Modular Chemical Descriptor Language (MCDL): Composition, Connectivity, and Supplementary Modules

Andrei A. Gakh* and Michael N. Burnett

Chemical and Analytical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831

Received July 22, 2000

The Modular Chemical Descriptor Language (MCDL) was developed to address the need for linear representation of structural and other chemical information for chemical databases, E-journals, and the Internet. The current paper describes in detail two major modules of the language: the composition and connectivity modules, which provide a representation of chemical structure. These modules are created using simple hierarchical principles based on ASCII codes and are unique except for stereoisomers and a few special cases (e.g., valence isomers, knot-type compounds). The MCDL also provides for additional information (such as atom coordinates, bond orders, spectra, and physical-chemical characteristics) to be included as a set of supplementary modules.

INTRODUCTION

Linear descriptors of molecular structure play a significant role throughout the printed chemical literature and have gained even greater importance in the cataloging of chemical information in electronic form, including Internet applications. Several approaches exist for representing molecular structures as strings of characters. The most obvious of these are the IUPAC[1] and CAS[2] nomenclature systems. Unfortunately, both of these require numerous and often complex rules to create a compound name, and these "systematic" names are not always unique, which complicates name searching. To illustrate the difficulty in naming compounds, the structure shown in Figure 1 was submitted to three different computer naming programs, and the result was three different names, which are shown below the structure.

Other linear descriptors of molecular structure include Wiswesser Line Notation (WLN),[3] SYBYL line notation (SLN),[4] and others.[5−8] Perhaps the most advanced example of molecular descriptors is the SMILES code (Daylight Chemical Information Systems).[9] SMILES notation is based on a "valence scheme" approach. It provides a fairly clear representation of the molecular composition and connectivity, but a molecule may often be depicted by more than one SMILES string. For example, both CNC(NC1(COC(C)(C)OC1)CN(C)C)═O and CC(C)(OC1)OCC1(CN(C)C)NC(═O)-NC are valid SMILES representations of the structure shown in Figure 1. Daylight has developed an algorithm for creating a unique SMILES for a structure using proprietary computer software.[10] The code can also be extended by adding sterochemical descriptors.

In an attempt to design simple text-based chemical information descriptors, primarily for E-journals, Internet, and database purposes, we developed the Modular Chemical Descriptor Language (MCDL). Advantages of this new approach include its ability to create unique descriptors, its modularity, and the flexibility to extend the language to include any compound information that is needed for a particular task. As an example, the unique MCDL descriptor (composition and connectivity modules) for the structure in Figure 1 is the following: 2C;3CHH;5CHHH;CO;N;2NH;2O[3,4,5,13;6,7,15,16;12;15;16;;;12;12;14;13,14]. This descriptor will be defined below.

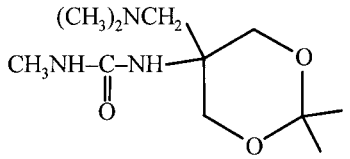

Figure 1. *N*-{5-[(Dimethylamino)methyl]-2,2-dimethyl-1,3-dioxan-5-yl}-*N*′-methylurea (ACD/IUPAC NAME). 1-(5-Dimethylaminomethyl-2,2-dimethyl-[1,3]dioxan-5-yl)-3-methyl-urea (AUTONOM). *N*-{5-[(Dimethylamino)methyl]-2,2-dimethyl(1,3-dioxan-5-yl)}-(methylamino)carboxamide (NOMENCLATOR).

METHODOLOGY

The first part of any MCDL linear descriptor includes two modules that uniquely describe the basic molecular structure. These are the composition and connectivity modules. Supplementary information about the compound (including information such as ID number, bond order information, atom coordinates, physical properties, spectra, etc.), which may or may not be unique, may follow the unique portion of the descriptor.

Composition Module. In this notation, the composition module provides a set of structural fragments, each consisting of a nonterminal atom and all terminal atoms attached to it. A structural fragment is represented in the linear descriptor by listing the nonterminal atom symbol followed by the terminal atom symbols in ASCII order (dictionary order in English). The structural fragments are listed in ASCII order and separated by semicolons. Multiple fragments of the same type are listed once with a prefix number. Thus, the four

* Corresponding author e-mail: gakhaa@ornl.gov.

10.1021/ci000108y CCC: $20.00
Published on Web 10/30/2001

structural fragments in 2-bromobutane are CHHH, CBrH, CHH, and CHHH, which become CBrH;CHH;2CHHH in the linear descriptor. The composition module of 3-methylphenol is 2C;4CHH;CHHH;OH.

