SEARCHING SPACES OF DISCRETE SOLUTIONS: THE DESIGN OF MOLECULES POSSESSING DESIRED PHYSICAL PROPERTIES

Kevin G. Joback[1] and George Stephanopoulos

Laboratory for Intelligent Systems in Process Engineering
Department of Chemical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

I. Introduction 258
A. Brief Review of Previous Work 260
B. General Framework for the Design of Molecules 264
II. Automatic Synthesis of New Molecules 267
A. The Generate-and-Test Paradigm 267
B. The Search Algorithm 271
C. Case Study: Automatic Design of Refrigerants 283
D. Case Study: Automatic Design of Polymers as Packaging Materials 284
III. Interactive Synthesis of New Molecules 290
A. Illustration of Interactive Design 291
B. Case Study: Interactive Design of Refrigerants 296
C. Case Study: Interactive Design of an Extraction Solvent 299
D. Case Study: Interactive Design of a Pharmaceutical 301
IV. The *Molecule-Designer* Software System 304
A. General Description 304
B. Interactive-Design-Relevant Sections 305
V. Concluding Remarks 307
References 309

Strings of letters form words. From words to verses and stanzas, a poet composes a work with its own dynamic behavior, such as emotional impact on the reader, that transgresses the character of its components. In an analogous manner, atoms form functional groups, and these, in turn, yield molecules with distinct behavior, e.g., physical properties. It takes a

[1] Present address: Molecular Knowledge Systems, Inc., Nashua, New Hampshire, USA.

Homeric or Shakespearean genius to convert letters to an epic with a predefined desired impact. It suffices to efficiently search a space of combinatorial alternatives, in order to identify the molecules that satisfy the desired constraints on a set of physical properties. Often, the requisite scientific knowledge is fragmented, dispersed, and nonformalized, making the deductive search for the desired molecules inefficient or impossible. The inductive "genius" of a scientist or engineer is needed to break the impasse in such cases. By evolution or revolution one needs to respond to tighter and shifting product specifications and identify new solvents, pharmaceuticals, imaging chemicals, herbicides and pesticides, refrigerants, polymeric materials, and many others. In this chapter we will sketch the characteristics of an intelligent, computer-aided tool to support the synthetic search for the desired molecules. With functional groups as the "letters" of an alphabet, automatic and interactive procedures compose and screen classes of potential molecules. The automatic synthesis algorithm defines and searches the space of discrete solutions (molecules) through a hierarchical sequence of the space's representations. At each level of detail, a set of explicit constraints is used to depict restrictions on the structure of molecules that can be generated from various combinations of functional groups. In addition, interval arithmetic is employed to test the satisfaction of design specifications, leading to the elimination of large classes of infeasible molecules. One, though, should never overestimate the effectiveness of search algorithms in locating the desired solution(s). Quite frequently we need to resort to human-driven, abductive jumps. In this chapter we will also describe how the automatic search can become interwoven with effective human–machine interaction. Thus, the resulting computer-aided tool, the *Molecule-Designer*, constitutes a paradigm of an intelligent system with two distinct but integrated and complementary capabilities. Examples on the synthesis of refrigerants, solvents, polymers, and pharmaceuticals will illustrate the logic and features of the design procedures in the *Molecule-Designer*.

I. Introduction

Physical properties have a major impact on the economics of many processes and the viability of many products. The refrigerant in a refrigeration cycle, the working fluid in a power cycle, and the solvent used in an azeotropic distillation all determine the physical and economic feasibility of the corresponding processes. Chemical products such as artificial sweeteners, lubricants, and textiles all must exhibit specific physical properties if they are to be accepted.

Until recently, the identification of compounds possessing desired physical property values required extensive experimental search through vast numbers of candidate molecules. Estimates indicate that 3000–5000 compounds need to be tested before a new, useful pharmaceutical is identified, and 5000–8000 to find a new pesticide (Verloop 1972). Computational estimation of the physical property values for candidate molecules can reduce significantly the need for experimentation. In this spirit, Horvath (1992) published a tremendous thesaurus of techniques and approaches for molecular design, by focusing on the estimation of physical properties from molecular structures. Numerous techniques are available for the computational estimation of thermodynamic properties (Reid *et al.*, 1987), environmental properties (Lyman *et al.*, 1982), polymer properties (van Krevelen 1976), biological activity (Martin, 1978; Hansch and Leo, 1979; Franke, 1984), phase equilibria (Fredenslund *et al.*, 1977), and others. However, these analytic methods can not be directly used in a synthetic manner for the design of molecules, as the following example illustrates:

Consider the design of a molecule whose physical properties, PP_1, PP_2, and PP_3, should have the values α, β, and γ, respectively. If functions f_1, f_2, and f_3 relate the three physical properties to the molecular structure, then

$$\begin{aligned} PP_1 &= f_1(\textit{molecular structure}) = \alpha, \\ PP_2 &= f_2(\textit{molecular structure}) = \beta, \\ PP_3 &= f_3(\textit{molecular structure}) = \gamma. \end{aligned} \tag{1}$$

The molecular structure of the unknown chemical could be found by inverting these three relationships. However, an explicit inversion is not analytic (the molecular structure is described by integer variables denoting the presence or absence of specific atoms and bonds), and it accepts multiple solutions (there may be several molecules satisfying the constraints). Implicit inversion of Eqs. (1) is possible through the formulation of appropriate optimization problems. However, in such cases the complexity and nonlinear character of the functional relationships used to estimate the values of physical properties in conjunction with the integer variables description of molecular structures, yield very complex mixed-integer optimization formulations.

Thus it is not surprising that the first efforts in systematically designing molecules possessing desired physical properties were heuristic in character, focusing on specific classes of chemical products. Godfrey (1972) used an empirical miscibility scale to determine whether two liquids were miscible. Francis (1944) used critical solution temperatures to choose solvents for the selective extraction of hydrocarbons. Berg (1969) developed a hydrogen bond classification scheme used to identify azeotropic distillation solvents. In recent years, the sophistication of the methods has improved (Venkatasubramanian *et al.*, 1994; Constantinou *et al.*, 1994; Gani *et al.*, 1991; Gani and Fredenslund, 1993; Nielsen, *et al.*, 1995), but the basic character of the various approaches has remained the same, specifically, all design-oriented techniques, (1) are *problem-specific*, i.e., for solvents, polymers, or pharmaceuticals; (2) are based on the *generate-and-test* paradigm with experiential heuristics employed to reduce the search space of potential alternatives; and (3) cannot utilize efficiently all available knowledge in a given area of molecular design.

A. Brief Review of Previous Work

Research findings from the areas of physical property estimation and chemical products selection are applicable to the design of molecules. A brief review of research in these areas is presented along with previous work in molecular design.

1. Estimation of Physical Properties

Every approach developed for the design of molecules with desired properties requires the estimation of physical properties. Some property estimation techniques lend themselves easily to the identification of the requisite molecular structures, starting from the desired design specs, and others do not. Depending on how various approaches attempt to relate molecular structure to physical properties, they can be grouped into five categories: pattern recognition, topological, group contribution, equation-oriented, and molecular-modeling-based techniques.

a. Pattern Recognition. Discriminant analysis and classification are the two statistical techniques used most often in pattern recognition. Both are multivariate techniques concerned with *separating* distinct sets of objects into classes and *allocating* new objects to previously defined classes. A discriminant function is developed from a set of experimental data called the "training set." This function is then used to classify new compounds. In many applications the number of classes equals 2; e.g., carcinogenic or noncarcinogenic, toxic or nontoxic.

Pattern recognition techniques lend themselves to synthetic designs of molecules for those problems for which the design specs require that a molecule is a member of a certain class.

b. Topological Techniques. These techniques ignore the actual three-dimensional shape of a molecule, the nature and lengths of the chemical bonds connecting its atoms, the angles between the bonds, and sometimes even atom types (Rouvray, 1986). Typically only the number of atoms and their interconnections are considered. This information is reduced to an *index* such as the Wiener Path Number (Wiener, 1947), Alternburg Polynomial Index (Alternburg, 1966), Gordon–Scantlebury Index (Gordon and Scantlebury, 1964), Hosoya's *Z* Index (Hosoya and Murakami, 1975), or Randić's Branching Index (Randić, 1975). These indices are then used to correlate the values of physical properties. The applicability of these techniques for property estimation has been extensively discussed in Kier and Hall (1986). By their nature, topological techniques require detailed information about the molecular connectivity of a compound and are difficult to incorporate into synthetic design procedures.

c. Group Contribution Techniques. These assume that each fragment of a molecule contributes a certain amount to the value of its physical properties. Contributions for each group are statistically regressed from large sets of experimental data. Techniques can become very complex, including nonlinear effects and interactions among groups. They are very appropriate for the design of molecules with desired physical properties and they constitute the basis for both *interactive* and *automatic* design procedures described in this chapter. Group contribution techniques have been used by other researchers for molecular design, as we will see in the next section.

d. Equation-Oriented Techniques. These techniques correlate estimated physical properties to properties more easily available or measured using empirical or theoretical models. Not relating a compound's molecular structure to its properties, these techniques cannot be used directly for molecular design. However, used in conjunction with group contribution techniques, rendering the values of the correlated physical properties, they broaden the list of physical properties which yield the specifications of the desired molecule.

e. Molecular-Modeling-Based Techniques. These techniques start with an atomic model of a molecule and use quantum and statistical mechanics to estimate its physical properties. Many of these estimates are more accurate than those obtained by any other estimation technique. Molecular-modeling-based techniques offer fairly complex and implicit relationships between molecular structure and physical properties. A straightforward generate and test is the only way such techniques are employed.

2. *Selection of Desired Chemicals*

Selecting a chemical product from a set of candidates is a two-step procedure. The first step is the most critical and involves the identification of those physical properties that are important to the performance of the chemical product and their values that give optimal performance. The second step involves a search through a database for existing compounds that possess these physical properties values. Unknown property values must be estimated. Such an approach has the advantage of being fast. Additionally, compounds in the database are, typically, commercially available or can be readily synthesized. The drawback of such an approach is that new compounds cannot be found.

3. *Design of a Desired Chemical*

This is also a two-step procedure similar to that of selecting a desired chemical from a list of candidate molecules. Unlike the selection from an existing set of compounds, the design of compounds implies the synthetic stipulation of molecules for which there are no experimental data of their physical properties. Therefore, using some of the available estimation techniques, different approaches have been proposed in the past and will be discussed in the following paragraphs.

a. Design of Solvents. Gani and Brignole (1983) and Brignole *et al.* (1986) used the UNIFAC (Fredenslund *et al.*, 1977) group contribution method to synthesize molecular structures with specific solvent properties for separation processes. Their synthesis procedure is divided into three steps:

1. Select the groups considered to be suitable building blocks for the molecular structures.
2. Combine the groups into candidate molecules according to specified combination rules.
3. Screen the candidate molecules using UNIFAC to evaluate their usefulness for a particular separation task.

To reduce the number of potential group combinations to a tractable number, several additional constraints are placed on the candidate solvents. For example, a high boiling point is required in order to facilitate simple separation of the solvent by distillation. In similar spirit are the works of Gani *et al.* (1991) and Gani and Fredenslund (1993), but the efficiency of search has improved with heuristic knowledge, while techniques for discrete optimization have been used to design optimal solvent mixtures. The works of Macchietto *et al.* (1990) and Odele and Macchietto (1993) have focused on the selection (rather than design) of optimal solvents for extractive separation processes.

b. Design of Polymers. Derringer and Markham (1985) proposed a generate and test methodology for designing polymers possessing desired physical properties, using the van Krevelen (1976) group contribution estimation techniques. Recognizing that the number of candidate polymers may be large, Derringer and Markham devised a ranking procedure that includes a desirability measure for each predicted property.

c. Design of Polymer Coatings. Tortorello and Kinsella (1983a, b) used the solubility parameter concept to design high-performance aircraft coatings resistant to water, fuels, hydraulic fluids, and lubricating oils. Additional design specs included resistance to high temperature and flexibility at low temperature.

d. Design of Drugs. Drug design has been the most active area for the development of systematic procedures to identify new chemical products. Beginning with a small set of experimental data on the efficacy of candidate compounds, one derives a statistical relationship between the drug's potency and a set of physicochemical properties such as Hammett's constant (1935), Taft's (1956) steric parameter, and the octanol–water partition coefficient, whose use was made popular by Hansch and coworkers (1963). These physicochemical properties are then related to structural characteristics. Such statistical relationships are called *quantitative structure activity relationships* (QSARs).

QSARs provide great insight into the drug design problem. Examination of the derived relationships often indicates how a drug's potency is being affected by reactive, transport, and steric considerations. To improve the potency, a drug designer can search for substituents, which when added to the candidate molecule will reduce steric hindrances or increase the rate of transport. Extensive tabulations exist (Hansch and Leo, 1979) listing the effect that specific substituents have on the various physical-chemical properties, typically used in drug designs.

B. General Framework for the Design of Molecules

The previous works on selecting and designing molecules with desired properties share certain common characteristics. Beginning with these characteristics, a general methodology was developed for designing molecules (Joback and Stephanopoulos, 1990). The overall philosophy of the design methodology is described in the following six paragraphs.

1. Problem Formulation

The first step in any design is to identify the target (Stephanopoulos and Townsend, 1986). A molecular design target consists of physical property, chemical, and structural constraints. Physical property constraints are typically concerned with the performance of the chemical product, such as its ability to perform as an aircraft coating. Chemical constraints are often related directly to molecular structure, restricting or requiring the occurrence of functional groups, such as the desire to design a diol with desired properties. Structural constraints are required when a molecule is constructed by assembling functional groups. The groups must have the correct type and occurrence of bonds so that they can be assembled into feasible molecules. Brignole *et al.* (1986) developed an extensive set of rules to constrain the choice of groups and ensure the structural feasibility of the resulting molecule. The effective formulation of the molecular design problem is crucial to the success of the design.

