Similarity-Based Search and Evaluation of Environmentally Relevant Properties for Organic Compounds in Combination with the Group Contribution Approach

Axel Drefahl* and Martin Reinhard

Department of Civil Engineering, Western Region Hazardous Substance Research Center,
Stanford University, Stanford, California

Received February 17, 1993®

A novel knowledge-based data evaluation system for organic compounds, DESOC, has been developed. DESOC comprises a compound/property database and several modules for the prediction of properties. The database contains compounds that are represented by SMILES notations. Property data are Antoine coefficients for temperature-dependent vapor pressure calculations, aqueous solubility data at 20 and 25 °C, and partition coefficients for the systems air/water, 1-octanol/water, and soil/water. DESOC includes routines for data retrieval, similarity-based search, and property estimation. Estimation of a query property is based on (1) identification of database compounds structurally related to the query, (2) recognition of the structural difference between query and database compounds, and (3) translation of the structural difference into the corresponding property difference. A new approach, the group interchange method (GIM), is introduced for the representation and analysis of structural differences between similar compounds. Compounds are related to each other in terms of elementary group operations. These operations are encoded as linear notations using the grammer of SMILES. The performance of DESOC is illustrated by protocol files generated for selected queries.

INTRODUCTION

Methods employed for estimating environmentally relevant properties are based on thermodynamic principles,[1,2] quantitative property/property relationships (QPPR), quantitative structure/property relationships (QSPR), and group contribution methods (GCM).[3–8] The choice of an estimation method depends on the query compound and the specific property. Application of QSPRs and GCMs requires solely the input of molecular structure. QPPRs and thermodynamic methods typically need the input of structural information and one or several physicochemical properties. Some methods are limited to narrowly defined compound classes, whereas others apply to a more diverse set of compounds. In all cases, property data available for compounds that are structurally similar to the query are useful for verifying and improving estimation results with respect to their plausibility, accuracy, and confidence.

The concept of molecular similarity[9–11] has found diverse applications in the analysis of property data. Similarity-based data retrieval and classification depends on the definition of similarity. Several measurements of molecular similarity are known[12–16] and have been tested with different compound/property sets.[17] The correlation between molecular similarity and property similarity has been examined, for example, with boiling points of isomeric alkanes,[18] with protein binding constants of steroids,[19] in simulated property prediction experiments,[16a] and in studies based on trend vector and cluster analysis.[13,20,21] Similarity-based algorithms have been designed for qualitative or semiquantitative analysis of property data. A quantitative, knowledge-based approach to data analysis and property estimation, however, has to include a structural comparison which accounts for the particular structural difference between a set of different molecules.

Group contribution approaches[22–32] are readily amenable to similarity-based property estimation. The property of a query compound can be estimated from the property of a structurally similar compound, if the property difference associated with the structural difference between the query and database compound has been recognized and evaluated. The incorporation of this approach into a chemical information system requires the following steps: (1) unique representation of structural differences between sets of two molecules, (2) implementation of algorithms which recognize the difference between a query and selected database compounds, and (3) assignment of property differences associated with particular structural differences. This paper presents a new approach for defining structural differences between molecules as linear notations. It is based on the chemical language SMILES.[33,34] The approach has been implemented in a data evaluation system for organic compounds, DESOC, and currently performs similarity-based predictions for solubility and partitioning properties. DESOC automatically identifies candidate compounds in the database, which exhibit a high degree of structural similarity to a query compound. Recognition of the structural difference between a query and candidates and the assignment of the corresponding property difference yields an estimate for the particular property. DESOC includes routines which aid in verifying the plausibility and accuracy of the current values of the built-in property differences.

CONCEPT OF GROUP INTERCHANGE

The maximum common subgraph (MCS)[35] has been used for evaluating the degree of similarity between two molecular graphs. The MCS concept has been applied to both 2-D and 3-D representations of molecules.[16a] In contrast to using the MCS for evaluating relations between two molecules, this work uses the complementary graph of the two MCSs, which consist of those two or more subgraphs that cause discongruence between two molecules. This MCS-complementary graph is denoted as *delta_S*. Imbedded in the formalism of the group contribution approach, *delta_S* will be defined in terms of subgraphs, i.e. groups, interchanged between a query and a candidate molecule.

The example in Figure 1 illustrates the group interchange (GI) operations transforming possible candidate structures (C_1–C_6) into a Query compound (Q). GI operations consist

® Abstract published in *Advance ACS Abstracts,* October 1, 1993.

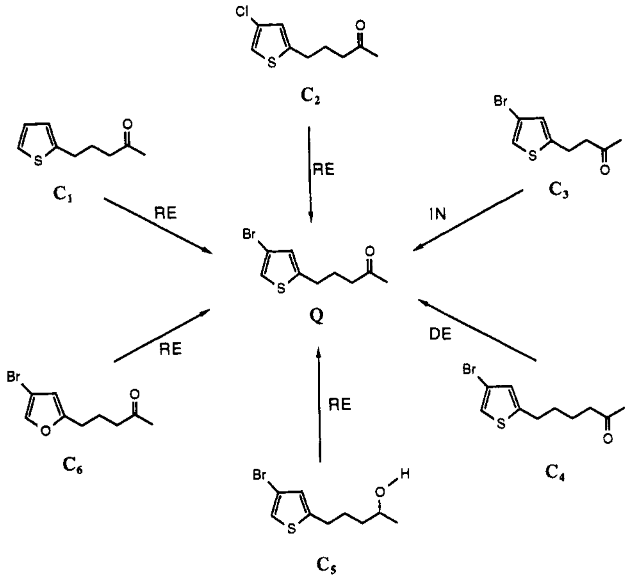


Figure 1. Group interchange between a query and six candidate compounds.

of one or several group operations, including insertion (IN) and deletion (DE) of a group. A group replacement (RE) represents either the replacement of one substituent, i.e. terminal group, by another or the composite operation of a DE followed by an IN. IN, DE, and RE are "elementary" GI operations.

Any two molecules can be transformed into each other by a finite number of elementary GI operations. The approach in DESOC, at this stage, is confined to molecules which can be transformed into one another by one IN, DE, or RE. A system of linear notations for GI operations (LNGI) has been developed to represent structural differences in a compact, unambiguous, and machine-readable code. LNGIs are one-dimensional strings which allow unique and efficient storage and retrieval of information associated with GI operations. The LNGI scheme imposes no *a priori* restriction on the size or composition of the groups. For instance, a group might be a single fluoro atom or a trifluoromethyl or a pentafluorophenyl group.

REPRESENTATION OF STRUCTURAL DIFFERENCES AS LINEAR NOTATIONS

The LNGI system denotes the operation necessary to transform a candidate molecule, C, into a query, Q, by identifying the groups that are interchanged (G_C, G_Q) and its unchanged neighbor groups (G_1, G_2). G_C represents a group that is deleted in the candidate molecule C and G_Q the group that is inserted, resulting in the query molecule Q.

A LNGI string is built as a sequence of symbols beginning with two characters representing the transformation type, IN, DE, or RE, followed by a colon. The interchangeable groups, G_C and G_Q, are included between two bars and are separated by a comma. Four different types of transformations are considered in DESOC:

IN:G_1|G_Q|G_2 insertion of a bivalent group G_Q between two adjacent groups G_1 and G_2

DE:G_1|G_C|G_2 deletion of a bivalent group G_C from a position between two groups G_1 and G_2

RE:G_1|G_C,G_Q| replacement of a terminal group G_C adjacent to a group G_1 by a group G_Q

RE:G_1|G_C,G_Q|G_2 replacement of a bivalent group G_C between G_1 and G_2 by a bivalent group G_Q

Group Representation and Specification. The SMILES[33] grammar is used to represent group structures. For example, the differences between 5-(*o*-bromophenyl)pentanal and 6-(*o*-bromophenyl)hexanal may be represented by either of the following LNGIs:

(a) IN:-C|C|C-; (b) IN:O=C|C|C-;
(c) IN:c1cccc(Br)c1|C|C- (1)

In all three LNGIs the net transformation is the insertion of a methylene group. However, the strings differ in their specification of G_1 and G_2. LNGI a indicates insertion of a methylene group between two methylene groups. LNGIs b and c indicate insertion of a methylene group between a methylene group and a formyl or an *o*-bromophenyl group, respectively.

