Generic Mathematical Programming Formulation and Solution for Computer-Aided Molecular Design

Zhang, Lei; Cignitti, Stefano; Gani, Rafiqul

Published in:
Computers & Chemical Engineering

Link to article, DOI:
10.1016/j.compchemeng.2015.04.022

Publication date:
2015

Document Version
Peer reviewed version

Title: Generic Mathematical Programming Formulation and Solution for Computer-Aided Molecular Design

Author: Lei Zhang Stefano Cignitti Rafiqul Gani

PII: S0098-1354(15)00123-4
DOI: http://dx.doi.org/doi:10.1016/j.compchemeng.2015.04.022
Reference: CACE 5175

To appear in: *Computers and Chemical Engineering*

Received date: 6-2-2015
Revised date: 15-4-2015
Accepted date: 22-4-2015

Please cite this article as: Zhang, L., Cignitti, S., and Gani, R.,Generic Mathematical Programming Formulation and Solution for Computer-Aided Molecular Design, *Computers and Chemical Engineering* (2015), http://dx.doi.org/10.1016/j.compchemeng.2015.04.022

Highlights

Ms. Ref. No.: CACE-D-15-00067

Title: Generic Mathematical Programming Formulation and Solution for Computer-Aided Molecular Design

(1) A generic model for formulation and solution of CAMD problems is presented
(2) The model uses first and second-order groups in the CAMD optimization problem
(3) A global optimum can be achieved without the use of property relaxation
(4) Three case studies from literature are solved to demonstrate the capabilities

Generic Mathematical Programming Formulation and Solution for Computer-Aided Molecular Design

Lei Zhang, Stefano Cignitti, Rafiqul Gani* (rag@kt.dtu.dk)

CAPEC-PROCESS, Department of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads, Building 229, DK-2800 Kgs. Lyngby, Denmark

Abstract

This short communication presents a generic mathematical programming formulation for Computer-Aided Molecular Design (CAMD). A given CAMD problem, based on target properties, is formulated as a Mixed Integer Linear/Non-Linear Program (MILP/MINLP). The mathematical programming model presented here, which is formulated as an MILP/MINLP problem, considers first-order and second-order molecular groups for molecular structure representation and property estimation. It is shown that various CAMD problems can be formulated and solved through this model.

Keywords: Molecular design, CAMD, chemical structure, group contribution, MILP/MINLP

1. Introduction

Computer-Aided Molecular Design (CAMD) is a method to design molecules with desired properties. That is, through CAMD, it is possible to generate molecules that match a specified set of target properties. CAMD has attracted much attention in recent years due to its ability to design novel as well as known molecules with desired properties. The attention is in particular targeted at the design of chemical based products, such as solvents, refrigerants, active pharmaceutical ingredients, polymers, surfactants, lubricants, and more (Gani, 2004).

Property prediction methods are needed in molecular design, as they enable the prediction of the target properties of the candidate molecules. Here, CAMD methods can be regarded as the reverse engineering approach to property prediction, as the target properties are known while the molecules that match them need to be determined. Typically, almost all CAMD methods use group contribution (GC) based property prediction methods (from Franklin, 1949 to Hukkerikar et al., 2012) to evaluate the generated compound with respect to the specified set of desirable target properties (Harper et al., 1999). The GC-based methods belong to a class known as additive methods (Hukkerikar et al., 2012).

$$F(p) = \omega_1 \sum_i N_i C_i + \omega_2 \sum_j M_j D_j + \omega_3 \sum_k O_k E_k + \cdots \quad (1)$$

In eq. (1), p is the desirable property, C_i is the contribution of first-order group i, N_i is the number of occurrences of first-order group i; D_i is the contribution of second-order group i, M_i is the number of occurrences of second-order group i; E_i is the contribution of third-order group i, O_i is the number of occurrences of third-order group i; ω_1, ω_2, ω_3 are weights that may be imposed on each of the additive terms. From a practical point of view, the highest order of eq. (1) is three (Marerro and Gani, 2001). Second and third order additive methods are able to distinguish some isomeric molecular structures in CAMD problems. In this paper, only first and second order groups are considered. Third order groups can also be considered using this new model, but it is not necessary for most CAMD problems.

