Pergamon

PII: S0098-1354(97)00253-6

Computers chem. Engng Vol. 22, No. 6, pp. 809–817, 1998
Copyright © 1998 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0098-1354/98 $19.00+0.00

A combinatorial approach for generating candidate molecules with desired properties based on group contribution

F. Friedler[1,2,*], L. T. Fan[2], L. Kalotai[1] and A. Dallos[3]

[1] Department of Computer Science, University of Veszprém, Veszprém, Egyetem u. 10, Hungary H-8200
[2] Department of Chemical Engineering, Kansas State University, 105 Durland, Manhattan, Kansas 66506-5102, U.S.A.
[3] Department of Physical Chemistry, University of Veszprém, Veszprém, Egyetem u. 10, Hungary H-8200

Abstract

The problem of designing molecules or compounds with desired properties is complex primarily because it is combinatorial. A novel approach is proposed here by resorting to a combinatorial analysis of the problem. It is supposed that the set of functional groups is available and that the intervals of values of the desired properties of the molecule or compound to be designed are known. The desired properties constitute constraints on the integer variables assigned to the functional groups. The feasible region defined by such constraints is determined by an algorithm involving a branching strategy. This algorithm generates those collections of the functional groups that can constitute structurally feasible molecules or compounds satisfying the constraints on the given properties. The efficacy of the approach has been demonstrated with several examples. © 1998 Elsevier Science Ltd. All rights reserved

1. Introduction

Designing chemical compounds having prespecified properties has enormous practical impact. Although a number of methods has been developed for this purpose, much remains to be done to devise new methods or to improve available methods for efficiently generating candidate molecular structures for such compounds.

The available methods can be divided into two major classes. In the first class, structures are composed exhaustively, randomly or heuristically, by resorting to expert systems (artificial intelligence) (Constantinou *et al.*, 1995; Joback and Stephanopoulos, 1995; Mavrovouniotis, 1995; Venkatasubramanian *et al.*, 1996) from a given set of groups; the resultant compound is subsequently examined to determine if it is endowed with the specified properties. This "generate-and-test" strategy is usually capable of taking into account only a small subset of feasible molecular structures of the compound of interest. While yielding promising results in some applications, the chance of reaching the target structure by the strategy can indeed be small for any complex problem, e.g. that involving a large number of groups. In the second class, a mathematical programming method is applied to a problem in which the objective function expresses the "distance" to the target (Macchietto *et al.*, 1990). Since the method for estimating the properties of the structure generated, e.g. group contribution, is not sufficiently precise, the assessment of the results on the basis of this objective function may be precarious.

The present work proposes a combinatorial approach for generating all feasible candidate molecular structures whose properties, determined by group contribution, fall within the given intervals. The final selection of the most appropriate structure or structures is carried out by further analysis of such candidate structures with available techniques.

2. Problem formulation

Suppose that the following information is given.

Set G of n groups of which a molecular structure can be composed;
The lower bounds, p_j's, and the upper bounds, P_j's, of the properties to be satisfied, where $j=1, 2, ..., m$;
Upper limit L_i ($i=1, 2, ..., n$) for the number of appearances of group i in a molecular structure to be determined; and
Function f_k ($k=1, 2, ..., m$) representing the value of

[*] Author to whom correspondence should be addressed ([1]).

property k estimated by the group contribution method as

$$f_k(x_1,x_2,...,x_n).$$

In the above expression, $x_1, x_2, ..., x_n$ are, respectively, the numbers of groups #1, #2, ..., #n contained in the molecular structure or compound. For convenience, a functional group in set G with one bond is called the terminator group, and that with three or more bonds, the brancher group.

The problem can now be formulated as follows: We are to search for all molecular structures formed from the given groups, #1, #2, ..., #n, whose numbers are $x_1, x_2, ..., x_n$, respectively, under the condition that the property constraints given below are satisfied.

$$p_j \le f_j(x_1,x_2,...,x_n) \le P_j \ (j=1,2,...,m). \qquad (1)$$

Throughout this paper, the constraints imposed by the molecular structures on feasible spatial configurations are relaxed, and the molecular structures are expressed by simple connected graphs whose vertices and edges represent, respectively, the functional groups from set G and the associated bonds. Thus, the set of such connected graphs need be generated from the set of functional groups G, which satisfies the property constraints by considering multiplicities of the functional groups.

In the conventional generate-and-test approach, all or some of the connected graphs, i.e. molecular structures, are generated from the available functional groups and are then tested against the property constraints. This usually yields an unnecessarily large number of structures. To illustrate the inefficiency of this approach, let the set of molecular structures be partitioned according to the set of functional groups of which they are composed; in other words, two structures are in the same partition if they contain the same groups with identical multiplicities. Naturally, all elements in one partition are either feasible or infeasible under the property constraints. Moreover, the graph or structure generation algorithm of the approach may produce all elements of this partition, even if an element of the partition has been found to be infeasible earlier under the property constraints; obviously, this is highly inefficient.

