Development of an Automated SMILES Pattern Matching Program To Facilitate the Prediction of Thermophysical Properties by Group Contribution Methods†

R. Jeremy Rowley, John L. Oscarson, Richard L. Rowley, and W. Vincent Wilding*

Department of Chemical Engineering, 350 CB, Brigham Young University, Provo, Utah 84602

Group-contribution methods produce estimates of thermophysical properties from correlations that include summative contributions for the functional groups that make up a molecule. Accurately dividing or parsing a molecule into its constituent functional groups, as intended by the authors of group-contribution methods, can be tedious and error prone. An automated parsing algorithm, which accepts SMILES formulas of compounds and accurately parses these into constituent functional groups, has been developed to facilitate the use, comparison, and development of group-contribution methods. The algorithm is described with particular attention to the difficulties inherent in parsing large, multiring compounds. The methodology will be useful to others who are developing, testing, and using prediction techniques.

Thermophysical Property Prediction

Accurate thermophysical property values are essential to chemical process design and development. Experimental measurements have always been the preferred source of these values, but because of the expense and difficulty of measurements and the nearly infinite combination of compounds and conditions encountered in process design, measurements have not kept pace with data needs. Engineers and scientists, therefore, must often rely on prediction techniques to obtain estimates of thermophysical properties.

Property estimation is usually based on the structure of the molecule, other known properties, or both. To perform most predictions, the molecule must be broken into constituent functional groups. This often requires some familiarity with the nuances of the particular prediction method and can be prone to calculation errors, particularly for large, multifunctional molecules. An automated prediction routine that, when fed a chemical structure, will output a reliable value for various prediction methods would greatly increase the efficiency of the prediction process (in speed and accuracy) and be of tremendous value to engineers and scientists. This paper discusses the development of an effective parsing algorithm. It is hoped that the methodology presented here will be helpful to others who are developing and using prediction packages. The algorithm is being incorporated into DIADEM (DIPPR Information and Data Evaluation Manager), a software interface to the DIPPR 801 database.

Structural Representation

The central component of an automated prediction program is the splitting or parsing algorithm that takes the specified structure and splits it into constituent chemical groups. Two key aspects of the parsing algorithm are apportioning a structural representation of a compound into chemically significant information and the building and recognition of chemical groups from that information. Though this structural representation can be either graphical or textual, the algorithm needs some sort of rule-based logical format to parse. In our work, this representation requirement is fulfilled through the use of SMILES formulas.[1,2] SMILES (Simplified Molecular Input Line Specification) is one of many ways to write a chemical structure in a linear format. Linear formulas have an advantage over graphical structures because they are easily read through computer code and parsed.

A SMILES formula depicts a valence model of a molecule and includes characters for each atom and for the structural arrangement within the molecule. These formulas can be used to describe single molecules but not mixtures. SMILES formulas are built on five basic rules:

(1) Atoms in SMILES formulas are represented by their atomic symbol, and compounds are created by combining atomic symbols in a string. Specification of valence hydrogens is not required. For example, CC represents ethane and CCC represents propane.

(2) The equal and pound signs are used to represent double and triple bonds between atoms. Single bonds can be represented with a dash or omitted. Aromatic bonds are represented by lower case letters. C═C is the SMILES representation of ethylene while C#C represents acetylene.

(3) Branches are represented using parentheses. CC(C)C represents *tert*-butane.

(4) Digits matching ring opening and closing positions specify rings. The rings can be reconnected by tracing a path from one digit to the other. Benzene is given by c1ccccc1.

(5) Conformational isomers are distinguished through the use of slashes. For example, Cl/C═C/Cl is *trans*-dichloroethylene and Cl/C═C\Cl is *cis*-dichloroethylene.

Formulas for most compounds can be created using these rules. In addition, most chemical drawing programs are able to convert graphical structure representations to SMILES formulas.[3]

† This contribution will be part of a special print edition containing papers presented at the Fourteenth Symposium on Thermophysical Properties, Boulder, CO, June 25−30, 2000.

10.1021/je0002336 CCC: $20.00
Published on Web 07/18/2001

SMILES Matching Algorithm

The algorithm presented here is designed to take a previously created SMILES formula and split it into Domalski−Hearing (DH) functional groups.[4] The DH method includes groups containing carbon, hydrogen, sulfur, nitrogen, oxygen, bromine, fluorine, chlorine, and iodine. Such an extensive element array provides a vast number of predictive groups, enabling the prediction of properties for a wide variety of compounds.

