Article

Optimal Computer-Aided Molecular Design: A Polymer Design Case Study

Costas D. Maranas

Ind. Eng. Chem. Res., **1996**, 35 (10), 3403-3414• DOI: 10.1021/ie960096z • Publication Date (Web): 08 October 1996

Optimal Computer-Aided Molecular Design: A Polymer Design Case Study

Costas D. Maranas

Department of Chemical Engineering, The Pennsylvania State University, 112A Fenske Laboratory, University Park, Pennsylvania 16802

This paper addresses the problem of optimally designing molecular products. A systematic analysis framework is presented for transforming a class of optimal computer-aided molecular design problems with nonlinear structure−property functionalities into equivalent mixed-integer linear (MILP) problems. While, in general, it is not possible to solve the original problem formulation for the best molecular design with mathematical certainty, the equivalent (MILP) reformulation can be solved efficiently with existing solvers and identify not only the best, but also the second, third, etc., best molecular designs. Two alternative design objectives are considered: (i) minimization of the scaled deviation of design properties from some target values, *property matching*, and (ii) minimization/maximization of a single property subject to lower and upper bounds on the rest of the properties, *property optimization*. The framework is applied to the design of polymers where thermophysical and mechanical properties are estimated using group contribution methods. Three case studies, including comparisons with existing methods, illustrate the computational efficiency and feasibility of the proposed methodology.

1. Introduction and Background

The search for new molecular products with optimal levels of thermophysical, mechanical, and optical properties is a primary objective in the chemical industries. It encompasses the design of a wide range of products including (a) polymeric materials (Allcock, 1992; Halpin, 1994; Judas *et al*., 1991); (b) extractants and solvents (Gani *et al*., 1991; Joback, 1989; Naser and Fournier, 1991; Odele and Macchietto, 1993); (c) optical multilayer filters (Epstein, 1952; Thelen, 1989; Dobrowolski, 1994); (d) refrigerants (Joback and Stephanopoulos, 1989; Kopko, 1990; Duvedi and Achenie, 1995); (e) lubricants (Ichiro, 1991; Dare-Edwards, 1991); (f) ceramics (Giannelis, 1989; Mehrotra, 1992; Babonneau, 1993), and many more. Traditional molecular design involves a protracted and costly series of experiments for synthesizing each product candidate and evaluating its desirability. Computer-aided molecular design (CAMD) methods expedite the design process by forecasting promising molecular designs. Product properties are typically estimated with group contribution methods (van Krevelen, 1990; Gani *et al*., 1989; Joback, 1984; Joback and Reid, 1987) establishing input−output relations between the type and number of molecular groups in a molecule or polymer repeat unit and various macroscopic properties.

Optimal molecular design (OMD) involves the identification of a single or mixture of compounds that optimizes one or multiple objectives while satisfying a number of macroscopic property specifications. Specifically, in polymer engineering it is often important to systematically identify the architecture of a polymer which best meets a number of performance requirements and design considerations. When there are only a few "loose" property constraints, a suitable candidate might be found from polymer property databases and tabulations (van Krevelen, 1990; Blackletter, 1988). However, when many "tight" property constraints must be satisfied simultaneously or a performance objective needs to be maximized, database searching, which is limited by the number of tabulated alternatives, is not always sufficient. The advantage of CAMD is that it is not restricted to an existing tabulation of polymer designs. Instead, by utilizing group contribution methods (van Krevelen, 1990) to estimate physical, chemical, and mechanical polymer properties, CAMD is capable of eliciting new, sometimes unexpected, molecular designs. This aids the preliminary screening process, which is followed by experimental testing and verification with published property data.

A number of CAMD methods have been proposed in the chemical engineering literature, primarily for the design of solvents, refrigerants, and polymers. These approaches are based on enumeration techniques (Stephanopoulos and Townsend, 1986; Joback, 1989; Joback and Stephanopoulos, 1989; Derringer and Markham, 1985), knowledge-based strategies (Brignole *et al*., 1986; Nagasaka *et al*., 1990; Nielsen and Gani, 1990; Gani *et al*., 1991), graph reconstruction methods (Gordeeva *et al*., 1990; Kier *et al*., 1993), multistage approaches (Naser and Fournier, 1991; Gani and Fredenslund, 1993), genetic algorithms (Venkatasubramanian *et al*., 1994a,b, 1995), artificial intelligence (Bolis *et al*., 1991), local MINLP optimization (Odele *et al*., 1990; Odele and Macchietto, 1993; Vaidyanathan and El-Halwagi, 1994; Duvedi and Achenie, 1995), and interval Newton implementations (Vaidyanathan and El-Halwagi, 1996). For the polymer design problem, Derringer and Markham (1985) first proposed a CAMD approach for finding viable polymer candidates satisfying a number of properties that are estimated with empirical equations (van Krevelen, 1976). The method was of a heuristic nature and was applied to a polymer design problem involving water absorption, glass transition temperature, and density specifications. Later, Joback and Stephanopoulos (1989) introduced an enumeration-based technique coupled with interval operations which was also applied to the identification of polymers satisfying constraints on properties such as glass transition temperature, volume expansivity, thermal conductivity, and permeability to oxygen. Recently, Venkatasubramanian *et al*. (1994a,b, 1995) introduced an interesting adaptation of genetic searches in the design of polymer repeat units which were dynamically evolved in an attempt to reach prespecified property targets. Finally, Vaidyanathan and El-Halwagi (1994) addressed the problem of opti-

S0888-5885(96)00096-6 CCC: $12.00

mally designing addition or condensation polymers that optimize a performance objective while satisfying other design specifications. This task was formulated as an MINLP problem and solved with a local optimizer GINO (Liebman *et al*., 1986). Preliminary efforts to solve these problems globally were reported in Vaidyanathan and El-Halwagi (1996). The encouraging results obtained by all these methods indicate the potential of computer-aided molecular design (CAMD) methods for expediently generating promising molecular design candidates. Despite the diversity of the aforementioned CAMD approaches and the intriguing ways that they propose to tackle the enormity of the search space, a number of questions remain unanswered.

1. Can the globally optimum molecular design be always reached with mathematical certainty and without excessive computational effort?

2. Is it possible to identify not only the global optimum molecular design but also a ranked list of the the second, third, etc., best molecular designs?

3. Given a quantitative description of the inherent property estimation uncertainty, what is the degree of confidence that the obtained molecular designs will optimally meet the design objectives?

In this paper, a systematic methodology is introduced for reformulating a class of optimal molecular design (OMD) problems with widely used nonlinear structure−property functionalities into equivalent mixed-integer linear (MILP) problems. This enables the efficient identification of not only the best but also the second, third, etc., best molecular designs with existing (MILP) solvers. By eliminating the caveat of convergence to suboptimal molecular designs, the chances of identifying novel, possibly counterintuitive, superior design alternatives are improved. The key issue of quantifying property prediction imprecision and deriving a probabilistic measure of confidence that the obtained molecular designs will optimally meet the design objectives is addressed in (Maranas, 1996).

In the following sections, first a mathematical description of OMD is provided and the structure−property prediction functionality is discussed. Next, it is shown how the original nonconvex mixed-integer nonlinear optimization MINLP representation of the (OMD) problem can equivalently be expressed as a much more tractable mixed-integer linear MILP optimization problem. Finally, three case studies are addressed to illustrate the proposed framework.

2. Mathematical Description

The problem of identifying the best molecular design based on some measure of performance can be expressed as the following mixed-integer nonlinear optimization problem.

$$\min \mathrm{MP}(p_j(\mathbf{n})) \qquad \text{(OMD)}$$

$$\text{subject to} \qquad p_j^{\mathrm{L}} \le p_j(\mathbf{n}) \le p_j^{\mathrm{U}}$$

$$n_i \in \{n_i^{\mathrm{L}}, n_i^{\mathrm{L}} + 1, ..., n_i^{\mathrm{U}}\}, \qquad i = 1, ..., N$$

$$n_{\min} \le \sum_{i=1}^{N} n_i \le n_{\max}$$

In formulation (OMD), $\mathbf{n} = (n_1, ..., n_N)$ is the vector of the integer variables $n_i \in \{0, 1, 2, ...\}$, $i = 1, ..., N$, describing the number of times the ith molecular group participates in the polymer repeat unit and n_i^{L}, n_i^{U} are the corresponding upper and lower bounds. Additional constraints in (OMD) may also be placed on the total number of groups composing the polymer repeat unit,

The expressions $p_j = p_j(\mathbf{n})$, $j = 1, ..., M$, established by group contribution methods, denote the functionality between polymer property j and the number of different molecular groups in the polymer repeat unit. p_j^{L} and p_j^{U} are prespecified lower and upper bounds on property p_j. Finally, the objective function MP is a measure of the performance of the molecular design and is typically a function of one or more polymeric properties, $\mathrm{MP} = \mathrm{MP}(p_j(\mathbf{n}), j = 1, ..., M)$. Because $p_j = p_j(\mathbf{n})$, MP is ultimately a function of only the decision variables n_i which fully specify the target molecule. The following two most widely used measures of performance are considered in this study:

(1) Minimization of the maximum scaled deviation of properties from some target values (*property matching* (PM)),

$$\min \mathrm{MP} = \max_j \frac{1}{p_j^{\mathrm{s}}} |p_j(\mathbf{n}) - p_j^{\mathrm{o}}|$$

where p_j^{o} is the target for property j and p_j^{s} the corresponding scale. Clearly, the selection of the property scales may affect but not completely overwhelm the selection and relative optimality order of the best molecular designs. Therefore, property scales must be carefully selected so that they truly reflect the relative importance of various property deviations from the target values. For example, if a 10 K deviation for the glass transition temperature target is as significant as a 0.1 g/cm^3 deviation from a density target, then an obvious selection for the corresponding property scales is:

$$T_{\mathrm{g}}^{\mathrm{s}} = 10, \qquad D^{\mathrm{s}} = 0.1$$

If such quantitative trade-off information does not exist, then the natural selection for property scales is the actual property targets. In this case, the same relative importance is assigned to all percent property violations from their target values.

