Pergamon

PII: S0098-1354(97)00257-3

Computers Chem. Engng Vol. 22. No. 6 pp. 747–763, 1998

0098-1354/98 $19.00 + 0.00

Optimization in product design with properties correlated with topological indices

V. Shankar Raman and Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, 112A Fenske Laboratory, University Park, PA 16802, U.S.A.

Abstract

This paper addresses the problem of incorporating topological indices as structural descriptors for correlating properties in the design of product molecules with fine-tuned or optimized property values. Three different types of topological indices are considered: Randić's molecular connectivity indices, Kier's shape indices and the Wiener Index. The adjacency matrix representation which provides a complete description of the connectivity of a molecule is utilized. Thus, complete molecular interconnectivity information is introduced in the optimization framework which, in principle, provides for more accurate property prediction than simple group contributions. The nonlinear expressions for the topological indices are systematically transformed into equivalent linear relations enabling the formulation of the molecular design problem as a Mixed Integer Linear Program (MILP). Two different examples are considered: The first involves the design of alkanes with target physical properties correlated with Kier's shape indices and the second the selection of the best substituent of a compound with desired fungicidal properties correlated with Randić's connectivity index. In both examples, uncertainty in the model regression coefficients is quantitatively taken into account. © 1998 Elsevier Science Ltd. All rights reserved

Keywords: Product design; topological indices; MILP

1. Introduction

The demand for application-specific products demonstrates the necessity and importance of molecular design. Traditional molecular design involves an iterative procedure of compound synthesis and property evaluation. Computer-aided molecular design (CAMD) reduces experimental effort by pointing out only promising compounds for synthesis and experimental verification. Examples of applications of CAMD in the chemical engineering literature include the design of polymers with desired thermophysical and mechanical properties (Vaidyanathan and El-Halwagi, 1996; Maranas, 1996), environmentally benign refrigerants (Duvedi and Achenie, 1996; Churi and Achenie, 1996), solvents and extractants (Odele and Macchietto, 1993; Gani *et al.*, 1991), and herbicides/pesticides (Reynolds *et al.*, 1995). A thorough review can be found in (Mavrovouniotis, 1996).

CAMD requires that the properties relevant to design be expressed as a function of molecular structure. The type of structure-property relations employed depend on both the desired level of accuracy and the particular property. The input-output relations employed in this paper are Quantitative Structure Property Relationships (QSPR) and Quantitative Structure Activity Relationships (QSAR) (Kier and Hall, 1986; de Waterbeemd, 1995). These relations are obtained by correlating the structural attributes of a set of compounds with their physicochemical properties or biological activities using statistical methods. The particular class of QSAR/QSPRs considered in this paper utilize topological indices as structural descriptors. Topological indices characterize compounds with a single number based on its interconnectivity and the types of atoms in the molecule. By taking into account the interconnectivity of the molecule, more accurate property prediction can, in principle, be achieved than with simple group contributions which largely neglect internal molecular architecture. It is important to keep in mind, however, that topological indices do not necessarily have

* Author to whom all correspondence should be addressed. Tel.: (814)863-9958, fax: (814)865-7846, e-mail: cdm8@psu.edu.

a causal relationship with the correlated property. Instead, they provide a convenient vehicle for providing an empirical correlation between structure and property. In this paper, the following three most popular topological indices are employed: (1) Randić's molecular connectivity index $^1\chi$; (2) Kier's shape indices, and (3) the Wiener index.

A review of the variety of applications of topological indices in QSAR/QSPR can be found in Trinajstic (1992). Some of the key contributions include the work of Gordeeva *et al.* (1990) and Skvortsova *et al.* (1992) who addressed the problem of generating structures from Randić indices. Baskin *et al.* (1990) solved the problem of structure generation from Wiener and Randić indices. The method is based on the exhaustive generation of graphs with a given distribution of vertices and edges. Skvortsova *et al.* (1993) used a similar approach to design molecules with target properties that are functions of Kier's shape indices. Kvasnička and Pospíchal (1990) proposed an algorithm for the generation of molecules with a given Randić index using graph theory. Kier *et al.* (1993) designed molecules with target molar volumes correlated with the first and second-order molecular connectivity index. Kier *et al.* (1993) and Hall *et al.* (1993a,b) derived relations between vertex degrees and distribution of edges to aid in structure generation. Skvortsova *et al.* (1996) used the number and nature of basic fragments comprising a molecule to generate structures with specified property values correlated with topological descriptors.

In essence, most of the methods proposed so far for generating molecules with specified target property values use vertex, edge distribution types and type and number of structural fragments to describe the molecule. However, these descriptors do not necessarily define a unique molecule because interconnectivity is not fully specified. In this paper, a complete representation of the molecular connectivity based on the vertex adjacency matrix (Trinajstic, 1992) is employed. The rows and columns of this matrix correspond to atoms in the molecule. An element of the matrix is one if the two atoms representing the particular row and column are connected by a bond and zero otherwise. The advantage of the vertex adjacency matrix representation is that it defines a unique molecule because the interconnectivity is fully specified. However, most topological indices exhibit a nonlinear functional dependence on the elements of the vertex adjacency matrix. This complicates the application of optimization-based techniques. To remedy this, the nonlinear functional terms are transformed into equivalent linear relations which enable the formulation of the molecular design problem as a Mixed Integer Linear Program (MILP). The resulting formulation involves a number of important features:

- The optimization problem can be solved to global optimality with commercially available solvers (e.g., CPLEX, OSL etc.).
- Multiple property targets correlated with different topological indices can be handled simultaneously and there is considerable flexibility in selecting the molecular design objective (e.g., property target matching or property value optimization).
- With the appropriate use of integer cuts, not only the optimal structure but also the second best, third best etc. structures are obtained. Also, a quantitative description of uncertainty in the structure-property relations can readily be incorporated (Maranas, 1997a).

In the following sections, the vertex adjacency matrix representation of molecular graphs is discussed and the proposed optimization framework is highlighted. Then, a basic set of relations in molecular graphs is presented. This is followed by a detailed description of Randić's molecular connectivity indices, Kier's shape indices and the Wiener index. The discussions of the topological indices are divided into two parts: Definition and method of calculation of the topological index, and proposed reformulation as an MILP. The discussion of the three topological indices is followed by a brief description of the formulations used to quantify uncertainties in the properties. Finally, two illustrative examples of the proposed methodology are presented. The first example deals with the design of alkane molecules with targeted physical properties correlated with Kier's shape indices. The second example involves the optimal active substituent selection for a compound to obtain desired fungicidal properties using the molecular connectivity index $^1\chi$ as the structural descriptor. For both examples, the effect of property prediction uncertainty is addressed following the developments of Maranas (1997a).

2. Molecular graph background

A graph is characterized by two sets (Harary, 1972; Trudeau, 1976):

1. Vertex Set $\mathscr{V} = \{1, 2, \ldots, N\}$
2. Edge Set $\mathscr{E} = \{(i, j) \mid \text{vertices } i \text{ and } j \text{ are connected by an edge}\}$

In particular, a *molecular graph* is the graph representation of a molecule where atoms and bonds correspond to vertices and edges respectively. Usually, molecular graphs are hydrogen-suppressed (Spialter, 1964) since hydrogen has a valency of one and cannot participate in the molecular backbone. The graphs which do not account for bond multiplicities (i.e., double or triple bonds) are referred to as *simple graphs*. For example, the conventional and the molecular graph representation of 2-methyl butane are shown in Figs 1 and 2 respectively. The vertex and edge sets for this molecular graph representation of 2-Methyl Butane are:

$$\mathscr{V} = \{1, 2, 3, 4, 5\}$$

$$\mathscr{E} = \{(1, 2), (2, 3), (2, 4), (4, 5)\}$$

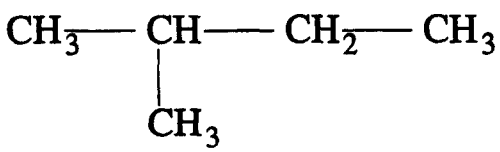

Fig. 1. Conventional representation of 2-methyl butane.

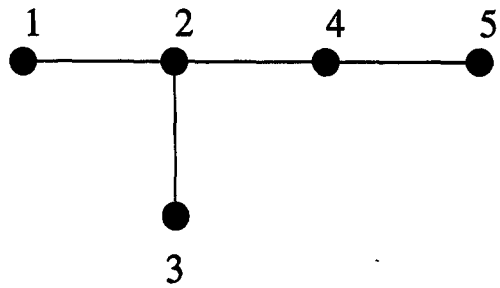

Fig. 2. Molecular graph of 2-methyl butane.

A molecular graph can be represented by a variety of matrices (Trinajstic, 1992) such as vertex adjacency matrix, edge adjacency matrix, incidence matrix, cycle matrix or 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 graph is a $N \times N$ matrix, where N is the number of vertices in the graph. It is given by:

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

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

Based on the definition it is clear that the adjacency matrix is symmetric and the diagonal elements are zero. Given the adjacency matrix, the connectivity of a molecule is uniquely and unambiguously determined. For example, the adjacency matrix of the molecular graph in Fig. 2 is:

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Because the adjacency matrix provides complete information about the connectivity of a molecule, it follows that topological indices are uniquely defined by the elements of the adjacency matrix a_{ij}.