Connectivity Module. For the connectivity module of an MCDL linear descriptor to be unique, the structural fragments must be numbered uniquely so that only one description of the connectivity is possible. The MCDL numbering scheme is based on the ASCII priorities of the structural fragments. The approach was developed with consideration of existing "canonical numbering" methodologies and combines both the effectiveness of hierarchical schemes with the thoroughness of "smallest binary code" approaches.[11−20]

For example, there are three structural fragment types in 2-bromobutane: CHHH, CHH, and CBrH. The ASCII (dictionary) order of these is CBrH, CHH, and CHHH. Thus, CBrH has the highest priority, CHH is next, and CHHH has the lowest priority. Since there is only a single CBrH fragment and it has the highest priority, CBrH is fragment #1 in the final numbering scheme. With similar reasoning, CHH is fragment #2. The two CHHH fragments are #3 and #4, which can be distinguished only after consideration of their immediate neighbors. One CHHH is connected to fragment #1, while the other is connected to fragment #2. The CHHH connected to the higher priority fragment #1 has a higher priority than the one connected to the lower priority fragment #2. Thus, CHHH connected to CBrH is fragment #3, leaving fragment #4 to the CHHH connected to CHH.

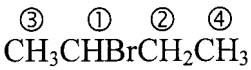

Figure 2.

Once all fragments have unique numbers, the connectivity module of the MCDL linear descriptor can be built. The process begins by listing the connections to each fragment, with multiple connections to a single fragment separated by commas.

fragment	1	2	3	4
connections	2, 3	1, 4	1	2

To eliminate redundant information, connection numbers smaller than the fragment number are removed.

fragment	1	2	3	4
connections	2, 3	4		

Finally, the highest numbered fragments having no connections (if any) are removed. [Occasionally, the removal of redundant information in the second step of this process leaves fragments other than the highest numbered ones without any connections. These are not eliminated when forming the descriptor, and consecutive semicolons are included to indicate that a particular structural fragment needs no connections listed. (See the 3-methylphenol example below.)]

fragment	1	2
connections	2, 3	4

The connectivity module of the MCDL linear descriptor is formed by placing the connections within square brackets [ ]. The connectivities of different structural fragments are separated by semicolons. A semicolon between the last fragment and a square bracket] is not used. Thus, the connectivity module of 2-bromobutane is [2,3;4].

In most cases, it will not be possible to resolve all numbering ambiguities in one consideration of the priority values of immediate neighbors. Then, an iterative approach is taken. As each ambiguity is resolved, additional structural fragments receive their priority numbers, and the priorities of the attachments are reexamined. This is illustrated by a more complex example, 3-methylphenol with eight structural fragments 2C; 4CH; CHHH; and OH (Figure 3).

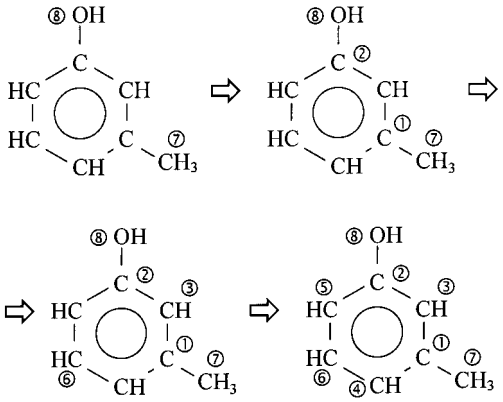

Figure 3.

The unique linear descriptor of 3-methylphenol is 2C; 4CH;CHHH;OH[3,4,7;3,5,8;;6;6]. (Note the absence of any connections for fragment #3 after removal of redundant information as described above).

There are some cases where priorities of the fragments cannot be resolved by this method. If these fragments are truly topologically equivalent, one is arbitrarily given the higher priority, which has the effect of resolving other ambiguities. However, if these fragments are topologically nonequivalent, all possibilities created by this random approach should be considered. Since this task is difficult to accomplish manually, application of LINDES software (see Supporting Information) is encouraged. The essence of the computer algorithm is outlined below.

The input to the program is the structure of the molecule in either FRAGCON format or a MOL file. In the former, the input file consists of a listing of the molecular structural fragments with an arbitrary numbering scheme followed by the connections of the fragments. Each connection must be included at least once. MOL files are converted into the FRAGCON format by LINDES prior to processing.

