Optimization-Based Framework for Computer-Aided Molecular Design

Apurva P. Samudra and Nikolaos V. Sahinidis
Dept. of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213

DOI 10.1002/aic.14112
Published online April 25, 2013 in Wiley Online Library (wileyonlinelibrary.com)

A new framework to automate, augment, and accelerate steps in computer-aided molecular design is presented. The problem is tackled in three stages: (1) composition design, (2) structure determination, and (3) extended design. Composition identification and structure determination are decoupled to achieve computational efficiency. Using approximate group-contribution methods in the first stage, molecular compositions that fit design targets are identified. In the second stage, isomer structures of solution compositions are determined systematically, and structure-based property corrections are used to refine the solution pool. In the final stage, the design is extended beyond the scope of group-contribution methods by using problem-specific property models. At each design stage, novel optimization models and graph theoretic algorithms generate a large and diverse pool of candidates using an assortment of property models. The wide applicability and computational efficiency of the proposed methodology are illustrated through three case studies.
 AIChE J, 59: 3686–3701, 2013
Keywords: molecular design, CAMD, group contribution methods, isomer generation, integer optimization, product design, refrigerant design, solvent design

Introduction

The chemical industry produces a vast array of materials used in most aspects of human life. These products, which can be categorized as basic chemicals, pharmaceuticals, specialty chemicals, and consumer products, are continuously evolving. The growth of the chemical industry over the last century is in part due to the constant search for "novel" and "improved" products. In the context of the chemical industry, a product can be defined as any chemical entity that enables a device or process to perform a desired function. The product and its properties are, therefore, deeply intertwined with its function. For instance, chemical product-function pairs include

- solvents, such as hexane and toluene, to enable industrial separations;
- additives, such as benzoic acid and sodium nitrate, to increase shelf life of foodstuffs;
- cosmetics, such as talc and titanium dioxide, to enhance the appearance of human body;
- pharmaceuticals, such as amoxicillin and tetracycline, to combat microbial infections;
- fuels, such as gasoline and methane, to generate mechanical work.

Market demand has traditionally provided much of the impetus for designing new products. Additionally, over the last few decades, health and safety concerns along with environmental impact awareness have led to a real thrust for product innovation. As a result, regulatory factors combined with economic incentives have fueled research work on product design.

Product design is the process of identifying the best set of candidates that fit desired product functional criteria. Experimental trial-and-error design involves generating new formulations, guided by previous products or expected performance of new ingredients, and testing these formulations for design criteria. The search space that can be practically investigated using trial-and-error methods is limited by the effort, cost, and time required. Thus, many researchers have developed methods that combine property predictive models with computer-assisted search.

Computer-aided molecular design (CAMD) has emerged as a powerful technique to identify promising molecules that meet predefined property targets. Early work by Gani and coworkers[1] was based on the UNIFAC[2] group-contribution method for solvent design. Kier and Hall developed molecular design techniques based on connectivity indices.[3–5] A generic molecular design framework was developed by Joback and Stephanopoulos[6,7] with an aim to include many properties. The first explicit optimization formulation for molecular design was presented by Odele and Macchietto.[8] Over the last two decades, applications of CAMD have included the design of alternative refrigerants,[9–14] repeating units of polymers,[15–20] solvents and extraction agents,[21–35] ionic liquids,[36,37] fuels,[38,39] molecular switches,[40] and pharmaceutical compounds.[41–44]

Although techniques based on enumeration perform competitively for the design of small molecules from a small set of submolecular building blocks, optimization is necessary to explore the complete search space of larger design spaces. However, a molecular design problem that involves many and complex properties is difficult to solve as a single-

Correspondence concerning this article should be addressed to N. V. Sahinidis at sahinidis@cmu.edu.

optimization problem. The complexity of property prediction models along with the large molecular solution space can impose severe limitations on the solution methods and may lead to inferior designs. Decomposition methods have naturally been proposed in this context to solve complex molecular design problems as a series of relatively easier subproblems.[45,46] Such decomposition methods allow gradual property-based screening, whereby complex property techniques are used at later stages in the process.

In this article, we propose a comprehensive molecular design framework using novel optimization methods for molecule identification and structure generation. We decompose the CAMD problem based on the structure of the molecule, determining molecular constitution and structure separately. Graph optimization techniques are used to obtain unique molecular structures, distinguishing only their unique structural isomers and discarding redundant structures. The task of estimating design properties is also decomposed based on the nature of property models. In early stages, approximate property models for a subset of design properties are used to explore the chemical space. In later stages, corrections are added to the estimated properties based on the generated structures to boost estimation accuracy. Subsequently, more detailed property models are used to predict any remaining complex properties. The proposed framework covers a wide variety of CAMD problems and is easily adaptable to many different property prediction methods, including empirical correlations, group-contribution models, and complex prediction techniques such as simulations.

The primary contributions of this work are as follows:

1. A novel optimization formulation is proposed to exploit linearity in a suitable coordinate set of property prediction models. In contrast to mixed-integer nonlinear programming (MINLP) approaches to CAMD, we are able to use efficient mixed-integer linear programming algorithms that can handle large solution sets easily. As we demonstrate in this article, our linear formulation results in orders of magnitude reduction in computing time requirements as compared to an equivalent MINLP model. In comparison to previous linearization techniques,[47] our linearization is not limited to nonlinear fractional terms and requires no additional variables to be introduced.

2. We propose a number of novel optimization and graph models that lead to fast solution of subproblems in our decomposition approach. The approach includes a systematic method to identify isomers and a novel modeling technique to deal with redundancy in chemical space and identify unique solutions.

3. In comparison to previous decomposition approaches that relied on enumeration or stochastic optimization,[25,45] the use of deterministic optimization techniques at each stage of our approach leads to an efficient algorithm that is not limited by the number of solutions desired or the size of the molecule sought.

The remainder of this article is structured as follows. In the next section, we present a brief overview of the CAMD problem and current approaches. In the section that follows the overview, we detail the proposed framework, along with its objectives, constituent models, and implementation. Subsequently, we present case studies that illustrate the applicability of the framework. While studying three previously studied molecular design problems, we identify many new solutions. Finally, we provide conclusions from this work.

State-of-the-Art in CAMD

CAMD is the problem of identifying molecular structures that satisfy predefined property targets. The problem can be stated as follows:

> Given a set of building blocks, a number of models of varying complexity and accuracy that correlate these building blocks to properties, and a set of property targets to match; determine molecules formed by combinations of the building blocks that match the design targets.

The property model, along with its descriptors, forms the basis of the problem statement. Molecules are to be generated by combining building blocks in ways that avoid chemically infeasible combinations. The property targets can be viewed as design constraints that narrow the space of all possible combinations of building blocks. We next consider different aspects of this design problem in more detail.

Property models

Property models are computational tools developed to predict molecular properties from structural descriptors, which quantify molecular structure. Structural descriptors, such as chemical bonds and molecular geometry, are most commonly used. In addition, many techniques use descriptors not directly related to structure, including critical properties and partial charges. The descriptors used in the property model determine the quality and quantity of chemical information that can be exploited by the overall CAMD approach. The property models used by CAMD approaches influence the accuracy and validity of these methods as well as the numerical solution techniques they use.

There exist many models capable of predicting, with varying degrees of accuracy, properties of pure compounds, mixtures, and polymers. Mechanical models are based on atomic-level simulations using *ab initio* quantum and molecular mechanics. Although very accurate, these methods are computationally intensive. Semiempirical models are regression-based models that fit proposed model equations to a large database of molecular properties. These models are usually easy to implement, but are approximate in nature. Empirical models are based on purely empirical techniques, such as pattern matching, chemometrics, and quantitative structure–activity relationships. Although simple to develop and widely reported for many properties, empirical models are limited to small classes of molecules.[48]

Group-contribution models

An important class of semiempirical property models is developed from the idea of additive contributions of molecular fragments to properties. These methods use groups of atoms present in the molecule as descriptors. In the related literature, the terms groups and descriptors are used interchangeably. The properties of the molecule are then expressed as functions of the frequency of each descriptor in the molecule

$$X = f\left(\sum_i c_i n_i\right) \tag{1}$$

Equation 1 collectively represents typical group-contribution methods, where X is the value of a property to be determined, c_i is the contribution of group i to the property, and

n_i is the frequency of group i in the molecule. The model parameters c_i are determined by regression over a large database of molecular properties. The function f is chosen to provide a good fit to the experimental data.

Early group-contribution models, such as the one proposed by Joback and Reid,[49] assumed the groups to be independent and nonoverlapping basic atomic constructs, such as >CH— and =O. These groups do not account for interactions between descriptors in a molecular structure. As proximity effects also contribute to molecular properties, many efforts have been made to develop models that capture interactions between atoms/groups located close to each other. In addition to not being able to capture proximity effects, simple group-contribution methods are unable to differentiate between isomers. Attempts to improve prediction accuracy through conjugate forms based on valence electrons[50,51] and connectivity indices[18,52] have been reported. Other approaches[53,54] use second-order groups to capture the interactions between descriptors. Such methods retain the simplicity of basic group-contribution methods and are fairly accurate.

GC$^+$ property model

Marrero and Gani[55] proposed a three-level property prediction model to predict key molecular properties. In the Marrero–Gani (MG) model, descriptors are divided in first-order and higher-order groups. The first-order groups form nonoverlapping building blocks of the model, whereas overlapping higher-order groups capture proximity effects. The contributions to the property are divided in compositional (first-order) and structural (higher-order) terms. The functional form of the model is similar to simple group-contribution models

$$X = f\left(\sum_{i\in F} c_i n_i + \sum_{i\in S} c_i n_i + \sum_{i\in T} c_i n_i\right)$$

Here, F is the set of first-order groups, S is the set of second-order groups, and T is the set of third-order groups. The original MG model involves 182 first-order, 122 second-order, and 66 third-order groups. The large number of descriptors in the model leads to its high accuracy. Knowledge of the molecular composition, that is, knowledge of what descriptors are present in the molecule and their frequencies, is necessary and sufficient for first-order property predictions. The second- and third-order groups are combinations of first-order groups and can be determined from the bonds between different first-order groups once the structure is known.

The MG model was developed for important physical and critical properties, including boiling point, melting point, critical temperature, pressure, and volume. Group contributions for standard Gibbs energy, enthalpy of vaporization, enthalpy of fusion, and enthalpy of formation have been reported.[55] The original MG model has also been expanded to include additional properties such as octanol-water partition coefficient and aqueous solubility of molecular species,[56] heat capacity,[57,58] phase equilibria,[59] viscosity, and surface tension.[60] These models use slightly different basis sets of molecular descriptors than the original MG method. The MG method has been extended using connectivity indices to predict contributions of groups absent in the original model.[61] The original MG method, along with expanded properties and connectivity indices, is known as the GC$^+$ method.