3. Problem definition

The basic question addressed in this paper is defined as follows:

Generate a ranked list of all molecules whose property values either match some prespecified targets or are optimized, when the available QSAR/QSPR employ topological indices as structural descriptors.

This problem can be recast in an optimization framework as proposed by Maranas (1996). The variables describing the molecular structure are the elements of the vertex adjacency matrix a_{ij}. Additional variables in the include the properties p_l which depend on the topological indices TI_k which in turn are functions of a_{ij}. The objective function to be optimized is a function of the properties p_l. The minimization of the maximum scaled deviation from some target values is formulated as (Problem (I)),

$$\min \max_l \frac{1}{p_l^{target}} | p_l - p_l^{target} |$$

$$\text{subject to } p_l = p_l(TI_1, TI_2, \ldots, TI_T), \quad l = 1, 2, \ldots, P$$

$$TI_k = TI_k(a_{ij}), \quad k = 1, 2, \ldots, T$$

where $p_l(TI_k)$ are the QSAR/QSPR's with p_l denoting activities/properties and TI_k are the topological indices serving as structural descriptors. $TI_k(a_{ij})$ relate the elements of the adjacency matrix to the topological index. Note that property p_l can be a function of any number of topological indices TI_k.

The maximization/minimization of a given property subject to lower and upper bounds is formulated as (Problem (II)),

$$\max/\min p_j$$

$$\text{subject to } (p_l)^L \le p_l \le (p_l)^U, \quad l = 1, 2, \ldots, P$$

where $(p_l)^L$ and $(p_l)^U$ are lower and upper bounds on values of p_l respectively.

4. Basic relations

In this section, the fundamental relations pertaining to the description of molecular graphs with binary variables are discussed. These relations are applicable to all formulations irrespective of the type of topological indices being considered. The order of the adjacency matrix is denoted as N^U, which is the maximum number of atoms (vertices) allowed to participate in a molecule. The valency of a vertex i, V_i, is equal to the number of edges originating from it. The valency of the ith vertex is the sum of the elements in the ith row of the adjacency matrix:

$$V_i = \sum_{j=1}^{N^U} a_{ij}, \; i = 1, 2, \ldots, N^U$$

Due to the symmetry of the adjacency matrix only the elements of the upper triangular part of the adjacency

matrix ($i < j$) need to be considered. This implies that

$$V_i = \sum_{j=1}^{N^U} a_{ij} = \sum_{j=1}^{i-1} a_{ij} + \sum_{j=i}^{N^U} a_{ij}$$
$$= \sum_{j=1}^{i-1} a_{ji} + \sum_{j=i}^{N^U} a_{ij}, \quad i = 1, 2, \ldots, N^U$$

The maximum allowed valency of a vertex in the molecular graphs considered here is four. Because the valency of a vertex can take only integral values from zero (no atom) through four, it can be expressed with binary variables (Glover, 1975):

$$V_i = \sum_{k=1}^{4} k\delta_i^k, \quad i = 1, 2, \ldots, N^U$$

$$\sum_{k=1}^{4} \delta_i^k \leq 1, \quad i = 1, 2, \ldots, N^U$$

where δ_i^k is a binary variable which is equal to one only if vertex i has a valency of k and is equal to zero otherwise. In general, any function $g(V_i)$ of V_i can be recast as (Glover, 1975):

$$g(V_i) = \sum_{k=1}^{4} g(k)\delta_i^k, \quad i = 1, 2, \ldots, N^U$$

$$\sum_{k=1}^{4} \delta_i^k \leq 1, \quad i = 1, 2, \ldots, N^U$$

The above expressions aid in the transformation of the nonlinear functional dependences on valencies into equivalent linear relations.

An additional useful relation in subsequent developments is Euler's Polyhedral formula as applied to planar graphs (Trinajstic, 1992; Harary, 1972; Trudeau, 1976):

$$N + R = ({}^1P) + 1$$

where N is the number of vertices, R is the number of rings and ${}^1P$ is the number of edges or 1-length paths in the molecular graph. A formal proof for this formula can be found in Trudeau (1976). A value of one for $\sum_{k=1}^{4} \delta_i^k$ indicates the presence of vertex (i.e., atom) i. Hence the total number of vertices is given by the expression:

$$N = \sum_{i=1}^{N^U} \sum_{k=1}^{4} \delta_i^k$$

5. Topological indices

Each topological index requires a different set of transformations for recasting it into an MILP form. These transformations are provided for the three topological indices addressed in this work starting with the molecular connectivity indices χ.

5.1. Molecular connectivity indices χ

The molecular connectivity indices χ provide a quantitative assessment of the degree of branching of molecules. Randić (1975) first addressed the problem of relating the physical properties of alkanes to the degree of branching across an isomeric series. The degree of branching of a molecule was quantified using a branching index which subsequently became known as first-order molecular connectivity index ${}^1\chi$. Kier and Hall (1986) extended this to higher orders and introduced modifications to account for heteroatoms. Currently, molecular connectivity indices are the most popular class of indices (Trinajstic, 1992) employed in QSAR/QSPR. They have been used in a wide spectrum of applications ranging from predicting physicochemical properties such as boiling point, solubility, partition coefficient etc. (Murray *et al.*, 1975; Kier and Hall, 1976) to predicting biological activities such as antifungal effect, anesthetic effect, enzyme inhibition etc., (Kier *et al.*, 1975; Kier and Murray, 1975).

Molecular connectivity indices are characterized by their order. The nth order molecular connectivity index is equal to,

$${}^n\chi = \sum_{(i_0, i_1 \ldots i_n) \in \mathscr{E}_n} \frac{1}{\sqrt{V_{i_0} \cdot V_{i_1}, \ldots, V_{i_n}}}$$

where $\mathscr{E}_n$ is the set of n consecutive edges in a molecule, $i_0, i_1, \ldots, i_n$ denote the $n + 1$ vertices forming the n consecutive edges, and $V_{i_k}, k = 0, \ldots, n$ is the valency of the i_k vertex. The order of a connectivity index χ defines the number of consecutive edges forming the path involved in each term of the summation. Specifically, the first-order molecular connectivity index is defined as:

$${}^1\chi = \sum_{(i,j) \in \mathscr{E}} \frac{1}{\sqrt{V_i V_j}}$$

where $\mathscr{E}$ is the edge set of the graph. The calculation of ${}^1\chi$ and ${}^2\chi$ for 2,3-dimethyl hexane is next illustrated. The molecular graph along with the 1-path and 2-path fragments used for this calculation are shown in Figs 3 and 4 where the numbers denote the valency of each vertex. Based on the fragments shown in Figs 3 and 4 the first and second-order connectivity indices are equal to:

$${}^1\chi = \frac{1}{\sqrt{1\cdot3}} + \frac{1}{\sqrt{3\cdot1}} + \frac{1}{\sqrt{3\cdot3}} + \frac{1}{\sqrt{3\cdot1}} + \frac{1}{\sqrt{3\cdot2}} + \frac{1}{\sqrt{2\cdot2}} + \frac{1}{\sqrt{2\cdot1}}$$
$$= 3.6807$$

$${}^2\chi = \frac{1}{\sqrt{1\cdot3\cdot1}} + \frac{1}{\sqrt{1\cdot3\cdot3}} + \frac{1}{\sqrt{1\cdot3\cdot3}} + \frac{1}{\sqrt{3\cdot3\cdot1}} + \frac{1}{\sqrt{3\cdot3\cdot2}} + \frac{1}{\sqrt{1\cdot3\cdot2}} + \frac{1}{\sqrt{3\cdot2\cdot2}} + \frac{1}{\sqrt{2\cdot2\cdot1}}$$
$$= 3.01$$

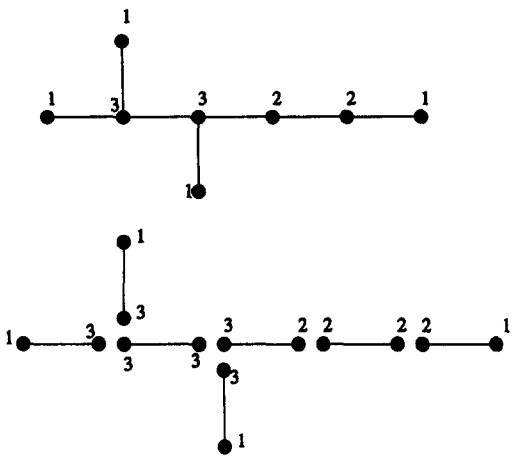

Fig. 3. Fragments of one-length paths for 2,3-dimethyl hexane.

