Molecular Design Using Quantum Chemical Calculations for Property Estimation

Andreas Lehmann and Costas D. Maranas*

Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802

In this paper, we examine the combination of quantum chemical methods with optimization techniques for molecular design. A simple hydrofluorocarbon refrigerant design example and a solvent design example illustrate the proposed framework. The hydrofluorocarbon compounds are optimized for their heats of formation, and the potential solvents are searched for capacity, selectivity, and environmental safety. In both examples, a genetic algorithm is applied to generate and screen candidate molecules. The molecular properties are evaluated using a combination of quantum chemical calculations and group contribution methods. We assess the feasibility of the proposed approach for small molecules and find that establishing a proper tradeoff between the accuracy of the quantum chemical method and computational expense is vital.

1. Introduction

Molecular design is an iterative process of finding molecules that possess one or more desired properties. To save time and resources, it is of great interest to carry out part of this search for potentially new molecules on a computer rather than experimentally. For computational searching, the desired property must be calculated from a model describing the structure−property relationship. Previous molecular design studies[1−4] typically employed group contribution methods (GCMs)[5−7] for this purpose. In these methods, the parameters for the contribution of molecular groups to the property of the entire molecule (or polymer repeat unit) are obtained by fitting the group contribution model to experimental data for a set of chemical compounds. However, if a certain atom type, molecular group type, or type of chemical bonding is not present in the experimental set, the GCM will not account for this chemical information, thus limiting the predictive capabilities of these methods. Quantum chemical ab initio methods (calculations from first principles) obtain molecular properties from the most fundamental level of molecular information: the location of the nuclei and the number of electrons. From this input, in principle, molecular level information about any system can be predicted (e.g., molecular energies, electronic charge distributions, dipole and higher moments, vibrational frequencies, or molecular structure). Quantum chemical methods can provide molecular level properties with an accuracy that lies within the limits of experimental error. Although at the price of high computational cost, quantum chemical methods offer intriguing advantages for molecular design. They do not depend on a particular class of compounds, and as more methods for accessing properties (the combination of quantum calculations and subsequent evaluations) become available, the potential to predict the properties of *unknown* molecules is growing. Frequently, the accuracy of the results is limited by the computational time rather than by the chosen method. In this paper, we explore the use of ab initio methods for molecular design through two small example problems.

In the first problem, we address the molecular design of hydrofluorocarbons by minimizing the deviation between a target property and the property of the designed molecules. The ideal gas heat of formation ($\Delta H_f°$) at standard conditions is chosen as the target property, motivated by an interest in chemically stable hydrofluorocarbon refrigerants. The second example focuses on the design of solvents for liquid−liquid extraction. The optimized properties are the solvent limiting capacity $C_{\infty,A}$, the limiting selectivity $S_{\infty,A,B}$, and the environmental safety of the solvent, represented by the octanol−water partition coefficient K_{OW}. In this example, K_{OW} is evaluated by a quantum chemical method[8] while $C_{\infty,A}$ and $S_{\infty,A,B}$ are evaluated using the UNIFAC GCM.[7,9] We emphasize that these examples were selected primarily as benchmark problems to explore the applicability of our approach.

The solution of the quantum chemical problem as part of the property evaluation of a molecule is found by numerical methods and not by analytical structure−property relationships. Because of the ability to handle nonanalytical functions, a genetic algorithm (GA) was chosen as the optimization procedure. The GA generates molecules and treats the quantum chemical evaluation subroutine of the molecules as a black box. It is important to note that, because GAs are directed *random* search methods, one cannot claim global optimality. For this reason and because ab initio calculations are very time-consuming, it is important to tune the performance of the GA before engaging in costly quantum chemical property evaluations. Performance is governed by a number of adjustable parameters such as probabilities of crossover and mutation, population size, and number of generations.

For GA tuning, it is favorable to have a list of the globally best, second best, etc., molecules for evaluating how many of these globally best candidates the GA is able to find with a particular set of parameters. Generating such a list using quantum chemistry, however, is exactly the problem to be solved. It is much less time-consuming to generate the list when using GCMs for rapid property evaluation. In this study, we use two

* To whom correspondence should be addressed. Tel.: (814) 863-9958. Fax: (814) 865-7846. E-mail: costas@psu.edu.

