United States Environmental Protection Agency | Environmental Research Laboratory Duluth, MN 55804

Research and Development | EPA/600/M-87/021 | August 1987

EPA ENVIRONMENTAL RESEARCH BRIEF

SMILES: A Line Notation and Computerized Interpreter for Chemical Structures

Eric Anderson, Gilman D. Veith, and David Weininger

Introduction

As the use of structure-activity relationships matures in the search for cost-effective molecular design and chemical safety evaluation, interaction with advanced computational methods on small computers is unavoidable. Methods for specifying chemical structures through an interactive construction vary widely and include specifying line notations, atom and bond list matrices, and graphical building blocks of substructures. Line notations such as Wiswesser Line Notation (WLN) are rapid but require extensive knowledge and experience by the user (Smith, 1968; Granito et al., 1972, Elkins *et al.* 1974), making WLN of limited value to the non-chemist computer user The input of atom and bond lists requires a minimum of knowledge by the user but is tedious and slow (Kaufmann, 1981). Moreover, lists are not efficient representations of chemical structures in computer memory or storage devices if large sets of structures are desired. Graphically building structures from menus of substructures (Kao *et al.*, 1985) is user-friendly but requires more software overhead and hardware costs. This paper presents a convention for chemical structure notation which has the advantages of line notation and minimizes the chemical knowledge of the user by programming many rules of chemistry into the line notation interpreter.

SMILES notation (**S**implified **M**olecular **I**dentification and **L**ine **E**ntry **S**ystem) was developed by the Environmental Research Laboratory-Duluth QSAR Research Program to facilitate storage, retrieval, and modeling of chemical structures and chemical information. This notation provides a flexible and unambiguous method for specifying the topological structure of molecules, and interfaces with additional software to specify the geometry of molecules. SMILES notation reduces the difficulty of translating structure into appropriate notation for humans and software by focusing on common chemical conventions and only five simple "rules"

SMILES Notation

To prepare SMILES notation, it is generally best to draw chemical structures on paper, although experienced users often write SMILES directly from the structure envisioned in their minds. Chemical structures are most often hydrogen-suppressed because the location of hydrogens is implicit for normal valences of atoms. Hydrogen atom can be explicitly designated in heterocyclic systems to avoid ambiguities. Each atom in the structure is represented by its atomic symbol. Atomic symbols having two characters, such as lead, are designated with the first character "upper case" and the second character "lower case" (i.e., Pb). The bond symbols are designated by special characters - explicit single bonds are designated by hyphen (-), double bonds by equal sign (=), and triple bonds by (#). The bond between two atoms is represented by placing its symbol between the atomic symbols for the atoms which are connected by the bond. If a bond is not specified, it is interpreted as a single bond. Examples are.

Chloroethane	CH_3CH_2Cl	C-C-Cl or CCCl
Acetaldehyde	$CH_3CH=O$	CC=O
1-Propyne	$CH\equiv CCH_3$	C#CC

The SMILES interpreter reads the SMILES string from the left to right and identifies atoms with two-character symbols first. There is then, no ambiguity in the notation for chloroethane with respect to the character "C" in the chlorine atom.

To represent branched structures where atoms have more than two atoms connected to them, the additional connections, or branches, are enclosed in parentheses. The left parenthesis is interpreted to mean that all atoms in the

string until the corresponding right parenthesis are connected to the preceding atom:

2-Methylbutane	CC(C)CC
Di-n-butylphosphorate	O=P(OCCCC)OCCCC
or	CCCCOP(=O)OCCCC

Note that atoms such as oxygen which are double-bonded to the central atom are designated by placing the double bond just inside the opening parenthesis

There are no limits to the number of branches a structure can have in SMILES notation, although associated software in the SMILES interpreter will detect disallowed valence states (too many connections) for the atoms. Conceptually, there are no limits to the number of branches that can be designated within other branches because the pairs of parentheses are interpreted in logical order This permits very complex topological structures to be written in a simple linear string of characters. The SMILES interpreter does, however, include some practical software limitations For example, there is usually a limit to the number of atoms allowed in a structure.