Generally, different levels of specification for G_1 and G_2 are possible, depending on the atoms included to define these groups. A level-1 specification is based on the key atom A_K of G_1 or G_2. A_K is the atom that is connected to an atom in either G_C or G_Q. All bonds and hydrogen atoms attached to A_K are included into the specification. A level-2 representation accounts additionally for the first neighborhood sphere of A_K. This sphere includes all non-hydrogen atoms attached to A_K along with their attached bonds and hydrogen atoms. Specifications of higher level are obtained in the same manner by accounting for higher neighborhood spheres. The level of specification, LESP, for a LNGI is defined as

$$\mathrm{LESP} = g_1{:}g_2 \qquad (2)$$

where g_1 and g_2 are the specification levels of G_1 and G_2, respectively. For the LNGIs in (1), the LESP is (a) 1:1, (b) 2:1, and (c) 4:1. If an LNGI is based solely on G_1, the notation "g_1:–" applies. If G_1 and G_2 are not considered at all, the LESP is 0:0 or 0:–.

Selection of the LESP depends on the level of desired information attributed to a LNGI. For example, the use of RE:-C|O,=O| with LESP = 1:– is insufficient to specify structural differences associated with conjugated systems. LNGIs of higher degree allow a more distinctive representation. In this case, possible LNGIs with LESP = 3:– are

RE:-CCC|O,=O|; RE:-C=CC|O,=O|;
RE:-N=CC|O,=O|; RE:-C#CC|O,=O| (3)

that represent the same structural difference in different structural environments.

The rules for specifying atoms, bonds, branches, cycles, and aromatic-ring atoms are the same as those used in the SMILES grammar, while some exceptions apply. The atomic symbols "E", "G", and "J" are used instead of "Br", "Cl", and "I", respectively. The single- and the aromatic-bond symbol may be omitted. As already indicated in some of the examples above, the single bond symbol can be required. Its significance is illustrated with the following LNGIs:

(a) IN:-C|C|O-; (b) IN:C|C|O-; (c) IN:-C|C|O (4)

which represent the insertion of a methylene group between (a) a methylene and an oxa group, (b) a methyl and an oxa group, and (c) a methylene and a hydroxy group.

A few symbols in addition to the original SMILES symbols are needed in the LNGI system to represent atoms with unspecified neighboring atoms. The symbols "<" and ">"

denote two single bonds:

IN:>C|C|CN<; IN:=C<|C|CN<; IN:>C<|C|CN< (5)

The symbols ">." and ".<" represent one aromatic and one single bond. For example, the transformation of 1,3-xylene into 2,6-dimethylpyridine can be described by "RE:>.clc,-nlc.<".

The following notation is used to represent the structural difference between a query and a candidate compound

delta_S{S$_C$,S$_Q$} ← LNGI (6)

where S$_C$ and S$_Q$ denote their SMILES notations and "<-" reads as "is denoted by". For example, the difference between 2-iodo-3-methyl-1-pentanol and 2-iodo-3-methyl-1-pentanal can be represented as follows:

delta_S{CCC(C)C(I)CO,CCC(C)C(I)C=O}
← RE:-C|O,=O| (6a)

delta_S{CCC(C)C(I)CO,CCC(C)C(I)C=O}
← RE:>CC|O,=O| (6b)

delta_S{CCC(C)C(I)CO,CCC(C)C(I)C=O}
← RE:>CC(J)C|O,=O| (6c)

with LESP = 1:– for (6a), 2:– for (6b), and 3:– for (6c).

Generation of Unique LNGIs. The above rules can result in different LNGI notations for the same *delta_S*, even if the same LESP is considered. Depending on the form of the LNGI and the complexity of the groups, none, one, or several of the following rules have to be applied to generate a unique notation:

Rule 1. A group consists of at least one non-hydrogen atom. Substitution of a hydrogen atom is represented by including between the bars the atom to which the hydrogen atom is attached. Examples:

RE:-C|C,CJ| instead of RE:-CC|H,J|

RE:|c,c(C(F)(F)F)| instead of RE:c|H,C(F)(F)F|

Rule 2. IN or DE are used instead of RE whenever possible. Examples:

DE:-C|S|SC- instead of RE-C|SS,S|C-

IN:>CC|NC(=S)N|CC-
instead of RE:>C|CC,CNC(=S)NC|C-

Rule 3. The smallest possible number of atomic symbols appears between the bars. Example:

RE:-CC(C)|G,E| instead of RE:-C|C(G)C,C(E)C|

However, use

RE:>C|C=,C#|CC< instead of RE:>CC|=,#|CC<

since bond symbols alone are not group representations.

Rule 4. If G$_1$ and G$_2$ are not equal, then the group with the lower level, i.e. the smaller value for g, appears on the left side of the bars. If g_1 equals g_2, the group with the lower rank appears on the left side of the bars. The group rank is derived from the rank of the atom, A$_K$.

Table I. LNGIs and Applied Rules for Compound Pairs in Figure 1

compound pair representation, notation (6)	LESP	rule(s)
delta_S{C$_1$,Q} ← RE:c|c,c(E)|c	1:1	1
delta_S{C$_2$,Q} ← RE:c|G,E|	1:–	–
delta_S{C$_3$,Q} ← IN:-C|C|C-	1:1	2
delta_S{C$_4$,Q} ← DE:-C|C|C-	1:1	2
delta_S{C$_5$,Q} ← RE:CC(C-)|O,=O|	2:–	3, 6
delta_S{C$_6$,Q} ← RE:>.cc|o,s|c(C-)c	2:2	4, 6

The atom rank is determined by using atomic invariants. The RANK is calculated with the following equation

$$\text{RANK} = 10000N_{\text{con}} + 1000N_{\text{nHb}} + 10N_{\text{A}} + N_{\text{H}} \quad (7)$$

where N_{con} is the number non-hydrogen connections, N_{nHb} is the number of non-hydrogen valence bonds, N_{A} is the atomic number, and N_{H} is the number of attached hydrogen atoms. The following examples illustrate applications of rule 4:

IN:-C|C|C(=S)O- $g_1 < g_2$; LESP = 1:2

IN:-SC|C|SC-
$g_1 = g_2 = 2$; RANK of C in "-SC" = 22 062;
RANK of S in "SC-" = 22 160

RE:=C|C,N|C<
$g_1 = g_2 = 1$; RANK of C in "=C" = 23 061;
RANK of C in "-C<" = 33 061

Rule 5. If G$_1$ and G$_2$ are equal and the notation for G$_C$ or G$_Q$ is asymmetric, then the pairs of atoms in G$_C$ or G$_Q$, having the same distance to the A$_K$ atoms of G$_1$ or G$_2$, are used to determine the direction of G$_C$ and G$_Q$. Starting with the smallest distance, the first pair of atoms with different ranks are selected. G$_Q$ or G$_C$ is written from left to right with the lower ranked atom occurring first. In the following examples, the atom pairs used for ranking are underlined:

IN:-C|$\underline{\text{C}}$COCC$\underline{\text{O}}$|C-
distance = 1; RANK of "C" = 22 062;
RANK of "O" = 22 080

IN:-C|C$\underline{\text{CC}}$(Br)C|C-
distance = 2; RANK of C in "C" = 22 062;
RANK of C in "C(Br)" = 33 061

If the notation of both G$_Q$ and G$_C$ are asymmetric in a RE-type LNGI, rule 5 is applied to G$_C$.