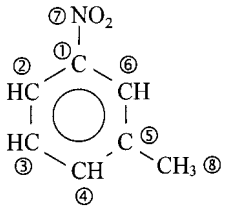

Figure 4.

Thus, the FRAGCON input file of 3-nitrotoluene, *arbitrarily numbered* as in Figure 4, might look like

1	C
2	CH
3	CH
4	CH
5	C
6	CH
7	CHHH
8	NOO
1	2
1	6
1	7
2	3
3	4
4	5
5	6
5	8

Initially, the program scans the structural fragments, determines the different types of fragments, and assigns each type a priority value (pn) based on the ASCII values of the fragment strings. The value p1 is the highest priority. Initial priorities for 3-nitrotoluene example are as follows: C = p1, CH = p2, CHHH = p3, NOO = p4. Each structural fragment then receives the priority of its type. (Legend: pn = priority n; fn = fragment with initial number n.)

fragment # (fn)	priority (pn)
f1	p1
f2	p2
f3	p2
f4	p2
f5	p1
f6	p2
f7	p3
f8	p4

The program then examines the priorities of the immediate neighbors for each fragment, starting with fragments with priority p1 (f1 and f5). The following connection table is constructed:

	p1	p2	p3	p4
f1	0	2	1	0
f5	0	2	0	1

This says f1 has 0 connections to fragments with priority p1, 2 connections to p2 fragments, 1 to p3, and 0 to p4, resulting in the connection string "0210". Similarly, f5 has 0 connections to p1 fragments, 2 connections to p2 fragments, 0 to p3, and 1 to p4, resulting in "0201".

The program then compares "0210" and "0201" as strings. The larger string receives the higher priority. Thus, f1 retains priority p1 and f5 becomes p2. All fragments with original priority p2 and greater have their priorities increased by one since one new priority has been determined. Now the fragments have the priority numbers shown below.

fragment # (fn)	priority (pn)
f1	p1
f2	p3
f3	p3
f4	p3
f5	p2
f6	p3
f7	p4
f8	p5

Looping through priorities continues. Since there is only one fragment with p2, processing continues with the four p3 fragments, and the following connection table is derived:

	p1	p2	p3	p4	p5
f2	1	0	1	0	0
f3	0	0	2	0	0
f4	0	1	1	0	0
f6	1	1	0	0	0

The connection string order is as follows: "11000" > "10100" > "01100" > "00200". Thus, f6 retains priority p3; f2 becomes p4; f4 becomes p5; and f3 becomes p6. Priorities p4 and p5 before this step are increased by three since three new priorities were determined.

fragment # (fn)	priority (pn)
f1	p1
f2	p4
f3	p6
f4	p5
f5	p2
f6	p3
f7	p7
f8	p8

Once the number of priority values equals the number of fragments, processing is complete. If, during the course of the processing, a distinction among multiple fragments with the same priority value cannot be made, the looping continues with the next higher priority value until all values have been considered. Then the loop starts over with p1 and looping continues until the process completes. Thus, the unique MCDL linear descriptor for 3-nitrotoluene is as follows: 2C;4CH;CHHH;NOO[3,4,7;3,5,8;;6;6].

If the process will not complete because two or more fragments are in indistinguishable environments, one is arbitrarily assigned the higher priority and the others in the group are assigned the same lower priority. This set is then put through the process described above. If this will go to completion, the connectivity portion of the linear descriptor is retained. Then another fragment in the indistinguishable group is assigned the higher priority, and the process runs again. The priority table from this run is compared to the first. If all priority tables determined in this manner are the same, the linear descriptor is returned as unique.

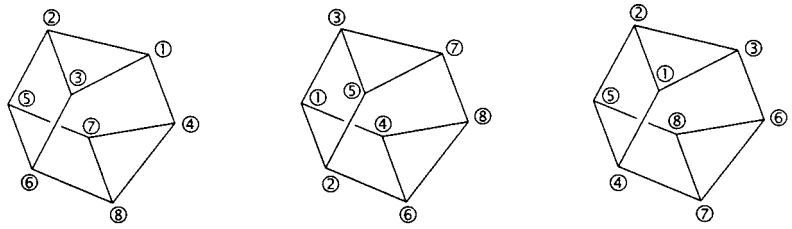

Table 1

	p1	p2	p3	p4	p5	p6	p7	p8
f1	0	1	1	1	0	0	0	0
f2	1	0	1	0	1	0	0	0
f3	1	1	0	0	0	1	0	0
f4	1	0	0	0	0	0	1	1
f5	0	1	0	0	0	1	1	0
f6	0	0	1	0	1	0	0	1
f7	0	0	0	1	1	0	0	1
f8	0	0	0	1	0	1	1	0