Formalized CAMD problem

A mathematical abstraction of the CAMD problem can be formulated using group contribution methods and integer optimization. The molecule is constructed as a collection of molecular descriptors, represented by connected nodes in a molecular graph. Solutions to the CAMD problem, that is, candidate molecules, are represented by their composition and structure. The frequencies of descriptors/groups in the molecule characterize the composition and edges between nodes in the molecular graph model the structure. The following integer optimization model captures the essence of the problem

$$\begin{aligned} \max \quad & \phi(x,n,y) \\ \text{s.t.} \quad & g(x,n,y) \le 0 \\ & h(n,y) \le 0 \\ & x \in \mathbb{R}^m, n \in \mathbb{Z}^N, y \in \{0,1\}^{R\times R} \end{aligned}$$

In this above formulation, n is a vector of integer variables used to denote the composition of the molecule. In particular, $n_i,\ i \in \mathcal{N} = \{1,\ldots,N\}$, denotes the frequency of the i-th descriptor in the molecule, $\mathcal{N}$ denotes the set of descriptors used by the property model, and $N = |\mathcal{N}|$ is the total number of descriptors in the model. R is the total number of descriptors in the molecule. Binary variables y model the presence or absence of bonds between descriptor nodes. The vector of continuous variables x denotes the design properties. The constraint set $g(x,n,y)$ consists of the property model, along with bounds placed on the properties by design criteria. The objective $\phi(x,n,y)$ is a function of property targets and structural components of the molecule to be optimized. Finally, $h(n,y) \le 0$ represents structural feasibility constraints. These constraints eliminate combinations of descriptors that do not satisfy chemical bonding requirements. The nature of these constraints depends on the types of molecular descriptors used by the model. Most of these constraints are based on graph theory and are applicable across different descriptor sets used by different models. Many basic structural feasibility constraints were developed by Joback and Stephanopoulos,[7] including constraints that enforce structural integrity for acyclic bonds, cyclic rings, aromatic rings, and mixed descriptors, that is, descriptors with both acyclic and ring (cyclic or aromatic) bonds. Odele and Macchietto[8] and Sahinidis et al.[12] introduced additional structural feasibility constraints, which further reduce the combinatorial search space. For simplicity, the above formulation was stated in terms of predefined property bounds. It is easy to incorporate this model, though, in a larger model with variable property bounds for simultaneous product and process design.

Solution methods

The nature of the property model and structural feasibility constraints determine the most suitable numerical solution strategy. The major effort in solving a CAMD problem is directed toward: (1) generating combinations of descriptors defining compositions, and (2) identifying feasible structures from the chosen descriptor sets.

Enumeration-based methods evaluate each possible combination of descriptors for feasibility and design criteria. The number of such possible combinations increases exponentially with the number of descriptors used by the property models. For a group-contribution method with N descriptors

and maximum size of molecule R, the number of possible group combinations is[7]

$$\sum_{i=2}^{R} \frac{(N+i-1)!}{i!(N-1)!}$$

This number roughly equals 10^{23} for the GC$^+$ method with $N = 182$ (first-order) descriptors and a maximum molecular size of $R = 15$ descriptors. Enumeration of these solutions, calculation of the GC$^+$ properties, and testing the feasibility of each combination for each property would require over 1.2×10^{27} floating point operations, that is, approximately 1.2 trillion peta-FLOPS. This estimate of the computational time for enumeration does not include structural feasibility tests and isomer generation steps, for each feasible combination. Thus, a brute force approach is impractical for large descriptor sets and is restricted to small classes of target compounds. Methods based on enumeration can be made more efficient by rule-based filtering and selection strategies. Due to its simplicity, enumeration has been used extensively.[6,45,62]

Deterministic optimization-based methods for CAMD have also been used for a wide range of property models.[11,12,21,23,24,27,31,37,40,63] Many of these efforts use MINLP algorithms. These MINLP models are often difficult to solve due to their high degree of nonlinearity. Stochastic methods do not offer optimality guarantees but have been used as fast means to generate candidate solutions.[38,41,64–67]

Apart from simple group-contribution-based models, problem formulations using complex property models, including process simulations and quantum chemistry calculations within an optimization and/or enumeration framework, have been reported in the literature.[23,28,31,65,68,69]

Challenges for a comprehensive CAMD framework

The implementation of any comprehensive CAMD approach faces challenges in its treatment of property models and its representation of molecules. We list the most important challenges in the development of a CAMD algorithm below.

- **Combinatorial explosion:** first-order property estimation does not depend on the structure of the molecule. Consequently, first-order estimations are easy to incorporate in molecular design using a compact formulation. At the same time, knowledge of molecular structure is essential for detailed property estimation and extended product design. Direct inclusion of structural interactions in CAMD leads to a large number of integer variables. The number of these variables poses a significant challenge to optimization solvers.
- **Nonlinearity:** the property estimation functions (f) in Eq. 1 are usually nonlinear and lead to a difficult MINLP problem for CAMD. A few examples of the functions used in the GC$^+$ model are

$$\text{Melting point } T_{\mathrm{m}} = T_{\mathrm{m}_0} \ln\left(\sum_{i=1}^{N} c_i^{T_{\mathrm{m}}} n_i\right)$$

$$\text{Critical pressure } P_{\mathrm{c}} = P_{c_1} + \left(\sum_{i=1}^{N} c_i^{P_{\mathrm{c}}} n_i + P_{c_2}\right)^{-2}$$

where $T_{\mathrm{m}_0}, c_i^{T_{\mathrm{m}}}, P_{c_1}, c_i^{P_{\mathrm{c}}}$, and P_{c_2} are constants. In generic MINLP approaches, solvers face numerical challenges due to these nonlinearities and may miss many good designs.

- **Redundancy in molecular space:** molecules can be easily represented as planar graphs to allow systematic inclusion of structural constraints along with higher-order property constraints. However, such a representation is plagued by redundancy due to identical nodes in the molecular graph corresponding to the same descriptor. For example, in the case of n-pentane ($CH_3{}^d$—$CH_2{}^a$—$CH_2{}^b$—$CH_2{}^c$—$CH_3{}^e$), the graph obtained by relabeling the groups ($CH_3{}^d$—$CH_2{}^b$—$CH_2{}^a$—$CH_2{}^c$—$CH_3{}^e$) is different from the original graph but corresponds to an identical molecule. Generating large solution sets with sufficient diversity and zero redundancy is a formidable hurdle.
- **Complex or little understood design targets:** molecular design is an initial step in the product development pipeline. Many design factors are often difficult to account for in early design. This includes the cost of production, intellectual property conflicts, alternative synthesis routes, and so forth. As these factors may eventually limit applicability of proposed molecules in the final execution stages, the output solution pool from the CAMD phase needs to be sufficiently large and diverse. The design process also needs to be flexible to handle feedback from subsequent steps in product design. Thus, an ability to generate and work with a large number of solutions is a necessity for any molecular design approach.

Approaches taken to counter the above issues include (1) limiting the search space to small molecules, (2) restricting the size of the solution set, and (3) using a small subset of the available descriptors. These strategies investigate a small fraction of the complete chemical search space, defeating, in part, the purpose of a CAMD method. In the next section, we propose a systematic framework for generic molecular design based on optimization algorithms that are devised with an explicit aim to address the above challenges.

Proposed Framework

We propose a comprehensive framework to automate, augment, and accelerate steps in CAMD. The proposed methodology tackles the challenges faced by CAMD methods that are based on group-contribution techniques by adopting property-based decomposition as well as structural decomposition over three design steps

1. *Phase 1: Composition design:* determine a large number of compositions matching relaxed design criteria based on first-order property estimates.
2. *Phase 2: Structure design:* find all unique structures for feasible compositions and add higher-order property corrections.
3. *Phase 3: Extended design:* use problem-specific property models to refine the solution pool.

An overview of the proposed framework is shown in Figure 1. As seen in this figure, the design space (property requirements) is first relaxed to account for possible inaccuracies in the first-order prediction techniques that are used in the first stage of the algorithm. As shown in the next subsection, the composition design step based on first-order property predictions can be posed through an integer linear formulation for CAMD. Linearity permits the rapid generation of a potentially large set of candidate designs. These candidates are further analyzed in the second step, where higher-order group corrections are used based on molecular

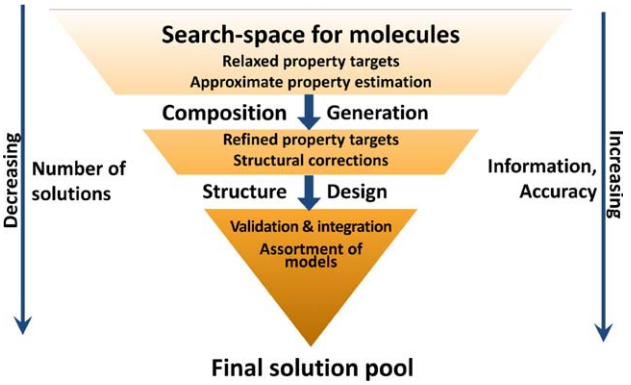

Figure 1. Overview of CAMD framework.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

structure. Using more accurate/complex property models at subsequent stages of the design process reduces the pool of potential candidates. Critical to the success of this approach are the specific optimization models and graph techniques used to tackle each step of the decomposition process. The overall approach is modular and permits incorporation of problem-specific property models to restrict the design space further. Furthermore, the molecular compositions obtained from Phase 1 of the algorithm are independent of each other. Hence, massive parallelization of Phases 2 and 3 is also possible. The details of each phase are presented in the following sections. A design example is used throughout the discussion to illustrate the proposed approach.

Composition design

The composition design step screens the vast molecular space to find potentially feasible molecular compositions with favorable properties. In this step, we seek a large and diverse solution pool of candidate molecules, characterized only by molecular composition, that is, the set of descriptors present in the molecule and their frequency. The GC$^+$ method is used to predict key molecular properties, whereby we use only first-order groups as the building blocks for the molecule. Although we are only concerned with composition and not structure in this stage, we still have to overcome the nonlinearity of property models. We do so through a transformation technique as follows.

Transforming the model

Consider the k-th property, X_k, with specified design criteria, $X_k \in \left[X_k^L, X_k^U\right]$. The typical property constraint is nonlinear in the integer variables

$$X_k^L \le X_k = f_k\left(\sum_{i\in F} c_i^k n_i\right) \le X_k^U \tag{2}$$

where c_i^k denote the first-order group contributions to X_k. However, the nonlinear constraint $X_k^L \le X_k \le X_k^U$ in the X_k-space is equivalent to the following linear constraint in the f_k^{-1}-space

$$\kappa_k \le \sum_{i\in F} c_i^k n_i \le \pi_k$$

where κ_k and π_k denote predetermined lower and upper bounds, respectively, on $f_k^{-1}(X_k)$ over the domain $\left[X_k^L, X_k^U\right]$

$$\kappa_k = \min_{X_k\in\left[X_k^L, X_k^U\right]} f_k^{-1}(X_k); \qquad \pi_k = \max_{X_k\in\left[X_k^L, X_k^U\right]} f_k^{-1}(X_k)$$

Determination of κ_k and π_k is straightforward after inversion for the monotone functions used in GC$^+$. For multimodal functions, this transformation would require the solution of two one-dimensional global optimization problems, which can be solved with very little computational effort by the global nonlinear optimization solver BARON.[70,71] We relax the corresponding bounds on f_k to allow for first-order estimation errors, that is, instead of $\left[X_k^L, X_k^U\right]$ we allow the property X_k to lie in the expanded interval $\left[0.9X_k^L, 1.1X_k^U\right]$. This relaxation is justified by the fact that the average errors in first-order property estimation of the GC$^+$ model rarely exceed 10%.[55,61] Larger ranges can be used in case of properties with higher estimation errors. As property constraints are enforced gradually to this relaxed representation in our decomposition approach, the use of sufficiently larger property ranges in this stage of our methodology will guarantee that the decomposition approach does not overlook structures that would have been found by a simultaneous approach. As we demonstrate in the following sections, the generation of solutions is quick, thus permitting us to use large property ranges for this purpose. For simplicity, from now on, we use $\left[X_k^L, X_k^U\right]$ to denote the range of the suitably relaxed properties.

By exploiting the linearity in the descriptor space of the GC$^+$ method, we can transform the property ranges imposed by design targets to simple linear constraints. Equipped with these linear equations, we can now proceed to formulate a linear integer model for composition design.