Rule 6. If two or more branches are following the same atom, A$_X$, then these branches are ordered from left to right with increasing rank of the branch atom that is attached to A$_X$. In the following examples, A$_X$ is underlined and the rank of the relevant branch atoms are shown:

IN:-C|$\underline{\text{C}}$(C)(CC)|C- RANK of C in "C" = 11 063;
RANK of C in "CC" = 22 062

IN:cc(c)|C(=$\underline{\text{C}}$(G)(E)|c(c)c
RANK of G = 11 170; RANK of E = 11 350

If the LESP values g_1 or g_2 are increased, a unique LNGI may not be obtained unless in-depth ranking is applied to a deeper neighborhood sphere of the atom to be ranked. However, such cases are not considered at this point. In Table I LNGIs that apply to the compound pairs of Figure 1 are shown.

Examples are given for different LESP, and the applied rules are notified.

DIFFERENCE BETWEEN MOLECULAR FORMULAS

The difference between the molecular formula of a query and a candidate compound is represented by the numbers for each atom type, counting the numbers of atoms that have to be added to or subtracted from the candidate formula to get the query formula. Molecular formula differences (MFD) can also be derived from LNGIs by comparing G_C and G_Q. An example is given for a LNGI representing the replacement of a cyano by a chlorocarbonyl group:

LNGI = RE:-C|C#N,C(G)=O|

atomic symbol	MFD vector	atomic symbol	MFD vector
H	0	Cl	+1
B	0	Br	0
C	0	I	0
N	−1	Si	0
O	+1	P	0
F	0	S	0

MFD vectors are particularly useful in classifying LNGIs. This classification is important for an efficient, automatic recognition of LNGIs. On the basis of the MFD vector, all those LNGIs that do not meet the MFD condition of a given compound pair can be rejected. Thus, only a small set of LNGIs have to be probed in detail.

GROUP INTERCHANGE MODEL (GIM)

The group contribution approach relies on the principle of group additivity.[22] The fundamental equation used in GCMs is

$$P = C_0 + \sum_{i=1}^{M} g(i) \qquad (8)$$

where P is the dependent property, C_0 is a GCM-specific constant, $g(i)$ is the contribution for the ith group in the molecule, and M is the total number of groups in the molecule. The contribution $g(i)$ depends on the type of the ith group. A GCM for a particular property consists of a list of group types with their associated $g(i)$ values.[22-32] Deviations from simple group additivity will be discussed below.

Considering a candidate compound, C, transformed into a query compound, Q, by a RE operation, the compound properties according to eq 8 are

$$P_C = C_0 + \sum_{i=1, i\neq G_C}^{M} g(i) + g(G_C)$$

$$P_Q = C_0 + \sum_{i=1, i\neq G_Q}^{M} g(i) + g(G_Q)$$

The summation terms in either equation refers to the MCS of C and Q. Since these terms are equal, the property difference, *delta_P*, for the replacement operation between Q and C is

$$\textit{delta\_P}\{\mathrm{RE{:}G_1|G_C,G_Q|G2}\} = P_Q - P_C = g(G_Q) - g(G_C) \qquad (9a)$$

The corresponding equations for the IN and DE operations are derived in analogy:

$$\textit{delta\_P}\{\mathrm{IN{:}G_1|G_Q|G2}\} = P_Q - P_C = g(G_Q) \qquad (9b)$$

$$\textit{delta\_P}\{\mathrm{DE{:}G_1|G_C|G2}\} = P_Q - P_C = -g(G_C) \qquad (9c)$$

With eqs 9a–c, the property of Q, related to C by *delta_S*{C,Q}, becomes

$$P_Q = P_C + \textit{delta\_P}\{\mathrm{LNGI}\} \qquad (10)$$

Specific values for *delta_P*{LNGI} can be derived either from $g(i)$ values in known GCM schemes or by using the program DELTAS, introduced below.

COMPOUND/PROPERTY DATABASE

The compound/property database used in DESOC is a factual database consisting of separate data files for each property. The properties included are the 1-octanol/water partition coefficient (K_{OW}), the soil/water partition coefficient (K_{OC}), the air/water partition coefficient (K_{AW}), the solubility in water (S_W), and the coefficients of the Antoine equation to calculate temperature-dependent vapor pressure data and boiling points. The files of property data in DESOC are listed in Table II, along with the number of compounds, N_{cmpd}, and the number of data entries, N_{dent}. A data entry is a property value including its units, accuracy, and reference. The compounds are represented in SMILES. Each SMILES notation is connected with one or several data entries. The data have been critically evaluated and prioritized with respect to their accuracy and whether they are from primary or secondary sources.

LOGKOW. The majority of compounds with K_{OW} values are those with recommended values in the compilation of Sangster.[36] These refer to temperatures between 20 and 25 °C or identified as "ambient". The compounds are hydrocarbons, halogenated hydrocarbons, and monofunctional compounds containing N, O, or S atoms. In addition, data for chlorobenzenes,[37] chloroanilines,[37] and PCBs,[37] thioureas,[38] *s*-triazines,[39] and other pesticides[45] are also included.

LOGKOC. K_{OC} data for hydrocarbons, halogenated organic compounds, and various classes of pesticides have been complied and evaluated by Sabljic.[40] These data have been included in the database along with recent data for alkylbenzenes, chlorobenzenes, and PCBs.[41-43]

AWPC20, AWPC25. The air/water partition coefficient (AWPC) is defined with the equation

$$K_{AW} = H/RT \qquad (11)$$

where K_{AW} is dimensionless, H is the Henry's law constant in Pa (mol/m^3), R is the gas constant (8.314 Pa m^3/(mol K)), and T is the temperature in K.[45] K_{AW} or H data have been taken from critical reviews[44,45] and compilations.[28] These data include values at 20 and 25 °C and are either experimental results or results derived from vapor pressure and solubility data at the same temperature. In addition, more recent experimental data have been compiled.[46-52] To enlarge the range of compound diversity in the database, K_{AW} values have been calculated from vapor pressure and aqueous solubility data, including hydrocarbons, halogenated hydrocarbons, and various O- and N-containing compounds as well as several classes of pesticides. The same compound types are considered in the compilation of S_W and Antoine data.

SWAT20, SWAT25. S_W data were obtained from different compilations[44,45,53,54] and from original sources[55-60] with S_W measured at 20 or 25 °C. The data range from "miscible in any proportions" to very low solubilities.

Table II. Properties in the DESOC Database

property file	property notation	N_{cmpd}	N_{dent}
LOGKOW	$\log(K_{OW})$	688	713
LOGKOC	$\log(K_{OC})$	174	206
AWPC20	K_{AW} or H at 20 °C	155	172
AWPC25	K_{AW} or H at 25 °C	390	506
SWAT20	S_W at 20 °C	205	227
SWAT25	S_W at 25 °C	222	331
ANTOINE	Antoine coeff: A, B, C	384	384

Table III. DESOC Modules and Their Descriptions

module	description
SWATOC	retrieval of solubility in water (S_W at 20 and 25 °C)
PARCOC	retrieval of partition coefficients (K_{AW}, K_{OW}, K_{OC})
ANTOC	vapor pressure and boiling point calculation using Antoine equation
SIMOC	Similarity-based selection of candidate compounds
GIMOC	group interchange model for organic compounds
DELTAS	structural comparison and verification of *delta_P*

ANTOINE. The Antoine parameters are taken from *Lange's Handbook*.[61] The data entry for an Antoine equation includes the coefficients A, B, and C along with the temperature range given for the applicability of the equation.

IMPLEMENTATION OF DESOC

DESOC programs are based on two high-level languages, FORTRAN and C. The *Microsoft* FORTRAN (version 5.1) and the *Microsoft* QuickC (version 2.5) compiler has been used to implement, test, and refine the modules in DESOC. To run DESOC programs, an IBM XT/AT compatible microcomputer with 640K of available RAM, a color monitor, a hard disk, and one floppy disk drive is needed along with DOS version 3.1 or higher.

DESOC is a package of programs designed to access the various property data files and to perform different, user-selected tasks. The modules and their overall functions are shown in Table III. All programs that require the input of query compounds include an interface to enter SMILES notations or to select SMILES files that come with DESOC. The SMILES algorithm, implemented in FORTRAN, has been used to represent molecular structures[21] and to derive connection tables. Screen interfaces have been designed to display intermediate and final results. The modular approach of Pinson,[62] based on C, has been applied to integrate the structure-based retrieval and estimation routines into a user-friendly environment, allowing interactive handling of compounds and evaluation of their properties.