With the advent of connectivity-based prediction methods, several researchers have developed new strategies for embedding it with CAMD method. Constantinou et al. (1996) proposed a systematic strategy for generating isomers from a set of groups. Harper et al. (1999) proposed a framework for CAMD method, where the pre-design phase defines the basic needs, the design phase determines the feasible candidates (generates molecules and tests for desired properties) and the post-design phase performs higher level analysis of the molecular structure and the final selection of the product. Samudra and Sahinidis (2013) proposed a new optimization model using relaxed property targets and refined property targets with structural corrections. It is usually difficult to model and solve the MILP/MINLP problem with structure information considered due to the increased size of the mathematical problem and number of alternatives. Thus, alternative solution strategies have been proposed to ensure that solution can be found and that also a global optimum can be found. Harper et al. (1999) used a generate and test approach to decompose the CAMD problem; selection of building blocks (functional groups), combination of groups into chemically feasible molecules, estimation of the specified set of properties for the generated molecules, selection as candidate compounds, and finally, determination of those that match the specified set of properties. Samudra and Sahinidis (2013) decomposed the problem into three design steps: composition design, structure design and extended design. In composition design, a large number of compositions (molecules composed of groups) matching relaxed design criteria based on first-order property estimates are determined. Thus, the GC^+ property estimation model is relaxed (only considering the first-order groups) to obtain the building blocks, then the property model is refined with second-order groups (structure design information) based on the results of the first step. However, this may result in the possibility of an optimal solution being excluded. Second-order groups refine the property prediction and molecules that wrongfully lie outside the search space are neglected. As seen in Figure 1, the solid line box is the feasible region of the decomposed model; the dash line box is the real feasible region of the CAMD problem. If decomposed approach method is used, the global optimal point is excluded from the feasible region. That means the optimal point obtained from the decomposed method might be a local optima. Samudra and Sahinidis (2013) used property relaxation method to avoid this situation. That is, instead of property interval $[X_k^L, X_k^U]$, they allow the property X_k to lie in the expanded interval $[0.9X_k^L, 1.1X_k^U]$. This relaxation is justified by the fact that the average errors in first-order property estimation of the GC^+ model rarely exceed 10% (Samudra & Sahinidis, 2013). But it is not always easy for the users to find the appropriate relaxations. On the other hand, the feasible region of the optimization problem will become larger when relaxations applied, which makes the solution of the problem harder.

Figure 1. Feasible region of CAMD problems using different modeling approach

In this short communication, a new model for CAMD problems is proposed. The models considers both first and second order groups simultaneously in the MILP/MINLP formulation, and the molecular structure is obtained from the solution of the adjacency matrix. This will avoid the possible situation in Figure 1, where a possible optimal point may be excluded from the feasible region, and ensures the obtainability of a global optimal solution. This short communication is structured as follows. Section 2 gives a detailed description of the methodology and the mathematical formulation of CAMD problems with the proposed model; section 3 gives three case studies; section 4 draws some conclusions from the presented results.

2. Methodology

The computer-aided molecular design framework is presented here in Figure 2. The framework has four steps (Cignitti et al., 2015), (1) problem definition: product needs, target properties and desired product type are defined here; (2) CAMD formulation: the needs, properties and product types are converted to a CAMD problem in which objective function and constraints related to molecular structure, product needs (property model) and

process models are defined; (3) MILP/MINLP formulation: the CAMD problem from step two is set-up as a MILP/MINLP formulation; (4) solution of MILP/MINLP problem: the MILP/MINLP formulation is solved directly or through a decomposed strategy depending on the problem type, linearity and size.

Figure 2. Computer-aided molecular design framework

If the needs and target properties of the designed molecule are defined in the design problem, the CAMD problem can be posed as a mathematical program in which the number of binary and continuous variables defines the search space.

The general MILP/MINLP problem is formulated as follows (Karunanithi et al., 2005). **N** is a vector of integer variables, which are related to the numbers of the building blocks (1st and 2nd order groups). **Y** is adjacency matrix which is related to the description of the molecular structure. **X** is a vector of continuous variables, which are related to the process variables. In Karunanithi et al. (2005) the adjacency matrix was not included in the MINLP.

$$\min/\max f_{obj}\left(\mathbf{X},\mathbf{N}\right) \tag{2}$$

$$\text{structural constraints: } g_1\left(\mathbf{N},\mathbf{Y}\right)\le 0 \tag{3}$$

$$\text{property constraints: } g_2\left(\mathbf{N}\right)\le 0 \tag{4}$$

$$\text{process model constraints: } g_3\left(\mathbf{X},\mathbf{N}\right)=0 \tag{5}$$

$$\mathbf{X}\in\Re^n,\mathbf{N}\in Z^m,\mathbf{Y}\in\{0,1\}^q$$

In this formulation, the structural constraints $g_1(\mathbf{N},\mathbf{Y})$ are always linear. The property constraints $g_2(\mathbf{N})$ are linear for most primary properties. For other properties, the property constraints can be non-linear or linear depending on the property model used. The linearity of the process model constraints $g_3(\mathbf{X},\mathbf{N})$ depends on the specific problem.