Composition design model

The model used to solve for molecular compositions is formulated as an MILP using the linear constraints developed above

$$\min \sum_k w_k \left[\frac{\sum_i c_i^k n_i - \kappa_k}{\pi_k - \kappa_k}\right] \tag{3}$$

$$\text{s.t.} \quad \kappa_k \le \sum_{i\in F} c_i^k n_i \le \pi_k, \qquad \forall k \in \mathcal{K}$$

$$2 \le \sum_{i\in F} n_i \le R \tag{4}$$

$$3N^{\text{cycl}} \le \sum_{i\in F^{\text{cycl}}} n_i \le 8N^{\text{cycl}} \tag{5}$$

$$6N^{\text{arom}} = \sum_{i\in F^{\text{arom}}} n_i \tag{6}$$

$$\sum_{i\in F^{\text{acyc}}} \left(v_i^{\text{acyc}} - 2\right) n_i + \sum_{i\in F^{\text{cycl}}} v_i^{\text{acyc}} n_i - 2N^{\text{cycl}} + \sum_{i\in F^{\text{arom}}} v_i^{\text{acyc}} n_i - 2N^{\text{arom}} + 2 = 0 \tag{7}$$

$$Y^{\text{arom}} \le N^{\text{arom}} \le R^{\text{arom}} Y^{\text{arom}} \tag{8}$$

$$Y^{\text{cycl}} \le N^{\text{cycl}} \le R^{\text{cycl}} Y^{\text{cycl}}$$
$$n_i^L \le n_i \le n_i^U, \qquad \forall i \in F \tag{9}$$

The integer variable n_i in this formulation denotes the frequency of the descriptor i in the molecule. The numbers of acyclic, aromatic, and cyclic bonds of descriptor i are represented by $v_i^{\text{acyc}}, v_i^{\text{arom}}$, and v_i^{cycl}, respectively. The sets of cyclic, aromatic, and purely acyclic descriptors are denoted

by $F^{\text{cycl}}, F^{\text{arom}}$, and F^{acyc}, respectively. The maximum number of descriptors in the design is given by R. The maximum numbers of aromatic and cyclic rings are given by R^{arom} and R^{cycl}, respectively. The binary variables Y^{arom} and Y^{cycl} represent the existence of aromatic and cyclic rings, respectively. The integer variables N^{arom} and N^{cycl} represent the number of aromatic and cyclic rings, respectively.

The objective function is a normalized weighted sum of the design properties that are indexed over $k \in \mathcal{K}$. The weights can be positive or negative depending on the design criteria. Equation 3 represents the GC$^+$ method for property prediction and design property targets $X^L \leq X \leq X^U$ in its linear form discussed in the previous subsection. Equations 4–9 represent the various structural constraints used. By exploiting the nature of GC$^+$ descriptors, the connectivity constraint (7) enforces acyclic tree structure in the space of acyclic bonds. In this equation, only the descriptors that possess acyclic external bonds are considered. This constraint states that the sum of all acyclic valencies of descriptors in a molecule equals twice the number of acyclic bonds in the molecule. The number of acyclic bonds in the molecule without fused rings is given by

$$\sum_{i \in F^{\text{acyc}}} n_i + N^{\text{arom}} + N^{\text{cycl}} - 1.$$

The above equation is a simplified representation of our computational implementation for designing compositions with multiple fused rings. These structural constraints provide solutions that can potentially form feasible molecular structures. Connectivity constraints are handled in the structure generation stage.

The properties included in the above model are all those for which GC$^+$ contributions have been developed, namely

- melting point (T_{m}),
- boiling point (T_{b}),
- critical temperature (T_{c}),
- critical pressure (P_{c}),
- critical volume (V_{c}),
- standard Gibbs energy at 298 K (G_{f}),
- standard enthalpy of formation at 298 K (H_{f}),
- standard enthalpy of vaporization at 298 K (H_{v}),
- standard enthalpy of fusion at 298 K (H_{fus}),
- octanol-water partition coefficient (K_{ow}),
- aqueous solubility (W_s),
- heat capacity of liquids (c_{p}^l),
- surface tension at 298 K (σ),
- viscosity at 300 K (η), and
- Hansen solubility parameters–dispersion (δ_D), dipolar (δ_P), and hydrogen bond (δ_H).

The estimates for all of these properties are available from the MILP formulation at no extra computational cost. Thus, even properties that do not have explicit target ranges can be used to analyze the results of the CAMD problem. In addition, it is trivial to integrate additional properties as GC$^+$ models for these become available in the future. Properties whose prediction methods involve correlations of GC$^+$ properties allow us to translate property targets to the GC$^+$ property space by interval arithmetic. This, for instance, is the case for the flash point correlation developed by Catoire and Naudet[72]

$$T_{\text{f}} = 1.477 T_{\text{b}}^{0.79686} H_{\text{v}}^{0.16845} n_{\text{C}}^{-0.05948}$$

Patterns in property–descriptor relationships can be identified from complex models used to estimate properties not suited for group-contribution methods. These patterns can then be used to set targets on GC$^+$ properties or the molecular structure. Such analysis is facilitated in our computational implementation of the proposed approach via utilities that generate GC$^+$ descriptors and property estimates for any given molecule.

For any molecular design problem, the size of the above MILP formulation is independent of the size of the molecule. Instead, the total number of integer variables in the model equals the number of descriptors used in the group-contribution method plus four (for variables $N_{\text{cycl}}, N_{\text{arom}}, Y_{\text{cycl}}$, and Y_{arom}). For the GC$^+$ method, in particular, the number of integer variables in the MILP equals the number of first-order descriptors, that is $|F| = 182$, plus four. MILPs of this size are easily solvable by solvers such as CPLEX and BARON. Both of these solvers can provide not just an optimal but all feasible solutions to this MILP. This capability facilitates application of the proposed approach to progressively narrow down a large solution set. The proposed MILP formulation is orders of magnitude faster than an equivalent MINLP formulation. This will be demonstrated with three case studies.

Let $\mathcal{C}$ denote the set of the solutions obtained by the MILP. Each solution is a composition vector $n^c = [n_1^c, \ldots, n_i^c, \ldots, n_N^c], c \in \mathcal{C}$ and may correspond to many isomers. Structures for all isomers are required for further screening based on structural property corrections. This topic is addressed in the next section.

Example

A fictitious molecular design example, previously used to demonstrate other CAMD methods,[73] is solved to illustrate the proposed approach. The design targets for the example are

Molecular weight	M_{w}	$\geq$	300 g mol^{-1}
Boiling point	T_{b}	$\geq$	500 K
Octanol-water coefficient	$\log K_{\text{ow}}$	$\leq$	4
Melting point	T_{m}	$\leq$	300 K
Total solubility parameter	δ	$\in$	[18, 22] MPa$^{0.5}$

As group-contribution methods for δ are not available, only molecular weight, melting point, boiling point, and $\log K_{\text{ow}}$ are used to formulate the first-stage MILP. In other words, the first-stage design space is relaxed by entirely ignoring the requirement on the solubility parameter. This requirement will be dealt with in a subsequent stage of the design process. The combined first-stage objective function considered is to maximize the liquid range of the molecule $T_{\text{b}} - T_{\text{m}}$.

For this example, the top 20,000 molecular compositions were generated within 234 s by solving the MILP in the composition design step using BARON. The maximum number of solutions was set to 20,000 for illustrative purposes. These solutions are diverse and many of them are novel. The solution reported in Ref. 73 with the molecular composition

$$n = [4(\text{—CH}_3), 2(\text{—CH}_2\text{—}), 1(> \text{C} <), 3(\text{—CH}_2\text{COO—}), 1(\text{—CH}_2\text{O—})]$$

was also found as the 19,994-th best solution of the MILP. This solution is used to describe the structure determination step in the next section.

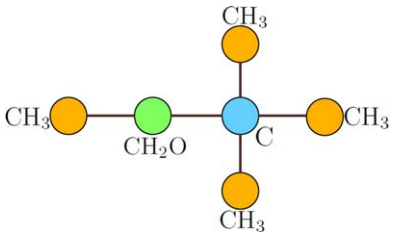

Figure 2. Molecular tree graph of ethyl tert-butyl ether.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary. com.]

Structure generation

Every composition from the candidate set obtained in the first design step may represent many isomers with different properties. Yet, these differences in properties are not captured by first-order group-contribution methods. To increase accuracy in property prediction, we need to quantify the precise structures of these isomers. For isomer characterization, we begin by considering the molecule as a graph, using GC^+ descriptors as the nodes of this graph. This graph is a tree. For example, Figure 2 shows the molecular graph for the common gasoline additive ethyl tert-butyl ether.

The structure generation problem is then as follows

> Given the frequency of a set of molecular descriptors in the molecule, generate all possible distinct molecular tree graphs representing unique isomers.

Let υ_i represent the number of bonds of descriptor i. For instance, $\upsilon_{-CH_2-} = 2$. For each node j, the parameter ID_j denotes the descriptor i corresponding to the node j. The set $\mathcal{J} = \{j_0, j_1, ..., j_{J-1}\}$ represents the set of nodes in the graph. To simplify the presentation, all descriptors i' with $\upsilon_{i'} = 1$ are collapsed into a single node. In our computational implementation, each incidence of a single-valence descriptor is present as a separate node along with j_0. The molecular graph is represented by an $M \times M$ adjacency matrix A. The elements $A_{jj'}$ of A are integers that represent the number of edges between nodes j and j'. Each unique adjacency matrix represents an isomer of the molecular composition.

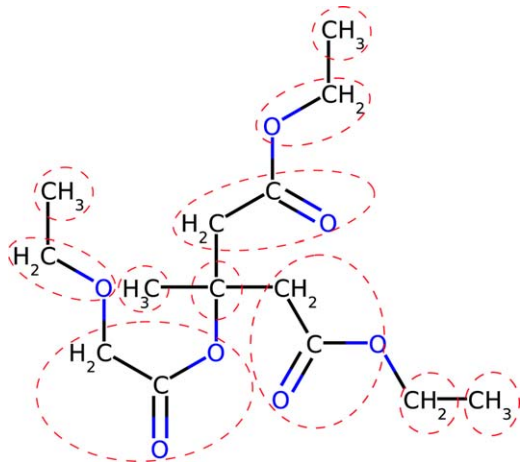

Figure 3. Structure of one of many possible isomers corresponding to example composition.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary. com.]

Properties of aromatic and cyclic compounds are difficult to predict using simple group-contribution methods. In the GC^+ model, many aromatic and cyclic groups are included to boost prediction accuracy. Such compounds are also difficult to represent as a graph, and increase the size of the structure generation problem. In the proposed model, a ring is treated as a supernode in the graph. The adjacency matrix of ring compounds is split into two parts, the cyclic and the aromatic one, and results in a tree representation. As a result, the tree description shown above will still apply to this case.