Property Data Representation. Both the compound structure and the compound property information are represented as linked lists of C structures, allowing fast retrieval or modification of either compound or property data. The linked list approach[63,64] has been applied in all procedures scanning the compound/property data.

Similarity Measurement. The Tanimoto coefficient is used to measure the structural similarity between a query and the database compounds. The Tanimoto SI is defined as follows:[20]

$$\mathrm{SI} = \frac{\sum f_Q(k)\, f_C(k)}{\sum f_Q^{\,2}(k) + \sum f_C^{\,2}(k) - \sum f_Q(k)\, f_C(k)} \qquad (12)$$

where the summation in each case is over all different substructures occurring in one or both compounds, and f_Q and f_C are the frequencies with which the kth substructure occurs in the query or database compound, respectively. The atoms are taken as substructures, resulting in a SI with low discrimination capability. However, this simple substructure type separates structural isomers, between which SI = 1, from all other database compounds with SI in the range $0 < \mathrm{SI} < 1$. Currently, the application of additional substructure generation algorithms is tested for goal-oriented, user-supported selection of candidates.

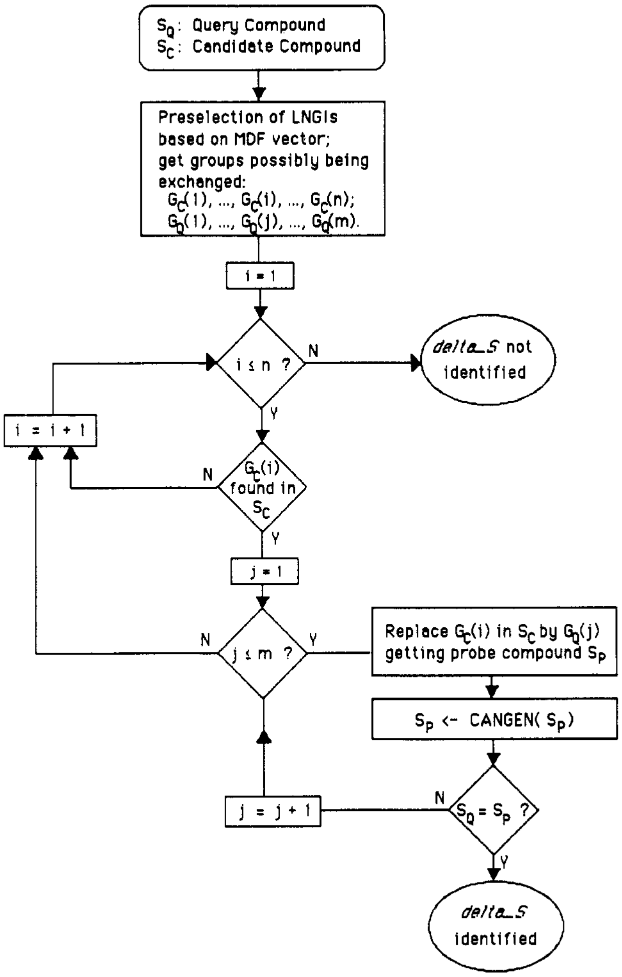

S_Q: Query Compound
S_C: Candidate Compound
Preselection of LNGIs based on MDF vector; get groups possibly being exchanged:
$G_C(1), ..., G_C(i), ..., G_C(n)$;
$G_Q(1), ..., G_Q(j), ..., G_Q(m)$.
i = 1
i ≤ n ?
N
delta_S not identified
Y
i = i + 1
N
$G_C(i)$ found in S_C
Y
j = 1
N
j ≤ m ?
Y
Replace $G_C(i)$ in S_C by $G_Q(j)$ getting probe compound S_P
S_P <- CANGEN(S_P)
N
$S_Q = S_P$?
j = j + 1
Y
delta_S identified

Figure 2. Flow diagram of *delta_S* identification algorithm (Y = yes, N = no).

Recognition of LNGIs. The algorithm used to recognize *delta_S* is indicated in Figure 2 for the RE case. The MDF vector derived for a given query, and a selected candidate compound defines the set of LNGIs possible to describe the difference. This set of LNGIs constitutes all the groups G_C in the candidate that may be replaced by G_Q to obtain the query compound. A loop over this limited number of possible replacement operations is performed. Each replacement leads to a SMILES notation S_P, which has to be probed for equality with the query SMILES after its unique notation is generated with the CANGEN[34] algorithm. If S_P and S_Q are equal, then the LNGI for the probed S_Q/S_C pair is identified. If *delta_S* is of type DE, then $j = 0$, and comparison between S_Q and S_C is performed directly after deletion of G_Q. For IN, the loop over all groups $G_C(i)$ is replaced by a loop to find possible G_1–G_2 substructures.

Built-in *delta_P* Values. At this stage, DESOC employs *delta_P* values, which have been derived from known GCM schemes for log K_{OW}[4] and log K_{AW}.[28,29] Currently, about 50 different LNGIs, most of them with LESP = 1:1 or 1:–, are incorporated into DESOC.

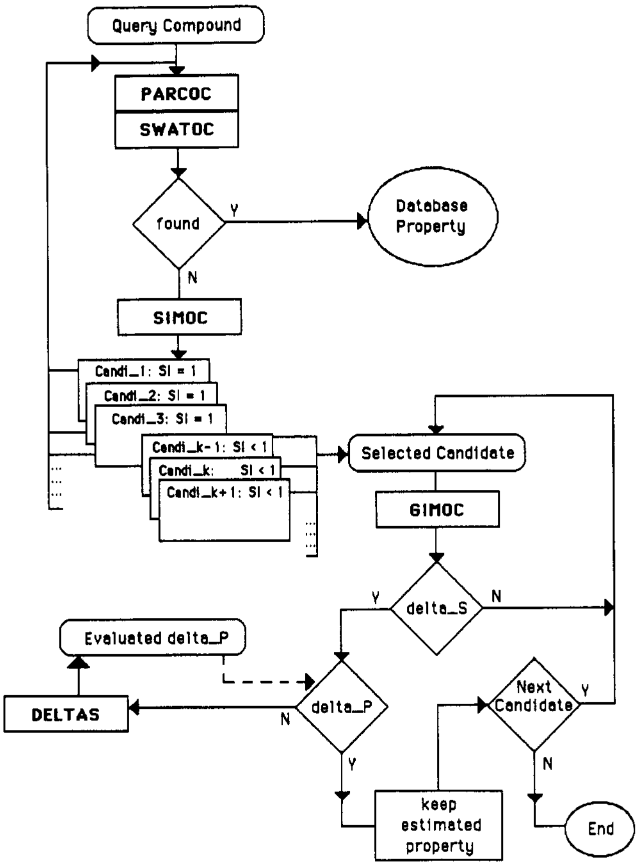
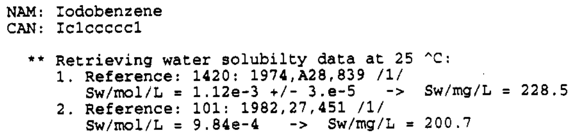


Figure 3. DESOC flow diagram (Y = yes, N = no).



```
NAM: Iodobenzene
CAN: Iclccccc1

  ** Retrieving water solubilty data at 25 ^C:
     1. Reference: 1420: 1974,A28,839 /1/
        Sw/mol/L = 1.12e-3 +/- 3.e-5   ->  Sw/mg/L = 228.5
     2. Reference: 101: 1982,27,451 /1/
        Sw/mol/L = 9.84e-4   ->  Sw/mg/L = 200.7
```

Figure 4. Retrieval of S_W for iodobenzene.