2. Target Transformation

For the computer to evaluate the performance of a candidate molecule, it must be able to estimate the values for those physical properties identified during problem formulation. The target transformation step develops *estimation procedures*, which enable the evaluation of the target constraints' physical properties in terms of the values of new physical properties (i.e., the transformed target). For example, if we want to design a molecule with vapor pressure P_{vp}, in a given range of values, we can use the Riedel–Plank–Miller correlation [Eq. (2)] and transform the target into these new properties, as shown by Equations (3a), (3b), and (3c):

$$P_{\mathrm{vp}} = P_{\mathrm{vp}}(T_{\mathrm{b}}, T_{\mathrm{c}}, P_{\mathrm{c}}) \qquad \text{(correlation by Riedel–Plank–Miller)}, \quad (2)$$

where

$$T_\mathrm{b} = T_\mathrm{b}(\text{molecular structure})$$

$$\text{(group-contribution technique by Joback)}, \quad (3a)$$

$$P_\mathrm{c} = P_\mathrm{c}(\text{molecular structure})$$

$$\text{(group-contribution technique by Lydersen)}, \quad (3b)$$

$$T_\mathrm{c} = T_\mathrm{c}(\text{molecular structure})$$

$$\text{(group-contribution technique by Fedors)}. \quad (3c)$$

Starting with a compound's molecular structure, the T_b, P_c, and T_c are estimated first, using the three group contribution techniques indicated above. These values are then used in an equation-oriented technique to yield the final estimate for P_vp.

3. *Design Procedure*

The previous design approaches are based on the generate-and-test paradigm. This paradigm consists of two parts: the *generator* and the *tester*. The generator enumerates candidate solutions, whereas the tester evaluates each candidate and either accepts or rejects it. When the number of candidates becomes very large, exhaustive enumeration becomes impractical. Although not all previous molecular design approaches have explored them, several strategies are available to manage the search space (Hayes-Roth *et al.*, 1983), such as (1) move the tester into the generator, (2) prune partial solutions, and (3) abstract the search space. The last strategy is extremely powerful in managing the combinatorics of the design problem and constitutes the basis of the automatic design methodology to be discussed in Section II of this chapter.

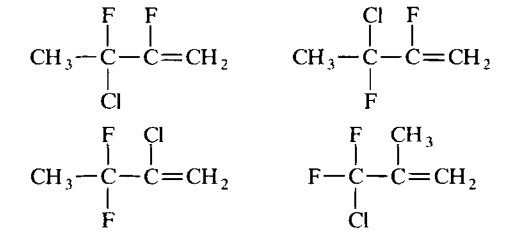
4. *Representation and Enumeration of Molecules*

Designing molecules through the use of group-contribution estimation techniques results in candidate molecules which are represented as a collection of functional groups. To form complete molecules, it is necessary to connect these groups together. At times more than one way of connecting the groups is possible. For example, the following collection of groups

$$-\mathrm{CH_3} \quad >\mathrm{C}< \quad -\mathrm{F} \quad -\mathrm{F} \quad -\mathrm{Cl} \quad >\mathrm{C}= \quad =\mathrm{CH_2}$$

TABLE I

FOUR ENUMERATED MOLECULES



can be combined, ignoring stereoisomers, to form the four different molecules shown in Table I. Molecule enumeration has been extensively investigated by researchers doing work in structure elucidation (Gray, 1986). Before more rigorous estimation techniques, such as molecular modeling, or chemical constraints can be used, it is necessary to provide techniques for the representation and enumeration of all potential molecular structures.

5. *Screening of Molecules*

Chemical constraints are applied once the satisfactory candidate molecules have been enumerated. Typically, chemical constraints prevent the generation of unstable substructures within the structure of generated molecules. For example, if the substructure —O—O— occurs in a compound desired to be stable, that compound is pruned. For design procedures based on group-contribution techniques, the application of chemical constraints must occur after enumeration, since the relative locations of the various groups within a molecule remain unspecified at earlier stages.

6. *Final Evaluation*

It is sometimes necessary to modify the physical properties estimation techniques employed by generate-and-test design procedures. Often this modification is introduced in order to remove computational steps in the estimation techniques that require knowledge of the global molecular structure. Using groups as the design basis, only the partial and local structure of a molecule is known during the design. However, once the candidate molecules are enumerated and screened, global molecular structure is known and more accurate estimation techniques can be used to further prune the candidates.

II. Automatic Synthesis of New Molecules

This section presents an algorithmic strategy for the automatic generation of molecules and their screening against a set of constraints, which represent the physical properties' values that the desired molecules should satisfy. *Functional groups* are the essential building blocks for the construction of molecules, allowing the use of *group-contribution estimation techniques* for the testing of physical property constraints. These techniques are simple, fast, and yield estimates of sufficient accuracy for preliminary screening purposes. The use of more complicated estimation techniques, such as molecular modeling, is unwise at this stage of design. The probability of an ab initio automatic identification of satisfactory molecules at this stage is low, implying that the effort expended for the examination of each candidate molecule should be kept to a minimum. Thus, a hierarchical approach has been adopted for the synthesis of desired molecules:

Phase 1. Simple estimation techniques are used to rapidly screen large number of candidate molecules, generated by a hierarchical search algorithm.

Phase 2. Candidate molecules satisfying the design constraints are reported to the human designer, who orders them using subjective preferences.

Phase 3. The retained molecules are evaluated through the use of more detailed estimation techniques, e.g. molecular modeling, complex equations of state.

In this chapter we will deal only with phase 1.

A. The Generate-and-Test Paradigm

The generate-and-test search paradigm, used by the automatic design algorithm for the synthesis of molecules, is composed of two modules. The first module, the *generator*, enumerates candidate molecules. The second, the *tester*, evaluates each molecule, estimating its physical properties and checking for structural feasibility, and either accepts or rejects it. We represent molecules as collections of groups, e.g., chloropropane is represented as ($-$Cl $-CH_3$ $-CH_2-$ $-CH_2-$). The generator simply constructs candidate molecules by selecting a collection of groups from an initial set. The representation of groups allows the tester to use group-contribution techniques to estimate physical property constraints and check the design constraints.

The combinations of groups that can be selected is infinite. However, from practical considerations, molecules for a typical application fall within some size range, which can be translated into an upper limit on the number of groups chosen. For example, refrigerants are generally of a small molecular weight. Placing a limit of 15 on the number of groups that can be used to form a molecule is a reasonable bound. A lower limit is established from the fact that at least two groups must be used to form a structurally feasible molecule. With limits on the minimum and maximum number of groups that can be chosen, the generator selects collections of groups beginning with all combinations of two groups, then all combinations of three groups, and so on until the upper limit is reached.

1. Design Constraints

The tester module checks each candidate molecule for satisfaction of the design constraints. Three types of constraints are used: (a) physical property constraints, (b) structural constraints, and (c) chemical constraints.

a. Physical Property Constraints. Estimation procedures are established for each physical property used in the design constraints. An estimation procedure is a collection of estimation techniques that determine physical property values when only the molecular structure is known. It is composed of group-contribution and equation-oriented estimation techniques. For example, using the constraint on vapor pressure, we obtain

$$P_{vp}(273\ \mathrm{K}) > 1.01\ \mathrm{bar}.$$

We need to employ an estimation procedure for P_{vp} and the associated independent physical properties, such as those described in Section I.B [see Eqs. (2) and (3a–c)], in order to determine each candidate molecule's vapor pressure at 273 K. Starting with a compound's molecular structure, T_b, T_{br}, and P_c are first estimated using the three group-contribution techniques [Eqs. (3a), (3b), and (3c), respectively]. These values are then used in an equation-oriented technique to yield the final estimate for P_{vp} [Eq. (2)].

b. Structural Constraints. Structural constraints determine whether a collection of groups can be connected in some manner to form a feasible molecule. Three requirements define the conditions for structural feasibility:

1. All groups in the collection should be able to be joined into a single connected component. The collection of groups (—F —F —F —F) is not feasible because it does not form a single molecule.
2. The single connected molecule formed from a set of groups cannot have any unconnected bonds. Connecting the groups (—CH_2— —F) gives us a single structured entity with one single bond unconnected.
3. The connections made in the single connected entity, formed from a set of groups, must all be between bonds of the same type. Single bonds may only connect with single bonds, double with double, etc.

Given a set of groups, it is possible to enumerate all ways in which they could be connected verifying that at least one candidate satisfies the three requirements. However, a graph theoretic examination of molecular structures provides a set of structural constraints that are much easier to apply. Such structural constraints have been developed and are shown in Table II.

c. Chemical Constraints. Chemical constraints are heavily dependent on the global connection of atoms within a molecule. Representing molecules as collections of groups does not provide knowledge about global connectivity. Chemical constraints are thus better used at later stages of the search methodology, where the complete structure of a molecule is known.

2. *Combinatorial Explosion*

Given a reasonable number of groups from which molecules can be constructed, the number of candidates that can be generated is extremely large. Allowing repetition of groups and ignoring the order of selection, the number of candidate molecules that can be generated by selecting n groups from a set of k groups is given by Eq. (4):

$$C^{\mathrm{R}}(k,n) = \frac{(k+n-1)!}{n!(k-1)!}. \tag{4}$$

The total number of candidate molecules that can be selected from a set of k groups in which each candidate molecule has between 2 and $n_{\max}$

TABLE II

STRUCTURAL CONSTRAINTS ON FORMING FEASIBLE MOLECULES

1. If G is a collection of n groups, then $n \geq 2$.
2. If G is a collection of n groups with n_c cyclic groups, n_m mixed groups, and n_a acyclic groups, and $n_a > 0$ and $n_c > 0$, then $n_m > 0$.
3. If G is a collection of n groups with n_c cyclic groups, n_m mixed groups, and n_a acyclic groups, then $n_m > 0$ implies that $n_a > 0$ or $n_c > 0$.
4. If G is a collection of n groups with n_c cyclic groups and n_m mixed groups, then either $n_c + n_m \geq 3$ or $n_c + n_m \geq 2$.
5. If G is a collection of groups, then the number of groups having an odd number of free bonds must be even.
6. If G is a collection of n groups with b free bonds, then $\frac{b}{2} \geq n - 1$.
7. If G is a collection of n groups with b free bonds, then $\frac{b}{2} \leq \frac{1}{2}n(n-1)$.
8. If a collection of groups contains more than one bond type, then there must be a transition group containing each bond type. A transition group is one that contains more than one bond type.
9. If G is a collection of groups with $n_{a,i}$ denoting the number of acyclic groups with a valence i and v_{mj} denoting the valence of some jth mixed group, then $n_1 \leq \Sigma_{\text{mixed}}(v_{m,j} - 2) + n_{a,3} + 2n_{a,4} + \cdots + (i-2)n_{a,i} + \cdots$.
10. If G is a collection of n groups with n_i denoting the number of groups with a global valence i and all n groups are acyclic, then $n_1 = 2 + n_3 + 2n_4 + \cdots + (i-2)n_i + \cdots$.
11. The number of occurrences of each bond type in a collection of groups must be even.

groups is given by Eq. (5):

$$\text{Total candidates} = \sum_{n=2}^{n_{\max}} C^{R}(k,n) = \sum_{n=2}^{n_{\max}} \frac{(k+n-1)!}{n!(k-1)!}. \tag{5}$$

Table III shows how this total number of candidates can quickly grow to very large values. Managing this combinatorial explosion is the major focus of the automatic design algorithm.

TABLE III

COMBINATORICS OF GROUP SELECTION
($k = 40$ GROUPS)

$n_{\max}$	# Molecules
4	135,710
5	1,221,718
6	9,366,778
7	62,891,458
8	377,348,953
9	2,054,455,593

TABLE IV

Initial Set of Groups

$>CH_3$	$-CH_2-$	$>CH-$	$>C<$
$=CH_2$	$=CH-$	$=C<$	$=C=$
$\equiv CH$	$\equiv C-$	$-F$	$-Cl$
$-Br$	$-I$	$-OH$	$-O-$
$>CO$	$-CHO$	$-COOH$	$-COO-$
$=O$	$-NH_2$	$>NH$	$>N-$
$-CN$	$-NO_2$	$-SH$	$-S-$

B. The Search Algorithm

The magnitude of the combinatorial problem, resulting from a large number of functional groups, can be reduced only by reducing the number of groups. This reduction is done by abstracting the groups into families of groups, called *metagroups*. Table V shows the groups of Table IV clustered into four metagroups.

Instead of generating molecules by choosing from an initial set of groups, we choose from an initial set of metagroups. The candidate molecules formed from a collection of metagroups are called *metamolecules*, which are sets of molecules. Using the metagroups from Table V, the metamolecule (2 1 0 0) is the set of all molecules that can be formed by taking any two groups from *metagroup 1* and any one group

TABLE V

Example Metagroups

Metagroup 1	$-CH_3$	$=CH_2$	$\equiv CH$	$-F$	$-Cl$	$-Br$
	$-I$	$-OH$	$-CHO$	$-COOH$	$=O$	$-NH_2$
	$-NO_2$	$-CN$	$-SH$			
Metagroup 2	$>CH_2$	$=CH-$	$=C=$	$\equiv C-$	$>CO$	$-COO-$
	$-O-$	$>NH$	$-S-$			
Metagroup 3	$=C<$	$>CH-$	$>N-$			
Metagroup 4	$>C<$					

from *metagroup* 2. The number of molecules contained in metamolecule (2 1 0 0) is

$$C^{R}(15,2) \times C^{R}(9,1) = \frac{(15+2-1)!}{2!(15-1)!} \times \frac{(9+1-1)!}{1!(9-1)!},$$

or 1080 molecules.

1. Evaluation of Metamolecules

Abstracting groups into metagroups reduces the combinatorics of candidate molecule generation. However, we must be able to efficiently evaluate whether a metamolecule satisfies the design constraints.

a. Structural Constraints. As long as all the groups within each metagroup have a consistent molecular characteristic such as global valence, the structural constraints are still applicable. Metagroups 1 and 2 are consistent in ring class and global valence. Structural constraint 10 of Table II is thus applicable. Applying the constraint to metamolecule (2 1 0 0) yields

$$2 = 2 + 0 + 2(0) = 2,$$

i.e. the constraint is satisfied. This implies that each of the 1080 molecules contained in (2 1 0 0) satisfies the constraint.

b. Physical Property Constraints. Associated with each of the groups in a metagroup is a contribution toward a particular physical property. The contribution of a metagroup is called a *metacontribution* and is defined by a set of values. Table VI shows the contributions toward the value of the boiling point, T_b, for each of the groups in metagroup 2 (see Table V) toward T_b. The metacontribution of metagroup 2 toward T_b is thus the following set of values:

$$(22.42 \quad 22.88 \quad 24.96 \quad 26.15 \quad 27.38 \quad 50.17 \quad 68.78 \quad 76.75 \quad 81.10).$$

To use metacontributions in the calculation of physical properties, it is necessary to find a representation that can capture the set value of the metacontributions and can be manipulated by mathematical operators. Interval numbers were chosen as the representation.