Example continued

For the solution found above for the example, the set of nodes $\mathcal{J} = \{j_0, \ldots, j_{J-1}\}$ corresponds to the composition vector

$$n = [n_{-CH_3}, n_{-CH_2-}, n_{>C<}, n_{-CH_2COO-}, n_{-CH_2O-}] = [4, 2, 1, 3, 1].$$

Figure 3 shows one of the many possible isomer structures corresponding to this composition vector. The corresponding molecular graph for this isomer is shown in Figure 4. The identities of the nodes of the graph are

ID_0 $-CH_3$
ID_1 $ID_2 = -CH_2-$
ID_3 $>C<$
ID_4 $ID_5 = ID_6 = -CH_2COO-$
ID_7 $-CH_2O-$

In this graph, all the nodes corresponding to single valence groups ($-CH_3$) are collapsed in one node j_0. The adjacency matrix capturing all the necessary bond information for the molecule in Figure 3 is as follows

		j_0	j_1	j_2	j_3	j_4	j_5	j_6	j_7
CH_3	j_0	0	1	1	1	0	0	0	1
CH_2	j_1	1	0	0	0	1	0	0	0
CH_2	j_2	1	0	0	0	0	0	1	0
C	j_3	1	0	0	0	1	1	1	0
CH_2COO	j_4	0	1	0	1	0	0	0	0
CH_2COO	j_5	0	0	0	1	0	0	0	1
CH_2COO	j_6	0	0	1	1	0	0	0	0
CH_2O	j_7	1	0	0	0	0	1	0	0

Second- and third-order property estimation requires detailed information about bonds between different descriptors. For example, all three bonds of $-CH<$ are identical, whereas $-CH_2O-$ exhibits two different types of bonds. This information is lost when each descriptor is treated as a single bond. Therefore, in practice, we generate multiple nodes in the molecular graph for descriptors with different bonds. Simple descriptors, such as $-CH<$ and $-CH_2O-$, are represented by a single node. For descriptors with different bonds, different nodes represent different types of bonds. Thus, $-CH_2O-$ would correspond to two connected nodes j_1 and j_2, both with degrees equal to one, that is,

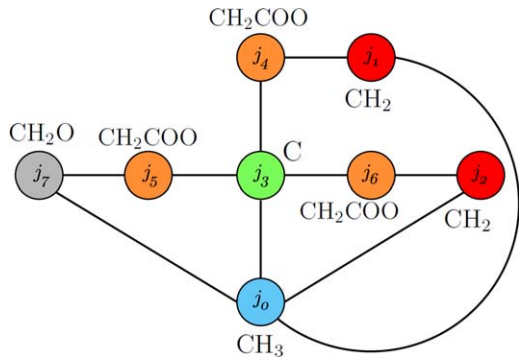

Figure 4. Descriptor node graph for the isomer.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$-^{j_1}\mathrm{CH_2O}-^{j_2}$. This modification allows easy calculation of second- and third-order properties.

Graph generation model

All molecular isomers represented by a molecular composition vector n^c are generated by solving an optimization model that identifies all molecular graphs, which correspond to feasible isomers. Given descriptors j and j' in the composition vector, let the binary variable $y_{jj'}$ model the presence or absence of a bond between these two descriptors. In contrast to previous adjacency matrix-based CAMD models,[63,74] we have complete information about the composition of the molecule in the structure generation step. Thus, we directly include the entries of the adjacency matrix ($y_{jj'}$) as variables in structure generation. The basic model for graph generation is as follows

$$\begin{aligned} &\min \quad f = 0 \\ &\text{s.t.} \quad \sum_{j' \neq j} y_{jj'} = \upsilon_j \qquad \forall j \in \mathcal{J} \qquad (10) \\ &\qquad y_{j'j} = y_{jj'} \qquad \forall j, j' \in \mathcal{J} \qquad (11) \\ &\qquad y_{jj'} \leq 1 \qquad \forall j, j' \in \mathcal{J} \backslash \mathcal{J}_0 j' \neq j \qquad (12) \\ &\qquad y_{jj'} \leq \upsilon_j - 1 \qquad \forall j \in \mathcal{J} \backslash \mathcal{J}_0 j' \in \mathcal{J}_0 \qquad (13) \\ &\qquad y_{jj} = 0 \qquad \forall j \in \mathcal{J} \qquad (14) \end{aligned}$$

where $\mathcal{J}_0$ represents the set of the nodes with valency/degree equal to one. Equation 10 ensures that the degree of each node in the graph is satisfied. Equation 11 enforces symmetry in the adjacency matrix. Equation 12 implies that any two nodes with degree greater than one can be connected by only one bond. The objective function is set to zero to find all feasible solutions to the problem.

The requirement of a completely connected tree graph needs to be included in the model to generate feasible structures. Equation 13 is a simplified connectivity constraint, which ensures that no disconnected tree components like a star graph are formed. Although necessary, this constraint is not sufficient to generate only completely connected graphs. A naive approach to generate connectivity constraints would simply ensure that no tree graph of size less than $|\mathcal{J}|$ exists. However, the number of such constraints increases exponentially with the number of nodes as every subset of nodes requires a separate constraint. The maximum number of nodes in the graph is $2R$, thus leading to a number of constraints of the order 2^{2R}. The problem is beyond reach of all current MILP solvers if all naive connectivity constraints are included in the model. For this reason, we exploit the tree nature of the molecular graph to add a smaller number of tree and cycle constraints in the structure generation model

$$\sum_{\forall j', j'' \in \hat{\mathcal{J}}} y_{j'j''} \leq 2|\hat{\mathcal{J}}| - 3, \forall \hat{\mathcal{J}} \subset (\wp(\mathcal{J}) \backslash \varnothing), |\hat{\mathcal{J}}| \leq \left\lceil \frac{|\mathcal{J}| + 1}{2} \right\rceil \qquad (15)$$

$$\sum_{\forall j', j'' \in \hat{\mathcal{J}}} y_{j', j''} \leq 2|\hat{\mathcal{J}}| - 1, \forall \hat{\mathcal{J}} \subset (\wp(\mathcal{J} \backslash \mathcal{J}_0) \backslash \varnothing), |\hat{\mathcal{J}}| \leq \left\lceil \frac{|\mathcal{J} \backslash \mathcal{J}_0| + 1}{2} \right\rceil \qquad (16)$$

In these constraints, the set $\wp(\mathcal{J})$ is the set of all subsets of $\mathcal{J}$. These cycle (tree) constraints ensure that there exist no cycle (tree) of size up to $(|\mathcal{J}| + 1)/2$ in the graph. Thus, the combination of tree and cycle constraints is satisfied only by completely connected graphs of size $|\mathcal{J}|$. The number of naive connectivity constraints scales with $2^{|\mathcal{J}|}$, whereas the number of tree and cycle constraints (15–16) scales with $2^{\lceil \frac{|\mathcal{J}|}{2} \rceil}$.

We solve the graph model repeatedly, adding uniqueness and redundancy cuts for each structure found to generate unique isomers. The following linear uniqueness cut is used to eliminate previously found solutions

$$\sum_{\substack{j,j' \\ A^s_{j,j'} = 0}} y_{jj'} + \sum_{\substack{j,j' \\ A^s_{j,j'} = 1}} (1 - y_{jj'}) \geq 1 \quad \forall s \in \mathcal{S} \text{ and } j, j' \notin \mathcal{J}_0 \qquad (17)$$

Here, the set $\mathcal{S}$ denotes the set of all solutions found previously. Efficiency of the structure generation step is a direct consequence of refined additions to the basic graph model. Apart from constraint (17) that eliminates previously found solutions, additional cuts are required to eliminate all redundant solutions. These are detailed in the next section. The complete structure generation algorithm to generate all the structures efficiently is expressed as follows

Step 0	Define set $\mathcal{J}$ from the solution vector n^c Initialize set $\mathcal{S} = \{\}, s = 0$. Select the number of structures to be found $S^{\max}$. Generate the basic graph constraint model with connectivity constraints (10)–(16).
Step 1	Solve the structure generation problem. If the problem is infeasible or $\lvert S \rvert = S^{\max}$, STOP; else, set $s = s + 1$. Add the adjacency matrix of the current solution to the solution set: $S = S \cup A$.
Step 2	Generate cuts that eliminate the current solution and its redundant variations and add them to the model. GO TO 1.

Handling redundancy

Redundancy in the context of molecular representation is caused by the presence of identical nodes in the molecular graph. For example, for the molecule in Figure 4, identical nodes are the two —CH_2O— nodes and the three —CH_2COO— nodes present in the graph. The two —CH_2— groups are represented by nodes j_1 and j_2. The graph obtained by interchanging nodes j_1 and j_2 has a different adjacency matrix than the original graph but represents the same molecule. The new adjacency matrix can be obtained from the original one by interchanging the rows and columns

corresponding to nodes j_1 and j_2. Many such redundant adjacency matrices represent the same isomer and must be eliminated from the search space. We introduce special constraints to achieve this goal.

The number of adjacency matrices that correspond to the same molecule grows exponentially and creates computational challenges even for molecules of moderate size. Consider a molecule represented by the solution vector $n = [n_1, \ldots, n_i, \ldots, n_N]$. In this case, the number of adjacency matrices that satisfy all structural feasibility constraints but represent the same molecule equals the number of all possible permutations of the N groups, that is,

$$P(n) = n_1!n_2!\ldots n_N!$$

Thus, $P(n)$ adjacency matrices are variations of the same original structure. Let the set $\mathcal{P}$, where $|\mathcal{P}| = P(n)$, denote the set of all such permutations. We use $p \in \mathcal{P}$ to represent a particular permutation of all exchangeable nodes. For instance, the three —CH_2COO— groups present in the molecule in Figure 4 are denoted by nodes j_4, j_5, j_6 in set $\mathcal{J}$. Considering the permutations of only these groups,

$$\mathcal{P} = \{\underbrace{(j_4j_6j_5)}_{p=1}, \underbrace{(j_5j_4j_6)}_{p=2}, \underbrace{(j_5j_6j_4)}_{p=3}, \underbrace{(j_6j_5j_4)}_{p=4}, \underbrace{(j_6j_4j_5)}_{p=5}, \underbrace{(j_5j_5j_6)}_{p=6}\}$$

The second element of set $\mathcal{P}$ ($p = 2$) corresponds to a permutation with node j_4 interchanged with j_5, representing the same molecule but a different adjacency matrix.

Let A^p represent the adjacency matrix derived from A corresponding to a permutation of nodes $p \in \mathcal{P}$. In the structure generation step, we solve the graph model and generate integer cuts for each structure found. These cuts are formulated using graph invariants, $G\prime(A)$, to exclude all variations A^p and are added to the model to obtain distinct isomers. To capture redundancy, an ideal graph invariant $G\prime(A)$ should be

- equivalent for A^p and $A^{p\prime}, \forall p, p\prime \in \mathcal{P}$
- different for different backbone structures of the molecule
- easily calculable from elements of the matrix A.

Many graph invariants can be used to generate cuts of the form

$$G\prime(A) \neq G'(A^s) \qquad \forall s \in \mathcal{S} = \{A^1, A^2, \ldots\},$$

where $\mathcal{S}$ is the set of all previously found structures. If $G\prime(A)$ is chosen correctly, each of these cuts will eliminate all the redundant solutions corresponding to each structure in set $\mathcal{S}$.