The transformation of the nonlinear expression for ${}^1\chi$ into a set of linear inequalities is addressed next. The value of a_{ij} indicates the presence or absence of an edge, therefore ${}^1\chi$ can alternatively be expressed as:

$$ {}^1\chi = \sum_{i=1}^{N^V} \sum_{j=i}^{N^V} \frac{a_{ij}}{\sqrt{V_i V_j}} \tag{1} $$

Furthermore,

$$ \frac{1}{\sqrt{Vi}} = \sum_{k=1}^{4} \frac{1}{\sqrt{k}} \delta_i^k, \quad i = 1,2, \ldots, N^V $$

$$ \sum_{k=1}^{4} \delta_i^k \leq 1, \quad i = 1,2, \ldots, N^V $$

Consequently the equation for ${}^1\chi$ transforms to:

$$ {}^1\chi = \sum_{i=1}^{N^V} \sum_{j=i}^{N^V} \left(\sum_{k=1}^{4} \frac{\delta_i^k}{\sqrt{k}} \right) \left(\sum_{l=1}^{4} \frac{\delta_j^l}{\sqrt{l}} \right) a_{ij} $$

$$ = \sum_{i=1}^{N^V} \sum_{j=i}^{N^V} \sum_{k=1}^{4} \sum_{l=1}^{4} \frac{\delta_i^k \delta_j^l a_{ij}}{\sqrt{kl}} $$

After replacing $\delta_i^k \delta_j^l a_{ij}$ with

$$ c_{ij}^{kl} = \begin{cases} 1, & \text{if } \delta_i^k = 1 \text{ and } \delta_j^l = 1 \text{ and } a_{ij} = 1 \\ 0, & \text{otherwise} \end{cases} $$

we have:

$$ {}^1\chi = \sum_{i=1}^{N^V} \sum_{j=i}^{N^V} \sum_{k=1}^{4} \sum_{l=1}^{4} \frac{c_{ij}^{kl}}{\sqrt{kl}} $$

Note that c_{ij}^{kl} is equal to one if and only if all of the following relations are satisfied,

1. Both vertices i and j must exist.
2. Vertex i must have valency k and vertex j valency l.
3. Vertices i and j must be joined by an edge.

Though c_{ij}^{kl} is a binary variable, it can be treated as a continuous variable by adding constraints that force c_{ij}^{kl} to take only zero or one values. If an edge between vertices i and j exists, $(a_{ij} = 1)$, then the c_{ij}^{kl} element with the correct valency combination (k, l) is equal to one and the rest of them are equal to zero. Alternatively, if there is no edge connecting vertices i and j, $(a_{ij} = 0)$ then all c_{ij}^{kl} elements are equal to zero. Both of these arguments are cast mathematically as:

$$ \left. \sum_{k=1}^{4} \sum_{l=1}^{4} c_{ij}^{kl} = a_{ij} \right\} \begin{array}{l} i = 1,2, \ldots, N^V \\ j = i+1, \ldots, N^V \end{array} $$

If a vertex i has valency k, then exactly k elements of c_{ij}^{kl}, corresponding to the total number of vertices j of any valency l connected with vertex i, assume the value of one:

$$ \left. k\delta_i^k = \sum_{j=1}^{i-1} \sum_{l=1}^{4} c_{ji}^{lk} + \sum_{j=i}^{N^V} \sum_{l=1}^{4} c_{ij}^{kl} \right\} \begin{array}{l} i = 1,2, \ldots, N^V \\ k = 1,2,3,4 \end{array} $$

Based on the relations described above the problem of finding a molecular graph which matches a given target for the molecular connectivity index ${}^1\chi$ can be expressed as an MILP problem. This formulation was solved for a target value of ${}^1\chi = 3.5$ using GAMS/CPLEX (Brooke *et al.*, 1988) on a IBM RS6000 43P-133 workstation with an absolute convergence tolerance of 10^{-6}. The molecular graphs

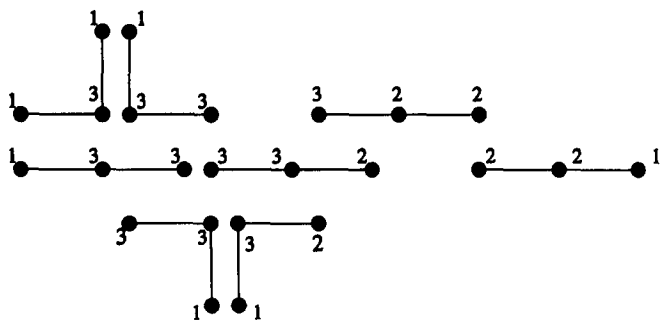

Fig. 4. Fragments of two-length paths for 2,3-dimethyl hexane.

Table 1. Compounds generated for ${}^1\chi^{target} = 3.5$

Rank	${}^1\chi$	Molecule
1	3.5040	2,3,3-trimethyl pentane
2	3.4814	2,2,3-trimethyl pentane
3	3.5534	2,3,4-trimethyl pentane
4	3.5607	2,2-dimethyl hexane
5	3.4165	2,2,4-trimethyl pentane

closest to the target index ${}^1\chi = 3.5$ and corresponding ${}^1\chi$ values are shown in Table 1. Next, the Kier's shape indices are addressed.

5.2. Kier's shape indices

Kier (1985, 1986) proposed three indices to quantitatively characterize the shape of a molecule. These three indices are the first, second and third order shape indices and are denoted by ${}^1\kappa, {}^2\kappa$ and ${}^3\kappa$ respectively. Index ${}^1\kappa$ quantifies the cyclicity of a molecule. For molecules with the same number of atoms, ${}^1\kappa$ decreases as the number of rings in the molecule increases. Index ${}^2\kappa$ quantifies the star-like attributes of a molecule. As the isomers in an acyclic isomeric series go from linear to star-like shape the values of ${}^2\kappa$ decrease. Index ${}^3\kappa$ quantifies the place in the chain where the branching occurs. The values of ${}^3\kappa$ for acyclic isomeric molecules reduce as the branching occurs closest to the center of the main chain. The main idea of the shape indices is to characterize a molecule by the number of n-length paths and normalize it with respect to two reference structures. The general formula for the nth order shape index is given by:

$$ {}^n\kappa = \frac{f({}^nP_{max})({}^nP_{min})}{({}^nP)^2} $$

where ${}^nP$ is the number of n-length paths in a molecule, ${}^nP_{max}$ and ${}^nP_{min}$ are the number of n-length paths in the two reference structures, and f is a normalizing factor. Kier has considered shape indices of up to order 3. The details of the derivation of the shape indices of first and third order can be found in Kier (1986) and of the second-order in Kier (1985).

The first-order shape index is given by (Kier, 1986),

$$ {}^1\kappa = \frac{N(N-1)^2}{({}^1P)^2} $$

where the number of 1-length paths is equal to:

$$ {}^1P = \sum_{i=1}^{N^v} \sum_{j=i}^{N^v} a_{ij} $$

The second-order shape index is given by the relation (Kier, 1985):

$$ {}^2\kappa = \frac{(N-1)(N-2)^2}{({}^2P)^2} $$

The number of 2-length paths (i.e., two consecutive edges), with vertex i at the center is equal to the number of ways of selecting two out of the V_i edges emanating from i.

$$ {}^2P = \sum_{i=1}^{N^v} \binom{V_i}{2} = \sum_{i=1}^{N^v} \frac{V_i(V_i-1)}{2} $$

Following the analysis of Section 4 the number of 2-length paths transforms to the following linear expression:

$$ {}^2P = \sum_{i=1}^{N^v} \sum_{k=1}^{4} \frac{k(k-1)}{2} \delta_i^k $$

The expression for the third-order shape index (Kier, 1986) is:

$$ {}^3\kappa = \frac{(N-3)(N-2)^2}{({}^3P)^2}, \quad \text{if } N \text{ is even} $$

$$ {}^3\kappa = \frac{(N-1)(N-3)^2}{({}^3P)^2}, \quad \text{if } N \text{ is odd} $$

The number of 3-length paths in a molecular graph can be identified as follows: Consider an edge (i,j) incident to vertices i and j. The edge (i,j) can be reached through vertex i in $V_i - 1$ ways and through vertex j in $V_j - 1$ ways. Each set of these three consecutive edges, (i.e., the edge (i,j) and the edges through which (i,j) is reached from i and j), constitutes a 3-length path. The number of 3-length paths with (i,j) as their middle edge is given by $(V_i - 1)(V_j - 1)$. After summing over all the edges of the graph the number of 3-length paths in a molecule is equal to (Hall *et al.*, 1993b):

$$ {}^3P = \sum_{(i,j)\in\mathcal{E}} (V_i - 1)(V_j - 1) $$

Consequently, the number of 3-length paths transform to:

$$ {}^3P = \sum_{i=1}^{N^v} \sum_{j=i}^{N^v} \sum_{k=1}^{4} \sum_{l=1}^{4} (k-1)(l-1)\delta_i^k \delta_j^l a_{ij} $$

After replacing $\delta_i^k \delta_j^l a_{ij}$ with c_{ij}^{kl}, the expression for ${}^3P$ yields:

$$ {}^3P = \sum_{i=1}^{N^v} \sum_{j=i}^{N^v} \sum_{k=1}^{4} \sum_{l=1}^{4} (k-1)(l-1)c_{ij}^{kl} $$

The following additional constraints, (same as those used for the ${}^1\chi$ index) must be added:

$$ \left.\sum_{k=1}^{4}\sum_{l=1}^{4} c_{ij}^{kl} = a_{ij}\right\} \begin{array}{l} i = 1,2,\ldots,N^U \\ j = i, i+1,\ldots,N^U \end{array} $$

$$ k\delta_i^k = \left.\sum_{j=1}^{i-1}\sum_{l=1}^{4} c_{ji}^{lk} + \sum_{j=i}^{N^v}\sum_{l=1}^{4} c_{ij}^{kl}\right\} \begin{array}{l} i = 1,2,\ldots,N^U \\ k = 1,2,3,4 \end{array} $$