Although all the first-order groups of Morrero and Gani (2001) are considered in the model, the groups used in a specific problem is only a subset of all the first-order groups. The first step of the problem formulation is the selection of groups for a specific problem based on the type of molecules to be generated. The number of selected groups is usually less than 20 in a specific CAMD problem. Note that here specific molecule types (acyclic, cyclic, and aromatic) are generated separately.

To specify the structure of the target molecule, vertex adjacency matrix is used for the description of the connectivity of groups. A number (ID) is assigned to each group to avoid duplicate names of the repeat groups in the molecule. For instance, the set of groups (4 CH_3, C, CH_2O) with vertex adjacency matrix below (Table 1) describes ethyl tert-butyl ether (*SMILES:* CCOC(C)(C)C).

Table 1. Adjacency matrix for ethyl tert-butyl ether

We now introduce the variables, constraints and objective function of the new CAMD model. First, we define the following sets:

G_1 = {i | i is a first-order group}; G_2 = {j | j is a second-order group}; ID = {id | id is the ID number of each groups}

Several binary variable representations are adopted in this model. Binary variable y_{i_1,id_1,i_2,id_2} denotes whether group i_1 with id id_1 (i_1, id_1) is connected to group i_2 with id id_2 (i_2, id_2), where $i_1, i_2 \in G_1, id_1, id_2 \in ID$. In this formulation, different bond type are considered within the structure of first-order groups, and all the second-order groups in Morrero and Gani (2001) are the connection of first-order groups using single bond.

$$y_{i_1,id_1,i_2,id_2} = \begin{cases} 1 & \text{group } (i_1, id_1) \text{ is connected to group } (i_2, id_2) \\ 0 & \text{otherwise} \end{cases}$$

Binary variable z_{i_1,id_1} is used to describe the existence of group (i_1, id_1).

$$z_{i_1,id_1} = \begin{cases} 1 & \text{group } (i_1, id_1) \text{ exists in the molecule} \\ 0 & \text{otherwise} \end{cases}$$

The constraints of the MILP/MINLP problem consists of structural constraints, property constraints and process constraints.

a) Structural constrains

Through classification of the different structural groups on the basis of their valencies (number of free attachments), the octet rule provides a simple relation for the structural feasibility of a collection of groups (Odele and Macchietto, 1993).

$$\sum_{i \in G_1} (2 - v_i) n_i^{(1)} = 2q \tag{6}$$

$$\sum_{i_1 \neq i_2; i_1, i_2 \in G_1} n_{i_1}^{(1)} \geq n_{i_2}^{(1)} (v_{i_2} - 2) + 2 \quad \forall i_2 \in G_1 \tag{7}$$

In eq. (6) and (7), $n_i^{(1)}$ is the number of first-order group i in the target molecule, v_i is the valency of group i, q is assigned the value of 1, 0 or -1 for acyclic, monocyclic or bicyclic groups, respectively.

In Churi and Achenie (1996), Eqs. (8)−(12) are added to ensure that only one molecule is formed.

$$\sum_{i_2 < i_1} \sum_{j_2} y_{j_1,i_1,j_2,i_2} + \sum_{j_2 < j_1} y_{j_1,i_1,j_2,i_1} \geq w_{j_1,i_1} \quad \forall i_1 > 1, j_1 > 1 \tag{8}$$

$$\sum_{i \in G_1} n_i^{(1)} + \sum_{j \in ID} \sum_{i \in G_1} w_{j,i} = n^{\max} \tag{9}$$

$$w_{1,1} = 0 \tag{10}$$

$$w_{j_1,i_1} \geq w_{j_2,i_2} \quad \forall i_1 > i_2; i_1, i_2 \in G_1; j_1, j_2 \in ID \tag{11}$$

$$w_{j_1,i_1} \geq w_{j_2,i_1} \quad \forall j_1 > j_2; i_1 \in G_1; j_1, j_2 \in ID \tag{12}$$