We use weights-based redundancy cuts derived from weights assigned to each type of bond. In the molecule shown in Figure 4, all the bonds between —CH_2COO— and >C< are assigned the same weight, different from the weight assigned to bonds between —CH_3 and —CH_2O—. For the cuts to be graph invariants, the weights W_1, W_2, … must satisfy a special property given by

$$\sum_{w \in \mathcal{W}_1} W_w = \sum_{w\prime \in \mathcal{W}_2} W_{w\prime} \quad \text{if and only if } \mathcal{W}_1 = \mathcal{W}_2,$$

where $\mathcal{W}_1, \mathcal{W}_2$ are any subsets of the weights. Powers of integers form a series with the above property, thus allowing these powers to be used in a number representation, such as binary, decimal, and hexadecimal systems. The set of alternate numbers from the Fibonacci series also satisfy this requirement. However, in both of these cases, the magnitude of the weights increases rapidly. Hence, we chose logarithms of prime numbers as weights that satisfy the requirement for graph invariants. The first-order graph invariant for each group is the sum of the weights of bonds between all of the group's corresponding nodes and their neighbors. The graph invariant assigned to each descriptor i and the cut derived from the invariant are

$$L_i = \sum_{\substack{j \\ \mathrm{ID}_{j=j}}} \sum_{\substack{j\prime \\ j\prime \neq j}} W_{jj\prime} y_{jj\prime} \quad \forall i \in \mathcal{N}, n_i > 0 \tag{18}$$

$$\sum_{\substack{i \\ n_i > 0}} (L_i^s - L_i)^2 \geq \epsilon^2 \quad \forall s \in \mathcal{S} \tag{19}$$

where $\epsilon \geq \max_{i,i\prime,\bar{i},\hat{i}} \left(W_{ii\prime} - W_{\bar{i}\hat{i}}\right)$. Constraint (18) defines the first-order graph invariant as a linear combination of bond weights. Constraint (19) forces at least two groups in the graph to have different neighbors than they had in all previous structures, thus eliminating redundant isomers. These nonlinear cuts can be readily linearized at the expense of introducing additional variables. The use of first-order graph invariants, though simple, leads to loss of some structural information. For example, the isomers 1-ethoxybutane (CH_3—CH_2—CH_2—CH_2—O—CH_2—CH_3) and 1-propoxypropane (CH_3—CH_2—CH_2—O—CH_2—CH_2—CH_3) are indistinguishable with first-order invariants. Thus, we enumerate a small number of such isomers in a post processing step.

The complete structure generation problem is defined by constraints (10)–(16), along with the uniqueness cut (17) and the redundancy cut (19). The model is implemented in GAMS and solved using BARON.

Additional considerations

The models and algorithms described in previous subsections form the core machinery of the proposed CAMD approach. Additional issues are addressed in this subsection, to enhance the applicability of the framework.

Higher-order corrections

The higher-order descriptors in GC^+ methods are overlapping groups which consist of first-order groups. In the structure generation step of the proposed framework, these higher-order groups are used to correct properties that were estimated based on first-order groups. Corrections identifiable by a single bond between different first-order groups (e.g., HO—CH_2COO—) are used in the structure generation algorithm to guide the search toward feasible solutions via the objective function. Higher-order groups that span two or more first-order groups (e.g., HO—$(CH_2)_m$—OH, $m \geq 3$) are identified by traversing adjacency matrices, a highly efficient operation. Inclusion of both second- and third-order groups facilitates accurate estimation of properties. The feasibility of the molecules is revisited after the addition of second- and third-order corrections. For example, consider the constraint

$$T_b \in [300, 450].$$

The first-order estimate for the normal boiling point of 4-aminobutanol is 448.54 K. Thus, 4-aminobutanol is feasible in the first stage. However, addition of third-order corrections leads to a revised estimate of 478.42 K, making 4-aminobutanol infeasible.

Including groups with missing contributions

Although the GC$^+$ method has been developed with a large descriptor basis, the contributions of a few descriptors to some properties cannot be determined with statistical confidence and are, therefore, unavailable in the GC$^+$ model. Apart from these missing contributions, there exist many group fragments that are entirely absent from the GC$^+$ descriptor base. To enable property prediction for such groups, the GC$^+$ method has been extended to calculate the missing group contributions using connectivity indices.[61] Predictions based on connectivity indices require knowledge of the bonds/structure of the molecule. For example, the missing contribution of the descriptor $-D-$ in $a-D-a'$, depends on the connectivity indices of atoms a, a' connected to the descriptor. The identities of a and a' and their connectivity indices are available only when the structure of the molecule has been determined. As a result, these missing contributions cannot be used in the first-order molecular design stage. We address this issue as follows.

We split the missing group contributions in approximate first-order compositional estimation groups and second-order structural correction groups, both of which are incorporated in the property model. The first-order estimate is calculated by considering all possible atoms in the descriptor set. The approximate contribution is calculated as a weighted-average over all possible atomic connections. The weights used are the scaled frequencies of all possible atoms in the descriptor set. The deviation from the averaged contribution, usually rather small in magnitude, is treated as a second-order correction and is calculated once the structure has been determined. All missing groups in the basis set of GC$^+$ method have been analyzed and approximate missing property contributions have been generated. The structure generation step calculates the actual contributions, thus leading to an accurate accounting of missing contributions in the composition design step.

Simplified molecular-input line-entry specification notation and implementation features

Once the solutions have been generated, we convert them to simplified molecular-input line-entry specification (SMILES) notation.[75] The line notation describes the structure of chemical molecules using short strings in a linear format. For example, the molecule shown in Figure 2 can be conveniently represented in SMILES format as CC(OC(=O)COCC)(CC(=O)OCC)CC(=O)OCC. The ability to produce SMILES structures facilitates analysis of the solutions through computational chemistry toolkits and automated database searches for each structure produced by the proposed methodology.

Extended design

Although group-contribution methods cover a variety of molecules and properties, their applicability is limited to the types of properties they consider. For instance, reactivity and manufacturability properties are not currently possible to predict accurately with group contributions. Hence, a comprehensive CAMD framework that relies on group-contribution methods for preliminary screening of the search space needs to be extendable to incorporate design criteria not captured by group-contribution models.

Complex property models that complement group-contribution methods can be easily assimilated in the final design stage, where a solution pool of molecular candidates along with their complete structure and accurate group-contribution property estimates is available. The methods used in this stage can be empirical correlations such as the Watson relation,[76] different classes of group-contribution models such as UNIFAC,[2] or more complex prediction techniques such as simulations. Correlations that rely on GC$^+$ properties to predict other design criteria can be used directly. The modular nature of the proposed framework also allows stagewise use of an assortment of property models for a specific application. Models that do not require explicit structure information can be used to prune the solution set before structure generation begins. Additionally, analytical models can also be used in the extended design stage. This process will be illustrated through specific case studies. Models that do not require explicit structural information are used to prune the solution set before structure generation. This allows early removal of infeasible structures and speeds up the solution process.

Example continued

In the composition design stage, only properties available in the GC$^+$ method, that is, molecular weight, melting point, boiling point, and octanol-water coefficient are considered. However, we would like to trim the solution set based on solubility. We achieve this in the extended design phase of the proposed approach using the total solubility parameter (δ)

$$\delta = \left[\frac{10^3 H_v - RT}{V_m}\right]^{0.5} \tag{20}$$

The estimate for H_v, which is required for this purpose, is available via the GC$^+$ method. However, that is not the case for the molar volume V_m. To estimate the latter, we use the Gunn-Yumada correlation in conjunction with the Ambrose-Walton correlation[77] for the accentric factor (ω)

$$V_m = V_c(0.29056 - 0.08775\omega)^{(1-T/T_c)^{2/7}} \tag{21}$$

$$\omega = \frac{\theta \ln P_c + 5.9762\bar{\theta} - 1.2987\bar{\theta}^{1.5} + 0.6039\bar{\theta}^{2.5} + 1.0684\bar{\theta}^{5}}{-5.0336\bar{\theta} + 1.1150\bar{\theta}^{1.5} - 5.4121\bar{\theta}^{2.5} - 7.4668\bar{\theta}^{5}} \tag{22}$$

where $\theta = T_b/T_c$ and $\bar{\theta} = 1-\theta$ can be estimated from the GC$^+$ predictions for boiling point and critical temperature. The 20,000 solutions identified earlier yield 12,120 solutions with favorable δ values.

Case Studies

Three case studies are presented in this section to illustrate the functionality and features of the proposed CAMD framework.

Designing R12 alternatives

Designing an alternative to R12 refrigerant was one of the earliest problems addressed in the CAMD literature. The need to replace chlorofluorocarbons due to their ozone depletion potential has posed a perfect application area for the emerging CAMD techniques over the past three decades.

The problem has been extensively studied by various molecular design techniques.[6,10–12,63,74,78] The objective of this case study is to design molecules that satisfy process constraints optimizing economic and environmental objectives. The details of the model are presented elsewhere.[74,78] The operating parameters are the refrigeration cycle's condensing temperature (T_{cnd}) and evaporating temperature (T_{evp}). Design constraints are placed on the refrigerant vapor pressure ($P_s^{\mathrm{T_x}}$) at these temperatures ($\mathrm{x} \in \{\mathrm{evp}, \mathrm{cnd}\}$). Important process design variables are the refrigerant's enthalpy of vaporization at evaporating temperature ($H_{\mathrm{v}}^{\mathrm{T_{evp}}}$) and liquid heat capacity (c_{p}^{l}) at average temperature, which is defined as $T_{\mathrm{avg}} = \left(T_{\mathrm{evp}} + T_{\mathrm{cnd}}\right)/2$. The design requirements for the problem are defined based on the refrigerant cycle as follows.

Design targets

The property targets for the replacement refrigerant are

Heat of vaporization at T_{evp}	$H_{\mathrm{v}}^{\mathrm{T_{evp}}}$	$\geq$	18.4 kJ mol^{-1}
Vapor pressure at T_{evp}	$P_s^{\mathrm{T_{evp}}}$	$\geq$	1.4 bar
Vapor pressure at T_{cnd}	$P_s^{\mathrm{T_{cnd}}}$	$\leq$	14 bar
Heat capacity at T_{avg}	$c_p^l\left(T_{\mathrm{avg}}\right)$	$\leq$	113.4 J mol^{-1} K^{-1}

Other property bounds are set to be consistent with design targets used in previous studies.[78] It should be noted that we take care to minimize the number of heteroatoms in the molecule, while the objective function is used to rank molecules rather than discard them.

Property models and implementation details

The problem is decomposed and solved in stages. The properties included in the first stage are $c_p^l\left(T_{\mathrm{avg}}\right), T_{\mathrm{m}}$, and T_{b}. These are calculated through the following five sets of equalities

$$\left.\begin{aligned} T_{\mathrm{c}} &= T_{\mathrm{c}_0} \ln\left(\sum_i c_i^{T_{\mathrm{c}}} n_i\right) \\ T_{\mathrm{b}} &= T_{\mathrm{b}_0} \ln\left(\sum_i c_i^{T_{\mathrm{b}}} n_i\right) \\ P_{\mathrm{c}} &= P_{c_1} + \left(P_{c_2} + \sum_i c_i^{P_{\mathrm{c}}} n_i\right)^{-2} \\ H_v^{298} &= H_{v_0}^{298} + \left(\sum_i c_i^{H_{\mathrm{v}}} n_i\right) \\ \eta &= \exp\left(\sum_i c_i^{\eta} n_i\right) \end{aligned}\right\} \tag{23}$$

$$\left.\begin{aligned} c_{\mathrm{p}}^l|_{T_{\mathrm{avg}}} &= c_{p_0}^l|_{T_{\mathrm{avg}}} + \sum_i n_i c_{p_i}^l|_{T_{\mathrm{avg}}} \\ c_{p_i}^l|_{T_{\mathrm{avg}}} &= a_i + b_i\left(\frac{T_{\mathrm{avg}}}{100}\right) + d_i\left(\frac{T_{\mathrm{avg}}}{100}\right)^2 \end{aligned}\right\} \tag{24}$$

$$H_{\mathrm{v}}^{T_{\mathrm{evp}}} = H_{\mathrm{v}}^{298}\left(\frac{T_{\mathrm{c}} - T_{\mathrm{evp}}}{T_{\mathrm{c}} - 298}\right)^{0.38} \tag{25}$$

$$\left.\begin{aligned} h &= \frac{T_{\mathrm{br}} \ln\left(P_{\mathrm{c}}/1.013\right)}{1 - T_{\mathrm{br}}} \\ G &= 0.4835 + 0.4605h \\ k &= \frac{h/G - (1 + T_{\mathrm{br}})}{(3 + T_{\mathrm{br}})(1 - T_{\mathrm{br}})^2} \\ T_{\mathrm{br}} &= T_{\mathrm{b}}/T_{\mathrm{c}} \\ \ln P_s^{T_{\mathrm{x}}} &= \frac{-G}{T_{\mathrm{xr}}}\left[1 - T_{\mathrm{xr}}^2 + k(3 + T_{\mathrm{xr}})(1 - T_{\mathrm{xr}})^3\right], \mathrm{x} \in \{\mathrm{cnd}, \mathrm{evp}\} \\ T_{\mathrm{xr}} &= T_{\mathrm{x}}/T_{\mathrm{c}}, \mathrm{x} \in \{\mathrm{cnd}, \mathrm{evp}\} \end{aligned}\right\} \tag{26}$$

Here, $T_{\mathrm{c}}, P_{\mathrm{c}}$ are the critical temperature and pressure, respectively, of the candidate molecule. In the case study, the replacement refrigerant was forced to be chlorine-free. The GC$^+$ model for predicting these properties is given by equation set (23). Equation set (24) is a group-contribution technique similar to the GC$^+$ model developed for heat capacity of the liquid as a function of temperature.[58] The heat of vaporization at evaporating temperature is estimated as a function of standard heat of vaporization using the Watson relationship,[76] as shown in Eq. 25. We use the Reidel–Plank–Miller correlation (26)[79] to obtain the reduced vapor pressure at the condensing and evaporating temperatures in the extended design phase.