Table 2

	p1	p2	p3	p4	p5	p6	p7	p8
f1	0	1	1	1	0	0	0	0
f2	1	0	0	0	1	1	0	0
f3	1	0	0	0	1	0	1	0
f4	1	0	0	0	0	1	0	1
f5	0	1	1	0	0	0	1	0
f6	0	1	0	1	0	0	0	1
f7	0	0	1	0	1	0	0	1
f8	0	0	0	1	0	1	1	0

Table 3

	p1	p2	p3	p4	p5	p6	p7	p8
f1	0	1	1	1	0	0	0	0
f2	1	0	1	0	1	0	0	0
f3	1	1	0	0	0	1	0	0
f4	1	0	0	0	1	0	0	1
f5	0	1	0	1	0	0	1	0
f6	0	0	1	0	0	0	1	1
f7	0	0	0	0	1	1	0	1
f8	0	0	0	1	0	1	1	0

Figure 5.

When several different priority tables for a structure are produced using the rules described thus far, the software compares the final priority tables line-by-line until a distinction can be made. (In these final tables fragments are renumbered according to their priority values: f1 has priority p1, f2 − p2, and so on.)

For example, three possible priority tables can be created using the above-mentioned rules for pentacyclooctane, C_8H_8 (Figure 5).

The connection priority strings for fragment f1 are the same in all three tables (01110000). The strings for fragment f2 are the same in the first and third tables (10101000) and larger—*higher numeric value*—than the string in the second table (10001100). This eliminates the second table. The strings for fragment f3 are the same (11000100). Finally, for fragment f4, the third table has the larger string (10001001 versus 10000011). Therefore, the third table will be chosen to create the canonical connectivity module. The unique MCDL linear descriptor arising from it is 8CH[2,3,4;3,5;6;5,7;8;7,8;8].

Supplementary Modules. The supplementary part of the descriptor may contain both standardized (defined within the MCDL) and nonstandardized (undefined) data types. Each module consists of a two-character header followed by a colon and the data itself, all enclosed in braces {}. The order of supplementary data modules in the linear descriptor is arbitrary. Many of them are developed for adequate translation of MOL file information into MCDL format.

Cartesian Coordinates. The basic format for the atomic Cartesian coordinates module {CC} is shown in the example {CC:0,2.42Br;1.4,2.42C;2.1,1.21C;3.5,1.21C;1.4,0O} where the numeric coordinates precede the atomic symbol. This example shows only *x* and *y* coordinates. If *z* coordinates are included, the format for each atom is 2.1,1.21,0.42C, for example. The order of atoms in the CC module is arbitrary.

Bond Descriptors. The bond descriptor module {BB} provides the means to include bond order information that is absent in the unique portion of the linear descriptor. This module has the format shown in the example {BB:7a8;3d4;11d12;1s2;5t6} where the numbers refer to the atom sequence provided in the CC module, if present, or to the fragment sequence in the connectivity module, if the CC module is absent. The bond between the atoms (or fragments) may be indicated as *s* for single, *d* for double, *t* for triple, *a* for aromatic, and *u* for undefined. The order of individual bond descriptors in the BB module is arbitrary. For example, MCDL notation (with bond descriptor module) for 3-nitrotoluene is as follows: 2C;4CH;CHHH;NOO[3,4,7;3,5,8;;6;6]{BB:2a3;5a6;6a4;1a3;1s7;1a4;2s8}.