Additional constraints may be placed on the number ($n_i^{(1)}$) of groups of group i to keep it within lower and upper bounds, n_i^L and n_i^U, respectively.

$$n_i^L \le n_i^{(1)} \le n_i^U \quad \forall i \in G_1 \tag{13}$$

Another constraint may be imposed on the total number of groups making up a molecule.

$$n^{\min} \le \sum_{i \in G_1} n_i^{(1)} \le n^{\max} \tag{14}$$

From eq. (13), the adjacency matrix of target molecular can be established as follows (Table 2):

In the adjacency matrix, same groups with the same ID (diagonal) cannot connected.

$$y_{i_1,id_1,i_1,id_1} = 0 \quad \forall i_1 \in G_1, id_1 \in ID \tag{15}$$

If group (i_1, id_1) connects to group (i_2, id_2), then (i_2, id_2) must connect to (i_1, id_1).

$$y_{i_i.id_1,i_2,id_2} = y_{i_2,id_2,i_1,id_1} \quad \forall i_1, i_2 \in G_1; id_1, id_2 \in ID \tag{16}$$

The constraints between binary variables z and y is shown in eq. (17) and (18).

$$\sum_{i_2 \in G_1} \sum_{id_2 \in ID} y_{i_1,id_1,i_2,id_2} = \nu_{i_1} z_{i_1,id_1} \quad \forall i_1 \in G_1, id_1 \in ID \tag{17}$$

$$\sum_{id_1 \in ID} z_{i_1,id_1} = n_{i_1}^{(1)} \quad \forall i_1 \in G_1 \tag{18}$$

Table 2. Adjacency matrix of target molecular

The other equations in the structural constraints restricts the number of second-order groups ($n_j^{(2)}$) from the adjacency matrix. There are limited number of second-order groups in the property estimation method. For instance, there are 122 second-order groups considered in Morrero and Gani (2001). For any second-order group J, constraints can be established based on its chemical structure (connection of first-order groups) to obtain $n_J^{(2)}$ as eq. (19) and (20) shows. N_B^J is the number of bonds in second-order group J, b^J is binary variable, T is an integer parameter, and it depends on the structure of the second-order group (examples are listed below). M is a big number for big-M method (Griva et al., 2009). In Big-M method, appropriate value of M should be selected. The value of M should be the smallest values that work in the context of the model, because large values of M can cause branch-and-bound solvers to make slow progress solving the MIP model. In this formulation, the value of M = 20, because in all second-order groups, the number of bonds never larger than 20.

$$N_B^J - M\left(1 - b_{id_1,id_2,\ldots,id_{N_B}}^J\right) \le \sum_{\substack{\text{if group } (i_1,j_1) \text{ and } (i_2,j_2) \text{ are} \\ \text{connected in 2nd group } J}} y_{i_1,j_1,i_2,j_2} \le N_B^J - 1 + M\left(b_{id_1,id_2,\ldots,id_{N_B}}^J\right) \quad \forall J \tag{19}$$

$$n_J^{(2)} = \frac{1}{T} \sum_{(id_1,id_2,id_{N_B})} b_{id_1,id_2,\ldots,id_{N_B}}^J \quad \forall J \tag{20}$$

Equations of several second-order groups are listed below as examples.

- $(CH_3)_2CH$:

$$2-M\left(1-b_{id_1,id_2,id_3}^{(CH_3)_2CH}\right)\le y_{CH,id_1,CH_3,id_2}+y_{CH,id_1,CH_3,id_3}\le 1+Mb_{id_1,id_2,id_3}^{(CH_3)_2CH}\quad \forall id_1,id_2,id_3\in ID \tag{21}$$

$$n_{(CH_3)_2CH}^{(2)}=\frac{1}{2}\sum_{id_1\in ID}\sum_{\substack{id_2\neq id_3\\ id_2\in ID}}\sum_{\substack{id_3\neq id_2\\ id_3\in ID}} b_{id_1,id_2,id_3}^{(CH_3)_2CH} \tag{22}$$

In eq. (21), if and only if $y_{CH,id_1,CH_3,id_2}=y_{CH,id_1,CH_3,id_3}=1$, there exist a second-order group $(CH_3)_2CH$, and $b_{id_1,id_2,id_3}^{(CH_3)_2CH}=1$. Since the two CH_3 groups in group $(CH_3)_2CH$ are counted twice as they have different ID, the number of the second-order group $(CH_3)_2CH$ equals to ½ times the summary of $b_{id_1,id_2,id_3}^{(CH_3)_2CH}$ as eq. (22) shows.