DH groups were chosen as the basis for the group recognition algorithm for several reasons. First, the algorithm includes an extensive number of functional groups, making it applicable to most molecules. Second, it is a second-order prediction method, accounting for next-to-nearest-neighbor group effects. Finally, there already exists a large number of prediction examples completed by the authors of the method against which the parsing algorithm can be tested. This test set includes over 500 different groups that include ring, isomer, and other corrections. For each of these groups, many different example calculations have been performed (well over 1500). Because the group recognition algorithm is based on a second-order method, the parser is extendable easily to first-order methods and third-order methods. Second-order groups constitute the minimal structure linkage set necessary to completely reconstruct the molecule’s structure.

The algorithm operates by making program indices out of the atomic symbols contained within the SMILES formulas. These indices are used to refer to the different atoms that make up the structure. The index structure is composed of two separate indices. The first, called the primary index, is a numerical representation of the atom’s location within the SMILES formula. All information about the elements of the SMILES formula can be retrieved by accessing the primary index. This information includes the atomic symbol, its position in the SMILES formula, adjacent atoms, its position within rings, and the conformation around double bonds. The second index points from an atom’s primary index to each of the neighboring atom’s primary indices. This facilitates identification of neighboring atoms. Although complicated by the fact that neighboring atoms may be spaced several characters away in the SMILES formula, because of inserted branch information, the SMILES formula can be completely reconstructed from the information stored in the secondary index.

As the algorithm analyzes the SMILES formula, it must link all bonded atoms to the primary atom under consideration. The patterns within a SMILES formula used to determine the indices for the bonded atoms can be classified into three different areas: direct bonds, branch bonds, and ring bonds.

Direct attachments (bonded atoms) are the easiest to locate within a SMILES formula. They are any of the atomic symbols and bonds located next to the primary atom outside of parentheses (which indicate branches) and ring openings or closings. In the mock SMILES formula CBC, the B atom has two direct attachments−both C. Direct attachments also include those atoms separated by sets of parentheses. Sets of parentheses (branches) are dealt with later. As an example, the SMILES formula for 2-chloropropane can be written CC(C)Cl. The chlorine atom in this formula is considered a direct attachment to the second C.

The second step used to locate bonded atoms treats branch attachments. Branch attachments are designated in the SMILES formula with parentheses. A different way of writing the SMILES formula of 2-chloropropane is CC(Cl)C. With this format, Cl is considered a branch attachment instead of a direct attachment.

At this point, all of the primary atoms are linked to each of their neighboring atoms. For prediction methods without ring corrections or for straight-chained compounds, the task of parsing the compound is complete. However, many group-contribution prediction techniques include corrections for ring strain, and in order to accommodate this feature, the number, size, and type of rings within the molecule must be determined.

Ring Determination

Atoms attached to the primary atom through a ring opening or closing are found in the SMILES formula before the matching digit. In benzene (c1ccccc1), the c preceding the first 1 is matched to the c immediately after the 1 and vice versa. Lower case letters are used to indicate aromaticity in SMILES formulas. If more than one pair of the same digit are found, the atoms are matched in the order in which they appear in the SMILES formula. In biphenyl (c1ccccc1c1ccccc1), for example, the first c1 would be matched to the second c1 and the third c1 would be matched to the fourth. Another example is naphthalene (c12ccccc1cccc2), in which the c12 is linked to both the c1 and c2 primary carbon atoms. The parser records the size of each ring structure within the formula along with the constituent atoms, connections with other rings, and connection order. The parser stores information on interconnected rings using the smallest possible rings. Interconnection positions are stored so that the ring structure may be reconstructed if necessary. For example, naphthalene is composed of two six-member rings. The rings are stored as two separate rings along with enough information that they may be reconstructed to form naphthalene if necessary. Adamantane, a C10 hydrocarbon, is split into three interconnected six-member rings.

The first step in ring evaluation is to determine the ring opening and closing atoms. The ring-opening atom is found by searching the SMILES formula for the atom that precedes the ring digit. If parentheses precede the ring number, then the first atom found outside the set of parentheses is considered the ring-opening atom. Consider the SMILES formula X(NC(C)C)1CC... The ring-opening atom is X in this formula because it is the first atom found prior to the ring number and outside of the parentheses.

This process is repeated to add the ring’s closing atom. The parser begins searching the SMILES formula for the closing atom from right to left at the character that precedes the ring’s closing digit. This is the first atom not located within a set of parentheses.

The parser moves through the SMILES formula examining each letter between the ring opening and closing to determine whether the selected letter is part of the ring. If the atom falls inside the opening and closing of a ring, is not part of another ring, and is not contained within parentheses, then the atom is considered a member of the ring under consideration. In the example of toluene (c1c(C)cccc1), the parser would examine each member of the c(C)cccc string. All of the c’s would be added because they all meet the test criteria. (C) would not be considered part of the ring because it is set off with parentheses.