10.1021/ie030120q CCC: $27.50
Published on Web 05/26/2004

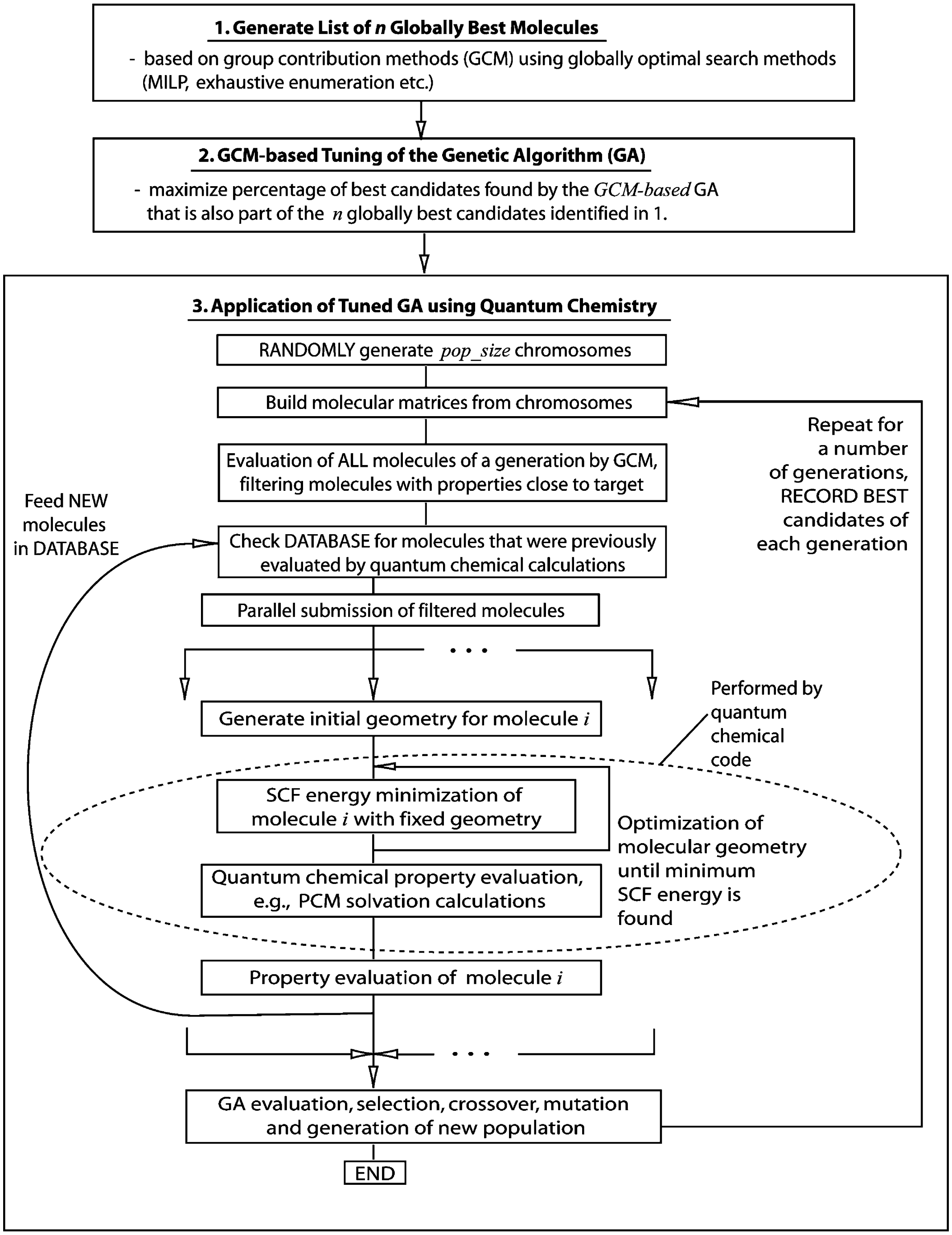
1. Generate List of n Globally Best Molecules
- based on group contribution methods (GCM) using globally optimal search methods (MILP, exhaustive enumeration etc.)
2. GCM-based Tuning of the Genetic Algorithm (GA)
- maximize percentage of best candidates found by the GCM-based GA that is also part of the n globally best candidates identified in 1.
3. Application of Tuned GA using Quantum Chemistry
RANDOMLY generate pop_size chromosomes
Build molecular matrices from chromosomes
Repeat for a number of generations, RECORD BEST candidates of each generation
Evaluation of ALL molecules of a generation by GCM, filtering molecules with properties close to target
Feed NEW molecules in DATABASE
Check DATABASE for molecules that were previously evaluated by quantum chemical calculations
Parallel submission of filtered molecules
. . .
Performed by quantum chemical code
Generate initial geometry for molecule i
SCF energy minimization of molecule i with fixed geometry
Optimization of molecular geometry until minimum SCF energy is found
Quantum chemical property evaluation, e.g., PCM solvation calculations
Property evaluation of molecule i
. . .
GA evaluation, selection, crossover, mutation and generation of new population
END

Figure 1. Flowchart of the proposed approach.

different approaches for GCM-based GA tuning. If the GCM is given in linear form or it can be transformed into a linear form,[2] a list of molecules can be generated by solving a GCM-based mixed-integer linear programming (MILP) form of the problem to global optimality and comparing the results to those of a *GCM-based* GA for the same molecular search space as in the original problem. This approach is applied to the first example. If the GCM cannot be given in linear form, as in our second example, one can sample either the entire search space or a reduced search space and find the GCM-based globally optimal molecules by exhaustive enumeration or other GCM-based search techniques.