Other second-order groups are formulated in a similar way. Another two examples are illustrated here.

- $(CH_3)_3C$

$$3-M\left(1-b_{id_1,id_2,id_3,id_4}^{(CH_3)_3C}\right)\le y_{C,id_1,CH_3,id_2}+y_{C,id_1,CH_3,id_3}+y_{C,id_1,CH_3,id_4}\le 2+Mb_{id_1,id_2,id_3,id_4}^{(CH_3)_3C}\quad \forall id_1,id_2,id_3,id_4\in ID \tag{23}$$

$$n_{(CH_3)_3C}^{(2)}=\frac{1}{6}\sum_{id_1\in ID}\sum_{\substack{id_2\neq id_3\\ id_2\neq id_4\\ id_2\in ID}}\sum_{\substack{id_3\neq id_2\\ id_3\neq id_4\\ id_3\in ID}}\sum_{\substack{id_4\neq id_2\\ id_4\neq id_3\\ id_4\in ID}} b_{id_1,id_2,id_3,id_4}^{(CH_3)_3C} \tag{24}$$

- CHCHO

$$y_{CH,id_1,CHO,id_2}=b_{id_1,id_2}^{CHCHO}\quad \forall id_1,id_2\in ID \tag{25}$$

$$n_{CHCHO}^{(2)}=\sum_{id_1\in ID}\sum_{id_2\in ID} b_{id_1,id_2}^{CHCHCO} \tag{26}$$

All second-order groups are formulated in this way to obtain their number from the adjacency matrix. These constraints are not needed to be modified for different problems. Thus, these second-order constraints can be stored separately for all CAMD problems.

b) Property constraints

The property constraints are represented in eq. (27). P is the set of all target properties of the molecule. All the target properties should be in its given range $[p_k^L, p_k^U]$.

$$p_k^L\le p_k\le p_k^U\quad \forall k\in P \tag{27}$$

The target properties p_k may be obtained from the molecular structural variables or the combination of other properties. Gani and Constantinou (1996) proposed a classification of properties as primary (pure component properties that can be determined only from the molecular structural variables as eq. (28) shows), secondary

(pure component properties that are dependent on primary properties) and functional (pure component properties dependent on temperature and/or pressure).

$$p_k = \sum_{i \in G_1} n_i^{(1)} p_k^{(1)} + \sum_{j \in G_2} n_j^{(2)} p_k^{(2)} \quad \forall k \in P \tag{28}$$

c) Process constraints

The process constraints are the process model which contains continuous and discrete variables. These constraints integrate the product design problem with process design problem.

The new CAMD model is established as MILP/MINLP model using the user defined objective function and constraints introduced here. This model is tested using several pure compound/polymer design case studies.

3. Case study

Three relevant case studies from literature (polymer design, solvent design and surfactant design) are presented in this communication to highlight the application of the mathematical programming model. Detailed information for the case studies can be found in the supplementary material. Design of other molecules (Samudra and Sahinidis, 2013; Harper et al., 2001; Karunanithi et al., 2005) such as refrigerant design and various types of solvents can be obtained from the authors.

a) Polymer design

This case study revisits the polymer design problem solved by Derringer and Markham (1985). It is required to identify viable polymer repeat unit structures that satisfy the property constraints based on the density ρ (g/cm^3) and the glass transition temperature T_g (K):

$$1 \le \rho \le 1.5 \qquad T_g \ge 298$$

The basis set of groups (G_1) is selected to include: {CH_2, CH, OH, CH_2CO, CH_2O, CHCl, COO, $CONHCH_2$ and R}. The set for *ID* is: {ID_1, ID_2 and ID_3}. The set of second order groups (G_2) is from Marrero & Gani (2001). All the second-order groups in Marrero & Gani (2001) are selected in G_2, but the number of second-order groups are restricted by set G_1 and constraints (19) and (20) (the same applies to case study (b) and (c)). Polymer repeat unit with a minimum of three and a maximum of five groups are allowed, with the same group appearing a maximum of three times.

The optimal polymer -[COO-CH_2O-CH_2CO]$_n$- (ρ = 1.45; T_g = 351.02) is among the feasible solution in Chelakara et al. (2009).