3. Algorithmic generation of feasible partitions

The feasible partitions of molecular structures can easily be generated for the problem defined in the preceding section by a tree search algorithm similar to the branch-and-bound framework. The approach proposed in the present work is, however, substantially different from the generate-and-test approach. It first identifies the feasible partitions satisfying the property constraints as well as some structural constraints; this is followed by the generation of different molecular structures for each of the resultant partitions. The proposed approach is more effective than the generate-and-test approach because each partition need be considered only once, and the algorithm for generating molecular structures is applied only to the feasible partitions. In addition, the approach can be conveniently implemented by means of tree search.

To exploit the forms of functions f_k $(x_1, x_2, ..., x_n)$ $(k=1, 2, ..., m)$ in generating the candidate molecular structures, these functions are classified according to the following.

Class 1: f_k is a linear function of $x_1, x_2, ..., x_n$ for $k=1, 2, ..., m_1$ $(m_1 \le m)$. It is given in the form of

$$\sum_{i=1}^{n} a_{ki}x_i \ (k=1,2,...,m_1). \qquad (2)$$

Class 2: f_k does not belong to class 1, and it is an invertible function of a linear combination of $x_1, x_2, ..., x_n$ for $k=m_1+1, m_1+2, ..., m_2$ $(m_2 \le m)$. In other words, it can be given in the form of

$$f_k(\sum_{i=1}^{n} a_{ki}x_i) \ (k=m_1+1,m_1+2,...,m_2). \qquad (3)$$

Class 3: f_k belongs to neither class 1 nor class 2, and a linear outer approximation is available for $k=m_2+1, m_2+2, ..., m_3$ $(m_3 \le m)$; thus, there exist coefficients a_{ki} and a_{ki}', such that

$$\sum_{i=1}^{n} a_{ki}x_i \le f_k(x_1,x_2,...,x_n) \le \sum_{i=1}^{n} a_{ki}'x_i \ (k=m_2+1,m_2+2,...,m_3) \qquad (4)$$

Class 4: f_k does not belong to any of the classes, 1 through 3.

Note that it is unnecessary for functions of class 3 to exist; however, the solution procedure is usually facilitated if some or all functions excluded from classes 1 and 2 have linear outer approximations. Although the sharpness of any of such approximations is usually difficult to estimate, the closer the approximation, the more efficient the search.

We are to seek the set of vectors $(x_1, x_2, ..., x_n)$ that satisfy constraint (1). This is accomplished by gradually reducing the search space defined by the constraints,

$$0 \le x_i \le L_i \ (i=1,2,...,n). \qquad (5)$$

The procedure can be illustrated by an enumeration tree in which each node of the tree represents a partition of the search space defined by constraints (5); this partition is termed a partial problem. Certain integer variables of the partial problem are fixed at an integer value. In particular, no values are associated with the variables of the partial problem corresponding to the root of the tree, and each variable is fixed for a partial problem belonging to a leaf of the enumeration tree. Every partial problem generated is tested against the structural and property constraints where the domain of the unfixed variables are extended to real numbers. This test can be performed, for example, by the first phase of the simplex algorithm. Any partial problem failing the test is pruned; otherwise, it is branched on the basis of the possible values of an unfixed integer variable. The leaves of the tree generated by the procedure outlined so far represent the solutions of the problem, i.e. the feasible partitions of the molecular structures. Stated formally, the partial problem represented by the root of the tree is tested against the constraints given below where x_i (i = 1, 2, ..., n) is considered to be a real value.

$$0 \le x_i \le L_i \; (i=1,2,..,n) \qquad (6)$$

$$p_j' \le \sum_{i=1}^{n} a_{ji}x_i \le P_j' \; (j=1,2,...,m_2) \qquad (7)$$

$$p_j \le \sum_{i=1}^{n} a_{ji}'x_i \; (j=m_2+1,m_2+2,..,m_3) \qquad (8)$$

$$\sum_{i=1}^{n} a_{ji}x_i \le P_j \; (j=m_2+1,m_2+2,...,m_3) \qquad (9)$$

$$\sum_{i\in I_1} x_i - \sum_{i\in I_2} x_i \ge -2 \qquad (10)$$

where p_j' and P_j' denote $f_j^{-1}(p_j)$ and $f_j^{-1}(P_j)$ (j = 1, 2, ..., m_2), respectively; constraint (10) forms the necessary condition for a partition to contain a connected graph; and sets I_1 and I_2 comprise the indices of the brancher and terminator groups, respectively (see, e.g. Macchietto *et al.*, 1990). Note that constraints of class 4 are not considered here.

Suppose that the values of variables x_1, x_2, ..., x_k ($k \le n-1$) are fixed a priori as l_1, l_2, ..., l_k, respectively, at an intermediate phase of the procedure; then, the problem is branched to $L_{k+1}+1$ partial problems for l_{k+1}=0, 1, 2, ..., L_{k+1} according to the following two cases.

