Article

# Optimization in Polymer Design Using Connectivity Indices

Kyle V. Camarda, and Costas D. Maranas

*Ind. Eng. Chem. Res.*, **1999**, 38 (5), 1884-1892• DOI: 10.1021/ie980682n • Publication Date (Web): 23 March 1999

MATERIALS AND INTERFACES

# Optimization in Polymer Design Using Connectivity Indices

**Kyle V. Camarda and Costas D. Maranas***

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

This work addresses the design of polymers with optimal levels of macroscopic properties through the use of topological indices. Specifically, two zeroth-order and two first-order connectivity indices are for the first time employed as descriptors in structure−property correlations in an optimization study. Based on these descriptors, a set of new correlations for heat capacity, cohesive energy, glass transition temperature, refractive index, and dielectric constant are proposed. These correlations are incorporated into an optimization framework. The proposed mathematical description, utilizing the concept of a basic group, accounts fully for molecular interconnectivity. When the nonconvex terms are appropriately recast in the formulation as convex inequalities, a convex mixed-integer nonlinear (MINLP) representation is obtained. Three example problems highlight the proposed molecular description, structure−property correlations, convex MINLP optimization formulation, and solution technique.

## 1. Introduction and Background

The design of new polymer products with optimal levels of thermophysical, mechanical, and optical properties is an important goal in polymer engineering. While traditional polymer design requires an extensive series of experiments to synthesize and evaluate candidate molecules, computer−aided molecular design (CAMD) methods expedite this process by identifying promising polymer designs in advance, thereby limiting the number of experiments which must be performed. CAMD requires that the polymer properties of interest be expressed as a function of the structure of the polymer repeat unit.

In previous work,[1−6] group contribution methods have been extensively used to establish input−output relations between the type and number of molecular groups in a polymer repeat unit and various macroscopic properties. The additivity assumption is invoked, and contributions from each molecular group are combined to give a property estimation which is an interpolation of property values of similar compounds. One of the limitations of group contribution estimation is that the internal molecular structure of the polymer repeat unit is only partially taken into account. For example, both polypropylene, $-CH_2CH(CH_3)CH_2CH(CH_3)-$, and head to head polypropylene, $-CH_2CH(CH_3)CH(CH_3)CH_2-$, have the same molecular group representation. Furthermore, the selection of the molecular groups, at least in the context of polymer design, is somewhat arbitrary. Larger molecular groups yield more accurate property prediction; however, a larger number of them is required to ensure that every possible candidate repeat unit can be constructed as their linear combination. On the other hand, smaller molecular groups provide less accurate estimates, but a smaller number of them are required. In addition, extrapolations, meaning estimations outside the range of the experimental data on which the group contribution parameters were derived, can be of problematic accuracy.[7]

We propose to alleviate these shortcomings with the use of property correlations involving topological indices as structural descriptors. These indices are numerical values which identify the polymer repeat unit and contain information about the atomic and electronic structure. Examples of topological indices include Kier's shape indices,[8,9] which encode information about cyclicity and branching, the Wiener index,[10] which is a path number representing the number of bonds between all pairs of atoms in the molecule, and Randić's molecular connectivity indices,[11] which contain information about the bonding environment of each non-hydrogen atom, the electronic structure of each atom, and larger scale features such as rings and branches. Specifically, Bicerano[7] used the zeroth- and first-order molecular connectivity indices to correlate a wide range of polymer properties, including density, glass transition temperature, bulk modulus, and heat capacity.

A review of many of the applications of topological indices is given by Trinajstic.[12] Many studies[13−15] have explored structure generation based on these indices. Skvortsova et al.[16] used an exhaustive generation of molecular graphs to design molecules whose properties are functions of Kier's shape indices. Kier et al.[17] designed molecules with a target molar volume and correlated this volume with the first- and second-order connectivity indices. All of these methods, however, use as descriptors structural features (e.g., number of bonds between atoms at a specific valency state) that do not necessarily define a feasible or unique molecular graph. Raman and Maranas[18] first incorporated connectivity indices within an optimization framework. A number

* Author to whom all corespondence should be addressed. Phone: (814)863-9958. E-mail: costas@psu.edu. Fax: (814)-865-7846.

10.1021/ie980682n CCC: $18.00 
Published on Web 03/23/1999

of hydrocarbon properties were correlated with different connectivity indices, and a mixed-integer linear problem (MILP) representation providing full connectivity information was derived. The developed MILP representation, however, was applicable only to hydrocarbon systems.

In this work, a complete representation of the molecular structure of the repeat unit is used which accounts for heteroatoms. The concept of a *basic group* is defined as any non-hydrogen atom at an allowable electronic configuration (i.e., valency state) bonded to a given number of hydrogen atoms (e.g., $CH_3-$, $-CH_2-$, O=, HO−, Cl−, etc.). A *basis set* is defined as a prepostulated set of basic groups which serve as building blocks for constructing polymer repeat units. Given that all non-hydrogen elements participating in the polymer repeat unit are included in the basis set at allowed valency states, any polymer repeat unit can then be constructed with the basic groups contained in the basis set. In this work, the basic groups employed have been chosen from a basis set proposed by Bicerano.[7] This set includes elements O, C, N, S, Cl, Br, and F at all their possible valency states. Any polymer containing only those elements can be created with the elements of this basis set. The advantage of forming the basis set exclusively from basic groups is that (i) any arbitrariness in the selection of molecular groups is avoided, (ii) the size of the basis set remains manageable (less than 20) even in the presence of many heteroatoms, and (iii) no bias is introduced in the type of polymer repeat units because of the internal structure of the adopted molecular groups. The loss of accuracy by using small groups is offset by the incorporation of topological and in particular connectivity indices which encode connectivity information. The following additive property predictive form is utilized:

$$\text{property prediction} = \text{basic group contribution} + \text{connectivity indices contribution}$$

Unlike most of the property prediction relations introduced by Bicerano[7] where numerous correction terms are included, the proposed prediction framework does not include such terms, which makes it amenable to mathematical optimization. Therefore, new correlations for polymeric properties such as heat capacity and glass transition temperature are developed and along with the linear ones by Bicerano[7] are utilized in our optimization formulation. Two different design philosophies are explored here, as in work by Maranas.[6] In the first, referred to as property matching, the maximum scaled deviation from a set of target property values is minimized:

$$\min \max_m \frac{1}{P_m^{\text{scale}}} |p_m - P_m^{\text{target}}|$$

where $P_m^{\text{target}}$ is the target for property $m$ and $P_m^{\text{scale}}$ the corresponding scale. The second design option is to minimize or maximize one property, subject to bounds on the remaining ones.