So far, the linear representation of the expressions for the number of i-length, $i = 1,2,3$ paths ${}^1P$, ${}^2P$ and ${}^3P$ is presented. Next, the linear equivalent representation of the shape indices ${}^n\kappa, n = 1,2,3$ is addressed. The general expression for the nth order

shape index is considered first since some of the transformations are common to all orders. The expression for the nth order shape index is given by,

$$ {}^{n}\kappa = \frac{f({}^{n}P_{max})({}^{n}P_{min})}{({}^{n}P)^2} $$

which can equivalently be written as:

$$ ({}^{n}P)({}^{n}P)({}^{n}\kappa) = f({}^{n}P_{max})({}^{n}P_{min}) \tag{2} $$

Because ${}^{n}P$ is an integer variable, it can be expressed as the sum of the binaries in the following way (Nemhauser and Wolsey, 1988; Salkin and Mathur, 1989):

$$ {}^{n}P = \sum_{i=0}^{I} 2^i pu_i $$

$$ pu_i \in \{0,1\} $$

where I is the smallest integer satisfying the relation:

$$ 2^{I+1} - 1 \geq ({}^{n}P)_{max} $$

Next, the product $({}^{n}P)({}^{n}\kappa)$ is replaced by a new variable pk_n,

$$ pk_n = ({}^{n}P)({}^{n}\kappa) = \left(\sum_{i=0}^{I} 2^i pu_i\right)({}^{n}\kappa) $$

where pk_n consists of terms involving products of binaries and continuous variables. The general method of transforming a product of a binary variable y with a variable x into a set of linear inequalities was introduced by Glover (1975). Specifically, the product of pu_i and ${}^{n}\kappa$ in the expression for pk_n is transformed as follows,

$$ (pu_i)({}^{n}\kappa) = pv_i $$

$$ pu_i({}^{n}\kappa)^L \leq pv_i \leq pu_i({}^{n}\kappa)^U $$

$$ {}^{n}\kappa - ({}^{n}\kappa)^U(1 - pu_i) \leq pv_i \leq {}^{n}\kappa - ({}^{n}\kappa)^L(1 - pu_i) $$

where $({}^{n}\kappa)^L$ and $({}^{n}\kappa)^U$ are lower and upper bounds on ${}^{n}\kappa$.

Next, the product $({}^{n}P)({}^{n}P)({}^{n}\kappa)$ is replaced by variable q_n and the following relations are introduced:

$$ q_n = ({}^{n}P)({}^{n}P)({}^{n}\kappa) = ({}^{n}P)(pk_n) = \sum_{i=0}^{I} 2^i pu_i pk_n $$

where $pw_i = pu_i pk_n$

$$ pu_i(pk_n)^L \leq pw_i \leq pu_i(pk_n)^U $$

$$ pk_n - (pk_n)^U(1 - pu_i) \leq pw_i \leq pk_n - (pk_n)^L(1 - pu_i) $$

The right handside of Eq. (2) is treated separately for indices of different orders. They contain terms such as N^2 and N^3 which are transformed using the relations introduced by Glover (1975). Because N is an integer variable, it can be expressed as the sum of binaries in the following way,

$$ N = \sum_{k=0}^{K} 2^k nu_k $$

$$ nu_k \in \{0,1\} $$

where K is the smallest integer such that

$$ 2^{K+1} - 1 \geq N^U $$

Next, N^2 is replaced by variable n_2 and the following relations are introduced:

$$ n_2 = N^2 = \left(\sum_{k=0}^{K} 2^k nu_k\right)N = \sum_{k=0}^{K} (2^k nu_k N) $$

where $nu_k N = nv_k$

$$ nu_k(N^L) \leq nv_k \leq nu_k(N^U) $$

$$ N - N^U(1 - nu_k) \leq nv_k \leq N - N^L(1 - nu_k) $$

The following expressions are obtained after substituting n_3 for N^3.

$$ n_3 = N^3 = \sum_{k=0}^{K} (2^k nu_k N^2) = \sum_{k=0}^{K} (2^k nu_k n_2) $$

where $nw_k = nu_k n_2$

$$ nu_k(n_2)^L \leq nw_k \leq nu_k(n_2)^U $$

$$ n_2 - (n_2)^U(1 - nu_k) \leq nw_k \leq n_2 - (n_2)^L(1 - nu_k) $$

Eq. (2) yields for the first-order shape index:

$$ ({}^{1}P)^2({}^{1}\kappa) = N(N-1)^2 = N^3 - 2N^2 + N $$

Based on the proposed transformations this can be written as:

$$ q_1 = n_3 - 2n_2 + N $$

Equivalently, for the second-order shape index we have,

$$ ({}^{2}P)^2({}^{2}\kappa) = (N-1)(N-2)^2 = N^3 - 5N^2 + 8N - 4 $$

which transforms to:

$$ q_2 = n_3 - 5n_2 + 8N - 4 $$

The third-order shape index takes different forms depending on whether N is odd or even. When N is even we have,

$$ ({}^{3}P)^2({}^{3}\kappa) = (N-3)(N-2)^2 = N^3 - 7N^2 + 16N - 12 $$

which is equivalently written:

$$ q_3 = n_3 - 7n_2 + 16N - 12 \tag{3} $$

When N is odd Eq. (2) yields:

$$ ({}^{3}P)^2({}^{3}\kappa) = (N-1)(N-3)^2 = N^3 - 7N^2 + 15N - 9 $$

which means that

$$ q_3 = n_3 - 7n_2 + 15N - 9 \tag{4} $$

The two different expressions for q_3 can be recast within the same form by introducing an integer variable ne and a binary variable ny.

$$ N - 2ne = ny $$

Table 2. Structures generated for ${}^1\kappa^{target} = 8.0$, ${}^2\kappa^{target} = 3.1$, ${}^3\kappa^{target} = 2.8$

Rank	${}^1\kappa$ Found	${}^2\kappa$ Found	${}^3\kappa$ Found	% Max.Violation	Structure
1	8.000	3.111	2.813	0.5	2,3,4-trimethyl pentane
2	7.000	3.061	2.667	12.5	2,3-dimethyl pentane
3	8.000	2.520	2.813	18.7	2,2,3-trimethyl pentane
4	8.000	3.111	2.222	20.6	3-methyl,3-ethyl pentane
5	8.000	2.520	2.222	20.6	2,3,3-trimethyl pentane

Eqs. (3) and (4) representing "either-or" type of constraints are modeled using binary variables as explained in Ravindran *et al.* (1987).

$$-By \leq q_3 - (n_3 - 7n_2 + 16N - 12) \leq By \quad (5)$$

$$-B(1-y) \leq q_3 - (n_3 - 7n_2 + 15N - 9) \leq B(1-y) \quad (6)$$

where B is a large positive number.

The linear transformations described above for ${}^1\kappa, {}^2\kappa$ and ${}^3\kappa$ recast the problem of finding molecular structures with shape indices as close as possible to some target values as an MILP problem. For example, the results for the five molecular graphs closest to the targets ${}^1\kappa = 8.0$, ${}^2\kappa = 3.1$ and ${}^3\kappa = 2.8$ are summarized in Table 2.

5.3. *The Wiener index*

The Wiener index was introduced by Wiener (1947a) as path number representing the number of bonds between all pairs of atoms in a molecule. Wiener (1947a, b, c, 1948a, b) used the path number to predict the boiling point, heat of vaporization, molar refraction and other properties of alkanes. The Wiener index was first defined in graph theoretical terms as function of the elements of the distance matrix by Hosoya (1971) and since then has found applications in a wide range of QSARs and QSPRs applications (Gutman, 1993).

Evaluation of the Wiener index requires the definition of the distance matrix. The distance matrix of a molecular graph is defined as:

$$D = (d_{ij})$$

where d_{ij} is the number of edges in the shortest path connecting vertices i and j. It is evident from the definition that the distance matrix is symmetric and the diagonal terms are zero. For example, the distance matrix of 2-Methyl Butane shown in Fig. 2 is:

$$D = \begin{pmatrix} 0 & 1 & 2 & 2 & 3 \\ 1 & 0 & 1 & 1 & 2 \\ 2 & 1 & 0 & 2 & 3 \\ 2 & 1 & 2 & 0 & 1 \\ 3 & 2 & 3 & 1 & 0 \end{pmatrix}$$

The Wiener index is given by the sum of the elements of the upper (or lower) triangular part of the distance matrix. This can be written mathematically as:

$$W = \sum_{i=1}^{N} \sum_{j=i+1}^{N} d_{ij}$$

Using this expression, the Wiener index for 2-Methyl Butane is equal to:

$$W = 1 + 2 + 2 + 3 + 1 + 1 + 2 + 2 + 3 + 1 = 18$$

The problem of generating the distance matrix from the adjacency matrix has been addressed by (Bersohn, 1983; Müller *et al.*, 1987; Senn, 1988). The algorithm of Müller *et al.* (1987) is used in this work to encompass the problem of generating the adjacency matrix from the distance matrix within the optimization framework. The algorithm of Müller *et al.* (1987) can be summarized with the following steps:

Step 1: A new matrix A_0 is generated from the adjacency matrix A by replacing all non-diagonal elements which are zero with N. The value of N is chosen because an acyclic molecule does not contain a path of length longer than $N-1$.

Step 2: Set $l = 1$.

