# Novel Sampling Approach to Optimal Molecular Design Under Uncertainty

**Manish C. Tayal**
Dept. of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213

**Urmila M. Diwekar**
Dept. of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213

*For a reliable optimal molecular design, imprecision associated with property-prediction models cannot be neglected. This study presents a novel sampling approach to stochastic optimization, incorporating property-prediction uncertainty effects in a robust, generalized optimization framework. Detailed uncertainty analysis addresses, through nine case studies, various issues in computer-aided molecular design under uncertainty. Results indicate that property-prediction uncertainty can significantly impact the optimal molecular designs. Additional complex cases with nonlinear or black-box models, nonlinear objective function and constraints, and nonstable distribution for model parameter uncertainty representation highlight the flexibility and versatility of this approach. Sensitivity analysis of uncertainties in the model parameters was also made possible in this generalized framework. Uncertainties, the focus of any future research, were identified through this critical model. Increased computational efficiency of this approach and wider applicability to solve problems involving various kinds of objective functions and constraints, and different forms of uncertainties, is illustrated in the context of polymer design case studies.*

## Introduction

The search for new molecules possessing the desired physical, chemical, biological, and health properties is an important and ongoing process in chemical and pharmaceutical industries. Such a search encompasses problems of designing polymers, refrigerants, solvents, composites and blends, drugs, agricultural chemicals, paints, varnishes, and perfumes, among others. It has a major role to play in the overall economics of these industries. Further stringent environmental regulations day by day, renders several existing chemicals useless, and thus creates an enormous need of replacing these chemicals with other promising environmentally friendly molecules with similar properties. Such a search can elicit new, sometimes unexpected, counterintuitive molecules of enormous superiority never comprehended before. It has many industrial applications and can have immense impact on the overall performance of the industry. Some examples include designing environmentally friendly aircraft deicing fluids, replacing existing solvents with environmentally benign solvents, and designing polymers for use as integrated circuit (IC) encapsulants. Computer-aided molecular design (CAMD) techniques are being increasingly employed in these applications. Given a property-prediction model, CAMD solves the inverse problem of searching for new and promising molecules satisfying the desired properties. It forecasts the most promising molecules, which possess desired properties, from a plethora of possible designs that easily run into billions even for a smaller-sized design molecule. CAMD can reduce the search time drastically by pruning off futile experiments. The idea is to make most of the mistakes on paper and learn from them before even performing any experiments, so as to maximize the throughput of the experimental work and of any future research efforts that follow. Furthermore, useful insights gained through these efforts can redirect experiments to unknown areas and fuel the chemist's intuition. The ever increasing computational power of modern computers and their low cost, coupled with a better under-

Correspondence concerning this article should be addressed to U. M. Diwekar.

standing of the molecular behavior through effective modeling, has accelerated the development of computer-aided design methodology and tools for challenging optimal molecular design problems.

For any CAMD approach to be successful, widely applicable and reliable property-prediction models are required. In the last two decades significant advances have been made in the improvement of existing models as well as in the development of the new models. These models fall into various categories: empirical, semiempirical, theoretical, and hybrid. These models are of varying complexity, and several of them are limited by their applicability to specific cases. For the case of polymers, however, van Krevelen presented an extensive compilation of structure−property relations using group additive parameters for the thermal, mechanical, optical, electrical, and rheological properties of polymers (van Krevelen, 1976), also called the group contribution methods (GCM). Even though there are several types of case-specific models, GCM models have still remained widely used due to their simplicity, versatility, and wider applicability. These models are essentially empirical in nature and the various group contribution parameters are determined from regression of the experimental data. With such simple GCM models, property computations are very rapid, though at the expense of prediction accuracy. An error of 5−10% or even more between the GCM property predictions and experimental values is quite common. On the other side of the spectrum, we have more sophisticated models based on molecular-modeling techniques, like statistical mechanics models. Such models are more accurate, but are fairly complex and computationally intensive. They are very promising for a one-time accurate property prediction for a given molecule, but are not deemed suitable for the challenging inverse problem of CAMD, due to their huge computational costs and limited applicability to specific cases. Other case-specific models, using approaches from connectivity, pattern-recognition, equation-oriented, and topological indices have also been reported. These models are essentially a trade-off between model development effort, computational time, property-prediction accuracy, and wider applicability. A general overview of the existing models can be found in Joback and Stephanopoulos (1995).

The always challenging problem of optimal molecular design has been approached in several ways in the literature. Some initial approaches include generate-and-test, exhaustive, and enumeration based techniques (Joback and Stephanopoulos, 1995; Joback, 1989; Friedler et al., 1998). These approaches all generate feasible candidate molecules that satisfy given property constraints, and so are very limited by the size of the problem. Another approach involves searching an existing property database for molecules with properties in a desired range. Their efficacy is, however, bounded by the size of the database and the efficiency of the engine. Further database search is limited to known molecules, for which property data are available. Industries often face three possible scenarios for property requirements: (1) pure components, for which property values are known experimentally and are tabulated in the literature and commercial databases; (2) pure components, for which property values are unknown and so models like GCM are used for predictions; and (3) a mixture of components for which property values are unknown. A database search can be helpful for scenario (1) only, though again uncertainties due to experimental and statistical errors still exist and simple search engines could not handle such uncertainties. Thus searching for an optimal molecule in a search space consisting of both *known* and *unknown* molecules, the true essence of an optimal molecular design, is not possible with the database search. Also the sizes of commercial databases are much limited compared to the enormous size of the unexplored territory comprising the unknown molecules.

The problem of "optimal" molecular design was approached in the true sense by knowledge-based strategies, heuristics-based methods, graph reconstruction methods, multistage approaches, an approach from artificial intelligence using genetic algorithms, and a mathematical programming approach. A thorough review of these techniques was given in Mavrovouniotis (1996). Some of these techniques are applicable and have been demonstrated in the area of polymer design (Venkatasubramanian et al., 1994; Vaidyanathan and El-Halwagi, 1996; Maranas, 1996), which is the primary focus of this study. Venkatasubramanian et al. (1994) developed a novel framework using newly developed genetic operators in an evolutionary approach of genetic algorithms for CAMD. Vaidyanathan and El-Halwagi (1996) have used a mathematical programming algorithm for design of both addition and condensation polymers. Maranas (1996), however, used a mathematical programming approach to demonstrate a linear reformulation for some specific nonlinear structure−property relations, which resulted in an MILP formulation. All these varied approaches have their own importance and appeal, but simultaneously also suffer from some crucial drawbacks. For any real-life molecular design problem, these methodologies are limited in their applicability due to one or more of the following: the combinatorial complexity of the molecular-design problem; difficulties dealing with nonlinear structure−property relations; nonlinear search spaces with discontinuities; the possibility of getting trapped in several local minimum traps; problems in incorporating nonlinear objective function and constraints; and difficulty in design knowledge acquisition.

However the most crucial drawback, common in all these approaches, is that they all ignore property-prediction model uncertainties. Reliability of any CAMD initiative will, without considering prediction model uncertainties, always be questioned. This hinders its full-fledged use in large-scale real-life design problems of high impact. Given that the property-prediction model uncertainty cannot be immediately avoided, there is an obvious need to develop a generalized framework incorporating uncertainty issues in an optimization framework for molecular design. Such a framework can not only help quantify uncertainties but also aid decision making in the face of property-prediction uncertainties. Recently a systematic method to quantitatively assess the effect of property-prediction uncertainty on optimal molecular design was reported (Maranas, 1997). This was a very important step in uncertainty incorporation in optimal molecular design. In this study, the original stochastic constraint obtained in a chance-constrained framework was transformed into deterministic equivalent expressions. This results in an MINLP formulation. Further, such a transformation is only possible for stable distribution functions of uncertain parameters (Billingsley, 1995; Breiman, 1968). This strongly limits the choice to stable distributions like normal distribution. Additionally, the convexity of deterministic equivalent expressions requires that the deterministic variables appear linearly in the constraint (Maranas, 1997). However, complex property-prediction models with additional model uncertainties of various types, like lognormal distributions, complicate the problem further and render the MINLP formulation inappropriate.

Owing to the various limitations of approaches to CAMD, with and without uncertainty considerations, there is a need to develop a robust, very generalized, and computationally efficient framework. Such a framework should not be limited by the choice of (1) objective function and constraints, (2) simplified models, and (3) stable distribution for parameter uncertainty representation. In this article, we propose such a generalized approach to optimal molecular design under uncertainty. This approach is based on the robust stochastic annealing-nonlinear programming framework coupled with a novel sampling technique used to represent the model parameter uncertainties. The flexibility of this approach allows for incorporating nonlinear and black-box models, nonlinear objective function and constraints, and model uncertainties of various kinds, including nonstable distributions. It can effectively handle search-space discontinuities and avoid local minima. Sensitivity analysis of the uncertain parameters was also made possible in this generalized framework.

Key questions that are addressed in this study are:

1. How confident can we be in the GCM models, given model parameter uncertainties?

2. How drastic could be the impact of the model uncertainties and what is the risk associated with ignoring such uncertainties in any industrially reliable CAMD?

3. Which are the most sensitive model uncertainties that future research can target to reduce to achieve an even more reliable and robust molecular design?

4. How could chemist intuition and other heuristic knowledge, gathered from time to time, be fed back into the CAMD framework?

5. Can pattern recognition and trend searching in CAMD through data visualization be used to acquire design knowledge and rules?

This article focuses on the problem of CAMD under uncertainty applied to optimal polymer design for the desired properties. Strong emphasis has been on the uncertainty analysis within the optimization framework. The first section provides a mathematical description of the optimal molecular design problem. It further explains crucial aspects in problem formulation, which include objective function formulation and GCM model parameter uncertainty representation. The next section describes the proposed methodology of optimization under uncertainty using novel sampling technique. It further draws a parallel of this approach to the chance constrained optimization approach. Finally, flexibility and robustness of the proposed methodology is demonstrated by applying it to polymer design case studies. The results from nine cases are included that address key issues in CAMD under uncertainty: (1) parameter uncertainty representation; (2) objective function formulation; (3) sensitivity analysis of uncertain parameters; (4) combining heuristics with stochastic optimization; and (5) choice of target properties and its effect on the optimal molecular design. The results from using this approach also expose the possibility of knowledge acquisition through trend searching. We conclude with a discussion of fruitful insights gained through the application of the proposed framework.

## Computer-Aided Molecular Design for Polymers

CAMD refers to the general methods for expediting the molecular design by forecasting promising molecular designs. Thus various methods, such as database search, enumeration techniques, knowledge-based techniques, among others, come under the common terminology of CAMD. However, optimal molecular design (OMD) refers to a particular subset of methods that formulate an optimal design problem for optimizing a set of desired target properties or for some bounds on the target properties. This optimization is then done over an exhaustive molecular-design search space. OMD then aims at screening out promising candidate molecules from such a search space, which may include both *known* and *unknown* molecules.

The problems in OMD often fall into two types of design categories. In *property matching* (PM) a molecule is designed to meet several property targets simultaneously. Sometimes, however, we want to maximize/minimize some key property while maintaining other properties within some tolerable bounds, which falls within the *property optimization* (PO) framework (Maranas, 1996). These formulations are presented in the subsection below. Furthermore, when property-prediction uncertainties are considered in a stochastic framework, these become *stochastic property matching* (SPM) and *stochastic property optimization* (SPO) formulations, respectively (Maranas, 1997). These stochastic formulations are presented later, in the next section. In polymer engineering it is often more desirable to identify polymers that best meet a number of property constraints simultaneously. Thus PM/SPM is of more direct relevance in polymer design, and so is the focus in this article. However, flexibility of the proposed stochastic approach makes it equally applicable and effective in other problem formulations, such as SPO.

### *Mathematical formulation for CAMD*

The problem of property matching (PM) for optimal polymer design can be formulated as follows:

$$
\begin{aligned}
\text{Optimize:}\quad & F_{\text{det}}(\boldsymbol{P}, \boldsymbol{P}_o) \\
& \boldsymbol{P} = (P_1, \ldots, P_j, \ldots, P_m) \\
& \boldsymbol{P}_o = (P_{1o}, \ldots, P_{jo}, \ldots, P_{mo}) \\
& P_j = P_j(\boldsymbol{N}, A_{1j}, \ldots, A_{Nj}, B_{1j}, \ldots, B_{Nj}), \\
& \qquad j = 1, \ldots, m \qquad\qquad \text{(GCM Model)} \\
& \boldsymbol{N} = (n_1, \ldots, n_i, \ldots, n_N) \qquad \text{(Polymer Molecule)} \\
& n_i \in \{n_i^L, n_i^L + 1, \ldots, n_i^U\}, \qquad i = 1, \ldots, N \\
\text{Subject to:}\quad & f = \sum_{i=1}^{N} (\upsilon_i - 2) n_i + 2 \\
& \qquad\qquad \text{(Structural Feasibility Constraint)} \\
& n_{\min} \le \sum_{i=1}^{N} n_i \le n_{\max}. \\
& \qquad\qquad \text{(Polymer Length Constraint)}
\end{aligned}
$$

Here, $F_{\text{det}}$ is the *deterministic* objective function (for details, refer to the subsection on objective function formulation given below), which is a function of model-predicted properties $\boldsymbol{P}$ and target properties $\boldsymbol{P}_o$. This objective function is a performance measure of a given molecular design. Each property $P_j$ for a given molecule is predicted by the GCM model, which is a function of the molecular structure of the molecule represented by $\boldsymbol{N}$ (for details, see the next subsection). It also depends on the GCM model parameters $A_{ij}$ and $B_{ij}$, which are associated with a specific molecular group $i$ within that molecule, and a specific property $j$. A polymer molecule is then represented by $\boldsymbol{N}$, which is a vector of integer variables $n_i$, where $n_i$ describes the number of times the $i$th molecular group participates in the polymer repeat unit; $n_i^L$ and $n_i^U$ are the corresponding upper and lower bounds: $\upsilon_i$ represents the valency of the $i$th molecular group; and $f$ represents the number of free attachments available for bonding in the polymer repeat unit. Further, $n_{\min}$ and $n_{\max}$ are the lower and upper bounds on the polymer repeat unit. In all, there are $N$ number of participating molecular groups and $m$ number of property constraints. The aim then is to find the appropriate $\boldsymbol{N}$ representing an optimal polymer design that optimizes the deterministic objective function $F_{\text{det}}$, and thus best matches the desired target properties.

### Group contribution structure – property relations for polymers

GCM models are based on the additive principle that each constituting group contributes to the molecule property. They have been utilized especially extensively in polymer studies due to their simplicity, versatility, and wide applicability to a large number of polymeric properties. The general expression for the property prediction of polymers for most of the properties can be represented as (van Krevelen, 1976)

$$P_j(N) = \frac{\sum_{i=1}^{N} A_{ij} n_i}{\sum_{i=1}^{N} B_{ij} n_i}, \qquad j = 1, \ldots, m.$$

As also explained earlier, $A_{ij}$ and $B_{ij}$ are the GCM model parameters, and are unique to a specific molecular group $i$ and a specific property $j$. These parameter values do not change, and contribute equally in any monomer where they are present. Further it can be observed that GCM models do not consider the way these molecular groups are interconnected to form the monomer. These methods therefore cannot differentiate between isomers. GCM property-prediction expressions are mostly nonlinear, with even more complicated expressions for some specific properties. The proposed approach is flexible enough to incorporate any such complicated prediction model. More comprehensive details about the GCM models associated with various thermophysical, optical, mechanical, and other polymeric properties, and the corresponding parameter values can be found in van Krevelen (1990).