The only remaining topological feature of chemical structures is that of rings or cycles The simple SMILES rule for cyclic structures is that one bond broken in each ring will result in a structure which can be expressed in a linear string. To identify the "broken" bond in each ring, the two atoms connected by the bond are each labeled with a digit, termed the ring-closure pair The digit is placed immediately following the two atoms connected by the "broken" bond and the SMILES interpreter restablishes the bond in the internal connection matrix for the structure For example:

Cyclohexane	ClCCCCCl
Benzene	Cl=CC=CC=Cl

To avoid having to draw Kekule structures and designate conjugated double bonds in aromatic structures, SMILES notation includes the convention of designating atoms in aromatic rings with lower case atom symbols

Benzene	clcccccl
Naphthalene	clcc2ccccc2ccl
4-chlorobenzoic acid	O=C(O)clccc(C)ccl

More complicated structures generally require that the structure be drawn on paper first to facilitate "bookkeeping" as illustrated in Figure 1. It is obvious that the SMILES notation can begin at any atom in a structure and still be valid. We have developed software which rapidly plots the structure entered by the user to provide a visual verification of the structure Moreover, we have developed a conical ordering algorithm which translates all variations of possible SMILES notation for a given structure into a "unique" SMILES notation for that structure. This algorithm and its use in storage and retrieval of chemical information will be the subject of a subsequent paper.

The convention of using lower case symbols for aromatic atoms introduces the possibility of ambiguous SMILES notation in the case where an aromatic atom has a double-character symbol. For example, "Sn" designates the atom tin; however, it could be interpreted as an aliphatic sulfur singly bonded to an aromatic nitrogen Also, if tin were to be designated as aromatic, the lower case "sn" could be interpreted as an aromatic sulfur connected to an aromatic nitrogen. Wa have found these ambiguities to occur infrequently and can be omitted by designating double-character atoms as aromatic by placing the exclamation point (!) as a suffix for aromatic atoms immediately following the atom (e.g , Sn! designates aromatic tin). In general, aliphatic atoms connected to aromatic rings most frequently would be designated as a branch using parentheses. Finally, if a user wished to designate an aliphatic sulfur connected to an aromatic carbon, the use of the explicit single bond "S-c" would remove ambiguity and prevent the notation from being interpreted as scandium.

A summary of SMILES notation is as follows:

1) Represent atoms with their atomic symbols using hydrogen suppression in most cases.
2) Represent bonds between atoms using "-" for single bonds, "=" for double bonds, and "#" for triple bonds. Single bonds are implicit if not designated
3) Enclose branches from a central atom in parentheses.
4) Linearize cyclic structures by removing one bond in each ring and designating the atoms as ring closure pairs with a digit immediately after the atoms to be reconnected
5) Aromatic ring atoms can be represented using lower case symbols for single character atoms and by appending an "!" suffix to double character atoms.

SMILES Interpreter

SMILES notation is logically interpreted by a syntax interpreter which parses the character string from left to right and assumes that.

SMILES = piece ([bond] piece) space

Piece = atom (digression)

Digression = [bond] label |
left_paren <[bond] piece> right_paren

Label = "1" | "2" | "3" | "4" | "5" | "6" | "7" | "8" | "9"

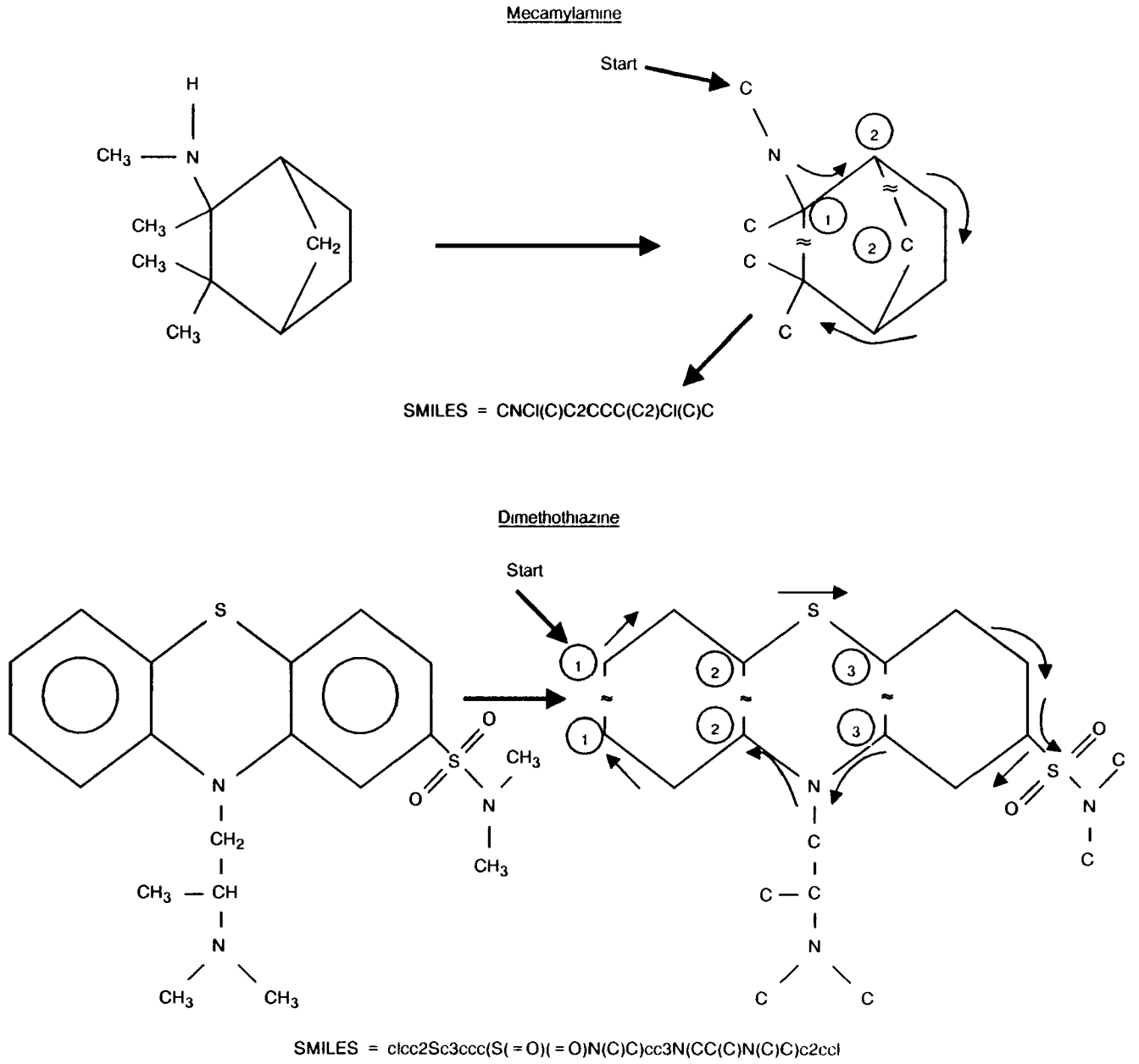
Bond = bond_symbol [bond_qualifiers]