Step 3: The matrix A_l is recursively generated from matrix A_{l-1} in the following way:

$$a_l(i,j) = \min_{k=1,\ldots,N} (a_{l-1}(i,k) + a_{l-1}(k,j)) \left.\begin{matrix} i = 1,2,\ldots,N \\ j = 1,2,\ldots,N \end{matrix}\right\} \quad (7)$$

Step 4: Set $l = l + 1$

Step 5: If $2^{l-1} < N - 1$ go to Step 3

Step 6: A_L is the distance matrix of the molecule where L is the smallest integer for which $2^L \geq N - 1$.

The variables that are implied by the above described algorithmic procedure and need to be added in the optimization formulation are:

$$\left.\begin{matrix} a_0(i,j) \\ a_1(i,j) \\ \vdots \\ a_{L-1}(i,j) \\ a_L(i,j) \end{matrix}\right\} \begin{matrix} i = 1,2,\ldots,N \\ j = 1,2,\ldots,N \end{matrix}$$

Based on the above definitions, the Wiener index is equal to:

$$W = \sum_{i=1}^{N} \sum_{j=i+1}^{N} a_L(i,j)$$

Step 3 of the algorithmic procedure is accomplished through the use of the following constraints:

$$\left.\begin{array}{l} a_l(i,j) \le a_{l-1}(i,k) + a_{l-1}(k,l) \\ a_l(i,j) \ge a_{l-1}(i,k) + a_{l-1}(k,j) \\ \quad - 2N(1 - ay_l(i,k,j)) \\ ay_l(i,k,j) \in \{0,1\} \end{array}\right\} \begin{array}{l} i = 1,2,\ldots,N \\ j = 1,2,\ldots,N \\ k = 1,2,\ldots,N \\ l = 1,2,\ldots,L \end{array}$$

$$\left.\sum_{k=1}^{N} ay_l(i,k,j) = 1\right\} \begin{array}{l} i = 1,2,\ldots,N \\ j = 1,2,\ldots,N \\ l = 1,2,\ldots,L \end{array}$$

where $ay_l(i,k,j)$ is an additional binary variable which selects the minimum term.

Given a target Wiener index of 18 the structure of 2-methyl butane was correctly generated when this MILP formulation was solved using GAMS/CPLEX on a IBM RS6000 43P-133 workstation with an absolute convergence tolerance of 10^{-6}.

6. Stochastic formulation

The estimation of the QSAR/QSPR model parameters is typically accomplished through multilinear regression. This implies that the employed values are only "best" estimates and in reality fluctuations around them must be expected. This uncertainty in model parameters can be quantified using the method described in Maranas (1997a). The stochastic formulation and the deterministic equivalent formulation for the molecular design problems **(I)** and **(II)** are described in this section. Only the final expressions for the formulation are given, the details of the formulation can be found in Maranas (1997a). The stochastic formulation of problem **(I)** is:

$$\text{Max } \alpha$$

$$\text{subject to } Pr\left[s_0 \ge \frac{1}{p_j^0}|P_j - p_j^0|\right] \ge \alpha, \ \forall j \in J$$

The properties being considered are denoted by the set J. The property j is denoted by the random variable P_j and p_j^0 is the target value of property j. s_0 is the maximum allowable scaled violation from the target property values. Formulation **(I)** finds the maximum probability α for which all scaled property violations are less than s_0. The deterministic equivalent formulation of **(I)** is:

$$\text{Max } \alpha$$

$$\text{subject to } p_j^0(1 - s_0) - \mu(P_j) \le -t_j^l \sigma(P_j), \ \forall j \in J$$

$$p_j^0(1 + s_0) - \mu(P_j) \ge t_j^u \sigma(P_j), \ \forall j \in J$$

$$\Phi(t_j^l) + \Phi(t_j^u) \ge 1 + \alpha, \ \forall j \in J$$

where $\mu(P_j)$ and $\sigma(P_j)$ are the mean and standard deviation of the realization of property P_j respectively, and Φ denotes the standardized normal cumulative distribution.

The stochastic formulation of **(II)** is:

$$\text{Max } p_k$$

$$\text{Subject to } Pr[P_k \ge p_k] \ge \alpha$$

$$Pr[l_j \le P_j \le u_j] \ge \beta, \ \forall j \in J$$

where l_j, u_j are the imposed lower and upper bounds on the value of property j. The deterministic equivalent formulation of **(II)** is:

$$\text{Max } p_k$$

$$\text{Subject to } p_k - \mu(P_k) \le -\Phi^{-1}(\alpha)\sigma(P_k)$$

$$l_j - \mu(P_j) \le -t_j^l \sigma(P_j), \ \forall j \in J$$

$$u_j - \mu(P_j) \ge t_j^u \sigma(P_j), \ \forall j \in J$$

$$\Phi(t_j^l) + \Phi(t_j^u) \ge 1 + \beta, \ \forall j \in J$$

Subsequent discussion in this section addresses the estimation of the mean and variance of the random variable P denoting the property value. This is accomplished using the method outlined in Snedecor and Cochran (1989). The employed regression model is:

$$P = b_0 + \sum_{i=1}^{M} b_i x_i + \varepsilon$$

where

(i) P is the random variable whose realization is the property value.
(ii) x_i, $(i = 1,2,\ldots,M)$ are the structural descriptors (i.e., Wiener index, Kier's shape indices, etc.).
(iii) b_i, $(i = 0,1,\ldots,M)$ are the unknown parameters of the model.
(iv) ε is a normal random variable with mean zero and variance $\sigma_{p.x}^2$ which is independent of the compound under consideration.

Based on the above definition the mean of P is equal to:

$$\mu(P) = b_0 + \sum_{i=1}^{M} b_i x_i$$

The model parameters are estimated based on the known property values of a set of K compounds. These values, denoted by p_k^{exp}, $k = 1,2,\ldots,K$, are realizations of the random variables,

$$P_k = b_0 + \sum_{i=1}^{M} b_i x_{ki} + \varepsilon_k, \quad k = 1,2,\ldots,K$$

respectively. Using matrix notation, the column vector $\mathbf{p}^{exp}$ is a realization of the random vector $\mathbf{Xb} + \bar{\varepsilon}$ where:

(i) $\mathbf{X}$ is the augmented $(K \times (M+1))$ data matrix (x_{ki}) in which $x_{k0} = 1$ and x_{ki}, $i \neq 0$ is the structural descriptor i of compound k from the data set.
(ii) $\mathbf{b} = (b_0, b_1, \ldots, b_M)^T$
(iii) $\bar{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_K)^T$

The random vector,

$$\mathbf{a} = (a_0, a_1, \cdots a_M)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{Xb} + \bar{\varepsilon})$$

is an unbiased estimator of the unknown model parameter vector $\mathbf{b}$. This relation implies that,

(i) the mean of the random vector $\mathbf{a}$ is equal to the unknown model parameter vector $\mathbf{b}$, and
(ii) the variance-covariance matrix of $\mathbf{a}$ is given by:

$$[Cov(a_i, a_j)] = \sigma^2_{p.x}(\mathbf{X}^T\mathbf{X})^{-1}$$

Since $\mathbf{p}^{exp}$ is a realization of the random vector $\mathbf{Xb} + \bar{\varepsilon}$, an unbiased estimate of the unknown model parameter vector $\mathbf{b}$ is given by:

$$\underline{a} = (\underline{a_0}, \underline{a_1}, \cdots \underline{a_M})^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{p}^{exp}$$

Therefore, an unbiased estimate of the mean of a property with structural descriptors x_i is:

$$\hat{\mu}(P) = \underline{a_0} + \sum_{i=1}^{M} \underline{a_i} x_i$$

The variance of the predicted property value for a new compound consists of two terms:

1. Uncertainty in the estimation of $\hat{\mu}(P)$. This is specific to the compound under consideration and is quantified by the variance

$$\sigma^2_{p.x} \sum_{i=0}^{M} \sum_{j=0}^{M} S_{ij} x_i\, x_j$$

where S_{ij} is the (ij)th element of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.
2. Uncertainty due to the random error ε, which according to the model is independent of the compound under consideration. The variance of ε is $\sigma^2_{p.x}$.

Therefore, the variance of P is given by:

$$Var(P) = s^2_{p.x} + s^2_{p.x} \sum_{i=0}^{M} \sum_{j=0}^{M} S_{ij} x_i x_i$$

where

$$s_{p.x} = \sqrt{\frac{\sum_{k=1}^{K} (p_k^{exp} - \hat{\mu}(P))^2}{K - M - 1}}$$

is an unbiased estimator of $\sigma_{p.x}$. Note that the denominator $(K - M - 1)$ indicates that $(M + 1)$ degrees of freedom have been eliminated in estimating the $(M + 1)$ parameters of the model.