(2) Minimization/maximization of a single property j^* (*property optimization* (PO)),

$$\min/\max \mathrm{MP} = p_{j^*}(\mathbf{n})$$

Note that it is important to ensure that the molecular designs which optimize these measures of performance are structurally feasible. Specifically, if v_i, $i = 1, ..., N$, is the valency (i.e., number of free attachments) of the ith molecular group (e.g., the valency of $-CH_2-$ is two, while the valency of $>CH-$ is three), then there must be enough free attachments to interconnect all molecular groups in a structurally feasible way. Also, the number of the remaining free attachments must be equal to zero for molecules and two for polymer repeat units. To maintain the structural feasibility of the molecule, a number of linear constraints on $\mathbf{n}$ must be included in the problem (OMD). These structural feasibility constraints define the necessary conditions under which a set of molecular groups can be interconnected so that there is no shortage or excess of free attachments. For the sake of simplicity and without loss of generality we assume that any two molecular groups may not be linked with more than a single bond, implying that possible double and triple bonds must be

Table 1. Molecular Groups for Case Study 1

index	1	2	3	4	5	6	7
group	$-CH_2-$	$-CO-$	$-COO-$	$-O-$	$-CONH-$	$-CHOH-$	$-CHCl-$

Table 2. Group Contribution Parameter Values in Case Study 1

index	group	Y_i	V_i	H_i	M_i
1	$-CH_2-$	2 700	15.85	3.3×10^{-5}	14
2	$-CO-$	27 000	13.40	0.11	28
3	$-COO-$	8 000	23.00	0.075	44
4	$-O-$	4 000	10.00	0.02	16
5	$-CONH-$	12 000	24.90	0.75	43
6	$-CHOH-$	13 000	19.15	0.75	30
7	$-CHCl-$	20 000	29.35	0.015	48.5

Table 3. Molecular Groups in Case Study 2

index	mol. group	index	mol. group	index	mol. group
1	CH_3	5	CH_3COO-	9	$\rangle C_6H_4$
2	$-Cl$	6	$\rangle CH_2$	10	$\rangle CH-$
3	$-C_6H_5$	7	$-CH_2COO-$	11	$\rangle C_6H_3-$
4	$-COOH$	8	$\rangle CHNH_2$	12	$\rangle C\langle$

fully contained within the molecular groups. In this case, the total number of free attachments in a molecule can be specified as follows: Each time a molecular group i is added to a molecule, ν_i attachments are contributed while spending two of them to form the connection (starting with the second group). Therefore, the total number of free attachments available for bonding in the molecule or repeat unit is given by:

$$f = \sum_{i=1}^{N} (\nu_i - 2) n_i + 2$$

For example, the unit $-CH_2CH_2CH_2-$ has $(2 - 2)3 + 2 = 2$ free attachments. Clearly, in the design of a molecule the number of free attachments must be equal to zero, $f = 0$, whereas in the case of polymer repeat unit this number must be $f = 2$.

Problem (OMD) is, in general, very difficult to solve due to the nonlinearities in the property−structure relations $p_j = p_j(\mathbf{n})$ and the large number of different ways that a set of molecular groups can be interconnected in a structurally feasible manner. It has been shown that the total number of distinct molecular designs containing between K_{min} and K_{max} molecular groups selected from a pool of N molecular groups is equal to (Joback and Stephanopoulos, 1989),

$$\sum_{K=K_{min}}^{K_{max}} \frac{(N + K - 1)!}{K!(N - 1)!}$$

assuming that different permutations of the same molecular groups in the molecule are indistinguishable. This corresponds to all different combinations of N different objects, K at a time for $K = K_{min}, ..., K_{max}$ with repetitions. Table 6 summarizes the total number of distinct molecular group interconnections for representative values of K_{max} and N with $K_{min} = 1$. Table 6 clearly demonstrates that simple search techniques cannot cope with the explosive growth of molecular alternatives as more realistic problem sizes are addressed. Instead, a method is needed which can efficiently eliminate many suboptimal designs at a time without explicit one-by-one enumeration. While the potential for reaching computational tractability for arbitrary property-estimation models is questionable, by recognizing, recovering, and utilizing the prevailing mathematical features of widely used group contribution methods, this task becomes more manageable.

Table 4. Molecular Groups in Case Study 3

index	mol. group	index	mol. group	index	mol. group
1	$\rangle C\langle$	11	$-NH-$	21	$-C_2H_5$
2	$-CH_2-$	12	$-CONH-$	22	$-nC_3H_7$
3	$-CH\langle$	13	$-OCONH-$	23	$-iC_3H_7$
4	$-S-$	14	$-NHCONH-$	24	$-tC_4H_9$
5	$-SO_2-$	15	$-aC_6H_4-$	25	$-F$
6	$-O-$	16	$-bC_6H_4-$	26	$-Cl$
7	$-CO-$	17	$-cC_6H_4-$	27	$-Br$
8	$-COO-$	18	$\rangle C_6H_3-$	28	$-OH$
9	$-OCOO-$	19	$\rangle C_6H_2\langle$	29	$-C_6H_5$
10	$-COOCO-$	20	$-CH_3$	30	$-CN$

Table 5. Group Contribution Parameters in Case Study 3

index	ν_i	M_i	V_{ai}	V_{wi}	C_{pi}	U_{Ri}
1	4	12.01	4.70	3.33	6.2	40
2	2	14.03	16.37	10.23	25.35	880
3	3	13.02	9.65	6.78	15.90	460
4	2	32.06	17.30	10.80	24.05	550
5	2	64.06	32.50	20.30	50.00	1250
6	2	16.00	8.00	3.71	16.80	400
7	2	28.01	18.50	11.70	23.05	875
8	2	44.01	24.60	15.20	46.00	1225
9	2	60.01	31.00	18.90	63.00	1575
10	2	72.02	40.00	27.00	63.00	2150
11	2	15.02	6.40	8.08	14.25	875
12	2	43.03	25.00	19.56	46.00	1750
13	2	59.03	30.00	23.00	58.00	2100
14	2	58.04	30.00	27.60	50.00	2000
15	2	76.09	65.50	43.32	78.80	4100
16	2	76.09	69.00	43.32	78.80	4050
17	2	76.09	65.50	43.32	78.80	4000
18	3	75.08	63.34	40.80	71.85	3700
19	4	74.08	58.91	38.28	65.00	3300
20	1	15.03	23.00	13.67	30.90	1400
21	1	29.06	38.50	23.90	56.25	2280
22	1	43.09	55.50	34.13	81.60	3160
23	1	43.09	55.50	34.12	77.40	3250
24	1	57.11	74.00	44.34	99.00	4250
25	1	19.00	10.00	6.00	21.40	530
26	1	35.46	18.40	12.20	27.10	1265
27	1	79.92	20.95	14.60	26.30	1300
28	1	17.01	11.54	8.04	17.00	630
29	1	77.10	64.65	45.84	85.60	5000
30	1	26.02	21.09	14.70	25.00	1400

Table 6. Total Number of Distinct, Structurally Feasible Molecular Group Interconnections for Different Values of K_{max} and N

K_{max}	N	distinct designs
5	5	251
5	10	3,002
5	20	53,129
5	30	324,631
5	40	1,221,758
10	5	3,002
10	10	184,755
10	20	30,045,014
10	30	847,660,527
10	40	10,272,278,169
20	5	53,129
20	10	30,045,014
20	20	137,846,528,819
20	30	47,129,212,243,959
20	40	4,191,844,505,805,494

3. Structure−Property Relations

Group contribution methods (GCM) (Franklin, 1949) provide popular, versatile, and relatively accurate (Hor-

vath, 1992) ways for estimating properties based on the number and type of molecular groups participating in a molecule or repeat unit. GCM are based on the additivity principle of the groups constituting the molecule under investigation and have been extensively utilized in the estimation of a wide spectrum of polymeric properties including volumetric, calorimetric, thermophysical, optical, electromagnetic, and mechanical properties. An extensive compilation of these estimation methods along with the corresponding parameters can be found in (van Krevelen, 1990). Note that while some quantities are either exactly (i.e., repeat unit molecular weight) or approximately (e.g., molar volume, molar heat capacity) additive with respect to the individual molecular group contributions, this is not the case for almost all properties of interest in molecular design. In fact, it appears that the estimation of most properties pertinent to engineering design is given by the ratio of two linear expressions in n_i. We denote this special nonlinear structure−property functionality as type I.

$$p_j(\mathbf{n}) = \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i}, \qquad j = 1, ..., M \text{ (type I)}$$

Note that A_{ij} and B_{ij} are given parameters associated with a specific molecular group i and property j and independent of the molecular architecture. Table 7 summarizes a number of polymeric properties whose estimating formula is of type I. Note that the widespread use of type I functionality in property estimation is not a mere coincidence. It stems from the fact that while most properties of interest are not additive on the individual molecular group contributions, their products with either molar volume or monomer molecular weight are. Not all polymeric properties can be predicted based on an estimating formula of type I. A number of properties, in particular, mechanical properties, require an estimating formula which is derived by raising a formula of type I to some real positive power d_j. We denote this estimating type of formula as type II.

$$p_j(\mathbf{n}) = \left(\frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \right)^{d_j}, \qquad j = 1, ..., M \text{ (type II)}$$

Table 8 summarizes a number of polymeric properties and their estimating formulas which are of type II.

