SMILES

Marvin imports and exports SMILES strings with the following specification rules:

Atoms

- Atoms are represented by their atomic symbols.
- Isotopic specifications are indicated by preceding the atomic symbol.
- Any atom but not hydrogen is represented with "*".
- '[Z]' symbols are imported as R-group attachment points. The attachment orders are ascending with atom indexes. Exporting R-group attachment points is possible in ChemAxon Extended SMILES (CXSMILES).
- Since radicals are not stored in SMILES format, they are calculated during SMILES import for atoms that tend to have radicals. It is done in the case when no implicit Hydrogens can be added because of the SMILES definition and the valence of the atom has to be corrected. E.g. for the SMILES string [Cl] no Hydrogens are allowed (because of the brackets), thus Chlorine is imported with a monovalent radical. No radicals are added to metals, for them a valence property is set if their valence differs from the usual. E.g. [AsH2] is imported with valence property 2 and two implicit Hydrogen atoms, since its usual valence would be different (it would be 3). No radicals are added to the following atoms:
  - Helium(He), Lithium(Li), Neon(Ne) and Sodium(Na)
  - All atoms above Chlorine(Cl) except Bromine(Br) and Iodine(I).
Radicals are stored in ChemAxon Extended SMILES (CXSMILES) format, for cases when the radical would be lost in SMILES, please, use CXSMILES.

Bonds

- Single, double, triple, and aromatic bonds are represented by the symbols -, =, #, and :, respectively.
- Single and aromatic bonds may be omitted.
- Branches are specified by enclosing them in parentheses. The implicit connection to a parenthesized expression (a branch) is to the left.
- Cyclic structures are represented by breaking one single (or aromatic) bond in each ring and the missing bond is denoted by connection placeholder numbers

Disconnected structures:
• Disconnected compounds are written as individual structures separated by a period.

Isomeric specification

• Configuration around double bonds is specified by "directional bonds": / and \.
• Configuration around tetrahedral centers may be indicated by a simplified chiral specification (parity) @ or @@.

Stereochemistry

• Parity is a general type of chirality specification based on the local chirality.

Cis-trans isomerism
The default stereoisomers in small rings (size < 8) are cis, which are not written explicitly.
See import option c to override this feature.

Reactions

• syntax: reactant(s)>agent(s)>product(s), where
  reactants = reactant1 . reactant2.....
  agents = agent1.agent2 . ....
  products = product1.product2 . ...

Agents are molecular structures that do not take part in the chemical reaction, but are added to the reaction equation for informative purpose only.

All of the above sections are optional. For example:

• a reaction with no agents: reactant(s)>>product(s)
• a reaction with no agents and no products (mainly used in reaction search): reactant(s)>>
• a reaction with no agents and no reactants (mainly used in reaction search): >> product(s)
• atom maps

Unique SMILES

The "unique" name can be sometimes misleading when dealing with compounds with stereo centres.

Daylight's SMILES specification (3.1.SMILES Specification Rules) defines generic, unique, isomeric and absolute SMILES as:

1. generic SMILES: representing a molecule (there can be many different representations)
2. unique SMILES: generated from generic SMILES by a certain algorithm [1].
3. **isomeric SMILES**: string with information about isotopism, configuration around double bonds and chirality
4. **absolute SMILES**: unique SMILES with isomeric information - in Marvin during graph canonicalization the isomeric information is also considered as an atom invariant

The name **canonical SMILES** is used for absolute or unique SMILES depending whether the string contains isomeric information or not (both strings are "canonicalized" where the atom /bond order is unambiguously).

**Marvin** generates always canonical SMILES with isomerism info if it is possible to find out from the input file. The molecule graph is always canonicalized using the algorithm in article [1] but it is not guaranteed to give absolute SMILES for all isomeric structures. The unique SMILES generation (option **u**) currently uses an approximation to make the SMILES string as **absolute** (unique for isomeric structures) as possible. In this case the form of any aromatic compound is aromatized before SMILES export. For correct exact (perfect) structure searching **MolSearch** and **JChemSearch** classes of **JChem Base** or the **jc_equals** SQL operator of the **JChem Cartridge** are suggested.

The initial ranks of atoms for the canonicalization are calculated using the following atom invariants:

1. number of connections
2. sum of non-H bond orders (single=1, double=2, triple=3, aromatic=1.5, any=0)
3. atomic number (list=110, any atom=112)
4. sign of charge: 0 for nonnegative, 1 for negative charge
5. formal charge
6. number of attached hydrogens
7. isotope mass number

See ref. [1] for details.

With option **u** it is possible to include chirality into graph invariants. This option must be used with care since for molecules with numerous chirality centres the canonicalization can be very CPU demanding [2]. SMILES canonicalization algorithm is not generic, it depends on the software package, so it is most useful to compare SMILES strings within a software package.

**Not supported SMILES features:**

- Branch specified if there is no atom to the left.
- General chiral specification: Allene like, Square-planar, Trigonal-bipyramidal, Octahedral.

**Reference**


[2] *A New Effective Algorithm for the Unambiguous Identification of the Stereochemical Characteristics of Compounds During Their Registration in Databases* ; T. Cieplak and
™: SMILES, SMARTS, and SMIRKS are trademarks of Daylight Chemical Information Systems.