been entered until it is closed, a previous search must identify and register all ring-bonds.

From the paper cited above it is not clear, how the Daylight USMILES procedure determines that a branch decision coincides with ring entering. A secure method, used by MOSFECCS, is to temporarily delete each bond and search by DFS if the resulting structure is still a spanning tree of all atoms. If yes, the temporarily deleted bond is a ring bond.

Resolving symmetry-induced ties
(atoms with the same class at branching points)

If the molecule has topological symmetry, as many atoms as the symmetry number indicates can have the same class. If the first atoms of two or more branches have the same class (so-called tie), a procedure ("tie breaking") is needed to decide on the order in which they will be searched in the DFS. Weininger (*J. Chem. Inf. Comput. Sci.* **1989**, *29*,97-101) used a tie-breaking procedure where the *first* atom of the lowest degenerate class is chosen arbitrarily and made unique by doubling all class numbers and lowering the class of the selected atom by one. Then classification and tie-breaking are used iteratively until the number of classes is the same as the number of atoms. However, *this procedure is not unique because it depends on the original numbering of the atoms (“first” meaning the first in the stored connectivity list) and depends on the drawing/deleting history of the structure* (*see* paper discussing cases where the tie-breaking produre of Weinberg fails by Neglur et al. DILS 2005, LNBI 3615, pp 145-157, 2005, Ludäscher and Raschid, Eds.).

The second problem of the procedure of Weinberg is that the method with prime-products of classes reflects only the constitutional equivalence of subtrees (branches) in the graph but does not consider stereochemical differences inside the compared branches.

MOSFECCS uses the following approach to cope with the tie-break problem induced by symmetry:

1. If several atoms are of the lowest class 1, each one is used as a starting point of the DFS to generate a SMILES string. Exception: if all atoms of the molecule have class 1 (e.g. cyclohexane), any one of them can be used as a single starting point without losing completeness.

2. If branching points with two or more degenerate branches (same class for the first atom in each branch) are encountered during the search, all possible permutations at each branching point are combined into a permutation list and all combined permutations on the list are searched by DFS.
3. Any SMILES generated this way is added to the collection of alternative SMILES, if it is new (different from all previously collected SMILES).
4. A lexicographic sorting of all different SMILES in the final collection is carried out and the first SMILES in the sorted list is returned as the canonical SMILES.

It can be shown that searching the degenerate branches at the first branching point in different order leads to identical SMILES if there are no stereochemical configurations at any atoms or double bonds. In this case, one of the branches at the first breaking point can be given arbitrary preference (as in Weinbergs procedure) without losing completeness. If stereo configurations are present (E/Z double bonds or chirality centers), all permutations have to be searched including those at the first degenerate branching point.

In terms of computation time, this "brute force" procedure is more expensive than the (more involved) algorithm published by McKay (B.D. McKay, *Congressus Numerantium*, **1981**, *30,*45-87; B.D.MacKay, A.Piperno, *J. Symb. Comput.* **2014**,*60,*94-112), which identifies isomorphous branches during the search and eliminates them from further searching. In practice, however, the "brute force" approach is fast enough for the few molecules of modest size that are typically present in the editor in on-line learning environments. On modern computing devices such as laptops and tablets the SMILES calculation of MOSFECCS is practically instantaneous. If intended for use in the context of database work with thousands of molecules, the more efficient McKay algorithm should be implemented.

2. Canonicalized multi-SMILES

When several molecules are drawn on the editor canvas, the SMILES generator generates a canonical SMILES string for each one. According to the standard SMILES convention, these SMILES strings are concatenated into a multi-SMILES string with a period as the separator. The order in which the SMILES of individual molecules appear in the multi-SMILES is not specified by the published conventions.
MOSFECCS concatenates them in the order of decreasing length of the SMILES string and, if two or more have the same length, in the inverse alphabetical order (later in alphabet first) to keep the multi-SMILES canonical.