### GCM parameter uncertainty

Being empirical in nature, GCM methods are not sufficiently precise. A 5–10% or greater deviation in the property-prediction values from the experimental values is quite common. This discrepancy in property predictions is attributed to the uncertainty in the GCM parameters like $A_{ij}$ and $B_{ij}$. These parameters are obtained by regression, and hence are prone to measurement errors. For any industrially reliable CAMD, such uncertainties cannot be neglected and need to be incorporated into the optimization framework. However, these uncertainties are rarely measured, and so the exact nature of such a model or model parameter uncertainties is unknown. A probability distribution is then assumed for these uncertain variables to reflect the associated uncertainties. In most cases, parameter values obtained from the literature are then taken as the mean of these probability distributions. Further, a fixed $x\%$ of scatter around this mean value is assumed. Variance value for the distribution function of the uncertain parameter, say $A_{ij}$, is then selected such that 99% of the possible realizations of $A_{ij}$ are within $\pm x\%$ from the mean value:

$$\text{Var}(A_{ij}) = \left( \frac{x \times \mu(A_{ij})}{2.58} \right)^2, \qquad i = 1, \ldots, N \quad \text{and} \quad j = 1, \ldots, m.$$

Detailed case studies of optimization under such model parameter uncertainties in a stochastic framework are given in the later section.

### Objective function formulation for property matching

An optimization formulation, as given earlier in the subsection on mathematical formulation for CAD, provides an automatic and systematic way of finding the best molecules. However, the driving force behind any optimization algorithm is the objective function that drives the algorithm toward the target. In the case of OMD, the characteristic that a molecule has properties close to (if not exact) a specified set of target properties, with or without a set of bounds, has to be mathematically represented by such an objective function. An appropriate objective function is therefore used as the criterion to determine the optimality among various possible molecules. It reflects the desired performance criterion to be optimized, which in this case is a minimum deviation from the target properties.

Several types of performance measure can be used to decide the best molecular design. For a PM problem, two objective functions were considered in this article: (1) root-mean-squared deviation (RMSD) and (2) Gaussian fitness function (GAUSSIAN). RMSD is defined for the case when we need a set of target properties to be matched, while GAUSSIAN is defined for the case when we additionally have strict upper and lower bounds for the desired properties. Both essentially measure deviations from the desired target properties. The smaller the deviations, the closer the match. If the deviations are zero, then all the desired properties are exactly matched simultaneously. This can be possible for deterministic problems, but cannot be achieved for stochastic formulations. The robust optimization framework uses such an optimization function to search for the best molecule with the smallest deviations. These objective functions tend to be highly nonlinear, but their effectiveness is demonstrated in various case studies later.

*Root-Mean-Squared Deviation.* RMSD is defined as the square root of the sum of the squared-scaled deviation from the target properties. It aims at matching all the properties simultaneously. Its expression is given by

$$F_1 = \sqrt{\sum_{i=1}^{m} \omega_i \left( \frac{P_i - P_{io}}{P_{io}} \right)^2},$$

where $P_i$ is the $i$th predicted property, $P_{io}$ is the target value for the $i$th property, and $\omega_i$ is the weight given to the $i$th property. In this case study, all the properties were given equal weight. Depending on the case study considered, however, we could give preference to meeting one specific property target over meeting others. RMSD is equivalent to the weighted distance of a molecule's property from the target properties. The smaller the RMSD, the closer the molecule is to the target properties. Further, by definition $F_1$ values are all positive and the target molecule that satisfies all the target properties corresponds to a zero $F_1$ value. With such an objective function it is a problem of minimization, where minimizing $F_1$ gives priority to molecules closer to the property targets.

*Gaussian Fitness Function.* A GAUSSIAN was earlier chosen in the literature to represent the objective function in a genetic algorithm approach for optimal molecular design (Venkatasubramanian et al., 1994). This objective function is designed for the case when we have strict upper and lower bounds for the desired properties. In terms of genetic search, this fitness function, $F_2$, determines the probability with which a particular molecule participates in the evolution. With such an objective function, if a molecule's properties fall outside of the desired property range, then its corresponding fitness decreases drastically. This objective function is given by

$$F_2 = \exp\left( -\lambda \sum_{i=1}^{m} \omega_i \left( \frac{P_i - P_{io}}{P_{i,\max} - P_{i,\min}} \right)^2 \right),$$

where $P_i$ is the $i$th predicted property, $P_{io}$ is the target value for the $i$th property, and $\omega_i$ is the weight associated with the $i$th property. For the $i$th property, $P_{i,\max}$ and $P_{i,\min}$ form the maximum and minimum acceptable property values, respectively, with $P_{io}$ as the mean. The parameter $\lambda$, also called the *fitness decay rate*, determines the steepness of the fitness curve near the target molecule. As compared to RMSD, GAUSSIAN is here viewed as an inverse function of the weighted distance from the target molecule. The smaller the distance, the higher the GAUSSIAN value and the fitter the molecule. The $F_2$ value ranges from 0 to 1, with 1 being the fitness for the target molecule, that is, the molecule for which all the property values meet the target properties. With such an objective function, the optimal design problem is of maximizing the molecule's fitness, $F_2$, as opposed to minimizing the RMSD in $F_1$.

Since the nature of the objective function might also govern the speed of convergence of the algorithm, such an objective function holds quite a promise in CAMD. The success of this objective function in a genetic algorithm framework (Venkatasubramanian et al., 1994), which is also an essentially stochastic optimization procedure based on natural selection and evolution, was the motivation behind using it in the stochastic optimization approach of this study. The proposed methodology is shown to be easily applicable for such highly nonlinear objective functions as well.

### Design constraints

In the PM formulation of the CAMD problem, there are structural-feasibility and polymer-length constraints. To ensure the structural feasibility of the polymers formed, an equality constraint was imposed. This constraint makes sure that the monomer formed has only two free ends. In addition, any upper and lower bounds on the length of the monomer unit add another inequality constraint in the problem formulation. These constraints can also be seen in the mathematical formulation given earlier.

### Polymer design case study

Polymers that can simultaneously meet a number of physical property targets are of prime industrial importance, and thus are the focus of any polymer design initiative. Along similar lines, the polymer design problem considered in this article aims at simultaneously meeting property constraints on the density, water absorption, and glass transition temperature of the polymer. The case study and the choice of target properties were taken from an earlier work by Derringer and Markham (1985), and are used as test bed problems. The empirical GCM relations used to predict these properties are taken from the literature (van Krevelen, 1976) and are given by

$$\text{Water Absorption } (\text{g H}_2\text{O/g polymer}) \qquad W = \frac{\sum_{i=1}^{N} 18 H_i n_i}{\sum_{i=1}^{N} M_i n_i}$$

$$\text{Glass Transition Temperature (K)} \qquad T_g = \frac{\sum_{i=1}^{N} Y_i n_i}{\sum_{i=1}^{N} M_i n_i}$$

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

The aim of this case study is to design polymers that simultaneously match the following target properties

$$D_o = 1.50 \text{ g/cm}^3, \quad W_o = 0.005 \text{ g H}_2\text{O/g polymer}, \quad T_{go} = 383\ K.$$

The corresponding objective function and design constraints are as formulated in the previous two subsections.

Although the target polymer properties were chosen from the study by Derringer and Markham (1985), they could be useful for designing *barrier polymers*. They act as encapsulants for integrated chips (ICs). A very low water-absorption capacity is obviously desired, as it minimizes any outside moisture, which affects the chip performance. Also a higher density increases the encapsulants' thermal conductivity, which in turn facilitates faster heat dissipation. This process is instrumental in cooling the chips, as they heat up over time. In addition, the glass transition temperature has to be moderately high so as to sustain higher temperatures. Alternative uses can also be found in *food packaging* materials that have similar property requirements. These properties can be customized further for the specific uses of the polymer. Further, additional properties, such as permeability, can be added in the set of target properties, depending on the use. The aim here is to show, through this case study, the usefulness of the proposed method in addressing key uncertainty issues in the optimization framework.

A pool of seven molecular groups is chosen to participate in the polymer repeat units. The same molecular group is allowed to participate up to three times in the polymer repeat units, $n_i \in \{0, 1, 2, 3\}$, $i = 1, 2, \ldots, 7$. No upper bound was imposed on the length of the polymer repeat unit. The search space then consists of a total number of ($4^7 - 1$, that is, 16,383) possible designs of polymer repeat units. Further restricting the size of the molecule, as an additional target property, will definitely reduce this search space significantly, which would make it a simpler problem to solve. This restricting was not considered, because we would not want to miss out on any promising molecule of a different size. Furthermore, as shown later in this article, it was possible to use this methodology to solve this problem in a computationally efficient manner. However, for large-scale problems involving a huge pool of functional groups, with billions of possible molecules, a bound on the size of the repeat monomer unit (if a reasonable estimate of molecule size is known) can also be added in this framework.

All of the seven participating molecular groups and the corresponding GCM model parameters are taken from the literature (van Krevelen, 1976) and are tabulated in Table 1. In the stochastic framework, uncertainty is considered in parameters $Y_i$, $V_i$, and $H_i$. No uncertainty is needed in the $M_i$ values, which represent the molecular weights of each molecular group, and so are rigorously additive. With only seven functional groups, it seems like a smaller case study, but the presence of 21 uncertain GCM parameters makes it a substantially bigger problem in a stochastic optimization framework. Also uncertainties of various types, such as normal, lognormal, and mixed distributions, have been considered, as is described in the later sections.

**Table 1. Molecular Groups and Their Contribution Parameter Values**

| No. | Group | $Y_i$ | $V_i$ | $H_i$ | $M_i$ |
|---|---|---|---|---|---|
| 1 | —$CH_2$— | 2,700 | 15.85 | $3.3 \times 10^{-5}$ | 14 |
| 2 | —CO— | 27,000 | 13.40 | 0.110 | 28 |
| 3 | —COO— | 8,000 | 23.00 | 0.075 | 44 |
| 4 | —O— | 4,000 | 10.00 | 0.020 | 16 |
| 5 | —CONH— | 12,000 | 24.90 | 0.750 | 43 |
| 6 | —CHOH— | 13,000 | 19.15 | 0.750 | 30 |
| 7 | —CHCl— | 20,000 | 29.35 | 0.015 | 48.5 |

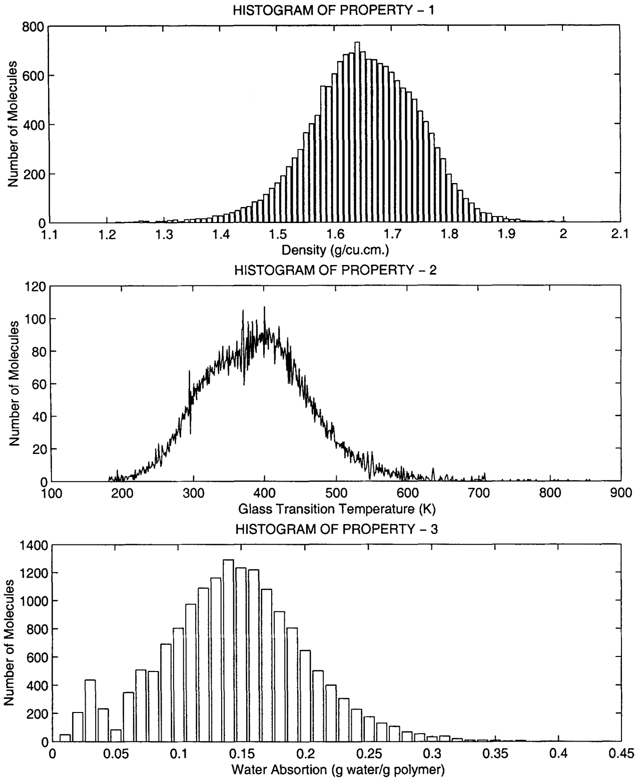

**Figure 1. Exhaustive search on properties of molecules for case study 1.**

### *Characterization of the search space*

As mentioned earlier, this case study considers a pool of seven molecular groups, with up to a maximum of three repetitions for each group. The search space then consists of a total number of ($4^7 - 1$, that is, 16,383) possible designs of polymer repeat units. Useful insights about the space topology can be derived from knowledge of the search space. An exhaustive search was therefore performed, assuming the model parameter values to be constants without any uncertainties. Figure 1 shows the possible range of values and frequencies of occurrence of the three properties, that is, density, water absorption, and glass transition temperature, in these ranges. It can be seen that very few molecules exist that have a water-absorption capacity as low as 0.005 g $H_2O$/g polymer, one of the target properties. The optimization framework has to search for those polymer designs that, besides satisfying such a low water-absorption capacity, also meet the other two desired properties.

Objective function formulation also affects the search for space topology. It plays an essential role in governing the effectiveness of the optimization algorithm. To study this role two objective function formulations, $F_1$ and $F_2$, are used and compared. Figure 2 shows a frequency diagram of the objective function $F_1$ by exhaustive search. As can be observed, there are highly probable regions of local optima near the

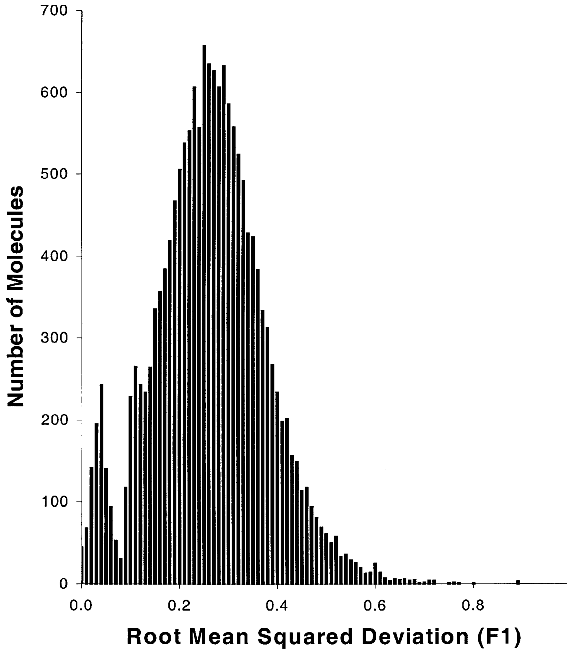

**Figure 2. Exhaustive search of molecules: objective function I.**

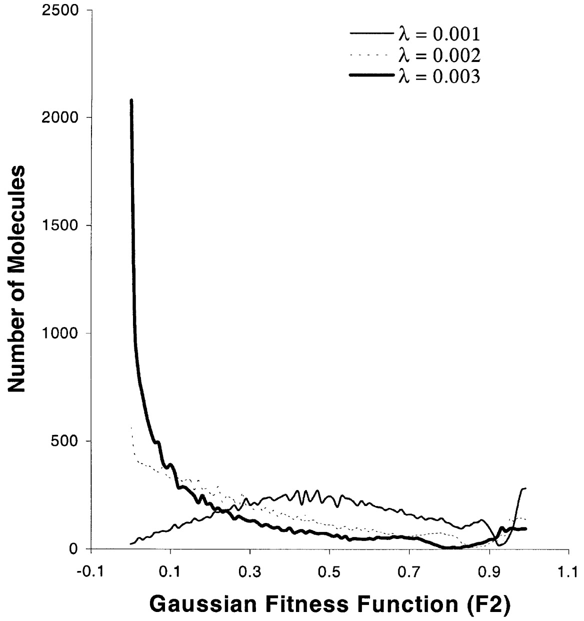

**Figure 3. Exhaustive search of molecules: objective function II.**

optimum, where this optimum occurs at an $F_1$ value of zero. Figure 3 shows a similar graph for objective function $F_2$ with varying $\lambda$ values. In this case, however, it does not show many high regions of local optima near the optimum, which occurs at an $F_2$ value of one. Clearly, compared to $F_1$, $F_2$ is a much better choice of objective function, although both give the same optimal solutions. Additionally, with $F_2$ as the objective function, the steepness of the search space can be tuned by the $\lambda$ values. Therefore, although $F_2$ is just a nonlinear scaling of $F_1$ values, it can make a big difference in the convergence speed of the algorithm. Results to this effect are shown in the various case studies later.

