SMILES Tutorial

What is SMILES?

SMILES (**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem) is a chemical notation that allows a user to represent a chemical structure in a way that can be used by the computer. **SMILES** is an easily learned and flexible notation. The **SMILES** notation requires that you learn a handful of rules. You do not need to worry about ambiguous representations because the software will automatically reorder your entry into a unique **SMILES** string when necessary.

SMILES was developed through funding from the U.S. Environmental Protection Agency, Mid-Continent Ecology Division-Duluth, (MED-Duluth) Duluth, MN to the Medicinal Chemistry Project at Pomona College, Claremont, CA and the Computer Sciences Corporation, Duluth, MN. Several publications discuss **SMILES** in more detail, including Anderson et al. 1987, Weininger 1988, Weininger et al. 1989, and Hunter et al., 1987.

SMILES has five basic syntax rules which must be observed. If basic rules of chemistry are not followed in **SMILES** entry, the system will warn the user and ask that the structure be edited or reentered. For example, if the user places too many bonds on an atom, a **SMILES** warning will appear that the structure is impossible. The rules are described below and some examples are provided. The rules below allow for the representation of a two-dimensional structure of a chemical. For the ASTER system, a two-dimensional depiction is adequate. Other rules are available for chemicals that are structural isomers, but will not be discussed in this basic tutorial.

Rule One: Atoms and Bonds

SMILES supports all elements in the periodic table. An atom is represented using its respective atomic symbol. Upper case letters refer to non-aromatic atoms; lower case letters refer to aromatic atoms. If the atomic symbol has more than one letter the second letter must be lower case.

Bonds are denoted as shown below:

-	Single bond
=	Double bond
#	Triple bond
*	Aromatic bond
.	Disconnected structures

Single bonds are the default and therefore need not be entered. For example, 'CC' would mean that there is a non-aromatic carbon attached to another non-aromatic carbon by a single bond, and the computer would identify the structure as the chemical ethane. It is also assumed that the bond between two lower case atom symbols is aromatic. A blank terminates the **SMILES** string.

Rule Two: Simple Chains

By combining atomic symbols and bond symbols simple chain structures can be represented. The structures that are entered using **SMILES** are hydrogen-suppressed, that is to say that the molecules are represented without hydrogens. The **SMILES** software understands the number of possible connections that an atom can have. If enough bonds are not identified by the user through **SMILES** notation, the system will automatically assume that the other connections are satisfied by hydrogen bonds.

Some examples:

CC	CH_3CH_3	Ethane
C=C	CH_2CH_2	Ethene
CBr	CH_3Br	Bromomethane
C#N	$C{\equiv}N$	Hydrocyanic acid
Na.Cl	NaCl	Sodium chloride

The user can explicitly identify the hydrogen bonds, but if one hydrogen bond is identified in the string, the **SMILES** interpreter will assume that the user has identified **all** hydrogens for that molecule.

HC(H)=C(H)(H)	Ethene

Because **SMILES** allows entry of all elements in the periodic table, and also utilizes hydrogen suppression, the user should be aware of chemicals with two letters that could be misinterpreted by the computer. For example, 'Sc' could be interpreted as a sulfur atom connected to an aromatic carbon by a single bond, or it could be the symbol for scandium. The **SMILES** interpreter gives priority to the interpretation of a single bond connecting a sulfur atom and an aromatic carbon. To identify scandium the user should enter [Sc].

Rule Three: Branches

A branch from a chain is specified by placing the **SMILES** symbol(s) for the branch between parenthesis. The string in parentheses is placed directly after the symbol for the atom to which it is connected. If it is connected by a double or triple bond, the bond symbol immediately follows the left parenthesis. Some examples:

CC(O)C	2-Propanol
CC(=O)C	2-Propanone
CC(CC)C	2-Methylbutane
CC(C)CC(=O)	2-Methylbutanal
c1c(N(=O)=O)cccc1	Nitrobenzene
CC(C)(C)CC	2,2-Dimethylbutane

Rule Four: Rings

SMILES allows a user to identify ring structures by using numbers to identify the opening and closing ring atom. For example, in C1CCCCC1, the first carbon has a number '1' which connects by a single bond with the last carbon which also has a number '1'. The resulting structure is cyclohexane. Chemicals that have multiple rings may be identified by using different numbers for each ring. If a double, single, or aromatic bond is used for the ring closure, the bond symbol is placed before the ring closure number. Some examples:

	C=1CCCCC1	Cyclohexene
	C*1*C*C*C*C*C1	
or	c1ccccc1	Benzene
	C1OC1CC	Ethyloxirane
	c1cc2ccccc2cc1	Naphthalene

Rule Five: Charged Atoms

Charges on an atom can be used to override the knowledge regarding valence that is built into **SMILES** software. The format for identifying a charged atom consists of the atom followed by brackets which enclose the charge on the atom. The number of charges may be explicitly stated ({-1}) or not ({-}). For example:

	CCC(=O)O{-1}	Ionized form of propanoic acid
or	CCC(=O)O{-}	
	c1ccccn{+1}1CC(=O)O	1-Carboxylmethyl pyridinium