Case 1: $k \le n-2$, i.e. the search is at an intermediate level on the enumeration tree. For this case,

$$0 \le x_i \le L_i \; (i=k+2,k+3,...,n) \qquad (11)$$

$$p_j' \le \sum_{i=1}^{k+1} a_{ji}l_i + \sum_{i=k+2}^{n} a_{ji}x_i \le P_j' \; (j=1,2,...,m_2) \qquad (12)$$

$$p_j \le \sum_{i=1}^{k+1} a_{ji}'l_i + \sum_{i=k+2}^{n} a_{ji}'x_i \; (j=m_2+1,m_2+2,...,m_3) \qquad (13)$$

$$\sum_{i=1}^{k+1} a_{ji}l_i + \sum_{i=k+2}^{n} a_{ji}x_i \le P_j \; (j=m_2+1,m_2+2,...,m_3) \qquad (14)$$

$$\sum_{i\in I_1} x_i - \sum_{i\in I_2} x_i \ge -2 \qquad (15)$$

where the values of variables x's are extended from integers to real values; and constraint (15) forms the necessary condition for a partition to contain a connected graph.

Case 2: $k=n-1$, i.e. the partial problem belongs to a leaf of the tree. For this case, a test must be performed by simple substitution to determine if *conditions (1) and (2)* and constraints (16) and (17) given below are satisfied for l_1, l_2, ..., l_n

Condition 1. If the partition specified by l_1, l_2, ..., l_n contains functional groups with different types of bonds, e.g. single and double bonds, then there must be a group contained in the partition, which has at least two different types of bonds, and each type belongs to at least one functional group in the partition containing another type of bonds.

Condition 2. The number of bonds identical in type is even.

$$p_j \le f_j(l_1,l_2,...,l_n) \le P_j \; (j=1,2,...,m) \qquad (16)$$

$$\sum_{i\in I_1} l_i - \sum_{i\in I_2} l_i \qquad (17)$$

is an even number not less than -2.

If the partial problem under consideration passes the test, it represents a feasible partition of the candidate molecular structures. In other words, each molecular structure composed of l_1, l_2, ..., l_n numbers of functional groups #1, #2, ..., #n, respectively, satisfies the property constraints.

Similar to the general branch-and-bound framework, different strategies may be devised for searching the feasible partial problems; however, the imposition of some rules is usually advantageous. For example, it is usually effective to order the groups so that the branching is performed on the brancher groups at the top levels of the enumeration tree.

Once the feasible partitions are generated, the feasible molecular structures can be constructed by available computer programs. The present procedure is most advantageous when applied to problems involving large numbers of constraints on the predicted properties, especially if most of them are linear or can be sharply bounded by linear functions. The procedure is illustrated first by a simple problem.

4. Illustration

Suppose that the four functional groups, >CH–, –CH$_3$, –F, and –CH$_2$–, are available to compose molecular structures satisfying the constraints on the boiling (T_b) and melting (T_m) points, i.e.

$$311(K) \le T_b \le 313(K)$$

$$116(K) \le T_m \le 119(K)$$

The normal boiling points, T_b's, and the normal melting points, T_m's, of the compounds are calculated by the group-contribution method of Joback (Horvath, 1992), respectively, as follows:

$$T_b(K) = 198.18 + \sum_i n_i \Delta T_{b,i} \qquad (18)$$

$$T_m(K) = 122.5 + \sum_i n_i \Delta T_{m,i} \, . \qquad (19)$$

Table 1 lists the group contributions; it is assumed that each group may appear at most twice. The enumeration tree is given in Fig. 1. The root of the tree represents the set of constraints given by inequalities (6) through (10); specifically,

Table 1. Group contributions for T_b and T_m of the functional groups investigated in the Illustration

Groups	ΔT_b†(K)	ΔT_m‡(K)
–CH$_3$	23.58	−5.10
–CH$_2$–	22.88	11.27
>CH–	21.74	12.64
–F	−0.03	−15.78

† Joback's increments for T_b (Horvath, 1992).

‡ Joback's increments for T_m (Horvath, 1992).