A key challenge after establishing reliable structure−property relations is addressing the large complexity encountered in optimization studies. Joback and Stephanopoulos[19] showed that the total number of distinct ways of selecting between $K_{\min}$ and $K_{\max}$ molecular groups from a list of $N$ molecular groups is equal to

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

For example, this yields 10 272 278 169 distinct ways of selecting up to 10 molecular groups from a list of 40. On top of this complexity, the number of ways that a selected set of molecular groups can be interconnected in a feasible manner increases exponentially with the total number of basic groups in the molecule.[12] Therefore, the increased complexity associated with the incorporation of connectivity information implies that optimization approaches that were sufficient for group contribution may no longer be tractable.

## 2. Topological Indices in Polymer Design

Bicerano[7] first developed correlations for many macroscopic polymeric properties by utilizing molecular connectivity indices as some of the structural descriptors. The calculation of these connectivity indices requires the concept of hydrogen-suppressed molecular graphs. A *hydrogen−suppressed molecular graph* is the graph representation of a molecule (or repeat unit) where all non-hydrogen atoms are represented as vertices and all bonds between non-hydrogen atoms are modeled as edges. Each one of the non-hydrogen atoms defines an indivisible molecular group which is referred to as a *basic group*. Therefore, a one to one mapping exists between basic groups and vertices in the molecular graph.

A molecular graph can be described with a variety of matrices[12] such as a vertex adjacency matrix, an edge adjacency matrix, an incidence matrix, a cycle matrix, or a distance matrix. The vertex adjacency matrix representation of a molecular graph is employed in this paper and is simply referred to as the adjacency matrix in subsequent sections. The adjacency matrix of a simple graph is a $N \times N$ matrix, where $N$ is the number of vertices (basic groups) in the graph. It is defined as

$$\mathbf{A} = (a_{ij})$$

$$a_{ij} = \begin{cases} 1 & \text{if vertex } i \text{ is connected to vertex } j \\ 0 & \text{otherwise} \end{cases}$$

where $i, j = 1, ..., N$ is the set of vertices. The adjacency matrix is symmetric, and the diagonal elements are zero. Note that the adjacency matrix defined above cannot distinguish between different types of bonds (e.g., single vs double). To remedy this shortcoming, we define a family of three complementary adjacency matrices $\mathbf{A}^k$, $k = 1$, 2, and 3 to account for single, double, and triple bonds, respectively. Specifically, an element $a_{ijk}$ of this family of adjacency matrices is equal to 1 if basic groups $i$ and $j$ are bonded with a $k$-type bond. Otherwise, $a_{ijk}$ is equal to 0. While the binary variables $a_{ijk}$ uniquely and unambiguously determine the connectivity of the molecule, no information is provided regarding the types of basic groups participating at vertices $i = 1, ..., N$. To this end, a type-vertex binary variable $y_{il}$ is defined which is equal to 1 if vertex $i$ is occupied by a basic group of type $l$ and 0 otherwise.

In this work, both zeroth- and first-order connectivity indices are employed as structural descriptors. The zeroth- and first-order connectivity indices are defined as the weighted sum over all of the vertices and edges in the graph, respectively. These weights depend on the

**Table 1. Basic Group Coefficients for the Two Weighting Schemes**

| basic group | $\delta$ | $\delta^v$ | basic group | $\delta$ | $\delta^v$ |
|---|---|---|---|---|---|
| $-CH_3$ | 1 | 1 | $-N<$ | 3 | 5 |
| $-CH_2-$ | 2 | 2 | $-OH$ | 1 | 5 |
| $-CH<$ | 3 | 3 | $=O$ | 1 | 6 |
| $>C<$ | 4 | 4 | $-O-$ | 2 | 6 |
| $=CH-$ | 2 | 3 | $-F$ | 1 | 7 |
| $=C<$ | 3 | 4 | $-Cl$ | 1 | $^7/_9$ |
| $-NH-$ | 2 | 4 | | | |

specific electronic configuration of the basic groups. Two different such weighting schemes are considered which yield two zeroth-order and two first-order connectivity indices. The coefficients $\delta$ for the first weighting scheme are equal to the total number of edges emanating from vertex $i$. This is equal to the the total number of bonds that basic group $i$ forms with other basic groups (double and triple bonds are only counted once). The coefficients $\delta^v$ for the second weighting scheme incorporate information about the number of valence and inner-shell electrons of the non-hydrogen atoms of each basic group. Detailed information about these weights can be found in Bicerano.[7] Table 1 lists the values of the coefficients $\delta$ and $\delta^v$ for all basic groups employed in this work.

The two zeroth-order connectivity indices, ${}^0\chi$ and ${}^0\chi^v$, are related to the bonding configuration and the electronic structure of each non-hydrogen atom in the polymer repeat unit. They contain information about valence shell hybridization, inner-shell electrons, and lone pairs. These indices are defined by the relations

$$ {}^0\chi = \sum_{i \in N} \frac{1}{\sqrt{\delta_i}} $$

$$ {}^0\chi^v = \sum_{i \in N} \frac{1}{\sqrt{\delta_i^v}} $$

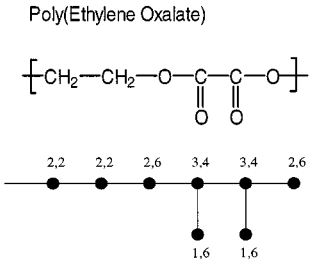
where $N$ is the vertex set in the molecular graph. The two first-order connectivity indices, ${}^1\chi$ and ${}^1\chi^v$, encode information about the electronic structure of each bonded set of atoms, including $\sigma$ and $\pi$ electrons. They are computed using the following equations

$$ {}^1\chi = \sum_{(i,i') \in \varepsilon} \frac{1}{\sqrt{\delta_i \delta_{i'}}} $$

$$ {}^1\chi^v = \sum_{(i,i') \in \varepsilon} \frac{1}{\sqrt{\delta_i^v \, \delta_{i'}^v}} $$

where $\varepsilon$ is the molecular graph edge set.