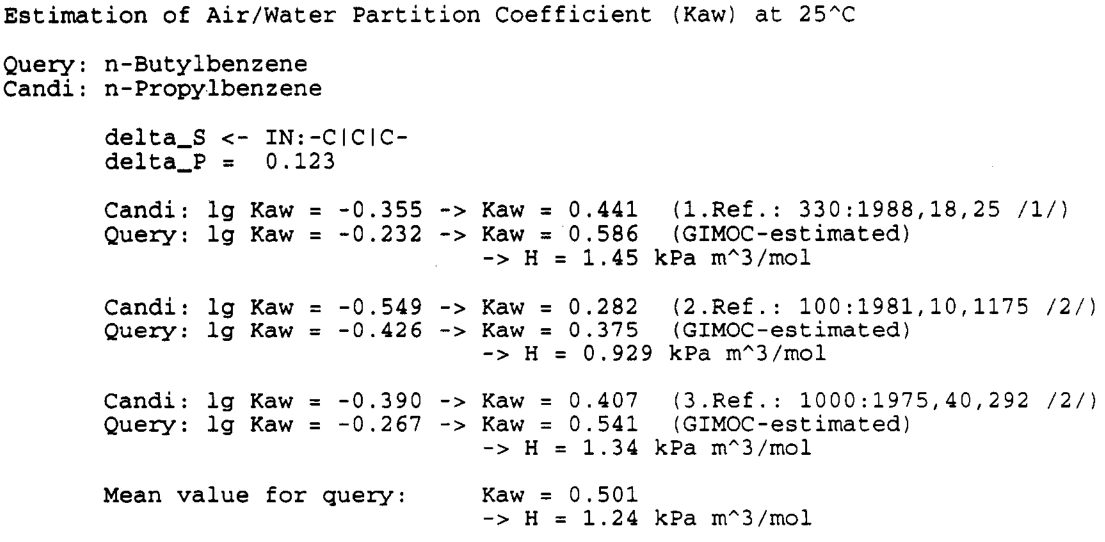
USE AND APPLICATION OF DESOC

A flow chart for DESOC is presented in Figure 3, showing the main functions and their interrelations. Although the programs for which the names are given in bold face are individually executable, the diagram emphasizes their logical, integrated use in property evaluation. Starting with a query compound, PARCOC and SWATOC access the available property data. Depending on the query property, SIMOC rearranges the database compounds with respect to their structural similarity to the query compound. The candidate compounds are structural isomers with SI = 1 followed by the remaining database compounds in the sequence of descending SI. Then data evaluation continues with either data retrieval for the isomers or with GIM-based property estimation. The latter needs the selection of a candidate from the SI-ordered list. GIMOC tries to recognize *delta_S* for the query and the selected candidate. If *delta_S* cannot be identified, another candidate might be selected. Otherwise GIMOC continues by assigning the corresponding *delta_P* to *delta_S*. According to eq 10, the query property is estimated and the procedures can be repeated with additional candidates. Employing DELTAS, an assigned, built-in *delta_P* can be verified or an unassigned *delta_P* can be evaluated.

Each program begins by displaying a bar menu on top of the screen that allows selection of functions for the SMILES input, or searching of compounds, and evaluation of their property data. A compound entry screen supports editing a compound name and the SMILES notation. A list file with an arbitrarily large number of compounds can be generated. The function CANGEN[34] has to be called to transform all SMILES notations in the list file into unique notations. A query compound can be selected either from this user-generated list file or from DESOC files containing CANGEN-processed SMILES notations for various compound classes.

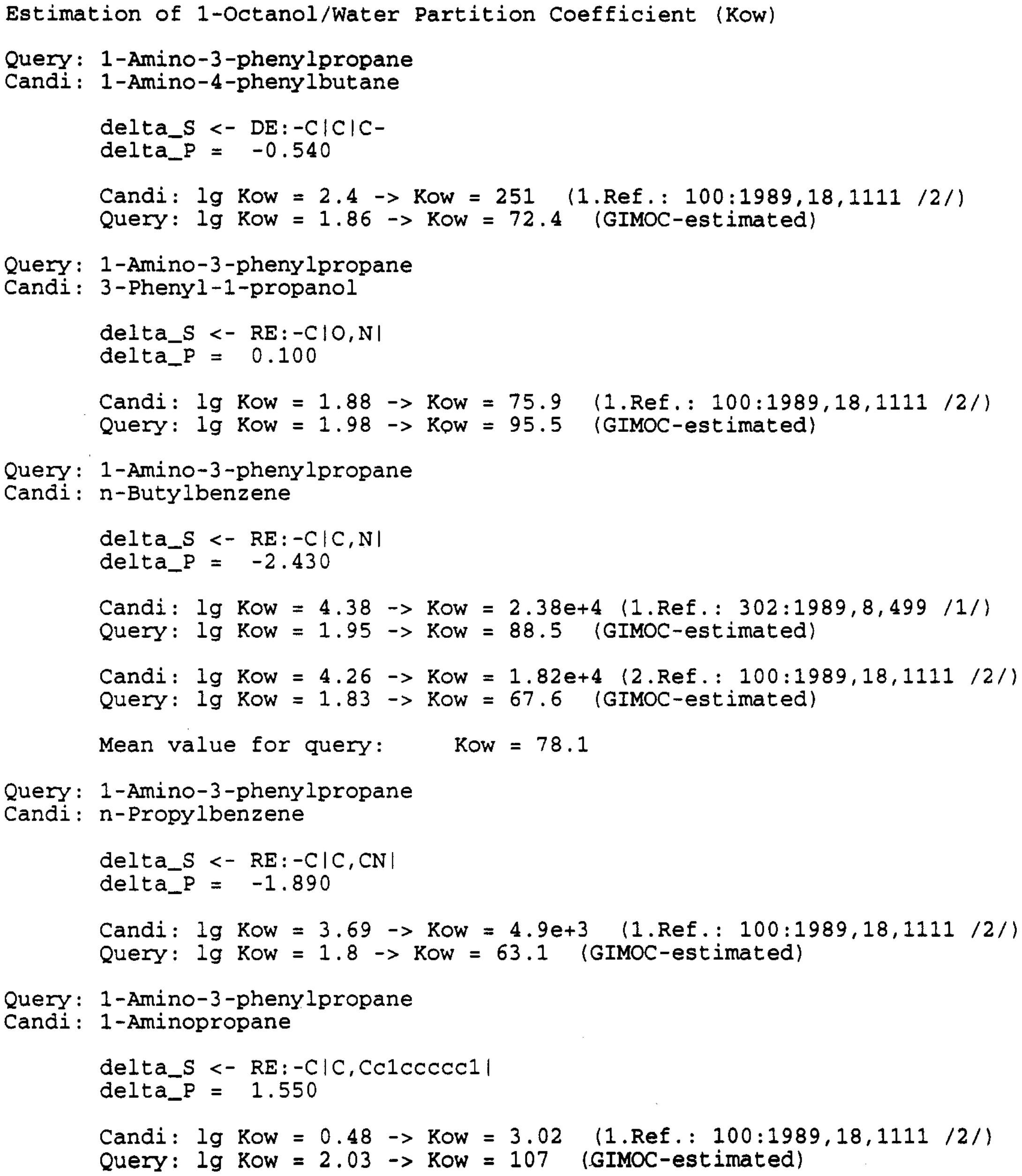
Results are displayed in any of the DESOC programs either in screen windows or in protocol files, as shown in Figures 4–7. An example of data retrieval is shown in Figure 4. Two references for the water solubility of iodobenzene at 25 °C have been found.[55,60] The reference notation contains a journal code, year, volume, and page and an integer indicating if the reference is primary (/1/) or otherwise (/2/). Solubilities are given in mol/L and mg/L. The value in the first reference has been published with its uncertainty.