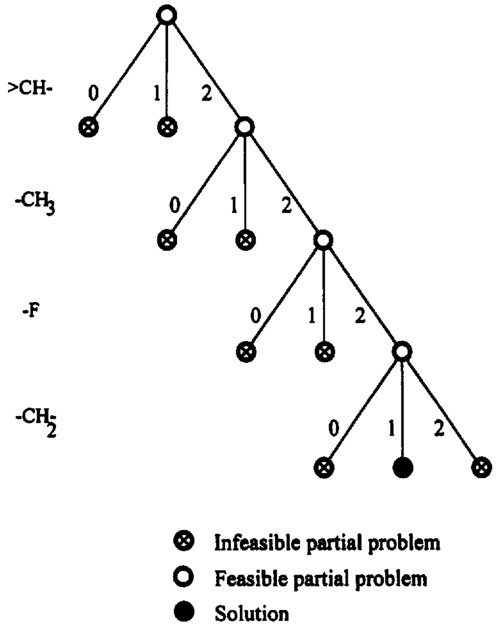


Fig. 1. Enumeration tree for the illustration

$$0\le x_1\le 2, 0\le x_2\le 2, 0\le x_3\le 2, 0\le x_4\le 2$$

$$21.74x_1+23.58x_2-0.03x_3+22.88x_4\le 114.82$$

$$21.74x_1+23.58x_2-0.03x_3+22.88x_4\ge 112.82$$

$$12.64x_1-5.1x_2-15.78x_3+11.21x_4\ge -3.5$$

$$12.64x_1-5.1x_2-15.78x_3+11.21x_4\ge -6.5$$

$$x_1-x_2-x_3\ge -2.$$

This problem is found to be feasible by extending the values of x_i's from integers to real numbers; thus, it is branched into partial problems (P1), (P2), and (P3). If no >CH– is contained in the molecular structure, i.e. if $x_1=0$, we have:

$$\begin{aligned}&0\le x_2<2, 0\le x_3<2, 0\le x_4<2\\&23.58x_2-0.03x_3+22.88x_4\le 114.82\\&23.58x_2-0.03x_3+22.88x_4\ge 112.82\\&-5.1x_2-15.78x_3+11.21x_4\le -3.5\\&-5.1x_2-15.78x_3+11.21x_4\ge -6.5\\&-x_2-x_3\ge -2.\end{aligned}\tag{P1}$$

Similarly, $x_1=1$, expressing that one >CH– appears in the molecular structure, yields:

$$\begin{aligned}&0\le x_2<2, 0\le x_3<2, 0\le x_4\le 2\\&23.58x_2-0.03x_3+22.88x_4\le 93.08\\&23.5x_2-0.03x_3+22.88x_4\ge 91.08\\&-5.1x_2-15.78x_3+11.21x_4\le -16.14\\&-5.1x_2-15.78x_3+11.21x_4\ge -19.14\\&-x_2-x_3\ge -3.\end{aligned}\tag{P2}$$

The case of $x_1=2$, representing two >CH–'s in the structure, leads to:

$$\begin{aligned}&0\le x_2<2, 0\le x_3<2, 0\le x_4<2\\&23.58x_2-0.03x_3+22.88x_4\le 71.34\\&23.58x_2-0.03x_3+22.88x_4\ge 69.34\\&-5.1x_2-15.78x_3+11.21x_4\le -28.78\\&-5.1x_2-15.78x_3+11.21x_4\ge -31.78\\&-x_2-X_3\ge -4.\end{aligned}\tag{P3}$$

Now, by extending the values of x_2, x_3, and x_4 from integers to real numbers, only (P3) is found to be feasible. Hence, it is branched further by regarding x_2 as 0, 1, or 2. The procedure is continued until the search encompasses all x_i's. Every branching step is indicated in the complete enumeration tree exhibited in Fig. 1. Note that this simple problem has a single feasible partition with two >CH–, two –CH_3, two –F, and one –CH_2– functional groups. These groups give rise to various molecular structures, e.g.

$CH_3CHFCHFCH_2CH_3$
$CH_3CH(CH_3)CH_2CHF_2$
$CH_3CHFCH_2CHFCH_3$
$CHF_2CH(CH_3)CH_2CH_3$
$CH_2FCHFCH(CH_3)_2$.

5. Examples

The proposed approach has been tested with several examples. The first and second examples are from the literature. The third example is concerned with the selection of compounds with specific properties, such as molecular weight (mass), melting point, boiling point, crystal density, and liquid volume; the fourth example, the identification of compounds of polyhalogenated biphenyls with plausible molecular structural features by means of a gas chromatographic retention index search window; and the fifth example, the design of pyrazine derivatives as flavor constituents having desired odor-threshold values on the basis of a structural-odor relationship.

5.1. Example 1

The problem given in Joback and Stephanopoulos (1995) has been re-examined under the following conditions. We are to search for compounds, each of which has the estimated normal boiling point, T_b (18), and the estimated normal melting point, T_m (19), in the intervals, $330(K)\le T_b\le 340(K)$ and $130(K)\le T_m\le 140(K)$, respectively, and is composed of the six functional groups given in Table 2; the maximal number of multiplicities of a functional group is ten, i.e. $L_i=10$ ($i=1, 2, ..., 6$).

The algorithm has generated only one feasible partition, $x_1=3$, $x_2=0$, $x_3=2$, $x_4=1$, $x_5=0$, $x_6=3$. Some of the feasible molecular structures contained in this partition are listed below.

Table 2. Group contributions for T_b and T_m of the functional groups investigated in Example 1

Groups	ΔT_b†(K)	ΔT_m‡(K)
$-CH_3$	23.58	−5.10
$-CH_2-$	22.88	11.27
$>CH-$	21.74	12.64
$>C<$	18.25	46.43
$-Cl$	38.13	13.55
$-F$	−0.03	−15.78

†Joback's increments for T_b (Horvath, 1992).
‡Joback's increments for T_m (Horvath, 1992).