An example of the computation of connectivity indices is given in Figure 1. The molecular structure of the repeat unit of poly(ethylene oxalate) is shown, along with the hydrogen-suppressed graph for the repeat unit. The numbers above each vertex of the graph correspond to the values of $\delta$ and $\delta^v$, respectively. Values for the connectivity indices ${}^0\chi$, ${}^0\chi^v$, ${}^1\chi$, and ${}^1\chi^v$ are also listed. Note that the leftmost $-CH_2-$ is bonded to the $-O-$ group of the next repeat unit. This is modeled by treating the two end groups of the repeat unit as if they were bonded. This implies that the connectivity index representation of a polymer repeat unit corresponds to a ring structure.




$${}^0\chi = 5.9831 \qquad {}^0\chi^v = 4.0472$$

$${}^1\chi = 3.8045 \qquad {}^1\chi^v = 2.1438$$

**Figure 1.** Structure and hydrogen-suppressed graph for the repeat unit of poly(ethylene oxalate).

In optimization studies, the number, type, and labeling order of atomic groups are unknown. Based on the definition of the decision binary variable $y_{il}$, the linear defining relations for the zeroth-order connectivity indices ${}^0\chi$ and ${}^0\chi^v$ are as follows:

$$ {}^0\chi = \sum_{i=1}^{N} \sum_{l=1}^{L} \frac{y_{il}}{\sqrt{\delta_l}} $$

$$ {}^0\chi^v = \sum_{i=1}^{N} \sum_{l=1}^{L} \frac{y_{il}}{\sqrt{\delta_l^v}} $$

The corresponding relations for the first-order connectivity indices are

$$ {}^0\chi = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\sum_{k=1,2,3} a_{ijk}}{\sqrt{\Delta_i \Delta_j}} $$

$$ {}^1\chi^v = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\sum_{k=1,2,3} a_{ijk}}{\sqrt{\Delta_i^v \, \Delta_j^v}} $$

where

$$ \Delta_i = \sum_{l=1}^{L} \delta_l y_{il}, \qquad i = 1, \ldots, N $$

and

$$ \Delta_i^v = \sum_{l=1}^{L} \delta_l^v y_{il}, \qquad i = 1, \ldots, N $$

Unlike the case for the zeroth-order connectivity indices, the linking relation between the decision variables $a_{ijk}$ and $y_{il}$ and the first-order connectivity indices is nonlinear. In a later section it is shown that these nonlinear linking relations can equivalently be recast in the form of convex inequalities. Next, the use of these connectivity indices in structure−property relations for polymeric properties is discussed.

## 3. Structure−Property Correlations

Bicerano[7] first correlated a number of polymeric properties using connectivity indices and correction

**Table 2. Coefficients for Structure−Property Correlations**

| | properties | | | | |
|---|---|---|---|---|---|
| descriptors | $Cp$ ($\times 10^3$ J/kg·K) | $E_{coh}$ ($\times 10^4$ J/mol) | $T_g$ ($\times 10^2$ K) | $n$ | $\epsilon$ |
| $^0\chi$ | 4.26 | −279 | 0.0861 | 1.16 | 1.15 |
| $^0\chi^v$ | −4.02 | 143 | −0.0186 | −0.294 | 0.000 |
| $^1\chi$ | −1.47 | 0.812 | −0.00620 | −0.436 | 7.54 |
| $^1\chi^v$ | 2.32 | −0.297 | 0.152 | 1.22 | −2.21 |
| $-CH_3$ | 2.37 | 137 | −0.244 | −0.139 | −1.46 |
| $-CH_2-$ | 1.50 | 96.7 | 0.180 | 0.0248 | −1.50 |
| $-CH<$ | 0.487 | 78.9 | 0.888 | 0.270 | −1.26 |
| $>C<$ | −0.628 | 68.2 | 1.59 | 0.592 | 1.27 |
| $=CH-$ | 1.05 | 115 | 0.221 | 0.343 | −1.35 |
| $=C<$ | 0.290 | 90.0 | 0.978 | 0.532 | 1.03 |
| $-NH-$ | 2.80 | 128 | 1.23 | 0.711 | 3.92 |
| $-N<$ | 0.000 | 97.8 | 1.34 | 1.28 | 3.26 |
| $-OH$ | 0.000 | 218 | 0.000 | 0.000 | 6.20 |
| $-O-$ | 0.834 | 139 | 0.193 | 0.424 | −0.64 |
| $=O$ | 0.00 | 222 | −0.0787 | 0.0447 | 1.63 |
| $-F$ | −0.730 | 225 | −0.376 | 0.190 | 1.99 |
| $-Cl$ | 1.49 | 118.2 | 0.179 | 0.0059 | 0.647 |

**Table 3. Statistics for Structure−Property Correlations**

| | | % | | $r$ | |
|---|---|---|---|---|---|
| property | no. of polymers | Bicerano | this work | Bicerano | this work |
| $Cp$ | 74 | 5.0 | 4.5 | 0.9938 | 0.9049 |
| $E_{coh}$ | 113 | 3.9 | 4.4 | 0.9974 | 0.9984 |
| $T_g$ | 330 | 6.7 | 14.2 | 0.9749 | 0.9049 |
| $n$ | 167 | 1.0 | 2.55 | 0.9770 | 0.8906 |
| $\epsilon$ | 58 | 3.0 | 4.16 | 0.9788 | 0.9799 |

terms. These correction terms can be classified as atomic, which include only the number of atoms in a given electronic configuration, or structural, which refer to functional groups and their locations within the repeat unit. While the atomic terms are usually linear, the structural correction terms are nonlinear and cumbersome to express as the function of the design variables $y_{il}$ and $a_{ijk}$. In the interest of maintaining a standardized form for the structure−property relations, only the linear correlations of Bicerano[7] are used. These are as follows:

Molar Volume ($cm^3$/mol):[7]

$$V_w = 3.8618\,{}^0\chi + 13.748\,{}^1\chi^v$$

Molar Diamagnetic Susceptibility ($10^{-6}$ $cm^3$/mol):[7]

$$\zeta_m = 14.4516\,{}^0\chi^v$$

For the rest of the polymeric properties studied, structure−property correlations of the following form are developed:

$$P = \frac{1}{n}\left(\sum_{l=1}^{L} C^l n_l + C^0\,{}^0\chi + C^{0v}\,{}^0\chi^v + C^1\,{}^1\chi + C^{1v}\,{}^1\chi^v\right)$$

where

$$n_l = \sum_{i=1}^{N} y_{il}$$

Here $L$ is the number of different types of basic groups used, $n_l$ is the number of basic groups of type $l$ participating in the polymer repeat unit, $n$ is the total number of basic groups, and $C^l$, $C^0$, $C^{0v}$, $C^1$, $C^{1v}$ are the correlation coefficients. Note that the denominator is set equal to 1 for the correlating expression for the cohesive energy (extensive property). The presence of the total number of basic groups in the denominator of the correlating expressions for intensive properties is necessary to ensure that the property estimates are independent of the size of the polymer repeat unit. The correlation coefficients are calculated based on a linear least-squares fit of experimental data found in the software package SciPolymer[20] by SciVision, Inc. The correlating coefficients for the heat capacity, cohesive energy, glass transition temperature, refractive index, and dielectric constant are listed in Table 2. The size of the data set, standard deviation as a percentage of average property value (% sd), and the correlation coefficient $r$, both for the correlations developed here and those of Bicerano,[7] are listed in Table 3. The standard deviation sd and the standard deviation as a percentage of the average property value % sd are defined as follows:

$$\mathrm{sd} = \frac{1}{N-K-1}\sqrt{\sum_{i=1}^{N}(P_i^{\mathrm{pred}} - P_i^{\mathrm{exp}})^2}$$

$$\%\ \mathrm{sd} = 100\frac{\mathrm{sd}}{P^{\mathrm{ave}}}$$

where

$$P^{\mathrm{ave}} = \frac{1}{N}\sum_{i=1}^{N} P_i^{\mathrm{exp}}$$

Here $i = 1, \ldots, N$ is the set of experimental values used in the correlation, $K$ is the number of descriptors, $P_i^{\mathrm{pred}}$ is the predicted property value, $P_i^{\mathrm{exp}}$ is the experimentally measured property value, and $P^{\mathrm{ave}}$ is the average over all of the experimentally measured property values. Comparisons of the correlations derived in this work with those of Bicerano[7] indicate that their accuracy is slightly inferior despite the elimination of all nonlinear terms. While the correlations derived in this work include more terms, all of these terms are linear and thus lead to linear constraints within our optimization framework. However, because a smaller set of polymers were used to derive the correlations, incorrect values can be generated for polymers which are structurally very different from those used in the correlation. Also, the correlation for the glass transition temperature is less accurate than that of Bicerano, because an accurate quantification of chain stiffness is required to characterize this behavior. Zeroth- and first-order connectivity indices alone do not capture this phenomenon well. Instead, higher order connectivity indices become necessary to improve accuracy.

## 4. Convex Representation of Nonlinear Terms

The relations which link the design variables $y_{il}$ and $a_{ijk}$ and the proposed structure−property correlations through the use of first-order connectivity indices are nonlinear. These nonlinearities arise because the defining expressions for $^1\chi$ and $^1\chi^v$ involve trilinear terms. where

$$ {}^1\chi = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\sum_{k=1,2,3} a_{ijk}}{\sqrt{\Delta_i \Delta_j}} $$

$$ {}^1\chi^{\mathrm{v}} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{\sum_{k=1,2,3} a_{ijk}}{\sqrt{\Delta_i^{\mathrm{v}} \Delta_j^{\mathrm{v}}}} $$

$$ \Delta_i = \sum_{l=1}^{L} \delta_l y_{il}, \qquad i = 1, ..., N $$

and

$$ \Delta_i^{\mathrm{v}} = \sum_{l=1}^{L} \delta_l^{\mathrm{v}} y_{il}, \qquad i = 1, ..., N $$

When the new variables

$$ X_{ij} = \frac{1}{\sqrt{\Delta_i \Delta_j}}, \qquad i = 1, ..., N, \quad j = i + 1, ..., N $$

are introduced, the trilinear terms, summing up to ${}^1\chi$, can be rewritten as the product of a continuous variable $X_{ij}$ times a binary variable $\sum_{k=1,2,3} a_{ijk}$ (only one type of bond is allowed between two basic groups). The continuous times binary variable products can then be exactly linearized using the Glover[21] transformation. The defining relation for $X_{ij}$ can be written as the following two-sided inequality:

$$ \frac{1}{\sqrt{\Delta_i \Delta_j}} \le X_{ij} \le \frac{1}{\sqrt{\Delta_i \Delta_j}}, \quad i = 1, ..., N, \quad j = i + 1, ..., N $$

In Appendix B of Maranas and Floudas[22] it is shown that the term $x^a y^b$ (for $x, y \ge 0$ ) is convex if $a, b \le 0$ and concave if $a, b \ge 0$ and $a + b \le 1$. This means that the left-hand side of the above inequality defines a convex set and the right-hand side a nonconvex set. Convexity of the right-hand side inequality can thus be induced by changing the exponents from $a = b = -1/2$ to $a, b \ge 0$, $a + b \le 1$ (for example, $a = b = 1/2$ ). This can be accomplished by introducing copies $\bar{\Delta}_i$ of the $\Delta_i$ variables which by definition ensure that the exponents are positive and their sum is less than or equal to 1. These new variables $\bar{\Delta}_i$ are defined as

$$ \bar{\Delta}_i = \sum_{l=1}^{L} \frac{y_{il}}{\delta_l}, \qquad i = 1, ..., N $$

When the new variables are used, the definition of $X_{ij}$ is equivalently recast as a two-sided convex constraint set.

