Optimal Molecular Design under Property Prediction Uncertainty

Costas D. Maranas
Dept. of Chemical Engineering, The Pennsylvania State University, University Park, PA 16802

A systematic method that quantitatively assesses property prediction uncertainty (imprecision) on optimal molecular design problems is introduced. Property–structure relations are described with specific nonlinear functionalities based on group contribution methods. Property prediction uncertainty is explicitly quantified by using multivariate probability density distributions to model the likelihood of different realizations of the group contribution parameters. Assuming stability of these probability distributions, a novel approach is introduced for transforming the original nonlinear stochastic formulation into a deterministic MINLP problem with linear binary and convex continuous parts with separability. The resulting convex MINLP formulation is solved to global optimality for molecular design problems involving many uncertain group contribution parameters. Results indicate the computational tractability of the method and the profound effect that property prediction uncertainty may have in optimal molecular design. Specifically, trade-off curves between performance objectives, probabilities of meeting the objectives, and chances of satisfying design specifications offer a concise and systematic way to guide optimal molecular design in the face of property prediction uncertainty.

Introduction

The systematic identification of molecular products with optimal values of thermophysical, mechanical, and/or biological properties is a key objective in the chemical industries. Financial success and market share are ultimately tied to a company's ability to continuously introduce novel, successful products. Computer-aided molecular design (CAMD) techniques are increasingly being employed to elicit promising candidates from the astounding plethora of different potential molecular alternatives. One can already find in the literature success stories of CAMD in identifying novel compounds substantially superior to existing ones. These include the antibacterial compound norfloxacin, designed by Kyorin Pharmaceutical Company in Japan, that is 500 times more potent than previously marketed compounds; the fungicide myclobutanil, developed by Rohm and Haas, and the herbicides Metamitron by Bayer AG, and bromobutide by Sumitomo (Boyd, 1990).

In the chemical engineering literature, a considerable amount of work has been devoted to the derivation of structure–property relations (Fredenslund et al., 1975; Joback and Reid, 1987; Constantinou and Gani, 1994; Mavrovouniotis, 1990). These relations have been utilized in the computer-aided design of molecular products such as polymers (Venkatasubramanian et al., 1994b; Vaidyanathan and El-Halwagi, 1995), extractants (Gani and Brignole, 1983; Odele et al., 1990; Naser and Fournier, 1991), solvents (Brignole et al., 1986; Odele and Macchietto, 1993), and refrigerants (Duvedi and Achenie, 1996; Churi and Achenie, 1996). The employed search algorithms include enumeration techniques (Stephanopoulos and Townsend, 1986; Joback, 1989; Joback and Stephanopoulos, 1989; Derringer and Markham, 1985), knowledge-based strategies (Brignole et al., 1986; Nagasaka et al., 1990; Nielsen and Gani, 1990; Gani et al., 1991), graph reconstruction methods (Gordeeva et al., 1990; Kier et al., 1983), multistage approaches (Naser and Fournier, 1991; Gani and Fredenslund, 1993), genetic algorithms (Venkatasubramanian et al., 1994a,b, 1995), artificial intelligence (Bolis et al., 1991), local MINLP optimization (Odele et al., 1990; Odele and Macchietto, 1993; Vaidyanathan and El-Halwagi, 1994; Duvedi and Achenie, 1996), interval Newton implementations (Vaidyanathan and El-Halwagi, 1995), mixed-integer linear optimization for linear structure–property

relations (Constantinou et al., 1996), and exact linear reformulations for specific nonlinear structure–property relations (Maranas, 1996). A thorough review of the latest developments in the area can be found in (Mavrovouniotis, 1996).

The present state of the art involves primarily search techniques of either a local or opportunistic nature that may or may not find the mathematically best design. Preliminary efforts at finding the best molecular design with mathematical certainty include the work of Vaidyanathan and El-Halwagi (1995), Constantinou et al. (1996), and Maranas (1996). Identifying the best molecular design with mathematical certainty is important because by eliminating the caveat of convergence to suboptimal molecular designs, the chances of identifying novel, possibly counterintuitive, superior design alternatives are improved. Furthermore, by removing the possibility of unknowingly converging to suboptimal solutions, possible discrepancies between obtained optimal solutions and experimentally derived designs can explicitly and unequivocably be attributed to uncertainty.

While group contribution methods (GCM) (Franklin, 1949) provide popular, versatile, and relatively accurate (Horvath, 1992) ways for estimating properties based on the number and type of molecular groups participating in a molecule or repeat unit, they provide only estimates for different properties and may only be in partial agreement with experimental values. The same is true even for sophisticated methods based on connectivity indices (Kier and Hall, 1986), factor analysis (Cattell, 1952), molecular similarity (Horvath, 1988), pattern recognition (Kowalski and Bender, 1972), and so forth. In fact, 5–10% (or even higher) discrepancies between experimental values and group contribution predictions are common. The relative accuracy of these predictions depends on the particular property (e.g., polymer density estimates are typically more accurate than glass transition temperature estimates), on molecular complexity, and on the particular prediction method. The frequency and magnitude of these discrepancies in most instances of property prediction provide ample motivation for quantifying their effect on optimal molecular design. Failure to handle, within a quantitative framework, the inherently approximate nature of structure–property predictive techniques may affect the quality and weaken confidence in the molecular designs obtained.

The effect of imprecision (uncertainty) in process systems was recognized early and has been the subject of extensive work. Grossmann and coworkers (Grossmann and Sargent, 1978; Halemane and Grossmann, 1983; Swaney and Grossmann, 1985a; Floudas and Grossmann, 1987; Pistikopoulos and Grossmann, 1988) pioneered the concept of the flexibility index to quantify the ability of a plant to operate over a range of conditions .while satisfying performance specifications. The concept was later extended to account for stochastic uncertainty measuring the probability of feasible operation by Straub and Grossmann (1990) and Pistikopoulos and Mazzuchi (1990). These ideas were mainly applied to problems in heat-exchanger networks (Floudas and Grossmann, 1987), multiproduct batch-plants design (Straub and Grossmann, 1992), and operational planning (Ierapetritou and Pistikopoulos, 1994). Other approaches for quantifying uncertainty in process systems include the early work of Friedman and Reklaitis (1975), two-stage expectation optimization (Pai and Hughes, 1987), and Monte Carlo simulations (Diwekar and Rubin, 1991; Reed and Whiting, 1993). While the problem of quantifying uncertainty in process systems has received considerable attention, there is so far, in the chemical engineering literature, little or no work on the effect of uncertainty in optimal molecular design.

It is the objective of this article to establish the necessary theoretical foundation and provide a tractable computational framework for quantifying the effect of uncertainty in optimal molecular design and to address questions such as:

1. Which molecular architecture is the *most likely* to meet a given design objective?
2. What are the chances that the optimal molecular design will indeed meet the performance target and design specifications?
3. How is the selection of the mathematically best molecular design affected by increasing/decreasing the desired probability of meeting the design target?

First, we provide a brief mathematical description of the specifics of the optimal molecular design problem. This is followed by a description of how property-prediction uncertainty can be quantified within a stochastic framework. Next, we show how one can equivalently express the resulting stochastic optimization formulation, based on notions from chance-constraint programming, as a deterministic mixed-integer nonlinear (MINLP) optimization problem with convex-continuous and linear-integer parts. Two case studies illustrate the proposed theoretical and computational framework and demonstrate the significant effect property-prediction uncertainty has on optimal molecular design.

Mathematical Description

The basic features of optimal molecular design can be captured mathematically in the following mixed-integer nonlinear optimization problem (Maranas, 1996):

$$\begin{aligned} \min \quad & \mathfrak{M}\mathcal{P}[p_j(\boldsymbol{n})] \qquad \text{(OMD)} \\ \text{subject to} \quad & p_j^L \le p_j(\boldsymbol{n}) \le p_j^U, \quad j=1,\ldots,M \\ & n_i \in \{n_i^L, n_i^L+1, \ldots, n_i^U\}, \quad i=1,\ldots,N. \end{aligned}$$

In formulation OMD, $\boldsymbol{n}=(n_1, \ldots, n_N)$ is the vector of the integer variables $n_i \in \{0, 1, 2, \ldots\}$ describing the number of times the ith molecular group participates in the molecule. The expressions $p_j = p_j(\boldsymbol{n})$, $j=1, \ldots, M$, established by group contribution methods, denote the functionality between property j and the type and number of different molecular groups. The objective function $\mathfrak{M}\mathcal{P}$ is a measure of performance of the molecular design and is typically a function of one or more properties, $\mathfrak{M}\mathcal{P} = \mathfrak{M}\mathcal{P}[p_j(\boldsymbol{n})]$. Additional constraints may be placed on OMD to ensure structural feasibility, such as,

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

where v_i is the valency (number of possible attachments) of the ith molecular group and f is the total number of remaining attachments available for bonding in a molecule ($f=0$) or polymer repeat unit ($f=2$). Chemical feasibility con-

straints may be added using developments of Pretel et al. (1994).

The following two most widely used measures of performance reflecting two distinct design philosophies are considered (see Maranas, 1996):

1. Optimal design of a molecular product with properties that match some prespecified targets. This objective is formulated as the minimization of the maximum allowable scaled deviation of properties from some target values (property matching):

$$\min \ \mathfrak{M}\mathcal{P}, \quad \text{where} \quad \mathfrak{M}\mathcal{P} = \max_j \frac{1}{p_j^s} |p_j(\boldsymbol{n}) - p_j^0|.$$

Here p_j^0 is the target value for property j and p_j^s is an appropriate scale. If the maximum percent property deviation is minimized, then $p_j^s = p_j^0$.

2. Identification of a molecular product with the largest or smallest value for one property while maintaining the other property values within some lower and upper bounds. This is formulated as the minimization/maximization of a single property j^* subject to lower and upper bounds on the rest of them (property optimization):

$$\min/\max \ \mathfrak{M}\mathcal{P}, \quad \text{where} \quad \mathfrak{M}\mathcal{P} = p_{j^*}(\boldsymbol{n}).$$

In the present study, emphasis is placed on the optimal design of polymers with optimized or customized property values. Based on the property compilation by van Krevelen (1990), most of the thermophysical, optical, electromagnetic, and mechanical property-estimating formulas for polymers conform or can be transformed to the following general functionality, as shown in Maranas (1996):

$$p_j(\boldsymbol{n}) = \left(\frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \right), \quad j = 1, \ldots, M.$$

Here A_{ij} and B_{ij} are given group contribution parameters associated with a specific molecular group i and property j and are independent of the particular molecular architecture.