Results

- The first-stage MILP identified 994 solutions in 10 s. One thousand seven hundred ninety seven unique structures corresponding to 675 feasible compositions were generated at the rate of 0.12 s per structure.
- The proposed approach identified the results from previous works when the descriptors for the solutions were present in the GC$^+$ method. All the hydrofluorocarbons reported in prior CAMD literature were found, including the target molecule 1,1,1,2-tetrafluoroethane (R134a). Moreover, the proposed methodology identified the following additional structures
- Molecules already in use as commercial refrigerants, including isobutane (R-600a), fluoropropane (R-281), trifluoropropane (R-263), 2,2,2-trifluoroethyl methyl ether (R-E143a), and 1,1,2,3,3-pentafluoropropane (R-245ea).
- 2,3,3,3-tetrafluoropropene (HFO-1234yf), currently being developed as fourth-generation replacement and slated to be industry standard by year 2013.
- Several novel compounds, which could be investigated for lower global-warming potential.

Metal degreasing solvent design

This case study was initially introduced by Shelley and El-Halwagi.[80] The objective of the problem is to design a solvent that will strip metal parts in a degreaser and capture light organics in an absorption column. The solvent is expected to reduce organic flare, resulting in economic gains and low pollution. Simultaneous process and product design were used in Ref. 30 to formulate the solvent design problem by identifying suitable property targets. These targets are based on process requirements of zero sulfur content, limited molar volume, and low-vapor pressure. The corresponding requirements were translated in terms of targets on

boiling point, molar volume, and heat of vaporization of the solvent.

Design targets

The design targets for the solvent are

Boiling point	T_b	$\in$	[418, 457] K
Molar volume	V_m	$\in$	[90.1, 720.8] $cm^3\ mol^{-1}$
Heat of vaporization	H_v	$\in$	[50, 100] $kJ\ mol^{-1}$

Property models and implementation details

Properties H_v and T_b are considered in the composition design phase using Eq. 23. The molar volume is predicted using Eq. 21. Correlations based on critical properties estimated by GC^+ are included in the extended design phase. The descriptor set used for the problem is chosen to be the same one used in Ref. 30, that is, $[-CH_3, >CH_2, -CH_2O-, -CH_2N<, CH_3N<, CH_3C(=O)<, -COOH]$.

Results

- The composition design problem is solved in 4 s identifying 105 compositions. The structure generation stage along with extended design generated 890 molecules taking 0.06 s on average per structure.
- The solution set obtained by the proposed algorithm includes all solutions reported previously in Ref. 30. Solutions such as 2,5-hexadione and 2-octanone were found, along with their homologues.
- The families of solutions found are
  - straight chain mono and diethers (C6–C9);
  - straight chain secondary amines (C7–C8).
- Many existing industrial solvents not reported in previous CAMD literature for this problem are also found by the proposed approach, including 2-heptanone, diamyl ether, and decane. Only H_v, T_b, and V_m were used to design the solvent. Therefore, these solutions need further scrutiny using properties such as LC_{50} and toxicity.
- Viscosity values are essential to solvent performance and allow us to rank the solutions obtained. At this stage, the identified solutions can be compared on the basis of viscosity predictions available from the model, even though no direct constraint on viscosity was included.

Solvent design for crystallization

Crystallization is a major unit operation involved in purification of a variety of solid products. The pharmaceutical sector is among the industries in which crystallization plays a dominant role. Many high quality active ingredients are obtained via commercial crystallization. Crystal morphology is controlled by solvent and the process conditions in the crystallizer. Morphology affects processing steps, filtering, packaging, handling, and so forth. It can also affect quality of the active ingredient when it is to be added to a tablet.

Karunanithi et al.[46] introduced the problem of crystallization solvent selection and design. They formulated the problem for solvent design based on properties such as potential recovery of solute, solubility, and crystal morphology, while satisfying process and safety constraints. Their model requires calculation of a number of properties, including pure species properties and solute–solvent interaction properties. For this problem, we will rely on a UNIFAC calculation for phase equilibria for the organic solute sebacic acid ($HOOC(CH_2)_8COOH$).[32] The solvent design problem will then be handled by adding complex property models at the extension stage.

Design targets

The design targets for properties of the crystallizer solvent are as follows

Total solubility parameter	δ	$\in$	[22, 27] $MPa^{0.5}$
Melting point	T_m	$\leq$	270 K
Viscosity	η	$\leq$	3.5 cP
Lethal concentration	$-\log LC_{50}$	$\leq$	2
Flash point	T_f	$\geq$	273 K
Boiling point	T_b	$\geq$	340 K
Solvent hydrogen bonding solubility parameter	δ_H	$\geq$	15 $MPa^{0.5}$

Property models and implementation details

The properties of the solvent, boiling point, melting point, and viscosity are estimated using the GC^+ model as shown in Eq. 23. The total solubility parameter is estimated using Eqs. 20 and 21. The Hansen H-bond solubility parameters are predicted using two different group-contribution methods, both of which include higher-order groups for accurate estimation. The first method is an extension of the GC^+ model[81]

$$\delta_h = \sum_i c_i^{\delta_h} n_i$$

The second method was developed by Stefanis and Panayiotou[82]

$$\delta_h^{SP} = \sum_i c_i^{\delta_h^{SP}} n_i + \delta_{h*}^{SP}$$

These methods include models for dispersion and polar solubility parameters. The GC^+-based extension is used in the composition design phase and the Stefanis and Panayiotou method is used in the extended design stage. The toxicity and safety of the solvent are captured by LC_{50} and flash point, respectively. For flash point prediction, we use a correlation developed by Catoire and Naudet[72]

$$T_f = 1.477 T_b^{0.79686} H_v^{0.16845} n_C^{-0.05948},$$

where n_C is the number of carbon atoms in the molecule. LC_{50} values are estimated by using the Toxicity Estimation Software Tool (TEST) developed by the Environmental Protection Agency (EPA).[83] This tool includes group-contribution methods for LC_{50} estimation.[84] The solution molecules are exported to the estimation tool via SMILES format.

In the extended design phase, we evaluate the efficiency of the identified molecules by optimizing the crystallizer's operating conditions for potential recovery using UNIFAC calculations to predict the equilibrium properties. Thus, for each of the potential solvents identified above, the following nonlinear programming problem is solved to determine the potential recovery of the solute

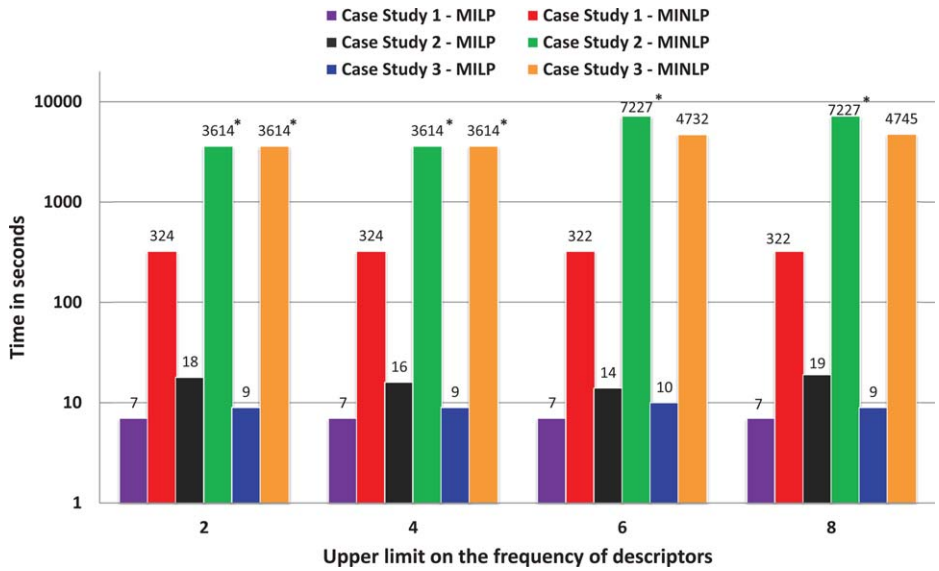

Figure 5. Comparison of MILP and MINLP approaches to composition design.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\begin{aligned}
\max \quad & PR\% = \frac{100}{100 - X_{\mathrm{L}}}\left(\frac{X_{\mathrm{H}} - X_{\mathrm{L}}}{X_{\mathrm{H}}}\right) \\
\text{s.t.} \quad & \ln \gamma_1^{\mathrm{sat}}|_T - \frac{H_{\mathrm{fus}}}{\mathrm{R}T_{\mathrm{m}}}\left(1 - \frac{T_{\mathrm{m}}}{T}\right) \\
& + \ln \gamma_1^{\mathrm{sat}}|_T = 0, \quad T \in \{T_{\mathrm{L}}, T_{\mathrm{H}}\} \\
& \chi_1 + \chi_2 = 1 \\
& 260 \le T_{\mathrm{L}}, T_{\mathrm{H}} \le 320,
\end{aligned}$$

where $\gamma_1^{sat}|_T$ are the activity coefficients estimated by the UNIFAC method as a function of solute, solvent composition, and temperature. The higher and lower-temperatures (T_{H} and T_{L}) denote the crystallizer's operating temperatures. The melting point of sebacic acid is denoted by T_{m} and its heat of fusion by H_{fus}. Subscript 1 corresponds to the solute, whose mole fraction is given by χ_1. The mass solubilities of the solute (X_{L} and X_{H}) correspond to lower- and higher-temperature. Finally, *PR* is the potential recovery in percentage. Sixty-one acyclic GC^+ descriptors that correspond to UNIFAC groups were included in the case study. Using the computationally intensive UNIFAC calculations only at the extended design stage leads to efficient solution generation and simpler NLPs for equilibrium properties.

Results

- In the first stage, we identified 312 compositions in 7 s. The structure generation step provided 1612 feasible structures with 0.06 s on average per structure. UNIFAC-based NLPs were then solved for each of these structures, and LC_{50} values were estimated using the TEST software.
- 2-methoxy ethanol ($CH_3OCH_2CH_2OH$) was reported in the original paper[32] as the optimal solvent for maximizing potential recovery. We also identified this compound as one of the solutions. The results include a series of monohydric alcohols (CH_3OROH) and alkoxy alcohols of varying chain lengths. In addition, we identified many novel solutions that fit the targets.
- Results from a few database searches for molecules that fit the property ranges were also reported in Ref. 32. As hydrogen bonding requirements are included in the model, these solvents are expected to be alcohols. The solvents 1-propanol, 1-butanol, and allyl alcohol that were found in Ref. 32 were also found by the proposed approach.
- Analyzing other solutions after relaxing bounds, we identified many solution families, including diamines, thioethers, and ketones, that match most of the constraints except for one or two. These solutions can be subject to further experimental scrutiny and crystal morphology studies.