$$ \frac{1}{\sqrt{\Delta_i \Delta_j}} \le X_{ij} \le \sqrt{\bar{\Delta}_i \bar{\Delta}_j}, \quad i = 1, ..., N, \quad j = i + 1, ..., N $$

After the continuous $X_{ij}$ variables have been defined in terms of two convex inequalities, the remaining binary times continuous variable products

$$ Z_{ij} = X_{ij} \sum_{k=1,2,3} a_{ijk} $$

can be rewritten as a set of four linear inequalities (Glover[21]).

$$ \left.\begin{aligned} X_{ij} - X_{\mathrm{U}}\left(1 - \sum_{k=1,2,3} a_{ijk}\right) \le Z_{ij} \le X_{ij} - X_{\mathrm{L}}\left(1 - \sum_{k=1,2,3} a_{ijk}\right) \\ X_{\mathrm{L}} \sum_{k=1,2,3} a_{ijk} \le Z_{ij} \le X_{\mathrm{U}} \sum_{k=1,2,3} a_{ijk} \end{aligned}\right\} \begin{aligned} &\forall\, i = 1, ..., N \\ &\forall\, j = i + 1, ..., N \end{aligned} $$

Here $X_{\mathrm{L}}$ and $X_{\mathrm{U}}$ are lower and upper bounds, respectively, on the variables $X_{ij}$ computed based on the entries of Table 1. When exactly the same transformations are performed, the nonconvex defining relation for ${}^1\chi^{\mathrm{v}}$ is converted into a set of convex nonlinear and linear constraints, at the expense of introducing the extra continuous variables $\bar{\Delta}_i^{\mathrm{v}}$, $X_{ij}^{\mathrm{v}}$, and $Z_{ij}^{\mathrm{v}}$.

The exact linearization of binary times continuous variable products is also applied to account for the presence of $n$ in the denominator of the structure−property correlating expressions. Specifically, the definition of the deviation between the value of a given property and a prespecified target,

$$ sn \ge \frac{1}{P_m^{\mathrm{scale}}} |p_m - P_m^{\mathrm{target}} n| $$

where

$$ p_m = \sum_{l=1}^{L} C_m^l n_l + C_m^0 {}^0\chi + C_m^{0\mathrm{v}} {}^0\chi^{\mathrm{v}} + C_m^1 {}^1\chi + C_m^{1\mathrm{v}} {}^1\chi^{\mathrm{v}}, \qquad m = 1, ..., M $$

where $s$ is the property deviation, $p_m$ is an extensive version of property $m$, and $M$ is the total number of properties being targeted, is rewritten as

$$ \sum_{i=1}^{N} \sum_{l=1}^{L} z_{il} \ge \frac{1}{P_m^{\mathrm{scale}}} (p_m - P_m^{\mathrm{target}} n) $$

$$ \sum_{i=1}^{N} \sum_{l=1}^{L} z_{il} \ge \frac{1}{P_m^{\mathrm{scale}}} (P_m^{\mathrm{target}} n - p_m) $$

The new variable $z_{il}$ is defined as

$$ \left.\begin{aligned} s - s^{\mathrm{U}}(1 - y_{il}) \le z_{il} \le s - s^{\mathrm{L}}(1 - y_{il}) \\ s^{\mathrm{L}} y_{il} \le z_{il} < s^{\mathrm{U}} y_{il} \end{aligned}\right\} \begin{aligned} &\forall\, i = 1, ..., N \\ &\forall\, l = 1, ..., L \end{aligned} $$

where $s^{\mathrm{L}}$ and $s^{\mathrm{U}}$ are upper and lower bounds on $s$.

The above-described transformations ensure that the linking expressions between $a_{ijk}$ and $y_{il}$ and the objective function involve only linear and convex nonlinear relations, yielding a convex MINLP problem. The complete problem formulation is given in the next section.

## 5. Problem Formulation

The complete listing of sets, parameters, variables, and constraints used in the optimization formulation is as follows:

### *Sets*

$i = 1, ..., N$: set of vertices in the molecular graph representation

$l = 1, ..., L$: set of all of the different basic groups employed

$m = 1, ..., M$: set of all of the macroscopic properties of interest
$k = 1, 2, 3$: set of single, double, and triple bonds, respectively

***Parameters***

$MW_l$: molecular weight of the $l$th basic group (g/mol)
$v_{lk}$: no. of bonds of multiplicity $k$ that the basic group of type $l$ forms
$\delta_l$: coefficient of the first weighting scheme for the basic group of type $l$
$\delta_l^v$: coefficient of the second weighting scheme for the basic group of type $l$
$R_{max}$: maximum number of allowed rings in a polymer repeat unit

***Variables***

$a_{ijk}$: adjacency matrix elements for bonds of multiplicity $k$
$y_{il}$: assignment variable for group $l$ to vertex $i$
$s$: maximum scaled property deviation
${}^0\chi$: zeroth-order simple molecular connectivity index
${}^0\chi^v$: zeroth-order valence molecular connectivity index
${}^1\chi$: first-order simple molecular connectivity index
${}^1\chi^v$: first-order valence molecular connectivity index
$n$: total number of basic groups in the polymer repeat unit
MW: molecular weight of the repeat unit (g/mol)

***Constraints***

Assignment of the weighting coefficients to each vertex $i$:

$$\Delta_i = \sum_{l=1}^{L} \delta_l y_{il}, \qquad \Delta_i^v = \sum_{l=1}^{L} \delta_l^v y_{il}, \qquad i = 1, ..., N$$

Definition of zeroth-order connectivity indices:

$${}^0\chi = \sum_{i=1}^{N}\sum_{l=1}^{L} \frac{y_{il}}{\sqrt{\delta_l}} \qquad {}^0\chi^v = \sum_{i=1}^{N}\sum_{l=1}^{L} \frac{y_{il}}{\sqrt{\delta_l^v}}$$

Definition of inverse weighting coefficients employed in the convexification:

$$\bar{\Delta}_i = \sum_{l=1}^{L} \frac{y_{il}}{\delta_l}, \qquad \bar{\Delta}_i^v = \sum_{l=1}^{L} \frac{y_{il}}{\delta_l^v}, \qquad i = 1, ..., N$$