A systematic analysis framework is presented in Maranas (1996) for transforming the original formulation with nonlinear structure–property functionalities of the aforementioned type into equivalent mixed-integer linear (MILP) problems that can be solved to global optimality. This removes the chance of converging to suboptimal solutions, and therefore any discrepancies between solutions obtained and experimental verifications can unequivocably be attributed to property-estimation imprecision. This article extends this optimization framework to account for property-prediction imprecision based on the probabilistic description of uncertainty.

Stochastic Formulation

Group contribution methods are based on the conjecture that property values for different molecules are uniquely defined by the type and number of the molecular groups composing the molecule (transferability assumption). However, discrepancies between model prediction and actual experimental data imply that the transferability assumption is not always rigorously satisfied. These discrepancies can be reconciled by recognizing that the contribution of molecular groups, for a given property, is not unchanged from one molecule to another (independent of molecular architecture), but varies slightly around some nominal value depending on the particular molecular architecture. This intuitive abstraction can be expressed mathematically by utilizing probability density distributions to describe the likelihood of different realizations of the group contribution parameters A_{ij}, B_{ij}. The proposed probabilistic description of uncertainty renders both design specifications and performance objective stochastic. Therefore, unless a particular realization of the uncertain parameters A_{ij}, B_{ij} is specified, questions regarding the satisfaction of design specifications and performance objectives cannot be fathomed. While the *a priori* identification of the realization of the uncertain parameters A_{ij}, B_{ij} and consequently of design specifications and performance objectives is impossible, the evaluation of the probability of meeting a performance target or maintaining feasibility of the design specification is computable since the uncertain parameters A_{ij}, B_{ij} assume values according to some known probability density distribution. This probabilistic description of performance objectives and constraints yields the following optimal molecular design problem under stochastic uncertainty on the group contribution parameters:

$$\begin{aligned} \max \quad & \mathfrak{M}\mathcal{P}^{\text{target}} && \text{(SOMD)} \\ \text{subject to} \quad & Pr\left\{ \mathfrak{M}\mathcal{P}[p_j(\boldsymbol{n})] \geq \mathfrak{M}\mathcal{P}^{\text{target}} \right\} \geq \alpha \\ & Pr\left\{ p_j^L \leq p_j(\boldsymbol{n}) \leq p_j^U \right\} \geq \beta, \quad j = 1, \ldots, M. \end{aligned}$$

Formulation SOMD involves a set of constraints imposing lower bounds on the probability of satisfying the performance objective and the imposed lower and upper bounds of the property. These constraints are called *chance constraints*. Formulation SOMD identifies the maximum value of the performance target $\mathfrak{M}\mathcal{P}^{\text{target}}$ that the stochastic performance objective $\mathfrak{M}\mathcal{P}$ can meet with probability of at least α (e.g., 90%), and at the same time maintain all property values within their respective lower and upper bounds with probability greater than or equal to β. Therefore, the solution of formulation SOMD will have at least an α chance of meeting the performance objective and at least a β chance of maintaining the property values within their designated bounds. By solving formulation SOMD for different values of α and β, trade-offs between the performance objective target $\mathfrak{M}\mathcal{P}^{\text{target}}$, the probability α of meeting this performance target, and the probability β of satisfying all property constraints can be established. These trade-offs can then provide a concise and systematic way of selecting the most promising molecular design in the face of property-prediction uncertainty. Next, two special instances of the general stochastic optimal molecular design problem SOMD are highlighted.

Property matching under uncertainty

After omitting, for the sake of succinctness, all deterministic linear constraints in $\boldsymbol{n}$ (i.e., structural feasibility require-

ments, variable bounds, etc.), the *stochastic property-matching* (SPM) problem under property-prediction uncertainty can be formulated as the following chance-constrained optimization problem:

$$\min \quad s$$

$$\text{subject to} \quad Pr\left[s \geq \frac{1}{p_j^s} \left| \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} - p_j^o \right| \right] \geq \alpha, \quad j = 1, \ldots, M \tag{SPM}$$

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

Formulation SPM identifies the type and number of molecular groups for which a scaled property-violation target s, guaranteed to be met by all properties j with probability of at least α, is minimized. For $\alpha = 0.5$, all properties j have at least a fifty-fifty chance of meeting the property-violation target s. Higher values of α reflect a more conservative attitude regarding meeting the property targets.

Property optimization under uncertainty

Often, rather than matching property values to some prespecified targets, the maximization or minimization of a single property j^* is sought while maintaining property values within some lower and upper bounds. This objective, under property-prediction uncertainty, can be expressed mathematically as the following *stochastic property optimization* (maximization) formulation SPO

$$\max \quad p_{j^*} \tag{SPO}$$

$$\text{subject to} \quad Pr\left[\frac{\sum_{i=1}^{N} A_{ij^*} n_i}{\sum_{i=1}^{N} B_{ij^*} n_i} \geq p_{j^*} \right] \geq \alpha$$

$$Pr\left[p_j^L \leq \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \leq p_j^U \right] \geq \beta, \quad j = 1, \ldots, M$$

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

Formulation SPO identifies the type and number of molecular groups for which a property target p_{j^*} for property j^*, met with probability α, is maximized, while lower and upper bounds on the rest of the properties are satisfied with probability of at least β. By manipulating the values of the probability levels α and β the relative importance of meeting a property target as opposed to satisfying the property bounds can be adjusted.

Deterministic Equivalent Representation

Formulations SPM and SPO involve probability terms whose evaluation for each realization of the deterministic variables requires the integration of multivariate probability density distributions. Many integration methods exist, but in general, they all exact a heavy computational burden, either in the form of additional variables (quadrature point integration), or excessive function and gradient evaluation (Monte Carlo integration) (Watanabe and Ellis, 1994) and thus are restricted to problems with only a few uncertain parameters. However, the number of uncertain parameters in optimal molecular design can be as high as $2MN$, where M and N are the total number of properties and molecular groups, respectively. Typically $M = 3$–5 and $N = 10$–30, therefore formulation OMD problems routinely involve from tens to hundreds of uncertain parameters. In this study, to deal with such a high number of possibly correlated uncertain parameters the exact transformation of the original stochastic constraints into equivalent deterministic ones is sought without having to resort to computationally intensive explicit or implicit multivariate integration. To this end, the deterministic equivalent representation of the chance-constrained formulations based on the ideas pioneered by (Charnes and Cooper, 1959, 1960, 1963) is pursued.

Basic results

Assuming that the uncertain parameters A_{ij} and B_{ij} follow *stable* (Allen et al., 1974) up to two-parameter probability density distributions (i.e., normal, Poisson, chi-square, binomial, etc.), chance constraints that are linear in terms of the uncertain parameters can be transformed into equivalent deterministic constraints. A probability density distribution F is stable if (1) it can be completely specified with only two parameters u, v, and (2) the convolution of any two distribution functions $F((x-u_1)/v_1)$ and $F((x-u_2)/v_2)$ is again of the form $F((x-u)/v)$ (Vajda, 1970). The normal distribution is selected in this work to describe the uncertainty associated with group contribution parameters. The normal distribution is a stable distribution with $u = \mu$ and $v = \sigma$, which adequately captures the qualitative trends of group contribution uncertainty. Specifically, it involves relatively high probability density around the mean, which gradually diminishes away from it.

To illustrate how the deterministic equivalent representation is obtained, the following general chance constraint is considered:

$$Pr\left[\sum_{i}^{n} a_i f_i(\boldsymbol{x}) \leq 0 \right] \geq \alpha.$$

Here a_i denotes the uncertain parameters whose realization follows a two-parameter stable probability distribution F (normal) and $f_i(\boldsymbol{x})$ are a set of functions of the deterministic variables $\boldsymbol{x}$. Let $\mu(a_i)$ denote the expected value of a_i; $\mathrm{Var}(a_i) = E[a_i - \mu(a_i)]^2$ the variance of a_i; and $\mathrm{Cov}(a_i, a_{i'}) = E[a_i - \mu(a_i)][a_{i'} - \mu(a_{i'})]$ the covariance between uncertain parameters a_i and $a_{i'}$. Here E represents the expectation operator. By subtracting the mean and dividing by the square root of the variance of $\sum_{i=1}^{n} a_i f_i(\boldsymbol{x})$, the chance constraint can equivalently be written as

$$Pr\left[\frac{\sum_i^n a_i f_i(x) - \mu\left[\sum_i^n a_i f_i(x)\right]}{\left\{\mathrm{Var}\left[\sum_i^n a_i f_i(x)\right]\right\}^{1/2}} \le \frac{-\mu\left[\sum_i^n a_i f_i(x)\right]}{\left\{\mathrm{Var}\left[\sum_i^n a_i f_i(x)\right]\right\}^{1/2}}\right] \ge \alpha.$$

Because the normal distribution is stable, the lefthand side expression,

$$\frac{\sum_i^n a_i f_i(x) - \mu\left(\sum_i^n a_i f_i(x)\right)}{\left[\mathrm{Var}\left(\sum_i^n a_i f_i(x)\right)\right]^{1/2}}$$

is a normally distributed random variable with a mean of zero and a variance of one (standardized form). Thus, if Φ is the standardized normal cumulative density distribution, then the chance constraint can be replaced by the following deterministic equivalent:

$$\Phi\left(\frac{-\mu\left[\sum_i^n a_i f_i(x)\right]}{\left\{\mathrm{Var}\left[\sum_i^n a_i f_i(x)\right]\right\}^{1/2}}\right) \ge \alpha.$$

By applying the inverse of the cumulative normal distribution function Φ^{-1} on both sides of the last relation, we get

$$\frac{-\mu\left[\sum_i^n a_i f_i(x)\right]}{\left\{\mathrm{Var}\left[\sum_i^n a_i f_i(x)\right]\right\}^{1/2}} \ge \Phi^{-1}(\alpha).$$