$CH_3C(CH_3)_2CHFCHF_2$

$CH_3CF(CH_3)CH(CH_3)CHF_2$

$CH_3CH(CH_3)CF(CH_3)CHF_2$.

5.2. *Example 2*

This is a modification of Example 1. Suppose that nine additional functional groups, given in Table 3, are also available for searching molecules having the properties specified in Example 1. The number of feasible partitions generated is five; see Table 4. Five of the molecular structures, one from each feasible partition, are given below.

Table 3. Group contributions for T_b and T_m of the functional groups investigated in Example 2

Groups	ΔT_b†(K)	ΔT_m‡(K)
$-CH_3$	23.58	−5.10
$-CH_2-$	22.88	11.27
$>CH-$	21.74	12.64
$>C<$	18.25	46.43
$=CH_2$	18.18	−4.32
$=CH-$	24.96	8.73
$=C<$	24.74	11.14
$>C=O$	76.75	61.2
$\equiv C-$	27.38	64.32
$-Br$	66.86	43.43
$-Cl$	38.13	13.55
$-F$	−0.03	−15.78
$-I$	93.84	41.69
$-O-$	22.42	22.23
$-OH$	92.88	44.45

†Joback's increments for T_b (Horvath, 1992).
‡Joback's increments for T_m (Horvath, 1992).

$CH_3CHFCH(OCH_3)CH_2F$

$CH_3CH(CH_3)CHFOCH_3$

$CHF_2OCH(OCH_2F)CHF_2$

$CHF_2OCH(OCH_3)CHF_2$

$CH_3CHFCF(CH_3)CHFCH_3$.

5.3. *Example 3*

We are to design the molecular structure of a compound with the molecular weight (mass), M, the normal boiling point, T_b; the normal melting point, T_m; the density of the crystalline solid, d_s; and the liquid volume, V_b; moreover, these properties fall in given ranges. All the explicit target properties can be estimated by group contribution methods which result in candidate molecular structures represented as collections of functional atomic groups. The molecular weight of a compound (M) is the sum of the molecular weights of its functional groups (ΔM_j), i.e.

$$M(\text{g mol}^{-1}) = \sum_i n_i \Delta M_i \qquad (20)$$

which can be calculated from atomic weights as recommended by IUPAC (1984). The normal boiling and melting points, can be calculated from equations (18) and (19).

The crystal volume of a single molecule can be predicted by the additivity method of Immirzi and Perini (Horvath, 1992) given below:

$$V_s(\text{nm}^3\ \text{molecule}^{-1}) = \sum_i n_i \Delta V_{s,i}\,. \qquad (21)$$

The density of crystalline solid, d_s can be derived from the molecular weight and crystal volume as

$$d_s(\text{g cm}^{-3}) = \frac{1.660M}{V_s} \qquad (22)$$

Although this formula is not linear, it has a sharp linear lower and upper estimator.

The molar liquid volume at the normal boiling point can be readily estimated by the simple additive method of Schroeder (Lyman *et al.*, 1990) by summing the incremental values of every atom, chemical structure or bond as follows:

$$V_b(\text{cm}^3\ \text{mol}^{-1}) = \sum_i n_i \Delta V_{b,i} \qquad (23)$$

The design specifications for selecting target molecular structures are given in Table 5. Table 6 lists increments of the target properties of functional groups

Table 4. Feasible partitions of Example 2

#	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1	2	1	2	0	0	0	0	0	0	0	0	2	0	1	0
2	3	0	2	0	0	0	0	0	0	0	0	1	0	1	0
3	0	1	3	0	0	0	0	0	0	0	0	5	0	2	0
4	1	0	3	0	0	0	0	0	0	0	0	4	0	2	0
5	3	0	2	1	0	0	0	0	0	0	0	3	0	0	0

Table 5. Target property constraints for substances of Example 3

Properties	Value ranges
Molar weight (g mol $^{-1}$)	217–330
Normal boiling point (K)	550–553
Normal melting point (K)	310–314
Density of crystalline solid (g cm $^{-3}$)	1.5–1.7
Molar liquid volume at T_b (cm^3 mol^{-1})	209–250

from which the molecular properties are to be estimated. The groups in this table are based on the functional group definition of Joback (Horvath, 1992).

The proposed algorithm has generated six feasible partitions for the constrains given in Table 5 under the condition that the maximal multiplicity of every functional group is four, i.e. L_i=4 (i=1, 2, ..., 15); these partitions are listed in Table 7. Table 8 lists one compound from each feasible partition together with their estimated properties.

5.4. *Example 4*

The GC retention indices (Kováts, 1961) and the retention indices for high-performance liquid chromatography (HPLC) (Möckel, 1984) are useful aids to the identification of organic compounds, especially, those regarded as environmental contaminants, e.g. polyhalogenated biphenyls (PHBs), depicted below.