Figure 1 illustrates our proposed approach. First, a list with *n* GCM-based globally best molecules is found either by solving an MILP model or by exhaustive enumeration (step 1). Next, we evaluate how many of these best solutions are found by the GA with a particular set of parameters (step 2). The GA is tuned such that it finds the maximal average number of globally best solutions when using the GCM for property evaluation. Step 3 describes the actual optimization procedure. All molecules are first screened using the GCM. Molecules with GCM values within a targeted range depending on the GCM's accuracy and the optimization goals are selected for further evaluation by the ab initio calculation. This reduces the number of candidate molecules that need to be evaluated in the computationally expensive ab initio step. The number of candidate molecules is further reduced by comparing them with a database and checking if a particular molecule was evaluated in a previous generation or GA run. In this case, the result is retrieved from the database. After their initial geometries are generated, the candidate molecules are submitted to a quantum chemical software. Note that the molecules are submitted individually, which results in a parallel execution of this most time-consuming step. After geometry optimization, the candidate molecules can be subjected to further quantum chemical evaluations as in the second example, where each candidate molecule is solvated in water and octanol, respectively. After the final properties are found, the fitness of the candidate molecules is evaluated by the GA. When this GA cycle is repeated for a number of generations, the best candidates are recorded. All program parts are linked. While the evaluation routines run, the GA pauses and waits for result files. If a candidate molecule causes problems, it can be treated individually and the result files are then submitted to the GA for continuation.

2. Theory

Quantum Chemistry. The objective of quantum chemistry is to find the best possible approximation of the wave function $\psi(r,\mathbf{R})$ in the electronic Schrödinger equation of a system containing one or more molecules.

$$\hat{H}\psi(r,\mathbf{R}) = E(\mathbf{R})\ \psi(r,\mathbf{R}) \qquad (1)$$

Here, $\hat{H}$ is the Hamiltonian operator, $E(\mathbf{R})$ is the energy of the system, $\mathbf{R}$ is the vector of the nucleic coordinates, and r refers to the coordinates of the electrons. The wave function $\psi(r,\mathbf{R})$ is postulated to contain all physical information of the system; once it is known, in principle, all other physical information can be obtained from it. Because of the complexity of eq 1, no analytical solution is known for systems of practical interest and ψ is approximated computationally. Most methods are based on the *variation principle*, which states that any trial wave function ϕ estimating the true ψ can only yield an energy $E \geq E_0$, where E_0 is the ground-state energy of the system. Variational methods such as the Hartree–Fock (HF) method or density functional theory (DFT) approximate ψ by varying ϕ until E converges to a local minimum, using the self-consistent-field (SCF) method. Following the *Born–Oppenheimer approximation*, nuclei are treated as fixed because they are about 1000 times heavier than electrons and, thus, move much slower than electrons. Therefore, E and ψ depend only parametrically on $\mathbf{R}$: for each $\mathbf{R}$, $\psi(r,\mathbf{R})$ is different. In a geometry optimization step, $\mathbf{R}$ is modified using a nonlinear gradient method until E converges to a local minimum. For each $\mathbf{R}$, the SCF procedure is solved again. Hence, there are two nested energy minimization problems (Figure 1, step 3), resulting in significant computational expense. Quantum chemical results from nonoptimized geometries are physically meaningless. For a thorough introduction to the large field of quantum chemistry, the reader is referred to the literature.[10–13]

Initial Geometry and Molecular Representation. For successful geometry optimization, nonlinear gradient methods require a good initial guess for $\mathbf{R}$. Most quantum chemical codes have a graphical user interface for building molecules to accompany the chemical intuition of the user. In an optimization problem, new molecules are generated by the solver; therefore, the initial guess for $\mathbf{R}$ must also be generated computationally. The quantum chemistry package Gaussian 98[14] provides the subroutine Model Builder,[15,16] which generates an initial geometry solely from molecular connectivity information by assigning standard bond lengths and bond angles to adjacent atoms according to their types. For representing molecules, a connectivity matrix $\mathbf{M}$ was chosen. In $\mathbf{M}$, each row contains information about one atom of the molecule. The first-column element m_{i1} represents the atom type through its atomic number. For example, for carbon, $m_{i1} = 6$. All other columns contain the connectivity information as binary variables: $m_{i,j+1} = 1$ if the atom in row i is connected to the atom in row j; $m_{i,j+1} = 0$ otherwise. For example, the molecular matrix for $F_2C{=}CFH$ is given by

$$\mathbf{M} = \begin{bmatrix} 6 & 0 & 1 & 1 & 1 & 0 & 0 \\ 6 & 1 & 0 & 0 & 0 & 1 & 1 \\ 9 & 1 & 0 & 0 & 0 & 0 & 0 \\ 9 & 1 & 0 & 0 & 0 & 0 & 0 \\ 9 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad (2)$$

Note that this representation only intrinsically encodes multiple bonds. This way of modeling is similar to the quantum chemical notion of the molecule as a collection of nuclei surrounded by an electron cloud. It is important to note that bonds are not an input but rather a result of the electronic structure calculation.