The original inequality sign is preserved because the inverse of the cumulative normal distribution is a monotonically increasing function (Abramowitz and Stegun, 1972). In Figure 1 the inverse of the cumulative standardized normal distribution is plotted vs. the probability level α. Rearranging terms yields,

$$\mu\left[\sum_i^n a_i f_i(x)\right] + \Phi^{-1}(\alpha)\left\{\mathrm{Var}\left[\sum_i^n a_i f_i(x)\right]\right\}^{1/2} \le 0$$

where

$$\mu\left[\sum_i^n a_i f_i(x)\right] = \sum_i^n \mu(a_i) f_i(x)$$

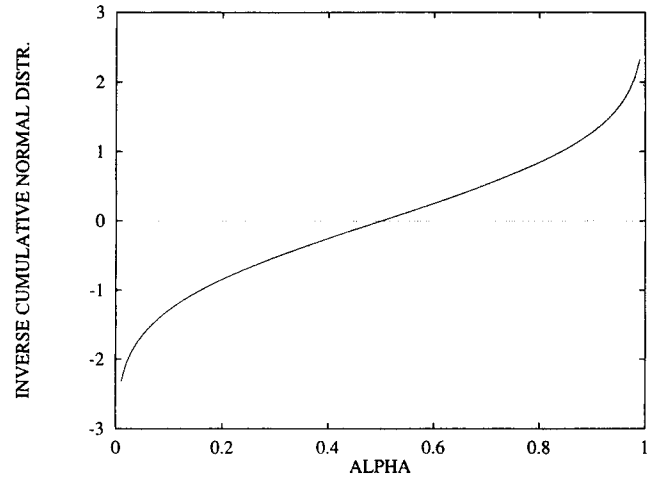

Figure 1. Inverse of the normal cumulative distribution.

and

$$\mathrm{Var}\left[\sum_i^n a_i f_i(x)\right] = \sum_{i=1}^n \mathrm{Var}(a_i) f_i(x)^2 + 2\sum_{i=1}^n \sum_{i'=i+1}^n f_i(x)\mathrm{Cov}(a_i, a_{i'}) f_{i'}(x).$$

Inspection of the deterministic equivalent constraint reveals that it is composed of the mean of the original constraint augmented by the square root of its variance times $\Phi^{-1}(\alpha)$. Typically, α is greater than 0.5, and thus $\Phi^{-1}(\alpha) \ge 0$. This implies that the variance term penalizes the deterministic constraint (for $\alpha \ge 0.5$), making it more restrictive than the mean of the original constraint. This is consistent with the intention of satisfying the original constraint not only at its expectation but also for all realizations with probability greater than or equal to α. As the imposed probability α of satisfying the constraint increases, $\Phi^{-1}(\alpha)$ increases as well, implying that the stricter (more conservative) the imposed probability levels are, the more difficult it will be to satisfy the chance constraint. In the limit of $\alpha = 1$, $\Phi^{-1}(\alpha)$ diverges to plus infinity and the chance constraint becomes rigorously infeasible for any values of the deterministic variables x. Kataoka (1963) showed that for $f_i(x) = x_i$, $i = 1, \ldots, n$, the square root of the variance,

$$\left[\mathrm{Var}\left(\sum_i^n a_i x_i\right)\right]^{1/2}$$

is a convex function in x. Therefore the deterministic equivalent constraint of

$$Pr\left(\sum_i^n a_i x_i \le 0\right) \ge \alpha$$

is convex for $\alpha \ge 0.5$ and concave for $\alpha \le 0.5$. Convexity of the deterministic equivalent representation of the chance constraint is very important in optimization studies because it

implies a single optimum solution that can be found efficiently with existing solvers.

In this subsection, we describe how a *single* general chance constraint with linearity in the stochastic parameters can be transformed into an equivalent deterministic constraint provided that the stochastic parameters follow stable distributions. If in addition, the deterministic variables x_i appear linearly in the constraint and the probability level of constraint satisfaction α is greater than or equal to 0.5, then the deterministic equivalent constraint is convex in x. In the following subsection, we address the deterministic equivalent representation of a specific joint chance constraint that captures all the essential features of the chance constraints found in formulations SPM and SPO.

Deterministic equivalents of joint constraints

The absolute-value constraint of formulation SPM and the upper and lower property bounds in formulation SPO give rise to joint chance constraints of the form

$$Pr\left[L_j \le \frac{\sum_i^N A_{ij} n_i}{\sum_i^N B_{ij} n_i} \le U_j \right] \ge \alpha, \quad j = 1, \ldots, M,$$

measuring the probability of maintaining the ratio of two linear expressions in n_i between a lower and an upper bound. Here A_{ij}, B_{ij} (group contribution parameters) are normally distributed stochastic parameters; n_i are deterministic variables; and L_j, U_j are given deterministic parameters.

After lumping all the stochastic parameters into a single stochastic variable

$$y_j = \sum_i^N A_{ij} n_i \Big/ \sum_i^N B_{ij} n_i,$$

we have $Pr[L_j \le y_j \le U_j] \ge \alpha$. Because $Pr[L_j \le y_j \le U_j] = 1 - Pr[y_j \le L_j] - Pr[y_j \ge U_j]$, $Pr[y_j \le L_j] = 1 - Pr[y_j \ge L_j]$, and $Pr[y_j \ge U_j] = 1 - Pr[y_j \le U_j]$, the original joint chance constraint can be decomposed into the sum of two nonjoint chance constraints,

$$Pr[y_j \ge L_j] + Pr[y_j \le U_j] - 1 \ge \alpha.$$

If β_j^1 and β_j^2 are the probabilities of satisfying $L_j \le y_j$ and $y_j \le U_j$, respectively, we have

$$\left.\begin{aligned} Pr[y_j \ge L_j] &\ge \beta_j^1 \\ Pr[y_j \le U_j] &\ge \beta_j^2 \\ \beta_j^1 + \beta_j^2 &\ge 1 + \alpha \end{aligned}\right\} \quad j = 1, \ldots, M.$$

After substituting the expression for y_j and rearranging, we obtain

$$\left.\begin{aligned} Pr\left[L_j \sum_i^N B_{ij} n_i - \sum_i^N A_{ij} n_i \le 0 \right] &\ge \beta_j^1 \\ Pr\left[-U_j \sum_i^N B_{ij} n_i + \sum_i^n A_{ij} n_i \le 0 \right] &\ge \beta_j^2 \\ \beta_j^1 + \beta_j^2 &\ge 1 + \alpha \end{aligned}\right\} \quad j = 1, \ldots, M.$$

Next, the parameter J_{jl} is defined as

$$J_{jl} = \begin{cases} L_j, & \text{for } l = 1 \\ -U_j, & \text{for } l = 2. \end{cases}$$

This enables the recasting of both chance constraints in the same form:

$$Pr\left[J_{jl} \sum_i^N B_{ij} n_i + (-1)^l \sum_i^N A_{ij} n_i \le 0 \right] \ge \beta_j^l, \quad j = 1, \ldots, M, \quad l = 1, 2.$$

Based on the analysis presented earlier, the equivalent deterministic constraint of the "decoupled" joint constraint is

$$\left[J_{jl} \sum_i^N \mu(B_{ij}) n_i + (-1)^l \sum_i^N \mu(A_{ij}) n_i \right] + \Phi^{-1}(\beta_j^l) \left\{ \mathrm{Var}\left[J_{jl} \sum_i^N B_{ij} n_i + (-1)^l \sum_i^N A_{ij} n_i \right] \right\}^{1/2} \le 0, \quad j = 1, \ldots, M, \quad l = 1, 2$$

$$\sum_{l=1}^{2} \beta_j^l \ge 1 + \alpha, \quad j = 1, \ldots, M.$$

Calculation of the variance terms requires prior knowledge of stochastic parameter variances and covariances. In this analysis, covariances between stochastic parameters of only the same type are considered; however, the extension for other types of correlations is straightforward:

$$\mathrm{Cov}(A_{ij}, A_{i'j}), \quad i, i' = 1, \ldots, N, \quad j = 1, \ldots, M.$$

These covariances measure how discrepancies in the value of a group contribution parameter of a molecular group i from its mean value biases similar deviations for a different group i' for the same property and molecule. A positive covariance element $\mathrm{Cov}(A_{ij}, A_{i'j})$ implies that when the mean value of A_{ij} for group i over(under)estimates the true value of A_{ij} for a given molecule, then more often than not the mean value for $A_{i'j}$ will also over(under)estimate the true value of $A_{i'j}$ for the same molecule. A negative value for $\mathrm{Cov}(A_{ij}, A_{i'j})$ implies the opposite. Based on the definition of variance we have:

$$\mathrm{Var}\left[J_{jl}\sum_{i}^{N}B_{ij}n_i+(-1)^l\sum_{i}^{N}A_{ij}n_i\right]$$
$$=\sum_{i=1}^{N}\mathrm{Var}(A_{ij})n_i^2+2\sum_{i=1}^{N}\sum_{i'=i+1}^{N}n_i\mathrm{Cov}(A_{ij},A_{i'j})n_{i'}$$
$$+J_{jl}^2\left[\sum_{i=1}^{N}\mathrm{Var}(B_{ij})n_i^2+2\sum_{i=1}^{N}\sum_{i'=i+1}^{N}n_i\mathrm{Cov}(B_{ij},B_{i'j})n_{i'}\right],$$
$$j=1,\ldots,M,\quad l=1,2.$$