Deterministic solutions for the best 20 molecules without considering any uncertainties are obtained through exhaustive search, and are listed in Table 2 for the two objective functions. The order of molecules differs from the best 10 molecules obtained in a previous study (Maranas, 1996), which considered maximum scaled property deviation as the objective function. Thus, besides governing the speed of the convergence of the algorithm, different objective functions also correspond to different optimal solutions.

**Table 2. Deterministic Solution for Best 20 Polymer Designs (Objective Functions I & II)**

| Rank | Repeating Monomer | $F_1$ RMSD | $F_2$ $\lambda = 0.001$ | $F_2$ $\lambda = 0.002$ | $F_2$ $\lambda = 0.01$ | $W$ | $T$ (K) | $D$ |
|---|---|---|---|---|---|---|---|---|
| 1 | —($CH_2(CHCl)_2$)— | 0.0218 | 1.0000 | 1.0000 | 1.0000 | 0.0049 | 384.68 | 1.4889 |
| 2 | —($CH_2(CHCl)_3$)— | 0.0405 | 1.0000 | 1.0000 | 1.0000 | 0.0051 | 393.10 | 1.5351 |
| 3 | —($(CH_2)_2(CHCl)_3$)— | 0.0708 | 1.0000 | 1.0000 | 0.9999 | 0.0047 | 376.95 | 1.4489 |
| 4 | —($CH_2CHCl$)— | 0.1685 | 1.0000 | 0.9999 | 0.9997 | 0.0043 | 363.20 | 1.3827 |
| 5 | —($CHCl$)— | 0.1750 | 1.0000 | 0.9999 | 0.9997 | 0.0056 | 412.37 | 1.6525 |
| 6 | —($(CH_2)_3O(CHCl)_3$)— | 0.1894 | 1.0000 | 0.9999 | 0.9997 | 0.0058 | 354.30 | 1.3977 |
| 7 | —($(CH_2)_3O(CHCl)_2$)— | 0.2301 | 0.9999 | 0.9999 | 0.9995 | 0.0058 | 336.13 | 1.3333 |
| 8 | —($(CH_2)_2O(CHCl)_3$)— | 0.2454 | 0.9999 | 0.9999 | 0.9994 | 0.0062 | 366.23 | 1.4605 |
| 9 | —($(CH_2)_3(CHCl)_2$)— | 0.2721 | 0.9999 | 0.9999 | 0.9993 | 0.0039 | 346.04 | 1.3082 |
| 10 | —($(CH_2)_2O(CHCl)_2$)— | 0.2995 | 0.9999 | 0.9998 | 0.9991 | 0.0064 | 350.35 | 1.4044 |
| 11 | —($CH_2O(CHCl)_3$)— | 0.3412 | 0.9999 | 0.9998 | 0.9989 | 0.0067 | 380.06 | 1.5408 |
| 12 | —($(CH_2)_3OCHCl$)— | 0.3336 | 0.9999 | 0.9998 | 0.9989 | 0.0059 | 301.41 | 1.2255 |
| 13 | —($(CH_2)_2CHCl$)— | 0.3672 | 0.9999 | 0.9997 | 0.9987 | 0.0035 | 332.03 | 1.2531 |
| 14 | —($(CH_2)_3(O)_2(CHCl)_3$)— | 0.4154 | 0.9998 | 0.9997 | 0.9983 | 0.0070 | 346.70 | 1.4107 |
| 15 | —($CH_2O(CHCl)_2$)— | 0.4219 | 0.9998 | 0.9996 | 0.9982 | 0.0071 | 367.72 | 1.5021 |
| 16 | —($(CH_2)_2OCHCl$)— | 0.4195 | 0.9998 | 0.9996 | 0.9982 | 0.0068 | 317.84 | 1.3019 |
| 17 | —($O(CHCl)_3$)— | 0.4521 | 0.9998 | 0.9996 | 0.9979 | 0.0072 | 396.28 | 1.6471 |
| 18 | —($(CH_2)_3CHCl$)— | 0.4922 | 0.9998 | 0.9995 | 0.9976 | 0.0030 | 310.50 | 1.1769 |
| 19 | —($(CH_2)_2(O)_2(CHCl)_3$)— | 0.5049 | 0.9998 | 0.9995 | 0.9976 | 0.0075 | 357.18 | 1.4705 |
| 20 | —($(CH_2)_3(O)_2(CHCl)_2$)— | 0.5103 | 0.9997 | 0.9995 | 0.9974 | 0.0074 | 328.07 | 1.3545 |
| | Target molecule properties: | | | | | **0.005** | **383.0** | **1.500** |

Overall observations indicate that the search space has numerous local minima, where optimal solutions can get trapped using traditional methods. The search space also shows a very small fraction of molecules that are highly fit, which is mainly due to the stricter property targets. Also a correlation between the structural topology of the search space and the speed of convergence of the algorithm can be established (refer to the section on case studies later).

## New Framework for Optimization Under Uncertainty

Because of the high risk involved in neglecting the unavoidable property-prediction uncertainties in CAMD, a new generalized stochastic framework was developed. In the next subsection we present a mathematical reformulation of the optimal molecular design problem under uncertainty. An efficient stochastic optimization algorithm coupled with a new sampling technique was used, and also is explained in the following subsection. In a later subsection, this approach is compared with a recent chance constrained formulation approach to OMD under uncertainty. A detailed study and comparisons of various approaches to stochastic optimization in large-scale, real-life applications of diverse complexities is given in Birge (1997).

### *Mathematical reformulation for CAMD under uncertainty*

When we consider the property-prediction uncertainties in a stochastic framework, the PM formulation given earlier can be extended to a stochastic property matching (SPM) formulation, which is as shown below:

$$
\begin{aligned}
\text{Optimize:}\quad & F_{\text{stoch}}(\boldsymbol{P},\boldsymbol{P}_o) \\
& F_{\text{stoch}}(\boldsymbol{P},\boldsymbol{P}_o) = P\left[F_{\text{det}}(\boldsymbol{P},\boldsymbol{P}_o)\right] + b(t)\epsilon_P\left[N_{\text{samp}}\right] \\
& \boldsymbol{P} = \left(P_1,\ldots,P_j,\ldots,P_m\right) \\
& \boldsymbol{P}_o = \left(P_{1o},\ldots,P_{jo},\ldots,P_{mo}\right) \\
& P_j = P_j\left(N,U_{1j},\ldots,U_{Nj},C_{1j},\ldots,C_{Nj}\right), \\
& \qquad\qquad j=1,\ldots,m \qquad\qquad \text{(GCM model)} \\
& \boldsymbol{N} = (n_1,\ldots,n_i,\ldots,n_N) \qquad \text{(Polymer molecule)} \\
& n_i \in \left\{n_i^L, n_i^L+1,\ldots,n_i^U\right\}, \qquad i=1,\ldots,N \\
& U_{ij}\colon f_U(u,\mu_{ij},\sigma_{ij}) \qquad \text{(Uncertain variable pdf)} \\
\text{Subject to:}\quad & f = \sum_{i=1}^{N}(v_i-2)n_i+2 \\
& \qquad\qquad \text{(Structural feasibility constraint)} \\
& n_{\min} \le \sum_{i=1}^{N} n_i \le {}_{\max} \quad \text{(Polymer length constraint)}.
\end{aligned}
$$

In addition to the details given in the formulation in the previous section, here $F_{\text{stoch}}$ is the *stochastic* objective function, as explained in the next subsection, which is also a function of the model-predicted properties $\boldsymbol{P}$ and target properties $\boldsymbol{P}_o$. Each property $P_j$ is given by the GCM model, which is a function of the molecular structure represented by $\boldsymbol{N}$, and also depends on the model parameters. Further, $U_{ij}$ and $C_{ij}$ divides the GCM parameters into uncertain and constant model parameters, respectively. The uncertainty of the uncertain parameter $U_{ij}$ is then represented by the probability distribution function $f_U$ with a mean value of $\mu_{ij}$ (often taken as the literature value of that parameter) and a variance of $\sigma_{ij}$. The aim, then, is to find the appropriate $\boldsymbol{N}$, representing an optimal polymer design that optimizes the stochastic objective function $F_{\text{stoch}}$.

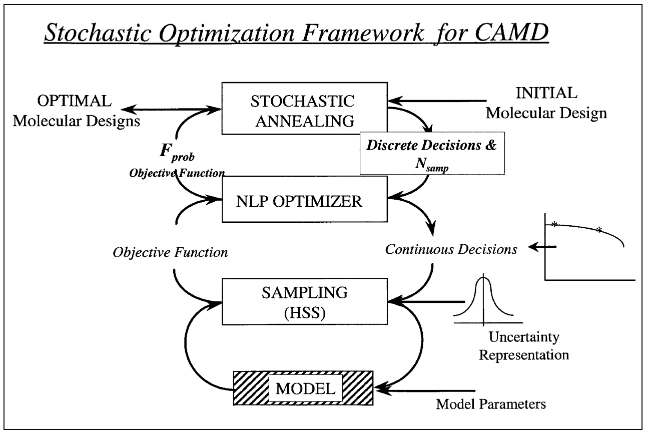

**Figure 4. New framework for stochastic optimization.**

### *Efficient stochastic optimization framework*

The generalized approach to molecular design under uncertainty is to formulate it as a stochastic optimization problem, which involves optimization of a probabilistic function, obtained by sampling over the uncertain variables. Figure 4 shows the outline of the stochastic optimization framework involving three loops: (1) the outer two loops are for obtaining discrete and continuous decision variables, and (2) the inner sampling loop. Recently, Chaudhuri and Diwekar (1996) proposed a new approach to circumvent the computational intensity of these loops based on a new sampling technique. Further, an error characterization of the sampling technique is used to improve the interaction between the optimization and the sampling loop. This approach is described briefly below.

The efficiency of the inner sampling loop is greatly enhanced by using the novel sampling technique named Hammersley sequence sampling (HSS) as opposed to the costly Monte Carlo samples. HSS is based on the Hammersley sequence. Diwekar and Kalagnanam (1997) showed that this new sampling technique is 3 to 100 times faster than the stratified Latin hypercube sampling (LHS). For example, for the continuous stirred-tank reactor (CSTR) problem in Diwekar and Kalagnanam (1997), the HSS technique required about 150 points to converge, an order of magnitude less than the 6500 points required by the LHS and 12,000 Monte Carlo samples. The basic idea in this new sampling technique is to replace the Monte Carlo method by a quasi-Monte Carlo (QMC) scheme. Given the intensive computational burden in a stochastic framework, these techniques are quite appealing over Monte Carlo, and have been increasingly used in varied applications (Sudjianto et al., 1997). Both LHS and HSS sampling techniques are used in the nine case studies given later.

The choice was based on whether it is a one-dimensional or a $k$-dimensional uncertainty domain.

To improve the efficiency of the optimization algorithms for undertaking problems involving uncertainties, a new stochastic annealing-nonlinear programming algorithm is used. The stochastic annealing not only decides the discrete variables but also finds the optimal number of samples, $N_{\text{samp}}$, for each iteration of the inner sampling loop. The NLP algorithm finds the continuous variables. Since in the posed CAMD problem we are only dealing with discrete variables, we will concentrate on the discrete optimization loop. However, in other molecular-design problems, like that of designing a mixture of solvents, we can also encounter continuous variables.

The stochastic annealing algorithm proposed in an earlier work by Chaudhuri and Diwekar (1996), is an algorithm designed to efficiently optimize a probabilistic objective function. This algorithm provides a trade-off between accuracy and efficiency by selecting an increasing number of samples as one approaches the optimum. In stochastic annealing, the cooling schedule is used to decide the weight on the penalty term for imprecision in the probabilistic objective function. The choice of the penalty term itself, on the other hand, must depend on the error bandwidth of the function that is optimized, and must incorporate the effect of the number of samples. Monte Carlo techniques have the advantage that the precision of the estimated parameters can be calculated based on a particular sample size. This is because one can apply standard statistical techniques to analyze the output from a Monte Carlo run, as the sampled values of each output variable is a "random" sample from a true probability distribution of that variable. For example, from the central limit theorem (Morgan and Henrion, 1990), it can be proved that the error band width $\epsilon_\mu$ for the mean $\mu_y$ of the parameter $y = f(x_1, x_2, \ldots)$ calculated using $N_{\text{samp}}$ number of samples, is proportional to $\sqrt{N_{\text{samp}}}$. However, the error-bandwidth expression presented earlier is only applicable to Monte Carlo simulations and not for any other sampling technique. Furthermore, theoretical discussions, such as the preceding one, are not possible for other techniques, as a result of which a systematic quantification of accuracy is lacking for non−Monte Carlo techniques for the various performance indices (and for Monte Carlo techniques for indices other than mean and variance). Chaudhuri and Diwekar (1999) used the fractal dimension approach to characterize the accuracy of the new sampling technique. The fractal dimension approach was used to determine the penalty term for this new variant of the simulated annealing algorithm, where the error is characterized as

$$\epsilon_\mu \propto \frac{\sigma_y}{\left(N_{\text{samp}}\right)^{1.8}}.$$

The new *stochastic objective function* $\boldsymbol{F}_{\textbf{stoch}}$ in stochastic annealing therefore consists of the sum of a probabilistic objective measure $P$ and the penalty function, and is represented as follows:

$$\text{Optimize: } F_{\text{stoch}} = P\left[F_{\text{det}}(x,u)\right] + b(t)\epsilon_P.$$

In the preceding equations, the first term represents the real objective function, which is a probabilistic function in terms of the decision variables $x$ and uncertain variables $u$, and all other terms following the first term signify the *penalty function*. Here, $P$ is a probability measure (that is, mean, variance), over the various $F_{\text{det}}$ values corresponding to different samples. Also $\epsilon_P$ is the error bandwidth corresponding to the probabilistic objective measure $P$ and the sampling technique used. The weighting function $b(t)$ can be expressed in terms of the annealing temperature levels $t$. At high temperatures, the sample size can be small, since the algorithm explores the functional topology or the configuration space, to identify regions of optima. As the system gets cooler, the algorithm searches for the global optimum; consequently, it is necessary to take more samples to get more accurate and realistic objective function evaluations.

The stochastic annealing algorithm minimizes the CPU time by balancing the trade-off between computational efficiency and solution accuracy by the introduction of a penalty function in the objective function. This is necessary, since at a high temperature the algorithm mainly explores the solution space and does not require precise estimates of any probabilistic function. The algorithm must select a larger number of samples, as the solution is near the optimum. The weight of the penalty term, as mentioned before, is governed by $b(t)$, and is based on the annealing temperature.