Atom = atomic_symbol [atom_qualifiers]

In this notation, (.) enclose items repeated zero or more times, [.] enclose an optional item, | separates alternatives, and <..> enclose items repeated one or more times. From this definition of SMILES, it can be seen that a space is used to designate the end of the SMILES string. The labels are used to close rings Consequently, a given label must appear an even number of times in pairs.

A syntax diagram for the purpose of software development and explanation is presented in Figure 2. Although recursive implementations of the syntax diagram are certainly possible, we have constructed a non-recursive implementation of SMILES interpretation in FORTRAN 77. The first subroutine in the software is designed to identify atoms by matching the characters to the atomic symbols. This routine also records whether the atom is aliphatic or aromatic as described above. The second subroutine connects two atoms In addition to identifying the explicit bond types, the routine must identify implicit bonds Implicit bonds within a ring or fused ring system are identified as

Figure 1. Writing SMILES for branched, cyclic structures.



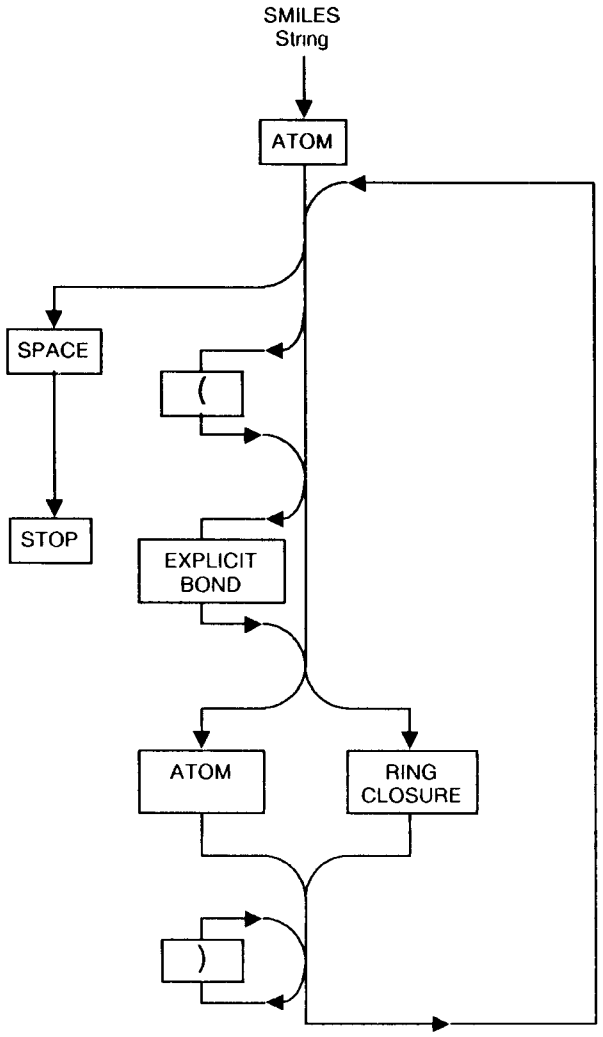
aromatic if both atoms connected by the bond are designated aromatic.