MOLfile Data. Several additional supplemental data types have been defined in MCDL to permit the linear descriptor to store the majority of information contained in an MDL Information Systems MOL file.[21] The three lines in the header block of a MOL file are stored in the three modules {Z1:}, {Z2:}, and {Z3:}. Data in the properties block are stored in {MM:} modules. For example, the MOL file line "M CHG 2 11 1 21 −1" becomes

{VE:?}	version number of MCDL; e.g. {VE:MCDL1.5}
{NA:n}	number of atoms in {CC}; e.g. {NA:28}
{NB:n}	number of bonds in {BB}; e.g. {NB:16}
{ZV:?}	indicates if *z* coordinates are included in {CC} yes: ?=Y or no: ?=N
{ID:?}	identification number; e.g. {ID:ORNL/0001023}
{RE:?}	reference; e.g. {RE:C. A.,1996,124,317152h}
{CN:?}	chemical name; e.g. {CN:CAS;propane,2-bromo-} or {CN:IUPAC;2-bromopropane}
{MF:?}	molecular formula, e.g. {MF:C2H6O}
{EA:?}	elemental analysis data, e.g. {EA:46.3C;2.3H;16.2N;C10H6F3N3S;46.7C;2.4H;16.3N}
{YD:?}	yield; e.g. {YD:66−68%}
{BP:?}	boiling point; e.g. {BP:41−45C;1000BAR}
{MP:?}	melting point; e.g. {MP:381−383K;METHANOL}
{SN:?}	NMR spectra; e.g. {SN:1H,CDCL3;4.28(D,*J* = 14 Hz,2H,CH2); 5.34(T,*J* = 14 Hz,1H,CH);PPM} or {SN:13C,D2O;94.2(CH);43.4(CH2);PPM} or {SN:19F,MeOH-D4;114.2(S,1F,CF);PPM}
{SM:?}	mass-spectra; e.g. {SM:EI;124(+M,100%);106(+M−H2O,14%);M/Z}
{SI:?}	IR-spectra; e.g. {SI:KBr;3400(NH);1540(NO2);CM-1}
{SU:?}	UV−Vis spectra, e.g. {SU:H2O;305(14500);350(8600);NM}
{DE:?}	definition of nonstandardized data, e.g. {DE:1A,EnolFormContent;9P,DielectricConstant}

{MM:CHG,2,11,1,21,−1}. The full MCDL descriptor generated from MOL file for 3-nitrotoluene is as follows: 2C;4CH;CHHH;NOO[3,4,7;3,5,8;;6;6]{Z1:}{Z2:}{Z3:}{NA:10}{NB:10}{ZV:N}{CC:0.02,1.24C;0.02,0.41C;0.73,0C;1.45,0.41C;1.44,1.24C;0.73,1.65C;0.72,2.48N;1.44,2.90O;0,2.89O;2.16,0.01C}{BB:2s3;5d6;6s1;1d2;6s7;3d4;7s8;7d9;4s5;4s10}{MM:CHG,2,7,1,8,−1}. Please note that the numbers in the BB and MM modules refer to the atom sequence provided in the CC module and not to the fragment sequence in the connectivity module.

Miscellaneous Data. Several additional standardized supplementary data types are defined in MCDL, mostly to accommodate MOL file transcription and CAS compatibility and to provide some additional information (name, ID number) as well as simple physical and physical-chemical characteristics (boiling point, melting point, spectra, and so on).

Nonstandardized Data. The number of standardized supplemental data types in MCDL is very small, but that does not limit the types of supplemental data that may be included in the linear descriptor. Until specific data types are standardized, they may be included with nontype-specific two-digit or digit-and-letter numeric headers: {01:...}, {02:...}; or {1A:...}, {2B:...}, etc.

RESULTS AND DISCUSSION

The major difference between the MCDL and SMILES is the reduced number of atom descriptors. For example, carbon atoms in SMILES are represented as **C** (an aliphatic carbon) or **c** (an aromatic carbon). The additional atom descriptors in SMILES allow reduction in the size of the descriptor and provide information regarding the atom environment. At the same time, the smaller set of atom descriptors in MCDL simplifies the interpretation of the overall descriptor. In addition, these extended atom descriptors in SMILES might be difficult to generate in some cases (cyclical polyene structures, which can be either aromatic or antiaromatic, depending on the ring size).

Another difference between SMILES and MCDL is the explicit placement of hydrogens. In SMILES code the hydrogens are removed, which leads to a more compact descriptor notation compared to MCDL. At the same time, explicit positioning of hydrogen atoms in an MCDL notation could provide a more distinct structure recognition in some complicated cases.

Finally, MCDL presents composition, connectivity, and bond information as separate modules. SMILES provides this information in a single module. There are no obvious advantages and disadvantages associated with these format differences.

Version 1.5 of the Modular Chemical Descriptor Language software provides translation only for covalently bonded organic molecules containing the elements C, H, O, N, P, S and halogens in which all atoms have their "normal" valences. This excludes salts, such as CH_3COO^- NH_4^+, and coordination compounds. On the other hand, species with formal charges (arising simply as a consequence of how the bonds are drawn in the structure) can be represented. Thus, a nitro group, often drawn as $O{=}N^+{-}O^-$, becomes the structural fragment NOO.