The complete STA-NLP framework, along with the HSS sampling, has been successfully applied to various problems, such as HDA synthesis and synthesis of optimal waste blends. It was found that for a large-scale problem like synthesis, this framework reduced the computational intensity from 20 days to 18 h (Chaundhuri and Diwekar, 1999).

Stochastic annealing-parameter values govern the robustness of the algorithm, and thus are very crucial for the success of this algorithm. They determine the algorithm's speed of convergence and its ability to avoid local minima. Table 3 lists all the parameter values chosen for the stochastic annealing algorithm. These values are based on guidelines given in the literature (Painton and Diwekar, 1995). The parameters $T_i$, $T_f$, $N_T$, and $\beta$ correspond to various stochastic an-

**Table 3. Specifications and Parameter Values of Stochastic Annealing Algorithm**

| Parameter | Description | Values Cases 1−7 and 9 | Values Case 8 |
|---|---|---|---|
| $T_i$ | Initial temp. | 100 | 100 |
| $T_f$ | Freezing temp. | 1 | 1 |
| $\beta$ | Temp. decrement factor | 0.95 | 0.80 |
| $b(t)$ | Weighting function | $b_o/(k^T)$ | $b_o/(k^T)$ |
| $b_o$ | Empirical constants | 0.001 | 0.001 |
| $k$ | Empirical constants | 0.9 | 0.9 |
| $N_b$ | No. of bit changes in move generator | 5 | 3 |
| $N_T$ | No. of steps at each temp. level | 50 | 25 |
| $F_{\text{stoch}}$ | Stochastic objective function | Mean$(F_{\text{det}}(x,u))$ + penalty | |
| | Penalty function | $b(t)\ \epsilon_P$ | |
| $\epsilon_P$ | Error bandwidth | $f[(N_{\text{samp}})^{-0.5}]$ (LHS)<br>$f[(N_{\text{samp}})^{-1.8}]$ (HSS) | |

nealing schedule parameters, and the parameter $N_b$ decides the number of bit changes performed on the polymer molecule $\boldsymbol{N}$ to obtain the next perturbed molecule.

### *Parallel with the chance constrained approach*

Recently, a chance constrained optimization approach was applied to molecular design under uncertainty (Maranas, 1997). In this approach, the objective function and constraints are satisfied with some probability (Charnes and Cooper, 1959). These probabilistic constraints are then converted into equivalent deterministic constraints in a mathematical programming framework.

Chance constrained formulation for the case of SPM is shown below:

$$\text{Min} \quad s$$
$$\text{Subject to} \quad \text{Prob}\left(\left|\frac{P_j(N)-P_{jo}}{P_{jo}}\right| \le s\right) \ge \alpha \qquad j=1,\ldots,m,$$

where $s$ represents the maximum-scaled deviations of a molecule's properties from the target properties; and $\alpha$ refers to the probability of meeting the objective function with that value of $s$. The impact of uncertainty is then reflected in the trade-off curve between the probability of meeting the objective function value and the objective function value itself. In this study, trade-off curves are obtained between $s$ and $\alpha$. The larger the maximum scaled property violation $s$ that can be tolerated, the higher the probability of maintaining deviations within this tolerable limit.

A similar trade-off curve can also be drawn using the proposed stochastic optimization approach. To draw a comparison between the two approaches, random samples are drawn from the uncertainty distributions of model parameters using the same type of uncertainty distribution, scatter values, and objective function as taken in the recent study by Maranas (1997). These samples are then passed through the prediction model, and corresponding values of the maximum scaled property violation $s$ are obtained for each sample. Through these values, the probability $\alpha$ of meeting different values of $s$ can be obtained. This trade-off curve is then compared with the data values taken from the trade-off curve from a recent study (Figure 5) for the same molecule —$(CH_2)_2$—$(CHCl)_3$—. Matching trends of a similar nature are obtained in the two cases.

Thus, even though the two approaches differ in problem representation, formulation, and solution methodology, the two solutions are essentially equivalent. In both the approaches uncertainty is assumed in the model parameters, in the form of some known distribution functions. This uncertainty is then propagated through the model and further through the optimization framework, though differently in the two approaches.

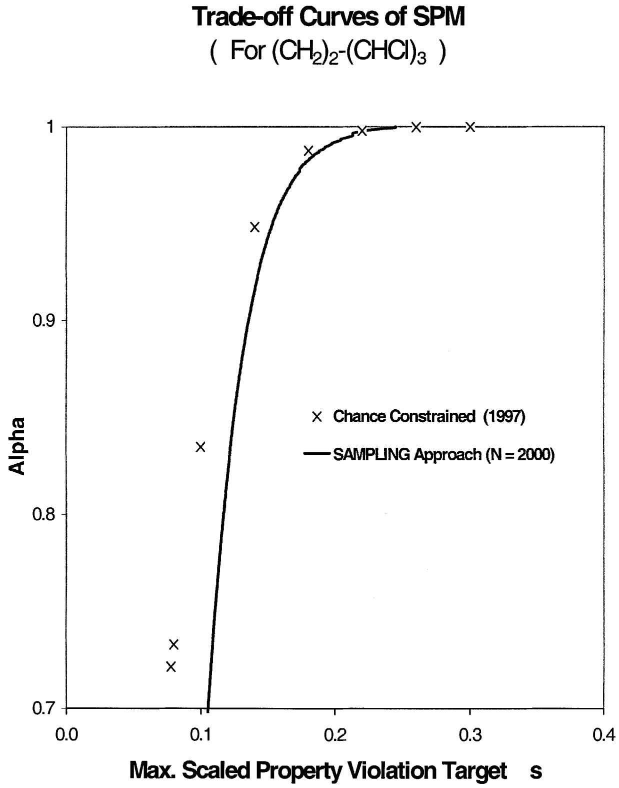

**Figure 5. Stochastic optimization (through sampling) vs. chance constrained optimization approach.**

## Polymer Design Case Studies

Nine case runs were performed on the polymer design case study, as explained earlier, using the proposed stochastic optimization approach. These runs looked at the various crucial aspects of optimal polymer design under uncertainty. There are four important aspects of *uncertainty analysis* in the problem of optimal molecular design: (1) appropriate uncertainty representation of the uncertain parameters, (2) incorporating uncertainties into the optimization framework, (3) uncertainty propagation through the model, and (4) quantifying the impact of uncertainty in the final design. The development of a new stochastic optimization framework in the preceding section addressed aspects (2) and (3). Aspects (1) and (4) are addressed in the various case studies from case 1 to case 9 given below.

These case studies are chosen keeping the following objectives in mind: (1) to address all the crucial uncertainty issues in CAMD in a simplified and a very thorough manner, (2) to be able to draw a parallel with other CAMD approaches with and without uncertainty in the literature (such as Maranas, 1996, 1997), and (3) to highlight the key advantages and novelties unique to our new sampling framework for molecular design under uncertainty. However, these cases represent just a small portion of the possible problems that can be solved by this methodology.

### *Quantifying impact of uncertainties in a stochastic optimization framework*

Quantifying the impact of uncertainty in the final optimal molecular designs is very important. In the context of stochastic optimization, we define Uncertainty Impact (UI) as the percentage deviation of the stochastic objective function $\boldsymbol{F}_{\mathbf{stoch}}$ from the deterministic objective function $\boldsymbol{F}_{\mathbf{det}}$, which are explained in the previous section. Thus, it is a measure of the deviations in the objective function caused by the model uncertainties, normalized over the deterministic objective function value:

$$\mathrm{UI} = \left( \frac{F_{\mathrm{stoch}} - F_{\mathrm{det}}}{F_{\mathrm{det}}} \right) \times 100.$$

Such a definition of UI helps significantly in quantifying the effect of uncertainty in the stochastic optimization approaches through sampling. In the context of sensitivity analysis, the model parameter corresponding to the highest UI value is also the most sensitive parameter (for case studies, refer to the various subsections below).

### *Effect of types of uncertainty representation (cases 1–3)*

The literature often has little or no information on the uncertainty associated with the various model parameters. As is shown below in these studies, uncertainty representation can play a key role in optimal CAMD. Cases 1, 2, and 3 focus on the various forms of uncertainty representation and their impact on the optimal molecular designs.

*Case 1: All Stable Distribution (Normal).* This first case study considers all the parameters to be normally distributed, with mean values taken as the actual parameter values obtained from the literature. This is a similar case to that in Maranas (1997), where all the model parameters are represented by stable distribution, such as normal distribution. A scatter of 20% around the mean was assumed such that a 99% confidence interval lies within this scatter.

Table 4 lists the best 20 molecular designs obtained under uncertainty. It also lists the $F_{\mathrm{det}}$, $F_{\mathrm{stoch}}$, and the UI values for these designs. The relative ranking of molecules was unchanged as compared to the deterministic solutions given in Table 2. UI values were very high for the top few optimal designs, but were not very high for suboptimal molecules. In addition, Figure 6 will show the best 200 molecules obtained, from top to bottom, in a gray color-coded format for better visualization. There are 200 rows, where each row signifies one polymer design. For example, third row shows the third best polymer for this case. In addition, three different levels of gray color density signify whether that particular group was repeated 1, 2 or 3 times in the monomer unit. The higher the density, the more the number of repetitions of that molecular group. The white color signifies the complete absence of that group. From Figure 6 it can be clearly observed that groups —CHOH— and —CHNH— are not desired in the monomer having targeted properties.

*Case 2: All Unstable Distribution (Lognormal).* GCM models are based on the group-contribution parameters, where each group "contributes" to the property value predicted. Hence, essentially all GCM parameters are positive. Uncertainty associated with such parameters can be realistically and more appropriately represented via lognormal distribution, which spans only the positive real values of the random variable. Furthermore, a phenomenon that stems from the multiplicative effect of a large number of uncorrelated factors tends to have a lognormal distribution (Kottegoda and Rosso, 1997). Several phenomena in nature, such as size distribution of small particles in sediment transport, have been best represented by a lognormal distribution. GCM model parameters are also regression coefficients, which implicitly represent the effect of several contributing factors governing the molecule property. These parameters reflect the cumulative effect of various molecular-level interactions, in the form of a lumped parameter. The aforementioned reasons collectively suggest that lognormal is possibly a better representation of the property-prediction GCM model parameter uncertainty.

**Table 4. Stochastic Solution for Best 20 Polymer Designs (Case 1: All Normal Distributions)**

| | Repeating | | | *UI* | Deterministic Properties | | |
|---|---|---|---|---|---|---|---|
| Rank | Monomer | $F_{\mathrm{stoch}}$ | $F_{\mathrm{det}}$ | (%) | *W* | *T* (K) | *D* |
| 1 | —$(CH_2(CHCl)_3)$— | 0.1308 | 0.0405 | 222.65 | 0.0051 | 393.10 | 1.5351 |
| 2 | —$((CH_2)_2(CHCl)_3)$— | 0.1385 | 0.0708 | 95.60 | 0.0047 | 376.95 | 1.4489 |
| 3 | —$(CH_2CHCl)$— | 0.1899 | 0.1685 | 12.73 | 0.0043 | 363.20 | 1.3827 |
| 4 | —$(CHCl)$— | 0.2944 | 0.1750 | 68.25 | 0.0056 | 412.37 | 1.6525 |
| 5 | —$((CH_2)_3(O)_2(CHCl)_3)$— | 0.4255 | 0.4154 | 2.44 | 0.0070 | 346.70 | 1.4107 |
| 6 | —$((CH_2)_3(O)_2(CHCl)_2)$— | 0.5179 | 0.5103 | 1.49 | 0.0074 | 328.07 | 1.3545 |
| 7 | —$(CH_2(O)_2(CHCl)_3)$— | 0.6130 | 0.6019 | 1.86 | 0.0080 | 369.19 | 1.5456 |
| 8 | —$(CH_2OCHCl)$— | 0.6538 | 0.6126 | 6.73 | 0.0080 | 340.13 | 1.4221 |
| 9 | —$((CH_2)_2(O)_2CHCl)$— | 0.8660 | 0.8500 | 1.89 | 0.0091 | 307.83 | 1.3387 |
| 10 | —$((CH_2)_2(COO)(CHCl)_3)$— | 1.0021 | 0.9873 | 1.50 | 0.0099 | 337.47 | 1.5236 |
| 11 | —$((CH_2)_3(COO)O(CHCl)_3)$— | 1.0570 | 1.0517 | 0.51 | 0.0102 | 323.64 | 1.4680 |
| 12 | —$((CH_2)_2(O)_3CHCl)$— | 1.2005 | 1.2028 | −0.19 | 0.0109 | 300.40 | 1.3674 |
| 13 | —$((CH_2)_3(COO)\ (O)_2\ (CHCl)_3)$— | 1.2083 | 1.1918 | 1.38 | 0.0109 | 319.17 | 1.4754 |
| 14 | —$(CH_2(COO)(O)\ (CHCl)_3)$— | 1.3126 | 1.3066 | 0.46 | 0.0115 | 340.32 | 1.6034 |
| 15 | —$(CH_2(COO)(O)_2\ (CHCl)_3)$— | 1.4620 | 1.4473 | 1.02 | 0.0122 | 334.18 | 1.6031 |
| 16 | —$(CH_2(COO)(O)_3\ (CHCl)_3)$— | 1.5941 | 1.5878 | 0.40 | 0.0129 | 328.83 | 1.6029 |
| 17 | —$((CH_2)_2(COO)\ CHCl)$— | 1.7201 | 1.7229 | −0.16 | 0.0135 | 277.18 | 1.4337 |
| 18 | —$((O)_3CHCl)$— | 1.8215 | 1.8069 | 0.80 | 0.0140 | 331.61 | 1.6259 |
| 19 | —$((CH_2)_3(CO)\ (O)_2\ (CHCl)_3)$— | 1.8512 | 1.8422 | 0.49 | 0.0142 | 416.57 | 1.4645 |
| 20 | —$(CH_2(COO)_2(CHCl)_3)$— | 1.8585 | 1.8506 | 0.43 | 0.0142 | 317.98 | 1.6511 |
| | | Target molecule properties: | | | **0.005** | **383.0** | **1.500** |

For this case, lognormal distribution was assumed for all the variables. The parameter value obtained from the literature was considered to be the median of the lognormal distribution. The variance of the lognormal distribution was taken to be equivalent to a scatter of 20% in the transformed normal distribution function.

Table 5 lists the best 20 molecular designs obtained under uncertainty. Drastic changes in the relative ranking of molecules were observed as compared to the deterministic solutions given in Table 2. For example monomer —[(CH$_2$)$_3$O]— faired better than —(CH$_2$CHCl)— in this case, with uncertainty incorporated on the basis of $F_{\text{stoch}}$ values. On the other hand —(CH$_2$CHCl)— was a much better design in deterministic solutions on the basis of $F_{\text{det}}$ values. UI values were very high throughout for all the molecules, signifying a high state of uncertainty with such unstable distributions as lognormal distribution.