Computational efficiency of composition design model

Above we claimed that, in comparison to previous MINLP approaches, the MILP composition model facilitates much faster generation of large solution sets. This point is demonstrated in this subsection. Figure 5 shows the time required to find the top 500 solutions for all three case studies. For each case study, the proposed MILP model is compared to an equivalent MINLP formulation based on Eq. 2. All optimization problems were solved with GAMS/BARON 10.2, with default options using XPRESS as the linear programming subsolver and SNOPT as the nonlinear programming subsolver. In the figure, time is plotted in log scale on the Y axis. All times were obtained on an Intel i7 3.4-GHz processor. On the X axis, we plot the upper limit placed on the frequency of each of the 182 first-order GC^+ descriptors (n_i^U). For $n_i^U = 2$ or 4, the CPU time a solver was allowed to take was 3600 s. For $n_i^U = 4$ or 6, this time was set to 7200 s. Cases in which the solver hit the maximum allowed CPU time are marked by an asterisk in the figure. As seen from the figure, the MILP formulation took less than 20 s to solve each of these problems, whereas the MINLP formulation hit the time limit in half the instances. In each case, the MILP formulation is orders of magnitude faster than the MINLP model. Furthermore, its computational requirements are independent of the upper limit imposed on the number of descriptors allowed in the molecule. These

computational results demonstrate that our MILP-based approach makes possible the effective utilization of computational resources to identify molecular compositions. This computational efficiency accelerates the design process and facilitates repetitive design using feedback from later stages of the product design pipeline.

Conclusions

This article presents an efficient, flexible, and extendable CAMD framework for identifying promising molecules for a variety of applications. Through modeling group-contribution methods in a suitable coordinate space, we were able to use linear integer programming techniques to tackle combinatorial aspects of the problem. It is also orders of magnitude faster than MINLP formulations and complete enumeration of the search space. An equally important feature of the proposed methodology is the incorporation of several graph theoretic and linear integer programming models, which we developed to address the systematic generation of isomers and avoid redundancy in molecular graphs. Finally, the decomposition nature of the proposed framework provides a natural and systematic setting for gradual incorporation of increasingly complex design objectives, along with integration with molecular modeling software, database searches, and simulation-optimization engines. Compared to earlier nonlinear optimization approaches, the proposed approach is more reliable and guarantees global solutions. The applicability and potential of our approach was illustrated through a number of case studies previously reported in the CAMD literature. In all cases, we identified all previously reported solutions, along with many new ones.

Acknowledgment

Portions of this work were performed in support of the National Energy Technology Laboratory's research under the RDS Contract, along with financial support from the National Science Foundation under award 1030168, and the Dow Chemical Company.

Notation

δ = total solubility parameter of the molecular species, $Mpa^{0.5}$
δ_D = Hansen solubility parameter dispersion, $Mpa^{0.5}$
δ_H = Hansen solubility parameter hydrogen bonds, $Mpa^{0.5}$
δ_P = Hansen solubility parameter dipolar, $Mpa^{0.5}$
η = dynamic viscosity of the molecule at 300 K
v_j = the valency of node j, that is, number of edges incident to node j
κ_k = global minimum value of property function $f_k^{-1}(X)$ over interval $X_k \in \left[X_k^L, X_k^U\right]$
π_k = global maximum value of property function $f_k^{-1}(X)$ over interval $X_k \in \left[X_k^L, X_k^U\right]$
σ = surface tension of molecule at 298 K
υ_i = number of bonds of descriptor i
ϕ = the objective function to optimize in the CAMD problem
c = index running over set $\mathcal{C}$ representing a solution to composition design problem
$\mathcal{C}$ = set representing solutions to composition design problem
C^{max} = number of solutions to be found in the composition design step
c_p = liquid heat capacity of molecular species, J/mol K
c_i = contribution of group i in group-contribution method
F = set of first-order groups in GC^+
F^{acyc} = set of purely acyclic first-order groups in GC^+
F^{cycl} = set of cyclic first-order groups in GC^+. These group may or may not have acyclic bonds
F^{arom} = set of aromatic first-order groups in GC^+. These group may or may not have acyclic bonds
f_k = group-contribution function to predict property X_k
g = constraint set including group-contribution property model and property constraints
G_f = standard Gibbs energy of molecular species at 298 K, kJ/mol
$G\prime(A)$ = graph invariant of adjacency matrix A
h = structural constraints on descriptors in CAMD problem
H_{fus} = standard enthalpy of fusion of molecular species, kJ/mol
H_v = standard enthalpy of vaporization of molecular species at 298 K, kJ/mol
i = index running over set $\mathcal{I}$ representing a descriptor
ID_j = scalar denoting the identity of the descriptor represented by node j
j = index running over set $\mathcal{J}$ representing a node in the molecular graph
$\mathcal{J}$ = set representing the nodes in the molecular graph
$\mathcal{J}_0$ = subset of $\mathcal{J}$ representing the nodes in the molecular graph with degree equal to one
k = index running over set $\mathcal{K}$ representing a property estimated by the GC^+ model
$\mathcal{K}$ = set of properties estimated by the GC^+ model
K_{ow} = octanol-water partition coefficient of molecular species
$K(n)$ = the number of permutations of similar nodes in the molecular graph
L_j^s = weight assigned to node j for solution s in set $\mathcal{S}$
L_j = weight assigned to node j in weights-based graph invariant
LC_{50} = lethal concentration
M = total number of nodes generated to represent the molecular graph
n = vector of integer variables used to denote the composition of the molecule
$\mathcal{N}$ = set representing the descriptors/groups
N = the total number of descriptors in group-contribution model
n_i = frequency of group i in the molecule
n_i^L = lower limit on the frequency of descriptor i in composition design problem
n_i^U = upper limit on the frequency of descriptor i in composition design problem
N^{arom} = integer variable representing the number of aromatic rings
N^{cycl} = integer variable representing the number of cyclic rings
p = index running over set $\mathcal{P}$ representing a permutation of exchangeable nodes in the molecular graph
$\mathcal{P}$ = set representing all possible permutations generated by interchanging nodes representing the same descriptor
$\wp(\mathcal{J})$ = the set of all subsets of set $\mathcal{J}$ (the power set of set $\mathcal{J}$)
P_c = critical pressure of molecular species, bar
R = maximum size of the molecule, that is, maximum number of descriptors allowed in the molecule
R^{arom} = maximum number of aromatic rings allowed in the molecule
R^{cycl} = maximum number of cyclic rings allowed in the molecule
$\mathcal{S}$ = set representing solutions to the structure determination problem
S = set of second-order groups in GC^+
s = index running over set $\mathcal{S}$ representing a solution to structure determination problem
S_v = entropy of vaporization at the normal boiling temperature of molecular species
T = set of third-order groups in GC^+
T_b = normal boiling point of molecular species, K
T_c = critical temperature of molecular species, K
T_f = flash point temperature of molecular species, K
T_m = normal melting point of molecular species, K
V_m = molar volume of molecular species, cm^3/mol
w = vector of weights used to formulate optimization objective function in the composition design step
$W_{jj'}$ = weight assigned to the bond between nodes j and j' in weights-based graph invariant

$\mathcal{W}_1, \mathcal{W}_2$ = subsets of weights used in the redundancy cuts
x = the vector of continuous variables denoting the design properties
X = the property to be determined by group contribution
X_L = lower bound on property X inferred from design criteria
X_U = upper bound on property X inferred from design criteria
Y^{arom} = binary variable representing the existence aromatic rings
Y^{cycl} = binary variable representing the existence cyclic rings
$y_{jj'}$ = integer variable denoting the number of edges between node j and j'