More details are given as supplementary material.

b) Solvent design

This solvent design case study is taken from Karunanithi et al. (2005). The goal is to design the optimum extractant which has the lowest enthalpy of formation (H_f) and the following properties (T_m: melting point; T_b: boiling point; S_p: Hildebrand solubility parameter):

$$T_m \ge 270 \qquad T_b \le 430 \qquad S_p \ge 20$$

The basis set of groups (G_1) is selected to include: {CH_3, CH_2, CH, C, OH, CHO, CH_3COO, CH_2COO, HCOO, CH_3CO, CH_2CO, COOH and COO}. This allows the formation of thousands of acyclic organic molecules containing C, H and O atoms. The set for *ID* is: {ID_1, ID_2 and ID_3}. The set of second order groups (G_2) is selected from Marrero & Gani (2001). Molecules with a minimum of three and a maximum of seven groups are allowed, with the same group appearing a maximum of three times.

The optimal molecule obtained from the MILP model is $C_5H_{10}O_3$ (*SMILES*: O=C(C)OCC(O)C; H_f = -462.08 KJ/mol; T_m = 266.65 K; T_b = 441.15; S_p = 21.32). The first-order groups contained in the design molecule are: 1 CH_3, 1 CH_2, 1 CH, 1 OH and 1 CH_3COO; the second-order group contained in the molecule is: 1 CHOH.

More details are given as supplementary material.

c) Surfactant design

This case study is taken from Mattei et al. (2012). The aim of this case study is the design of a UV sunscreen, in the emulsified form, with a high sun protection factor. The benzene ring is fixed as backbone structure in this case study. The designed surfactant should have the following properties (*lc*50: Fathead Minnow 96-hr LC50; *Sp*: Hildebrand solubility parameter; *clp*: cloud point):

$$lc50 \geq 3.16 \qquad Sp \leq 25 \qquad clp \geq 343.15$$

The basis set of groups (G_1) is selected to include: {CH_3, CH_2, CH, C, aCH, aC-OH, CH_2COO, CH_3O, CH_2O, aC-O and OCH_2CH_2OH}. The set for *ID* is: {ID_1, ID_2, ID_3, ID_4 and ID_5}. The set of second order groups (G_2) is selected from Marrero & Gani (2001). Molecules with a minimum of 10 and a maximum of 15 groups are allowed, with the same group appearing a maximum of five times.

The optimal molecule obtained from the MILP model is $C_{15}H_{22}O_3$ (*SMILES*: CCCCC(OC(=O)CC)Cc1ccc(O)cc1; *lc*50: 5.40; *Sp*: 20.80; *clp*: 347.64). The first-order groups contained in the design molecule are: 2 CH3, 5 CH2, 1 CH, 4 aCH, 1 aC-OH, 1 CH2COO, 1 aC-O; the second-order group contained in the molecule is: 1 AROMRINGs1s4 (benzene ring with 1,4-free attachment).

More details are given as supplementary material.

4. Conclusion

A new mathematical model for Computer-Aided Molecular Design (CAMD) problems is proposed in this paper. With this new mathematical model, first-order and second-order groups are considered in the MILP/MINLP formulation simultaneously, ensuring optimal solution is found and not a local optima. Chemical structure information and more precise target molecular properties are obtained from this model. This model can also be expanded with UNIFAC equations for the mixture/blend design problems easily.

References:

Churi, N., & Achenie, L. E. (1996). Novel mathematical programming model for computer aided molecular design. *Industrial & engineering chemistry research*, 35(10), 3788-3794.

Cignitti S., Zhang L., & Gani R. (2015). Computer-aided framework for design of pure, mixed and blended products. *Computer Aided Chemical Engineering*, 37, 2093-2098.

Constantinou, L., Bagherpour, K., Gani, R., Klein, J. A., & Wu, D. T. (1996). Computer aided product design: problem formulations, methodology and applications. *Computers & Chemical Engineering*, 20(6), 685-702.

Derringer, G. C., & Markham, R. L. (1985). A computer-based methodology for matching polymer structures with required properties. *Journal of Applied Polymer Science*, 30(12), 4609-4617.

Franklin, J. L. (1949). Prediction of heat and free energies of organic compounds. *Industrial & Engineering Chemistry*, *41*(5), 1070-1076.

Gani, R. (2004). Chemical product design: challenges and opportunities. *Computers & Chemical Engineering*, *28*(12), 2441-2457.

Gani, R., & Constantinou, L. (1996). Molecular structure based estimation of properties for process design. *Fluid Phase Equilibria*, 116(1), 75-86.