*Case 3: Mixed Distributions (Normal, Uniform, and Lognormal).* A marked difference can be seen in Table 1 in the order of magnitude of the various parameter values. While $H_i$ values lie in the (0,1) range, $V_i$ is in the (10,30) range, and $Y_i$ is in the (1,000, 30,000) range. This difference is only natural since these parameters correspond to different physical properties, which in turn are of different orders of magnitude. However, since these parameters are regression coefficients, a small measurement of the regression errors in these values can translate into large errors in the property predictions. A lot depends on the accuracy of the measurement performed on these regressions. Of further importance is the fact that this error propagation would be relatively more noticeable for parameters with values of much smaller orders of magnitude. For example, for the —CH$_2$— group, with an $H_i$ value of $3.3 \times 10^{-5}$, even a measurement error of the order of $10^{-2}$ is quite significant compared to a similar error in the $Y_i$ value of 2700. This case study has used different distribution types to represent the uncertainty in various parameters in order to see their effect on the optimal design and on the UI values. The choice of distribution type for each parameter is based on the relative order of magnitude of that parameter, as well as on the knowledge of accuracy associated with the different parameter values reported in the literature. Flexibility in the proposed stochastic optimization approach through sampling can accommodate all such possibilities.

Since the $H_i$ values are very close to zero, a lognormal distribution is chosen to represent their uncertainty. Also normal and uniform distribution are chosen as representative of the uncertainty in $V_i$ and $Y_i$ values. For the normal and lognormal distributions, mean and variance values were the same as in case 1 and case 2, respectively. For the uniform distribution, the literature values were taken as the mean, and a 20% scatter around the mean value formed the two ends of the uniform distribution.

Table 6 lists the best 20 molecular designs obtained under uncertainty with mixed distributions. The relative ranking of molecules remains mostly unchanged (barring a few exceptions) as compared to the deterministic solutions given in Table 2. The UI values were very high for the top few optimal designs, but were not very high for suboptimal molecules.

*Comparison between Different Distributions.* The UI values of the top 200 molecules were taken as an indicator of the uncertainty impact for different distribution types. Figure 7 shows the probability distribution function of the UI values of the top 200 molecules obtained in cases 1, 2, and 3, respectively. Clearly, lognormal distribution had a dramatic effect on the optimal solutions, as compared to the mixed and normal distributions. Mixed distributions from case 2 also had more impact than the normal distribution in case 1.

Studies on factor analysis show that a number of physical properties are correlated (Joback, 1995), which, in turn, implies that the model parameters corresponding to these prop-

**Table 5. Stochastic Solution for Best 20 Polymer Designs (Case 2: All Lognormal Distributions)**

| | Repeating | | | *UI* | Deterministic Properties | | |
|---|---|---|---|---|---|---|---|
| Rank | Monomer | $F_{\text{stoch}}$ | $F_{\text{det}}$ | (%) | *W* | *T* (K) | *D* |
| 1 | —((CH$_2$)$_3$O)— | 0.7722 | 0.6104 | 26.51 | 0.0062 | 208.62 | 1.0078 |
| 2 | —((CH$_2$)$_3$OCHCl)— | 0.8000 | 0.3336 | 139.81 | 0.0059 | 301.41 | 1.2255 |
| 3 | —(CH$_2$CHCl)— | 0.8671 | 0.1685 | 414.65 | 0.0043 | 363.20 | 1.3827 |
| 4 | —((CH$_2$)$_3$O(CHCl)$_2$)— | 0.8809 | 0.2301 | 282.92 | 0.0058 | 336.13 | 1.3333 |
| 5 | —((CH$_2$)$_2$OCHCl)— | 0.9506 | 0.4195 | 126.60 | 0.0068 | 317.84 | 1.3019 |
| 6 | —((CH$_2$)$_3$(O)$_2$(CHCl)$_2$)— | 1.0058 | 0.5103 | 97.12 | 0.0074 | 328.07 | 1.3545 |
| 7 | —((CH$_2$)$_3$ (O)$_2$CHCl)— | 1.0658 | 0.6799 | 56.76 | 0.0081 | 294.69 | 1.2642 |
| 8 | —(CH$_2$(O))— | 1.7233 | 1.4782 | 16.58 | 0.0120 | 223.33 | 1.1605 |
| 9 | —((CH$_2$)$_2$(COO)O(CHCl)$_2$)— | 1.7909 | 1.4525 | 23.30 | 0.0122 | 310.27 | 1.4992 |
| 10 | —(CH$_2$(COO)O(CHCl)$_2$)— | 1.9262 | 1.6494 | 16.78 | 0.0132 | 319.88 | 1.5900 |
| 11 | —((CH$_2$)$_2$ (O)$_3$)— | 2.0714 | 1.8919 | 9.49 | 0.0142 | 228.95 | 1.2318 |
| 12 | —(CH$_2$(COO)$_2$(CHCl)$_3$)— | 2.1151 | 1.8506 | 14.29 | 0.0142 | 317.98 | 1.6511 |
| 13 | —((CH$_2$)$_3$(COO) (O)$_3$ CHCl)— | 2.1717 | 1.9857 | 9.37 | 0.0148 | 263.56 | 1.4049 |
| 14 | —((CH$_2$)$_3$(CO)(CHCl)$_3$)— | 2.1874 | 1.6076 | 36.07 | 0.0130 | 441.30 | 1.4463 |
| 15 | —((CH$_2$)$_3$(CO)O(CHCl)$_3$)— | 2.2291 | 1.7243 | 29.28 | 0.0136 | 428.08 | 1.4560 |
| 16 | —(CH$_2$(COO)$_2$ O (CHCl)$_3$)— | 2.3085 | 1.9509 | 18.33 | 0.0147 | 313.85 | 1.6479 |
| 17 | —(CH$_2$ (O)$_2$)— | 2.3512 | 2.1805 | 7.83 | 0.0157 | 232.61 | 1.2831 |
| 18 | —((COO) (O)$_2$(CHCl)$_2$)— | 2.3576 | 2.0304 | 16.12 | 0.0151 | 323.70 | 1.7011 |
| 19 | —((CH$_2$)$_3$(CO) O (CHCl)$_2$)— | 2.5541 | 2.1447 | 19.09 | 0.0157 | 432.24 | 1.4115 |
| 20 | —((CH$_2$)$_3$(O)$_3$CHCl)— | 2.6325 | 0.9999 | 163.29 | 0.0098 | 289.53 | 1.2956 |
| | Target molecule properties: | | | | **0.005** | **383.0** | **1.500** |

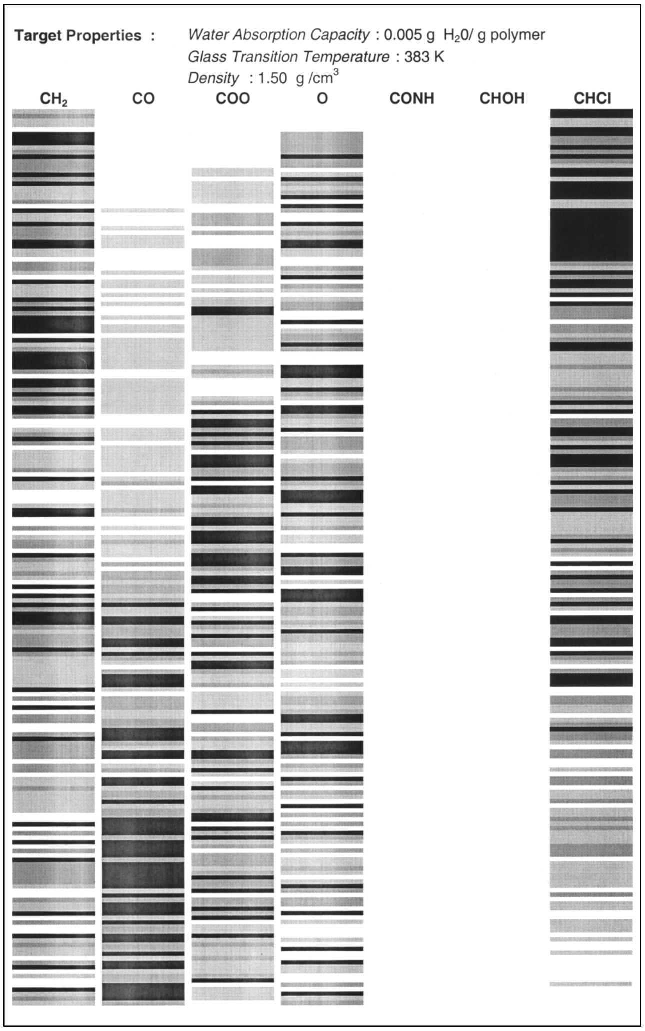

**Figure 6. Best 200 molecules: stochastic solution (all normal distributions).**

erties are correlated as well. However, the exact nature of these correlations is unknown. The proposed methodology is equally applicable to *correlated* and other *user-defined* distribution functions. Such cases have not been considered here due to space constraints.

### *Sensitivity of uncertain parameters (cases 4–6)*

The importance of sensitivity analysis of uncertainties in the physical-property estimation has long been realized (Duvedi and Achenie, 1997; Venkatasubramaniam et al., 1994; Gani et al., 1991). As seen in the previous subsection, the numerical values of the parameters differ by orders of magnitude. Besides, different parameters govern different properties. For example, $Y_i$ affect glass transition properties, $V_i$ affect polymer density, and $H_i$ affect the water-absorption property. In addition, the objective function represents the cumulative effect of the scaled deviation from the target values of all these predicted properties values. Thus uncertainties in some parameters might have significantly higher impact in their corresponding property predictions, and thus eventually in the final optimal molecular designs. There is a relative order in the UI values for different parameters. This study emphasizes which parameter is the most crucial, and also highlights which parameters are desired to be more accurate for a reliable CAMD. This, in turn, will depend on the target properties and the nature of the objective function. Such a sensitivity analysis can be easily performed in the proposed framework by comparing the UI values.

Unlike case 1, in these case studies uncertainty was assumed only in one parameter while keeping other parameters at constant values, as reported in the literature. Parameters $H_i$, $V_i$, and $Y_i$ were assumed uncertain in cases 4, 5, and 6, respectively. Normal distribution was assumed for the only uncertain parameter, with the mean and variance values calculated as described in the earlier cases. Additionally, the same scatter of 20% around the mean value was assumed for

**Table 6. Stochastic Solution for Best 20 Polymer Designs (Case 3: Mixture of Distributions)**

| | | | | | Deterministic Properties | | |
|---|---|---|---|---|---|---|---|
| Rank | Repeating Monomer | $F_{stoch}$ | $F_{det}$ | *UI* (%) | *W* | *T* (K) | *D* |
| 1 | —($CH_2CHCl$)— | 0.3102 | 0.1685 | 84.14 | 0.0043 | 363.20 | 1.3827 |
| 2 | —($CH_2OCHCl$)— | 0.7217 | 0.6126 | 17.81 | 0.0080 | 340.13 | 1.4221 |
| 3 | —(($CH_2$)$_3$(O)$_3$(CHCl)$_2$)— | 0.8705 | 0.7620 | 14.24 | 0.0087 | 321.39 | 1.3725 |
| 4 | —((O)CHCl)— | 1.1309 | 0.9649 | 17.21 | 0.0098 | 372.09 | 1.6391 |
| 5 | —($CH_2$(O)$_3$(CHCl)$_2$)— | 1.1482 | 1.0451 | 9.87 | 0.0102 | 344.03 | 1.5208 |
| 6 | —($CH_2$)— | 1.1818 | 1.1897 | −0.67 | 0.0000 | 192.86 | 0.8833 |
| 7 | —($CH_2$(O))— | 1.5911 | 1.4782 | 7.64 | 0.0120 | 223.33 | 1.1605 |
| 8 | —(COO(CHCl)$_2$)— | 1.7751 | 1.6904 | 5.01 | 0.0134 | 340.43 | 1.7258 |
| 9 | —(($CH_2$)$_2$(COO) (O)$_3$ (CHCl)$_2$)— | 1.8539 | 1.7530 | 5.75 | 0.0137 | 301.38 | 1.5132 |
| 10 | —(($CH_2$)$_3$(CO) (O)$_3$ (CHCl)$_3$)— | 2.0319 | 1.9411 | 4.68 | 0.0147 | 406.45 | 1.4721 |
| 11 | —((COO)$_2$ (CHCl)$_3$)— | 2.0987 | 2.0121 | 4.30 | 0.0150 | 325.48 | 1.7419 |
| 12 | —(($CH_2$)$_2$(CO) O (CHCl)$_2$)— | 2.5681 | 2.4068 | 6.70 | 0.0170 | 452.07 | 1.4851 |
| 13 | —($CH_2$ (O)$_3$)— | 2.6872 | 2.5110 | 7.02 | 0.0174 | 237.10 | 1.3522 |
| 14 | —(($CH_2$)$_2$(COO)$_3$(O)$_3$(CHCl)$_2$)— | 2.8266 | 2.7378 | 3.24 | 0.0186 | 266.89 | 1.6103 |
| 15 | —($CH_2$(CO)(COO) (O)$_3$(CHCl)$_3$)— | 2.8296 | 2.7417 | 3.20 | 0.0187 | 392.49 | 1.6412 |
| 16 | —(($CH_2$)$_2$(CO)(COO)$_2$(CHCl)$_3$)— | 2.8890 | 2.8012 | 3.14 | 0.0190 | 374.44 | 1.616 |
| 17 | —((CO) (COO) (CHCl)$_3$)— | 2.9412 | 2.8084 | 4.73 | 0.0190 | 436.78 | 1.7477 |
| 18 | —((CO) (COO) (O) (CHCl)$_3$)— | 2.9469 | 2.8663 | 2.81 | 0.0193 | 423.98 | 1.7367 |
| 19 | —($CH_2$(CO) (O)$_3$ (CHCl)$_2$)— | 2.9624 | 2.8640 | 3.44 | 0.0193 | 436.90 | 1.5854 |
| 20 | —(($CH_2$)$_3$(CO)(O)CHCl)— | 2.9821 | 2.8857 | 3.34 | 0.0194 | 439.41 | 1.341 |
| | Target molecule properties: | | | | **0.005** | **383.0** | **1.500** |