The resulting set of equivalent deterministic constraints is nonconvex because unlike α, which is a fixed parameter, β_j^l are variables, and thus the products of $\Phi^{-1}(\beta_j^l)$ with the square roots of the variances introduce nonconvex terms. These nonconvexities can be eliminated at the cost of introducing extra continuous variables. First, two new sets of variables are defined, $t_j^l=\Phi^{-1}(\beta_j^l)$, which implies that $\Phi(t_j^l)=\beta_j^l$, and $nt_{ij}^l=n_i\cdot t_j^l$. After incorporating the new variables and variance expressions into the deterministic equivalent representation, we obtain

$$J_{jl}\sum_{i}^{N}\mu(B_{ij})n_i+(-1)^l\sum_{i}^{N}\mu(A_{ij})n_i$$
$$+\left\{\sum_{i=1}^{N}\mathrm{Var}(A_{ij})nt_{ij}^{l^2}+2\sum_{i=1}^{N}\sum_{i'=i+1}^{N}nt_{ij}^l\mathrm{Cov}(A_{ij},A_{i'j})nt_{i'j}^l\right.$$
$$\left.+J_{jl}^2\left[\sum_{i=1}^{N}\mathrm{Var}(B_{ij})nt_{ij}^{l^2}+2\sum_{i=1}^{N}\sum_{i'=i+1}^{N}nt_{ij}^l\mathrm{Cov}(B_{ij},B_{i'j})nt_{i'j}^l\right]\right\}^{1/2}$$
$$\le 0,\qquad j=1,\ldots,M,\quad l=1,2$$
$$\sum_{l=1}^{2}\Phi\left(t_j^l\right)\ge 1+\alpha,\quad j=1,\ldots,M$$
$$nt_{ij}^l=n_i\cdot t_j^l,\quad i=1,\ldots,N,\quad j=1,\ldots,M,\quad l=1,2.$$

Kataoka (1963) showed that the square root of the variances is a convex function in nt_{ij}^l. Also, the cumulative density distributions $\Phi(t_j^l)$ are concave for $t_j^l\ge 0$, which is satisfied for $\alpha\ge 0.5$. This implies that $\Sigma_{l=1}^2\Phi(t_j^l)\ge 1+\alpha$ is a convex constraint for $\alpha\ge 0.5$.

The only remaining source of nonconvexity stems from the definition of nt_{ij}^l involving the products between n_i and t_j^l. However, nonlinear products of continuous and integer variables can be expressed equivalently with a linear set of equations at the expense of introducing extra continuous variables. Specifically, the integer variables n_i are first expressed as linear combinations of binary variables y_{ik} as follows:

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

where

$$K=\left[\frac{\log_2(n_i^U-n_i^L)}{\log_2 2}\right].$$

Next, the products $t_j^l\cdot y_{ik}$ between continuous t_j^l and binary y_{ik} variables are equivalently expressed with four linear inequality constraints (Glover, 1975),

$$t_j^l-t_j^{l,U}(1-y_{ik})\le yt_{ijk}^l\le t_j^l-t_j^{l,L}(1-y_{ik})$$
$$t_j^{l,L}y_{ik}\le yt_{ijk}^l\le t_j^{l,U}y_{ik},$$
$$i=1,\ldots,N,\quad j=1,\ldots,M,\quad k=0,\ldots,K,\quad l=1,2.$$

Here yt_{ijk}^l are additional continuous variables that account for the products between binary and continuous variables and are related with nt_{ij}^l as follows:

$$nt_{ij}^l=n_i^L t_j^l+\sum_{k=0}^{K}2^k yt_{ijk}^l,\quad i=1,\ldots,N,\quad j=1,\ldots,M,$$
$$l=1,2.$$

In this subsection, we showed how chance constraints following the mathematical formalism,

$$Pr\left[L_j\le\frac{\sum_{i}^{N}A_{ij}n_i}{\sum_{i}^{N}B_{ij}n_i}\le U_j\right]\ge\alpha,\quad j=1,\ldots,M$$

can be recast into a deterministic equivalent representation with linear binary and convex continuous part. Next, based on this analysis the deterministic equivalent representation of formulation SPM is derived.

Deterministic equivalent representation of SPM

Formulation SPM, presented earlier, identifies the type and number of molecular groups for which a scaled property-violation target s, guaranteed to be met by all properties j with probability of at least α, is minimized. To conform with the joint chance-constraint formalism discussed earlier, instead of minimizing s for a given target on α, the equivalent formulation of maximizing α for a given target s_o on s is solved. Thus, the alternative formulation of the stochastic optimization problem is

$$\max\ \alpha$$
$$\text{subject to}\quad Pr\left[s_o\ge\frac{1}{p_j^s}\left|\frac{\sum_{i=1}^{N}A_{ij}n_i}{\sum_{i=1}^{N}B_{ij}n_i}-p_j^o\right|\right]\ge\alpha,\quad j=1,\ldots,M,$$

where s_o is given and α is a deterministic variable between zero and one. After rearrangement the chance constraints are brought in the standard joint form discussed in the previous section:

$$Pr\left[p_j^o - s_o p_j^s \le \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \le p_j^o + s_o p_j^s \right] \ge \alpha, \quad j = 1, \ldots, M.$$

By setting $L_j = p_j^o - s_o p_j^s$ and $U_j = p_j^o + s_o p_j^s$, the chance-constraint formalism discussed in the previous section is recovered. Thus, the deterministic equivalent formulation for the stochastic property-matching problem (SPM) is

$$\max \quad \alpha$$

$$\text{subject to} \quad \left[J_{jl} \sum_{i}^{N} \mu(B_{ij}) n_i + (-1)^l \sum_{i}^{N} \mu(A_{ij}) n_i \right]$$

$$+ \left\{ \sum_{i=1}^{N} \mathrm{Var}(A_{ij}) n t_{ij}^{l^2} + 2 \sum_{i=1}^{N} \sum_{i'=i+1}^{N} n t_{ij}^{l} \mathrm{Cov}(A_{ij}, A_{i'j}) n t_{i'j}^{l} \right.$$

$$\left. + J_{jl}^2 \left[\sum_{i=1}^{N} \mathrm{Var}(B_{ij}) n t_{ij}^{l^2} + 2 \sum_{i=1}^{N} \sum_{i'=i+1}^{N} n t_{ij}^{l} \mathrm{Cov}(B_{ij}, B_{i'j}) n t_{i'j}^{l} \right] \right\}^{1/2}$$

$$\le 0, \qquad j = 1, \ldots, M, \quad l = 1,2$$

$$1 + \alpha - \sum_{l=1}^{2} \Phi\left(t_j^l\right) \le 0, \quad j = 1, \ldots, M$$

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

$$n t_{ij}^l = n_i^L t_j^l + \sum_{k=0}^{K} 2^k y t_{ijk}^l, \quad i = 1, \ldots, N, \quad j = 1, \ldots, M, \quad l = 1,2$$

$$t_j^l - t_j^{l,U}(1 - y_{ik}) \le y t_{ijk}^l \le t_j^l - t_j^{l,L}(1 - y_{ik})$$

$$t_j^{l,L} y_{ik} \le y t_{ijk}^l \le t_j^{l,U} y_{ik},$$

$$i = 1, \ldots, N, \quad j = 1, \ldots, M, \quad k = 0, \ldots, K, \quad l = 1,2$$

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

In this subsection, we showed how the original chance-constrained stochastic property-matching formulation (SPM) can be first transformed into a deterministic equivalent representation and then recast into a mixed-integer nonlinear programming MINLP formalism with a linear discrete and convex continuous part. The resulting MINLP formulations can be solved to global optimality with existing algorithms such as OA (Viswanathan and Grossmann, 1990) or GOS (Floudas et al., 1989). This allows the exact solution of optimal molecular design problems under property-prediction uncertainty. In the next subsection, the deterministic equivalent representation of the SPO problem is presented.

Equivalent deterministic representation of SPO

Formulation SPO, presented earlier, involves two different sets of chance constraints, each with a different mathematical structure. The first one, composed by a single constraint, models the probability α of meeting the property objective that is optimized. The second set comprises M joint chance constraints, maintaining that all lower and upper property bounds are met with probability of at least β.

After defining $np_{ij^*} = n_i \cdot p_j{}^*$ and rearranging terms, the first chance constraint can be written as

$$Pr\left[\sum_{i=1}^{N} B_{ij^*} np_{ij^*} - \sum_{i=1}^{N} A_{ij^*} n_i \le 0 \right] \ge \alpha$$

with

$$np_{ij^*} = n_i \cdot p_j{}^*, \quad i = 1, \ldots, N.$$

Based on the analysis of the previous section, the deterministic equivalent representation is

$$\sum_{i=1}^{N} \mu(B_{ij^*}) np_{ij^*} - \sum_{i=1}^{N} \mu(A_{ij^*}) n_i$$

$$+ \Phi^{-1}(\alpha) \left\{ \sum_{i=1}^{N} \mathrm{Var}(B_{ij^*}) np_{ij^*}^2 \right.$$

$$+ 2 \sum_{i=1}^{N} \sum_{i'=i+1}^{N} np_{ij^*} \mathrm{Cov}(B_{ij^*}, B_{i'j^*}) np_{i'j^*} + \sum_{i=1}^{N} \mathrm{Var}(A_{ij^*}) n_i^2$$

$$\left. + 2 \sum_{i=1}^{N} \sum_{i'=i+1}^{N} n_i \mathrm{Cov}(A_{ij^*}, A_{i'j^*}) n_{i'} \right\}^{1/2} \le 0$$

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

$$np_{ij^*} = n_i^L p_j{}^* + \sum_{k=0}^{K} 2^k y p_{ij^*k}, \quad i = 1, \ldots, N$$

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

$$p_{j^*}^L y_{ik} \le y p_{ij^*k} \le p_{j^*}^U y_{ik}, \quad i = 1, \ldots, N, \quad k = 0, \ldots, K$$

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

The second set of constraints,

$$Pr\left[p_j^L \le \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \le p_j^U \right] \ge \beta, \quad j = 1, \ldots, M,$$

has exactly the same form with the one addressed earlier after setting $L_j = p_j^L$ and $U_j = p_j^U$. Thus, its deterministic equivalent representation is identical with the one derived in that same subsection. Next, two molecular design case studies under property-prediction uncertainty are addressed to illustrate the proposed framework.