Optimization Procedure: GA. The molecular design problem is a discrete optimization problem: atom A is either connected to atom B or not, and it is of a given type (i.e., carbon) or not. Commonly used discrete optimization methods are mathematical programming methods or guided random search methods such as simulated annealing or GAs. Mathematical programming techniques[17] such as MILP require a complete analytical model of the optimization problem. Because the approximate solution of the Schrödinger equation is not available in analytical form, these methods cannot be used directly to solve the quantum chemical model of the molecular system. A GA, first introduced by Holland[18] and discussed by Michalewicz,[19] is used in this study. By maintaining a population of chromosomes and applying crossovers and mutations to them, a GA constantly samples the search space and is intrinsically parallel in nature.[19]

Chromosome Representation. The molecular candidates are represented as bit strings. These are one-dimensional fields of binary elements, i.e., chromosomes, on which the GA operates. For evaluating the fitness, chromosomes are translated into the molecular matrix $\mathbf{M}$ (eq 2). The combination of bit-string representation and translation procedure encodes the constraints of the optimization problem. After randomly creating a number of chromosomes, one obtains an initial population and can apply the GA, which consists of the following steps.

(a) Fitness Evaluation. For each chromosome, the fitness is a measure of how far this chromosome deviates from the optimization goal. In this study, the fitness expresses how far the property P_i of a particular chromosome i deviates from the given target P_{target}. For a vanishing deviation, the fitness F_i should approach 1, and it should decrease as the deviation increases. This behavior is achieved by defining F_i in form of a Gaussian

function:[3]

$$F_i = \exp\left[-\alpha \frac{(P_{\text{target}} - P_i)^2}{P_{\text{target}}{}^2}\right] \tag{3}$$

Here, α is a parameter of the algorithm, which determines how steeply the Gaussian curve decreases with larger deviations. If the design involves properties that have either a lower bound or an upper bound, we use a sigmoidal fitness function

$$F_i = \left\{1 + \exp\left[-\beta\left(\frac{P_i - P_{F=0.5}}{P_{\text{range}}}\right)\right]\right\}^{-1} \tag{4}$$

where $P_{F=0.5}$ is the property value for which the evaluated fitness is 0.5. The lower or upper limit of the property constraints is placed at this point.[3] Note that the limit is not a bound; i.e., values below the lower bound or above the upper bound can still yield acceptable fitness values, depending on the slope β of the resulting *S* curve. For multiple objectives, eqs 3 or 4 are formulated for each objective and the total fitness of each chromosome is obtained by averaging over all property objectives. From the fitness of each chromosome, the total fitness of the population is determined by

$$F_{\text{tot}} = \sum_{i=1}^{pop\_size} F_i \tag{5}$$

where *pop_size* is the size of the population.

(b) Selection. The selection process ensures that the fittest chromosomes have better chances for survival, while the unfit chromosomes do not die out immediately but only after a number of generations. A probability p_i for selection into the next generation is defined to be proportional[20] to the fitness of chromosome *i*:

$$p_i = F_i/F_{\text{tot}}$$

After calculation of a cumulative probability q_i,

$$q_i = \sum_{j=1}^{i} p_j$$

and generation of a random number *r* with $0 \le r < 1$, chromosome *i* is selected into the new population if $q_{i-1} < r \le q_i$. This is repeated *pop_size* times in order to obtain the population of the next generation.

(c) Crossover. Chromosomes of the new population are randomly selected for crossover with a probability p_c. During crossover, two chromosomes exchange their sequences of bits following a randomly selected position of the bit string.

(d) Mutation. Chromosomes of the new population are randomly selected for mutation with a probability p_m. Mutation is the change of one bit (0 → 1 or 1 → 0) of a chromosome at a randomly selected position of the bit string.