Convex representation of $X_{ij}$:

$$\left.\begin{aligned} \frac{1}{\Delta_i \Delta_j} \le X_{ij} \le \sqrt{\bar{\Delta}_i \bar{\Delta}_j} \\ \frac{1}{\Delta_i^v \Delta_j^v} \le X_{ij}^v \le \sqrt{\bar{\Delta}_i^v \bar{\Delta}_j^v} \end{aligned}\right\} \forall\, i = 1, ..., N \quad \forall\, j = i+1, ..., N$$

Exact linear representation of binary times continuous variable products $Z_{ij} = X_{ij}\sum_{k=1,2,3} a_{ijk}$ and $Z_{ij}^v = X_{ij}^v \sum_{k=1,2,3} a_{ijk}$:

$$\left.\begin{aligned} X_{ij} - X_U\Big(1 - \sum_{k=1,2,3} a_{ijk}\Big) \le Z_{ij} \le X_{ij} - X_L\Big(1 - \sum_{k=1,2,3} a_{ijk}\Big) \\ X_L \sum_{k=1,2,3} a_{ijk} \le Z_{ij} \le X_U \sum_{k=1,2,3} a_{ijk} \\ X_{ij}^v - X_U^v\Big(1 - \sum_{k=1,2,3} a_{ijk}\Big) \le Z_{ij}^v \le X_{ij}^v - X_L^v\Big(1 - \sum_{k=1,2,3} a_{ijk}\Big) \\ X_L^v \sum_{k=1,2,3} a_{ijk} \le Z_{ij}^v \le X_U^v \sum_{k=1,2,3} a_{ijk} \end{aligned}\right\} \begin{aligned} i &= 1, ..., N \\ j &= i+1, ..., N \end{aligned}$$

Definition of first-order connectivity indices:

$${}^1\chi = \sum_{i=1}^{N}\sum_{j=i+1}^{N} Z_{ij}, \qquad {}^1\chi^v = \sum_{i=1}^{N}\sum_{j=i+1}^{N} Z_{ij}^v$$

Definition of the total number of basic groups in a polymer repeat unit:

$$n = \sum_{i=1}^{N}\sum_{l=1}^{L} y_{il}$$

Definition of the molecular weight of a repeat unit:

$$MW = \sum_{l=1}^{L} MW_l \Big(\sum_{i=1}^{N} y_{il}\Big)$$

Extensive form of structure−property correlations:

$$p_m = C_m^0\, {}^0\chi + C_m^{0v}\, {}^0\chi^v + C_m^1\, {}^1\chi + C_m^{1v}\, {}^1\chi^v + \sum_{i=1}^{N}\sum_{l=1}^{L} C_m^l y_{il}, \qquad m = 1, ..., M$$

Exact linear representation of $z_{il} = s y_{il}$:

$$\left.\begin{aligned} s - s^U(1 - y_{il}) \le z_{il} \le s - s^L(1 - y_{il}) \\ s^L y_{il} \le z_{il} \le s^U y_{il} \end{aligned}\right\} \begin{aligned} i &= 1, ..., N \\ l &= 1, ..., L \end{aligned}$$

Definition of the maximum scaled property deviation using variables $z_{il}$:

$$\sum_{i=1}^{N}\sum_{l=1}^{L} z_{il} \ge \frac{1}{P_m^{scale}}(p_m - P_m^{target} n),$$

$$\sum_{i=1}^{N}\sum_{l=1}^{L} z_{il} \ge -\frac{1}{P_m^{scale}}(p_m - P_m^{target} n)$$

Valency balance constraint:

$$\sum_{j=1}^{i-1} a_{jik} + \sum_{j=i+1}^{N} a_{ijk} = \sum_{l=1}^{L} v_{lk} y_{il}, \quad i = 1, ..., N, \quad k = 1, 2, 3$$

This constraint equates the total number of bonds with $k$ multiplicity, linking vertex $i$ and the rest of the graph vertices with the total number of $k$-type bonds that a group of type $l$ at vertex $i$ may form. It ensures that all valency requirements for all bond types are satisfied for each basic group.

Assignment of a single basic group type $l$ to each vertex location $i$:

$$\sum_{l=1}^{L} y_{il} = 1, \qquad i = 1, ..., N$$

This constraint ensures that a single basic group type $l$ is assigned to each vertex $i$. This definition implies that the total number of groups participating in the polymer repeat unit is fixed and equal to $N$. However, in optimization studies this number is preferably variable, allowing for polymer repeat units with different numbers of basic groups. A straightforward correction would be to impose $\sum_{l=1}^{L} y_{il} \le 1$ instead of $\sum_{l=1}^{L} y_{il} = 1$. The problem with this correction is that it generates unoccupied vertices $i$ whose corresponding coefficients

$\Delta_i$ and $\Delta_i^v$ are equal to zero. This yields division by zero errors in the definitions for $X_{ij}$ and $X_{ij}^v$, respectively. To remedy this problem, the concept of dummy groups is introduced, allowing the total number of basic groups in the repeat unit to vary between a lower bound and $N$. Specifically, a dummy group is defined so that it (i) has zero valency with respect to all bond types, (ii) has coefficients $\delta = \delta^v = 1$, (iii) is excluded from the evaluation of the zeroth-order connectivity indices, and (iv) has no effect on the property correlations. Furthermore, the defining relation for $n$ is rewritten as

$$\sum_{l=1,l\neq\text{dummy}}^{L} \sum_{i=1}^{N} y_{il} = n$$

Bond-type selection between vertices $i$ and $j$:

$$\sum_{k=1,2,3} a_{ijk} \leq 1, \qquad i = 1, ..., N, \quad j = i+1, ..., N$$

This equation ensures that no two basic groups are bonded with more than one bond type (i.e., single, double, or triple).

Balance constraint between vertices and edges of the molecular graph:

$$\sum_{i=1}^{N} \sum_{j=i+1}^{N} \sum_{k=1,2,3} a_{ijk} \leq \sum_{i=1}^{N} \sum_{l\neq\text{dummy}}^{L} y_{il} + R_{\max}$$

This relation requires the repeat unit to have no more than $R_{\max}$ rings, excluding one cycle which models the periodicity.