Table 7. Stochastic Solution for Best 20 Polymer Designs (Case 4: Uncertainty only in $Y_i$)

| Rank | Repeating Monomer | $F_{\text{stoch}}$ | $F_{\text{det}}$ | $UI$ (%) | Deterministic Properties $W$ | $T$ (K) | $D$ |
|---|---|---|---|---|---|---|---|
| 1 | —(CHCl)— | 0.1861 | 0.1750 | 6.33 | 0.0056 | 412.37 | 1.6525 |
| 2 | —($(CH_2)_3$(O)CHCl)— | 0.3401 | 0.3336 | 1.95 | 0.0059 | 301.41 | 1.2255 |
| 3 | —($CH_2$(O)$(CHCl)_3$)— | 0.3428 | 0.3412 | 0.49 | 0.0067 | 380.06 | 1.5408 |
| 4 | —($(CH_2)_2$CHCl)— | 0.3672 | 0.3672 | 0.00 | 0.0035 | 332.03 | 1.2531 |
| 5 | —($(CH_2)_3$CHCl)— | 0.4969 | 0.4922 | 0.96 | 0.0030 | 310.50 | 1.1769 |
| 6 | —((O)$(CHCl)_2$)— | 0.6058 | 0.6079 | −0.35 | 0.0080 | 389.38 | 1.6448 |
| 7 | —($(CH_2)_3$(O))— | 0.6139 | 0.6104 | 0.58 | 0.0062 | 208.62 | 1.0078 |
| 8 | —($(CH_2)_2(O)_2(CHCl)_2$)— | 0.6221 | 0.6126 | 1.56 | 0.0080 | 340.13 | 1.4221 |
| 9 | —($(CH_2)_2(O)_3(CHCl)_2$)— | 0.8876 | 0.8911 | −0.39 | 0.0094 | 331.79 | 1.4369 |
| 10 | —($(CH_2)_3$(COO)$(O)_2(CHCl)_2$)— | 1.4483 | 1.4377 | 0.73 | 0.0121 | 298.14 | 1.4405 |
| 11 | —($(CH_2)_2$(COO)$(O)_3(CHCl)_3$)— | 1.4520 | 1.4491 | 0.20 | 0.0122 | 321.66 | 1.5369 |
| 12 | —($CH_2$(COO)$(O)_2(CHCl)_3$)— | 1.4550 | 1.4473 | 0.53 | 0.0122 | 334.18 | 1.6031 |
| 13 | —($CH_2(O)_3$CHCl)— | 1.4571 | 1.4514 | 0.39 | 0.0122 | 314.03 | 1.4694 |
| 14 | —($(CH_2)_2(COO)_2$(O)$(CHCl)_3$)— | 1.8025 | 1.8115 | −0.49 | 0.0140 | 307.75 | 1.5789 |
| 15 | —($CH_2$(COO)$(O)_2(CHCl)_2$)— | 1.8031 | 1.8100 | −0.38 | 0.0140 | 313.90 | 1.5908 |
| 16 | —($(CH_2)_2$(COO)(O)CHCl)— | 1.9250 | 1.9215 | 0.18 | 0.0145 | 273.99 | 1.4514 |
| 17 | —($CH_2$(COO)$(O)_3(CHCl)_2$)— | 1.9383 | 1.9307 | 0.39 | 0.0146 | 308.87 | 1.5915 |
| 18 | —($(CH_2)_2(COO)_2(CHCl)_2$)— | 2.0594 | 2.0554 | 0.20 | 0.0152 | 288.26 | 1.5616 |
| 19 | —($(CH_2)_2$(CO)$(O)_3(CHCl)_3$)— | 2.1068 | 2.1021 | 0.22 | 0.0155 | 418.44 | 1.5293 |
| 20 | —($CH_2(O)_2$)— | 2.1741 | 2.1805 | −0.29 | 0.0157 | 232.61 | 1.2831 |
| | Target molecule properties: | | | | **0.005** | **383.0** | **1.500** |

each parameter in the three cases. Also, as mentioned earlier, equal importance is given to all the properties (that is, $\omega_i = 1$ for $i = 1$, 2 and 3) in the objective function definition. These factors collectively form a common ground for comparisons and sensitivity analysis of model parameters. LHS sampling was used instead of the HSS.

Tables 7, 8, and 9 list the best 20 optimal molecular designs, under uncertainty, obtained in the three cases. Figure 8 shows the probability distribution function of the UI values of the top 200 molecules obtained in cases 4, 5, and 6, respectively. Clearly $H_i$ had the most significant uncertainty effect, and thus is the most sensitive parameter compared to $V_i$ and $Y_i$, which are of equal sensitivity, but are not as sensitive as $H_i$.

### *Effect of objective function formulation (case 7)*

*Case 7: Gaussian Fitness Function.* As was also mentioned earlier, the nature of the objective function governs the search space topology. This case study was performed to study the impact on the speed of convergence and on the overall optimal designs, of highly nonlinear objective functions like the

Table 8. Stochastic Solution for Best 20 Polymer Designs (Case 5: Uncertainty only in $V_i$)

| Rank | Repeating Monomer | $F_{\text{stoch}}$ | $F_{\text{det}}$ | $UI$ (%) | Deterministic Properties $W$ | $T$ (K) | $D$ |
|---|---|---|---|---|---|---|---|
| 1 | —($CH_2(CHCl)_3$)— | 0.0699 | 0.0405 | 72.54 | 0.0051 | 393.10 | 1.5351 |
| 2 | —($CH_2$(O)$(CHCl)_3$)— | 0.3421 | 0.3412 | 0.28 | 0.0067 | 380.06 | 1.5408 |
| 3 | —($(CH_2)_3(O)_2(CHCl)_3$)— | 0.4170 | 0.4154 | 0.39 | 0.0070 | 346.70 | 1.4107 |
| 4 | —($(CH_2)_2$(O)CHCl)— | 0.4266 | 0.4195 | 1.69 | 0.0068 | 317.84 | 1.3019 |
| 5 | —($(CH_2)_2$(O))— | 0.8337 | 0.8325 | 0.14 | 0.0082 | 213.64 | 1.0552 |
| 6 | —((O)CHCl)— | 0.9624 | 0.9649 | −0.25 | 0.0098 | 372.09 | 1.6391 |
| 7 | —($(CH_2)_2$(COO) $(CHCl)_3$)— | 0.9977 | 0.9873 | 1.05 | 0.0099 | 337.47 | 1.5236 |
| 8 | —($CH_2(O)_2$CHCl)— | 1.1096 | 1.1109 | −0.12 | 0.0105 | 324.87 | 1.4494 |
| 9 | —($CH_2$(COO) $(CHCl)_3$)— | 1.1325 | 1.1260 | 0.58 | 0.0106 | 347.42 | 1.6036 |
| 10 | —($(CH_2)_2$(COO) (O) $(CHCl)_3$)— | 1.1674 | 1.1679 | −0.04 | 0.0108 | 331.48 | 1.5286 |
| 11 | —($CH_2$)— | 1.1823 | 1.1897 | −0.63 | 0.0000 | 192.86 | 0.8833 |
| 12 | —($(CH_2)_3$(COO) $(O)_2$ $(CHCl)_3$)— | 1.2012 | 1.1918 | 0.79 | 0.0109 | 319.17 | 1.4754 |
| 13 | —($(CH_2)_3$(COO) (O) $(CHCl)_2$)— | 1.2833 | 1.2785 | 0.38 | 0.0113 | 302.01 | 1.4291 |
| 14 | —($(O)_2$ CHCl)— | 1.4682 | 1.4655 | 0.19 | 0.0123 | 347.83 | 1.6312 |
| 15 | —($(CH_2)_3$(CO) $(CHCl)_3$)— | 1.6002 | 1.6076 | −0.46 | 0.0130 | 441.30 | 1.4463 |
| 16 | —($(CH_2)_3$(COO) (O) CHCl)— | 1.6652 | 1.6703 | −0.31 | 0.0132 | 266.45 | 1.3694 |
| 17 | —((COO) $(CHCl)_2$)— | 1.6940 | 1.6904 | 0.22 | 0.0134 | 340.43 | 1.7258 |
| 18 | —($(CH_2)_2(COO)_2$ $(CHCl)_3$)— | 1.6988 | 1.6912 | 0.45 | 0.0134 | 311.28 | 1.5777 |
| 19 | —($(O)_3$CHCl)— | 1.8074 | 1.8069 | 0.02 | 0.0140 | 331.61 | 1.6259 |
| 20 | —($CH_2(COO)_2$ $(CHCl)_3$)— | 1.8498 | 1.8506 | −0.04 | 0.0142 | 317.98 | 1.6511 |
| | Target molecule properties: | | | | **0.005** | **383.0** | **1.500** |

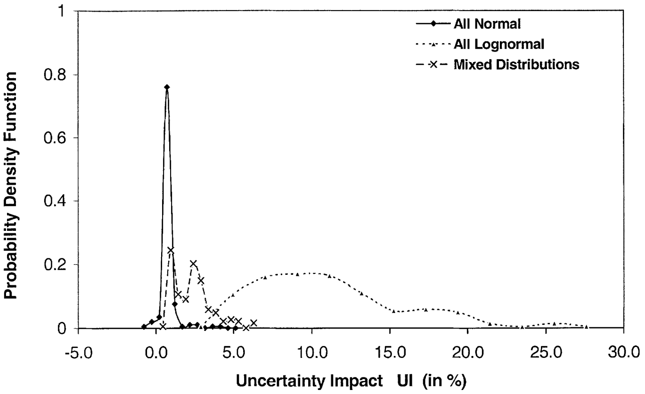

Figure 7. Sensitivity analysis of distribution type: PDF for uncertainty impact.

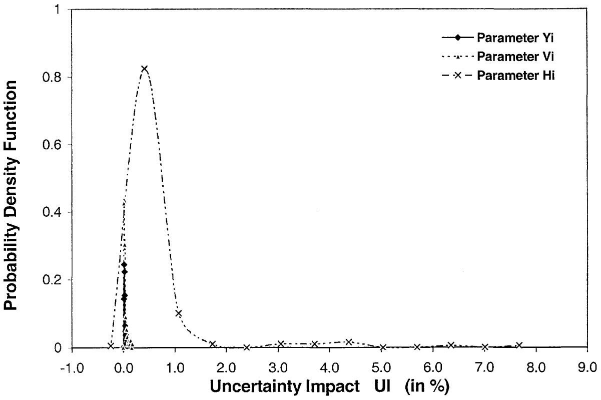

Figure 8. Sensitivity analysis of model parameters: PDF for uncertainty impact.

Gaussian fitness function $F_2$, where we can govern the steepness of the search space topology.

Case 1 was rerun with $F_2$ ($\lambda = 0.003$) as the objective function. Similar results were obtained and as shown in Table 10, the CPU time was significantly reduced from 85.3 to 56.1 s, leading to 34.2% computational savings. These savings were mainly due to a sharp decrease in the number of samples

Table 9. Stochastic Solution for Best 20 Polymer Designs (Case 6: Uncertainty only in $H_i$)

| Rank | Repeating Monomer | $F_{stoch}$ | $F_{det}$ | $UI$ (%) | Deterministic Properties $W$ | $T$ (K) | $D$ |
|---|---|---|---|---|---|---|---|
| 1 | —$(CH_2(CHCl)_3)$— | 0.0753 | 0.0405 | 85.68 | 0.0051 | 393.10 | 1.5351 |
| 2 | —$((CH_2)_2(CHCl)_3)$— | 0.0912 | 0.0708 | 28.80 | 0.0047 | 376.95 | 1.4489 |
| 3 | —$(CH_2CHCl)$— | 0.1684 | 0.1685 | −0.03 | 0.0043 | 363.20 | 1.3827 |
| 4 | —$(CHCl)$— | 0.1836 | 0.1750 | 4.91 | 0.0056 | 412.37 | 1.6525 |
| 5 | —$((CH_2)_3(CHCl)_2)$— | 0.2749 | 0.2721 | 1.00 | 0.0039 | 346.04 | 1.3082 |
| 6 | —$((O)(CHCl)_3)$— | 0.4609 | 0.4521 | 1.93 | 0.0072 | 396.28 | 1.6471 |
| 7 | —$(CH_2(O)_2(CHCl)_3)$— | 0.6239 | 0.6019 | 3.66 | 0.0080 | 369.19 | 1.5456 |
| 8 | —$((CH_2)_3(COO)\,(CHCl)_2)$— | 1.0878 | 1.0801 | 0.71 | 0.0103 | 306.56 | 1.4159 |
| 9 | —$(CH_2(COO)\,(CHCl)_3)$— | 1.1263 | 1.1260 | 0.03 | 0.0106 | 347.42 | 1.6036 |
| 10 | —$((CH_2)_3(COO)\,(O)\,(CHCl)_2)$— | 1.2862 | 1.2785 | 0.60 | 0.0113 | 302.01 | 1.4291 |
| 11 | —$((COO)\,(CHCl)_3)$— | 1.2914 | 1.2889 | 0.19 | 0.0114 | 358.84 | 1.7064 |
| 12 | —$((CH_2)_3(COO)\,(O)_3\,(CHCl)_3)$— | 1.3320 | 1.3319 | 0.01 | 0.0116 | 315.21 | 1.4820 |
| 13 | —$((CH_2)_2(COO)\,(O)_2\,(CHCl)_3)$— | 1.3666 | 1.3086 | 4.43 | 0.0115 | 326.25 | 1.5330 |
| 14 | —$(CH_2(COO)\,(CHCl)_2)$— | 1.4445 | 1.4486 | −0.28 | 0.0122 | 327.10 | 1.5889 |
| 15 | —$((CH_2)_3(COO)\,CHCl)$— | 1.4479 | 1.4548 | −0.48 | 0.0121 | 268.40 | 1.3463 |
| 16 | —$((CH_2)_2(COO)\,(O)_3\,(CHCl)_3)$— | 1.4505 | 1.4491 | 0.10 | 0.0122 | 321.66 | 1.5369 |
| 17 | —$(CH_2(COO)\,(O)_3\,CHCl)$— | 1.4583 | 1.4514 | 0.48 | 0.0122 | 314.03 | 1.4694 |
| 18 | —$((CH_2)_2(COO)\,(O)_2)$— | 1.4851 | 1.4782 | 0.47 | 0.0120 | 223.33 | 1.1605 |
| 19 | —$((CH_2)_2(COO)\,(O)_3\,(CHCl)_3)$— | 1.4997 | 1.4491 | 3.49 | 0.0122 | 321.66 | 1.5369 |
| 20 | —$(CH_2(COO)\,(CHCl)_2)$— | 1.6926 | 1.6904 | 0.13 | 0.0134 | 340.43 | 1.7258 |
| | Target molecule properties: | | | | **0.005** | **383.0** | **1.500** |

Table 10. CPU Time Comparisons for Different Case Runs

| Case No. | Property Targets $W$ | $T$ (K) | $D$ | No. of Groups | Objective Function | Parameter Uncertainties | Sampling | CPU (s) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.005 | 383 | 1.50 | 7 | $F_1$ | All Normal Distributions | HSS | 85.3 |
| 2 | 0.005 | 383 | 1.50 | 7 | $F_1$ | **All Lognormal Distributions** | HSS | 95.2 |
| 3 | 0.005 | 383 | 1.50 | 7 | $F_1$ | **Mixed Distributions** | HSS | 88.1 |
| 4 | 0.005 | 383 | 1.50 | 7 | $F_1$ | **Uncertainty only in $Y_i$ (Normal)** | **LHS** | 113.8 |
| 5 | 0.005 | 383 | 1.50 | 7 | $F_1$ | **Uncertainty only in $V_i$ (Normal)** | **LHS** | 64.3 |
| 6 | 0.005 | 383 | 1.50 | 7 | $F_1$ | **Uncertainty only in $H_i$ (Normal)** | **LHS** | 83.9 |
| 7 | 0.005 | 383 | 1.50 | 7 | $\boldsymbol{F_2}$ | All Normal Distributions | HSS | 56.1 |
| 8 | 0.005 | 383 | 1.50 | **5** | $F_1$ | All Normal Distributions | HSS | 12.1 |
| 9 | **0.150** | 383 | 1.50 | 7 | $F_1$ | All Normal Distributions | HSS | 75.9 |

*Note:* $F_1$: Root mean squared deviation; $F_2$: Gaussian fitness function; HSS: Hammersley sequence sampling; LHS: Latin hypercube sampling.

required to represent uncertainty at each temperature level in the annealing schedule. This was possible because of the self-adjusting sample-size feature of the proposed STA-NLP algorithm. Furthermore, a steeper search-space topology aids faster annealing as the algorithm proceeds. This highlights the importance of the appropriate choice of objective function and how it can affect the speed of convergence of the algorithm. Key advantages associated with such complex, typically nonlinear objective functions can be easily leveraged in the proposed framework, which does not impose any restrictions on the type of objective function to be used.