It is interesting to note that the property-predicting methods summarized in Tables 7 and 8 take into account only the type and number of molecular groups composing the molecule in question and not the way they are interconnected. Note that sometimes the specific molecular group interconnection has a significant effect on property values. For example, while the density (at 296 K) of polypropylene $-CH_2CH(CH_3)CH_2-CH(CH_3)-$ is 0.8504 (g/cm^3), the density of head-to-head polypropylene $-CH_2CH(CH_3)CH(CH_3)CH_2-$ is 0.8736 (g/cm^3) even though both repeat units have the same molecular group representation. In general, only partial information about the internal molecular architecture can be elicited based on knowledge of pertinent property

Table 7. Polymer Properties Estimating Formulas Following Functionality of Type I

density	$\rho = \dfrac{\sum_{i=1}^{N} M_i n_i}{\sum_{i=1}^{N} V_i n_i}$ (g/cm^3)
specific thermal expansivity	$e = \dfrac{0.45 \times 10^{-3} \sum_{i=1}^{N} V_{wi} n_i}{\sum_{i=1}^{N} M_i n_i}$ (cm^3/g K)
thermal expansion coefficient	$\alpha = \dfrac{0.45 \times 10^{-3} \sum_{i=1}^{N} V_{wi} n_i}{\sum_{i=1}^{N} V_i n_i}$ (1/K)
specific heat capacity	$c_p = \dfrac{\sum_{i=1}^{N} C_{pi} n_i}{\sum_{i=1}^{N} M_i n_i}$ (J/g K)
crystalline melting temperature	$T_m = \dfrac{\sum_{i=1}^{N} Y_{mi} n_i}{\sum_{i=1}^{N} M_i n_i}$ (K)
glass transition temperature	$T_g = \dfrac{\sum_{i=1}^{N} Y_{gi} n_i}{\sum_{i=1}^{N} M_i n_i}$ (K)
cohesive energy density	$e_{coh} = \dfrac{\sum_{i=1}^{N} E_{coh} n_i}{\sum_{i=1}^{N} M_i n_i}$ (J/cm^3)
solubility parameter	$\delta = \dfrac{\sum_{i=1}^{N} F_i n_i}{\sum_{i=1}^{N} V_i n_i}$ (J$^{1/2}$/cm$^{3/2}$)
refraction index	$n = \dfrac{\sum_{i=1}^{N} (R_{GDi} + V_i) n_i}{\sum_{i=1}^{N} V_i n_i}$ or $\dfrac{\sum_{i=1}^{N} R_{Vi} n_i}{\sum_{i=1}^{N} M_i n_i}$
dielectric constant	$\epsilon = \dfrac{\sum_{i=1}^{N} (V_i + 2P_{LLi}) n_i}{\sum_{i=1}^{N} (V_i - P_{LLi}) n_i}$ or $\dfrac{\sum_{i=1}^{N} P_{Vi} n_i}{\sum_{i=1}^{N} M_i n_i}$
water absorption	$W = \dfrac{\sum_{i=1}^{N} 18 H_i n_i}{\sum_{i=1}^{N} M_i n_i}$ (g of H_2O/g of polymer)

Table 8. Polymer Properties Estimating Formulas Following Functionality of Type II

surface tension	$\gamma = \left(\frac{\sum_{i=1}^{N} P_{si} n_i}{\sum_{i=1}^{N} V_i n_i}\right)^4$ (N/m)
unperturbed viscosity coefficient	$K_\Theta = \left(\frac{\sum_{i=1}^{N} (J_i + 4.2 Z_i) n_i}{\sum_{i=1}^{N} M_i n_i}\right)^2$ ($cm^3\ mol^{-1/2}\ g^{-3/2}$)
dipole moment	$\mu = \left(\frac{\sum_{i=1}^{N} (P_{LLi} - R_{LLi}) n_i}{20.6}\right)^{1/2}$ (D)
specific shear modulus	$G/\rho = \left(\frac{\sum_{i=1}^{N} U_{Hi} n_i}{\sum_{i=1}^{N} V_i n_i}\right)^6$ (J/g)
specific bulk modulus	$K/\rho = \left(\frac{\sum_{i=1}^{N} U_{Ri} n_i}{\sum_{i=1}^{N} V_i n_i}\right)^6$ (J/g)
activation energy of viscous flow	$E_\eta(\infty) = \left(\frac{\sum_{i=1}^{N} H_{\eta i} n_i}{\sum_{i=1}^{N} M_i n_i}\right)^3$ (J/mol)

values by utilizing group contribution methods. Exceptions to this rule are the estimating relations for the glass transition and crystalline melting temperatures which in some cases provide additional information on the internal molecule architecture. In particular, the suggested values of the parameters Y_{gi} and Y_{mi} for some molecular groups (e.g., $-C(CH_3)_2-$, $-COO-$, $-SO_2-$, etc.) are adjusted according to whether they are bonded with zero (nonconjugated), one (one-sided conjugation), or two (two-sided conjugation) aromatic groups. This can be accounted for by defining additional pseudo-molecular groups and modeling the same molecular group with a different integer variable depending on whether it has none, one-sided, or two-sided conjugation.

It is important to emphasize that the aforementioned group contribution methods provide only estimates for different polymeric properties and may only be in partial agreement with experimental values. In fact, 5−10% or even higher discrepancies between experimental values and group contribution predictions are not uncommon. The relative accuracy of these predictions depends on the particular property (e.g., density estimates are typically more accurate than glass transition temperature estimates), molecular complexity, and property prediction method. The effect of property prediction uncertainty on optimal molecular designs is addressed in (Maranas, 1996). Next, by utilizing the underlying mathematical functionalities of type I or II in polymeric property estimation (van Krevelen, 1990), the original mixed-integer nonlinear (MINLP) formation (OMD) is transformed into an equivalent MILP representation.

4. Reformulation

First, problems whose structure−property functionalities are all of type I are addressed. After omitting all linear constraints in **n** (i.e., feasibility structure requirements, size bounds, etc.), for the sake of clarity, the *property matching* (PM) problem is formulated as:

$$\min s \qquad \text{(PM)}$$

$$\text{subject to} \qquad s \geq \frac{1}{p_j^s}\left|\frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} - p_j^\circ\right|, \qquad j = 1, ..., M$$

$$n_i \in \{n_i^L, n_i^L + 1, ..., n_i^U\}, \qquad i = 1, ..., N$$

Here s is the maximum over all target properties scaled violation to be minimized, p_j° is the target value for property j, and p_j^s is an appropriate scale. Typically, the maximum percent property deviation is minimized, and therefore $p_j^s = p_j^\circ$. Note that this formulation is nonconvex since the expressions for $p_j(\mathbf{n})$ are, in general, nonlinear in **n**. The original formulation, however, can be simplified by first eliminating the absolute values. By observing that (i) an inequality of the form $|x| \leq y$ can equivalently be replaced by two inequalities $x \leq y$, $-x \leq y$, and (ii) A_{ij}, $B_{ij} > 0$, the *property matching* problem (PM) can equivalently be written as:

$$\min s \qquad \text{(PM)}$$

subject to

$$p_j^s\left(\sum_{i=1}^{N} B_{ij} n_i\right)s \geq \left(\sum_{i=1}^{N} A_{ij} n_i - p_j^\circ \sum_{i=1}^{N} B_{ij} n_i\right), \qquad j = 1, ..., M$$

$$p_j^s\left(\sum_{i=1}^{N} B_{ij} n_i\right)s \geq -\left(\sum_{i=1}^{N} A_{ij} n_i - p_j^\circ \sum_{i=1}^{N} B_{ij} n_i\right), \qquad j = 1, ..., M$$

$$n_i \in \{n_i^L, n_i^L + 1, ..., n_i^U\}, \qquad i = 1, ..., N$$

This transformation removes all absolute values but introduces nonlinear products between the continuous variable s and the linear sums of integer variables $\sum_{i=1}^{N} B_{ij} n_i$, $j = 1, ..., M$. To eliminate this nonlinearity, the integer variables **n** are first expressed as a linear combination of binary variables y_{ik} (Floudas, 1995). The integer variables $n_i \in \{n_i^L, ..., n_i^U\}$ can be expressed as a linear combination of a number of binary 0−1 variables y_{ik} as follows:

$$n_i = n_i^L + \sum_{k=0}^{K} 2^k y_{ik}, \qquad i = 1, ..., N$$

$$\text{where} \qquad K = \left\lceil \frac{\log(n_i^U - n_i^L)}{\log 2} \right\rceil$$