Figure 5 demonstrates property estimation with SIMOC and GIMOC. For the query-compound *n*-butylbenzene, *n*-propylbenzene has been selected as a candidate to estimate the air/water pollution coefficient at 25 °C. *Delta_S* has been recognized, and a *delta_P* of 0.123 is assigned. Three references are found for the candidate from the database. The K_{AW} values are printed for each reference together with the estimated K_{AW}. A mean value and H in (kPa m^3)/mol, according to eq 11, are additionally recorded. In this case



```
Estimation of Air/Water Partition Coefficient (Kaw) at 25^C

Query: n-Butylbenzene
Candi: n-Propylbenzene

       delta_S <- IN:-C|C|C-
       delta_P =  0.123

       Candi: lg Kaw = -0.355 -> Kaw = 0.441  (1.Ref.: 330:1988,18,25 /1/)
       Query: lg Kaw = -0.232 -> Kaw = 0.586  (GIMOC-estimated)
                                 -> H = 1.45 kPa m^3/mol

       Candi: lg Kaw = -0.549 -> Kaw = 0.282  (2.Ref.: 100:1981,10,1175 /2/)
       Query: lg Kaw = -0.426 -> Kaw = 0.375  (GIMOC-estimated)
                                 -> H = 0.929 kPa m^3/mol

       Candi: lg Kaw = -0.390 -> Kaw = 0.407  (3.Ref.: 1000:1975,40,292 /2/)
       Query: lg Kaw = -0.267 -> Kaw = 0.541  (GIMOC-estimated)
                                 -> H = 1.34 kPa m^3/mol

       Mean value for query:     Kaw = 0.501
                                 -> H = 1.24 kPa m^3/mol
```

Figure 5. Estimation of K_{AW} and H in (kPa m^3)/mol for *n*-butylbenzene using GIMOC.



```
Estimation of 1-Octanol/Water Partition Coefficient (Kow)

Query: 1-Amino-3-phenylpropane
Candi: 1-Amino-4-phenylbutane

       delta_S <- DE:-C|C|C-
       delta_P =  -0.540

       Candi: lg Kow = 2.4 -> Kow = 251  (1.Ref.: 100:1989,18,1111 /2/)
       Query: lg Kow = 1.86 -> Kow = 72.4  (GIMOC-estimated)

Query: 1-Amino-3-phenylpropane
Candi: 3-Phenyl-1-propanol

       delta_S <- RE:-C|O,N|
       delta_P =  0.100

       Candi: lg Kow = 1.88 -> Kow = 75.9  (1.Ref.: 100:1989,18,1111 /2/)
       Query: lg Kow = 1.98 -> Kow = 95.5  (GIMOC-estimated)

Query: 1-Amino-3-phenylpropane
Candi: n-Butylbenzene

       delta_S <- RE:-C|C,N|
       delta_P =  -2.430

       Candi: lg Kow = 4.38 -> Kow = 2.38e+4 (1.Ref.: 302:1989,8,499 /1/)
       Query: lg Kow = 1.95 -> Kow = 88.5  (GIMOC-estimated)

       Candi: lg Kow = 4.26 -> Kow = 1.82e+4 (2.Ref.: 100:1989,18,1111 /2/)
       Query: lg Kow = 1.83 -> Kow = 67.6  (GIMOC-estimated)

       Mean value for query:     Kow = 78.1

Query: 1-Amino-3-phenylpropane
Candi: n-Propylbenzene

       delta_S <- RE:-C|C,CN|
       delta_P =  -1.890

       Candi: lg Kow = 3.69 -> Kow = 4.9e+3  (1.Ref.: 100:1989,18,1111 /2/)
       Query: lg Kow = 1.8 -> Kow = 63.1  (GIMOC-estimated)

Query: 1-Amino-3-phenylpropane
Candi: 1-Aminopropane

       delta_S <- RE:-C|C,Cc1ccccc1|
       delta_P =  1.550

       Candi: lg Kow = 0.48 -> Kow = 3.02  (1.Ref.: 100:1989,18,1111 /2/)
       Query: lg Kow = 2.03 -> Kow = 107  (GIMOC-estimated)
```

Figure 6. Estimation of K_{OW} for 1-amino-3-phenylpropane using GIMOC.

experimental data are available for *n*-butylbenzene. The retrieved value for H is 1.30 ± 0.25 (kPa m^3)/mol.

The example in Figure 6 shows property prediction with the GIM approach using various candidate compounds. K_{OW} is estimated for 1-amino-3-phenylpropane from the K_{OW} of five candidates: 1-amino-4-phenylbutane, 3-phenyl-1-propanol, *n*-butylbenzene, *n*-propylbenzene, and 1-aminopropane. Two K_{OW} references are available for *n*-butylbenzene. *Delta_S* and *delta_P* are given for each candidate followed by the retrieved candidate K_{OW} and the estimated value for the query. The recommended value[36] is 1.83 ± 0.20 for log K_{OW}.

APPLICATION OF DELTAS TO EVALUATE OR VERIFY *DELTA_P*

DELTAS is used to evaluate *delta_P* for a user-selected *delta_S*. DELTAS performs a loop over all possible pairs of compounds in the database for a given property. If a compound pair matches the selected LNGI, the compound pair and the associated property information are printed into the protocol file. This is shown in Figure 7 for the air/water partition coefficient at 25 °C[28,44,48] and RE:-C|C,CE|. Six pairs have been found, where the first three involve alkane/monohaloalkane, and the remaining are monohaloalkane/dihaloalkane pairs. Whereas *delta_P* for the first class of pairs ranges from about −1.9 to −1.6, it ranges from −1.2 to −1.0 for the second. Clearly, *delta_P* depends on the degree of halogen substitution in the query and the candidate. The importance of multiple substitution factors has previously been recognized by Hine and Mookerjee,[28] who developed a group contribution and a bond contribution method for the estimation of K_{AW}. On the basis of recent data, their bond method has been reevaluated and updated by Meylan and Howard.[29] Hine and Mookerjee[28] introduced "distant polar interaction" terms to account for certain factors in multifunctional compounds. However, most of these factors are without statistical significance because they are drawn from only a small number of compounds. Similarly, Meylan and Howard[29] have used correction factors for particular chemical classes and multiple occurrences of certain functional groups. The calculation of *delta_P* for the aforementioned *delta_S* is shown in Figure 8, using the group and bond notations as given in the sources.

```
1. Pair of compounds:
   A. Propane
      lg(Kaw) = 1.461     (100: 1981,10,1175 /2/)
      lg(Kaw) = 1.460     (1000: 1975,40,292 /2/)
      meanA   = 1.460
   B. 1-Bromopropane
      lg(Kaw) = -0.410    (1000: 1975,40,292 /2/)
DELTA_P = -1.870

2. Pair of compounds:
   A. n-Butane
      lg(Kaw) = 1.588     (100: 1981,10,1175 /2/)
      lg(Kaw) = 1.580     (1000: 1975,40,292 /2/)
      meanA   = 1.584
   B. 1-Bromobutane
      lg(Kaw) = -0.300    (1000: 1975,40,292 /2/)
DELTA_P = -1.884

3. Pair of compounds:
   A. Isopentane
      lg(Kaw) = 1.746     (100: 1981,10,1175 /2/)
   B. 1-Bromo-3-methylbutane
      lg(Kaw) = 0.150     (1000: 1975,40,292 /2/)
DELTA_P = -1.596

4. Pair of compounds:
   A. Chloroethane
      lg(Kaw) = -0.306    (330: 1988,18,25 /1/)
      lg(Kaw) = -0.460    (1000: 1975,40,292 /2/)
      meanA   = -0.383
   B. 1-Chloro-2-bromoethane
      lg(Kaw) = -1.430    (1000: 1975,40,292 /2/)
DELTA_P = -1.047

5. Pair of compounds:
   A. Bromoethane
      lg(Kaw) = -0.510    (1000: 1975,40,292 /2/)
   B. 1,2-Dibromoethane
      lg(Kaw) = -1.576    (330: 1988,18,25 /1/)
      lg(Kaw) = -1.889    (100: 1981,10,1175 /2/)
      lg(Kaw) = -1.540    (1000: 1975,40,292 /2/)
      meanB   = -1.668
DELTA_P = -1.158

6. Pair of compounds:
   A. 1-Bromopropane
      lg(Kaw) = -0.410    (1000: 1975,40,292 /2/)
   B. 1,3-Dibromopropane
      lg(Kaw) = -1.440    (1000: 1975,40,292 /2/)
DELTA_P = -1.030
```

Figure 7. DELTAS result for RE:-C|C,CE| with AWPC25 compounds.

```
delta_S  <-  RE:-C|C,CE|

1. Using GCM of Hine and Mookerjee, 1975:

   delete  CH3(X)          - (+0.62)
   insert  CH2Br(C)        + ( -1.10)
   ---------------------------------
                delta_P  =   -1.72

2. Using GCM of Meylan and Howard, 1991:

   delete  C-H             - (+ 0.1197)
   insert  C-Br            + ( - 0.8187)
  -------------------------------------
                delta_P =     -0.94
```

Figure 8. Evaluation of *delta_P* using two different GCM schemes.