Case Study 1

The first case study involves the design of a polymer that meets constraints on density, water absorption, and glass

Table 1. Molecular Groups of the First Case Study

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

transition temperature (Derringer and Markham, 1985). The deterministic property matching (PM) and property optimization (PO) formulations of this problem were addressed in Maranas (1996). The molecular groups that are allowed to participate in the polymer repeat unit are shown in Table 1. The contribution of these molecular groups to the three properties of interest follow the empirical equations proposed by van Krevelen (1976),

$$\text{Density} \qquad \rho = \frac{\sum_{i=1}^{N} M_i n_i}{\sum_{i=1}^{N} V_i n_i} \quad (\text{g/cm}^3)$$

$$\text{Glass transition temperature} \qquad T_g = \frac{\sum_{i=1}^{N} Y_{gi} n_i}{\sum_{i=1}^{N} M_i n_i} \quad (\text{K})$$

$$\text{Water absorption} \qquad W = \frac{\sum_{i=1}^{N} 18 H_i n_i}{\sum_{i=1}^{N} M_i n_i}$$

$$(\text{g } H_2O/\text{g polymer}).$$

The mean values for the group contribution parameters H_i, M_i, Y_{gi}, and V_i for different molecular groups are the ones proposed by van Krevelen (1990) and tabulated in (Maranas, 1996). The same molecular group is allowed to participate up to three times in the polymer repeat unit, $n_i \in \{0, 1, 2, 3\}$, $i = 1, \ldots, 7$. The property targets are, respectively,

$$W^o = 0.005 \text{ (g } H_2O/\text{g polymer)},$$

$$\rho^o = 1.50 \text{ (g/cm}^3), \qquad T_g^o = 383 \text{ (K)}.$$

The property scales, W^s, ρ^s, T_g^s, are selected to be equal to the property targets W^o, ρ^o, and T_g^o, respectively. This implies that the same relative importance is assigned to all percent property violations from their target values.

The group contribution parameters are assumed to be independent random variables (covariances equal to zero), normally distributed and with mean values equal to their values reported in the literature. Their variances are chosen to reflect the relative accuracy of the group contribution methods. For instance, density estimates based on group contribution are typically more accurate than estimates of water absorption or glass transition temperature. Specifically, the variance $\text{Var}(M_i)$ of M_i is chosen to be zero because the repeat unit molecular weight is rigorously additive to the individual molecular group contributions. The variance of V_i is selected so that 99% of possible realizations of V_i are within $\pm 5\%$ from the mean value $\mu(V_i)$ (Kreyszig, 1993),

$$\text{Var}(V_i) = \left(\frac{0.05 \times \mu(V_i)}{2.58} \right)^2, \quad i = 1, \ldots, N.$$

A 10% scatter around the mean value for H_i and a 20% scatter for Y_{gi} are imposed, implying that

$$\text{Var}(H_i) = \left(\frac{0.10 \times \mu(H_i)}{2.58} \right)^2, \quad i = 1, \ldots, N$$

$$\text{Var}(Y_{gi}) = \left(\frac{0.20 \times \mu(Y_{gi})}{2.58} \right)^2, \quad i = 1, \ldots, N.$$

Note that the variance values are not rigorously estimated based on experimental data. They are arbitrarily chosen to provide reasonable estimates. A systematic procedure is described in the Appendix for rigorously obtaining not only the mean values of the group contribution parameters but also the variance–covariance matrix.

SPM formulation

The solution of the deterministic property matching problem (PM) (where property uncertainty was not considered), which was studied by Maranas (1996), yields the five best molecular designs, which are shown in Table 2 in decreasing order of optimality.

In this article, the deterministic equivalent formulation of SPM is solved using the GAMS/DICOPT interface on a RS6000 43P-133 workstation with an absolute convergence tolerance of 10^{-6}. The maximum scaled property violation target s_o ranges from 0.0163 (solution of PM) to 0.30. Because the resulting MINLP involves convex continuous and linear discrete components that are mutually separable, the GAMS/DICOPT interface is guaranteed to identify the global minimum. The results for different values of s_o are shown in Table 3.

These results imply that the higher the probability α, the larger the achievable maximum scaled property violation target is. The most promising polymer design for scaled property violations of less than about 0.065 (or α less than 0.6) is the design identified as the best for the deterministic model, $-(CH_2-(CHCl)_2)-$. However, for probabilities greater than 0.6 the best design becomes the third-best design of the deterministic model $-((CH_2)_2-(CHCl)_3)-$. This result demonstrates that property prediction uncertainty may affect the selection of the best molecular design by reversing the deterministic order of optimality for certain probability levels α. Furthermore, there is less than 18% chance of meeting the scaled property violation target 0.0163 predicted by the deterministic model. For a more likely target ($\alpha = 0.5$), a

Table 2. Five Best Molecular Designs for the PM Formulation

Alias	Repeating Unit	Violation	W	T	ρ
1-2	$-(CH_2\text{-}(CHCl)_2)-$	0.0163	0.0049	384.68	1.4889
1-3	$-(CH_2\text{-}(CHCl)_3)-$	0.0263	0.0051	393.10	1.5351
2-3	$-((CH_2)_2\text{-}(CHCl)_3)-$	0.0526	0.0047	376.95	1.4489
0-1	$-(CHCl)-$	0.1134	0.0056	412.37	1.6524
1-1	$-(CH_2\text{-}CHCl)-$	0.1169	0.0044	363.20	1.3827

Table 3. Pareto Optimum Solutions of Case Study 1

α	Max. Viol.	Repeat Unit	CPU (s)
0.1761	0.0163	$-(CH_2\text{-}(CHCl)_2)-$	32.72
0.2152	0.0200	$-(CH_2\text{-}(CHCl)_2)-$	21.31
0.4151	0.0400	$-(CH_2\text{-}(CHCl)_2)-$	21.31
0.5874	0.0600	$-(CH_2\text{-}(CHCl)_2)-$	9.27
0.7331	0.0800	$-((CH_2)_2\text{-}(CHCl)_3)-$	15.02
0.8349	0.1000	$-((CH_2)_2\text{-}(CHCl)_3)-$	8.82
0.9481	0.1400	$-((CH_2)_2\text{-}(CHCl)_3)-$	10.63
0.9876	0.1800	$-((CH_2)_2\text{-}(CHCl)_3)-$	8.58
0.9978	0.2200	$-((CH_2)_2\text{-}(CHCl)_3)-$	8.44
0.9997	0.2600	$-((CH_2)_2\text{-}(CHCl)_3)-$	8.88
0.9999	0.3000	$-((CH_2)_2\text{-}(CHCl)_3)-$	9.86

threefold increase in the value of the scaled property violation target is predicted. Finally, for $\alpha = 0.99$, the achievable scaled property violation target is approximately 0.19. This is an order of magnitude higher than the results of the deterministic model.

Figure 2 illustrates the trade-off curves between scaled property violations and probabilities for the five best molecular designs predicted by the deterministic model. Simple inspection of the trade-off curves reveals that the first three molecular designs with aliases 1-2, 1-3, 2-3 are superior over the entire probability range to those with 0-1, 1-1. Additionally, while for probability less than about 0.6, design 1-2 is superior; for higher probability values, design 2-3 becomes the optimum. This is shown more clearly in Figure 3. Note that the "crossing over" of the trade-off curves of designs 1-2 and 2-3 is not unique. For example, the trade-off curve for design 1-1 crosses over the one for design 0-1.

By definition, the trade-off curves of all molecular designs start from the solution derived by the deterministic model reflecting that *the quantitative effect of property-prediction uncertainty is to penalize the deterministic model predictions*. In fact, the higher ranked the molecular design is, the lower the probability of meeting the deterministic model suggestions appear to be. This indicates that random perturbations around a mean of the group contribution parameters have a more prominent effect on optimal designs than on "suboptimal" ones. Because the best molecular design is so "fine-tuned," any random fluctuations on the values of the group contribution parameters almost always worsen rather than improve the scaled property violation. The trade-off curves shown in Figure 2 provide a concise and systematic way for answering important questions pertinent to design. For example, if a scaled property violation of only up to 10% can be tolerated, then designs 2-3, 1-2, and 1-3 with probabilities of 0.87, 0.82, and 0.78, respectively, emerge as the most promising designs. Alternatively, if a 90% probability level of meeting the scaled property violation target is imposed, then molecular designs 2-3, 1-2, 1-3, 1-1, and 0-1 will involve scaled property violations of at most 0.11, 0.12, 0.14, 0.16 and 0.19, respectively.

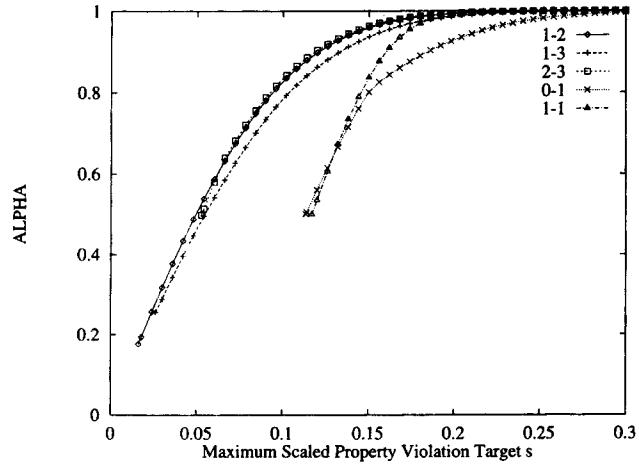

Figure 2. Trade-off curves of the five best molecular designs of SPM for the first case study.

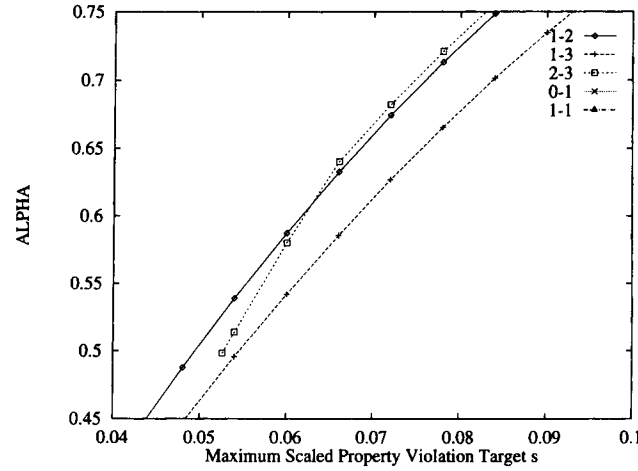

Figure 3. Magnification of the trade-off curves for the SPM problem around the crossover point in the first case study.