TABLE VI

T_b GROUP CONTRIBUTIONS FOR METAGROUP 2 (ACYCLIC GROUPS)

Groups	Contribution
$-CH_2-$	22.88
$=CH-$	24.96
$=C=$	26.15
$\equiv C-$	27.38
$-O-$	22.42
$>CO$	76.75
$-COO-$	81.10
$>NH$	50.17
$-S-$	68.78

2. *Interval Arithmetic and Meta-Contributions*

The generalization of ordinary arithmetic to closed intervals is known as *interval arithmetic*. An interval is defined as a closed bounded set of real numbers (Moore, 1979):

$$X = [\underline{X} \quad \overline{X}] = \{x | \underline{X} \le x \le \overline{X}\}. \tag{6}$$

Thus, intervals have a *dual* nature as both a number and a set. The basic interval arithmetic operations are

$$[\underline{X} \quad \overline{X}] + [\underline{Y} \quad \overline{Y}] \equiv [\underline{X} + \underline{Y} \quad \overline{X} + \overline{Y}],$$

$$[\underline{X} \quad \overline{X}] + [\underline{Y} \quad \overline{Y}] \equiv [\underline{X} - \overline{Y} \quad \overline{X} - \underline{Y}],$$

$$[\underline{X} \quad \overline{X}] * [\underline{Y} \quad \overline{Y}] \equiv \Big[\min\big(\underline{X} * \underline{Y}, \quad \underline{X} * \overline{Y}, \quad \overline{X} * \underline{Y}, \quad \overline{Y} * \overline{Y}\big),$$

$$\max\big(\underline{X} * \underline{Y}, \quad \underline{X} * \overline{Y}, \quad \overline{X} * \underline{Y}, \quad \overline{X} * \overline{Y}\big)\Big],$$

$$[\underline{X} \quad \overline{X}] \div [\underline{Y} \quad \overline{Y}] \equiv [\underline{X} \quad \overline{X}] * [1/\overline{Y} \quad 1/\underline{Y}] \quad \text{iff } 0 \notin [\underline{Y} \quad \overline{Y}].$$

The metacontribution of the metagroup 2 (see Table VI) in interval representation is [22.42 81.10]. Thus, we can construct Table VII, which shows the metacontributions for each metagroups, displayed in Table V for boiling point T_b, reduced boiling point T_{br}, and heat of vaporization ΔH_{vb} (Joback and Reid, 1987). Using the group-contribution estimation

TABLE VII

METACONTRIBUTIONS

Metagroup	T_b	T_{br}	ΔH_{vb}
1	[−10.50 169.09]	[0.0027 0.0791]	[−0.670 19.537]
2	[22.42 81.10]	[0.0020 0.0481]	[2.205 9.633]
3	[11.74 24.14]	[0.0117 0.0169]	[1.691 2.138]
4	[18.25 18.25]	[0.0067 0.0067]	[0.636 0.636]

models

$$T_b = 198.18 + \sum_{\text{all groups}} n_i \Delta_{i,T_b}, \tag{7}$$

$$T_{br} = 0.584 + 0.965 \sum_{\text{all groups}} n_i \Delta_{i,T_{br}} - \left(\sum_{\text{all groups}} n_i \Delta_{iT_{br}} \right), \tag{8}$$

$$\Delta H_{vb} = 15.30 + \sum_{\text{all groups}} n_i \Delta_{i,\Delta H_{vb}}, \tag{9}$$

we can estimate the values of T_b, T_{br}, and ΔH_{vb} for metamolecule (2 1 0 0) as follows:

$$\begin{aligned} T_b &= 198.18 + 2[-10.50 \quad 169.09] + [22.42 \quad 81.10] \\ &= [199.6 \quad 617.46]\text{K}, \end{aligned} \tag{10}$$

$$\begin{aligned} T_{br} &= 0.584 + 0.965(2[0.0027 \quad 0.0791] + [0.0020 \quad 0.0481]) \\ &\quad - (2[0.0027 \quad 0.0791] + [0.0020 \quad 0.0481])^2 \\ &= [0.549 \quad 0.783], \end{aligned} \tag{11}$$

$$\begin{aligned} \Delta H_{vb} &= 15.30 + 2[-0.670 \quad 19.537] + [2.205 \quad 9.633] \\ &= [16.165 \quad 64.007] \text{ kJ/mol}. \end{aligned} \tag{12}$$

The intervals given by Eqs. (10)–(12) span the range of physical property values possessed by each of the 1080 molecules in metamolecule (2 1 0 0).

Interval values for these fundamental properties can be used in equation-oriented estimation techniques. The Watson relation (Watson, 1943)

$$\Delta H_v = \Delta H_{vb} \left(\frac{1 - T/T_c}{1 - T_{br}} \right)^{0.38} \tag{13}$$

is used to estimate the enthalpy of vaporization at 250 K for metamolecule

(2 1 0 0), where T_c is obtained from

$$T_c = \frac{T_b}{T_{br}} = \frac{[199.6 \quad 617.46]}{[0.549 \quad 0.783]} = [254.9 \quad 1124.7].$$

Inserting the value of T_c into Eq. (13), we obtain the interval value of the enthalpy of vaporization:

$$\Delta H_v = [16.165 \quad 64.007]\left(\frac{1 - 250/[254.9 \quad 1124.7]}{1 - [0.549 \quad 0.783]}\right)^{0.38}$$

$$= [0.688 \quad 229.40]\ \text{kJ/mol}.$$

3. Searching through Successive Molecular Abstractions

The generate-and-test search paradigm described earlier must now be modified to deal with the abstractions introduced by the metamolecules. Instead of generating and testing individual molecules, we generate and test metamolecules. Those metamolecules satisfying the test are reduced in abstraction, by dividing a metagroup into more meta-groups. This refinement produces a new generation of metamolecules that are retested.

Let us demonstrate the logic of the procedure using the metagroups of Table V and the meta-contributions of Table VII. Consider the following constraint on boiling point: $T_b > 500$ K.

Limiting the number of groups contained in a molecule between 2 and 4, the following 65 metamolecules are generated:

(2 0 0 0) (0 2 0 0) (0 0 2 0) (0 0 0 2) (1 1 0 0)
(1 0 1 0) (1 0 0 1) (0 1 1 0) (0 1 0 1) (0 0 1 1)
(3 0 0 0) (0 3 0 0) (0 0 3 0) (0 0 0 3) (2 1 0 0)
(2 0 1 0) (2 0 0 1) (0 2 1 0) (0 2 0 1) (0 0 2 1)
(1 2 0 0) (1 0 2 0) (1 0 0 2) (0 1 2 0) (0 1 0 2)
(0 0 1 2) (1 1 1 0) (1 1 0 1) (1 0 1 1) (0 1 1 1)
(4 0 0 0) (0 4 0 0) (0 0 4 0) (0 0 0 4) (3 1 0 0)
(3 0 1 0) (3 0 0 1) (0 3 1 0) (0 3 0 1) (0 0 3 1)
(1 3 0 0) (1 0 3 0) (1 0 0 3) (0 1 3 0) (0 1 0 3)
(0 0 1 3) (2 2 0 0) (2 0 2 0) (2 0 0 2) (0 2 2 0)
(0 2 0 2) (0 0 2 2) (2 1 1 0) (2 1 0 1) (2 0 1 1)
(0 2 1 1) (1 2 1 0) (1 2 0 1) (1 0 2 1) (0 1 2 1)
(1 1 2 0) (1 1 0 2) (1 0 1 2) (0 1 1 2) (1 1 1 1)

TABLE VIII

T_b VALUES FOR FOUR METAMOLECULES

Metamolecule	T_b
(2 0 0 0)	[177.18 536.36]
(2 1 0 0)	[199.60 617.46]
(3 0 1 0)	[178.42 729.59]
(2 2 0 0)	[222.02 698.56]

Recall that the metamolecule (1 0 1 2) represents the set of all molecules that can be formed by taking any one group from metagroup 1, no group from metagroup 2, any one group from metagroup 3, and any two groups from metagroup 4.

Structural constraint 10 of Table II is used to prune the candidate metamolecules. The maximum valence any group has is 4. Therefore, each metamolecule is checked to ensure that it satisfies the constraint $n_1 = 2 + n_3 + 2n_4$; 61 metamolecules are pruned using this constraint. The remaining four metamolecules are

$$(2\,0\,0\,0) \quad (2\,1\,0\,0) \quad (3\,0\,1\,0) \quad (2\,2\,0\,0).$$

Physical constraints are applied next. The boiling point value T_b was estimated using the metacontributions of Table VII and Eq. (7). Table VIII shows these estimates for each of the remaining four metamolecules. These values show that all four metamolecules satisfy the constraint on boiling point T_b.

The next step of the search algorithm is to reduce the level of abstraction. Groups were abstracted into metagroups to reduce the combinatorics of the problem. However, this same abstraction reduced the effectiveness of property constraints. As the abstraction is reduced, this effectiveness is regained. Metagroup 1 is divided into two new metagroups:

$$\begin{matrix} -CH_3 & =CH_2 & =CH & -F & -Cl & -Br \\ -I & -OH & -CHO & =O & -NH_2 & -SH \end{matrix}, \quad \text{(metagroup 1,1)}$$

$$\{-COOH \quad -NO_2 \quad -CN\}. \quad \text{(metagroup 1,2)}$$

This division of metagroup 1 is propagated to the metamolecules. Metamolecule (2 0 0 0) was the set of all molecules that could be formed by taking any two groups from metagroup 1. With metagroup 1 divided into

TABLE IX

T_b VALUES FOR 13 METAMOLECULES

[(2 0) 0 0 0]	[177.18 385.86]
[(0 2) 0 0 0]	[449.50 536.36]
[(1 1) 0 0 0]	[313.34 461.11]
[(2 0) 1 0 0]	[199.60 466.96]
[(0 2) 1 0 0]	[471.92 617.46]
[(1 1) 1 0 0]	[335.76 542.21]
[(3 0) 0 1 0]	[178.42 503.84]
[(0 3) 0 1 0]	[586.90 729.59]
[(2 1) 0 1 0]	[314.58 579.09]
[(1 2) 0 1 0]	[450.74 654.34]
[(2 0) 2 0 0]	[222.02 548.06]
[(0 2) 2 0 0]	[494.34 698.56]
[(1 1) 2 0 0]	[358.18 623.31]

two new metagroups there are three possibilities:

1. Take any two groups from metagroup 1,1.
2. Take any two groups from metagroup 1,2.
3. Take any one group from metagroup 1,1 and any one group from metagroup 1,2.

These possibilities correspond to an expansion of the metamolecule (2 0 0 0) into three new metamolecules:

$$[(2\,0)\;0\,0\,0]\quad[(0\,2)\;0\,0\,0]\quad[(1\,1)\;0\,0\,0].$$

Table IX displays the 13 new meta-molecules resulting from the expansion of all four metamolecules.

The metacontributions toward T_b are also divided; e.g.

$$T_b \text{ of metagroup } 1,1 = [-10.50 \quad 93.84],$$

$$T_b \text{ of metagroup } 1,2 = [125.66 \quad 169.09].$$

T_b is estimated for each meta-molecule and the property constraint is applied. Table IX shows estimated T_b values for the 13 metamolecules.

Applying the property constraint prunes metamolecules [(2 0) 0 0 0] [(1 1) 0 0 0] and [(2 0) 1 0 0]. Additionally, the estimate of T_b for metamolecule (0 3 0 1 0) shows that all the molecules it contains have T_b values that satisfy the property constraint. None of the metamolecules resulting from further expansion of metamolecule [(0 3) 0 1 0] need to be checked.

The search continues with the expansion of metagroups until all metagroups contain only one group. At that point the abstraction has been removed, and the metamolecules generated represent individual molecules.

4. *Strategies for the Formation of Molecular Abstractions*

Metagroup division can be accomplished in many ways. Given a set of k objects, the number of ways these can be portioned into p sets is given by

$$S(k, p),$$

which is the Stirling number of the second kind. Starting with a set of hypothetical groups, $(a\ b\ c\ d)$, and dividing them into two metagroups, yields $S(4, 2) = 7$ possibilities. These are $[(a)(bcd)]$ $[(b)(acd)]$ $[(c)(abd)]$ $[(d)(abc)]$ $[(ab)(cd)]$ $[(ac)(bd)]$ $[(ad)(bc)]$. For a reasonable number of groups, the possible choices of metagroups is very large.

Two approaches can be used to add back detail: *expansion* and *division*. Expansion adds back knowledge about the metagroups, which is used by the structural constraints. Division focuses on reducing the width of metacontributions, thus improving the screening power of physical property constraints.

a. Division In Half. Dividing a metagroup in half is the simplest strategy. However, division without regard to the metacontributions could prove inefficient. Given the following set of groups with the corresponding contributions

Groups	$[g_1\quad g_2\quad g_3\quad g_4]$,
Contributions	$[10\quad 60\quad 11\quad 61]$,

our initial meta-group $[g_1\quad g_2\quad g_3\quad g_4]$ would have a metacontribution of $[10\quad 61]$. Dividing the meta-group in half would result in the two metagroups $[g_1\quad g_2]$ and $[g_3\quad g_4]$. The metacontributions for these new metagroups would be $[10\quad 60]$ and $[11\quad 61]$. These metacontributions are almost identical to the original. The division thus added to the combinatorics without improving the possibility for pruning.