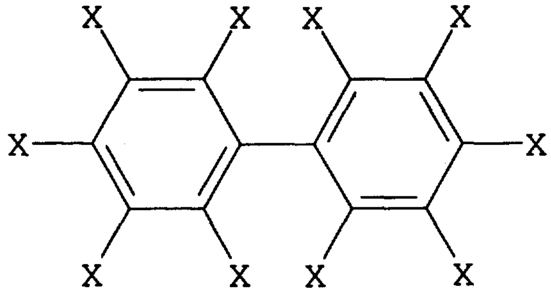


Table 6. Group contributions for M, T_b, T_m, V_s and V_b of the functional groups investigated in Example 3

Groups	ΔM†(g mol^{-1})	ΔT_b‡(K)	ΔT_m§(K)	ΔV_s¶(nm^3 molecule^{-1})	ΔV_b‖(cm^3 mol^{-1})
$-CH_3$	15.032	23.58	−5.10	0.0317	28.0
$-CH_2-$	14.025	22.88	11.27	0.0248	21.0
$>CH-$	13.018	21.74	12.64	0.0179	14.0
$>C<$	12.011	18.25	46.43	0.011	7.0
$=CH_2$	14.025	18.18	−4.32	0.0275	17.5
$=CH-$	13.018	24.96	8.73	0.0206	17.5
$=C<$	12.011	24.14	11.14	0.0137	10.5
$>C=O$	28.01	76.75	61.2	0.0277	21.0
$\equiv C-$	12.011	27.38	64.32	0.0153	14.0
-Br	79.904	66.86	43.43	0.033	31.5
-Cl	35.453	38.13	13.55	0.0267	24.5
-F	18.998	−0.03	−15.78	0.0128	10.5
-I	126.904	93.84	41.69	0.045	38.5
-O-	15.999	22.42	22.23	0.0092	7.0
-OH	17.006	92.88	44.45	0.0161	14.0

†Calculated on the basis of atomic weight as recommended by IUPAC (1984).

‡Joback's increments for T_b (Horvath, 1992).

§Joback's increments for T_m (Horvath, 1992).

¶Calculated on the basis of unit volumes of atomic elements (Lyman *et al.*, 1990).

‖Calculated on the basis of Scroeder's unit volumes of atomic elements (Lyman *et al.*, 1990).

Table 7. Feasible partitions of Example 3

#	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1	3	0	0	0	2	0	4	0	2	0	0	0	1	0	0
2	2	2	2	1	0	0	0	2	0	0	0	4	0	2	0
3	3	0	2	1	0	0	0	2	0	1	0	2	0	0	0
4	3	0	2	1	0	0	0	1	0	1	2	0	0	0	0
5	2	1	1	3	0	0	0	0	0	0	3	3	0	0	1
6	3	0	1	3	0	0	0	0	0	0	3	2	0	0	1

Table 8. Some of the compounds generated in Example 3 with their estimated properties

Compound	M(g mol^{-1})	T_b(K)	T_m(K)	d_s,(g cm^{-3})	V_b(cm^3 mol^{-1})
$CH_3C(CH_2)CCC(CH_3)C(CH_3)C(I)CH_2$	272.116	550.44	313.45	1.61	227.5
$CH_3CF_2CHFCHFC(O)CH_2OC(O)CH_2OCH_3$	260.171	551.05	310.29	1.51	231.0
$CH_3CF(COF)CH(CH_3)C(O)CHBrCH_3$	257.063	550.95	313.18	1.66	213.5
$CHBrClC(CH_3)_2C(O)CHClCH_3$	261.963	550.52	310.64	1.69	220.5
$CH_3CCl_2CCl(CHFCH_2F)CF(OH)CH_3$	273.499	551.89	313.26	1.65	231.0
$CH_3CCl_2CFClC(CH_3)_2CHF(OH)$	255.508	552.62	312.67	1.58	227.5

The GC and HPLC retentions of PHBs are determined by intermolecular forces between stationary phases and PHBs, and they depend on their molecular structures. Seybold and Bertrand (1993) have established a simple regression model describing both the different contributions of the halogens (F, Cl, Br, I) and their positional influences (ortho and non-ortho) on the retention indices measured on the DB-210-CB capillary column (I^{GC}) and on the ODS phases (Waters RAD PAK A Column) using HPLC (I^{HPLC}).

$$I^{GC}_{PCB} = 1645.3 + 227.5N_{Cl} + 315.1N_{Br} + 426.7N_I - 20.5N_{OrF} - 168.8N_{OrCl} - 177.0N_{OrBr} - 214.0N_{OrI} \tag{24}$$

$$I^{HPLC}_{PCB} = 381 + 144N_{Cl} + 171N_{Br} + 270N_I - 49N_{OrF} - 86N_{OrCl} - 61N_{OrBr} - 168N_{OrI} \tag{25}$$

Note that hydrogen and non-ortho fluor-substituents do not affect the values of I^{GC} and I^{HPLC}.