Significantly different *delta_P* values are obtained. DELTAS helps to clarify such differences with respect to the compound pair types. The example of Figures 7 and 8 demonstrates the need of higher level LNGIs that would consider both halogen substitutions, on the interchangeable substituents and the α-, β-, or γ-position of these substituents. Although the GIM concept has been designed for this task, the pool of experimental data is limited for full exploitation of its potential.

DISCUSSION

Applicability and usefulness of an estimated property value depends on the likely error associated with the estimate. The GCM and GIM approaches differ in their principles for assessing this error. The total error of a GCM estimate depends on the method error derived from the uncertainties associated with each contributing group occurring in a query molecule. In contrast, the total error of a GIM estimate relies only on the uncertainties of those groups determining *delta_S*. In addition, however, the GIM error depends on the propagated error for the candidate property P_C.

DESOC provides three types of arguments for assessing the validity of a GIM-based estimate. These arguments rely on (1) the accuracy of the candidate property value, P_C, (2) the verification of the current *delta_P* by employing DELTAS, and (3) the comparison of GIM estimates derived from different candidates. DESOC does not generate a final "best value" but supports the user to make an informed judgement. The GIM approach emphasizes a deductive line in the estimation process with each step confirmed by the currently available facts. GIM-based property estimation is the preferred method whenever appropriate candidates are available for a query. Especially when query compounds become structurally more complex, the application of GCMs gets increasingly involved. Many GCMs employ a scheme of extra contributions accounting for diverse, intramolecular factors associated with chain length, ring size, group-to-group interactions, and specific substitution patterns. The GIM approach reduces the complexity imposed by the variety of factors significantly. Only the factors applying to the interchanged groups have to be accounted for, whereas all those factors confined to the MCS do not have to be considered.

The modular design of DESOC supports the facile integration of additional modules and their alignment with user needs emerging with new types of queries.

CONCLUSIONS

1. DESOC is a chemical information system designed for the integration of group contribution models (GCM) and property-estimation approaches based on the comparison of the molecular structure of a query with the structure of database compounds.

2. DESOC aids in verifying and refining available GCMs.

3. The concept of group interchange (GI) is introduced to describe the difference between two molecules in terms of formal group deletion (DE), insertion (IN), or replacement (RE) operations.

4. A notation system for GI operations between molecules of similar structure has been proposed, verified, and tested for computer-assisted property-estimation purposes.

5. LNGIs are linear notations serving as compact, descriptive, and unambiguous names for particular GI operations.

6. Statistical parameter evaluation, problems with chance correlations and training set criteria which are faced with regression-based models are less relevant in the primarily cognitive GIM approach.

GLOSSARY

C	candidate
DE	deletion
GCM	group contribution method
GIM	group interchange method
IN	insertion
LESP	level of specification
LN	linear notation (=line notation)
LNGI	linear notation for group interchange
MCS	maximum common subgraph
Q	query
QPPR	quantitative structure/property relationship

QSPR	quantitative structure/property relationship
RE	replacement
S	SMILES notation
SI	similarity index

ACKNOWLEDGMENT

This work was supported in part by the Office of Research and Development, U.S. Environmental Protection Agency, Washington, D.C., under Agreement R-815738-01 through the Western Region Hazardous Substance Research Center, Stanford University, Stanford, CA, and through support from Cononco, Norway, through Aquateam, Oslo. The content of this paper does not necessarily represent the view of these organizations. We appreciate the valuable input of Frank Hiersekorn and Jack E. Barbash.

REFERENCES AND NOTES

(1) Reid, R. C.; Prusnitz, J. M.; Polling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill Book Co.: New York, 1987.

(2) Barton, A. F. M. *CRC Handbook of Solubility and Other Cohesion Parameters*, 2nd ed.; CRC Press, Inc.: Boca Raton, FL, 1991.

(3) Jochum, C.; Hicks, M. G.; Sunkel, J. *Physical Property Prediction in Organic Chemistry*; Springer-Verlag: Berlin, Heidelberg, 1988.

(4) Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handbook of Chemical Property Estimation Methods*, 3rd ed.; American Chemical Society: Washington, DC, 1990.

(5) Jørgensen, S. E. Estimation of Physical-Chemical Properties in Ecotoxicology. In *Modelling in Ecotoxicology*; Jørgensen, S. E., Ed.; Elsevier: Amsterdam, 1990.

(6) Neely, W. B.; Blau, G. E. *Environmental Exposure from Chemicals*; CRC Press, Inc.: Boca Raton, FL, 1985; Vol. I.

(7) Karcher, W.; Devillers, J. Practical Applications of Quantitative Structure-Activity Relationships (QSAR). *Environmental Chemistry and Toxicology*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990.

(8) Kaiser, K. L. E. *QSAR in Environmental Toxicology*; D. Reidel Publishing Co.: Dordrecht, The Netherlands, (a) 1984, Vol. I; (b) 1987, Vol. II.

(9) Johnson, M. A. A Review and Examination of the Mathematical Spaces Underlying Molecular Similarity Analysis. *J. Math. Chem.* **1989**, *3*, 117–145.

(10) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons, Inc.: New York, 1990.

(11) Heller, S. R. Similarity in Organic Chemistry: A Summary of the Beilstein Institute Conference. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 578–579.

(12) Randic, M.; Wilkins, C. L. Graph Theoretical Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31–37.

(13) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

(14) Jerman-Blazic, B.; Randic, M. Similarity Measures for Sets of Strings and Application in Chemical Classification. *J. Math. Chem.* **1990**, *4*, 217–225.

(15) Hagadone, T. R. Molecular Substructure Similarity Searching: Efficient Retrieval in Two-Dimensional Structure Databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515–521.

(16) (a) Willett, P. Algorithms for the Calculation of Similarity in Chemical Structure Databases. Reference 10, Chapter 3. (b) Bawden, D. Application of Two-Dimensional Chemical Similarity Measures to Database Analysis and Querying. Reference 10, Chapter 4.

(17) Willett, P.; Winterman, V. A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18–25.

(18) Randic, M. Design of Molecules with Desired Properties. Reference 10, Chapter 5.

(19) Rum, G.; Herndon, W. C. Molecular Similarity Concepts. 5. Analysis of Steroid–Protein Binding Constants. *J. Am. Chem. Soc.* **1991**, *113*, 9055–9060.

(20) Willett, P.; Winterman, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.

(21) Drefahl, A. Model Development for Predicting the Environmental Behavior of Organic Compounds Based on Computer Aided Structure/Property Transformations. Ph.D. dissertation, Department of Organic Chemistry I, Technical University Munich, Garching, Germany, 1988.

(22) Exner, O. Additive Physical Properties. I. *Collect. Czech. Chem. Commun.* **1966**, *31*, 3222–3251. Exner, O. Additive Physical Properties. II. *Collect. Czech. Chem. Commun.* **1967**, *32*, 1–22. Exner, O. Additive Physical Properties. III. *Collect. Czech. Chem. Commun.* **1967**, *31*, 24–54.

(23) Homologous Series and Homomorphs. Reference 2, Chapter 6.

(24) Lyman, W. J. Octanol/Water Partition Coefficient. Reference 4, Chapter 1.