Meta-contributions should be considered when dividing metagroups in half. The midpoint of the initial metacontribution is $(61 - 10)/2 = 25.5$. All groups whose contributions are less than $25.5 + 10 = 35.5$ are collected into the first new metagroup, and all those with contributions greater than 35.5 are collected into the second new metagroup. The resulting metagroups have tighter interval representation of their contributions, leading to more efficient screening.

b. Division by Largest Gap. The interval representation of metacontributions ignores their discrete nature. Dividing a metagroup at the largest gap in its contributions attempts to take advantage of the distribution of contributions and produce two new metagroups whose metacontributions are distributed over a much more narrow range than the original metacontribution. Using the same example set of groups and contributions given above, the initial metacontribution, [10 61], has a width of 51. Dividing the metagroup at the largest gap in the contributions produces two new metagroups with metacontributions [10 11] and [60 61]. The total width of these two intervals is 2, a considerable reduction from 51.

c. Division to Isolate Groups. Extreme values of the contributions by some groups can greatly affect the interval value of the calculated properties. The contribution toward the boiling point T_b, from Joback's method (Joback and Reid, 1987) for the group $=O$, is -10.5. If we are searching for low values of T_b, then it is desirable to have many $=O$ groups in our molecules. However, from chemical considerations it is unlikely that a molecule with a large number of $=O$ groups would be stable. Isolating $=O$ onto its own metagroup enables the designer to pose constraints on the maximum number of occurrences of the metagroup in any metamolecule.

5. *Evaluation of the Search Algorithm: Taming the Combinatorial Explosion*

Tables X and XI summarize the results of the application of the search algorithm in two case studies. Let us look at the highlights of each one of them:

Case 1. In this case study we want to synthesize molecules that have a vapor pressure, at 273 K, larger than 1.0 bar. The molecules are to be composed from a set of 44 functional groups, and they can contain up to 3 functional groups, i.e., $k = 44$ and $n = 3$. There exist 15,180 molecules that can be created from various combinations of the functional groups (Table X). The search algorithm generates 460 metamolecules and rejects 104 of them. The refinement of the metagroups and the pruning of infeasible molecules is guided by a series of constraints, as indicated in the footnotes of Table X.

TABLE X

PRUNING RESULTS FOR $k = 44$, $n = 3$ AUTOMATIC DESIGN [CONSTRAINT = $P_{vp}(273\text{ K}) > 1.0$ BAR]

# Metagroups	# Metamolecules	Kept	Pruned
1	1	1	0
3[a]	10	4	6
4[b]	7	4	3
10[c]	51	1	50
11	3	1	2
12[d]	3	2	1
13	5	3	2
14	6	3	3
15	6	4	2
16	8	7	1
17	11	11	0
18	12	11	1
19	21	18	3
20	28	26	2
21	33	29	4
22	44	44	0
27	109	85	24
28[e]	102	102	0
Total[f]	460	356	104

[a] Expanded by ring class.
[b] Isolated —COOH, —NO_2, and —CN.
[c] Expanded by global valence.
[d] Isolated =O. Restricted =O occurrences to 1.
[e] 12 metagroups never occurred in any metamolecules.
[f] There are 15,180 molecules contained in the search.

Case 2: The premises are the same as in case 1, but here we allow the formation of molecules with up to five functional groups. The total number of potential molecules is 1,712,304. The search algorithm has generated 4131 metamolecules and rejected 2,094 of them (Table XI). See footnotes of Table XI for constraints guiding the pruning of infeasible metamolecules.

From both these examples is clear the advantage of using a search with successive molecular abstractions; *the number of metamolecules needed to be evaluated is far smaller than the number of individual molecules.*

The algorithm was also analyzed in order to identify some of its "bounding" properties. Assume that an initial metagroup is divided into two children metagroups. The metamolecules formed from these metagroups are tested, and all those that contain occurrences of the second

TABLE XI

PRUNING RESULTS FOR $k = 44$, $n = 5$ AUTOMATIC DESIGN [CONSTRAINT $= P_{\mathrm{vp}}(273\ \mathrm{K}) > 1.0$ BAR]

# Metagroups	# Metamolecules	Kept	Pruned
1	1	1	0
3[a]	21	8	16
4[b]	23	8	15
5[c]	23	12	11
6[d]	30	27	3
7	58	27	31
8	58	32	26
9	37	35	2
10	53	35	18
11	71	42	29
12	87	86	1
13	110	86	24
14	165	107	58
15	206	179	27
24[e]	1675	185	1490
30[f]	625	479	146
37[g]	888	688	200
Total[h]:	4131	2037	2094

[a] Expanded by ring class.

[b] Isolated —COOH, —NO$_2$, and —CN.

[c] Isolated $=$O. Restricted $=$O occurrences to a maximum of 1.

[d] Isolated —F.

[e] Expanded by global valence.

[f] Expanded several metagroups containing two or three groups in half.

[g] Expanded all nonzero occurring metagroups to individual groups. Ten metagroups never occurred in any metamolecule.

[h] There are 1,712,304 molecules contained in the search.

group are pruned away. This scenario is repeated with the surviving metagroups, and so on.

Dividing a metagroup (MG) containing k groups into two metagroups, MG$_1$ and MG$_2$, containing k_1 and k_2 groups, respectively, allocates the

$$\frac{(k+n-1)!}{n!(k-1)!}$$

possible molecules into

$$\frac{(2+n-1)!}{n!(2-1)!} = n + 1$$

metamolecules. One metamolecule contains only occurrences of MG_1, one metamolecule contains only occurrences of MG_2, and the remaining $n - 1$ metamolecules contain occurrences of both metagroups. If no metamolecules containing MG_2 survive the testing, then the percentage of molecules pruned is given by

$$\left[1 - \frac{(k-1)!}{(k_1-1)!}\frac{(k_1+n-1)!}{(k+n-1)!}\right]*100\%.$$

Repeating this expansion and pruning process r times, until the final metagroup contains only one group, requires the generation and testing of

$$r(n+1)+1$$

metamolecules. The advantage of abstraction, as measured by the total number of molecules contained in the search divided by the number of metamolecules needed to be generated and tested, is given by

$$\text{Advantage of abstraction} = \frac{1}{r(n+1)+1}\frac{(k+n-1)!}{n!(k-1)!}.$$

Considering the worst-case scenario, in which MG_2 and all subsequent second metagroups contain only a single group, we have $r = k - 1$ leading to

$$\text{Advantage of abstraction} = \frac{1}{(k-1)(n+1)+1}\frac{(k+n-1)!}{n!(k-1)!}.$$

Table XII shows this advantage of abstraction for several values of k and

TABLE XII

ADVANTAGE OF ABSTRACTION

$k\backslash n$	2	3	4	5	6
15	5.6	11.9	43.1	136.8	391.5
20	7.2	20.0	92.2	369.6	1,321.6
25	8.9	30.2	169.2	819.0	3,513.5
30	10.6	42.4	280.3	1,590.0	7,956.7
35	12.2	56.7	431.7	2,808.6	16,060.2
40	13.9	73.1	629.6	4,621.3	29,726.5
45	15.6	91.6	880.5	7,195.8	51,426.2
50	17.2	112.2	1,190.3	10,720.4	84,272.3

n. In particular, Table XII shows that if we have an automatic design involving 40 groups with an occurrence value of 5, then the number of metamolecules needed for exhaustive searching will be 4621.3 times smaller than the number of potential molecules.

C. Case Study: Automatic Design of Refrigerants

Automotive air conditioners are a major source of refrigerant emissions, which contribute to the depletion of the Earth's protective ozone layer. This case study generates replacement refrigerants for automotive air conditioners.

Identifying the target set of constraints is the first step of the methodology. Constraints are derived from performance considerations and in an evolutionary manner attempting to find a compound with properties better than refrigerant 12. The design temperatures between which the refrigerant must operate are 110°F (43.3°C) maximum and 30°F (−1.1°C) minimum (Langley, 1986). The following constraints form the design target:

- $P_{vp}(T = -1.1°\mathrm{C}) > 1.4$ bar
  The lowest pressure in the cycle should be greater than atmospheric (Dossat, 1981). This reduces the possibility of air and moisture leaking into the system. Douglas (1988) recommends a safety factor of 5 psig (pounds per square inch gauge).
- $P_{vp}(T = 43.3°\mathrm{C}) < 14$ bar
  A high system pressure increases the size, weight, and cost of equipment (Dossat, 1981). A pressure ratio of 10 is considered to be the maximum for a refrigeration cycle (Perry and Chilton, 1973).
- $\Delta H_v(T = -1.1°\mathrm{C}) > 18.4$ kJ/g-mol
  The value for refrigerant 12's enthalpy of vaporization at −1.1°C is 18.4 kJ/g-mol [American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), 1972]. A higher value reduces the amount of refrigerant required.
- $C_{p_L}(T = 21.1°\mathrm{C}) < 32.2$ cal/g-mol · K
  It is desirable to have a low liquid heat capacity to reduce the amount of refrigerant that flashes on passage through the expansion valve (Dossat, 1981). Refrigerant 12's liquid heat capacity at 21.1°C is 32.2 cal/g-mol · K (ASHRAE, 1972).

Estimation procedures were developed for each physical property used in the target constraints: P_{vp}, H_v, C_{p_L}. These estimation procedures were based on correlations that require the evaluation of seven physical properties, given by group-contribution techniques:

(a) Reduced boiling point, T_{b_R}
(b) Normal boiling point, T_b
(c) Critical pressure, P_c
(d) Coefficients for cubic feet of ideal gas heat capacity with temperature, $C^o{}_{p,a}, C^o{}_{p,b}, C^o{}_{p,c}, C^o{}_{p,d}$

In the automatic design, 44 functional groups were used; molecules containing 2–7 groups were allowed; the number of group occurrences was limited to 7 to account for the fact that most refrigerants are of small molecular size, and 47 molecules were designed that satisfy the four physical property constraints and the structural constraints listed in Table II. Table XIII shows some of these 47 molecules along with their estimated values for P_{vp}, ΔH_v, and C_{p_L}.

Molecules 19 and 20 are of particular interest. These are two ringed compounds that possess physical properties satisfying the design constraints. Although the chemical stability of these compounds still needs to be verified, it is considered a success that the automatic design was able to design such not obvious compounds.

D. Case Study: Automatic Design of Polymers As Packaging Materials

In this second example we demonstrate the applicability of the automatic design methodology to the design of polymers with desired properties. The problem is to design polymers for use as integrated-circuit (IC) encapsulants.

Electronic packages are sealed to prevent gross contamination, handling damage, and the entry of corrosive gases (Mih, 1984). To package microelectronic circuitry so that it is useful and functions properly under various environmental conditions, it is essential to select the correct packaging materials (Fogiel, 1972). Polymeric coatings are widely used in the electronics industry because of their excellent properties and low cost (Goosey, 1985). Polymers used for semiconductor encapsulation must protect against moisture, chemical agents, wide temperature variations, and mechanical shock. The polymeric material must be able to satisfy these requirements with a minimal effect on device parameters over an

TABLE XIII

REFRIGERANT DESIGN—AUTOMATIC RESULTS

Molecule	P_{vp}(272.05)	P_{vp}(316.45)	H_v(272.05)	C_{p_L}(294.35)
1. 1(—CH$_3$)1(—Cl)	1.59	6.20	21.58	20.00
2. 2(—F)1(>NH)	2.69	10.96	19.06	22.26
3. 1(—Cl)1(—F)1(—CH$_2$—)	1.65	6.61	20.71	22.89
4. 1(≡CH)1(—CH$_3$)1(≡C—)	1.67	6.30	21.44	24.06
5. 2(—F)2(≡C—)	2.09	7.87	19.61	21.97
6. 1(≡CH)1(—F)1(—CH$_2$—) 1(≡C—)	1.74	6.72	20.56	26.95
7. 1(=O)2(—CH$_3$)1(=C<)	1.71	7.29	27.14	30.06
8. 1(—Cl)2(—F)1(>N—)	2.65	10.41	19.04	24.83
9. 1(—Cl)2(—F)1(>CH—)	1.74	6.96	19.45	25.28
10. 2(—CH$_3$)1(—F)1(>N—)	1.81	7.29	20.42	30.04
11. 3(—F)1(>NH)1(>N—)	1.70	8.12	20.88	29.98
12. 1(—CH$_3$)2(—F)1(—O—) 1(>N—)	1.96	8.35	19.70	31.60
13. 1(=CH$_2$)2(—F)1(=CH—) 1(>N—)	2.15	8.61	18.73	30.15
14. 1(=CH$_2$)2(—F)1(—CH$_2$—) 1(=C<)	1.41	5.78	19.59	30.78
15. 1(=CH$_2$)2(—F)1(=CH—) 1(>CH—)	1.42	5.82	19.12	30.62
16. 1(≡CH)2(—F)1(≡C—) 1(>N—)	2.77	10.54	18.90	28.88
17. 1(≡CH)2(—F)1(≡C—) 1(>CH—)	1.83	7.07	19.31	29.34
18. 3(—F)2(=CH—)1(>N—)	1.66	7.13	18.92	31.79
19. 2(${}^{r}_{r}{>}$CH—)1(=C${<}^{r}_{r}$)1(=O) 2(—F)	1.53	7.11	26.08	31.93
20. 3(${}^{r}_{r}{>}$CH—)3(—F)	1.40	5.80	18.67	31.11

extended period of time, and be relatively inexpensive and easy to process. Some of the important physical properties of packaging material are (Dillinger, 1988)

1. Permeability to water vapor at high temperatures.
2. Thermal conductivity.
3. Outgassing in plastics at elevated temperatures and the resulting impact on water vapor permeability.
4. Thermal expansion coefficient and mismatch between expansion coefficients of package and chip interconnect.

The following constraints are the design specifications of a good encapsulant:

- $T_g > 400°\text{C}$
  The glass transition temperature must be $> 400°\text{C}$. The encapsulant must keep its structural integrity during use. The high temperatures at which microelectronic circuits operate place a restriction on T_g.
- $R > 10^{16}$ Ω-cm
  The volume resistivity of the solid polymer must be $> 10^{16}$ Ω-cm. Since the packaging material will make contact with the metal leads of the microelectronics device, it is essential that the compound have a high-volume resistivity.
- $\lambda > 0.16$ w/m · K
  The thermal conductivity of the solid polymer is desired to be greater than the thermal conductivity of the currently used polyimide. A high thermal conductivity is desirable allowing the microelectronic circuitry to be cooled more effectively.
- $P(O_2) < 1.0$ cm^3-mil 100 in.$^{-2}$ day^{-1} atm^{-1}
  The permeability of the polymer to oxygen should be < 1.0 cm^3-mil 100 in.$^{-2}$ day^{-1} atm^{-1}. Diffusion of oxygen and water through the polymer to the microelectronic circuitry could cause corrosion and is thus undesirable. To establish a value for this physical property constraint, we examined polymers used as barriers. Polymers with a permeability to oxygen of ≤ 1.0 cm^3-mil 100 in.$^{-2}$ day^{-1} atm^{-1} are considered high-barrier materials.