On the basis of equations (24) and (25), we can select the PHBs satisfying the constraints imposed by the GC and HPLC retention indices as functions of their molecular structures, i.e. the number and position of each type of halogen substituents present.

The proposed algorithm has generated nine feasible partitions (see the constraints in Table 9); these partitions are listed in Table 10. Some of the molecular structures are illustrated in Table 11.

5.5. *Example 5*

The first step in any design of perfume is to select the flavor constituents having desired odors and odor-

Table 9. Target property constraints for substances of Example 4

Properties	Retention index value ranges
I^{GC}	2100–2110
I^{HPLC}	600–700

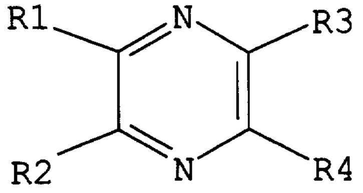
Table 10. Feasible partitions of Example 4

#	x_{Cl}	x_{Br}	x_I	x_H	x_{OrF}	x_{OrCl}	x_{OrBr}	x_{OrI}	x_{OrH}
1	0	0	2	4	0	0	1	1	2
2	0	0	3	3	0	0	1	3	0
3	0	2	0	4	0	1	0	0	3
4	1	0	1	4	1	0	1	0	2
5	1	0	2	3	1	0	1	2	0
6	1	2	0	3	2	0	2	0	0
7	2	0	0	4	0	0	0	0	4
8	3	0	0	3	2	0	1	0	1
9	4	1	0	1	0	2	0	2	0

Table 11. Some of the polyhalogenated biphenyls (B) generated in Example 4 with their estimated retention indices

Compounds	I^{GC}	I^{HPLC}
4,4′,6′-triiodo, 2-bromo-B	2107.7	692
2,3′,4′,5′,6,6′-hexaiodo, 2-bromo-B	2106.4	626
4,4′-dibromo, 2-chloro-B	2106.7	637
4-iodo, 4′-chloro, 2-fluoro, 6′-bromo-B	2102.0	685
2,3,5,6-tetraiodo, 2′-fluoro, 4′-chloro, 6′-bromo-B	2100.7	619
4′-chloro, 2,3,5,6′-tetrabromo, 2′,6′-dibromo-B	2108.0	647
4,4′-dichloro-B	2100.3	669
3,3′,5′-trichloro, 2,6-difluoro, 2′-bromo-B	2109.8	654
2,2′,3,3′,5,5′-hexachloro, 6,6′-diiodo, 4-bromo-B	2104.8	620

threshold (OT) values. These important olfactometric properties depend on the molecular structures of substances. For example, the pyrazine derivatives, illustrated below, are mostly responsible for the roasted and green odor characteristics of some natural products.



These pyrazine derivatives possess low OT values (Mihara and Masuda, 1988) and have widely served as perfume components. Their physico-chemical properties correlate well with the GC retention indices and molecular structures. The difference between the retention indices on the polar (P) and nonpolar (A) stationary phases ($\Delta I = I^P - I^A$) for a given pyrazine derivative is regarded as a measure of the polar interactions (Kováts, 1961). The effects of substituents on intermolecular forces are represented by subtracting the value of ΔI of the parent molecule (pyrazine), i.e. ΔI_0, from that of the substituted one, ΔI; this yields

$$\Delta\Delta I = \Delta I - \Delta I_0 \tag{26}$$

The values of $\Delta\Delta I$ of disubstituted pyrazines can be correlated to their OTs as (Mihara and Masuda, 1988)

$$\log(1/\mathrm{OT}) = 0.04(\sum_i \delta I_i - \Delta\Delta I) + 6.2 \tag{27}$$

where δI_i is an additional parameter for the substituent group. Based on the values of $\Delta\Delta I$'s of the disubstituted pyrazine derivatives (Mihara and Masuda, 1988), the group contribution parameters ($\delta\Delta I_i$) have been determined for several substituent groups. The $\Delta\Delta I$'s for other disubstituted pyrazine derivatives can be estimated from these parameters even without any experimental data on OT as

$$\Delta\Delta I = \sum_i \delta\Delta I_i \tag{28}$$

Table 12 lists the group contribution parameters (δI_i, $\delta\Delta I_i$) for disubstituted pyrazines. Substituted pyrazine derivatives with the desired OT values can be designed from (27) and (28). By assuming that log(1/OT) is in the interval, 7.8 – 7.85, the proposed approach yields the four feasible partitions given in Table 13. Some of the molecular structures are listed in Table 14 together with the estimated OT values; whenever available, the measured data are also listed.

Examples 1 through 5 have been solved on a PC in time ranging from half minute to fifteen minutes. This indicates that the proposed approach is highly efficient.