SPO formulation

The property optimization problem under uncertainty SPO involves the minimization of a target on water absorption subject to lower and upper bounds on glass transition temperature and density (Maranas, 1996):

$$298\ (\mathrm{K}) \le T_g \le 673\ (\mathrm{K}) \quad \text{and} \quad 1\ (\mathrm{g/cm^3}) \le \rho \le 1.5\ (\mathrm{g/cm^3})$$

The five best molecular designs in decreasing order of optimality identified in Maranas (1996) by utilizing the deterministic model PO (without considering property-prediction uncertainty) are shown in Table 4.

First, the maximum value of β is identified, for which the five best molecular designs 3-1, 2-1, 3-2, 1-1, and 2-3 satisfy the lower and upper bounds on density and glass transition

Table 4. Five Best Molecular Designs for the PO Formulation

Alias	Repeating Unit	W	T_g	ρ
3-1	$-((CH_2)_3\text{-}CHCl)-$	0.00318	310.49	1.1768
2-1	$-((CH_2)_2\text{-}CHCl)-$	0.00368	332.03	1.2531
3-2	$-((CH_2)_3\text{-}(CHCl)_2)-$	0.00401	346.04	1.3082
1-1	$-(CH_2\text{-}CHCl)-$	0.00441	363.20	1.3827
2-3	$-((CH_2)_2\text{-}(CHCl)_3)-$	0.00474	376.95	1.4488

Table 5. Maximum Values of β for the Five Best Molecular Designs

Alias	Repeating Unit	β_{max}
3-1	$-((CH_2)_3\text{-}CHCl)-$	0.7505
2-1	$-((CH_2)_2\text{-}CHCl)-$	0.9475
3-2	$-((CH_2)_3\text{-}(CHCl)_2)-$	0.9826
1-1	$-(CH_2\text{-}CHCl)-$	0.9954
2-3	$-((CH_2)_2\text{-}(CHCl)_3)-$	0.9878

temperature. For a given molecular design (n_i fixed) this requires the solution of the following convex (NLP) formulation:

$$\max \quad \beta$$

$$\text{subject to} \quad Pr\left[p_j^L \le \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i} \le p_j^U \right] \ge \beta, \quad j = 1, \ldots, M.$$

The solution of its deterministic equivalent representation yields the maximum values for probability levels β for the five best designs (see Table 5). These values indicate that molecular designs 2-1, 3-2, 1-1, and 2-3 are more likely to satisfy the lower and upper bounds on density and glass transition than design 3-1, which is the mathematically best according to the deterministic model.

Next, by fixing β to the values shown in Table 5 for each of the molecular designs 3-1, 2-1, 3-2, 1-1, and 2-3, the stochastic property optimization formulation is solved (n_i fixed) while varying the probability level α between 0.1 to 0.9999. This yields the trade-off curves between the minimum water absorption target and probability α of satisfaction of the target for the five best molecular designs (see Figure 4). Clearly, the optimality order derived by the deterministic model is maintained and no "crossover" points between trade-off curves are observed. The effect of uncertainty is not nearly as pronounced as in the property-matching problem. In fact, a scatter of only about 0.0002 of the water absorption target is observed for lower or higher values of α. In all problems studied so far, the effect of property-prediction uncertainty is much more pronounced in the property-matching problem than in the property-optimization case. The reason for this lies in the different nature of the chance constraints in the two cases. Specifically, in the SPM problem discrepancies of *all* properties from their target values are bounded from above and below. However, in the SPO problem a *single* property to be maximized is only bounded by its target from below.

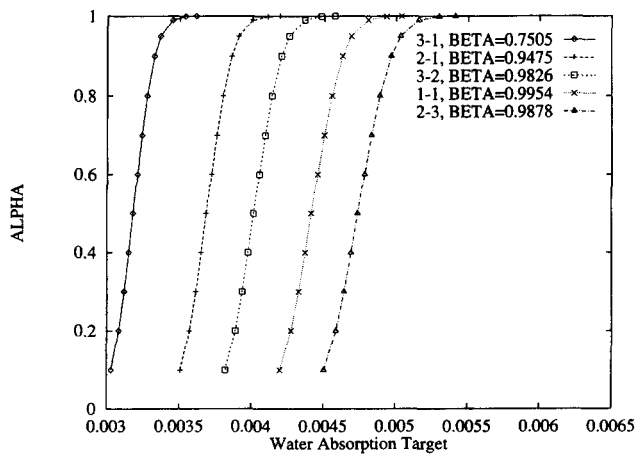

Figure 4. Trade-off curves of the five best molecular designs of SPO for the first case study.

Table 6. Property Targets for Second Case Study

Polymer Target	ρ (g/cm^3)	α (1/K)	c_p (J/gK)	K (N/m^2)
Kevlar	1.3170	$3.1338\ 10^{-4}$	1.0111	$9.6396\ 10^{9}$

Case Study 2

The second case study addresses the design of the polymer repeat unit that most closely matches given property targets on density ρ, thermal expansion coefficient α, specific heat capacity c_p, and bulk modulus K. The molecular groups composing the optimal polymer repeat unit are chosen from a set of thirty molecular groups and are tabulated in Maranas (1996) along with the utilized group contribution estimating formulas for density ρ, thermal expansion coefficient α, specific heat capacity c_p, and bulk modulus K. The five best molecular designs, according to the deterministic formulation PM, that most closely match the property targets shown in Table 6 are given in Table 7. Here n_i is the number of times the ith molecular group [see Table 13 in Maranas (1996)] participates in the polymer repeat unit.

Next, the effect of property estimation uncertainty on the prediction of the deterministic model PM is quantified. The group contribution parameters are again assumed to be independent random variables (covariances equal to zero), normally distributed, with mean values equal to those reported in the literature. The values of the variances of the group contribution parameters are again selected to reflect the relative accuracy of the estimating formulas:

$$\mathrm{Var}(M_i) = \left(\frac{0.00 \times \mu(M_i)}{2.58} \right)^2, \quad i = 1, \ldots, N$$

(molecular weight)

$$\mathrm{Var}(V_{ai}) = \left(\frac{0.05 \times \mu(V_{ai})}{2.58} \right)^2, \quad i = 1, \ldots, N$$

(molar volume)

Table 7. Five Best Designs in Second Case Study

Alias	Max. Viol.	Molecular Groups
1	0.0000	$n_7 = 1$, $n_{11} = 1$, $n_{12} = 1$, $n_{15} = 2$ (Kevlar)
2	0.0022	$n_3 = 1$, $n_6 = 1$, $n_7 = 1$, $n_{14} = 1$, $n_{15} = 2$, $n_{29} = 1$
3	0.0032	$n_3 = 1$, $n_6 = 1$, $n_{14} = 1$, $n_{15} = 1$, $n_{18} = 1$, $n_{28} = 1$, $n_{29} = 1$
4	0.0035	$n_{10} = 1$, $n_{11} = 3$, $n_{16} = 1$, $n_{18} = 1$, $n_{20} = 1$
5	0.0048	$n_{12} = 1$, $n_{13} = 1$, $n_{14} = 1$, $n_{16} = 2$, $n_{18} = 1$, $n_{29} = 1$

$$\mathrm{Var}(V_{wi}) = \left(\frac{0.05 \times \mu(V_{wi})}{2.58}\right)^2, \quad i = 1, \ldots, N$$

(VDW molar volume)

$$\mathrm{Var}(C_{pi}) = \left(\frac{0.02 \times \mu(C_{pi})}{2.58}\right)^2, \quad i = 1, \ldots, N$$

(molar heat capacity)

$$\mathrm{Var}(U_{Ri}) = \left(\frac{0.10 \times \mu(U_{Ri})}{2.58}\right)^2, \quad i = 1, \ldots, N$$

(bulk modulus)

Next, the trade-off curves between maximum property violations and probabilities of meeting them are constructed (see Figure 5) for each one of the five best molecular designs. The trade-off curves shown in Figure 5 indicate that the optimality order obtained from the deterministic formulation PM is almost completely reversed when group contribution uncertainty is considered. Designs 3, 4 and 5 (with 3 slightly better) clearly emerge as the most promising designs, with design 2 following fourth, and design 1 being the last. Likely values (for $\alpha = 0.5$) for the scaled property violation are in the range of 0.014 to 0.022, while conservative estimates ($\alpha = 0.9$) are between 0.035 and 0.055. As in the first case study, it is observed that property prediction uncertainty has a significant effect on which molecular design is the most promising and what is an achievable design target.

A closer inspection of the low probability region (see Figure 6) reveals that no discernible crossover between trade-off curves exists. This implies that the pareto optimum curve (max over all trade-off curves) is *discontinuous*. This partitions the scaled property-violation range into three subintervals. For scaled property violations of less than 0.0022, molecular design 1 is the only alternative. Between 0.0022 and 0.0032, molecular design 2 is superior to design 1. Finally, for values greater than 0.0033 molecular designs 3, 4, and 5 outperform designs 1 and 2. While the deterministic predictions for the scaled property violations are very small (less than 0.005), the probabilities of meeting these deterministic predictions are also small (less than 20%). This is in agreement with the observation made in the first case study that the lower the value the deterministic model predicts for the scaled property violation, the more sensitive it is to random perturbations of the group contribution parameter values.

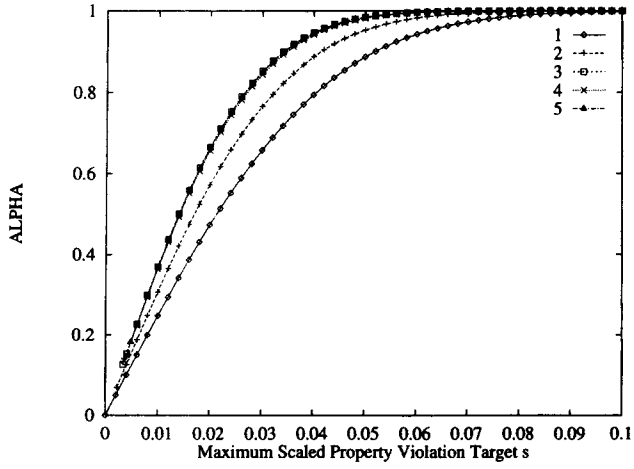

Figure 5. Trade-off curves of the five best molecular designs for the second case study.