(25) Chou, J. T.; Jurs, P. C. Computer-Assisted Computation of Partition Coefficients from Molecular Structures Using Fragment Constants. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 172–178.

(26) Broto, P.; Moreau, G.; Vandycke, C. Molecular structure: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem-Chim. Ther.* **1984**, *19*, 71–78.

(27) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure–Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

(28) Hine, J.; Mookerjee, P. K. The Intrinsic Hydrophilic Character of Organic Compounds. Correlations in Terms of Structural Contributions. *J. Org. Chem.* **1975**, *40*, 292–298.

(29) Meylan, W.; Howard, P. H. Bond Contribution Method for Estimating Henry's Law Constants. *Environ. Toxicol. Chem.* **1991**, *10*, 1283–1293.

(30) Polak, J.; Lu, B. C.-Y. Mutual Solubilities of Hydrocarbons and Water at 0 and 25 °C. *Can. J. Chem.* **1973**, *51*, 4018–4023.

(31) Korenman, I. M.; Gur'ev, I. A.; Gur'eva, Z. M. Solubility of Liquid Aliphatic Compounds in Water. *Russ. J. Phys. Chem.* **1971**, *45EE*, 1065–1066.

(32) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474–482.

(33) 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.

(34) 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.

(35) Brint, A. T.; Willett, P. Algorithms for the Identification of Three-Dimensional Maximal Common Substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152–158.

(36) Sangster, J. Octanol-Water Partition Coefficients of Simple Organic Compounds. *J. Phys. Chem. Ref. Data* **1989**, *18*, 1111–1229.

(37) De Bruijn, J.; Busser, F.; Seinen, W.; Hermens, J. Determination of Octanol/Water Partition Coefficients for Hydrophobic Organic Chemicals with the "Slow-Stirring" Method. *Environ. Toxicol. Chem.* **1989**, *8*, 499–512.

(38) Govers, H.; Ruepert, C.; Stevens, T.; van Leeuwen, C. J. Experimental Determination of Partition Coefficients of Thiourea and Their Toxicity to Photobacterium Phosphoreum. *Chemosphere* **1986**, *15*, 383–393.

(39) Finizio, A.; Di Guardo, A.; Arnoldi, A.; Vighi, M.; Fanelli, R. Different Approaches for the Evaluation of K_{ow} for s-Triazine Herbicides. *Chemosphere* **1991**, *23*, 801–812.

(40) Sabljic, A. On the Prediction of Soil Sorption Coefficients of Organic Pollutants from Molecular Structure: Application of Molecular Topology Model. *Environ. Sci. Technol.* **1987**, *21*, 358–366.

(41) Paya-Perez, A. B.; Riaz, M.; Larsen, Bo R. Soil Sorption of 20 PCB Congeners and Six Chlorobenzenes. *Ecotoxicol. Environ. Saf.* **1991**, *21*, 1–17.

(42) Abdul, S. L.; Gibson, T. L.; Rai, D. N. Statistical Correlations for Predicting the Partition Coefficient for Nonpolar Organic Carbons and Water. *Hazard. Waste Hazard. Mater.* **1987**, *4*, 211–222.

(43) Vowles, P. D.; Mantoura, R. F. C. Sediment-Water Partition Coefficients and HPLC Retention Factors of Aromatic Hydrocarbons. *Chemosphere* **1987**, *16*, 109–116.

(44) Mackay, D.; Shiu, W. Y. A Critical Review of Henry's Law Constants for Chemicals of Environmental Interest. *J. Phys. Chem. Ref. Data* **1981**, *10*, 1175–1199.

(45) Suntio, L. R.; Shiu, W. Y.; Mackay, D.; Seiber, J. N.; Glotfelty, D. Critical Review of Henry's Law Constants for Pesticides. *Rev. Environ. Contam. Toxicol.* **1988**, *103*, 1–59.

(46) Dunnivant, F. M.; Coates, J. T.; Elzerman, A. W. Experimentally Determined Henry's Law Constants for 17 Polychlorobiphenyl Congeners. *Environ. Sci. Technol.* **1988**, *22*, 448–453.

(47) Fendinger, N. J.; Glotfelty, D. E. Henry's Law Constants for Selected Pesticides, PAH's and PCBs. *Environ. Toxicol. Chem.* **1990**, *9*, 731–735.

(48) Ashworth, R. A.; Howe, G. B.; Mullins, M. E.; Rogers, T. N. Air-Water Partition Coefficients of Organics in Dilute Aqueous Solutions. *J. Hazard. Mater.* **1988**, *18*, 25–36.

(49) Gossett, J. M. Measurement of Henry's Law Constants for C_1 and C_2 Chlorinated Hydrocarbons. *Environ. Sci. Technol.* **1987**, *21*, 202–208.

(50) Nicholson, B. C.; Maguire, B. P.; Bursill, D. B. Henry's Law Constants for the Trihalomethanes: Effect of Water Composition and Temperature. *Environ. Sci. Technol.* **1984**, *18*, 518–521.

(51) Lalezary, S.; Pirbazari, M.; McGuire, M. J.; Krasner, S. W. Air Stripping of Taste and Odor Compounds from Water. *J. Am. Water Works Assoc.* **1984** (March), 83–87.

(52) Munz, C. Air Water Phase Equilibria and Mass Transfer of Volatile Organic Solutes. Ph.D. dissertation, Department of Civil Engineering, Stanford University, Stanford, CA, 1985.

(53) Freier, R. K. *Aqueous Solutions*; Walter de Gruyter: Berlin, 1976; Vol. 1.

(54) Suntio, L. R.; Shiu, W. Y.; Mackay, D. A Review of the Nature and Properties of Chemicals Present in Pulp Mill Effluents. *Chemosphere* **1988**, *17*, 1249–1290.

(55) Tewari, Y. B.; Miller, M. M.; Wasik, S. P.; Martire, D. E. Aqueous Solubility and Octanol/Water Partition Coefficient of Organic Compounds at 25.0 °C. *J. Chem. Eng. Data* **1982**, *27*, 451–454.

(56) Miller, M. M.; Ghodbane, S.; Wasik, S. P.; Tewari, Y. B.; Martire, D. E. Aqueous Solubilities, Octanol/Water Partition Coefficients, and Entropies of Melting of Chlorinated Benzenes and Biphenyls. *J. Chem. Eng. Data* **1984**, *29*, 184–190.

(57) Owens, J. W.; Wasik, S. P.; DeVoe, H. Aqueous Solubilities and Enthalpies of Solution of *n*-Alkylbenzenes. *J. Chem. Eng. Data* **1986**, *31*, 47–51.

(58) Yalkowsky, S. H.; Orr, R. J.; Valvani, S. C. Solubility and Partitioning. 3. The Solubility of Halobenzenes in Water. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 351–353.

(59) Chiou, C. T.; Freed, V. H. *Chemodynamic Studies on Bench Mark Industrial Chemicals*. Annual Report NSF/RA-770286; Oregon State University: Corvallis, 1977.

(60) Vesala, A. Linear Free Energy Correlations of Free Energy of Transfer with Solubility and Heat of Melting of a Nonelectrolyte. *Acta Chem. Scand.* **1974**, *A28*, 839.

(61) Dean, J. A. *Lange's Handbook of Chemistry*; McGraw-Hill Book Co.: New York, 1979.

(62) Pinson, J. *Designing Screen Interfaces in C*; Yourdon Press Computing Series; Yourdon Press: Englewood Cliffs, NJ, 1991.

(63) Schildt, H. *Advanced C*, 2nd ed.; Osborne McGraw-Hill: Berkeley, CA, 1988.

(64) Schildt, H. *Artificial Intelligence Using C*; Osborne McGraw-Hill: Berkeley, CA, 1987.

Registry No. Supplied by the Author: Ic1ccccc1, 591-50-4; CCCCc1ccccc1, 104-51-8; CCCc1ccccc1, 103-65-1; NCCCc1ccccc1, 2038-57-5; NCCCCc1ccccc1, 13214-66-9; OCCCc1ccccc1, 122-97-4; CCCN, 107-10-8.