Estimation techniques from van Krevelen (1976) and Salame (1986) were combined into estimation procedures for the properties of each target constraint: $T_g, R, \lambda, P(O_2)$. These procedures are detailed in the following paragraphs:

1. *Thermal conductivity*. This is estimated by the following correlation (van Krevelen, 1976)

$$\lambda(298\text{ K}) = \lambda\left(C_p^S, V, U\right),$$

where

C_p^s = C_p^s (groups)—specific heat by the van Krevelen group-contribution technique

V = V(groups)—specific volume by the modified van Krevelen group-contribution technique

U = U(groups)—Rao function by the van Krevelen group-contribution technique

2. *Electrical resistivity*. This is estimated by the following correlation (van Krevelen, 1976)

$$R = R(P_{LL}, V),$$

where

P_{LL} = P_{LL}(groups)—dielectric polarization by the van Krevelen group-contribution technique,

V = V(groups)—specific volume by the modified van Krevelen group-contribution technique.

This estimation procedure requires two properties to be estimated by group-contribution techniques: P_{LL} and V.

3. *Glass transition temperature*. This is estimated by the following correlation (van Krevelen, 1976)

$$T_g = T_g(Y_g, M),$$

where

Y_g = Y_g(groups)—glass transition function by the van Krevelen group-contribution technique,

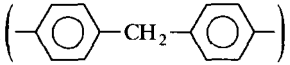
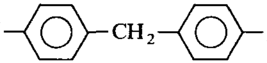
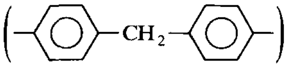
M = M(groups)— repeat unit weight by the van Krevelen group-contribution technique.

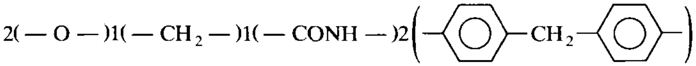
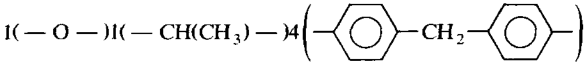
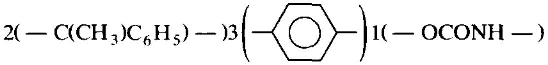
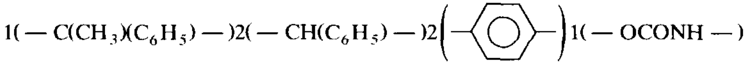
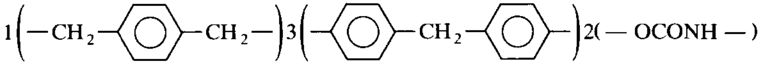
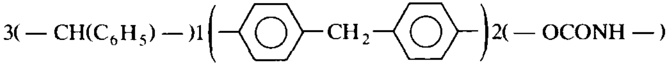
This estimation procedure requires two properties to be estimated by group-contribution techniques: Y_g and M.

4. *Permeability to oxygen*. This is estimated by the following correlation

TABLE XIV

Polymer Design—Automatic Results

Molecule	T_g	R	L	Pi
1. 1 	479.5	20.1	0.164	2.40e—03
2. 1(— OCONH —)	423.5	122.6	0.214	3.25e—15
3. 1(— CONH —)2(— C(CH$_3$)(C$_6$H$_5$) —)	445.7	19.6	0.172	6.42e—02
4. 1(— CONH —)1(— C(CH$_3$)(C$_6$H$_5$) —)1(— CH(C$_6$H$_5$) —)	408.8	19.4	0.181	1.74e—02
5. 1(— O —)1(— CHF —)2()1(— OCONH —)	453.7	19.6	0.163	1.33e—04
6. 1(— CH$_2$ —)1(— CHF —)2(— C(CH$_3$)(C$_6$H$_5$) —)1(— OCONH —)	442.6	19.8	0.161	1.62e—01
7. 3(— CH$_2$ —)1(— CONH —)2 	429.9	19.8	0.166	7.73e—02

8.		432.0	19.6	0.165	9.92e—03
9.		466.6	20.2	0.160	7.59e—02
10.		445.9	19.8	0.168	2.24e—01
11.		421.7	19.8	0.174	1.73e—01
12.		453.0	19.5	0.168	6.15e—07
13.		423.2	19.3	0.180	5.74e—05
14.	2(— $C(CH_3)(C_6H_5)$ —)2(— $CH(C_6H_5)$ —)2(— OCONH —)	434.3	19.4	0.176	1.27e—03

(Salame, 1986)

$$P = P(\pi),$$

where

$\pi = \pi(\pi_i, N_b)$—is given by the Salame correlation,

with

$\pi_i = \pi_i(\text{groups})$—permachor by the Salame group-contribution technique,

$N_b = N_b(\text{groups})$—number of backbone groups by the Salame group-contribution technique.

This estimation procedure requires two properties to be estimated by group-contribution techniques: π_i and N_b. These procedures result in seven physical properties that are estimated by group-contribution techniques: C_p^s, V, U, P_{LL}, Y_g, M, and π.

In the automatic design of new polymers, 21 groups were used. Polymers having between one and six group occurrences were allowed. The design procedure produced over 18,000 feasible polymers. Table XIV shows several polymers randomly selected from the set of feasible candidates along with the estimated values of their important physical properties. The large number of retained molecules indicates that the design specs are fairly "loose" and could be tightened by either tightening the bounding values of the physical property constraints or introducing additional physical properties constraints.

III. Interactive Synthesis of New Molecules

The automatic synthesis of molecules, described in the previous section, attempts to generate feasible candidates without resorting to the detailed, fragmented, often informal, but nevertheless sound and valuable knowledge possessed by the human designer. The search for new molecules is carried out efficiently, but it is entirely built on a limited amount of knowledge, which is represented by the sets of constraints on (1) physical properties, (2) the feasibility of molecular structures, and (3) chemical stability and other chemical properties of the resulting molecules.

The interactive design attempts to support the articulation of and incorporate the designer's informal knowledge and abductive capabilities in postulating "promising" molecular structures. In this manner, the resulting computer-aided tool can capture and utilize the best of two worlds: (a) the computer's ability to locate feasible solutions after extensive search of the solution space and (b) the human designer's "intelligence" in expanding and guiding the search in an efficient manner. Therefore, although this section will deal only with the features of the interactive design procedures, the *Molecule-Designer*, the computer-aided tool that implements both automatic and interactive procedures, has integrated both in a seamless manner (see Section IV).

The human-driven character of interactive design allows the use of (1) problem-specific subjective preferences, (2) informal, qualitatively stated scientific knowledge, (3) rapid evaluation of alternatives, and (4) evolutionary design of new molecules, starting from known and existing alternatives. It provides the designer with the following facilities:

1. Visualization of the target constraints, helping the designer in the selection of the most "promising" functional groups to be incorporated in a molecule.
2. Extensive databases for the extraction of patterns in the structural evolution of known molecules.
3. Interactive definition of new property estimation techniques and qualitative scientific rules, which are to be incorporated in the automatic design algorithm.

A. Illustration of Interactive Design

The framework of interactive design is entirely built on the premise of *additivity of group contributions* for the estimation of physical properties. Let us look at some typical illustrations. Table XV shows the estimation of normal boiling points T_b, and normal melting points T_m, using Joback's group contribution techniques (Joback and Reid, 1987), along with the

TABLE XV
Example of Linear Group Contribution Estimation Techniques

	Contributions[a]	
Groups	Δ_{i,T_b}	Δ_{i,T_m}
$—CH_3$	23.58	−5.10
$—CH_2—$	22.88	11.27
>CH—	21.74	12.64
>C<	18.25	46.43
—F	−0.03	−15.78
—Cl	38.13	13.55

[a] $T_b = 198.18 + \Sigma n_i \Delta_{i,T_b}$; $T_m = 122.5 + \Sigma n_i \Delta_{i,T_m}$.

contributions of a few select groups. The additivity of group contributions allows the molecules to be assembled group by group, while offering a "partial" estimate of the desired physical properties. Given a constraint, such as, $T_b > 300$ K, a group is added to the molecule and then the constraint is evaluated. Choosing one $—CH_3$ group results in $T_b = 221.76$ K. The constraint is not satisfied. Choosing a second $—CH_3$ group results in $T_b = 245.34$. The constraint is still not satisfied. Adding three $—CH_2—$ groups results in $T_b = 313.98$, which satisfies the constraint. Adding a second constraint, $T_m < 250$ K, makes the monitoring of the numerical values difficult. Graphical representations can simplify this problem.

Figure 1 shows a two-dimensional design space formed from T_b and T_m. The two imposed constraints form a feasible region denoted by the shaded area. The contributions of each group toward T_b and T_m form a two-dimensional (2D) vector. These are called *group vectors*. The intercepts of the group contribution models shown in Table XV establish the starting point for the first group vector. Beginning at the intercept point an appropriate set of group vectors are selected that terminate in the feasible region and produce a structurally feasible molecule. Figure 1 shows the group vectors for chloropropane.

Complex constraints on T_b and T_m are easily handled by the interactive design procedure. All that is required is to identify a feasible region or regions. The constraints need not be linear or convex.

The following heuristics on group selection are often helpful in selecting groups that satisfy the physical property constraints and are structurally feasible.

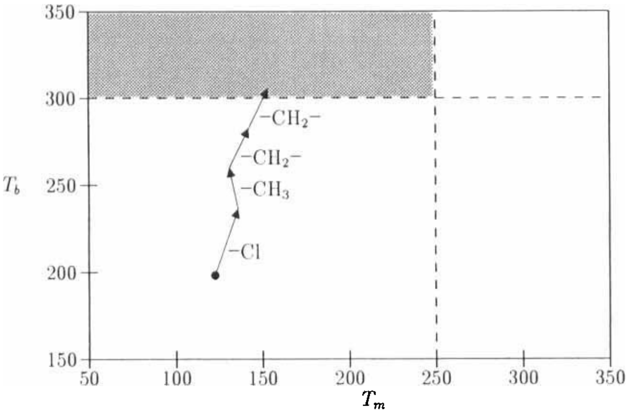


FIG. 1. An interactively designed molecule: chloropropane.

Rule 1. Separate the groups into three sets:

1. *Terminators*: all groups having one free bond.
2. *Extenders*: all groups having two free bonds.
3. Branchers: all groups having more than two free bonds.

Rule 2. Select the following initial group sets:

1. If a pure acyclic molecule is to be designed, choose two terminators.
2. If a pure cyclic molecule is to be designed, choose two cyclic extenders.

Rule 3. Continue by first considering extenders. If a brancher is to be added, follow it immediately with terminators.

These heuristics ensure that when a molecule reaches the feasible region, it is either structurally feasible or requires the addition of only one or two groups for feasibility.

1. Reduction in the Dimensionality of the Search Space

The dimension of the design space is equal to the number of fundamental physical properties needed to evaluate the property constraints. Using the estimation procedure for the vapor pressure P_{vp}, given by the Riedel–Plank–Miller correlation

$$P_{\mathrm{vp}} = P_{\mathrm{vp}}(T_{\mathrm{b}}, T_{\mathrm{br}}, P_{\mathrm{c}}),$$

where

$T_{\mathrm{b}} = T_{\mathrm{b}}(\text{groups})$ is given by Joback's group contribution,

$T_{\mathrm{br}} = T_{\mathrm{br}}(\text{groups})$ is given by Lydersen's group contribution,

$P_{\mathrm{c}} = P_{\mathrm{c}}(\text{groups})$ is given by Ambrose's group contribution,

we obtain three fundamental properties: T_{b}, T_{br}, and P_{c}. Designing for constraints on P_{vp} thus requires a three-dimensional (3D) design space. This is unfortunate because representing and manipulating 3D objects graphically is more complex than 2D manipulation.

The dimensionality of the design space can become more than a mere complication. One estimation procedure for the liquid heat capacity requires seven fundamental physical properties. To interactively design for constraints on $C_{\mathrm{p_L}}$ would require a seven-dimensional physical property space to display each of the seven fundamental properties.

Studies on *factor analysis* (Cramer, 1980a, b; Klincewicz, 1982; Joback, 1984) show that a number of physical properties are highly intercorrelated.

TABLE XVI

Physical Property–Factor Relationships

$1/\sqrt{P_c} =$	0.157	$-\ 0.019F_1$	
$V_c =$	296.1	$-\ 89.66F_1$	$-\ 36.68F_3$
$n_A =$	14.50	$-\ 5.35F_1$	$-\ 1.18F_3$
$C^o_{p,298} =$	25.70	$-\ 8.72F_1$	$-\ 2.44F_3$
$T_c =$	545.9	$-\ 24.65F_1$	$-\ 87.92F_3$
$T_b =$	358.4	$-\ 25.26F_1$	$-\ 64.94F_3$
$\Delta H_{vb} =$	7686.6	$-\ 432.3F_1$	$-\ 1614.3F_3$

High correlations between two properties indicate the possibility of replacing one with a function of the other. This would enable us to reduce the dimensionality of a design space. One study (Joback, 1984) found that nine physical properties were well approximated by three new properties called *factors*. Table XVI shows several of the derived physical property estimation techniques, using essentially two of these factors, F_1 and F_3. Group contribution estimation techniques were developed for both factors, F_1 and F_3.

Incorporating these equation-oriented estimation techniques into a new estimation procedure for P_{vp} yields the following correlation, proposed by Riedel, Plank, and Miller

$$P_{vp} = P_{vp}(T_b, T_c, P_c),$$

where the following properties can be estimated from their correlations to factors F_1 and F_3 (Table XVI):

$$T_b = T_b(F_1, F_3),$$

$$T_c = T_c(F_1, F_3),$$

$$P_c = P_c(F_1, F_3),$$

with F_1 and F_2 computed by group-contribution correlations established by Joback:

$$F_1 = F_1(\text{groups}),$$

$$F_3 = F_3(\text{groups}).$$

The fundamental properties are F_1 and F_3. Two fundamental properties enable design in a 2D physical property space.

2. *Utilization of Interactive Design*

The graphical representation of physical property constraints allows the designer to quickly gain insight into the feasibility of the problem and the relative importance of each constraint. Excessively large or small feasible regions, redundant constraints, or open feasible regions may indicate a need to respecify design constraints.