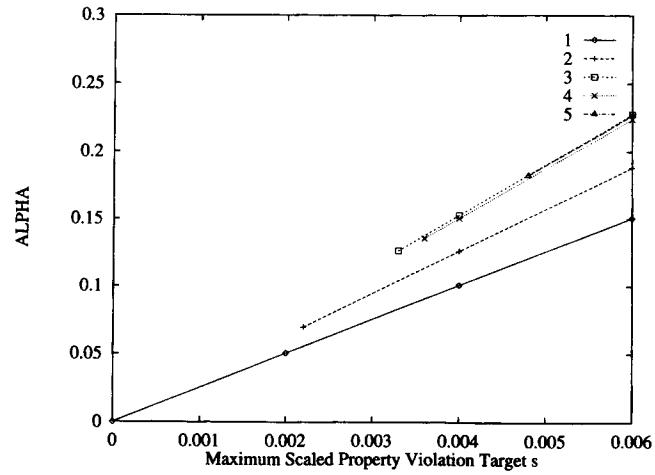

Figure 6. Magnification of the trade-off curves of the five best molecular designs for the second case study.

Summary and Conclusions

In this article, a systematic framework was introduced that for the first time quantitatively assesses the effect of property-prediction uncertainty in optimal molecular design. Property–structure relations with specific nonlinear functionalities were established based on group contribution methods compiled by van Krevelen (1990). Property prediction uncertainty was explicitly quantified by utilizing multivariate probability density distributions to model the likelihood of different realizations of the group contribution parameters employed. Two special instances of the general stochastic optimal molecular design problem were addressed in detail.

1. The stochastic property matching problem (SPM) that identifies the type and the number of molecular groups for which a scaled property violation target s, guaranteed to be met by all properties j with probability of at least α, is minimized.

2. The stochastic property optimization problem (SPO) that identifies the type and number of molecular groups for which a property target p_{j^*} for property j^*, together with probability α, is maximized, while lower and upper bounds for the rest of the properties are satisfied with probability of at least β.

Assuming that the probability density distribution modeling the stochastic nature of the group contribution parameters is stable (e.g., normal, Poisson, chi-square, and binomial), an approach based on the deterministic equivalent chance-constraint representation was introduced for transforming the original nonlinear stochastic formulation into a deterministic MINLP problem with linear binary and convex continuous parts. These MINLP formulations can be solved efficiently to global optimality (for $\alpha \geq 0.5$) with existing algorithms such as OA (Viswanathan and Grossmann, 1990) or GOS (Floudas et al., 1989). The proposed theoretical and computational

framework allowed the exact formulation and solution of optimal molecular design problems under property-prediction uncertainty. Note that for $\alpha < 0.5$ formulations SPM and SPO become nonconvex. Although multiple local minima may exist in theory, this was not observed in practice after employing multiple starting points. Of course, this is only indicative and not a rigorous proof for the existence of a unique minimum for $\alpha < 0.5$.

Two different case studies were addressed in this article. Trade-off curves between performance objectives, probabilities of meeting these objectives, and chances of satisfying design specifications were constructed establishing concise and systematic ways of selecting molecular designs in the presence of property-prediction uncertainty. For the property-matching problem (SPM), it was observed that property-prediction uncertainty had a dramatic effect on the selection of the best molecular design. This was manifested by the frequent reversal of the order of optimality obtained by the deterministic model, depending on the selection of the probability level α. This quantitatively demonstrated the intuitive expectation that the *answer to the question of what is the best molecular design depends on how frequently design target violations can be tolerated.* In fact, the best molecular designs for optimistic ($\alpha = 0.1$), most likely ($\alpha = 0.5$), and conservative ($\alpha = 0.9$) scenarios were usually different. In all cases, the trade-off curves of all molecular designs started from the solution derived by the deterministic model reflecting that *the quantitative effect of property prediction uncertainty in SPM was to penalize the deterministic model predictions.* The higher ranked the molecular design was (according to the deterministic model PM), the lower the probability of meeting the deterministic model predictions in the SPM formulation. This indicates that *random perturbations around the mean value of the group contribution parameters have a more prominent effect on optimal rather than suboptimal designs.* In the property-optimization formulation SPO, the effect of uncertainty was not even nearly as pronounced as in the SPM model. Computational results indicated that it was extremely unlikely to have the results of the deterministic model changed when property-prediction imprecision was superimposed. Of course, these conclusions are valid only for the addressed example and for the selected variance values.

While the variance values for the examples addressed in this article were selected somewhat arbitrarily, a systematic approach was proposed (see the Appendix) for calculating not only mean values but also the full variance–covariance matrix of group contribution parameters. Application of this procedure to an extensive set of polymer properties is currently under way. Although the proposed theoretical and computational framework was tailored only to polymer design, the same underlying mathematical features are present, to some extent, in (1) other molecular design problems, such as optimal design of agrochemicals, refrigerants, and solvents, and (2) more complex design instances involving process considerations and mixtures of compounds. Work is currently underway on extending the proposed framework in these directions.

Acknowledgments

Financial support by Du Pont's Educational Aid Grant 1996 is gratefully acknowledged.

Literature Cited

Abramowitz, M., and I. E. Stegun, *Handbook of Mathematical Functions*, 10th ed., Nat. Bur. Stand., Washington, DC (1972).

Allen, F. M., R. N. Braswell, and P. V. Rao, "Distribution-Free Approximations for Chance Constraints," *Oper. Res.*, **22**, 610 (1974).

Bolis, G., L. DiPace, and F. Fabrocini, "A Machine Learning Tool for Computer Aided Molecular Design," *Proc. Int. Conf. on Tools for Artificial Intelligence*, San Jose, CA (1991).

Boyd, D. B., *Successes of Computer-Assisted Molecular Design*, VCH Publishers, New York (1990).

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

Cattell, R. B., *Factor Analysis*, Harper, New York (1952).

Charnes, A., and W. W. Cooper, "Chance-constrained Programming," *Manage. Sci.*, **6**, 73 (1959).

Charnes, A., and W. W. Cooper, "Chance-constraints and Normal Deviates," *J. Amer. Stat. Assoc.*, **55**, 134 (1960).

Charnes, A., and W. W. Cooper, "Deterministic Equivalents for Optimizing and Satisfying under Chance Constraints," *Oper. Res.*, **11**, 18 (1963).

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

Constantinou, L., and R. Gani, "Group-contribution Method for Estimating Properties of Pure Compounds," *AIChE J.*, **40**(10), 1697 (1994).

Constantinou, L. K., R. Bagherpour, R. Gani, J. A. Klein, and R. C. Glen, "Computer-aided Product Design: Problem Formulations, Methodology, and Applications," *Comput. Chem. Eng.*, **20**(6), 685 (1996).

Derringer, G. C., and R. L. Markham, "A Computer-based Methodology for Matching Polymer Structures with Required Properties," *J. Appl. Polym. Sci.*, **30**, 4609 (1985).

Diwekar, U. M., and E. S. Rubin, "Stochastic Modeling of Chemical Processes," *Comput. Chem. Eng.*, **15**(2), 105 (1991).

Duvedi, A. P., and L. E. K. Achenie, "Designing Environmentally Safe Refrigerants Using Mathematical Programming," *Chem. Eng. Sci.*, **51**(15), 3727 (1996).

Floudas, C. A., A. Aggarwal, and A. R. Ciric, "Global Optimum Search for Nonconvex NLP and MINLP Problems," *Comput. Chem. Eng.*, **13**(10), 1117 (1989).

Floudas, C. A., and I. E. Grossmann, "Active Constraint Strategy for Flexibility Analysis in Chemical Processes," *Comput. Chem. Eng.*, **11**(6), 675 (1987).

Franklin, J. L., "Prediction of Heat and Free Energies of Organic Compounds," *Ind. Eng. Chem.*, **41**(51), 1070 (1949).

Fredenslund, A., R. L. Jones, and J. M. Prausnitz, "Group-contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures," *AIChE J.*, **21**(6), 1086 (1975).

Friedman, Y., and G. V. Reklaitis, "Flexible Solutions to Linear Problems under Uncertainty: Inequality Constraints," *AIChE J.*, **21**(1), 77 (1975).

Gani, R., and E. A. Brignole, "Molecular Design of Solvents for Liquid Extraction Based on UNIFAC," *Fluid Phase Equilibria*, **13**, 331 (1983).

Gani, R., and A. Fredenslund, "Computer-Aided Molecular and Mixture Design with Specified Property Constraints," *Fluid Phase Equilibria*, **82**, 39 (1993).

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

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

Gordeeva, E. V., M. S. Molcharova, and N. S. Zefirov, "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.*, **3**, 389 (1990).

Grossmann, I. E., and W. H. Sargent, "Optimum Design of Chemical Plants with Uncertain Parameters," *AIChE J.*, **24**, 1021 (1978).

Halemane, K. P., and I. E. Grossmann, "Optimal Process Design under Uncertainty," *AIChE J.*, **29**(3), 425 (1983).