6. Concluding remarks

The examples presented have amply demonstrated that the proposed combinatorial approach is capable of composing the molecular structures of chemical compounds in an optimal fashion by group contribution. The

Table 12. Group contribution parameters of the substituents of pyrazine for Example 5

R	δI_i†	$\delta\Delta I_i$‡	*R*	δI_i†	$\delta\Delta I_i$‡
H	0	0	$CH_2CH(CH_3)C_3H_7$	– 17	– 95
CH_3	– 27	– 32	$(CH_2)_3CH(CH_3)_2$	– 17	– 98
C_2H_5	– 45	– 60	C_7H_{15}	– 11	– 73
C_3H_7	– 45	– 68	C_8H_{17}	– 28	– 64
$CH(CH_3)_2$	– 45	– 103	$C_{10}H_{21}$	– 64	– 59
C_4H_9	– 32	– 70	$(CH_2)_2CH{=}CHCH_3$	28	– 15
$CH_2CH(CH_3)_2$	– 32	– 93	OCH_3	0	– 48
$CH(CH_3)C_2H_5$	– 32	– 105	OC_2H_5	– 33	– 78
$(CH_2)_3CH{=}CH_2$	28	– 23	OC_6H_5	200	203
C_5H_{11}	– 15	– 69	SCH_3	89	54
$CH_2CH(CH_3)C_2H_5$	-15	– 79	SC_2H_5	36	17
$(CH_2)_2CH(CH_3)_2$	– 15	– 80	SC_6H_5	298	304
C_6H_{13}	– 17	– 66	$COCH_3$	98	86

†Mihara and Masuda (1988).
‡Calculated from $\Delta\Delta I$ values of Mihara and Masuda (1988).

Table 13. Feasible partitions of Example 5

#	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

Table 14. Some of thę pyrazine derivatives generated in Example 5 with their estimated and measured odor-threshold values

R_1	R_2	R_3	R_4	$\log(1/OT)_{est.}$	$\log(1/OT)_{meas.}$
$(CH_2)_2CH{=}CHCH_3$	OC_6H_5	H	H	7.80	—
$C_{10}H_{21}$	OC_2H_5	H	H	7.80	7.22
CH_3	SCH_3	H	H	7.80	8.39
CH_3	H	SCH_3	H	7.80	7.22
CH_3	H	H	SCH_3	7.80	7.69
CH_3	C_8H_{17}	H	H	7.84	—

proposed approach will be a useful tool for expeditiously selecting compounds with desired properties and for identifying chemicals in the light of chromatographic data. It has been shown that the approach can facilitate the first step of the quantitative approach to the design of flavor constituents on the basis of the structure-odor relationship.

References

Constantinou, L., Bagherpour, K., Gani, R., Klein, J. A. and Wu, D. T. (1995) Computer aided product design: Problem formulation, methodology and applications. *Computers chem. Engng* **20**, 685–702.

Horvath, A. L. (1992) *Molecular Design Chemical Structure Generation from the Properties of Pure Organic Compounds*. Elsevier, Amsterdam.

IUPAC Inorganic Chemistry Division, (1984) Element by element review of their atomic weights. *Pure Appl. Chem.* **56**, 695–768..

Joback, K. G. and Stephanopoulos, G. (1995) Searching spaces of discrete solutions: The design of molecules possessing desired physical properties. In *Advances in Chemical Engineering, Vol. 21, Intelligent System in Process Engineering, Part I: Paradigms from Product and Process Design* (Edited by G. Stephanopoulos and C. Han), pp. 257-311. Academic Press, San Diego.

Kováts, E. sz. (1961) Relation between structure and gas-chromatographic data for organic compounds. *Z. Anal. Chem.*, **181**, 351

Lyman, W. J., W. F. Reehl and D. H. (1990) Rosenblatt, *Handbook of Chemical Property Estimation Methods*. ACS, Washington, DC.

Macchietto, S., Odele, O. and Omatsome, O. (1990) Design of optimal solvents for liquid–liquid extraction and gas absorption processes. *Trans. IChemE.*, **68**, 429–433.

Mavrovouniotis, M. L. (1995) Symbolic and quantitative reasoning: Design of reaction pathways through recursive satisfaction of constraints. In *Advances in Chemical Engineering, Vol. 21, Intelligent System in Process Engineering, Part I:Paradigms from Product and Process Design* (Edited by G. Stephanopoulos and C. Han), pp. 147-186. Academic Press, San Diego.

Mihara, S. and Masuda, H. (1988) Structure-Odor Relationships for Disubstituted Pyrazines. *J. Agric. Food. Chem.* **36**, 1242–1247.

Möckel, H. J. (1984) Retention of sulphur and sulphur organics in reversed phase liquid chromatography. *J Chromatogr.* **317**, 589–614.

Seybold, P. G. and Bertrand, J. (1993) A simple model for the chromatographic retentions of polyhalogenated biphenyls. *Anal. Chem.* **65**, 1631–1634.

Venkatasubramanian, V., Sundaram, A., Chan, K. and Caruthers, J. M. (1996) Computer-aided molecular design using neural networks and genetic algorithms. In *Genetic Algorithms in Molecular Modeling* (Edited by J. Devillers), pp. 271-302. Academic Press, London.