Connected molecular graph constraint:

$$\sum_{i=1}^{j-1} \sum_{k=1,2,3} a_{ijk} \geq \sum_{l=1,l\neq\text{dummy}}^{L} y_{jl}, \qquad j = 2, ..., N$$

This constraint, first proposed by Churi and Achenie,[23] forces each vertex $j > 1$ to be bonded to at least one vertex whose labeling is less than $j$. This imposes a bonding hierarchy, ensuring connectivity in the molecular graph.

Bond-exclusion constraints:

$$y_{il_1} + y_{jl_2} + \sum_{k=1,2,3} a_{ijk} \leq 2, \qquad i = 1, ..., N, \quad j = i+1, ..., N$$

For a given pair of basic group types $l_1$ and $l_2$, this relation ensures that they are not connected with a bond of any multiplicity. This set of constraints prevents the formation of unstable structures by disallowing groups such as −N< and −O− from bonding with themselves or with −OH, −F, and −Cl.

Bounds on $X$, $X^v$, and $n$:

$$X_L = \frac{1}{\max_l \delta_l}, \quad X_U = \frac{1}{\min_l \delta_l}, \quad X_L^v = \frac{1}{\max_l \delta_l^v}, \quad X_U^v = \frac{1}{\min_l \delta_l^v}, \qquad n \geq N^L$$

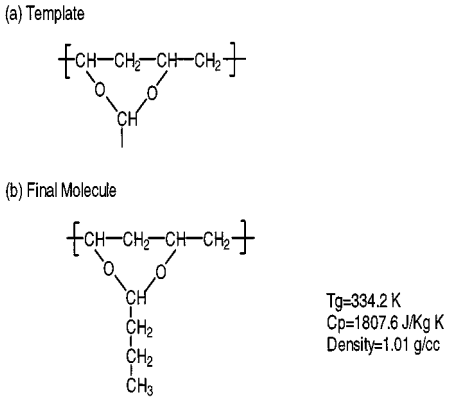



**Figure 2.** Template and optimal repeat unit for the first example.

These equations define a convex MINLP formulation which is solved using the algorithm Outer Approximation.[24]

## 6. Polymer Design Examples

Three examples are addressed to test the proposed polymer design methodology. The solver employed in this work is DICOPT++[25] implemented within the modeling language GAMS.[26] All computational results are reported on a RS6000 390 IBM workstation.

Higher level chemical knowledge about the structure of the repeat unit under investigation is incorporated through the use of a procedure which is referred to as *templating*. Specifically, this involves "fixing" part of the molecular graph which is expected or desired to remain unchanged. This improves solution tractability by reducing search complexity. Furthermore, when the search is constrained to repeat units with structures similar to those employed to generate the structure−property correlations, property estimation accuracy is expected to remain high. However, templating may eliminate some nonintuitive molecular topologies from consideration. Examples of the molecular templates employed for the example problems discussed are shown in Figures 2−4.

The first example tests whether the optimization formulation can correctly identify a prespecified target polymer, poly(vinylbutyral). This targeting is performed through matching the glass transition temperature and imposing bounds on the heat capacity and density of the polymer. The search template which fixes the backbone of the repeat unit is given in Figure 2. The number of basic groups in the polymer repeat unit is allowed to vary between 8 (the minimum possible based on the template used) and 11. The optimization problem solved with outer approximation converged to the poly(vinylbutyral) repeat unit after 280 CPU s. The final solution and its properties are shown in Figure 2.

After confirming the capability of the solution procedure to match a known structure, we then explored whether it would generate not only the best but also additional molecular graphs whose objective function is close to the optimal one. Three property targets on glass transition temperature, heat capacity, and density are matched simultaneously. The target values chosen are the ones from the second example of Venkatasubramanian et al.[4] using the correlations developed in this work. The backbone of the polymer is fixed and is shown in Figure 3, along with the target values of the properties. Integer cuts are used to derive multiple indepen-

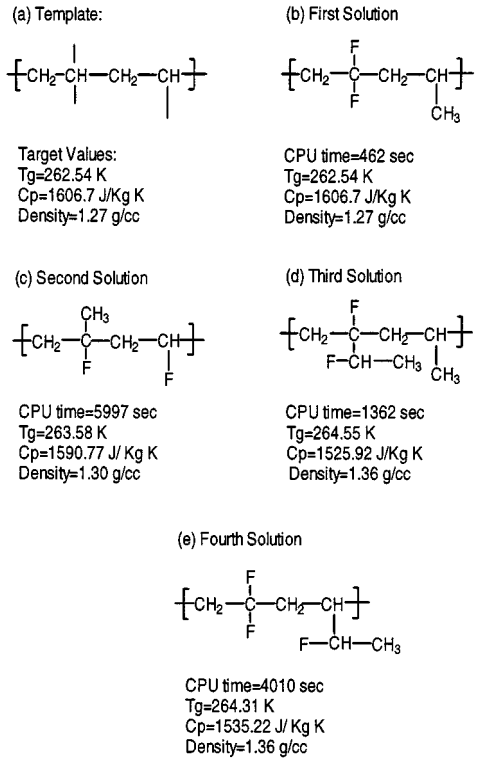



**Figure 3.** Template and optimal repeat units for the second example.

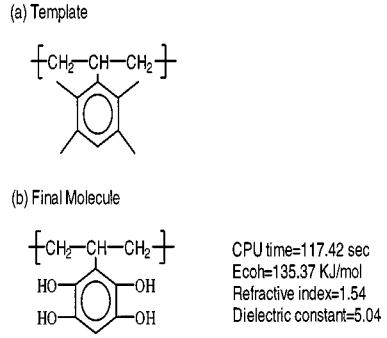



**Figure 4.** Template and optimal repeat unit for the third example.

dent solutions to the problem. The polymer repeat units derived by the algorithm, along with the property values for these structures and the CPU time required to generate them, are presented in Figure 3. High CPU requirements required for the additional solutions necessitate a limit on the number of MIP iterations. In the cases for which this limit is reached, only the best feasible solution is presented. As is seen in the figure, the first solution matches the targeted polymer repeat unit. Solutions 2−4 have structures which are incrementally different from the first and second solutions and, as expected, have a greater deviation from the property targets.