Once satisfied with the feasible region, design may proceed in one of the following three ways:

a. Evolutionary Design. Starting from an existing molecule, interactive design allows the evolution of the molecule to new altenatives through guided structural modifications. For example, Fig. 2a shows the feasible region defined by constraints on three physical properties and how two functional groups of the initial (infeasible) molecule were replaced by two

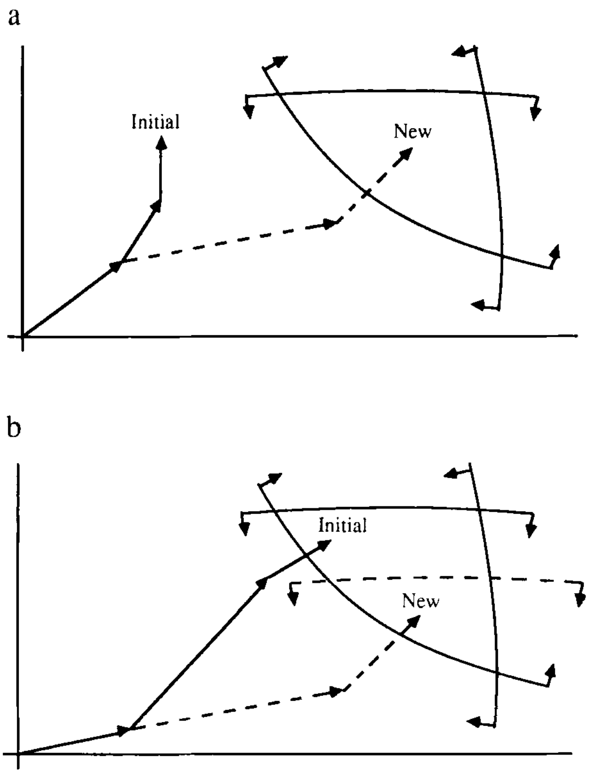


Fig. 2. Interactive design approaches: (a) evolutionary design of improved molecules; (b) evolutionary tightening of design constraints.

other groups to yield a feasible molecule. Figure 2b, on the other hand, shows how interactive design can be employed to tighten the specifications of a physical property constraint (e.g., by moving the location of a constraint; see dashed-line constraint), requiring the evolution of the initial molecule to a new one (satisfying the new set of constraints).

b. Grass-Roots Design. Building the molecule from scratch, group by group, using pure interactive or in conjunction with automatic search.

c. Combination of Evolutionary and Grass-Roots Designs. A molecule is designed from scratch so that it meets the initial specifications. Subsequently, improvements are searched for by tightening the design specifications and carrying out an evolutionary design. Figure 2b depicts such a situation.

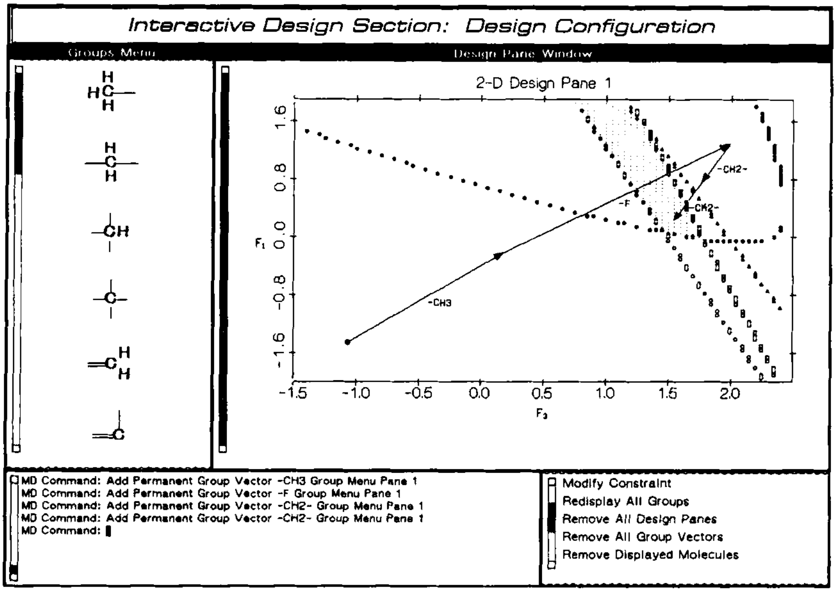
B. Case Study: Interactive Design of Refrigerants

The specifications of this case study are identical to those described in Section II.C and will not be reproduced here.

To design in a 2D space, the factor relationships shown in Table XVI are used to reduce the fundamental properties to the factors F_1 and F_3. Each physical property constraint, which is a function of F_1 and F_3, is plotted in a 2D $\{F_1$–$F_3\}$ design space shown in Fig. 3. The region in which all constraints are satisfied is shaded. The displayed symbols correspond to the constraints as follows:

Symbol	Constraints
Circle	P_{vp} ($T = -1.1$°C) > 1.4 bar
Triangle	P_{vp} ($T = 43.3$°C) < 14 bar
Square	ΔH_v ($T = -1.1$°C) > 18.kJ/ g-mol
Circle	C_{p_L} ($T = 21.1$°C) < 32.2 cal/ g-mol · K

The graphical representation used in the interactive design procedure immediately provides insights into the design problem. The first insight is that the chosen constraints yield a feasible solution space. Although each constraint was justified on its own, there was no guarantee that the set of four constraints would produce a feasible space. If the feasible region were too large or small, this would indicate the need to respecify the design target. Additionally, it is seen that for a major portion of the design space the P_{vp} ($T = 43.3$°C) < 14 bar constraint is redundant. It is superseded by the $\Delta H_v > 18.4$ constraint.



```
MD Command: Add Permanent Group Vector -CH3 Group Menu Pane 1
MD Command: Add Permanent Group Vector -F Group Menu Pane 1
MD Command: Add Permanent Group Vector -CH2- Group Menu Pane 1
MD Command: Add Permanent Group Vector -CH2- Group Menu Pane 1
MD Command:
```

FIG. 3. Refrigerant design: a solution.

TABLE XVII

ESTIMATED PROPERTY VALUES FOR DESIGNED REFRIGERANTS

Compound	P_{vp}		ΔH_v	C_{p_L}
	272.05 K	316.4 K	272.05 K	294.3 K
1. $CH_3 - CH_3$	2.710	9.486	18.67	22.49
2. $CH_3 - Cl$	1.590	6.193	21.58	19.99
3. $CH_3 - NH_2$	0.375	2.152	29.78	23.50
4. $CH_3 - CH_2 - CH_2 - F$	1.195	5.013	21.21	31.10
5. $CH_2 = CH - CH_3$	1.310	5.195	21.24	25.33
6. $CH_2 = CH - Cl$	0.750	3.340	24.14	23.01
7. $CH_2 = CH - O - CH_3$	0.543	2.648	24.83	29.83
8. $F - CH_2 - CH_2 - CH_2 - F$	1.236	5.330	20.34	34.03
9. $F - CH = CH - CH_2 - F$	1.045	4.573	20.54	29.89
10. CCl_2F_2	0.440	2.213	24.70	27.87
11. $CBrClF_2$	0.122	0.837	28.10	29.02
12. $CHBrF_2$	0.538	2.855	22.96	26.04
13. $CH(COOH)F_2$	0.002	0.041	42.56	37.58
14. $CH(HCO)F_2$	0.417	2.421	25.91	30.42
15. $NH_2 - NH_2$	0.027	0.311	41.41	26.03
16. $CH \equiv C - Cl$	0.968	4.084	24.33	21.72
17. $CH \equiv C - CH_3$	1.670	6.297	21.44	24.06
18. $CH \equiv C - NH_2$	0.214	1.353	32.56	25.42

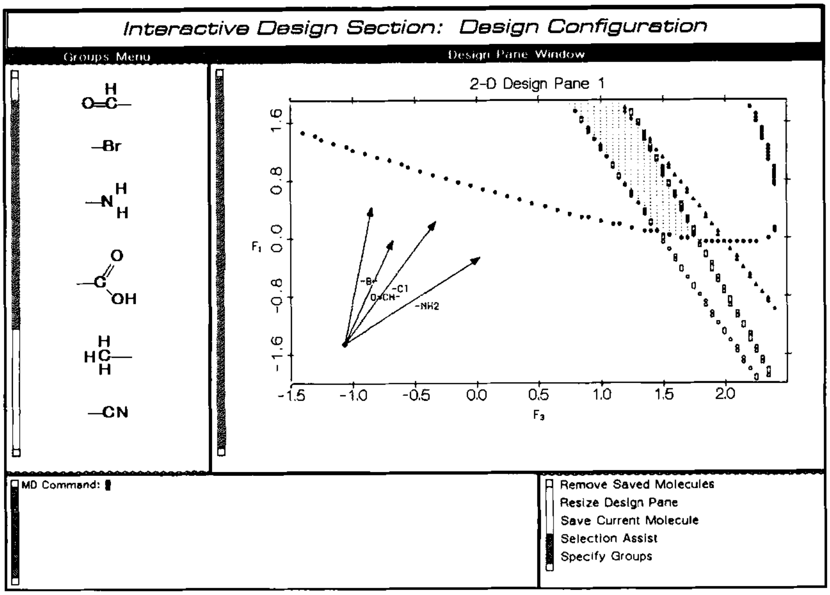


FIG. 4. Candidate groups for chlorine replacement.

Figure 3 also shows the group vectors for a designed molecule. The location of the vectors' end point can be directly interpreted into physical property information. For example, the molecule satisfies all constraints but has a liquid heat capacity near the constraint. Table XVII shows several interactively designed molecules with estimated values for their physical properties. Some of the molecules may be chemically unstable. At this stage of the design methodology only physical property and structural constraints are being explicitly considered. When choosing the groups, the designer implicitly considers chemical constraints. Once molecules are designed, the next steps of the methodology, molecule enumeration and screening, would identify and remove chemically unstable compounds.

1. Replacing Chlorine

Removing chlorine from current refrigerants would seem to reduce the hazard of ozone depletion. The interactive design procedure is well suited for searching for group replacements. The group vector for the chlorine group is shown in Fig. 4. The target is to find one or more groups making nearly equivalent contributions to that of chlorine.

The search for single group replacements is begun by restricting the possible groups to those with one single free bond. These groups are then sorted with respect to distance. The three closest groups, —CHO, —Br, and —NH_2, are displayed in the design space shown in Fig. 4.

Again the graphical representation used by the interactive design enables us to evaluate the effect of any substitution. Replacing, —Cl by O=CH—, would result in a compound having reduced vapor pressure, an increased enthalpy of vaporization, and an increased liquid heat capacity. These alterations of physical properties are derived from the relative positions of the group vectors and the locations of the constraints.

C. Case Study: Interactive Design of an Extraction Solvent

The success of a liquid–liquid extraction process depends on the selection of the most appropriate solvent (Lo *et al.*, 1983). This case study examines the design of a solvent for the extraction of acetic acid from water. Lo *et al.*'s (1983) procedure for solvent selection was adapted for interactive design use.

The physical properties of the solvent used to facilitate separation in liquid–liquid extraction have major impact on process performance. Two important physical properties are (1) solvent selectivity for the solute and (2) the solute's partition coefficient between the solvent and the parent liquor.

Lo used the three-term solubility parameter (Barton, 1983) and a graphical procedure to identify solvents for liquid–liquid extraction. In a 2D space constructed from the polar component of the solubility parameter δ_p and the hydrogen bonding component of the solubility parameter δ_H, the distribution coefficient of the solute B, m_B, is given by

$$m_B \propto r_{B,S}^{-2},$$

and the selectivity β_B is given by

$$\beta_B \propto \left(\frac{r_{A,S}}{r_{B,S}}\right)^2,$$

where $r_{i,j}$ is the distance between points i and j. The optimal solvent for extraction thus lies on a line passing through the solute and parent liquor close to the solute for large distribution but far from the parent liquor for high selectivity.

Figure 5 shows the design space for the problem of designing a solvent to extract acetic acid from water. The target area is the lower left-hand portion of the acetic acid–water line. Figure 5 shows the group vectors for acetone near the target region. Table XVIII shows several solvents that

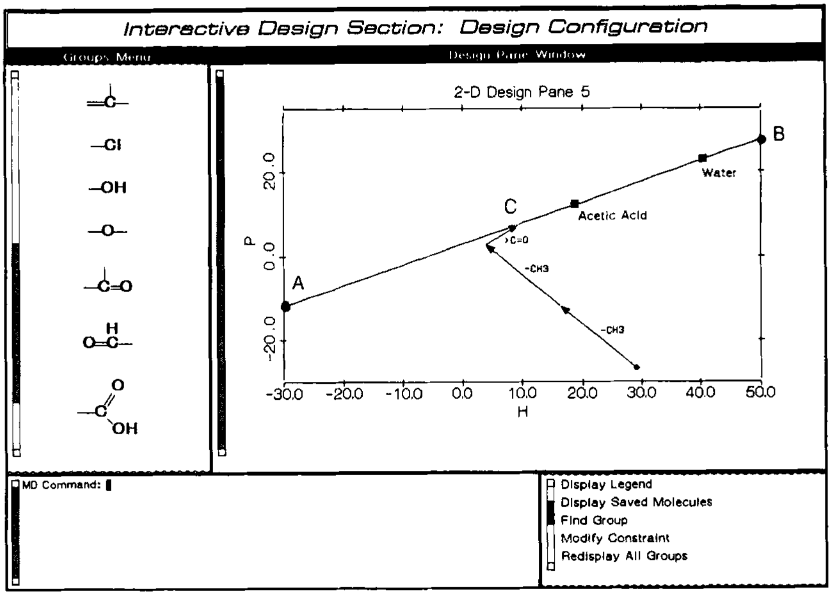


FIG. 5. Solvent design: a solution.

were interactively designed along with estimated values for their solubility parameters.

The —CH_2— group has a δ_p contribution of −.328, and δ_H contribution of −.512. The slope of the —CH_2— group vector in the interactive design space is .64. This is very close to the acetic acid–water target line slope of .49. Adding several —CH_2— groups to any of the solvents shown in Table XVIII thus results in an acceptable new solvent.