7. Examples

The application of the proposed linearization techniques enabling the incorporation of topological indices as structural descriptors within an MILP optimization framework is next highlighted with two example problems. The first example involves the identification of alkane molecules with targeted physical property values correlated with Kier's indices. In the second example, a search is conducted for an

Table 3. Abbreviations and units of properties in example 1

	Boiling point	Molar volume	Molar refractivity	Heat of vaporization
Units	°C	cm³/mol at 20°C	cm³/mol at 20°C	KJ/mol at 25°C
Abbr.	BP	MV	MR	HV
	Critical temperature	Critical pressure	Surface tension	
Units	°C	atm	dyn/cm at 20°C	
Abbr.	TC	PC	ST	

Table 4. Regression coefficients for property $= a_0 + a_1({}^1\kappa) + a_2({}^2\kappa) + a_3({}^3\kappa)$ in example 1

Property	a_0	a_1	a_2	a_3	r	s	n
BP	− 98.037	24.871	4.65	− 1.596	0.991	4.258	70
MV	39.982	14.571	− 0.226	1.242	0.995	1.724	69
MR	2.379	4.535	− 0.009	0.091	0.999	0.125	69
HV	2.683	3.971	1.285	− 0.253	0.996	0.48	69
TC	41.888	30.98	3.845	− 3.826	0.976	8.358	70
PC	43.498	− 1.791	− 0.219	− 0.426	0.947	0.858	70
ST	9.534	1.417	0.478	− 0.453	0.974	0.447	68

Table 5. Target values of properties in example 1

Target						
BP	MV	MR	HV	TC	PC	ST
80.0	144.0	34.0	32.0	260.0	30.0	20.0

agrochemical molecule which maximizes affinity while satisfying lower and upper bounds on mobility and retention.

7.1. Design of alkane molecules with targeted physical properties

This example addresses the design of alkanes with customized physical properties. The employed topological indices which serve as descriptors in the QSPRs are the Kier's shape indices of order one, two and three (see Skvortsova *et al.* (1993)). As mentioned earlier Kier's shape indices quantify the shape of a molecule by accounting for different degrees and location of branching. The seven properties considered in this example are boiling point, molar volume, molar refraction, heat of vaporization, critical temperature, critical pressure and surface tension. The data units and abbreviated acronyms for these properties (taken from Needham *et al.* (1988)), are given in Table 3. The regression equations of the physical properties, correlated with Kier's shape indices of order one, two and three, are given in Table 4. The target values of the properties are given in Table 5. Based on the discussion of Section 3, this example corresponds to the optimization problem **(I)**. The mathematical formulation is given by:

$$\text{Minimize } s$$

$$\text{subject to} \left.\begin{aligned} s &\geq \frac{1}{p_l^{target}}(p_l - p_l^{target}) \\ s &\geq -\frac{1}{p_l^{target}}(p_l - p_l^{target}) \end{aligned}\right\}, \quad l = 1, 2, \ldots, 7$$

where $p_l = a_{l0} + a_{l1}({}^1\kappa) + a_{l2}({}^2\kappa) + a_{l3}({}^3\kappa)$, $l = 1, 2, \ldots, 7$ and a_{l0}, a_{l1}, a_{l2} and a_{l3} are the regression coefficients. The equivalent linear representations for the Kier's shape indices discussed in subsection 5.2 and the basic graph relations discussed in Section 4 complete the formulation. This example is solved using the GAMS/CPLEX interface on an IBM 43P–133 RS6000 workstation with an absolute converge tolerance of 10^{-4}. The five structures which most closely match the imposed property targets are summarized in Tables 6–8. The predicted and experimental property values of the compounds found using the formulation are given in Table 6. Table 7 shows the values of the shape indices of the compounds generated. Table 8 summarizes the generated compounds names along with the CPU time used.

From the target property violations of predicted and experimental values, listed in Table 8, it is clear

Table 6. Predicted and experimental values for properties in example 1

Compound number	BP		MV		MR		HV		TC		PC		ST	
	Found	*Exptl.*	*Found*	*Exptl*	*Found*	*Exptl*	*Found*	*Exptl*	*Found*	*Exptl*	*Found*	*Exptl*	*Found*	*Exptl*
1	82.702	86.064	144.761	144.530	34.346	34.332	32.817	33.02	257.557	263.00	29.312	30.00	19.365	19.59
2	80.415	80.882	144.872	145.191	34.350	34.374	32.185	32.04	255.666	258.30	29.419	29.75	19.130	18.76
3	86.039	89.784	144.599	144.153	34.339	34.324	33.739	34.24	260.316	264.60	29.155	29.20	19.708	19.96
4	85.859	90.052	148.489	147.656	34.633	34.591	34.316	34.80	251.813	257.90	27.493	27.20	18.727	19.29
5	80.719	80.500	148.739	148.949	34.643	34.619	32.896	32.88	247.562	247.10	27.735	27.40	18.198	18.15

Table 7. Shape indices of the five compounds in example 1

Cmpd.	$^1\kappa$	$^2\kappa$	$^3\kappa$
1	7.000	2.344	2.667
2	7.000	1.852	2.667
3	7.000	3.061	2.667
4	7.000	4.167	6.000
5	7.000	3.061	6.000

that the predicted ranking of compounds after ignoring uncertainty does not perfectly reflect reality. Specifically, the compound which is ranked as number one is in fact the second best based on the experimental information. This discrepancy is due to errors (uncertainties) in the parameters of the regression models. This uncertainty can be quantified using the stochastic formulation for problem **(I)** described in Section 6. The deterministic equivalent of the problem is solved as an MINLP using GAMS/DICOPT on a IBM RS6000 43P-133 machine for different values of s_0. The results are summarized in Table 9. It can be seen that as the value of allowable maximum scaled deviation s_0 increases, the probability α of satisfying it becomes larger. Most notably by selecting a high enough probability, (i.e., $\alpha \geq 0.7$) *the true ranking of the best three solutions, as identified from experimental data, is preserved by the stochastic optimization formulation.* This is not the case when uncertainty is ignored (see Table 8). The trade-off curves between probabilities and maximum scaled deviations are given in Fig. 5. The labels for the curves in Fig. 5 refer to the ranking of the compound in the deterministic formulation. These trade-off curves give a systematic way to choose a molecule based on the maximum target violation tolerated and the minimum probability of satisfying it.

7.2. Substituent selection for optimal fungicidal and insecticidal properties of dialkyldithiolanylidenemalonates

This example addresses the optimal substituent selection for Dialkyl Dithiolanylidenemalonates (DD) shown in Fig. 6 which protect rice plants against the blast disease. Uchida (1980) quantified the effects of DD in terms of the affinity, mobility and retention of the compound to the plant. Uchida (1980) correlated $log(V_E)$, $log(\mu)$ and $log[R/(1-R)]$ (referred to as affinity, mobility and retention) with the hydrophobic parameter $log(P)$. Murray *et al.* (1975) showed that $log(P)$ is correlated linearly with the first-order molecular connectivity index $^1\chi$. In this example the affinity, mobility and retention are correlated with the topological index $^1\chi$. The regression equations are given in Table 10. The $^1\chi$ value is computed only for that part of the compound which changes (active substituent). The objective in the optimization formulation is to maximize the affinity subject to lower and upper bounds on the values for mobility and retention. This corresponds to the optimization formulation **(II)** of Section 3 and is expressed mathematically as:

$$\max \log(V_E)$$

$$\text{subject to } m^L \leq \log(\mu) \leq m^U$$

$$r^L \leq \log\left(\frac{R}{1-R}\right) \leq r^U$$

$$\log(V_E) = 0.5751(^1\chi) - 0.2942$$

$$\log(\mu) = -0.6983(^1\chi) + 2.0143$$

$$\log\left(\frac{R}{1-R}\right) = 0.787(^1\chi) - 2$$

The lower and upper bounds on mobility are $m^L = -0.3$ and $m^U = 0.3$ and those for retention $r^L = -0.3$ and $r^U = 1.0$ respectively. This formulation is solved using the GAMS/CPLEX interface on

Table 9. Probabilities α and best compounds for different scaled property violations s_0

s_0	α	Compound	CPU (sec)
0.0338	0.3937	3,3-Dimethyl pentane	57.6
0.0500	0.6118	2,2,3-Trimethyl butane	232.55
0.0750	0.8295	2,2,3-Trimethyl butane	218.67
0.1000	0.9324	2,2,3-Trimethyl butane	53.97
0.1250	0.9776	2,2,3-Trimethyl butane	151.29
0.1500	0.9939	2,2,3-Trimethyl butane	152.23
0.1750	0.9986	2,2,3-Trimethyl butane	226.39
0.2000	0.9997	2,2,3-Trimethyl butane	19.35

Table 8. Generated structures and CPU times

Cmpd. no.	Cmpd. name	CPU (sec)	% Maximum target violation of	
			Found value	Exptl. value
1	3,3-Dimethyl pentane	78	3.4	7.6
2	2,2,3-Trimethyl butane	59	4.4	6.2
3	2,3-Dimethyl pentane	91	7.6	12.2
4	2-Methyl hexane	95	8.4	12.6
5	2,4-Dimethyl pentane	506	9.0	9.3

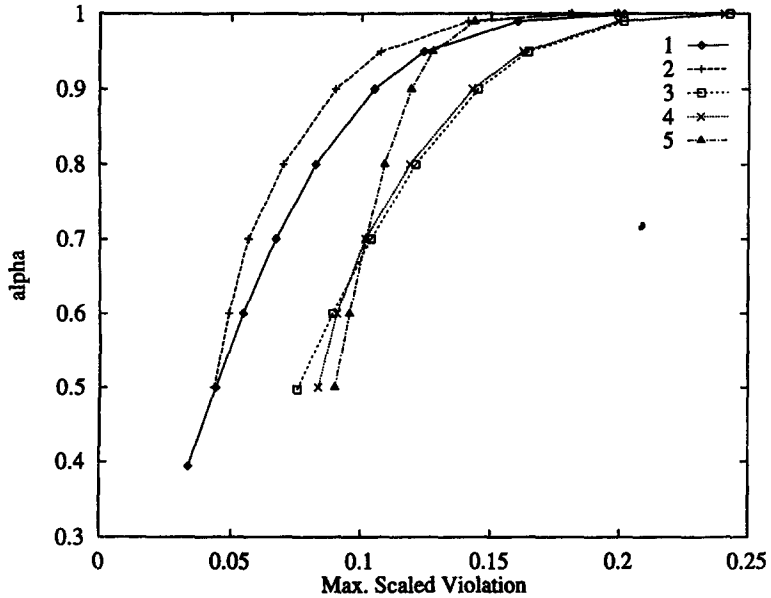

Fig. 5. Trade-off curves between α and s_0.