### *Heuristic combined with stochastic optimization (case 8)*

An important observation from the results of cases 1–7 is that the best 200 molecules obtained in all these cases did not contain either the —CONH— or the —CHOH— group. Also, out of these 200 molecules, the top 20 obtained did not

**Table 11. Stochastic Solution for Best 20 Polymer Designs (Case 8: Heuristic Combined with Stochastic Optimization)**

| | Repeating | | | *UI* | Deterministic Properties | | |
|---|---|---|---|---|---|---|---|
| Rank | Monomer | $F_{\text{stoch}}$ | $F_{\text{det}}$ | (%) | *W* | *T* (K) | *D* |
| 1 | —$(CH_2(CHCl)_2)$— | 0.1213 | 0.0218 | 456.96 | 0.0049 | 384.68 | 1.4889 |
| 2 | —$(CH_2(CHCl)_3)$— | 0.1282 | 0.0405 | 216.29 | 0.0051 | 393.10 | 1.5351 |
| 3 | —$(CH_2CHCl)$— | 0.1905 | 0.1685 | 13.06 | 0.0043 | 363.20 | 1.3827 |
| 4 | —$(CHCl)$— | 0.2142 | 0.1750 | 22.41 | 0.0056 | 412.37 | 1.6525 |
| 5 | —$((CH_2)_2(O)(CHCl)_3)$— | 0.2667 | 0.2454 | 8.68 | 0.0062 | 366.23 | 1.4605 |
| 6 | —$((CH_2)_3\ (CHCl)_2)$— | 0.2849 | 0.2721 | 4.67 | 0.0039 | 346.04 | 1.3082 |
| 7 | —$((CH_2)\ (O)\ (CHCl)_3)$— | 0.3551 | 0.3412 | 4.07 | 0.0067 | 380.06 | 1.5408 |
| 8 | —$((CH_2)_2\ (O)\ CHCl)$— | 0.4356 | 0.4195 | 3.83 | 0.0068 | 317.84 | 1.3019 |
| 9 | —$((O)(CHCl)_3)$— | 0.4809 | 0.4521 | 6.36 | 0.0072 | 396.28 | 1.6471 |
| 10 | —$(CH_2\ (O)\ CHCl)$— | 0.6293 | 0.6126 | 2.73 | 0.008 | 340.13 | 1.4221 |
| 11 | —$((CH_2)_2\ (O)_3\ (CHCl)_3)$— | 0.7243 | 0.7056 | 2.65 | 0.0085 | 349.44 | 1.4791 |
| 12 | —$(CH_2\ (O)_3\ (CHCl)_3)$— | 0.8375 | 0.8229 | 1.78 | 0.0091 | 360.00 | 1.5497 |
| 13 | —$((CH_2)_2\ (O)_2CHCl)$— | 0.8649 | 0.8500 | 1.76 | 0.0091 | 307.83 | 1.3387 |
| 14 | —$((CH_2)_3(COO)\ (CHCl)_3)$— | 0.8888 | 0.8720 | 1.93 | 0.0093 | 328.73 | 1.4596 |
| 15 | —$((CH_2)_2\ (O)_3\ (CHCl)_2)$— | 0.8947 | 0.8911 | 0.41 | 0.0094 | 331.79 | 1.4369 |
| 16 | —$((O)CHCl)$— | 0.9678 | 0.9649 | 0.30 | 0.0098 | 372.09 | 1.6391 |
| 17 | —$((CH_2)_2(COO)\ (CHCl)_3)$— | 1.0034 | 0.9873 | 1.63 | 0.0099 | 337.47 | 1.5236 |
| 18 | —$((CH_2)_3(COO)\ (O)\ (CHCl)_3)$— | 1.0599 | 1.0517 | 0.78 | 0.0102 | 323.64 | 1.468 |
| 19 | —$((CH_2)_3(COO)\ (CHCl)_2)$— | 1.0973 | 1.0801 | 1.59 | 0.0103 | 306.56 | 1.4159 |
| 20 | —$((CH_2)_2\ (O)_3\ CHCl)$— | 1.2026 | 1.2028 | −0.02 | 0.0109 | 300.40 | 1.3674 |
| | Target molecule properties: | | | | **0.005** | **383.0** | **1.500** |

**Table 12. Stochastic Solution for Best 20 Polymer Designs (Case 9: Effect of Target Properties)**

| | Repeating | | | *UI* | Deterministic Properties | | |
|---|---|---|---|---|---|---|---|
| Rank | Monomer | $F_{\text{stoch}}$ | $F_{\text{det}}$ | (%) | *W* | *T* (K) | *D* |
| 1 | —$((CH_2)_3(CO)(COO)(O)_3\ (CHOH)_3(CHCl))$— | 0.1212 | 0.0093 | 1210.01 | 0.1504 | 379.70 | 1.4969 |
| 2 | —$((CH_2)_3(CO)\ (CONH)_3(CHOH)(CHCl)_3)$— | 0.1258 | 0.0304 | 314.10 | 0.1516 | 384.78 | 1.5421 |
| 3 | —$((CH_2)_3(CO)(COO)\ (O)_3(CONH)(CHOH)_3(CHCl)_2)$— | 0.1258 | 0.0367 | 242.73 | 0.1504 | 372.70 | 1.5373 |
| 4 | —$((CH_2)_3(CO)\ (O)(CONH)_2(CHOH)_2(CHCl)_3)$— | 0.1282 | 0.0375 | 242.09 | 0.1514 | 394.97 | 1.5277 |
| 5 | —$((CH_2)_3(CO)\ (O)_3(CONH)_2(CHOH)(CHCl))$— | 0.1288 | 0.0516 | 149.66 | 0.1552 | 368.50 | 1.4927 |
| 6 | —$((CH_2)_3(CO)\ (O)_3(CONH)_2(CHOH)_2(CHCl)_3)$— | 0.1321 | 0.0621 | 112.73 | 0.1413 | 383.64 | 1.5331 |
| 7 | —$((CH_2)_3(CO)(COO)\ (O)(CONH)(CHOH)_2(CHCl))$— | 0.1341 | 0.0587 | 128.37 | 0.1579 | 373.36 | 1.5094 |
| 8 | —$((CH_2)_3(CO)_2(COO)_2\ (O)_3(CONH)(CHOH)_3(CHCl))$— | 0.1358 | 0.0586 | 131.68 | 0.1492 | 387.73 | 1.5856 |
| 9 | —$((CH_2)_3(CO)(COO)\ (O)_2(CONH)_2(CHOH)_3(CHCl)_3)$— | 0.1361 | 0.0592 | 129.82 | 0.1548 | 372.41 | 1.5622 |
| 10 | —$((CH_2)_3(CO)\ (O)_3(CONH)_3(CHOH)_2(CHCl)_3)$— | 0.1381 | 0.0657 | 110.05 | 0.1577 | 373.70 | 1.5497 |
| 11 | —$((CH_2)_3(CO)_2(COO)_2(O)_2(CONH)_2(CHOH)_2(CHCl))$— | 0.1394 | 0.0679 | 105.36 | 0.1495 | 378.42 | 1.6001 |
| 12 | —$((CH_2)_3(CO)(COO)\ (O)\ (CHOH)_3(CHCl)_2)$— | 0.1395 | 0.0710 | 96.43 | 0.1411 | 397.79 | 1.5088 |
| 13 | —$((CH_2)_2(CO)(COO)\ (O)_3(CONH)(CHOH)_2(CHCl))$— | 0.1414 | 0.0607 | 133.06 | 0.1509 | 368.61 | 1.5709 |
| 14 | —$((CH_2)_3(CO)\ (O)_2(CONH)_3(CHOH)(CHCl)_2)$— | 0.1415 | 0.0786 | 79.99 | 0.1599 | 368.99 | 1.5332 |
| 15 | —$((CH_2)_3(CO)_2(COO)_2\ (O)(CONH)(CHOH)_2)$— | 0.1422 | 0.0696 | 104.34 | 0.1558 | 393.77 | 1.5758 |
| 16 | —$((CH_2)_2(CO)\ (O)_2(CONH)(CHOH)(CHCl))$— | 0.1435 | 0.0795 | 80.42 | 0.1431 | 407.64 | 1.5126 |
| 17 | —$((CH_2)_3(CO)\ (O)(CONH)(CHOH)(CHCl))$— | 0.1438 | 0.0866 | 66.14 | 0.1427 | 405.30 | 1.4375 |
| 18 | —$((CH_2)_3(CO)(COO)(O)(CONH)_2(CHOH)_2(CHCl)_3)$— | 0.1447 | 0.0890 | 62.53 | 0.1388 | 372.72 | 1.5605 |
| 19 | —$((CH_2)_2(CO)\ (O)_2(CONH)(CHOH)(CHCl))$— | 0.1447 | 0.0795 | 81.99 | 0.1431 | 407.64 | 1.5126 |
| 20 | —$((CH_2)_3(CO)_2(COO)(CONH)_3(CHOH)(CHCl)_2)$— | 0.1450 | 0.0757 | 91.50 | 0.1504 | 399.75 | 1.5926 |
| | Target molecule properties: | | | | **0.150** | **383.0** | **1.500** |

contain any of the —CO— and —COO— groups. This implies that these groups are not desired for the targeted properties. This is further verified by the fact that all these groups are highly polar among the pool of molecular groups considered. Thus, these groups are not suitable for the very low water-absorption capacity of 0.005 g $H_2O$/g polymer, which was one of the target properties.

When this was known, case 1 was rerun with a pool of only five molecular groups—highly polar groups —CONH— and —CHOH— were removed—all with the same property targets. This reduced the dimensionality of the search space drastically. The stochastic annealing algorithm parameters were tuned for such reduced search space. The modified parameter values are shown in Table 3. Similar optimal designs were obtained and CPU time was drastically reduced by a factor of 7 from 85.3 to 12.1 s, leading to 85.8% computational savings, as can be seen in Table 10. Thus by combining heuristics and other knowledge into the CAMD framework, we can not only gain crucial insight into the problem but can also reduce the computational burden significantly. Such feedback can be appropriately used in the proposed framework due to the inherent flexibility it offers.

The best 20 optimal molecular designs are tabulated in Table 11. No change in relative ranking of molecules was observed when compared to the deterministic solutions given in Table 2.

### *Effect of target properties (case 9)*

All of cases 1–8 targeted $\boldsymbol{D_o} = 1.50$ g/cm$^3$, $\boldsymbol{W_o} = 0.005$ g $H_2O$/g polymer, and $\boldsymbol{T_{go}} = 383$ K. The water-absorption capacity targeted was very low and is quite rare, as can be seen from the results of exhaustive search in Figure 3. In all these cases, the complete absence of highly polar groups, such as —CHOH— and —CONH—, was observed in the best 200 molecules, as was expected. In addition, molecular group —CO— was absent in the top 20, and —COO— was absent in the top 10 molecules. This can be explained by the fact that all these molecules are polar groups, and thus are not desired for such a low $\boldsymbol{W}$, because polar groups tend to increase the water-absorption capacity of the molecule. Thus a higher $H_i$ value for polar groups reflects an analogy between the two. In fact, in Table 1 the groups in decreasing $H_i$ values or equivalently of decreasing polarity are —CONH—, —CHOH—, —CO—, —COO—, —O—, —CHCl—, and —$CH_2$—. So, case 9 was conducted to show the correctness of the algorithm and reconfirm gained insights about case study 1. Thus, in this case study, case 1 was rerun with a very high water-absorption capacity of 0.150 g $H_2O$/g polymer as the target instead of 0.005 g $H_2O$/g polymer, keeping other property targets the same as originally in case 1.

Table 12 shows the best 20 optimal molecular designs obtained. Clearly all the polar groups absent in the earlier cases, appear in these designs. With the increased water-absorption capacity as the target, all the highly polar molecules appear in the optimal molecular designs. Figure 10 shows the best 200 molecular designs obtained. Compared to Figure 6 earlier, all the columns corresponding to polar groups—CONH —, —CHOH—, —CO—, and —COO— are filled here in Figure 10. The optimal molecules invariably have most of the seven distinct groups present. This reaffirms the belief that highly polar groups are required for higher water-absorption capacity and that the $H_i$ values are analogous to the polarity of the molecular groups. Additionally, the best 12 molecular designs had 3 units of —($CH_2$)—, and the best 8 molecular designs had 1 unit of —(CO)—. With these kind of data visualization of the best 200 (or even more) molecular designs, it is possible to find hidden trends. This kind of interactive learning, as we solve the problem, can be very informative for a larger real-life CAMD problem.

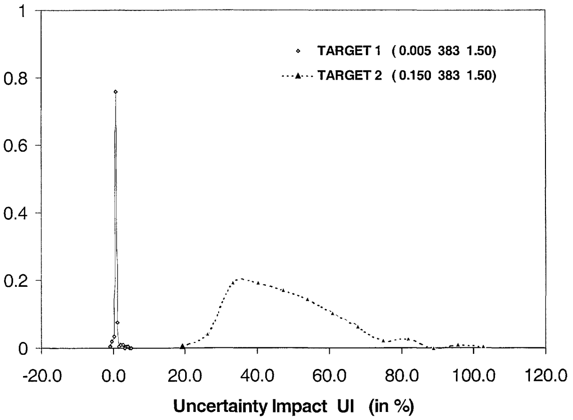

**Figure 9. Sensitivity analysis of target properties: PDF for uncertainty impact.**

Also Figure 9 compares the probability distribution function of the UI values of the top 200 molecules obtained in this case with the standard case (case 1). While UI values where in the 0–5% range in case 1, they jumped to 20–100% for this case, with modified target properties. Thus, for rarer property targets, the uncertainty impacts are found to be low, but for a more probable property target, like $\boldsymbol{W_o} = 0.150$ g $H_2O$/g polymer uncertainty impacts are quite high. In addition, comparing the best 20 molecular designs for case 1 and case 9 in Table 4 and Table 12, respectively, an important observation is made: Optimal molecules for case 1 have one to four distinct groups, while for case 9 they invariably have all seven distinct functional groups present. This observed phenomenon is explained by the fact that GCM methods can have a very large number of *local solutions*. This means that, with a 100% change in the parameter values, the same property value can be obtained while after only a 5% change, a very different property value can be obtained. This is especially likely for those molecules that have more than two distinct types of groups. Such cases have a higher possibility of having multiple local solutions of the GCM models. As was observed in this case, the optimal molecules indeed always have more than five distinct types of groups. The model parametric uncertainty is therefore more prominent in this case. The proposed optimization framework can efficiently handle such multiple local minima traps of the GCM models.

### *Overall results, observations, and comparisons with other approaches*

Several interesting and insightful observations obtained from these nine case studies are listed below.

*1. More Uncertainty Impact on the Optimal Than Suboptimal Solutions.* In all nine case studies, optimal solutions, as opposed to the lower ranked molecular designs, invariably exhibited very high UI values, ranging from 6.33% in case 4 to as high as 1210% in case 9. Similar behavior was also observed in a recent study on CAMD under uncertainty (Maranas, 1997), where chance-constrained optimization under a deterministic framework was used to get trade-off curves for the optimal solutions. In that case it also was observed that random perturbations around the mean value of group contribution parameters have a more prominent effect on the optimal rather than the suboptimal designs. This is due to the way the objective function is defined. Optimal solutions have a very small $F$ value, and the optimum solution occurs at an $F$ value of zero. Thus, even slight fluctuations in the parameter values transform into high UI values for the optimal solutions. In that sense, the optimal molecular design is fine-tuned to the effect of uncertainties.