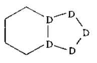
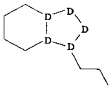
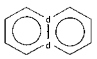
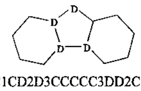
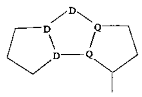
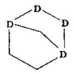
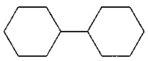
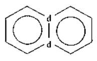
Ring Variations. Ring structures can be written in a variety of equivalent SMILES formulas. The ability of the parser to handle these variations is important, but it complicates the string parsing procedures. For example, occasionally a ring opens inside a set of parentheses but ends outside of the set. In a SMILES formula, a ring opening within a set of parentheses means that all of the atoms previous to the ring opening (up to where the parentheses opened) are part of the ring. To parse this kind of compound, the ring parser clears any ring-member information already stored for this ring atom and begins parsing anew at the ring's opening. This time, the parser loops backward through the ring instead of forward. Once the parentheses opening is reached, the parser adds the atom as the ring opening. This atom links the two parts, the part of the ring inside the parentheses and the part of the ring outside the parentheses, together. The parser then moves to the closing parenthesis, where it begins parsing the formula normally. The parser begins at the end parenthesis and loops through each atom in the SMILES formula, analyzing them until the ring's close is reached. Toluene from the previous example can be written also as Cc(ccc1)cc1. The ring parser begins parsing at the c before the first 1. This is the ring's opening atom and is, at first, considered part of the ring's members. Since the ring's end is outside the set of parentheses that contains the ring's opening, the parser begins moving backward through the ring, adding each atom it finds. In this case, the rest of the c's within the parentheses are added. When the parser encounters the opening parenthesis matching the closing parenthesis separating the ring's end and beginning, the parser adds the first atom found previous to the opening parenthesis. This atom connects the two halves of the ring. The parser then skips to the atom just after the closing parenthesis, where it resumes parsing normally. This part of the parsing adds the last two c's. This results in one six-member ring, the same result that was achieved using the other version of the SMILES formula.

Fused Rings. A difficult situation arises when the parser encounters multiple ring structures. If the ring finder finds a position within the ring where a second ring opening occurs, the parser must count how many potential ring atoms are held in common between the two rings. Atoms common to both rings are included in both ring SMILES structures. When it is known how many atoms are common between the two ring structures, the ring may be split into its smaller components.

The first atom in common opens the second ring, which is designated R2. This atom is added to the common count, after which the parser continues through the SMILES formula looking for additional atoms held in common. Atoms are considered to be in common if they are not set off by parentheses from one of the rings and if they do not open or close an additional ring. If either a set of parentheses opens or a third ring opens, the parser stops counting atoms in common until the parentheses/additional ring closes. In addition, a link atom in a ring (for cases in which a ring opens inside a set of parentheses but closes outside of the parentheses, as discussed above) is also in common. If the atom is held in common between R2 and the primary ring, R1, the atom is added to the common count. Several examples of atoms held in common and atoms not in common are provided in Table 1. In each of the examples the letter D is used to represent an atom in common.

After the common count is determined, the ring structure analysis can be completed. If only one or two atoms are found to be in common and no additional rings are opened, the atoms in common are added as members of both rings. However, if the common count is three or more, a special subroutine is called to finish the parsing. The operation of this subroutine is described below for the various types of SMILES formulas.

Table 1. Examples of Multiring Compounds and Atoms in Common

Structure	Common Count	Comments
 C1CD2DDDD2CC1	5	The common counter starts when a second ring is opened and stops when the ring closes.
 C1CD2D(CCC)DDD2CC1	5	The common counter stops until the end of the parentheses.
 d12d(cccc1)ccccc2	2	The common counter stops counting at the parentheses. Since the ring ends before the close of the parentheses, no additional atoms in common are found after the parentheses close.
 C1CD2D3CCCCC3DD2CC1	4	Once an additional ring is opened the common counter stops until the ring is closed.
 C1CD2DQ3Q(CCC3C)D2CC1	3	The common counter stops counting because of the additional ring opening, The counter does not start again until all of the parentheses sets and additional rings are closed. For ring number 2, the Q's are the atoms in common with ring 3.
 C1D2DDD1CC2	4	The common counter stops if R1's end is reached.
 C1CC(C2CCCCC2)CCC1	0	Rings set off in parentheses are not in common at all.
 c2ccd(cccc1)d1c2	2	A link atom is considered in common.