Any molecular descriptor based on connectivity tables (including MCDL) cannot be used to generate unequivocal representations of some topologically complex molecules such as molecular knots. For example, the two molecules presented below have the same composition and connectivity modules but different structures (Figure 6).

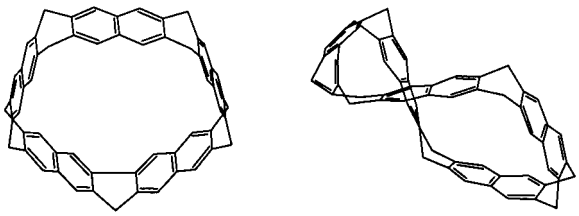

Figure 6.

In addition, valence isomers, such as the following two tetramethyl derivatives of cyclooctatetraene (Figure 7), cannot be uniquely represented using only composition and connectivity modules since these modules are identical. Fortunately, these molecular constructions are rare and can be easily resolved by adding a supplementary bond description module.

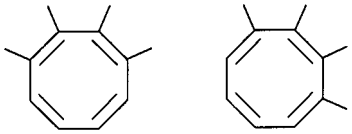

Figure 7.

Finally, the existing computer program LINDES (see Supporting Information, version 1.5) is limited to MOL files with no more than 99 atoms and 99 bonds. We plan to increase this number and to include a stereochemical module in the future versions of the MCDL software.

ACKNOWLEDGMENT

The research was sponsored by the IPP program, U.S. Department of Energy under contract DE-AC05-00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Battelle, LLC.

Supporting Information Available: The open source code of the C program "LINDES". This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) International Union of Pure and Applied Chemistry (IUPAC). *Nomenclature of Organic Chemistry*; Rigaudy, J., Klesney, S. P., Eds.; Pergamon Press: Oxford, U.K., 1979; http://www.iupac.org.
(2) Chemical Abstracts. *Index Guide 1992−1996*; Chemical Abstracts Service: Columbus, OH, 1997; http://www.cas.org. CAS names are designed to be unique.
(3) Wiswesser, W. J. *J. Chem. Doc.* **1968**, *8*, 146. Smith, E. G. *Wiswesser Line-Formula Chemical Notation Methods*; McGraw-Hill: New York, 1968.
(4) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL line notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 71−79.
(5) Bresmer, W. HOSE − a Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(6) Fujita, S.; Tanaka, N. XyM Notation for Electronic Communication of Organic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 903−914.
(7) Dietz, A. Yet Another Representation of Molecular-Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 787−802.
(8) Karabunarliev, S.; Ivanov, J.; Mekenyan, O. Coding of Chemical Structures Based on a Line Notation. *Computers Chem.* **1994**, *18*, 189−193.
(9) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36. SMILES Tutorial, Daylight

Chemical Information Systems, Inc. 1998; http://www.daylight.com/dayhtml/smiles/smiles-intro.html.
(10) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(11) Balaban, A. T.; Mekenyan, O.; Bonchev, D. Unique Description of Chemical Structures Based on Hierarchically Ordered Extended Connectivities (HOC Procedures). 1. Algorithms for Finding Graph Orbits and Canonical Numbering of Atoms. *J. Comput. Chem.* **1985**, *6*, 538−551.
(12) Randic, M. Unique Numbering of Atoms and Unique Codes for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 105−108.
(13) Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(14) Jochum, C.; Gasteiger, J. Canonical Numbering and Constitutional Symmetry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113−117.
(15) Shelley, C. A.; Munk, M. E. An approach to the assignment of canonical connection tables and topological symmetry perception. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 247−250.
(16) Herndon, W. C.; Leonard, J. E. Canonical numbering, stereochemical descriptors, and unique linear notaions for polyhedral clusters. *Inorg. Chem.* **1983**, *22*, 554−557.
(17) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.
(18) Hu, C.-Y.; Xu, L. A New Scheme for Assignment of a Canonical Connection Table. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 840−844.
(19) Agarwal, K. K.; Gelernter, H. L. A Computer-Oriented Linear Canonical Notational System for the Representation of Organic Structures with Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 463−479.
(20) Strokov, I. A Compact Code for Chemical Structure Storage and Retrieval. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 939−944.
(21) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

CI000108Y