Next, single property optimization is examined. A polymer repeat unit is sought for which the cohesive energy is maximized while bounds are enforced on the refractive index and dielectric constant. The repeat unit for polystyrene is used as a template, with an added methylene group in the backbone. This template is shown in Figure 4. The optimization procedure is called to choose substituents to add to the aromatic ring of the repeat unit. The optimal molecular design for this problem, along with the property values computed for that structure, are presented in Figure 4. Four −OH groups were added to the aromatic ring, which causes the cohesive energy to increase significantly. This effect is also predicted by Bicerano,[7] who uses a large correction term for −OH groups bonded to aromatic rings in his cohesive energy correlation. Furthermore, the repeat unit is seen to have a large dielectric constant, which is consistent with the idea that four highly polarizable groups have been chosen.

These three examples show the effectiveness of the given formulation in the derivation of complete descriptions of polymer repeat units. Both the identity of each basic group and the interconnections between these groups can be found, for either a property matching or property optimization design task. The correlations derived here are found to predict polymer properties effectively and follow general trends predicted from the chemistry of the repeat units. Also, this algorithm can find multiple polymer structures near the optimum, although the integer cuts required for this have a significant cost in terms of solution time.

## 7. Conclusions

This study investigated the use of topological and, in particular, connectivity indices as structural descriptors in optimal polymer design. Through the use of two sets of variables, $y_{il}$ and $a_{ijk}$, a complete description of the molecular graph was achieved, providing the means for molecular modeling accounting for interconnectivity in optimization studies beyond polymer design. Both zeroth- and first-order connectivity indices using two separate weighting schemes were employed. The accuracy of the correlations for heat capacity, cohesive energy, refractive index, and dielectric constant was comparable with those of Bicerano[7] with no correction terms but more regression coefficients. The only exception was the correlation for the glass transition temperature whose accuracy was lower than the one by Bicerano.[7] Because the glass transition depends on longer range interactions, connectivity indices of higher than second order, or correction terms, are required in this case. The extension of the proposed mathematical description to handle higher order indices is straightforward. When the generalized polynomial structure of the nonconvex equalities is exploited, a convex representation of the optimization problem was derived, yielding convex MINLP formulations. This formulation was then solved for three polymer design examples involving both property matching and property optimization. Molecular templates were employed to constrain the search space toward desired topologies. Computational requirements were relatively high, motivating the need for exploring customized decomposition approaches. In particular, when the $y_{il}$ part of the formulation is decoupled from the $a_{ijk}$, smaller subproblems can be obtained.

## Acknowledgment

Useful discussions with Dr. Bicerano as well as financial support by the NSF Career Award CTS-9701771 and computing hardware support by the IBM Shared University Research Program 1996, 1997, and 1998 are gratefully acknowledged.

## Literature Cited

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

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

(3) 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.

(4) 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.

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

(6) Maranas, C. D. Optimal Computer-Aided Molecular Design: A Polymer Design Case Study. *Ind. Eng. Chem. Res.* **1996**, *35*, 3403.

(7) Bicerano, J. *Prediction of Polymer Properties*; Marcel Dekker: New York, 1996.

(8) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.−Act. Relat.* **1985**, *4* (109).

(9) Kier, L. B. Shape Indexes of Order One and Three from Molecular Graphs. *Quant. Struct. Act. Relat.* **1986**, *5* (1).

(10) Weiner, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17.

(11) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609.

(12) Trinajstic, M. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1975.

(13) 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.

(14) Baskin, I. I.; Gordeeva, E. V.; Devdariani, R. O.; Zefirov, N. S.; Palyulin, V. A.; Stankevich, M. I. Methodology of Solution of the Inverse Problem for the Structure−Property Relationship for the case of Topological Indices. *Dokl. Chem. (Transl. of Dokl. Akad. Nauk)* **1990**, *307*, 217.

(15) Kvansnička, V.; Pospíchal, J. Canonical Indexing and Constructive Enumeration of Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 99.

(16) Skvortsova, M. I.; Baskin, I. I.; Slovokhotova, O. L.; Palyulin, V. A.; Zefirov, N. S. Inverse Problem in QSAR/QSPR Studies for the Case of Topological Indices Characterizing Molecular Shape (Kier Indices). *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 360.

(17) 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.

(18) Raman, V. S.; Maranas, C. D. Optimization in Product Design with Properties Correlated with Topological Indices. *Comput. Chem. Eng.* **1998**, *22* (6), 747.

(19) Joback, K. G.; Stephanopoulos, G. Designing Molecules Possessing Desired Physical Property Values. *Proceedings of the Third International Conference on Foundations of Computer-Aided Process Design*, Snowmass Village, CO, July 10−14, 1989; Siirola, J. J., Grossman, I. E., Stephanopoulos, G., Eds.; Elsevier: Amsterdam, New York, 1990; p 363.

(20) SciVision. *SciPolymer User Guide*; SciVision, Inc.: Lexington, MA 1996.

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

(22) Maranas, C. D.; Floudas, C. A. Finding All Solutions of Nonlinearly Constrained Systems of Equations. *J. Global Opt.* **1995**, *7*, 143.

(23) Churi, N.; Achenie, L. E. K. Novel Mathematical Programming Model for Computer Aided Molecular Design. *Ind. Eng. Chem. Res.* **1996**, *35* (10), 3788.

(24) Duran, M. A.; Grossmann, I. E. An Outer Approximation Algorithm for a Class of Mixed-Integer Nonlinear Programs. *Math. Prog.* **1986**, *36*, 307.

(25) Viswanathan, J.; Grossmann, I. E. A Combined Penalty Function and Outer-Approximation method for MINLP Optimization. *Comput. Chem. Eng.* **1990**, *14*, 769.

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

*Received for review* October 23, 1998
*Revised manuscript received* February 4, 1999
*Accepted* February 5, 1999

IE980682N