For example, the integer $n \in \{0, ..., 7\}$ can be written as $n = y_1 + 2y_2 + 4y_3$ where $y_1, y_2, y_3 \in \{0, 1\}$. This implies that products between continuous and integer variables can be decomposed into the sum of a number of products of continuous and binary variables. Such nonlinear products of continuous and binary variables can equivalently be expressed with four linear inequality constraints eliminating the nonlinearity at the expense of introducing only a single extra variable

(Glover, 1975). Let xy be such a product where x is a continuous variable with $x \in [x^L, x^U]$ and y is a binary variable $y \in \{0, 1\}$. This product can be replaced by a single additional continuous variable $z = xy$ which must satisfy the following constraints.

$$x - x^U(1 - y) \le z \le x - x^L(1 - y)$$

$$x^L y \le z < x^U y$$

Note that if $y = 0$ these constraints become $x - x^U \le z \le x - x^L$ and $0 \le z \le 0$, forcing z to zero. Alternatively, if $y = 1$, $x \le z \le x$, and $x^L \le z \le x^U$, then z is equal to x. In both cases z assumes the correct value xy. Based on (i) the expression of integer variables as a linear combination of binary variables and (ii) the replacement of continuous and binary variable products with linear inequality contraints, the original nonlinear formulation can equivalently be rewritten as the following mixed-integer linear programming (MILP) problem.

$$\max s \qquad \text{(PM)}$$

subject to

$$p_j^s(\sum_{i=1}^{N} B_{ij} n s_i) \ge (\sum_{i=1}^{N} A_{ij} n_i - p_j^\circ \sum_{i=1}^{N} B_{ij} n_i), \qquad j = 1, ..., M$$

$$p_j^s(\sum_{i=1}^{N} B_{ij} n s_i) \ge -(\sum_{i=1}^{N} A_{ij} n_i - p_j^\circ \sum_{i=1}^{N} B_{ij} n_i), \qquad j = 1, ..., M$$

$$n_i = n_i^L + \sum_{k=0}^{K} 2^k y_{ik}, \qquad i = 1, ..., N$$

$$n s_i = n_i^L s + \sum_{k=0}^{K} 2^k y s_{ik}, \qquad i = 1, ..., N$$

$$s - s^U(1 - y_{ik}) \le y s_{ik} \le s - s^L(1 - y_{ik})$$

$$s^L y_{ik} \le y s_{ik} \le s^U y_{ik}, \qquad i = 1, ..., N, \quad k = 1, ..., K$$

$$n_i \in \{n_i^L, n_i^L + 1, ..., n_i^U\}, \qquad i = 1, ..., N$$

Here, variables ns_i and ys_{ik} represent the products $n_i s$ and $y_{ik} s$, respectively. Also s^L and s^U are lower and upper bounds on the scaled maximum property violation (typically $s^L = 0$ and $s^U = 10-20\%$).

Often in practice, rather than designing a polymer whose properties match some prespecified targets, the maximization or minimization of a single property j^* is sought while maintaining property values within some lower and upper bounds. This objective is expressed mathematically by the *property optimization* formulation (PO). By again omitting linear constraints in n_i, this problem is formulated as follows:

$$\max/\min p_{j^*} = \frac{\sum_{i=1}^{N} A_{ij^*} n_i}{\sum_{i=1}^{N} B_{ij^*} n_i} \qquad \text{(PO)}$$

$$\text{subject to} \qquad p_j^L \le \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \le p_j^U, \qquad j = 1, ..., M$$

$$n_i \in \{n_i^L, n_i^L + 1, ..., n_i^U\}, \qquad i = 1, ..., N$$

where variable p_{j^*} denotes the j^*th property to be minimized or maximized and p_j^L and p_j^U are the corresponding lower and upper bounds. As described earlier in detail for the (PM) formulation, by (i) expressing all integer variables as linear combinations of binary variables and (ii) replacing nonlinear products of continuous and binary variables with linear inequalities (Glover, 1975), the nonlinear (PO) formulation can equivalently be rewritten as the following (MILP) problem.

$$\max/\min p_{j^*} \qquad \text{(PO)}$$

$$\text{subject to} \qquad \sum_{i=1}^{N} B_{ij^*} n p_{ij^*} = \sum_{i=1}^{N} A_{ij^*} n_i$$

$$(\sum_{i=1}^{N} B_{ij} n_i) p_j^L \le \sum_{i=1}^{N} A_{ij} n_i \le (\sum_{i=1}^{N} B_{ij} n_i) p_j^U, \qquad j = 1, ..., M$$

$$n_i = n_i^L + \sum_{k=0}^{K} 2^k y_{ik}, \qquad i = 1, ..., N$$

$$n p_{ij^*} = n_i^L p_{j^*} + \sum_{k=0}^{K} 2^k y p_{ij^*k}, \qquad i = 1, ..., N$$

$$p_{j^*} - p_{j^*}^U(1 - y_{ik}) \le y p_{ij^*k} \le p_{j^*} - p_{j^*}^L(1 - y_{ik})$$

$$p_{j^*}^L y_{ik} \le y p_{ij^*k} \le p_{j^*}^U y_{ik}, \qquad i = 1, ..., N, \quad k = 1, ..., K$$

Here, variables np_{ij^*} and $yp(ij^*k)$ represent the products $n_i p_{j^*}$ and $y_{ik} p_{j^*}$, respectively. So far it has been shown how formulations (PM) and (PO) can be transformed into (MILP) problems when all structure−property functionalities are of type I. Next, the linearization of problems where some of the structure−property functionalities are of type II is addressed.

The introduction of functionalities of type II for some properties gives rise to nonlinear terms of the form

$$(\sum_{i=1}^{N} A_{ij} n_i)^{d_j}, \qquad (\sum_{i=1}^{N} B_{ij} n_i)^{d_j}, \qquad s(\sum_{i=1}^{N} B_{ij} n_i)^{d_j}$$

in both formulations (PM) and (PO). The linearization of these terms is not as straightforward as the one for the nonlinear terms resulting from functionalities of only type I. This linearization is complex and requires the introduction of a large number of extra continuous variables. Nevertheless, one can avoid this computational burden by (i) appropriately rearranging equations and (ii) invoking monotonicity principles. Specifically, because the function $f(x) = x^d$ ($d \ge 0$) is strictly monotonically increasing in x, the maximum/minimum point of $f(x)$ subject to a set of constraints in x is the same with the maximum/minimum point of x subject to the same set of contraints. Thus, the maximization/minimization of

$$p_{j^*} = \left(\frac{\sum_{i=1}^{N} A_{ij^*} n_i}{\sum_{i=1}^{N} B_{ij^*} n_i} \right)^{d_{j^*}}$$

in the (PO) formulation can equivalently be replaced by the maximization/minimization of

$$p_{j^*}^{1/d_{j^*}} = \frac{\sum_{i=1}^{N} A_{ij^*} n_i}{\sum_{i=1}^{N} B_{ij^*} n_i}$$

This yields an objective function definition identical with that derived for type I structure−property relations. Also the nonlinear property-bounding constraints

$$p_j^{\mathrm{L}} \le \left(\frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \right)^{d_j} \le p_j^{\mathrm{U}}$$

can equivalently be replaced with the linear constraints

$$(p_j^{\mathrm{L}})^{1/d_j} \sum_{i=1}^{N} B_{ij} n_i \le \sum_{i=1}^{N} A_{ij} n_i \le (p_j^{\mathrm{U}})^{1/d_j} \sum_{i=1}^{N} B_{ij} n_i$$

Formulation (PM) involving some type II structure−property relations can also be converted to one with only type I structure−property formulas. Instead of attempting to match a target p_j° with a scale p_j^{s} to the jth property of type II,

$$s \ge \frac{1}{p_j^{\mathrm{s}}} \left| \left(\frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \right)^{d_j} - p_j^\circ \right|$$

it is equivalent to match p_j^{1/d_j} to the target $(p_j^\circ)^{1/d_j}$ with a scale $(p_j^{\mathrm{s}})^{1/d_j}$,

$$s \ge \frac{1}{(p_j^{\mathrm{s}})^{1/d_j}} \left| \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} - (p_j^\circ)^{1/d_j} \right|$$

The resulting mathematical formulation is again identical to that with only type I structure−property relations.

The key advantage of reformulating (PM) and (PO) as an MILP problem is that now efficient MILP solvers such as CPLEX and OSL can locate the global optimum molecular design with mathematical certainty. Moreover, by incorporating appropriate integer cuts in the formulations (Floudas, 1995), not only can the best solution be found but also the second best, third best, etc., can be generated successively. For example, if y_{ik}^{sol} is the best solution and the generation of the second best solution is sought, then the following integer cut,

$$\sum_{k=1}^{K} \left[\sum_{(i,k):y_{i,k}^{\mathrm{sol}}=1} y_{ik} + \sum_{(i,k):y_{i,k}^{\mathrm{sol}}=0} (1 - y_{i,k}) \right] \le KN - 1$$

makes y_{ik}^{sol} infeasible, forcing the solver to converge to the second best solution. By accumulating integer cuts and resolving the problem multiple times, a sequence with the n-best solutions can be generated. In this section, it was shown how the original nonlinear formulations (PM) and (PO) involving structure−property relations of types I and II can be reformulated as mixed-integer linear programming problems. Next, a number of polymer design case study problems are addressed.