Tuning the GA. Several parameters influence the performance of the GA: the population size *pop_size*, the probability of crossover p_c, the probability of mutations p_m, and the number of generations $N_{\text{generations}}$. Before the GA is used to operate on a population of molecules that are evaluated by costly quantum chemical calculations, it is important to tune these parameters for optimal performance of the algorithm, which is a compromise between rapid convergence (evaluation of few molecules) and extensive sampling of the search space (evaluation of many molecules). If the algorithm converges too fast, the confidence in the solution is smaller. After finding the *n* best molecules using a rapid property evaluation method such as GCM, information about the performance of the GA can be obtained by evaluating how many of the best solutions the GA finds and at what computational expense. For this purpose, the GA is run for a number of generations while the same GCM is used for the property evaluation. Using the fitness F_i as a measure, a record of *n* all-time-best (GA) individuals is maintained. The GA is a random search technique, and it is unlikely that it finds all globally best molecules. Let *N* be the set of the *n* globally best molecules. The percentage $P_{n,\text{best}}$ of the *n* globally best molecules that the GA finds is given by

$$P_{n,\text{best}} = 100\nu/n \tag{6}$$

where ν is the number of molecules found by the GA that are elements of *N*. When the GCM-based GA is run *k* times, $P_{n,\text{best}}$ is averaged by

$$P_{n,\text{best,avg}} = 1/k\sum_{i=1}^{k} P_{n,\text{best}}(i) \tag{7}$$

In this study, we chose $n = 10$. For $k \ge 1000$, $P_{10,\text{best,avg}}$ fluctuates only in the first decimal place. Another important performance measure of the GA is how many function evaluations the algorithm needs to achieve a certain $P_{n,\text{best}}$, i.e., how many candidate molecules pass the filter discussed below. This is of interest because, ultimately, costly quantum chemical calculations are to be used for these function evaluations. The objective of the empirical tuning procedure is to maximize $P_{10,\text{best,avg}}$.

Filter. To avoid unnecessary evaluations and by application of a GCM, molecules are screened for proximity to the optimization goal if the property is to be evaluated by quantum chemistry. A filter only accepts molecules with

$$P_{\text{target}} - \Delta P \le P_{i,\text{GCM}} \le P_{\text{target}} + \Delta P \tag{8}$$

from being evaluated by the quantum chemical method. In eq 8, $P_{i,\text{GCM}}$ is the property *P* of chromosome *i*, based on a GCM calculation, and ΔP is a tolerance that ensures that candidate molecules whose $P_{i,\text{GCM}}$ lies outside of this tolerance are highly unlikely to have a P_i based on quantum chemistry within the tolerance. The choice of ΔP depends on the accuracies of both the GCM and the quantum chemical method.

3. Case Study 1: Design of Hydrofluorocarbons

3.1. Problem Definition. The first example is motivated by the search for alternative, chlorine-free refrigerants. One of the properties of interest here is the heat of formation $\Delta H_f°$ of the candidate substance, which is used as an indicator for the stability of the compound. To this end, molecules with $\Delta H_f°$ closest to a target value $\Delta H^\circ_{f,\text{target}}$ are identified using the GA-based optmization procedure.

Heat of Formation. The heat of formation is obtained using the method of Curtiss et al.,[21] by first evaluating the dissociation (atomization) energy $\sum D_0$, i.e., the energy difference between the molecule and its dissociated atoms. For example, for the molecule $A_xB_yH_z$,

$$\sum D_0 = x\epsilon_0\,(\mathrm{A}) + y\epsilon_0\,(\mathrm{B}) + z\epsilon_0\,(\mathrm{H}) - \epsilon_0(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z) - \epsilon_{\mathrm{ZPE}}(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z) \quad (9)$$

where $\epsilon_0(X)$ is the energy of the particle (atom or molecule) X and $\epsilon_{\mathrm{ZPE}}(A_xB_yH_z)$ is the zero-point energy of the molecule. Variables x, y, and z represent the number of atoms of A, B, or H in the example molecule. The zero-point energy is a part of the ground-state energy that accounts for molecular vibrations persisting even at 0 K. All values in eq 9 are calculated using Gaussian 98.[14] Note that the quantum chemical model must be the same for all particles X. The enthalpy of formation of the molecule at 0 K is given by

$$\Delta H_{\mathrm{f}}°(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z;\,0\ \mathrm{K}) = x\,\Delta H_{\mathrm{f}}°(\mathrm{A},\,0\ \mathrm{K}) + y\,\Delta H_{\mathrm{f}}°(\mathrm{B},\,0\ \mathrm{K}) + z\,\Delta H_{\mathrm{f}}°(\mathrm{H},\,0\ \mathrm{K}) - \sum D_0 \quad (10)$$