Horvath, A. L., "Estimate Properties of Organic Compounds," *Chem. Eng.*, **95**(11), 155 (1988).
Horvath, A. L., *Molecular Design*, Elsevier, Amsterdam (1992).
Ierapetritou, M. G., and E. N. Pistikopoulos, "Simultaneous Incorporation of Flexibility and Economic Risk in Operational Planning Under Uncertainty," *Comput. Chem. Eng.*, **18**(3), 163 (1994).
Joback, K. G., "Designing Molecules Possessing Desired Physical Properties," PhD Thesis, MIT, Cambridge, MA (1989).
Joback, K. G., and R. C. Reid, "Estimation of Pure-component Properties from Group Contributions," *Chem. Eng. Commun.*, **57**, 233 (1987).
Joback, K. G., and G. Stephanopoulos, "Designing Molecules Possessing Desired Physical Property Values," *FOCAPD*, Snowmass, CO, p. 363 (1989).
Kataoka, S., "A Stochastic Programming Model," *Econometrica*, **31**, 181 (1963).
Kier, L. B., and L. H. Hall, *Molecular Connectivity in Structure-Activity Analysis*, Wiley, New York (1986).
Kier, L. B., H. H. Lowell, and J. F. Frazer, "Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts," *J. Chem. Inf. Comput. Sci.*, **33**, 142 (1993).
Kowalski, B. R., and C. F. Bender, "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data," *J. Amer. Chem. Soc.*, **94**(16), 5632 (1972).
Kreyszig, E., *Advanced Engineering Mathematics*, 7th ed., Wiley, New York (1993).
Maranas, C. D., "Optimal Computer-aided Molecular Design: A Polymer Design Case Study," *Ind. Chem. Eng. Res.*, **35**, 3403 (1996).
Mavrovouniotis, M. L., "Estimation of Properties from Conjugate Forms of Molecular Structures: The ABC Approach," *Ind. Eng. Chem. Res.*, **29**. 1943 (1990).
Mavrovouniotis, M. L., "Product and Process Design with Molecular-level Knowledge," *Proc. Int. Conf. on Intelligent Systems in Process Engineering*, Snowmass, CO, Vol. 92, J. F. Davis, G. Stephanopoulos, and V. Venkatasubramanian, eds., published by AIChE Symp. Series No. 312, p. 133 (1996).
Nagasaka, K., H. Wada, H. Yoshimitsu, H. Yasuda, and T. Yamanouchi, "Expert System for Polymer Design," *AIChE Meeting*, Chicago (1990).
Naser, S. F., and R. L. Fournier, "A System for the Design of an Optimum Liquid-Liquid Extractant Molecule," *Comput. Chem. Eng.*, **15**(6), 397 (1991).
Nielsen, B., and R. Gani, "Computer-aided Molecular Design by Group Contribution," *Proc. Eur. Symp. on Computer Applications in Chemical Engineering*, The Hague, The Netherlands (1990).
Odele, O., and S. Macchietto, "Computer Aided Molecular Design: A Novel Method for Optimal Solvent Selection," *Fluid Phase Equilibria*, **82**, 47 (1993).
Odele, O., S. Macchietto, and O. Omatsone, "Design of Optimal Solvents for Liquid-Liquid Extraction and Gas Absorption Processes," *Trans. Ind. Chem. Eng.*, **68**, 429 (1990).
Pai, C. D., and R. R. Hughes, "Strategies for Formulating and Solving Two-stage Problems for Process Design under Uncertainty," *Comput. Chem. Eng.*, **11**, 695 (1987).
Pistikopoulos, E. N., and I. E. Grossmann, "Optimal Retrofit Design for Improving Process Flexibility in Linear Systems," *Comput. Chem. Eng.*, **12**, 719 (1988).
Pistikopoulos, E. N., and T. A. Mazzuchi, "A Novel Flexibility Analysis Approach for Processes with Stochastic Parameters," *Comput. Chem. Eng.*, **14**(9), 991 (1990).
Pretel, E. J., P. A. Lopez, S. B. Bottini, and E. A. Brignole, "Computer-aided Molecular Design of Solvents for Separation Processes," *AIChE J.*, **40**(8), 1349 (1994).
Reed, M. E., and W. B. Whitting, "Sensitivity and Uncertainty of Process Designs to Thermodynamic Model Parameters: A Monte Carlo Approach," *Chem. Eng. Commun.*, **124**, 39 (1993).
Snedecor, G. W., and W. G. Cochran, *Statistical Methods*, 8th ed., Iowa State Univ. Press, Ames (1989).
Stephanopoulos, G., and D. W. Townsend, "Synthesis in Process Development," *Chem. Eng. Res. Dev.*, **64**, 160 (1986).
Straub, D. A., and I. E. Grossmann, "Integrated Stochastic Metric of Flexibility for Systems with Discrete State and Continuous Parameter Uncertainties," *Comput. Chem. Eng.*, **14**, 967 (1990).
Straub, D. A., and I. E. Grossmann, "Evaluation and Optimization of Stochastic Flexibility in Multiproduct Batch Plants," *Comput. Chem. Eng.*, **16**(2), 69 (1992).
Swaney, R. E., and I. E. Grossmann, "An Index for Operational Flexibility in Chemical Process Design. Part I. Formulation and Theory," *AIChE J.*, **31**, 621 (1985).
Vaidyanathan, R., and M. El-Halwagi, "Computer-aided Design of High Performance Polymers," *J. Elastom. Plasti.*, **26**(3), 277 (1994).
Vaidyanathan, R., and M. El-Halwagi, "Computer Aided Synthesis of Polymers and Blends with Target Properties," *Ind. Eng. Chem. Res.*, **35**(2), 627 (1996).
Vajda, S., "Stochastic Programming," in *Integer Nonlinear Program*, J. Abadie, ed., p. 321 (1970).
van Krevelen, D. W., *Properties of Polymers*, 2nd ed., Elsevier, Amsterdam (1976).
van Krevelen, D. W., *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*, 3rd ed., Elsevier, Amsterdam (1990).
Venkatasubramanian, V., K. Chan, and J. M. Caruthers, "Computer-aided Molecular Design using Genetic Algorithms," *Comput. Chem. Eng.*, **18**(9), 833 (1994a).
Venkatasubramanian, V., K. Chan, and J. M. Caruthers, "On the Performance of Genetic Search for Large-Scale Molecular Design," *PSE '94*, Korea, p. 1001 (1994b).
Venkatasubramanian, V., K. Chan, and J. M. Carruthers, "Evolutionary Design of Molecules with Desired Properties using the Genetic Algorithm," *J. Chem. Inf. Comput. Sci.*, **35**, 188 (1995).
Viswanathan, J., and I. E. Grossmann, "A Combined Penalty Function and Outer-Approximation Method for MINLP Optimization," *Comput. Chem. Eng.*, **14**, 769 (1990).
Watanabe, T., and H. Ellis, "A Joint Chance-constrained Programming Model with Row Dependence," *Eur. J. Oper. Res.*, **77**, 325 (1994).

Appendix: Estimation of Means and Variances of the Group Contribution Parameters

The means and variance–covariances of the group contribution parameters can be more efficiently found by treating separately the numerators and denominators of the adopted property prediction functionality. This simplifies the analysis by maintaining the linearity of the estimation models without sacrificing accuracy.

Let $k = 1, \ldots, K$ denote a set of molecular compounds (or repeat units) and $\hat{p}_k$ the experimental measurement of property p for compound k. Assuming an additive linear group contribution relation between the number of times n_{ik} molecular group i participates in compound k, we have

$$p_k = \sum_{i=1}^{N} a_i n_{ik}, \quad k = 1, \ldots, K,$$

where p_k is the group contribution estimate of property p, and a_i are the group contribution parameters.

(Multi)linear regression is based on the assumption that for each specific X there is a normal distribution of Y that is (1) independent of X, (2) involves a mean that depends linearly on X, and (3) has the same variance from which realizations of Y are drawn at random. Multilinear regression can be utilized to identify unbiased estimators (means) and sample estimators of the variance–covariance matrix of the vector of uncertain group contribution parameters $\boldsymbol{a} = [a_1, \ldots, a_n]^T$. The minimization of the sum of the squares of the differences between the experimentally measured $\hat{p}_k$ and estimated values p_k

$$\min \sum_{k=1}^{K} (\hat{p}_k - p_k)^2 = \min \sum_{k=1}^{K} \left(\hat{p}_k - \sum_{i=1}^{N} a_i n_{ik} \right)^2$$

yields the unbiased estimators for the values of the group contribution parameters (Snedecor and Cochran, 1989)

$$\mu(\boldsymbol{a}) = (\boldsymbol{N}^T\boldsymbol{N})^{-1}\boldsymbol{N}^T\boldsymbol{p},$$

where $\boldsymbol{a}$ is the $(N \times 1)$ vector of the group contribution parameters,

$$\boldsymbol{a} = [a_1, a_2, \ldots, a_N]^T,$$

$\boldsymbol{N}$ is the $(K \times N)$ matrix whose elements are the integers n_{ik} counting how many times molecular group i participates in compound k, and $\boldsymbol{p}$ is the $(K \times 1)$ vector of the experimental measurements

$$\boldsymbol{p} = [p_1, p_2, \ldots, p_K]^T.$$

Based on the analysis described in detail in Snedecor and Cochran (1989), the $(N \times N)$ variance–covariance matrix of the group contribution parameter vector $\boldsymbol{a}$ is given by

$$\mathrm{Var}(\boldsymbol{a}) = (\boldsymbol{N}^T\boldsymbol{N})^{-1}s^2_{p\cdot N},$$

where

$$s^2_{p\cdot N} = \frac{1}{K-N}\sum_{k=1}^{K}(\boldsymbol{p}-\boldsymbol{N}\boldsymbol{a})^T(\boldsymbol{p}-\boldsymbol{N}\boldsymbol{a})$$

is the unbiased estimator of the variance of the experimental property values.

Alternatively, using indexed equations, the mean and the variance terms can be expressed as the solution of the following system of linear equations:

$$\sum_{i'=1}^{N}\left(\sum_{k=1}^{K} n_{ik}n_{i'k}\right)\mu(a_{i'}) = \sum_{k=1}^{K} n_{i_k}p_k, \quad i = 1, \ldots, N$$

$$\sum_{i'=1}^{N}\left(\sum_{k=1}^{K} n_{ik}n_{i'k}\right)\mathrm{Cov}(a'_i, a_{i''}) = s^2_{p\cdot N}\delta_{ii''}$$

where

$$s^2_{p\cdot N} = \frac{1}{K-N}\sum_{k=1}^{K}\left(p_k - \sum_{i=1}^{N} a_i n_{ik}\right)^2$$

$\delta_{ii''}$ = Kronecker's delta

Note that the identification of the vector of means and the variance–covariance matrix requires only the solution of linear systems of equalities, thus, very large volumes of data can be processed efficiently.

Manuscript received July 5, 1996, and revision received Dec. 11, 1996.