5. Polymer Design Case Studies

5.1. Case Study 1. The first case study involves the design of a polymer repeat unit which meets constraints and/or optimizes objectives on density, water absorption, and glass transition temperature (Derringer and Markham, 1985). The molecular groups which are allowed to participate in the polymer repeat unit are given in Table 1. The contribution of these molecular groups to the three properties of interest follows the empirical equations proposed by van Krevelen (1976) and all fall within type I. The expressions for the water absorption, density and glass transition temperature are given in Table 7. The values for the group contribution parameters H_i, M_i, Y_i, and V_i for different groups are given in Table 2. The same molecular group is allowed to participate up to three times in the polymer repeat unit, $n_i \in \{0, 1, 2, 3\}$, $i = 1, ..., 7$; however, no additional upper bound is imposed on the total number of groups in the repeat unit. The property targets in the (PM) formulation are

$$W^\circ = 0.005 \text{ (g of } H_2O\text{/g of polymer)}, \qquad D^\circ = 1.50 \text{ (g/cm}^3\text{)}, \qquad T_{\mathrm{g}}^\circ = 383 \text{ (K)}$$

and the property scaling factors are selected to be identical to the property targets.

$$W^{\mathrm{s}} = W^\circ, \qquad D^{\mathrm{s}} = D^\circ, \qquad T_{\mathrm{g}}^{\mathrm{s}} = T_{\mathrm{g}}^\circ$$

This ensures that percent deviations from the target values are penalized equally for all properties. The property bounds for the (PO) formulation are

$$0 \le W \le 0.18, \qquad 298 \le T_{\mathrm{g}} \le 673, \qquad 1 \le D \le 1.5$$

Both formulations are solved with the GAMS/CPLEX (Brooke *et al.*, 1988) combination on a HP-730 workstation with an absolute convergence tolerance of 10^{-8}.

Table 9 summarizes, in decreasing order of optimality, the 10 best solutions of the (PM) problem. Specifically, the polymer repeat unit, maximum scaled property violation, water absorption, glass transition temperature, density, and CPU in seconds are tabulated. The best repeat unit structure involves groups $-CH_2-$ and $-CHCl-$ in a 1:2 ratio. Note that, even in suboptimal solutions, these two groups in various ratios are dominant. The oxygen group $-O-$ only later appears in some of the suboptimal solutions. Note that the CPU requirements are very small (less than a second) and grow approximately linearly as more integer cuts accumulate (see Figure 1). This gives rise to only a quadratic increase in the cumulative CPU requirements (see Figure 2). This implies that much larger problems

Table 9. Results of Property Matching for Case Study 1

rank	repeating unit	violation	W	T (K)	D	CPU (s)
1	$-(CH_2(CHCl)_2)-$	0.0163	0.0049	384.68	1.4889	0.32
2	$-(CH_2(CHCl)_3)-$	0.0264	0.0051	393.10	1.5351	0.32
3	$-((CH_2)_2(CHCl)_3)-$	0.0526	0.0047	376.94	1.4489	0.32
4	$-(CHCl)-$	0.1134	0.0056	412.37	1.6524	0.38
5	$-(CH_2CHCl)-$	0.1169	0.0044	363.20	1.3827	0.44
6	$-((CH_2)_3O(CHCl)_3)-$	0.1674	0.0058	354.30	1.3977	0.50
7	$-((CH_2)_3O(CHCl)_2)-$	0.1843	0.0059	336.12	1.3333	0.55
8	$-((CH_2)_3(CHCl)_2)-$	0.1974	0.0040	346.04	1.3082	0.55
9	$-((CH_2)_3OCHCl)-$	0.2166	0.0061	301.41	1.2255	0.54
10	$-((CH_2)_2O(CHCl)_3)-$	0.2474	0.0062	366.23	1.4605	0.70

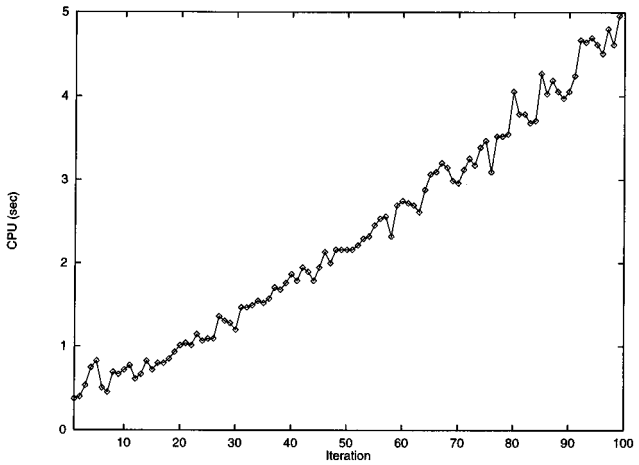

Figure 1. CPU requirements of the best 100 solutions of case study 1.

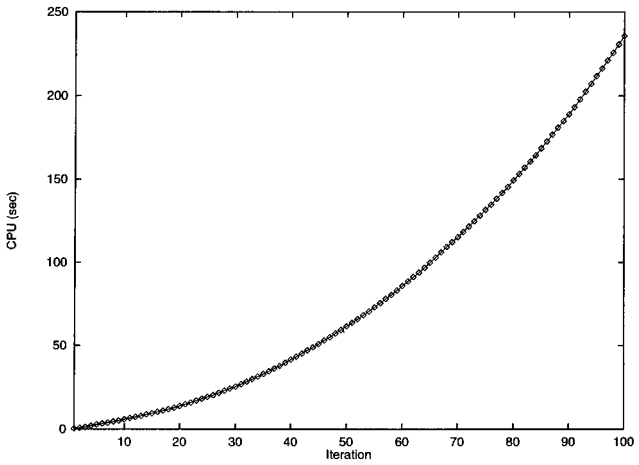

Figure 2. Cumulative CPU requirements of the best 100 solutions of case study 1.

and more complex property prediction models can be addressed. Clearly, the solution obtained depends on the adopted scaling; therefore, it is very important to select the scaling property factors in a way that truly reflects their relative importance.

Next, a (PO) problem formulation is addressed involving the minimization of the polymer water absorption W subject to bounds on the other properties. Table 10 summarizes, in increasing order of water absorption, the 10 best solutions of the water absorption minimization problem including the polymer repeat unit, water absorption, glass transition temperature, density, and CPU in seconds. The repeat unit structure which globally minimizes the water absorption involves again only groups $-CH_2-$ and $-CHCl-$ in a 3:1 ratio. In subsequent solutions, the same molecular groups again appear to dominate. The CPU requirements are again very small (about a second) and comparable with those of the property matching problem.

Table 10. Results of Water Absorption Minimization for Case Study 1

rank	repeating unit	W	T (°C)	D	CPU (s)
1	$-((CH_2)_3CHCl)-$	0.00318	37.49	1.1768	0.64
2	$-((CH_2)_2CHCl)-$	0.00368	59.03	1.2531	0.60
3	$-((CH_2)_3(CHCl)_2)-$	0.00401	73.04	1.3082	0.59
4	$-(CH_2CHCl)-$	0.00441	90.20	1.3827	0.81
5	$-((CH_2)_2(CHCl)_3)-$	0.00474	103.95	1.4488	0.77
6	$-(CH_2(CHCl)_2)-$	0.00492	111.68	1.4889	0.75
7	$-((CH_2)_3O(CHCl)_3)-$	0.00584	81.30	1.397	1.22
8	$-((CH_2)_3O(CHCl)_2)-$	0.00592	63.13	1.3333	0.97
9	$-((CH_2)_3OCHCl)-$	0.00608	28.41	1.2255	1.03
10	$-((CH_2)_2O(CHCl)_3)-$	0.00624	93.23	1.4605	0.88

5.2. Case Study 2. The second medium-size case study involves the optimal design of a polymer with property restrictions on the molecular weight M of the polymer repeat unit, glass transition temperature T_g, shear modulus G, density ρ, and specific heat capacity c_p. Three different design objectives are considered, involving the identification of the polymer repeat unit with (i) the maximum glass transition temperature T_g, (ii) maximum specific heat capacity c_p, and (iii) maximum specific shear modulus G/ρ.