Here, the dissociation energy is subtracted from the sum of the widely accepted values for the 0 K heats of formation of gaseous atoms $\Delta H_f°(X,\ 0\ K)$, which are tabulated in work by Chase et al.[22] $\Delta H_f°(A_xB_yH_z,\ 0\ K)$ is also corrected for the standard state (298 K):

$$\Delta H_{\mathrm{f}}°(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z,\,298\ \mathrm{K}) = \Delta H_{\mathrm{f}}°(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z,\,0\ \mathrm{K}) + [H°(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z,\,298\ \mathrm{K}) - H°(\mathrm{A}_x\,\mathrm{B}_y\,\mathrm{H}_z,\,0\ \mathrm{K})] - x\,[H°(\mathrm{A},\,298\ \mathrm{K}) - H°(\mathrm{A},\,0\ \mathrm{K})]_{\mathrm{st}} - y\,[H°(\mathrm{B},\,298\ \mathrm{K}) - H°(\mathrm{B},\,0\ \mathrm{K})]_{\mathrm{st}} - z\,[H°(\mathrm{H},\,298\ \mathrm{K}) - H°(\mathrm{H},\,0\ \mathrm{K})]_{\mathrm{st}} \quad (11)$$

In this equation, $H°(A_xB_yH_z,\ 298\ K) - H°(A_xB_yH_z,\ 0\ K)$ is evaluated by the quantum chemistry software and $[H°(X,\ 298\ K) - H°(X,\ 0\ K)]_{st}$ are obtained from tabulated values.[22]

Quantum Chemical Model. A number of preliminary runs were first carried out in order to find a model chemistry that balances acceptable accuracy and computational expense. The $\Delta H_f°$ values were predicted for several hydrofluorocarbons using B3LYP/6-31G(d) for both the geometry optimization and the final energy calculation. Bauschlicher[23] reports an average error of 5.18 kcal/mol for $\sum D_0$ using this model chemistry on the G2 test set of 55 molecules.[24] When comparing results for $\sum D_0$ and $\Delta H_f°$, we neglect the influence of the average error in $H°(A_xB_yH_z,\ 298\ K) - H°(A_xB_yH_z,\ 0\ K)$ in eq 11 on the average error in $\Delta H_f°$. For the set of 21 molecules that were calculated for this work, we found an average error of 4.53 kcal/mol. The results obtained in our work are summarized in Table 1. Curtiss et al.[21] reported average errors of 2.43 kcal/mol for the 55 molecules of the G2 test set. However, these results were obtained using the B3LYP/6-311+G(3df,2p) model chemistry. This basis set was deemed too computationally expensive for the purpose of using it as a subroutine within an optimization loop. For comparison, the heat of formation values as calculated from the GCM are provided in Table 1. The average error for this method was found to be 12.23 kcal/mol.

Bit String Representation and Objective Function for Case Study 1. For the problem at hand, the search space is constrained to hydrofluorocarbons with a maximum of three carbon atoms and possible multiple bonds between them. The bit strings contain 10 digits, each of which can assume the value 0 or 1 (hence, the name "bits"). For example, a randomly generated string S may look like

$$S = (1\ \ 0\ \ 0\ \ 1\ \ 1\ \ 1\ \ 1\ \ 0\ \ 0\ \ 1)$$

The elements of S are defined as follows. The first two positions define the number of carbon atoms in the molecule by the equation

$$\text{no. of carbons} = 1 + S(1) \times 2^{S(2)}$$

The third and fourth positions define the number of bonds between the first and second carbon atoms based on the formula

$$\text{no. of bonds between } \mathrm{C}_1 \text{ and } \mathrm{C}_2 = S(3) \times 2^0 + S(4) \times 2^1$$

The case $S(3) = S(4) = 0$ encodes a double bond on each side of the middle atom in the case of a molecule with three carbon atoms, and if the molecule has only two carbon atoms, they are also connected by a double bond between them. The last six positions of S define the number of fluorine atoms on each carbon:

$$\text{no. of fluorines on } \mathrm{C}_1 = S(5) \times 2^0 + S(6) \times 2^1$$

$$\text{no. of fluorines on } \mathrm{C}_2 = S(7) \times 2^0 + S(8) \times 2^1$$

$$\text{no. of fluorines on } \mathrm{C}_3 = S(9) \times 2^0 + S(10) \times 2^1$$

The translation of the bit strings into connectivity matrices is described next. First, all carbon atoms are connected in a molecular backbone. In the next step, the bond multiplicities between carbon atoms are assigned, which permits the calculation of the number of remaining bonds at each carbon atom (three at most). These bonds are then filled with the number of fluorine atoms assigned to each carbon atom through the bit string. When all fluorine atoms are connected, the remaining bonds of each carbon atom are filled with hydrogen atoms. On the basis of these definitions, one recognizes S as a bit string representation of $F_2C{=}CFH$. Note that one cannot attach three fluorine atoms to the first carbon atom as required by $S(5) = S(6) = 1$ because the carbon atoms are connected by a double bond, which has higher priority. Thus, in this problem representation, two bit strings can result in the same molecular connectivity matrix, producing some redundancy. However, this formulation ensures that no bit string can result in an infeasible molecule. The objective function in this example is written in the form of eq 3 with a target value of $\Delta H^{\circ}_{\mathrm{f,target}} = -150$ kcal/mol.