After processing the first atom, a WHILE loop will check for the presence of a terminating space. As long as the end has not been reached, each pass through the loop will process either one atom (attaching it to the preceding atom), or one end of a ring closure pair When two matching ends have been located the two atoms can be reconnected.

Within the loop, checks are made in the sequence indicated by the diagram for the optional presence of parentheses (indicating branching) or explicit bonds As opening parentheses are encountered, the last atom encountered is remembered on a stack (last in, first out). This allows the algorithm to return to the atom and proceed along a different branch after it has completed this current branch Later as the matching closing parenthesis is encountered, the atom is taken back off the stack and its status as the "current" or "last" atom is restored As each new atom is encountered, it is attached to the former "last" atom and in turn becomes the new "last" atom.

The routine also makes subsequent checks to make certain that the stack is empty (all parentheses were matched) and that all ring closure digits used were eventually matched.

Figure 2. Syntax diagram for SMILES notation.



Also, each aromatic atom should have at least two aromatic bonds associated with it. These conditions support proper syntax checking.

It is also important to perform various checks on the chemical meaningfulness of the structure described These include identifying improper valence states for a given atom The number of hydrogen atoms associated with each atom is computed as part of the valence checking software. The placement of hydrogens in heterocycles can be made explicit to avoid interpretation difficulties. In delocalized systems such as the nitro group, we have adopted the SMILES convention of double bonds to both oxygen atoms, e g , RN(=O)=O, to prevent the hydrogen connecting routine from adding a hydrogen on the oxygens These specialized bonding systems can be readily detected and modeled as delocalized bonds as the need arises

Semantic checks on the chemistry of the SMILES string are best kept in routines separate from the interpreter so they can be called sometime after interpreting a SMILES string Additions and refinements to such a routine can then proceed independently, without affecting the SMILES syntax interpreter. In the version used at ERL-D, the valence of carbon is 4, oxygen is 2, and halogens are all 1. Nitrogen can have valences of 3, 4, or 5 as well as a variety of special states. For example, an aliphatic nitrogen with 4 single bonds is considered charged as designated below. Nitrogens containing two double bonds decrease the neighboring atom hydrogen count by one. In addition to these common valence checks, simple checks to be certain aromatic atoms are located in a ring structure or that all atoms in an aromatic ring are so designated are made.

There are numerous other conventions which are being incorporated into the SMILES interpreter to designate other features of chemical structure. Simple conventions such as {+} or {-} locate explicit ionic charges on the preceding atom The use of braces have been a convenient method of designating specialized delocalized and tautomeric bonding in substructures and for special valences of inorganic structures in SMILES by expanding the list of qualifiers on atoms and bonds in the interpreter. The SMILES interpreter described herein is obviously capable only of the topology of structure and, in this simple form, cannot be used to designate conformation or other three-dimensional attributes of structure The addition of geometry to the SMILES conventions is beyond the scope of this paper and will be discussed separately Moreover, topological structures are adequate for models of many chemical properties and as input to conformational analysis programs. The primary advantages of SMILES is the simplicity of the conventions and the fact that the software can be implemented on almost any size computer using BASIC, FORTRAN, Pascal or C compilers The FORTRAN version of the SMILES interpreter is available upon written request to the Environmental Research Laboratory-Duluth

References

Smith, E.G The Wiswesser Line-Formula Chemical Notation; McGraw-Hill, New York, N.Y., 1968, pp 187.

Granito, C.E , Roberts, S ; Gilbson, G W J. Chem. Doc 12, 190-196 (1972).

Elkins, D ; Leo, A.; Hansch C , J. Chem Doc 14, 65-69 (1974)

Kaufmann, J J , Int J Quant. Chem 8, 419-439 (1981).

Kao, J , Eyerman, C , Walt, L , Maher, R., Leister, D , J. Chem. Inf. Comput. Sci. 25(4), 400-409 (1985).

Authors

Eric Anderson is with the Computer Sciences Corporation, Falls Church, VA

David Weininger is with the Medchem Project, Pomona College, Claremont, CA.

Address correspondence to:

Gilman D. Veith
U S. EPA
Environmental Research Laboratory
6201 Congdon Blvd.
Duluth, MN 55804