Participating molecular groups, group contribution parameters, and estimating formulas are taken from example 4.3 of (Vaidyanathan and El-Halwagi, 1996). The molecular groups are shown in Table 3 and the property estimating formulas are found in Tables 7 and 8 after replacing the molar volume estimating expression, $\sum_{i=1}^{N} V_{ai} n_i$ with $1.435 \sum_{i=1}^{N} V_{wi} n_i$. Unlike example 4.3 of (Vaidyanathan and El-Halwagi, 1996), where fairly tight bounds were selected for all polymer properties, which excluded all but two polymer designs, the following much wider property bounds are chosen in this study.

$$50 \le M \le 200 \text{ (g/mol of monomer)}$$

$$298 \le T_g \le 500 \text{ (K)}$$

$$500 \le G \le 20000 \text{ (MPa)}$$

$$0.8 \le \rho \le 1.4 \text{ (g/cm}^3\text{)}$$

$$1.0 \le c_p \le 1.5 \text{ (J/(g K))}$$

This increases considerably the allowable molecular diversity and thus the difficulty of the optimization problem. Apart from the estimating formulas for shear modulus,

$$G = \left(\frac{\sum_{i=1}^{N} M_i n_i}{1.435 \sum_{i=1}^{N} V_{wi} n_i} \right) \left(\frac{\sum_{i=1}^{N} U_{Hi} n_i}{1.435 \sum_{i=1}^{N} V_{wi} n_i} \right)^6$$

Table 11. Results of Maximization of T_g in Case Study 2

molecular groups	T_g (K)	CPU (s)
$n_1 = 3, n_2 = 1, n_{12} = 2$	447.85	3.12
$n_1 = 1, n_3 = 1, n_{12} = 1$	440.38	2.14
$n_1 = 3, n_3 = 1, n_{12} = 2$	434.93	4.10
$n_2 = 1, n_3 = 1, n_6 = 1, n_{12} = 1$	432.49	4.37
$n_1 = 1, n_2 = 1, n_8 = 1, n_{12} = 1$	424.04	2.73
$n_1 = 3, n_2 = 1, n_8 = 1, n_{12} = 2$	423.22	2.85
$n_1 = 1, n_2 = 1, n_9 = 1, n_{12} = 1$	423.11	2.87
$n_1 = 3, n_2 = 1, n_6 = 1, n_{12} = 2$	417.72	3.25
$n_1 = 1, n_3 = 1, n_8 = 1, n_{12} = 1$	417.29	5.12
$n_1 = 1, n_2 = 1, n_6 = 1, n_{12} = 1$	415.69	5.50

Table 12. Results of Maximization of c_p in Case Study 2

molecular groups	c_p (J/g K)	CPU (s)
$n_1 = 2, n_6 = 3, n_{12} = 1$	1.713	1.40
$n_1 = 2, n_6 = 2, n_{12} = 1$	1.694	1.67
$n_1 = 3, n_6 = 2, n_{10} = 1, n_{12} = 1$	1.685	2.31
$n_1 = 2, n_6 = 1, n_{12} = 1$	1.666	2.46
$n_1 = 3, n_6 = 1, n_{10} = 1, n_{12} = 1$	1.664	2.27
$n_1 = 3, n_{10} = 1, n_{12} = 1$	1.636	2.30
$n_1 = 2, n_6 = 3, n_8 = 1, n_{12} = 1$	1.596	3.50
$n_1 = 2, n_6 = 2, n_8 = 1, n_{12} = 1$	1.567	4.90
$n_1 = 3, n_6 = 1, n_8 = 1, n_{10} = 1, n_{12} = 1$	1.560	4.63
$n_1 = 3, n_8 = 1, n_{10} = 3$	1.557	4.62

Table 13. Results of Maximization of G/ρ in Case Study 2

molecular groups	G/ρ (J/g)	CPU (s)
$n_8 = 1$	16868	0.99
$n_6 = 1, n_8 = 3$	11614	2.29
$n_2 = 1, n_8 = 3, n_{10} = 1$	10447	3.73
$n_6 = 1, n_8 = 2$	9976	2.09
$n_1 = 1, n_8 = 3, n_{10} = 1$	9887	1.65
$n_1 = 1, n_2 = 1, n_8 = 3, n_{10} = 1$	8998	1.65
$n_2 = 1, n_8 = 2, n_{10} = 1$	8837	1.79
$n_6 = 2, n_8 = 3$	8723	1.87
$n_1 = 2, n_8 = 3, n_{12} = 1$	8592	3.05
$n_1 = 1, n_8 = 2, n_{10} = 1$	8250	3.05

which is neither of type I nor type II, all other property estimating formulas are of type I or simpler (i.e., molecular weight). However, the lower and upper bounds on G can equivalently be imposed with upper and lower bounds on G/ρ whose estimating function,

$$G^{L}/\rho^{U} \le G/\rho = \left(\frac{\sum_{i=1}^{N} U_{Hi} n_i}{1.435 \sum_{i=1}^{N} V_{wi} n_i} \right)^6 \le G^{U}/\rho^{L}$$

is of type II. This means that the maximization of T_g, c_p, or G/ρ subject to lower and upper bounds on molecular weight M, shear modulus G, density ρ, and specific heat capacity c_p falls within the scope of the *property optimization* (PO) problem whose linear reformulation is presented in section 4. Therefore, existing efficient (MILP) solvers (i.e., OSL, CPLEX) can be utilized and guarantee the identification of the global optimum without having to resort to local (MINLP) solvers or to typically computationally intensive global (MINLP) solvers.

The maximum total number of molecular groups allowed to participate in the polymer repeat unit is seven, and the maximum number of the same molecular groups is four. The GAMS/OSL combination is utilized on an IBM RS6000 43P-133 series workstation with an absolute convergence tolerance of 10^{-8}. Computational results on the maximization of T_g, c_p, and G/ρ are shown in Tables 11−13, respectively. These results include the 10 best solutions in decreasing order of optimality and the associated computational requirements in seconds. Note that in all cases the computational requirements are only a few seconds, indicating the computational efficiency and practical feasibility of the proposed approach.

The results of the maximization of the glass transition temperature T_g show that the 10 best solutions are all within 7% of the best molecular design, indicating the existence of many "good" molecular designs (see Table 11). The best molecular design involves a fully branched main chain with methyl $-CH_3$ and chlorine $-Cl$ side groups in a 3:1 rato. Additional solutions also involve highly branched main chains with primarily methyl and to a lesser extent aromatic, chlorine, and amino groups. These results are consistent with the experimentally observed fact that highly branched chains typically involve high T_g.

The 10 best solutions for the maximization of the specific heat capacity c_p are again fairly close in value to the global optimum (within 9%). It is also interesting that all six best solutions involve only hydrocarbon groups (see Table 12). The global optimum solution involves a main chain with three out of four carbon atoms saturated with hydrogens and only one out of four carbon atoms fully branched with two methyl groups. Additional hydrocarbon solutions involve increasingly more methyl side groups per main chain carbon, starting with a ratio of 2:3 of methyl groups to main-chain atoms for the second best solution and finishing with a ratio of 3:2 in the sixth solution. Finally, in the last four solutions a single $-CHNH_2-$ group appears in methyl-branched carbon main chains.

The values of the 10 best solutions for the maximization of the specific shear modulus G/ρ involve a much higher spread of over 50% (see Table 13). The methylamino $-CHNH_2-$ group completely overwhelms all other groups in all 10 best molecular designs. More specifically, the best design involves only the group $-CHNH_2-$. The second best design, involving a 31% drop in the specific shear modulus value, includes groups $-CHNH_2-$ and $-CH_2-$ in a 3:1 ratio. Additional solutions involve main chains composed primarily of methylamino groups with occasional methyl and chlorine side groups.

5.3. Case Study 3. This large-scale molecular design case study involves the optimal design of a polymer repeat unit which most closely matches given property targets on density ρ, thermal expansion coefficient α, specific heat capacity c_p, and bulk modulus K. This design objective can be formulated as the minimization of the scaled deviation of ρ, α, c_p, and K from given target values, *property matching problem* (PM). The molecular groups composing the optimal polymer repeat unit are chosen from a diverse pool of 30 molecular groups (see Table 4) which are essentially identical to those utilized in (Venkatasubramanian *et al.*, 1995) with only minor modifications to exclude molecular groups with parameters of low accuracy (i.e., $-H$). Figure 3 illustrates the location of the free attachments in the aromatic molecular groups. The estimating formulas for density ρ, thermal expansion coefficient α, specific heat capacity c_p, and bulk modulus K are given in Tables 7 and 8, and the corresponding parameters (van Krevelen, 1990) in Table 5. Note that while the estimating expressions for ρ, α, and c_p are of type I, the formula for K is neither of type I nor of type II. This problem can be overcome by observing that, instead of trying to match a bulk modulus target, one can equivalently match the ratio of bulk modulus over

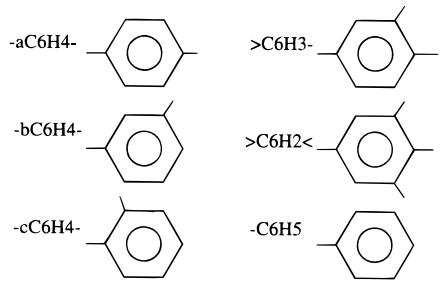

Figure 3. Molecular architectures of aromatic molecular groups of case study 3.

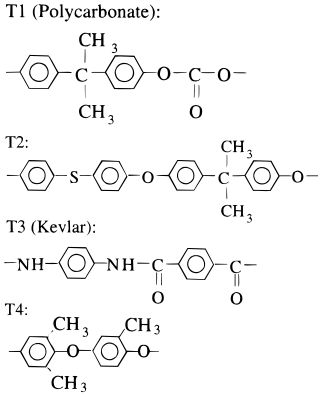

Figure 4. Molecular architectures of targets T1, T2, T3, and T4.

density, K/ρ, whose estimating formula,

$$K/\rho = \left(\frac{\sum_{i=1}^{N} U_{ri} n_i}{\sum_{i=1}^{N} V_i n_i} \right)^6$$

is of type II. Therefore, the minimization of the maximum scaled deviation of ρ, α, c_p, and K (alternatively K/ρ) from given target values falls within the scope of formulation (PM) with three property-estimating formulas of type I and one of type II. By selecting the property scales ρ^s, α^s, c_p^s, and K^s to be equal to the property targets $\rho°$, $\alpha°$, $c_p°$, and $K°$, the objective function to be minimized is the maximum percent property violation:

$$s = \max \left\{ \frac{\rho(\mathbf{n}) - \rho°}{\rho°}, \frac{\alpha(\mathbf{n}) - \alpha°}{\alpha°}, \frac{c_p(\mathbf{n}) - c_p°}{c_p°}, \frac{K(\mathbf{n})/\rho(\mathbf{n}) - K°/\rho°}{K°/\rho°} \right\}$$

Based on the linear reformulation of (PM), presented in section 4, an equivalent (MILP) representation is obtained.