3. Stereochemical indicators in SMILES

The published SMILES convention (*see* e.g. Daylight™-SMILES theory manual: http://www.daylight.com/dayhtml/doc/theory/index.pdf) uses @ and @@ to indicate the local sense of chirality at tetrahedral stereogenic centers (SC):
The convention is to look from the bonding partner listed before the stereogenic center (SC) in the SMILES towards the SC. The remaining three bonding partners are then ordered 1-2-3 *according to their order of appearance in the SMILES*. If the

sequence 1->2->3 forms a counterclockwise curve in space as seen from the incoming ligand, the SC receives the @ designator, if it forms a clockwise curve, the designator is @@.

Configurations at double bonds (E/Z or cis/trans) are designated by slashes / and \ before and after the double bond.

MOSFECCS uses the same "local" stereo-descriptors as Daylight™-SMILES theory manual.
(for details about how MOSFECCS determines the direction of the sequence 1->2->3 in 3D *see* appendix 3).

So far (MOSFECCS v6), the more General Chiral Specifications (TH, AL, SP, TB,OH) defined in the original SMILES convention and in section 3.3.4 of the Daylight™-SMILES theory manual are not implemented in MOSFECS. Chiral cumulenes (AL) are specified by @ and @@ like tetrahedral stereogenic centers (SC).

The following rules are applied hierarchically to decide whether an atom is designated as @ or@@:

1. Only atoms that have at least one wedge-bond (stereo-up or stereo-down) are considered as candidates for SC.
2. Only atoms in a predefined list of potentially stereogenic bonding configurations for tetrahedral four-coordinate and pyramidal three-coordinate centers are considered (*see* appendix 2). The bonding configurations are defined by the atomic number, the number of ligands, the number of implicit hydrogens, and the charge.
3. The stereo-descriptors @ or @@ are only applied to centers with four different ligand branches (at trigonal pyramidal centers, one of them is a lone pair and is assigned Nr. 1 in the sequence 1->2->3).

Ligand branches are compared pairwise as follows:

1. A directed breadth first search (BFS; no backtracking towards the SC) starting at the first atom of the branch and ending at a terminal atom is carried out: If the two ligand branches have a different number of atoms, they are different.
   If the number of atoms is equal, atoms in the same ligand sphere (n-bonds away from SC) are compared by calculating the prime product of their classes When the prime-product in the n^{th} sphere is different, the ligands are considered different and the search is terminated. If they are equal, the next sphere (n+1) is examined.
2. If the prime-products at each ligand sphere are identical (the two branches are constitutionally isomorphous) the configurations in the ligands are compared by a depth-first search and the stereo-designators (@,@, E/Z double bonds or none) of the n^{th} atoms in the search path of each ligand are compared. This allows to classify the ligands as homotopic (identical

stereodesignators at each n), enantiotopic (opposite stereo-descriptors at each n) or diastereotopic (some stereo-descriptors identical, others opposite). Homotopic ligands are considered as equal whereas enantiotopic and diastereotopic ligands are different.

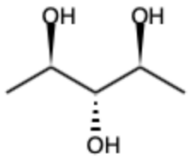
3. This procedure is carried out iteratively (with E/Z-double bond detection in each iteration as well) because step 3 (finding configurational differences in the ligand branches) depends on having successfully found stereogenic centers earlier in the process. The iteration is terminated when the number of stereo-designators remains constant.

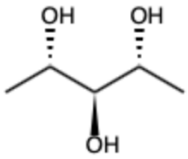
Double bonds are tested for configuration by an analogous procedure (steps 2 and 3 above) comparing the two ligand branches at each end. If the two branches are different at both ends, the / and \ stereo-designators are introduced in the SMILES at the appropriate place (see A1 below for the rules).

Pseudochirality

Pseudochirality centers (labeled *r* or *s* within the CIP system) are specified with @ and @@ by MOSFECCS because the reflection symmetry present in molecules with pseudochirality centers will lead to "enantiotopic" in the comparison of a pair of branches in step 2 above. The two ways of drawing the same molecule give identical SMILES codes. Actually, different SMILES are generated depending on whether the search starts on the left-side end or on the right-side end of the molecule but lexicographic sorting selects only one of them and always the same one. On the other hand, the two diastereoisomers in the example below give different SMILES.

Example:

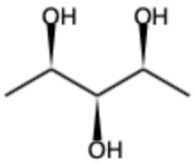


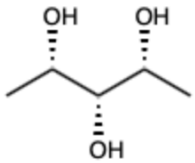


Two different SMILES are generated, depending on where the search starts.