Gani, R., Jiménez-González, C., & Constable, D. J. (2005). Method for selection of solvents for promotion of organic reactions. *Computers & Chemical Engineering*, 29(7), 1661-1676.

Griva, I., Nash, S. G., & Sofer, A. (2009). *Linear and nonlinear optimization*. Siam.

Harper, P. M. (2001). Computer Aided Molecular Design - Problem Formulation and Solution: Solvent Selection and Substitution. NATO/CCMS Pilot Study Clean Products and Processes, Copenhagen, Denmark.

Harper, P. M., Gani, R., Kolar, P., & Ishikawa, T. (1999). Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria*, 158, 337-347.

Hukkerikar, A. S., Sarup, B., Ten Kate, A., Abildskov, J., Sin, G., & Gani, R. (2012). Group-contribution^{+}(GC^{+}) based estimation of properties of pure components: Improved property estimation and uncertainty analysis. *Fluid Phase Equilibria*, 321, 25-43.

Karunanithi, A. T., Achenie, L. E., & Gani, R. (2005). A new decomposition-based computer-aided molecular/mixture design methodology for the design of optimal solvents and solvent mixtures. *Industrial & Engineering Chemistry Research*, 44(13), 4785-4797.

Marrero, J., & Gani, R. (2001). Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*, 183, 183-208.

Marrero, J., & Gani, R. ProPred Manual, PEC02-15. 2002. *CAPEC Internal Report* (Technical University of Denmark, Lyngby, Denmark).

Mattei, M., Kontogeorgis, G. M., & Gani, R. (2014). A comprehensive framework for surfactant selection and design for emulsion based chemical product design. *Fluid Phase Equilibria*, 362, 288-299.

Odele, O., & Macchietto, S. (1993). Computer aided molecular design: a novel method for optimal solvent selection. *Fluid Phase Equilibria*, 82, 47-54.

Samudra, A. P., & Sahinidis, N. V. (2013). Optimization‐based framework for computer-aided molecular design. *AIChE Journal*, 59(10), 3686-3701.

Satyanarayana, K. C., Abildskov, J., & Gani, R. (2009). Computer-aided polymer design using group contribution plus property models. *Computers & Chemical Engineering*, 33(5), 1004-1013.

Table 1. Adjacency matrix for ethyl tert-butyl ether

Groups	ID	CH_3 1	CH_3 2	CH_3 3	CH_3 4	C 1	CH_2O 1
CH_3	1						1
CH_3	2					1	
CH_3	3					1	
CH_3	4					1	
C	1		1	1	1		1
CH_2O	1					1	

Table 2. Adjacency matrix of target molecular

Groups	ID	i_1 1	i_1 2	…	i_1 $n_{i_1}^U$	i_2 1	i_2 2	…	i_2 $n_{i_2}^U$	…	i_k 1	i_k 2	…	i_k $n_{i_k}^U$
i_1	1	0		…				…		…			…	
i_1	2		0	…				…		…			…	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
i_1	$n_{i_1}^U$			…	0		$y_{i_1,n_{i_1}^U,i_2,2}$	…		…			…	
i_2	1			…		0		…		…			…	
i_2	2			…	$y_{i_2,2,i_1,n_{i_1}^U}$		0	…		…			…	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
i_2	$n_{i_2}^U$			…				…	0	…			…	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
i_k	1			…				…		…	0		…	
i_k	2			…				…		…		0	…	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
i_k	$n_{i_k}^U$			…				…		…			…	0