Four different molecular design targets are examined: T1 (polycarbonate), T2, T3 (Kevlar), and T4. Their molecular architectures are shown in Figure 4. Table 14 summarizes the molecular groups participating in the repeat units of these four targets and gives their density, thermal expansion coefficient, heat capacity, and bulk modulus target values as predicted by the estimating formulas of Tables 7 and 8. For each case, not only the molecular design with the global minimum scaled property violation, as defined in section 2, but additionally the five best molecular designs are sought. The total number of molecular groups allowed to participate in the polymer repeat unit is $1 \le \sum_{i=1}^{30} n_i \le 7$, apart from the T2 target where as many as 10 molecular groups are allowed. Also, the maximum number of occurrences of the same molecular group in the polymer repeat unit is five, $0 \le n_i \le 5$, $i = 1, ..., 30$.

Table 14. Molecular Design Targets and Architecture of Case Study 3

polymer target	ρ (g/cm^3)	α (1/K)	c_p (J/g K)	K (N/m^2)
T1	1.1954	2.8817×10^{-4}	1.1350	5.2027×10^9
T2	1.1864	2.8895×10^{-4}	1.0740	5.2688×10^9
T3	1.3170	3.1338×10^{-4}	1.0111	9.6396×10^9
T4	1.0917	2.7686×10^{-4}	1.1631	4.1135×10^9

polymer target	molecular groups
T1	$n_1 = 1, n_9 = 1, n_{15} = 2, n_{20} = 2$
T2	$n_1 = 1, n_4 = 1, n_6 = 2, n_{15} = 4, n_{20} = 2$
T3	$n_7 = 1, n_{11} = 1, n_{12} = 1, n_{15} = 2$
T4	$n_6 = 2, n_{18} = 1, n_{19} = 1, n_{20} = 3$

the developed linear reformulation is solved by utilizing the GAMS/OSL interface on a IBM RS6000 43P-133 series workstation with an absolute convergence tolerance of 10^{-5}. Tables 15 and 16 summarize the computational results for the four design targets. In all cases, not only the target polymer with zero property−target discrepancies is recovered, but also the second, third, fourth, and fifth optimum designs are generated. Note that in (Venkatasubramanian *et al.*, 1995) (where an additional glass transition temperature target is employed) the T3 target was never identified for any of the implemented variants of the proposed genetic algorithm and, furthermore, the targets T1 and T4 were sometimes missed. The difficulty in identifying the T2 target is manifested in our approach by the increased CPU requirements.

Differences between the molecular groups participating in the T1 (polycarbonate) target and the second through fifth best solutions include the presence of a variety of aromatic groups, −NH− containing molecular groups, and the persistence of the side group $-tC_4H_9$. For the target T2 design case study, these differences are even more pronounced including a host of different hydrocarbon side groups, halogens such as chlorine and fluorine, and amino-containing groups. In the T3 (Kevlar) design study, the basic aromatic, amino, and carbonyl groups remain in the second through fifth best solutions, and additional groups include mainly oxygen and aromatic side groups. In the T4 design study, most notable is the persistence of the $-C_3H_7$ group in the second through fifth best designs. Table 15 indicates that in all cases the second best molecular design involves only a very small maximum scaled property violation (0.1−0.6%). Also, in going from the second best to the fifth best molecular design, the drop in the objective function is in all cases less than 3×10^{-3} and sometimes even less than 10^{-4}, indicating the presence of a substantial number of molecular designs with objective values very close to the global optimum. However, despite this proximity in objective function values the corresponding molecular architectures are often very different from each other.

6. Summary and Conclusions

In this paper, OMD problems with nonlinear structure−property functionalities which are or can be transformed to type I and/or II were addressed. It was shown how they can be transformed into equivalent mixed-integer linear (MILP) problems. While, in general, it is not possible to solve the original nonlinear OMD problem formulation for the best molecular design with mathematical certainty, the equivalent MILP

Table 15. Molecular Architectures of Solutions in Case Study 3

max violation	molecular groups
	T1 (Polycarbonate)
0.0000	$n_1 = 1, n_9 = 1, n_{15} = 4, n_{20} = 2$
0.0027	$n_2 = 1, n_9 = 1, n_{13} = 1, n_{15} = 1, n_{16} = 1, n_{18} = 1, n_{24} = 1$
0.0028	$n_9 = 1, n_{13} = 1, n_{15} = 1, n_{18} = 2, n_{20} = 1, n_{24} = 1$
0.0028	$n_9 = 1, n_{13} = 1, n_{17} = 1, n_{18} = 2, n_{20} = 1, n_{24} = 1$
0.0032	$n_3 = 1, n_9 = 1, n_{13} = 1, n_{15} = 1, n_{16} = 2, n_{24} = 1$
	T2
0.0000	$n_1 = 1, n_4 = 1, n_6 = 2, n_{15} = 4, n_{20} = 2$
0.0013	$n_1 = 1, n_6 = 1, n_7 = 1, n_{15} = 2, n_{18} = 1, n_{20} = 1, n_{21} = 1, n_{26} = 1$
0.0013	$n_1 = 1, n_6 = 1, n_7 = 1, n_{12} = 1, n_{15} = 2, n_{16} = 2, n_{24} = 1, n_{26} = 1$
0.0013	$n_1 = 1, n_6 = 1, n_{15} = 3, n_{16} = 1, n_{18} = 1, n_{21} = 1, n_{25} = 2$
0.0013	$n_3 = 1, n_{11} = 1, n_{16} = 1, n_{18} = 3, n_{24} = 1, n_{25} = 2, n_{26} = 1$
	T3 (Kevlar)
0.0000	$n_7 = 1, n_{11} = 1, n_{12} = 1, n_{15} = 2$
0.0022	$n_3 = 1, n_6 = 1, n_7 = 1, n_{14} = 1, n_{15} = 2, n_{29} = 1$
0.0032	$n_3 = 1, n_6 = 1, n_{14} = 1, n_{15} = 1, n_{18} = 1, n_{28} = 1, n_{29} = 1$
0.0035	$n_{10} = 1, n_{11} = 3, n_{16} = 1, n_{18} = 1, n_{20} = 1$
0.0048	$n_{12} = 1, n_{13} = 1, n_{14} = 1, n_{16} = 2, n_{18} = 1, n_{29} = 1$
	T4
0.0000	$n_6 = 2, n_{18} = 1, n_{19} = 1, n_{20} = 3$
0.0067	$n_6 = 2, n_{16} = 2, n_{19} = 1, n_{20} = 1, n_{22} = 1$
0.0068	$n_3 = 1, n_6 = 2, n_{16} = 3, n_{22} = 1$
0.0072	$n_2 = 1, n_6 = 2, n_{16} = 2, n_{18} = 1, n_{23} = 1$
0.0072	$n_2 = 1, n_6 = 2, n_{16} = 2, n_{18} = 1, n_{22} = 1$

Table 16. Property Values of Solutions in Case Study 3

ρ (g/cm^3)	α (1/K)	c_p (J/g K)	K (N/m^2)	CPU (s)
		T1 (Polycarbonate)		
1.1954	2.8882×10^{-4}	1.1351	5.2027×10^9	1687
1.1954	2.8854×10^{-4}	1.1374	5.1185×10^9	2675
1.1920	2.8892×10^{-4}	1.1341	5.2730×10^9	2187
1.1920	2.8892×10^{-4}	1.1341	5.1228×10^9	895
1.1990	2.8821×10^{-4}	1.1341	5.1231×10^9	2491
		T2		
1.1864	2.8895×10^{-4}	1.0740	5.2688×10^9	2973
1.1879	2.8869×10^{-4}	1.0742	5.2471×10^9	5543
1.1877	2.8892×10^{-4}	1.07533	5.3118×10^9	5671
1.1862	2.8912×10^{-4}	1.0729	5.2999×10^9	1955
1.1870	2.8932×10^{-4}	1.0727	5.2746×10^9	4504
		T3 (Kevlar)		
1.3170	3.1338×10^{-4}	1.0112	9.6396×10^9	261
1.3153	3.1330×10^{-4}	1.0134	9.6366×10^9	1648
1.31533	3.1360×10^{-4}	1.0109	9.4431×10^9	1775
1.3204	3.1259×10^{-4}	1.0142	9.8725×10^9	1848
1.3233	3.1211×10^{-4}	1.0099	9.5564×10^9	1622
		T4		
1.0917	2.7686×10^{-4}	1.1631	4.1135×10^9	13
1.0857	2.7818×10^{-4}	1.1654	3.9294×10^9	237
1.0980	2.7843×10^{-4}	1.1616	3.9702×10^9	90
1.0939	2.7884×10^{-4}	1.1562	4.0991×10^9	143
1.0939	2.7886×10^{-4}	1.1695	3.9688×10^9	89

reformulation obtained in this study of even large-scale OMD problems was solved with commercially available MILP solvers for the best, second best, third best, etc., molecular designs. Two popular design objectives were addressed: (i) minimization of the maximum scaled deviation of design properties from some target values, *property matching* (PM), and (ii) minimization/maximization of a single property subject to lower and upper bounds for the remaining properties, *property optimization* (PO). The approach was customized to the design of polymers whose properties were estimated with widely used group contribution methods (van Krevelen, 1990).