C[C@H](O)[C@H](O)[C@@H](C)O C[C@H](O)[C@H](O)[C@@H](C)O SMILES selected lexicographically

C[C@@H](O)[C@@H](O)[C@H](C)O C[C@@H](O)[C@@H](O)[C@H](C)O





C[C@@H](O)[C@H](O)[C@H](C)O C[C@@H](O)[C@H](O)[C@H](C)O SMILES selected lexicographically

C[C@H](O)[C@@H](O)[C@@H](C)O C[C@H](O)[C@@H](O)[C@@H](C)O

Cis-trans substituents at rings
The SMILES convention does not include a special stereo-designator for cis/trans relations between substituents at rings. Therefore, relative configurations at rings have to be designated with @ and @@ even if the molecule as such is achiral.
MOSFECCS uses the @/@@ designators whenever a ring has two substituents attached to the ring through a wedge bond (stereo-up or stereo-down).

Achiral cyclic molecules with two or more substituents will nevertheless give the same SMILES for both ways of drawing. While starting the search from either methyl group will generate two different SMILES (C[C@H]1CC[C@H]1C and C[C@@H]1CC[C@@H]1C), the lexicographic sorting will always select one of the two.

Example:

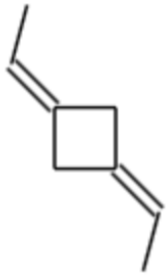
C[C@H]1CC[C@H]1C C[C@H]1CC[C@H]1C

Cumulenes
Cumulenes with an even number of cumulated double bonds are treated like tetrahedral stereogenic centers and assigned @ or @@ at the central cumulene atom in case of chirality.
Cumulenes with an odd number of cumulated double bonds are treated like double bonds and assigned / or \ if they have E/Z configurations.

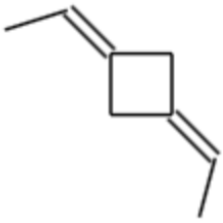
Configurationally linked exocyclic double bonds at rings
If two double bonds have E/Z configurations because a rigid spacer like a symmetrically substituted ring connects them, MOSFECCS takes this into account.

Example: quinodimethanes

C\C=C1\C=C\C(=C\C)C=C1 C\C=C1\C=C\C(=C/C)C=C1



C\C=C1\C\C(C1)=C\C



C\C=C1\C\C(C1)=C/C

4. "Aromaticity" and resonance structures

All published SMILES-conventions assign lower case letters to atoms in "aromatic" systems (benzene: c1ccccc1). This requires an algorithm to detect "aromatic systems"
although the concept of aromaticity is not really defined in terms of classical organic structure theory (Kekule).
The advantage of trying to use "aromaticity" when calculating SMILES codes is that database searches via SMILES will identify an aromatic molecule regardless of the drawn resonance structure thus making the search faster and simpler.
In the context of an e-learning environment, however, an editor telling the student, whether a molecule is aromatic or not will prevent the teacher from asking questions like "which of the following molecules are aromatic?"
For the same reasons, any algorithm that tries to "unify" different resonance structures of a given molecule (or even to unify tautomeric forms, which are different molecules) is didactically undesirable because the writing and use of correct resonance structures or understanding tautomeric isomerism are teaching aims of introductory organic chemistry courses.

For this reason, MOSFECCS deviates from the original SMILES convention for aromatic systems as follows:

No attempt is made to guess whether an unsaturated cyclic system is aromatic or not (e.g. via the Hückel rule).

Lower case element symbols for atoms in EMFU rings

(**E**ven **M**embered **F**ully **U**nsaturated):
To avoid having to draw many trivial resonance structures (and consider all of them in preprogrammed correct answers) for rings such as phenyl-substituents, atoms in *even-membered rings* with the maximal number of noncumulative double bonds are designated by small letters. This concept has nothing to do with aromaticity (benzene, cylco-octatetraene and cyclobutadiene are all *even-membered* and *fully unsaturated* but only benzene is *aromatic*). Odd membered fully unsaturated rings like cyclopentadiene do not have resonance structures differing from each other by shifted double bonds in the ring. For this reason, they are not assigned small letters in the SMILES string.