*2. Effect of Uncertainty Was to Penalize the Model Property Predictions.* In all nine case studies, the UI values obtained were positive, barring a few exceptions, as $F_{\text{stoch}}$ was invariably greater than $F_{\text{det}}$. This means that statistically uncertainties increased the weighted distance of the molecule's properties from the target properties. This increase was much larger for optimal than for suboptimal solutions, as was also mentioned in observation 1. Thus, the presence of uncertainties can significantly affect the choice of optimal molecular designs. This again agrees with a Maranas's study (1997), though in a deterministic framework. There it also was observed that the effect of property-prediction uncertainty was to penalize the deterministic property predictions.

*3. Factors Affecting Uncertainty Impact.* In the proposed framework, property-prediction uncertainty is reflected in the UI values. Surprisingly, the magnitude of UI, besides depending on the (1) uncertainty distribution type, (2) level of uncertainty (that is, variance around the mean parameter values), and (3) the objective function type, also seems to be affected by the (4) problem size and the (5) property targets. This can be seen by comparing the results in Tables 4, 11, and 12.

*4. Effect of Uncertainty Representation on the Optimal Design Predictions.* The lognormal distribution, which is a more realistic representation of the GCM uncertainty parameter, has a dramatic effect on the optimal designs. It changed the relative ranking of the optimal molecular designs obtained from the deterministic approach. The UI values were higher with lognormal distribution than with normal and mixed distributions. Thus the impact of the uncertainties can go beyond slight changes in ordering between, say, the best five molecules. Totally new molecules might reappear and some optimal molecules might vanish, as compared to the deterministic solutions.

*5. Comparison of Stochastic Solutions with Deterministic/Exhaustive Solutions.* Tables 4–9 and Table 11 tabulate stochastic solutions for the best 20 molecules for the various cases. On comparing with the deterministic solutions in Table 2, it can be seen that roughly 5–10 of the top 20 molecules from the deterministic solutions appear in the stochastic solutions. Their relative ranking was found to be markedly different from the lognormal distributions. Thus the algorithm yields promising molecules close to the target, and was able to obtain optimal molecules in the 99th percentile of the deterministic optimal solutions.

*6. Not only the objective function type but also the target properties together govern the search space topology.* Thus, in turn, they also govern the:

(a) *Optimal solutions*, which are essentially the peak points of this search space topology.

(b) *Speed of convergence* of the algorithm. As was also seen in case 8, where CPU time was reduced by 34% for the Gaussian fitness function as the objective function ($\lambda = 0.003$).

(c) *Sensitivity Analysis.*

*7. Sensitivity Analysis of the Model Parameters.* From Figure 9 the model parameter $H_i$ was found to be the most sensitive parameter and had a much higher level of sensitivity than both $Y_i$ and $V_i$. Parameters $Y_i$ and $V_i$ have similar levels of sensitivity, with $Y_i$ slightly more sensitive than $V_i$. Of course, this analysis is specific to the variance values and target properties chosen in these case studies.

*8. Confidence in the Model.* The target water absorption capacity $W$ was increased from the very low value of 0.005 g $H_2O$/g polymer in case 1 to 0.150 g $H_2O$/g polymer in case 9. It was observed that all the highly polar groups, such as —CONH— and —CHOH—, that were absent in the optimal molecular designs of case 1, appear in the optimal molecular designs for the latter case. This also builds *more confidence in the model*, which has been implicitly able to capture the physical behavior of the molecules in the form of GCM parameters.

*9. Good Molecular Designs that Stand the Test of Uncertainty.* Among others, the most promising molecular designs that could stand the test of various types of uncertainties are —($CH_2CHCl$)— and —($CH_2(CHCl)_3$)—. These correspond to the property targets: $\boldsymbol{D_o} = 1.50$ g/cm$^3$, $\boldsymbol{W_o} = 0.005$ g $H_2O$/g polymer, and $\boldsymbol{T_{go}} = 383$ K.

*10. Pool of Promising 200 Molecules in Much Less CPU Time.* The proposed framework does not guarantee global optimality, however, but can provide quite a promising pool of 200 molecules in significantly less CPU time, when considering uncertainty. However, to obtain the best 200 molecules in the chance-constrained optimization framework, 200 runs of the MINLP algorithm have to be performed with increasing constraints. A quadratic raise in cumulative CPU requirements has been reported (Maranas, 1996) for a similar case study, also without uncertainty. Enormous CPU time would be required to draw a pareto set from the trade-off curves of the best 200 molecules, with the same uncertainty approach. Figures 6 and 10 show the best 200 molecules obtained through stochastic optimization in case 1 and case 9. This can make sensitivity analysis and trend searching possible. Knowing this, we can further reduce the dimensionality of the problem, as also was demonstrated in this article. From a design point of view, we have a set of highly promising alternative structures to synthesize, with an associated UI value corresponding to each design.

*11. Reduced Computational Burden.* The main bottleneck with the sampling approach to stochastic optimization under uncertainty is the large number of computations required in drawing the samples representing the parameters' uncertainty distribution domain. In this study we have demonstrated that highly efficient sampling techniques like HSS require a great many fewer samples to represent the uncertainty domain. This drastically reduces the CPU time. Additionally, an appropriate choice of the objective function, even

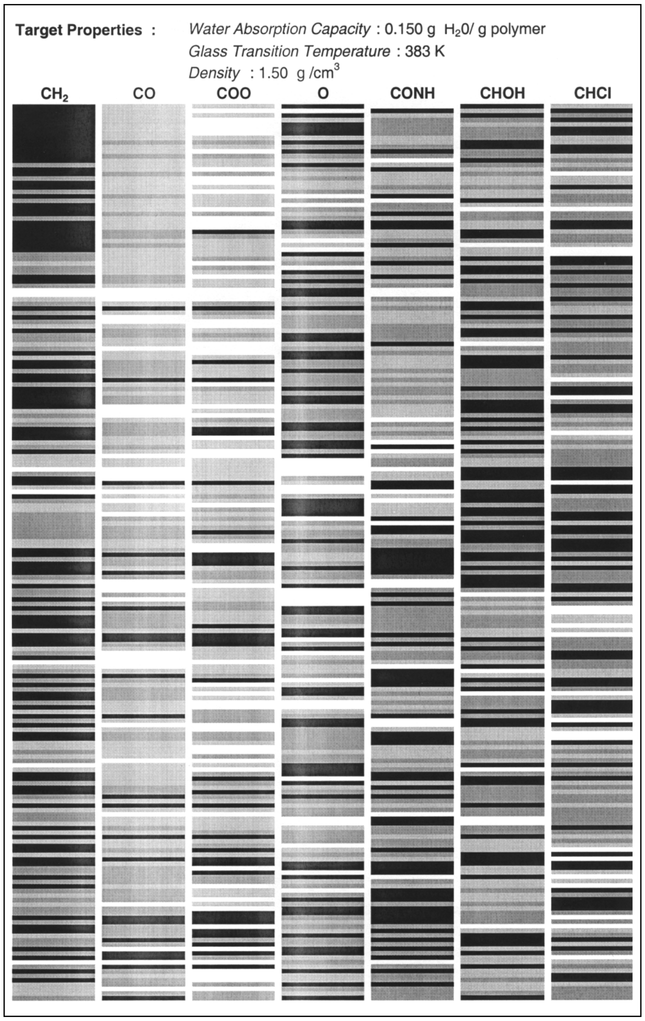

**Figure 10. Best 200 molecules: stochastic solution (different target properties).**

if it is nonlinear, governs the speed of convergence of the algorithm, and can save additional CPU time. Thus another advantage and the flexibility of the proposed approach can be exploited for more challenging and complex real-life problems.

*12. Flexibility and Versatility of the Approach.* The proposed framework does not impose any convexity restrictions on the type of objective function, constraints, and the structure−property model itself. It can also be applied to *nonlinear* objective functions, constraints, and structure−property relations. In addition, it can take several types of distributions for uncertainty representations besides *stable*, including *unstable*, and *mixture of distributions*. It can also be extended for user-defined and correlated distributions (useful for correlated parameter uncertainties). Although little information often exists on the type of uncertainty in the model parameters, this article highlights the flexibility and high versatility associated with the proposed approach. This approach can widen applicability to highly diverse and complex real-life CAMD problems under uncertainty.

*13. Comparison with Chance Constrained and Other Nonstochastic Approaches.* As opposed to the chance-constrained optimization approach, the proposed approach is highly flexible and can be applied to more complex nonlinear objective functions, constraints, and property-prediction models. The advantage with chance-constrained optimization is that it can guarantee global optimality, but that highly restricts its applicability to a few select cases where convexity can be guaranteed. These include linear constraints with variables $n_i$ and stable distribution for independent variables, among other restrictions.

*14. Applicability to Implicit and Statistical Mechanics Models.* This approach can be extended to highly nonlinear GCM models or even to more complicated models like those from the statistical mechanics model, which acts as a black box. Furthermore, several commercial packages, such as CRANIUM, are being developed for property predictions that use various embedded structure−property models for predictions. These packages have their own decision-tree networks to decide on the best prediction models specific to the molecule searched. Alternatively, this choice of appropriate prediction model might also rest with the user. Thus, different models often might be used for various property predictions at various levels of the tree, depending on the molecule type, physical property, or other user specifications. This creates a dynamic variant model environment. The proposed technique easily can be coupled with such packages, as *it requires the predicted values from the model* but *not the model itself*.

*15. Gain Insights.* As shown in case 8, this exercise can give crucial insights for the problem at hand. It can remove redundancies and focus research on the critical search space. One quite easily can integrate any heuristic knowledge one might gather or has about the problem, into the proposed stochastic framework, to speed up the design process. This knowledge can help reformulate the problem with fewer choices and reduce the dimensionality of the problem, in effect drastically reducing the CPU search time. Trend searching and pattern recognition can then be performed by careful observation of the best 200 (or even more) molecules for a given target. This can further assist chemists at the intuitive level. It also can highlight the most sensitive model parameter uncertainties that future efforts should try to reduce.

## Conclusions

In this article, we present a novel sampling approach to stochastic optimization for optimal molecular design under property-prediction uncertainty. We address quantifying the effect of property-prediction imprecision within a stochastic framework and study its effect on the choice of optimal molecular designs. The flexibility of this approach, coupled with the highly reduced computational burden, due to novel sampling techniques, such as Hammersley sequence sampling (HSS), makes this approach a very important tool for uncertainty analysis of optimal molecular-design problems of various complexities. Such a generalized framework makes it widely applicable to real-life CAMD problems of varied complexities, including (1) nonlinear or even black-box property-prediction models, (2) nonlinear objective function and constraints, and (3) nonnormal distribution functions for the uncertain parameters. Further, a parallel is drawn between the deterministic MINLP approach using chance-constrained formulation (Maranas, 1997), and the proposed approach of stochastic optimization using sampling. These two approaches represent two different ways of incorporating uncertainty in the optimization framework for the optimal molecular-design problem.

The article focuses on the Stochastic property matching (SPM) problem in the case of polymer design. Widely applicable group contribution methods (GCM) for property prediction of polymers have been employed, taking the model parameter values from the literature. Results from nine cases amply demonstrate that the proposed framework is capable of a detailed uncertainty analysis. These results highlight that uncertainty in model predictions can significantly affect the choice of optimal molecular designs. This study stresses the importance of (1) appropriate uncertainty representation, and (2) objective function formulation. Sensitivity analysis of the GCM uncertainty parameter was also made possible for the first time in the optimal molecular-design problem. This points to the crucial parameter uncertainties, which cause the biggest impact. Future research can then focus on reducing such uncertainties. Useful insights and conclusions drawn from these studies are listed in detail in the previous section.

A large pool of 200 (or even more) promising molecules can be obtained by such an approach for the given property targets, with a UI value associated with each polymer design. It also hints at the possibility of using such a generalized approach to conduct objective-specific trend searching. Useful insight into the problem can be gained that can fuel the chemist's intuition, making the search even faster.

## Acknowledgment

This work was partially supported by NSF CTS no. 9613561. We greatly acknowledge useful discussions with Dr. Kevin Joback from Molecular Knowledge Systems, Inc.

## Literature Cited

Billingsley, P., *Probability and Measure*, 3rd ed., Wiley, New York (1995).

Breiman, L., *Probability*, Addison-Wesley, Reading, MA, p. 84 (1968).

Birge, J. R., "Stochastic Programming Computation and Applications," *INFORMS J. of Comput.*, **9**, 111 (1997).

Chaudhuri, P. D., and U. M. Diwekar, "Process Synthesis Under Uncertainty: A Penalty Function Approach," *AIChE J.*, **42**, 742 (1996).

Chaudhuri, P. D., and U. M. Diwekar, "Synthesis Approach to the Determination of Optimal Waste Blends under Uncertainty," *AIChE J.*, **45**, 1671 (1999).

Charnes, A., and W. W. Cooper, "Chance Constrained Programming," *Management Sci.*, **6**, 73 (1959).

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

Diwekar, U. M., and J. R. Kalagnanam, "Efficient Sampling Technique for Optimization Under Uncertainty," *AIChE J.*, **43**, 440 (1997).

Duvedi, A., and L. E. K. Achenie, "On the Design of Environmentally Benign Refrigerant Mixtures: A Mathematical Programming Approach," *Comput. Chem. Eng.*, **21**, 915 (1997).

Friedler, F., L. T. Fan, L. Kalotai, and A. Dallos, "A Combinatorial Approach for Generating Candidate Molecules with Desired Properties on Group Contribution," *Comput. Chem. Eng.*, **22**, 809 (1998).

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

Joback, K. G., *Designing Molecules Possessing Desired Physical Properties*, PhD Diss., MIT, Cambridge, MA (1989).

Joback, K. G., and G. Stephanopoulos, "Searching Spaces of Discrete Solutions: The Design of Molecules Possessing Desired Physical Properties," *Adv. Chem. Eng.*, **21**, 257 (1995).

Kottegoda, N. T., and R. Rosso, *Probability, Statistics, and Reliability for Civil and Environmental Engineers*, McGraw-Hill, New York (1997).

Maranas, C. D., "Optimal Molecular Design Under Property Prediction Uncertainty," *AIChE J.*, **43**, 1250 (1997).

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

Mavrovouniotis, M. L., "Product and Process Design with Molecular-Level Knowledge," *Proc. Int. Conf. on Intelligent Systems in Process Engineering*, AIChE Symp. Series, J. F. Davis, G. Stephanopoulos, and V. Venkatasubramanian, eds., AIChE, New York (1996).

Morgan, G. M., and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge Univ. Press, Cambridge (1990).

Painton, L., and U. M. Diwekar, "Stochastic Annealing for Synthesis Under Uncertainty," *Eur. J. of Oper. Res.*, **83**, 489 (1995).

Sudjianto, A., L. Juneja, H. Agarwal, and M. Vora, "Computer Aided Reliability and Robustness Assessment," *Proc. Int. Conf. on Quality and Reliability*, Vol. 2, The Hong Kong Polytechnic University, Hong Kong, p. 277 (Sept. 1–3, 1997).

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).

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

Venkatasubramanian, V., K. Chan, and J. M. Caruthers, "Computer-Aided Molecular Design Using Genetic Algorithms," *Comput. Chem. Eng.*, **18**, 833 (1994).

*Manuscript received Sept. 20, 1999, and revision received July 6, 2000.*