TABLE XVIII

LIQUID–LIQUID EXTRACTION SOLVENTS

Solvent	δ_p	δ_H
CH_3-Cl	7.2	6.2
$CH_2=CH-CH_3$	4.7	4.0
$CH_2=CH-Cl$	9.2	6.3
CH_3-O-CH_3	3.8	5.7
$CH_3-CO-CH_3$	7.5	9.5
$CH_2=C(CH_3)_2$	4.6	4.1
$C(CH_3)_4$	0.5	1.2
$CCl_2(CH_3)_2$	9.5	5.8

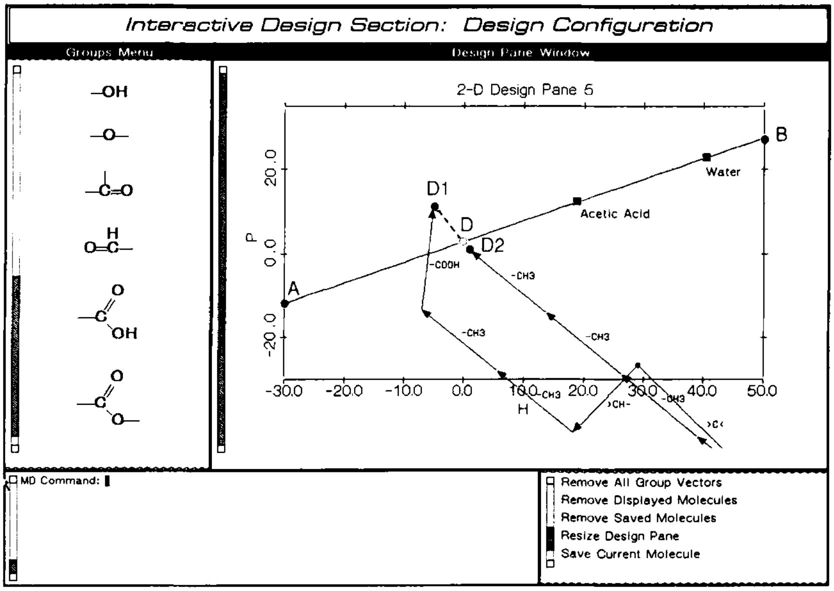


FIG. 6. Interactive design of solvent mixtures.

The direction of the $—CH_2—$ group vector is toward lower δ_p and δ_H values. This reduces the distribution coefficient with increasing number of $—CH_2—$ groups. This result is in accordance with experimental observation (Lo *et al.*, 1983).

Linear mixing rules for solubility parameters enable the design of solvent mixtures. In a δ_p–δ_H design space a binary mixture's solubility parameters lie on a straight line joining the components' solubility parameters. Figure 6 shows the group vectors for the $[CH_3—CH(COOH)—CH_3]$-$[C(CH_3)_4]$ solvent pair (point D resulting from the mixing of the two solvents represented by points D_1 and D_2). Such an approach can be used for any mixture property approximated by a linear mixing rule. Research is continuing on the use of nonlinear mixing rules for the interactive design of mixtures.

D. Case Study: Interactive Design of a Pharmaceutical

This case study presents an example taken from work done by Cramer (1980c; Cramer *et al.*, 1979). It demonstrates how the interactive design procedure assists in the second step of the quantitative structure activity relationship (QSAR) approach to drug design.

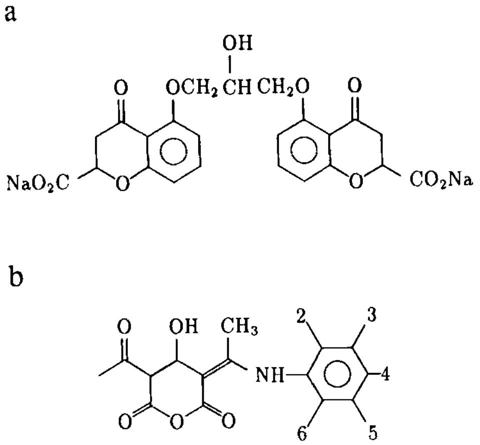


FIG. 7. Current and lead compounds for the drug design case study: (a) existing antiallergic, sodium chromoglycate; (b) a lead compound.

The QSAR approach to drug design takes a two-step approach to the correlation of biological activity with structure. The first step correlates a measure of biological activity, usually the reciprocal concentration having a 50% effect, with a number of physical properties. This correlation yields a model that is used to determine the optimal value of the physical properties giving the maximum potency. The second step of the approach is to identify group substitutions that lead to structures possessing these optimal physical property values.

Sodium chromoglycate (Fig. 7a) is effective at warding off asthmatic attacks. However, it must be administered by inhalation. Cramer *et al.* (1980c) performed a QSAR study to find a more potent pharmaceutical that could be administered orally. Searching through a database of approximately 1000 compounds they selected the pyranenamies (Fig. 7b) as their lead compound. The pyranenamines have biological properties sufficiently promising to merit a synthetic search for structurally related compounds having improved properties. The task was to develop highly active compounds by varying phenyl ring substituents.

Potency of the developed candidates was measured by the log of the concentration causing a 50% inhibition in asthmatic activity. This measure is represented as pI_{50}. To relate pI_{50} to molecular structure Cramer *et al.* (1980c) followed the two step QSAR approach. Nineteen compounds were synthesized with varying substituents. pI_{50} was measured for each analog. Values for the octanol–water partition logarithm, π, and Hammett's constant, σ, were computed for each of the substituent sets.

These data were regressed to develop a relationship between pI_{50} and the biochemical parameters π and σ. The model developed was

$$pI_{50} = -0.72 - 0.14\Sigma\pi - 1.35(\Sigma\sigma)^2. \qquad (14)$$

Group-contribution estimation techniques were developed for π and σ. The contributions were obtained by regressing the substituent constants tabulated by Hansch and Leo (1979). Overall the estimation techniques have considerable error and cannot be used for quantitative estimation. However, they provide correct overall trends and thus inform the designer of promising substitutions.

Equation (14) shows that to perform an interactive design we would use a two-dimensional σ vs. π physical property design space. The contours represent solutions of Eq. (14) with pI_{50} equal to -1.0, -0.5, and 0.0. The symbols correspond to the values:

Symbol	pI_{50}
Square	-1.0
Triangle	-0.5
Circle	0.0

The direction of maximum activity is thus near σ equal to zero and π large and negative. Figure 8 shows a pair of metasubstituents in our target

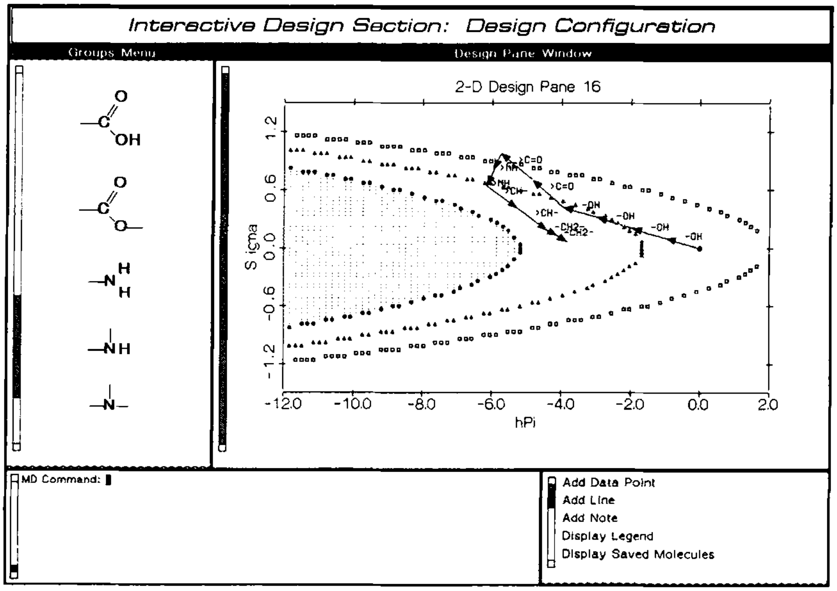


FIG. 8. Two high-activity metasubstituents during the interactive drug design.

TABLE XIX

Comparison of Experimental and Model Activities

			pI_{50}	
Substituents	π	σ_M	Model	Experimental
1. 3 — $NHCO(CHOH)_2H$ 5 — $NHCO(CHOH)_2H$	−3.812	0.064	−0.19	3.0
2. 3 — $NHCOCH_2CH_3$ 5 — $NHCOCH_2CH_3$	−0.964	0.269	−0.68	2.5
3. 3 — $NHCOCH_3$ 5 — $NHCOCH_3$	−1.552	0.391	−0.71	1.9
4. 3 — $NHCOCH_3$ 5 — OH	−1.763	0.301	−0.60	1.7
5. 3 — $NHCOCOOCH_2CH_3$ 5 — $NHCOCOOCH_2CH_3$	−1.890	0.838	−1.40	1.7
6. 3 — $NHCOCH_2CH_2CH_3$ 5 — $NHCOCH_2CH_2CH_3$	−0.376	0.147	−0.70	1.3
7. 3 — $NHCO(CHOH)_2H$	−1.906	0.032	−0.45	1.3
8. 3 — $NHCOCH_3$ 5 — NH_2	−1.653	0.157	−0.49	1.0
9. 3 — $NHCOCH_2CH_3$	−0.482	0.134	−0.68	0.7
10. 3 — $NHCOCH_3$	−0.776	0.196	−0.66	0.7

direction. These metasubstituents correspond to entry 1 of Table XIX. Cramer found these substituents to have one thousand times more activity than the original unsubstituted compound.

IV. The *Molecule-Designer* Software System

A. General Description

The *Molecule-Designer* is the software system constructed to implement the interactive and automatic procedures for the design of molecules discussed in Sections II and III. It consists of approximately 20,000 lines of LISP code with an additional 17,000-line databank. It is implemented in Common LISP on a LISP Machine. The system is divided into eight sections each corresponding to a section of the overall methodology. The eight sections of the system are (1) login section configuration, (2) database section, (3) interactive design section, (4) molecule evaluation section, (5) problem formulation section, (6) target transformation section, (7) automatic design section, and (8) group-contribution section.

B. Interactive-Design-Relevant Sections

The following sections are specifically relevant to interactive design.

1. Problem Formulation Section

The problem formulation section provides an interface with which the designer can enter physical property constraints. The system displays the properties stored in its database (currently about 40 properties are present) and provides a simple constraint editor for entering and modifying physical property constraints.

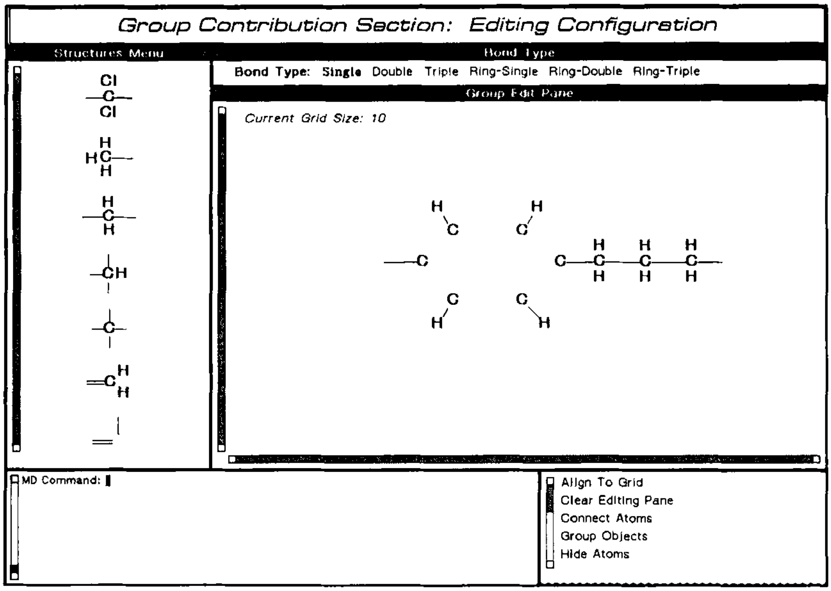
2. Group-Contribution Section

Physical property estimation procedures are at the heart of the design procedures. The group-contribution section provides facilities for entering group contribution and equation oriented correlations. Models in the form of LISP code are entered for both types of estimation techniques. Additionally, groups and their contributions are specified for group contribution techniques.

The section is divided into two configurations: (1) an editing configuration that provides facilities for entering new groups to the system's database and (2) a model entry configuration that provides facilities for entering contributions for these groups, models for group-contribution techniques, and models for equation oriented estimation techniques. Figure 9 shows an example screen of the editing configuration with a new group being constructed.

3. Target Transformation Section

Once physical property constraints are entered in the problem formulation section, it is necessary to instruct the system how to estimate the properties contained in these constraints. This involves the creation of estimation procedures for each physical property used in the design constraints. The target transformation section provides facilities for collecting estimation techniques into estimation procedures. The chosen



Group Contribution Section: Editing Configuration

Bond Type: **Single** Double Triple Ring-Single Ring-Double Ring-Triple

Current Grid Size: 10

MD Command:

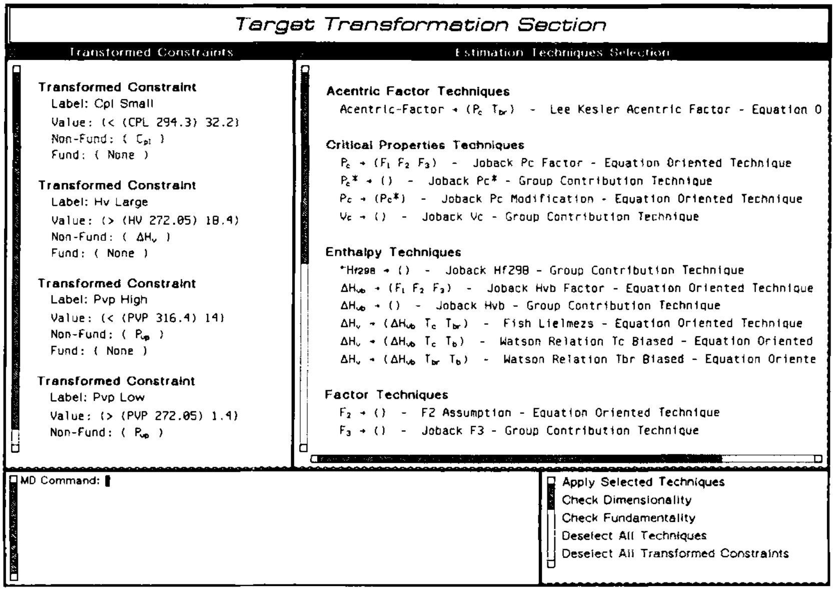
FIG. 9. Group contribution editing, configuration screen.

estimation techniques are used by the system to form an evaluation function to determine the feasibility of each point in a design space.