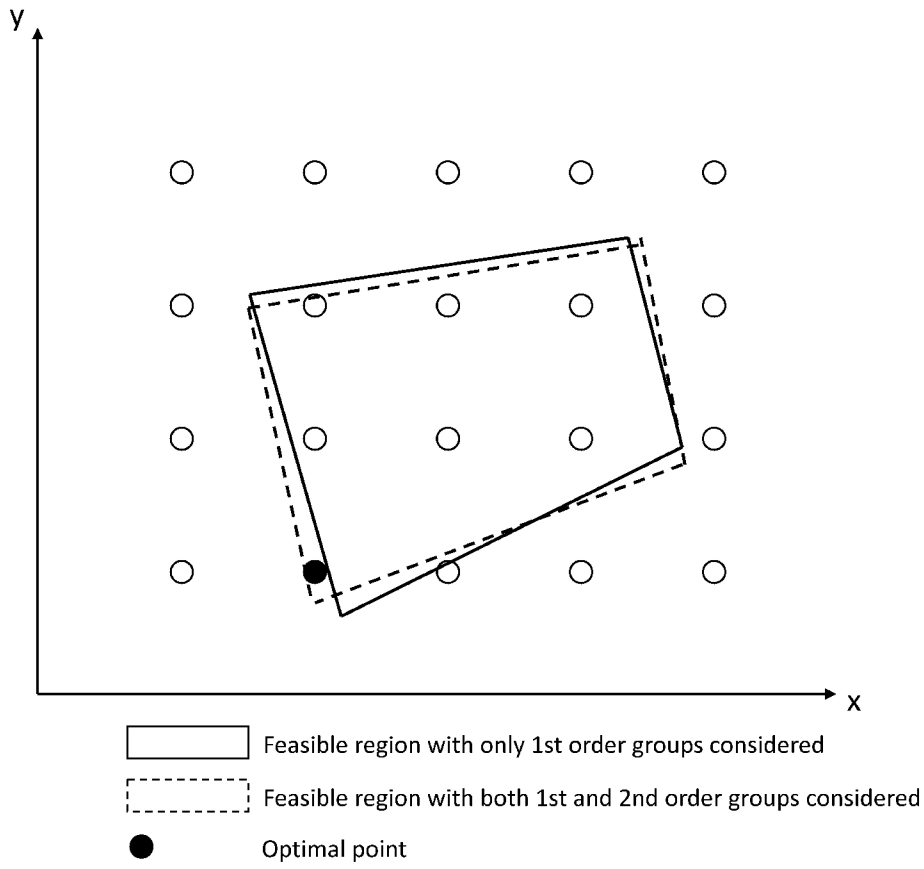

Figure 1. Feasible region of CAMD problems using different modeling approach

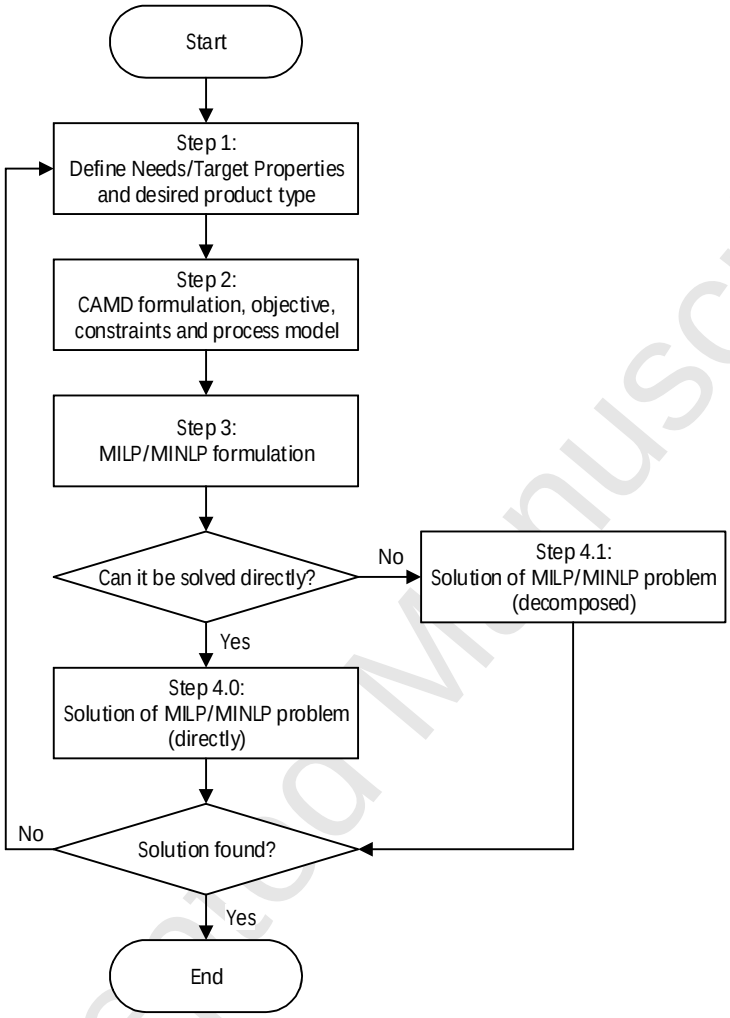

Figure 2. Computer-aided molecular design framework