MOSFECCS treats different resonance structures of a molecule as different and calculates different SMILES codes for them–with the exception of trivial resonance structures in EMFU rings and equivalent resonance structures related by reflectional or rotational symmetry (e.g. the two resonance structures of allyl anion or of a carboxylate group).

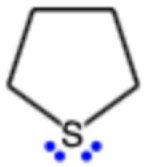
5. MOSFECCS-specific extensions of the SMILES convention for Lewis-structures, curved electron-shift arrows, and reaction schemes

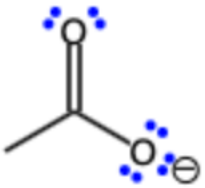
Sections of the SMILES string encoding these extensions are always placed at the end of the SMILES (after the standard part). For multismiles (several molecules), the whole standard SMILES is written out before the extensions.

5.1 Lewis Structures: encoding lone pairs

The position of explicitly drawn lone pairs is encoded in a section of the SMILES enclosed between two "!" characters. For each atom with explicitly drawn lone pair(s), the canonical number of the atom followed by ":" and the number of lone-pairs at this atom are given. ";" separates the entries for different atoms.

Example:



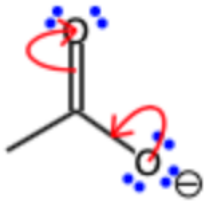


C1CCSC1.CC([O-])=O!4:2;8:3;9:2!

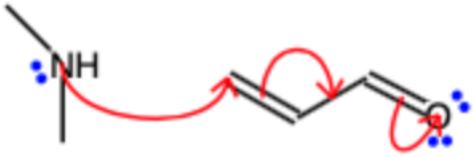
5.2. Mechanisms: curved arrows for "imminent" electron-pair or single electron shifts

Curved arrows start and end either at an atom or a bond. They are encoded in an extension section of the SMILES enclosed by "¿" characters.
Each curved arrow is specified by:
atom or bond where it starts ':' atom or bond where it ends ':' full ('f') or half ('h') ': the curvature ('l' for counterclockwise, 'r' for clockwise)
Atoms are characterized by their canonical number. Bonds are characterized by the canonical numbers of the two bound atoms (lower number first) joined by a dash. Entries for different curved arrows are separated by ";"

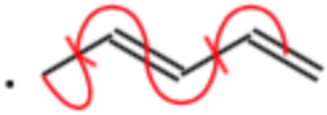
Examples:



CC([O-])=O¿2-4:4:f:l;3:2-3:f:r¿!3:3;4:2!



CNC.C=CC=O¿2:4:f:r;6-7:7:f:r;5-4:6-5:f:l¿!2:1;7:2!



[CH2]\C=C\C=C¿5-4:4-3:h:r;3-2:2-1:h:r;3-2:4-3:h:r;1:2-1:h:r¿

5.3. Reaction schemes: arrows for irreversible or reversible reactions and double headed arrows to connect resonance structures

MOSFECCS allows to draw reaction schemes with arrows (irreversible or equilibrium). In the drawing process, the groups of molecules constituting the reactants and products are defined for each arrow. Similarly, resonance structures can be connected with double-headed arrows.
Reaction arrows and resonance-structure-connectors are encoded in a section of the SMILES enclosed by '§' characters.

Each arrow is specified by the following fields:

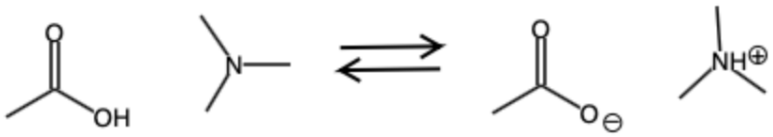
1. A list of molecule numbers of all molecules that are defined as "*reactants*" (irreversible arrows) or "left side of arrow" (for equilibrium or resonance structure arrows) for this arrow.
2. A list of molecule numbers of all molecules that are defined as "*products*" (irreversible arrows) or "right side of arrow" (for equilibrium or resonance structure arrows) for this arrow.

3. The type of the arrow: 1 for irreversible, 2 for equilibrium, 3 for resonance structures connector.
4. Text for annotation above/to the left of arrow
5. Text for annotation below/to the right of arrow

Numbering of molecules starts with 1 and follows the order in which they appear in the standard part of SMILES.
Resonance-structure-connectors can't have annotations.