The results obtained in three case studies illustrate the ability of the approach to uncover a plethora of molecular designs and diverse molecular architectures capable of meeting various design objectives in a globally optimum way. By removing the possibility of converging to suboptimal solutions, possible discrepancies between obtained optimal solutions and experimentally derived designs can be solely attributed to property estimation imprecision. Based on this, a prediction−correction loop for improving property prediction accuracy can be constructed. While in some cases the best molecular design is clearly superior over the second, third, etc., best molecular designs, in most cases such clear-cut distinction is absent. Considering this prevailing proximity in objective value of the second, third, or even tenth best molecular design to the global optimum one, it is equally important that the approach is guaranteed not to miss any of the *n*-best molecular designs with the identification of the global optimum molecular design with mathematical certainty.

Note that, although the proposed analysis framework was applied only to polymer design case studies, the same underlying mathematical features are present to some extent in other OMD problems such as optimal design of agrochemicals, refrigerants, and solvents. While in the case studies that were addressed in this paper all property-estimating expressions belonged or were transformed into type I or II, this may not always be the case. Nevertheless, even if full linearization is not possible, partial linearization of the OMD model, along the lines of the analysis presented in this paper, is bound to aid any mathematical optimization-based solution approaches. Furthermore, continuing efforts at solving globally OMD problems will greatly benefit from a standardization of the type of functionalities utilized in property-estimating formulas. This will greatly facilitate the algorithmic development for OMD problems.

A key issue in CAMD is the inherent discrepancies between property estimations and actual experimental data. In polymer design, sources of error can be due to polydispersity, cross-linking, and different grades of the same material in the final product. Incorporating information on molecular group interconnection in the property-estimating formulas is likely to reduce but not eliminate discrepancies with experimental data. Therefore, property prediction imprecision is likely to remain a key issue in OMD. Quantifying the effect of property prediction imprecision within a probabilistic framework and studying its effect on meeting design objectives is addressed in (Maranas, 1996).

Acknowledgment

Financial support by Du Pont's Educational Aid Grant is gratefully acknowledged.

Literature Cited

Allcock, H. R. Rational Design and Synthesis of New Polymeric Materials. *Science* **1992**, *255* (5048), 1106.

Babonneau, F. Molecular Design of Advanced Ceramics. EUROMAT 93: 3rd European Conference on Advanced Materials and Processes, Paris, France, 1993.

Blackletter, D. Computer-Aided Plastic Selection. *Mater. Eng.* **1988**, *34*.

Bolis, G.; DiPace, L.; Fabrocini, F. A machine learning tool for computer aided molecular design. Third International Conference on Tools for Artificial Intelligence, San Jose, CA, 1991.

Brignole, E. A.; Bottini, S.; Gani, R. Strategy for the Design and Selection of Solvents for Separation Processes. *Fluid Phase Equilib.* **1986**, *29*, 125.

Brooke, A.; Kendrick, D.; Meeraus, A. *GAMS: A User's Guide*; Scientific Press: Palo Alto, CA, 1988.

Dare-Edwards, M. P. Novel Family of Traction Fluids Deriving from Molecular Design. *J. Synth. Lubr.* **1991**, *8* (3), 197.

Derringer, G. C.; Markham, R. L. A Computer-Based Methodology for Matching Polymer Structures with Required Properties. *J. Appl. Polym. Sci.* **1985**, *30*, 4609.

Dobrowolski, J. A. *Usual and unusual applications of optical thin films*; Springer-Verlag, New York, 1994.

Duvedi, A. P.; Achenie, L. E. K. A MINLP model for the design of refrigerant mixtures. Annual Meeting of Chemical Engineers, Miami, FL, 1995.

Epstein, L. I. Design of Optical Filters. *J. Opt. Soc. Am.* **1952**, *42*, 806.

Floudas, C. A. *Nonlinear and Mixed-Integer Optimization*; Oxford University Press: New York, 1995.

Franklin, J. L. Prediction of heat and free energies of organic compounds. *Ind. Eng. Chem.* **1949**, *41* (51), 1070.

Gani, R.; Fredenslund, A. Computer-Aided Molecular and Mixture Design with Specified Property Constraints. *Fluid Phase Equilib.* **1993**, *82*, 39.

Gani, R.; Tzouvaras, N.; Rasmussen, P.; Fredenslund, A. A. Prediction of Gas Solubility and Vapor−Liquid Equilibria by Group Contribution. *Fluid Phase Equilib.* **1989**, *47* (2), 133.

Gani, R.; Nielsen, B.; Fredenslund, A. A Group Contribution Approach to Computer-Aided Molecular Design. *AIChE J.* **1991**, *37* (9), 1318.

Giannelis, E. P. Molecular Engineering of Ceramics. Chemical Approaches to the Design of Materials. *Eng.: Cornell Q.* **1989**, *23* (2), 15.

Glover, F. Improved Linear Integer Programming Formulations of Nonlinear Integer Problems. *Manage. Sci.* **1975**, *22* (4), 455.

Gordeeva, E. V.; Molcharova, M. S.; Zefirov, N. S. General Methodology and Computer Program for the Exhaustive Restoring of Chemical Structures by Molecular Connectivity Indices. Solution of the Inverse Problem in QSAP/QSPR. *Tetrahedron Comput. Methodol.* **1990**, *3*, 389.

Halpin, J. C. Evolution of design and material criteria for polymeric structures. *Compos. Struct.* **1994**, *27* (1), 3.

Horvath, A. L. *Molecular Design*; Elsevier Science Publishers B. V.: Amsterdam, The Netherlands, 1992.

Ichiro, M. Molecular Design of Lubricants. *J. Jpn. Soc. Tribol.* **1991**, *36* (4), 280.

Joback, K. G. A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques. M.S. Thesis, MIT, Cambridge, MA, 1984.

Joback, K. G. Designing Molecules Possessing Desired Physical Properties. Ph.D. Thesis, MIT, Cambridge, MA, 1989.

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

Joback, K. G.; Stephanopoulos, G. Designing Molecules Possessing Desired Physical Property Values. FOCAPD '89; Snowmass, CO, 1989; p 363.

Judas, D.; Germain, Y.; Biver, C.; Girault, S.; Taupin, C. Polymer design for high-performance materials. Liquid-crystal polymers. *Nature* **1991**, *350* (6319), 28.

Kier, L. B.; Lowell, H. H.; Frazer, J. F. Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 142.

Kopko, W. L. Beyond CFCs. Extending the search for new refrigerants. *Int. J. Refrig.* **1990**, *13* (2), 79.

Liebman, J.; Lasdon, L.; Schrage, L.; Waren, A. *Modeling and Optimization with GINO*; The Scientific Press: Palo Alto, CA, 1986.

Maranas, C. D. Optimal Molecular Design Under Property Prediction Uncertainty. **1996**, submitted to *AICHE J.*

Mehrotra, R. C. Molecular Design of Advanced Ceramic Materials. Materials Research Society Spring Meeting, San Francisco, CA, 1992.

Nagasaka, K.; Wada, H.; Yoshimitsu, H.; Yasuda, H.; Yamanouchi, T. Expert System for Polymer Design. AIChE Annual Meeting 39e, Chicago, IL, 1990.

Naser, S. F.; Fournier, R. L. A System for the Design of an Optimum Liquid−Liquid Extractant Molecule. *Comput. Chem. Eng.* **1991**, *15* (6), 397.

Nielsen, B.; Gani, R. Computer-Aided Molecular Design By Group Contribution. European Symposium on Computer Applications in Chemical Engineering, The Hague, The Netherlands, 1990.

Odele, O.; Macchietto, S. Computer Aided Molecular Design: A Novel Method for Optimal Solvent Selection. *Fluid Phase Equilib.* **1993**, *82*, 47.

Odele, O.; Macchietto, S.; Omatsone, O. Design of Optimal Solvents for Liquid−Liquid Extraction and Gas Absorption Processes. *Trans. IChemE* **1990**, *68*, 429.

Stephanopoulos, G.; Townsend, D. W. Synthesis in Process Development. *Chem. Eng. Res. Dev.* **1986**, *64*, 160.

Thelen, A. *Design of Optical Interference Coatings*; McGraw-Hill: New York, 1989.

Vaidyanathan, R.; El-Halwagi, M. Computer-aided design of high performance polymers. *J. Elastomers Plast.* **1994**, *26* (3), 277.

Vaidyanathan, R.; El-Halwagi, M. Computer Aided Synthesis of Polymers and Blends with Target Properties. *Ind. Eng. Chem. Res.* **1996**, *35* (2), 627.

van Krevelen, D. W. *Properties of Polymers*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 1976.

van Krevelen, D. W. *Properties of Polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 1990.

Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Computer-Aided Molecular Design using Genetic Algorithms. *Comput. Chem. Eng.* **1994a**, *18* (9), 833.

Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. On the Performance of Genetic Search for Large-Scale Molecular Design. PSE '94, Korea, 1994b; p 1001.

Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 188.

Received for review February 9, 1996
Revised manuscript received June 14, 1996
Accepted June 17, 1996[⊗]

IE960096Z

[⊗] Abstract published in *Advance ACS Abstracts,* September 15, 1996.