GA Tuning: An MILP Model Based on a Group Contribution Method for Rapid Property Calculation. For tuning of the GA, a significantly faster evaluation method is needed. In this case study, a GCM proposed by Joback and Reid[6] was used as a surrogate for the heat of formation predictions based on quantum chemical calculations. The advantage of Joback and Reid's method is its simplicity, which stems from the assumption that each group i contributes additively by a value $\Delta h^{\circ}_{\mathrm{f},i}$ to $\Delta H^{\circ}_{\mathrm{f}}$ of the molecule. The approximating expression is

$$\Delta H^{\circ}_{\mathrm{f}} = 68.29 + \sum n_i \Delta h^{\circ}_{\mathrm{f},i} \quad (12)$$

where n_i is the number of occurrences of group i in the

Table 1. Experimental and Calculated Data for ΔH_f° of Various Hydrofluorocarbons[a]

formula	name	$\Delta H^\circ_{f,exp}$ [kcal/mol]	$\Delta H^\circ_{f,B3LYP/6-31(d)}$(abs dev) [kcal/mol]	$\Delta H^\circ_{f,GCM}$(abs dev) [kcal/mol]
CH_3F	fluoromethane	−56.00[22]	−55.02 (0.98)	−62.16 (6.16)
		−59.00[44]	(3.99)	(3.16)
CH_2F_2	difluoromethane	−107.7[22]	−107.20 (0.50)	−109.03 (1.33)
		−108.08 ± 0.22[45]	(0.88)	(0.95)
CHF_3	trifluoromethane	−166.6[22]	−168.41 (1.81)	−157.17 (9.44)
		−165.1[46]	(3.31)	(7.94)
		−166.21[47]	(2.20)	(9.05)
$CF{\equiv}CH$	fluoroethyne	30.0[48]	28.45 (1.55)	2.67 (27.33)
$CFH{=}CH_2$	fluoroethene	−33.2[49]	−32.27 (0.93)	−37.12 (3.92)
		−32.4[50]	(0.13)	(4.72)
$CF_2{=}CH_2$	1,1-difluoroethene	−78.62[51]	−84.79 (6.17)	−86.33 (7.71)
		−82.19 ± 2.40[52]	(2.60)	(4.14)
		−77.7 ± 0.8[47]	(7.09)	(8.63)
		−79.83 ± 0.2[53]	(4.96)	(6.50)
$CF_2{=}CHF$	trifluoroethene	−113.3 ± 2.0[52]	−121.73 (8.43)	−135.16(21.86)
$CF_2{=}CF_2$	tetrafluoroethene	−157.4[49]	−166.38 (8.98)	−184.37 (26.97)
		−157.4[48]	(8.98)	(26.97)
		−157.9 ± 0.8[54]	(8.48)	(26.47)
		−157.9 ± 0.8[53]	(8.48)	(26.47)
		−164.0[55]	(2.38)	(20.37)
		−162.0 ± 1.0[56]	(4.38)	(22.37)
$CFH_2{-}CH_3$	fluoroethane	−63.19[57]	−63.76 (0.57)	−67.09 (3.90)
$CF_2H{-}CH_3$	1,1-difluoroethane	−118.8[58]	−120.12 (1.32)	−115.23 (3.57)
		−118.78 ± 0.95[59]	(1.34)	(3.55)
$CF_3{-}CH_3$	1,1,1-trifluoroethane	−176.0[51]	−182.01 (6.01)	−162.93 (13.07)
		−178.94 ± 0.76[60]	(3.07)	(16.01)
		−178.9 ± 0.4[61]	(3.11)	(15.97)
$CF_3{-}CFH_2$	1,1,1,2-tetrafluoroethane	−214.1[62,63]	−219.70 (5.60)	−209.80 (4.30)
$CF_2H{-}CF_2H$	1,1,2,2-tetrafluoroethane	−213.3[57]	−211.15 (2.15)	−210.23 (3.07)
$CF_3{-}CF_2H$	pentafluoroethane	−263.0[58]	−270.96 (7.96)	−257.93 (5.07)
$CF_3{-}CF_3$	hexafluoroethane	−321.2[48]	−329.06 (7.86)	−305.63 (15.57)
		−321.2 ± 1.2[64]	(7.86)	(15.57)
		−318[65]	(11.06)	(12.37)
		−318.0 ± 2.0[66]	(11.06)	(12.37)
		−321.22 ± 0.96[56]	(7.48)	(15.59)
$CFH{=}CH{-}CH_3$	*trans*-1-fluoro-1-propene	−41.3[67]	−38.64 (2.66)	−44.01 (2.71)
$CFH{=}CH{-}CH_3$	*cis*-1-fluoro-1-propene	−42.1[67]	−39.18 (2.92)	−44.01 (1.91)
$CH_2{=}CH{-}CF_3$	3,3,3-trifluoropropene	−146.8[58]	−150.04 (3.24)	−137.88 (8.92)
		−146.79 ± 1.6[68]	(3.25)	(8.91)
		−144.5 ± 1.6[69]	(5.54)	(6.62)
$CF_2{=}CF{-}CF_3$	hexafluoropropene	−275.26[70]	(7.55)	(9.88)
$CH_3{-}CF_2{-}CH_3$	2,2-difluoropropane	−129.8 ± 3.0[71]	−131.00 (1.20)	−120.99 (8.81)
$CF_3{-}CF_2{-}CF_3$	octafluoropropane	−426.2[72]	−430.88 (4.68)	−406.40 (19.80)
		−426.55 ± 2.10[73]	(4.32)	(20.15)