Examples:

C1C=CC=C1.C=C.C1CC2CC1C=C2§1,2:3:1:Diels-Alder:Δ§



CC(O)=O.CN(C)C.C[NH+](C)C.CC([O-])=O§1,2:3,4:2::§

CN(C)\C=C\C(C)=O.C\C([O-])=C\C=[N+](C)C¿2:4-2:f:l;5-4:6-5:f:r;8-6:8:f:l¿!2:1;8:2;11:3!§1:2:3::§

The order of appearance of the different types of extensions in the SMILES is curved arrows, lone pairs, reaction arrows.

Appendix

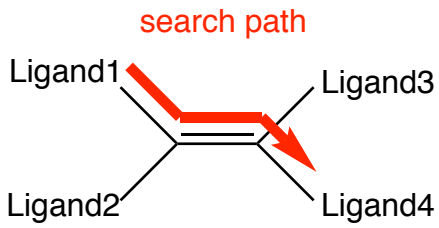
A1. Placement of slashes for specifying relative configuration at double bonds

The rules governing the placement of slashes are only described for simple examples in the Daylight-technical manual.

MOSFECCS uses rules that were chosen to lead to SMILES codes that are identical to Daylight-SMILES in most cases but there are also cases where the results are different.

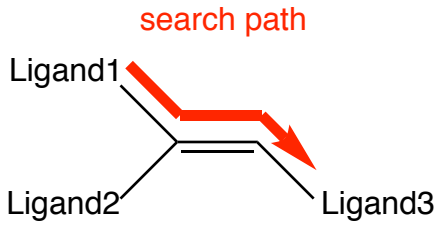
A double bond with (at most) four ligands receives slashes in the SMILES whenever the two ligands at both ends are different (for the comparison of ligands: see description above).

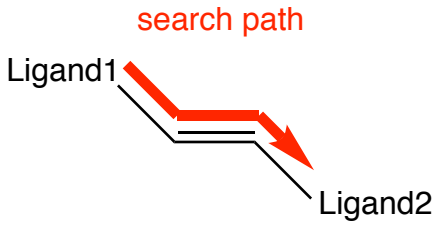
Isolated double bonds:



SMILES: Ligand1**\C**(Ligand2)**=C**(**/**Ligand3)Ligand4

In this simple case, the first slash is placed between Ligand1 and the first C of the double bond and the second slash is placed before ligand3 (inside the parenthesis). Two identical slashes (/ and / or \ and \) are used if Ligand1 and Ligand3 are *trans* to each other, two different slashes (/ and \ or \ and /) if they are *cis*.





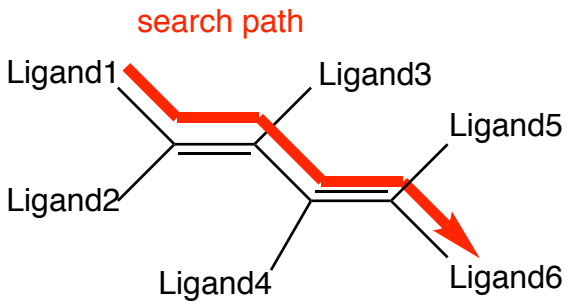
If there are only three ligands (one implicit H), the second slash is placed between the second double bond carbon and Ligand3 and reflects the relationship between Ligand1 and Ligand3.

SMILES: Ligand1**\C**(Ligand2)**=C**Ligand3

If there are only two ligands (two implicit H), the second slash is placed between the second double bond carbon and Ligand2 and reflects the relationship between Ligand1 and Ligand2.

SMILES: Ligand1**\C=C**Ligand2

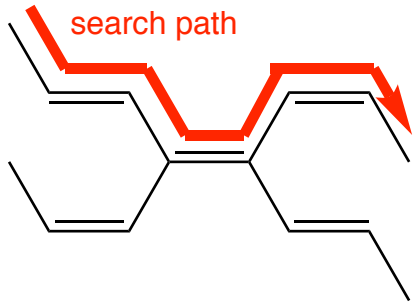
Linearly conjugated double bonds



SMILES: Ligand1****C(Ligand2)=C(Ligand3)****C(Ligand4)=C(**/**Ligand5)Ligand6

The second slash of the first double bond is at the same time the first slash of the second double bond. This "merged" slash is placed in front of the first carbon of the conjugated 2nd double bond and not before Ligand 3 in the parenthesis. This rule is applied for all double bonds in the conjugated system except the last one, where the second slash is placed in the sidechain as with an isolated double bond (here before Ligand5).