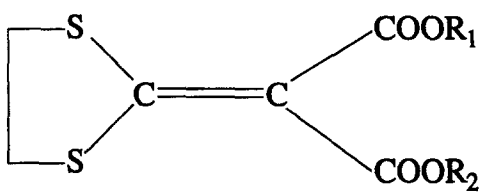

Fig. 6. Dialkyl dithiolanylidenemalonates.

a IBM 43P-133 RS6000 workstation with an absolute convergence tolerance of 10^{-6}. The results are summarized in Table 11. The experimental data set in Uchida (1980) mostly contained compounds for which R_1 and R_2 are the same. However, the solution of this formulation generates structures which do not always have the same substituents R_1 and R_2. In fact, most of the compounds shown in Table 11 were not in the initial data set. This validates the importance of molecular design in cases where experimental data are scarce. The compounds in this example may not necessarily give the predicted activity in practice, but it gives a possible direction of substituent selection for compound **(DD)**. Since there is uncertainty in the parameters of the regression estimates, the second best structure, third best structure etc. are also generated along with the optimal structure. The uncertainty in this example is quantified using the stochastic formulation of problem **(II)** discussed in Section 6. The analysis for this example is performed as described in (Maranas, 1997a). For each of the structures determined by the deterministic formulation, the maximum probability that all properties will satisfy their respective bounds is determined by solving the deterministic equivalent of the following problem.

$$\text{Max } \beta$$

$$\text{subject to } Pr[l_j \leq P_j \leq u_j] \geq \beta$$

In this example the properties which should remain within bounds are the mobility and the retention. The values of β obtained for the five structures of the deterministic formulation are listed in Table 12. Note that the probability β of satisfying the property bounds increases as the ranking of the compound decreases as observed elsewhere (Maranas, 1997a).

Table 10. Regression coefficients for property $= a_0 + a_1({}^1\chi)$ in example 2

Property	a_0	a_1	r	s
affinity: $\log(V_E)$	− 0.2942	0.5751	0.9844	0.1002
mobility: $\log(\mu)$	2.0143	− 0.6983	0.9801	0.1376
retention: $\log\left(\frac{R}{1-R}\right)$	− 2.0000	0.7870	0.9589	0.2269

Table 11. Five structures generated in example 2

Number	Found values of			CPU (sec)	R_1	R_2
	Mobility log (μ)	Affinity log (V_E)	Retention log ($\frac{R}{1-R}$)			
1	−0.2957	1.6083	0.6034	209	methyl methyl ethyl	3-methyl-butyl 2-pentyl *sec*-butyl
2	−0.2691	1.5864	0.5735	212	methyl ethyl n-propyl	*iso*-pentyl *iso*-butyl *iso*-propyl
3	−0.2068	1.5350	0.5032	244	methyl	2-methyl-2-butyl
4	−0.1685	1.5035	0.4601	193	*iso*-propyl	*iso*-propyl
5	−0.1653	1.5009	0.4565	281	methyl	*tert*-pentyl

Table 12. Probabilities β that property values are within bounds for five designs

Number	β
1	0.5116
2	0.5839
3	0.7395
4	0.8170
5	0.8227

8. Summary and conclusions

A new methodology was proposed for incorporating topological indices as structural descriptors for property correlation within an MILP optimization framework. The advantage of topological indices is that they encode information about molecular interconnectivity yielding, in principle, more accurate correlations of properties than simple group contributions. Three popular topological indices were considered: Randić's molecular connectivity indices, Kier's shape indices and the Wiener index. It was shown how to systematically recast the original nonlinear functional dependence of topological indices on the elements of the adjacency matrix with linear relations. This enabled the formulation of the problem as a Mixed Integer Linear Program (MILP). The proposed methodology was illustrated with an example involving the design of alkanes with target physical properties correlated with Kier's shape indices and a second example involving optimal substituent selection for a compound to obtain desired fungicidal properties correlated with the molecular connectivity index $^1\chi$. These examples illustrated the ability of the method to (i) simultaneously consider multiple target properties, (ii) generate not only the optimal structure but also the second best structure, third best structure etc., and (iii) quantify the effect of property prediction uncertainty.

The employed adjacency matrix description of molecular graphs provides a paradigm for the development of structure-property prediction methods beyond topological indices to supplement group contributions. For example, group contribution techniques provide information only about the number of different groups participating in the molecule. Using the principles described in this paper the next step of generating all structures consistent with the obtained group distribution can be accomplished. While in this paper heteroatoms and/or multiple bonds are not taken into account the extension of the proposed approach to account for the presence of multiple bonds and heteroatoms is necessary and conceptually straightforward. The penalty for the additional detail will be increasing the complexity of the problems to be solved. So far, no attempt has been made to take advantage of the structure of the resulting MILP formulations. This will become imperative for larger molecular design problems especially when process design considerations are embedded in the model. Additional issues which require further study include the problems of graph isomorphism and disconnected graphs which are briefly discussed in Appendix A.

Acknowledgements

Financial support by the NSF Career Award CTS-9701771 and Du Pont's Educational Aid Grant 1996/97 is gratefully acknowledged.

References

Baskin, I.I., Gordeeva, E.V., Devdariani, R.O., Zefirov, N.S., Palyulin, V.A. and Stankevich, M.I. (1990) Methodology of solution of the inverse problem for the structure-property relationship for the case of topological indices. *Doklady Akademii Nauk SSSR: Chemistry (English Translation)* **307**, 217.

Bersohn, M. (1983) A fast algorithm for calculation of the distance matrix of a molecule. *Journal of Computational Chemistry* **4**, 110.

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

Churi, N. and Achenie, L.E.K. (1996) Novel mathematical programming model for computer aided molecular design. *Industrial Engineering and Chemistry Research* **35**, 3788.

de Waterbeemd, H.V. (1995) *Chemometric Methods in Molecular Design.* VCH.

Duvedi, A.P. and Achenie, L.E.K. (1996) Designing Environmentally Safe Refrigerants Using Mathematical Programming. *Chemical Engineering Science* **51**, 3727.

Floudas, C.A. (1995) *Nonlinear and Mixed-Integer Optimization, Fundamentals and Applications.* Oxford University Press.

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

Glover, F. (1975) Improved linear integer programming formulations of nonlinear integer problems. *Management Science* **22**, 455.

Gordeeva, E.V., Molchanova, M.S. and Zefirov, N.S. (1990) General methodology and computer program for the exhaustive restoring of chemical structures by molecular connectivity indexes. Solution of the inverse problem in QSAR/QSPR. *Tetrahedron Computer Methodology* **3**, 389.

Gutman, I. (1993) A new method for the calculation of the wiener number of acyclic molecules. *Journal of Molecular Structure (Theochem)* **285**, 137.

Hall, L.H., Dailey, R.S. and Kier, L.B. (1993b) Design of molecules from quantitative structure-activity relaltionship models 3. Role of higher order path counts: Path 3. *Journal of Chemical Information and Computer Science* **33**, 598.

Hall, L.H., Kier, L.B. and Frazer, J.W. (1993a) Quantitative structure-activity relationship models 2. Derivation and proof of information transfer relating equations. *Journal of Chemical Information and Computer Science* **33**, 148.

Harary, F. (1972) *Graph Theory.* Addison-Wesley Series in Mathematics.

Hosoya, H. (1971) Topological index. A newly proposed quantity charecterizing the topological nature of structural isomers of saturated hydrocarbons. *Bulletin of Chemical Society of Japan* **44**, 2332.

Kier, L.B. (1985) A shape index from molecular graphs. *Quantitative Structure Activity Relationships* **4**, 109.

Kier, L.B. (1986) Shape indexes of order one and three from molecular graphs. *Quantitative Structure Activity Relationships* **5**, 1.

Kier, L.B. and Hall, L.H. (1976) Molecular connectivity VII: specific treatment of heteroatoms. *Journal of Pharmaceutical Sciences* **65**, 1806.

Kier, L.B. and Hall, L.H. (1986) Molecular connectivity in structure activity analysis. John Wiley and Sons.

Kier, L.B., Hall, L.H. and Frazer, J.W. (1993) Quantitative structure-ativity relationship models 1. Information transfer between path and vertex degree counts. *Journal of Chemical Information and Computer Science* **33**, 143.

Kier, L.B., Hall, L.H., Murray, W.J. and Randic, M. (1975) Molecular connectivity I: Relationship to nonspecific local anesthesia. *Journal of Pharmaceutical Sciences* **64**, 1971.