Literature Cited

1. Brignole EA, Bottini SB, Gani R. A strategy for the design and selection of solvents for separation processes. *Fluid Phase Equilib.* 1986;29:125–132.
2. Fredenslund A, Gmehling J, Michelsen ML, Rasmussen P, Prausnitz JM. Computerized design of multicomponent distillation columns using the UNIFAC group contribution method for calculation of activity coefficients. *Ind Eng Chem Process Des Dev.* 1977;16:450–462.
3. Kier LB, Hall LH. The generation of molecular structures from a graph-based QSAR Equation. *Quant Struct Act Rel.* 1993;12:383–388.
4. Kier LB, Hall LH, Dailey RS. Design of molecules from quantitative structure-activity relationship models. 3. Role of higher order path counts: path 3. *J Chem Inf Comput Sci.* 1993;33:598–603.
5. Kier LB, Hall LH, Frazer JW. Design of molecules from quantitative structure-activity relationship models. 1. Information transfer between path and vertex degree counts. *J Chem Inf Comput Sci.* 1993;33:143–147.
6. Joback KG, Stephanopoulos G. Designing molecules possessing desired physical property values. In: Proceedings of the 1989 Foundations of Computer-Aided Process Design Conference, Snowmass, CO. Amsterdam: Elsevier, 1990:195–230.
7. Joback KG Stephanopoulos G. Searching spaces of discrete solutions: the design of molecules possessing desired physical properties. *Adv Chem Eng.* 1995;21:257–311.
8. Odele O, Macchietto S. Computer aided molecular design: a novel method for optimal solvent selection. *Fluid Phase Equilib.* 1993;82:47–54.
9. Balaras CA, Jeter SM. A methodology for selecting and screening novel refrigerants for use as alternative working fluids. *Energy Convers Manag.* 1991;31:389–398.
10. Gani R, Nielsen B, Fredenslund A. A group contribution approach to computer-aided molecular design. *AIChE J.* 1991;37:1318–1332.
11. Duvedi AP, Achenie LEK. Designing environmentally safe refrigerants using mathematical programming. *Chem Eng Sci.* 1996;51:3727–3739.
12. Sahinidis NV, Tawarmalani M, Yu M. Design of alternative refrigerants via global optimization. *AIChE J.* 2003;49:1761–1775.
13. Eslick JC, Shulda SM, Spencer P, Camarda KV. Optimization-Based Approaches to Computational Molecular Design. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2011:173–193.
14. Warrier P, Sathyanarayana A, Patil DV, France S, Joshi Y, Teja AS. Novel heat transfer fluids for direct immersion phase change cooling of electronic systems. *Int J Heat Mass Transf.* 2012;55:3379–3385.
15. Derringer GC, Markham RL. A computer-based methodology for matching polymer structures with required properties. *J Appl Polym Sci.* 1985;30:4609–4617.
16. Vaidyanathan R, El-Halwagi M. Computer-aided design of high performance polymers. *J Elastom Plast.* 1994;26:277–293.
17. Maranas CD. Optimization accounting for property prediction uncertainty in polymer design. *Comput Chem Eng.* 1997;21:S1019–S1024.
18. Camarda KV, Maranas CD. Optimization in polymer design using connectivity indices. *Ind Eng Chem Res.* 1999;38:1884–1892.
19. Vaidyanathan R, Gowayed Y, El-Halwagi M. Computer-aided design of fiber reinforced polymer composite products. *Comput Chem Eng.* 1998;22:801–808.
20. Satyanarayana KC, Abildskov J, Gani R. Computer-aided polymer design using group contribution plus property models. *Comput Chem Eng.* 2009;33:1004–1013.
21. Macchietto S, Odele O, Omatsone O. Design of optimal solvents for liquid-liquid extraction and gas absorption processes. *Trans Inst Chem Eng.* 1990;68:429–433.
22. Naser SF, Fournier RL. A system for the design of an optimum liquid-liquid extractant molecule. *Comput Chem Eng.* 1991;15:397–414.
23. Pistikopoulos EN, Stefanis SK. Optimal solvent design for environmental impact minimization. *Comput Chem Eng.* 1998;22:717–733.
24. Sinha M, Achenie LEK, Ostrovsky GM. Environmentally benign solvent design by global optimization. *Comput Chem Eng.* 1999;23:1381–1394.
25. Marcoulaki EC, Kokossis AC. On the development of novel chemicals using a systematic optimisation approach. *Part II. Solvent design. Chem Eng Sci.* 2000;55:2547–2561.
26. Kim K, Diwekar UM. Efficient combinatorial optimization under uncertainty. 2. Application to stochastic solvent selection. *Ind Eng Chem Res.* 2002;41:1285–1296.
27. Wang Y, Achenie LEK. Computer aided solvent design for extractive fermentation. *Fluid Phase Equilib.* 2002;201:1–18.
28. Giovanoglou A, Barlatier J, Adjiman CS, Pistikopoulos EN, Cordiner JL. Optimal solvent design for batch separation based on economic performance. *AIChE J.* 2003;49:3095–3109.
29. Cismondi M, Brignole EA. Molecular design of solvents: an efficient search algorithm for branched molecules. *Ind Eng Chem Res.* 2004;43:784–790.
30. Eljack FT, Eden M, Kazantzi V, Qin X, El-Halwagi MM. Simultaneous process and molecular design—a property based approach. *AIChE J.* 2007;53:1232–1239.
31. Folić M, Adjiman CS, Pistikopoulos EN. Design of solvents for optimal reaction rate constants. *AIChE J.* 2007;53:1240–1256.
32. Karunanithi AT, Acquah C, Achenie LEK, Sithambaram S, Suib SL. Solvent design for crystallization of carboxylic acids. *Comput Chem Eng.* 2009;33:1014–1021.
33. Conte E, Gani R, Crafts PA, Sansonetti S. Efficient, reliable, and predictive solvent design for pharmaceutical processes. In: AIChE Annual Meeting, Minneapolis, USA, 2011.
34. Strübing H, Konstantinidis S, Karamertzanis PG, Pistikopoulos EN, Galindo A, Adjiman CS. Computer-Aided Methodologies for the Design of Reaction Solvents. Weinheim, Germany: Wiley-VCH Verlag GmbH and Co. KGaA, 2011:267–305.
35. Pereira FE, Keskes E, Galindo A, Jackson G, Adjiman CS. Integrated solvent and process design using a SAFT-VR thermodynamic description: high-pressure separation of carbon dioxide and methane. *Comput Chem Eng.* 2011;35:474–491.
36. McLeese SE, Eslick JC, Hoffmann NJ, Scurto AM, Camarda KV. Design of ionic liquids via computational molecular design. *Comput Chem Eng.* 2010;34:1476–1480.
37. Chávez-Islas LM, Vasquez-Medrano R, Flores-Tlacuahuac A. Optimal molecular design of ionic liquids for high-purity bioethanol production. *Ind Eng Chem Res.* 2011;50:5153–5168.
38. Sundaram A, Ghosh P, Caruthers JM, Venkatasubramanian V. Design of fuel additives using neural networks and evolutionary algorithms. *AIChE J.* 2001;47:1387–1406.
39. Hechinger M, Voll A, Marquardt W. Towards an integrated design of biofuels and their production pathways. *Comput Chem Eng.* 2010;34:1909–1918.
40. Knight JP, McRae GJ. A combinatorial optimization approach to molecular design. *Nanotechnology* 1991;2:142–148.
41. Douguet D, Thoreau E, Grassy G. A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm. *J Comput Aided Mol Des.* 2000;14:449–466.
42. Siddhaye S, Camarda KV, Topp E, Southard M. Design of novel pharmaceutical products via combinatorial optimization. *Comput Chem Eng.* 2000;24:701–704.
43. Kamphausen S, Höltge N, Wirsching F, Morys-Wortmann C, Riester D, Goetz R, Thrk M, Schwienhorst A. Genetic algorithm for the design of molecules with desired properties. *J Comput Aided Mol Des.* 2002;16:551–567.
44. Churchwell CJ, Rintoul MD, Martin S, Visco DP, Kotu A, Larson RS, Sillerud LO, Brown DC, Faulon J. The signature molecular descriptor: 3. Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides. *J Mol Graph Model.* 2004;22:263–273.
45. Harper PM, Gani R, Kolar P, Ishikawa T. Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilib.* 1999;158–160:337–347.
46. Karunanithi AT, Achenie LEK, Gani R. A computer-aided molecular design framework for crystallization solvent design. *Chem Eng Sci.* 2006;61:1247–1260.
47. Maranas CD. Optimal molecular design under property prediction uncertainty. *AIChE J.* 1997;43:1250–1264.

48. Kontogeorgis GM, Gani R. Introduction to computer aided property estimation. In: Kontogeorgis GM, Gani R, editors. Computer Aided Property Estimation for Process and Product Design, Vol.19. Amsterdam, The Netherlands: Elsevier, 2004:3–26.
49. Joback KG, Reid RC. Estimation of pure-component properties from group-contributions. *Chem Eng Commun*. 1987;57:233–243.
50. Constantinou L, Mavrovouniotis ML, Prickett SE. Estimation of thermodynamic and physical properties of acyclic hydrocarbons using the ABC approach and conjugation operators. *Ind Eng Chem*. 1993;32:1734–1746.
51. Constantinou L, Mavrovouniotis ML, Prickett SE. Estimation of properties of acyclic organic compounds using conjugation operators. *Ind Eng Chem*. 1994;33:395–402.
52. Bicerano J. Prediction of Polymer Properties. New York: Marcel Dekker, 2002.
53. Constantinou L, Gani R. New group contribution method for estimating properties of pure compounds. *AIChE J*. 1994;40:1697–1710.
54. Marrero-Morejón J, Pardillo-Fontdevila E. Estimation of pure compound properties using group-interaction contributions. *AIChE J*. 1999;45:615–621.
55. Marrero J, Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilib*. 2001;183–184:183–208.
56. Marrero J, Gani R. Group-contribution-based estimation of octanol/water partition coefficient and aqueous solubility. *Ind Eng Chem Res*. 2002;41:6623–6633.
57. Kolská Z, Ružička V, Gani R. Estimation of the enthalpy of vaporization and the entropy of vaporization for pure organic compounds at 298.15 K and at normal boiling temperature by a group contribution method. *Ind Eng Chem Res*. 2005;44:8436–8454.
58. Kolská Z, Kukal J, Zabransky M, Ružička V. Estimation of the heat capacity of organic liquids as a function of temperature by a three-level group contribution method. *Ind Eng Chem Res*. 2008;47:2075–2085.
59. González HE, Abildskov J, Gani R. Computer-aided framework for pure component properties and phase equilibria prediction for organic systems. *Fluid Phase Equilib*. 2007;261:199–204.
60. Conte R, Martinho A, Matos HA, Gani R. Combined group-contribution and atom connectivity index-based methods for estimation of surface tension and viscosity. *Ind Eng Chem Res*. 2008;47:7940–7954.
61. Gani R, Harper PM, Hostrup M. Automatic creation of missing groups through connectivity index for pure component property prediction. *Ind Eng Chem Res*. 2005;44:7262–7269.
62. Friedler F, Fan LT, Kalotai L, Dallos A. A combinatorial approach for generating candidate molecules with desired properties based on group contribution. *Comput Chem Eng*. 1998;22:809–817.
63. Churi N, Achenie LEK. Novel mathematical programming model for computer aided molecular design. *Ind Eng Chem Res*. 1996;35:3788–3794.
64. Tayal MC, Diwekar UM. Novel sampling approach to optimal molecular design under uncertainty. *AIChE J*. 2001;47:609–628.
65. Lehmann A, Maranas CD. Molecular design using quantum chemical calculations for property estimation. *Ind Eng Chem Res*. 2004;43:3419–3432.
66. Lin B, Chavali S, Camarda K, Miller DC. Computer-aided molecular design using tabu search. *Comput Chem Eng*. 2005;29:337–347.
67. Venkatasubramanian V, Sundaram A, Chan K, Caruthers JM. Computer-aided molecular design using neural networks and genetic algorithms. In: Devillers J, editor. Genetic Algorithms in Molecular Design, Vol. 35. San Diego, CA: Academic Press, 1996:271–302.
68. Buxton A, Livingston AG, Pistikopoulos EN. Optimal design of solvent blends for environmental impact minimization. *AIChE J*. 1999;45(4):817–843.
69. Strübing H, Karamertzanis PG, Pistikopoulos EN, Galindo A, Adjiman CS. Solvent design for a menschutkin reaction by using CAMD and DFT calculations. In: Pierucci S, Ferraris GB, editors. 20th European Symposium on Computer Aided Process Engineering, Vol. 28. Amsterdam, The Netherlands: Elsevier, 2010:1291–1296.
70. Sahinidis NV. BARON: a general purpose global optimization software package. *J Global Optim*. 1996;8:201–205.
71. Tawarmalani M, Sahinidis NV. A polyhedral branch-and-cut approach to global optimization. *Math Program*. 2005;103:225–249.
72. Catoire L, Naudet V. A unique equation to estimate flash points of selected pure liquids application to the correction of probably erroneous flash point values. *J Phys Chem Ref Data*. 2004;33:1083–1111.
73. Gani R. Computer aided methods and tools for chemical product design. *Chem Eng Res Des*. 2004;82:1494–1504.
74. Apostolakou A, Adjiman CS. Refrigerant design case study. In: Achenie LEK, Gani R Venkatasubramanian V, editors. Computer Aided Molecular Design: Theory and Practice; Computer Aided Chemical Engineering, Amsterdam, The Netherlands: Elsevier, Vol. 12. 2002:289–301.
75. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28:31–36.
76. Thek RE, Stiel LI. A new reduced vapor pressure equation. *AIChE J*. 1966;12:599–602.
77. Poling BE, Prausnitz JM, O'Connell JP. The Properties of Gases and Liquids, 5th ed. New York: McGraw-Hill, 2001.
78. Sahinidis NV, Tawarmalani M. Applications of global optimization to process and molecular design. *Comput Chem Eng*. 2000;24:2157–2169.
79. Riedel L. Kritischer koeffizient, dichte des gesattigen dampfes und verdampfungswarme. *Chemie Ingenieur Technik* 1954;26:679–683.
80. Shelley MD, El-Halwagi MM. Component-less design of recovery and allocation systems: a functionality-based clustering approach. *Comput Chem Eng*. 2000;24:2081–2091.
81. Modarresi H, Conte E, Abildskov J, Gani R, Crafts P. Model-based calculation of solid solubility for solvent selection: a review. *Ind Eng Chem Res*. 2008;47:5234–5242.
82. Stefanis E, Panayiotou C. Prediction of Hansen solubility parameters with a new group-contribution method. *Int J Thermophys*. 2008;29:568–585.
83. Toxicity Estimation Software Tool (TEST). Available at: http://www.epa.gov/nrmrl/std/cppb/qsar/.
84. Martin TM, Young DM. Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (Pimephales promelas) using a group contribution method. *Chem Res Toxicol*. 2001;14:1378–1385.

Manuscript received Jun. 16, 2012, revision received Feb. 2, 2013, and final revision received Apr. 1, 2013.