Cross conjugated double bonds



C\C=C\C(\C=C/C)=C(\C=C\C)/C=C\C

In this case, both ligands at each end of the cross conjugated double bond receive a slash. The slashes in the main chain (search path) reflect the cis/trans relationships in the same way as they would in a linearly conjugated system. In the sidechains (parenthesis), the first slashes are independent of those in the main chain and the second slashes reflect only the cis/trans arrangement inside this sidechain.

MOSFECCS uses the backslash (\) as the first slash except if conjugation fixes the type of the first slash because it is at the same time the second slash of a conjugated double bond earlier along the search path.

A2. Predefined "chemistry" constants used in SMILES calculation:

Element symbols that are allowed inside an EMFU ring:

```
const emfuAtoms = ["B","C", "N", "O", "P", "S", "As", "Se"];
// elements that can be part of an EMFU-ring
const emfuElesym = ["b","c", "n", "o", "p", "s", "as", "se"];
// symbol of emfu-Element in SMILES
```

Potentially stereocgenic centers (SC):

format of a pSC or pyrSC element:

"ElementSymbol : number of ligands : number of implicit hydrogens : charge"

tetrahedral:

```
const pSC=
["C:4:0:0","C:3:1:0","Si:4:0:0","Si:3:1:0","Ge:4:0:0","Ge:3:1:0","Sn:4:0:0","Sn:3:1:0","N:4:0:1",
"P:4:0:1","B:4:0:-1","B:3:1:-1","P:4:0:0","As:4:0:1","N:3:0:az","P:3:0:0","As:3:0:0","P:2:1:0","As:2:1:0",
"S:4:0:0","S:4:0:1","S:3:0:0","S:3:0:1","Se:4:0:0","Se:4:0:1","Se:3:0:0","Se:3:0:1"];
```

trigonal pyramidal:

```
const pyrSC =
["N:3:0:az","P:3:0:0","P:2:1:0","As:3:0:0","As:2:1:0","S:3:0:0","S:3:0:1","Se:3:0:0","Se:3:0:1"];
// H at pyramidal stereocenters must be drawn explicitly, az=aziridine N
```

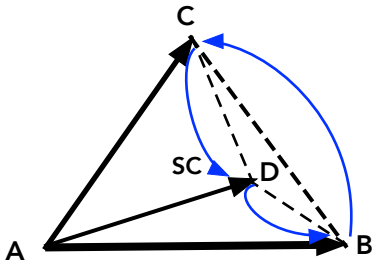
A3. Determination of the sense of chirality at tetrahedral stereogenic centers

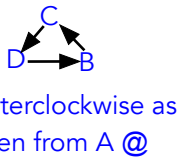
Atoms at the end of a wedge-bond (stereo-up or stereo-down) have a virtual z-coordinate of ±1 in the data structure whereas the x and y coordinates are real distances in 2D on the canvas.

After the algorithm for specifying the configuration at a stereogenic center (SC) has identified the four ligands and shown that they are all different, a tetrahedron is constructed from the coordinates of the four ligands by multiplying the virtual z-coordinates of end-of-wedge-atoms by half one standard bondlength. The bonds from the central atom (SC) to the four ligands are then normalized to the standard bondlength. Next, a check is made, whether the central atom is inside the tetrahedron spanned by the four ligands. If this is the case, the sign of the volume of the tetrahedron is determined by standard vector algebra as follows:

1. The ligand atom listed in the SMILES-string before the stereogenic center (SC) is labeled A, the other three ligands as B, C, D in the order of their

appearance in the SMILES after SC.





2. V = 1/6 [(ACxAD) • AB]
3. The sign of the volume V determines the sense of local configuration
   V > 0 for counterclockwise: @
   V < 0 for clockwise: @@