Kier, L.B. and Murray, W.J. (1975) Molecular connectivity. 4. Relationships to biological activities. *Journal of Medicinal Chemistry* **18**, 1272.

Kvasnička, V. and Pospíchal, J. (1990) Canonical indexing and constructive enumeration of molecular graphs. *Journal of Chemical Information and Computer Science* **30**, 99.

Maranas, C.D. (1996) Optimal computer-aided molecular design: A polymer design case study. *Industrial and Engineering Chemistry Research* **35**, 3403.

Maranas, C.D. (1997a) Optimal molecular design under property prediction uncertainty. *AIChE Journal* **43**(5), 1250.

Mavrovouniotis, M.L. (1996) Product and process design with molecular-level knowledge. *AIChE Symposium Series* **92**, 133.

Müller, W.R., Szymanski, K., Knop, J. V. and Trinajstić, N. (1987) An algorithm for construction of the molecular distance matrix. *Journal of Computational Chemistry* **8**, 170.

Murray, W.J., Hall, L.H. and Kier, L.B. (1975) Molecular connectivity III: Relationship to partition coefficients. *Journal of Pharmaceutical Sciences* **64**, 1978.

Needham, D.E., Wei, I. and Seybold, P.G. (1988) Molecular modeling of the physical properties of alkanes. *Journal of American Chemical Society* **110**, 4186.

Nemhauser, G.L. and Wolsey, L.A. (1988) *Integer and combinatorial optimization.* John Wiley and Sons.

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

Randić, M. (1975) On characterization of molecular branching. *Journal of American Chemical Society* **97**, 6609.

Randić, M. (1977) On canonical numbering of atoms in a molecule and graph isomorphism. *Journal of Chemical Information and Computer Science* **17**, 171.

Ravindran, A., Phillips, D.T. and Solberg, J.J. (1987) *Operations research, principles and practice.* John Wiley and Sons.

Reynolds, C.H., Holloway, M.K. and Cox, H.K. (1995) Computer-aided molecular design: Applications in agrochemicals, materials, and pharmaceuticals, *ACS Symposium Series* 589.

Salkin, H.M. and Mathur, K. (1989) *Foundations of integer programming.* Elsevier Publishing Co.

Senn, P. (1988) The computation of the distance matrix and the Wiener index for graphs of arbitrary complexity with weighted vertices and edges. *Computational Chemistry* **12**, 219.

Skvortsova, M.I., Baskin, I.I., Slovokhotova, O.L., Palyulin, V.A. and Zefirov, N.S. (1993) Inverse problem in QSAR/QSPR studies for the case of topological indices charecterizing molecular shape (Kier Indices). *Journal of Chemical Information and Computer Science* **33**, 630.

Skvortsova, M.I., Baskin, I.I., Slovokhotova, O.L., Palyulin, V.A. and Zefirov, N.S. (1996) The inverse problem of the structure-property relationship problem in the case of the correlation equation with arbitrary topological descriptors. *Doklady Akademii Nauk SSSR: Chemistry (English Translation)* **346**, 37.

Skvortsova, M.I., Stankevich, I.V. and Zefirov, N.S. (1992) Generation of molecular structures of polycondensed benzenoid hydrocarbons using the Randic index. *Journal of Structural Chemistry (English Translation)* **33**, 416.

Snedecor, G.W. and Cochran, W.G. (1989) *Statistical methods, eighth edition.* Iowa State University Press, Ames, Iowa.

Spialter, L. (1964) The atom connectivity matrix (ACM) and its characteristic polynomial (ACMCP). *Journal of Chemical Documentation* **4**, 261.

Trinajstic, N. (1992) *Chemical graph theory.* CRC Press.

Trudeau, R.J. (1976) *Dots and lines.* The Kent State University Press.

Uchida, M. (1980) Affinity and mobility of fungicidal dialkyl dithiolanylidenemalonates in rice plants. *Pesticide Biochemistry and Physiology* **14**, 249.

Vaidyanathan, R. and El-Halwagi, M. (1996) Computer aided synthesis of polymers and blends with target properties. *Industrial and Engineering Chemistry Research* **35**, 627.

Wiener, H. (1947a) Structural determination of paraffin boiling points. *Journal of American Chemical Society* **69**, 17.

Wiener, H. (1947b) Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *Journal of American Chemical Society* **69**, 2636.

Wiener, H. (1947c) Influence of interatomic forces on paraffin properties. *Journal of Chemical Physics* **15**, 766.

Wiener, H. (1948a) Vapor pressure-temperature relationships among the branched paraffin hydrocarbons. *Journal of Physical Chemistry* **52**, 425.

Wiener, H. (1948b) Relation of the physical properties of the isomeric alkanes to molecular structure. *Journal of Physical Chemistry* **52**, 1082.

Appendix A. Graph isomorphism and disconnected graphs

The importance of generating the second, third, etc. best structures was mentioned in many places in the paper. This is performed by adding constraints referred to as integer cuts to make the previous solution infeasible (Floudas, 1995). However, different labeling permutations of the same graph give rise to as many as $N!$ vertex adjacency matrices. Two graphs which are the same in all respects except for the labeling of vertices are called isomorphic graphs. Different labelings of the same graph have different adjacency matrices. Hence adding integer cuts may give rise to the same structure though with a different adjacency matrix corresponding to a different labeling of the same graph. The problem of graph isomorphism has attracted the attention of both mathematicians and chemists. In particular, considerable work has been expended into enumerating non-redundant chemical structures. The adjacency matrix of two isomorphic graphs are related by (Trinajstic, 1992):

$$P^T A_1 P = A_2 \tag{8}$$

where A_1 and A_2 are the adjacency matrices of the two isomorphic graphs. P is the $N \times N$ permutation matrix whose elements are:

$$p_{ij} = \begin{cases} 1 & \text{if vertex i in graph 1 is labeled as j in 2} \\ 0 & \text{otherwise} \end{cases}$$

In principle, the generation of isomorphic graphs can be avoided by adding constraints of the form:

$$\sum_{i=1}^{N} \sum_{j=i}^{N} |(P^T A_1 P)_{ij} - A_{ij}| \geq 1$$

where A_1 is the adjacency matrix of the previous solution. This constraint must be added for all possible permutation matrices which becomes prohibitively large even for moderate size problems. An alternative approach is to number the atoms of a molecule in a specific way referred to as canonical indexing. Randić (1977) proposed a numbering which gives rise to a minimal code of the adjacency matrix. The code of an adjacency matrix is found by writing all the rows of the adjacency matrix in a single line thus giving rise to a single binary number. The minimal code is the code with the smallest binary number. For example, the binary code of the adjacency matrix:

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

is found by arranging the elements of all the rows in a single line.

0111 1000 1000 1000

Other criteria include the maximal code of the whole adjacency matrix and the maximal/minimal code of the upper triangular/lower triangular part of the adjacency matrix (Kvasnička and Pospíchal, 1990). Gordeeva *et al.* (1990) used a labeling in which the valency of a vertex does not increase as the number of the label increases. In this framework considerable improvement has been obtained in generating non-redundant structures by incorporating two properties of the indexing proposed by Randić (1977):

1. Lower numbering is assigned to vertices with lower valencies.
2. The vertices with the lowest possible label are connected to vertices with highest possible label.

These two rules are incorporated in the MILP optimization framework in the form of the following constraints:

$$V_i \leq V_{i+1},\ i = 1, 2, \ldots, N-1$$

For vertices with the same valency the following constraints must hold:

$$\sum_{j=1}^{N} j a_{ij} \geq \sum_{j=1}^{N} j a_{(i+1)j},\ i = 1, 2, \ldots, N-1$$

This is ensured by introducing a binary variable y_i and transforming the two constraints as follows:

$$\left.\begin{aligned} &V_i \leq V_{i+1} - (1 - y_i) \\ &V_i \geq V_{i+1} - M(1 - y_i) \\ &\sum_{j=1}^{N} j a_{ij} \geq \sum_{j=1}^{N} j a_{(i+1)j} - M(1 - y_i) \end{aligned}\right\} i = 1, 2, \ldots, N-1$$

where M is a large positive number. Though this does not guarantee that the generation of equivalent structures will not happen, the occurrence of these have considerably reduced and no longer pose a significant problem in terms of the effort spent in enumerating them.

Another problem which arises in the MILP optimization framework is the generation of adjacency matrices which correspond to more than one distinct molecules (disconnected graph). Churi and Achenie (1996) proposed a straightforward way to avoid the occurrence of disconnected graphs. This is done by visualizing the construction of a molecule as a step by step process and specifying that the latest vertex (i) being added is connected to at least one of the already existing vertices ($1, 2, \ldots, i-1$). In the context of this work, this can be mathematically expressed as:

$$\left.\begin{aligned} &\sum_{j=1}^{i-1} a_{ji} \geq \sum_{k=1}^{4} \delta_i^k \\ &\sum_{k=1}^{4} \delta_{i-1}^k \geq \sum_{k=1}^{4} \delta_i^k \end{aligned}\right\} i = 2, 3, \ldots, N^U$$

However, these restrictions on the elements of the adjacency matrix are not compatible with the constraints imposed to prevent graph isomorphism. Hence they have not been incorporated in the optimization framework.