Figure 10 shows an example screen of the target transformation section. The constraints being transformed are shown in the window on the left. The estimation techniques that the designer can choose are shown in the window on the right. Currently the system has over 80 estimation techniques stored in its database.

4. *Interactive Design Section*

The interactive design section displays design spaces and group vectors. The configuration provides facilities to assist in selecting groups and in manipulating design constraints. Group vectors can be sorted by angle, magnitude, or closeness to another vector. The groups available can be restricted to those satisfying constraints on global valence, bond or atom types, and designer preferences. Each constraint can be interactively respecified enabling a designer to investigate the sensitivity of the feasible region. Example screens were previously shown when discussing the case studies in Sections III.B–D.


Target Transformation Section
Transformed Constraints
Estimation Techniques Selection
MD Command:
Apply Selected Techniques
Check Dimensionality
Check Fundamentality
Deselect All Techniques
Deselect All Transformed Constraints

Transformed Constraint
Label: Cpl Small
Value: (< (CPL 294.3) 32.2)
Non-Fund: (C_{pl})
Fund: (None)

Transformed Constraint
Label: Hv Large
Value: (> (HV 272.05) 18.4)
Non-Fund: (ΔH_v)
Fund: (None)

Transformed Constraint
Label: Pvp High
Value: (< (PVP 316.4) 14)
Non-Fund: (P_{vp})
Fund: (None)

Transformed Constraint
Label: Pvp Low
Value: (> (PVP 272.05) 1.4)
Non-Fund: (P_{vp})

Acentric Factor Techniques
Acentric-Factor → (P_c T_{br}) - Lee Kesler Acentric Factor - Equation O

Critical Properties Techniques
P_c → (F_1 F_2 F_3) - Joback Pc Factor - Equation Oriented Technique
P_c^* → () - Joback Pc* - Group Contribution Technique
P_c → (P_c^*) - Joback Pc Modification - Equation Oriented Technique
V_c → () - Joback Vc - Group Contribution Technique

Enthalpy Techniques
$^*H_{f298}$ → () - Joback Hf298 - Group Contribution Technique
ΔH_{vb} → (F_1 F_2 F_3) - Joback Hvb Factor - Equation Oriented Technique
ΔH_{vb} → () - Joback Hvb - Group Contribution Technique
ΔH_v → (ΔH_{vb} T_c T_{br}) - Fish Lielmezs - Equation Oriented Technique
ΔH_v → (ΔH_{vb} T_c T_b) - Watson Relation Tc Biased - Equation Oriented
ΔH_v → (ΔH_{vb} T_{br} T_b) - Watson Relation Tbr Biased - Equation Oriente

Factor Techniques
F_2 → () - F2 Assumption - Equation Oriented Technique
F_3 → () - Joback F3 - Group Contribution Technique

FIG. 10. Target transformation section, configuration screen.

5. *Molecule Evaluation Section*

The evaluation section provides facilities for estimating a molecules' physical properties. Estimating physical properties is especially useful when formulating the design target and evaluating designed molecules. During problem formulation we may need to estimate the properties of compounds currently in use. In final evaluation we would like to have the property profile of the designed compounds available for inspection.

One of the major objectives of the evaluation section is to provide estimation techniques of the highest accuracy. These estimation techniques may not be appropriate for use in the design procedures. They would thus serve as an additional check to verify the efficacy of any molecules designed.

V. Concluding Remarks

The identification of molecules with desired physical properties' values is becoming at an increasing frequency, a deliberate task with a dominant feedforward design philosophy and a decreasing number of feedback adjustments, coming from experimental results. This attitude will be reinforced as new theoretical and empirical physical property estimation techniques will provide improved estimates.

Using functional groups as the essential building blocks, one may compose any molecular structure and simultaneously employ group contribution techniques for the estimation of the molecule's properties. Clearly, a common alphabet for (1) the description of molecules and (2) the evaluation of their properties, is the most efficient but not necessarily the only alternative. The best representation for the design of molecules will be determined by the representation offering the best property estimation techniques, provided that it is at a higher-than-atomic level. Functional groups and group contribution techniques offer, at the present, the only available, consistent framework.

Once the representation of a molecular structure has been resolved, selecting the molecule(s), which satisfies a set of constraints, is simply a task of searching through a large space of discrete solutions. One could bring forth any technique for solving such problem, e.g., specific branch-and-bound algorithm to solve the underlying MINLP problem. Instead of *one, implicitly located solution*, we have opted for a search strategy that *explicitly locates all feasible solutions*. Such a preference has been motivated by the fact that during the initial design phase one is interested in all options rather than just one. Several additional considerations will be taken into account before a final choice is made; considerations that are not normally included during the initial phase of the design. To tame the combinatorial complexity of the design problem, we have used a hierarchical strategy with successive refinement of molecular representation. So, the clearly infeasible molecules can be easily screened out fairly early on. As more detail is added to the molecular representations, if the number of retained molecules is still very high, then, clearly, the initial design specs are fairly loose and should be tightened. Also, as the number of alternative molecules satisfying the set of property constraints decreases as a result of systematic search, more advanced techniques can be employed for the estimation of physical properties, and increased experimentation offers high returns.

As the limitations of the additive group-contribution techniques become more apparent, new representational models will be required to solve the product design problem. These models must maintain some simplicity in the *structure–property* relationships, which can be inverted in an efficient and explicit manner to yield the structure(s) of the feasible molecule(s). Such models have yet to be invented, but it is important to keep in mind the needs of the product design, as new theories and techniques are being written for the estimation of physical properties.

References

Alternburg, von K., Die Abhängigkeit der Siedetemperatur isomerer Kohlenwasserstoffe von der Form der Moleküle. *Brennst.-Chem.* **47**(11), 331–336 (1966).

American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), "Handbook of Fundamentals." ASHRAE, New York, 1972.

Barton, A. F. M., "CRC Handbook of Solubility Parameters and Other Cohesion Parameters." CRC Press, Boca Raton, FL, 1983.

Berg, L., Selecting the agent for distillation processes. *Chem. Eng. Prog.* **65**(9), 52–57 (1969).

Brignole, E. A., Bottini, S., and Gani, R., A strategy for the design and selection of solvents for separation processes. *Fluid Phase Equilib.* **29**, 125–132 (1986).

Constantinou, L., Gani, R., Fredenslund, Aa., Klein, J. A., and Wu, D. J., Computer-aided product design, problem formulation and application. *Proc. PSE'94*, Kyongju, Korea (1994).

Cramer, R. D., BC(DEF) Parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc.* **102**(6), 1837–1849 (1980a).

Cramer, R. D., BC(DEF) Parameters. 2. An empirical structure-based for the prediction of some physical properties. *J. Am. Chem. Soc.* **102**(6), 1849–1859 (1980b).

Cramer, R. D., A QSAR success story. *CHEMTECH*, December, pp. 744–747 (1980c).

Cramer, R. D., Snader, K. M., Willis, C. R., Chakrin, L. W., Thomas, J., and Sutton, B. M., Application of quantitative structure-activity relationships in the development of the antiallergic pyranenamines. *J. Med. Chem.* **22**(6), 714–725 (1979).

Derringer, G. C., and Markham, R. L., A computer-based methodology for matching polymer structures with required properties. *J. Appl. Polym. Sci.* **30**, 4609–4617 (1985).

Dillinger, T., "VLSI Engineering." Prentice-Hall, Englewood Cliffs, NJ, 1988.

Dossat, R. J., "Principles of Refrigeration." Wiley, New York, 1981.

Douglas, J. M., "Conceptual Design of Chemical Processes." McGraw-Hill, New York, 1988.

Fogiel, M., "Modern Microelectronics." Research and Education Association, New York, 1972.

Francis, A. W., Solvent selectivity for hydrocarbons. *Ind. Eng. Chem.* **36**(8), 764–771 (1944).

Franke, R., "Theoretical Drug Design Methods." Elsevier, Amsterdam, 1984.

Fredenslund, A., Gmehling, J., and Rasmussen, P., "Vapor-Liquid Equilibria using UNIFAC." Elsevier, Amsterdam, 1977.

Gani, R., and Brignole, E. A., Molecular design of solvents for liquid extraction based on UNIFAC. *Fluid Phase Equilib.* **13**, 331–340 (1983).

Gani, R., and Fredenslund, Aa., Computer-aided molecular and mixture design with specific property constraints. *Fluid Phase Equilib.* **82**, (1993).

Gani, R., Nielsen, B., and Fredenslund, Aa., A group contribution approach to computer-aided molecular design. *AIChE J.* **37**, 1318 (1991).

Godfrey, N. B., Solvent selection via miscibility number. *CHEMTECH*, June, pp. 359–363 (1972).

Goosey, M. T., Permeability of coatings and encapsulants for electronic and optoelectronic devices, *in* "Polymer Permeability" (J. Comyn, ed.), pp. 309–339. Elsevier, London, 1985.

Gordon, M., and Scantlebury, G. R., Non-random polycondensation, statistical theory of the substitution effect. *Trans. Faraday Soc.* **60**, 604–621 (1964).

Gray, N. A. B., "Computer-Assisted Structure Elucidation." Wiley, New York, 1986.

Hammett, L. P., Some relations between reaction rates and equilibrium constants. *Chem. Rev.* **17**(1), 125–136 (1935).

Hansch, C., and Leo, A., "Substituent Constants for Correlation Analysis in Chemistry and Biology." Wiley, New York, 1979.

Hansch, C., Muir, R. M., Fujita, T., Maloney, P. P., Geiger, F., and Streich, M., The correlation of biological activity of plant growth regulators and chloromycetin derivatives with Hammett constants and partition coefficients. *J. Am. Chem. Soc.* **85**, 2817–2824 (1963).

Hayes-Roth, F., Waterman, D. A., and Lenat, D. B., "Building Expert Systems." Addison-Wesley, Reading, MA, 1983.

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

Hosoya, H., and Murakami, M., Topological index as applied to p-electronic systems. II. Topological bond order. *Bull. Chem. Soc. J.* **48**(12), 3512–3517 (1975).

Joback, K. G., A unified approach to physical property estimation using multivariate statistical techniques. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA (1984).

Joback, K. G., and Reid, R. C. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* **57**, 233–243 (1987).

Joback, K. G., and Stephanopoulos, G. Designing molecules possessing desired physical property values. "Foundations of Computer-Aided Process Design." Elsevier, Amsterdam, 1990.

Kier, L. B., and Hall, L. H., "Molecular Connectivity in Structure-Activity Analysis." Wiley, New York, 1986.

Klincewicz, K. M., Prediction of critical temperatures, pressures, and volumes of organic compounds from molecular structure. Master's Thesis, Massachusetts Institute of Technology, Cambridge, MA, (1982).

Langley, B. C., "Refrigeration and Air Conditioning." Prentice-Hall, Englewood Cliffs, NJ, 1986.

Lo, T. C., Baird, M. H. I., and Hanson, C., "Handbook of Solvent Extraction." Wiley, New York, 1983.

Lyman, W. J., Reehl, W. F., and Rosenblatt, D. H., "Handbook of Chemical Property Estimation Methods." McGraw-Hill, New York, 1982.

Macchietto, S., Odele, O., and Omatsone, O., Design of optimal solvents for liquid–liquid extraction and gas absorption processes. *Trans. Inst. Chem. Eng.* **68**, 429 (1990).

Martin, Y. C., "Quantitative Drug Design: A Critical Introduction." Dekker, Inc., New York, 1978.

Mih, W. C., Catalysts for epoxy molding compounds in microelectronic encapsulation. *ACS Symp. Ser.* **242**, 273–283 (1984).

Moore, R. E., "Methods and Applications of Interval Analysis." Society for Industrial and Applied Mathematics, Philadelphia, 1979.

Nielsen, B., Gani, R., and Fredenslund, Aa., A group contribution approach to computer aided molecular design. *AIChE J.* (1995) (in press).

Odele, O., and Macchietto, S., Computer-aided molecular design, a novel method for optimal solvent selection. *Fluid Phase Equilib.* **82**, 47 (1993).

Perry, R. H., and Chilton, C. H., "Chemical Engineers' Handbook." McGraw-Hill, New York, 1973.

Randić, M., On characterization of molecular branching. *J. Am. Chem. Soc.* **97**(23), 6609–6615 (1975).

Reid, R. C., Prausnitz, J. M., and Poling, B. E., "The Properties of Gases and Liquid." McGraw-Hill, New York, 1987.

Rouvray, D. H., Predicting chemistry from topology. *Sci. Am.* **255**(3), 40–47 (1986).

Salame, M., Prediction of gas barrier properties of high polymers. *Polym. Eng. Sci.* **26**(33), 1543–1546 (1986).

Stephanopoulos, G., and Townsend, D. W., Synthesis in process development. *Chem. Eng. Res. Dev.* **64**, 160–174 (1986).

Taft, R. W., Separation of polar, steric, and resonance effects in reactivity. *In* "Steric Effects in Organic Chemistry" (M. S. Newman, ed.). Wiley, New York, 1956.

Tortorello, A., and Kinsella, M. A., Solubility parameter concept in the design of polymers for high performance coatings. I. *J. Coat. Technol.* **55**(696), 99–38 (1983a).

Torterello, A., and Kinsella, M. A., Solubility parameter concept in the design of polymers for high performance coatings, II. *J. Coat. Technol.* **55**(697), 29–38 (1983b).

van Krevelen, D. W., "Properties of Polymers, Correlations with Chemical Structure." Elsevier, Amsterdam (1976).

Venkatasubramanian, V., Chan, K., and Carathers, J. M., Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **18**, 883 (1994).

Verloop, A., The use of linear free energy parameters and other experimental constants in structure-activity studies. *In* "Drug Design" (E. J. Ariens, ed.) Academic Press, New York, 1972.

Watson, K. M., Thermodynamics of the liquid state. *Ind. Eng. Chem.* **35**, 398–405 (1943).

Wiener, H., Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **69**(11), 2636–2638 (1947).