(1) If the rings R1 and R2 do not contain any additional rings, the last atom in common encountered is added to both rings. Consider naphthalene, for example, which can be described by several different SMILES formulas. Three formulas are c1cc2ccccc2cc1, c12ccccc2cccc1, and c1cccc2ccccc21. The common count, with respect to R1, for each of these SMILES formulas is six. Since no additional rings are opened, the last atom in common is added to the ring. In the cases mentioned, the atoms added are the c's in positions 10, 8, and 12, respectively. Then the parser breaks the molecule into the smallest constituent rings, which in the case of naphthalene is two six-membered rings with two link carbons shared by the rings.

(2) If R2 closes on the same atom as a ring other than R1, a few additional checks are necessary before the parser can decide what should be done. If R1's end atom is set off by parentheses, precedes R2's end in the SMILES formula, and contains additional rings, then R2's end atom is added as a member of the ring. If the end is set off with parentheses, then an additional atom is added. Pyrene (c1cc2ccc3cccc4ccc(c1)c2c34) is a good example (See Figure 1a). R1 encounters R2, with which it has a common count greater than 2. Because R2 closes after R1 and the close of R1 is set off by parentheses, R2's ring end is added as a member of the R1 ring. The positions of the R1 ring elements are then 1, 3, 4, 17, 19, and 22.

(3) If R1's end precedes R2's end and R2's end is coupled with another ring's end that is also opened within R1, the atom that ends R2 is included as a member of R1. In fluoranthracene (see Figure 1b), c1ccc2c(c1)-

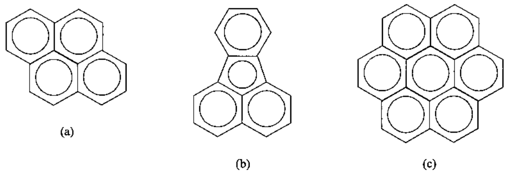


Figure 1. Structures of multiring compounds: (a) pyrene; (b) fluoranthracene; (c) coronene.

c3cccc4cccc2c34, R3 opens inside of R2. Since R3's end is after R2's end, and R3's end is coupled with R4's end (which is also opened inside of R2), the c, located as the 24th character in the SMILES string, is added as a member of R2.

(4) When R1's end and R2's end are separated by only one atom and multiple rings are opened at the same time, the SMILES formula adds the atom separating the two rings. In coronene (Figure 1c), c1cc2ccc3ccc4ccc5ccc6ccc1c7c2c3c4c5c67, R2 also contains rings 3, 4, 5, 6, and 7. R2's end is separated from R3's end by only one atom. The atom separating the two rings' ends is added as a member of R2.

(5) If none of the above cases describe the ring structure, then the parser begins at the end of R1 and loops backward through the SMILES formula, adding atoms until a stop point is reached. The stop point is a ring number in the SMILES formula. If the ring number is both a ring opening and R2, the SMILES formula stops. If the ring number is only a ring's end, then the closing atom is added as a member of the R1 ring. In the coronene example above, R1 opens R2. Since R1 does not meet any of the criteria in 1−4, the parser skips to the end of R1. Atoms are added as ring members as the parser moves backward through the SMILES formula. Once R6's opening is reached, the parser finishes. R2 also has R3 located inside it. After encountering R3, the parser loops to the end.

While parsing, the ring subroutine stores information as to whether each member of the ring has atoms not part of any ring attached to it. With this information, the ring finder is able to keep track of possible *ortho* and *meta* combinations.

Algorithm Results

The automated SMILES matching algorithm described above has been used to predict the heats of formation of over 1500 compounds in the DH test set. In every case the parser identified the correct groups and corrections. Although this does not ensure that the parser will work perfectly, success with over 1500 compounds is very encouraging. To further test the ring analyzer portion of the program, three different SMILES formulas were written (where possible) for about 75 compounds containing rings. The parser was able to find the correct ring groups and structures for all of the 200 ring formulas tested. As most of the rings were hydrocarbons, additional testing is in progress on molecules with interconnected rings containing heteroatoms.

When the automated SMILES matching algorithm is linked with a suite of prediction methods, it provides a powerful, reliable means of predicting thermophysical properties and comparing the values obtained from different methods. It is also a tremendously useful tool in the development of new prediction techniques.

Literature Cited

(1) 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.
(2) 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.
(3) "SMILES Tutorial", www.daylight.com, 1998.
(4) Domalski, E. S.; Hearing, E. D. Estimation of the Thermodynamic Properties of C−H−N−O−S−Halogen Compounds at 298.15 K. *J. Phys. Chem. Ref. Data* **1993**, *22*, 805−1159.

Received for review August 1, 2000. Accepted May 22, 2001.

JE0002336