[a] Values in parentheses show absolute deviations from experimental values.

molecule. This simple additivity form results in a relatively straightforward MILP formulation,[25,26] the details of which are given in the appendix. The key advantage of this MILP formulation is that efficient solvers such as CPLEX or OSL (accessed via GAMS[27,28]) can identify the globally optimal molecule that is closest to the given target value. In addition, the MILP framework can be used to generate a list of *n* best solutions,[2] knowledge of which provides the basis for GA tuning.

3.2. Results of Case Study 1. Solutions of the MILP Model. Solving the MILP model based on the GCM for a target of $\Delta H^\circ_{f,target} = -150$ kcal/mol (Figure 1, step 1) produced a list of 10 molecules (Table 2) that are closest to the target value. Table 2 provides the basis for evaluating the GA when using the GCM for property evaluation. Based on the additivity assumption, the GCM cannot predict if a particular conformation of atoms in a molecule will be energetically stable. Therefore, there is no certainty if all of the molecules in Table 2 can exist under standard conditions. The molecules ranked 3rd, 9th, and 10th, however, do exist as indicated in Table 1.

Table 2. Ranking of 10 Molecules Closest to a Target Value of $\Delta H^\circ_{f,target} = -150$ kcal/mol, Calculated from the GCM-Based MILP Model

rank	molecule	$\Delta H^\circ_{f,GCM}$ [kcal/mol]
1	$CF{\equiv}C{-}CF_3$	−149.651
2	$CF_2{=}C{=}CF_2$	−150.399
3	CF_3H	−157.165
4	$CF_2{=}CF{-}CH_3$	−142.433
5	$CF_2{=}CH{-}CFH_2$	−140.093
6	$CH_2{=}CF{-}CF_2H$	−139.393
7	$CFH{=}CH{-}CF_2H$	−139.015
8	$CFH_2{-}CF_2H$	−162.098
9	$CH_2{=}CH{-}CF_3$	−137.882
10	$CH_3{-}CF_3$	−162.928

GA Tuning Based on the GCM. Figures 2 and 3 show the evolution of the normalized total fitness of the population (F_{tot}, eq 5, divided by *pop_size*). In these runs, the parameters of the GA were varied. All runs exhibit the expected behavior: the normalized total fitness of the population is increasing. A GA is converged if the total fitness assumes a value that cannot be improved over a reasonable number of generations. The examples typically show convergence after 10−20 generations. The GA tries to maximize the total fitness of a population; it does not strive for a great diversity of specific individuals. Therefore, a converged population frequently has a high number of equal individuals with

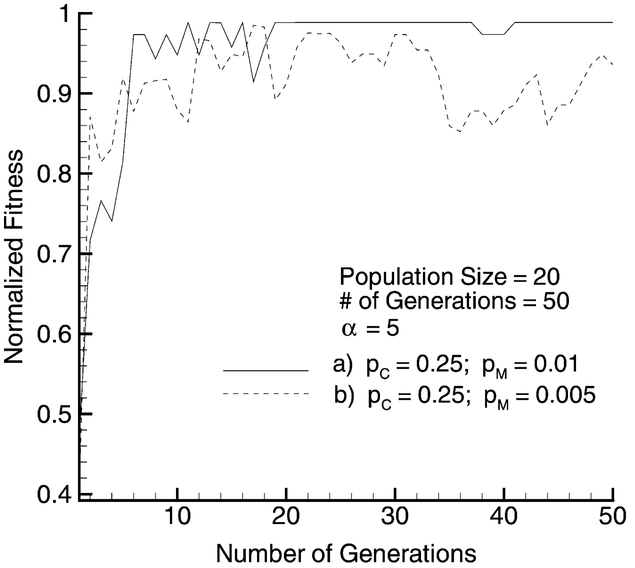

Figure 2. Development of the mormalized total fitness of a population of 20 molecules for different GA parameters. If p_m is too small (case b), the fitness of the population may deteriorate. The population of case a is converged with many identical molecules.

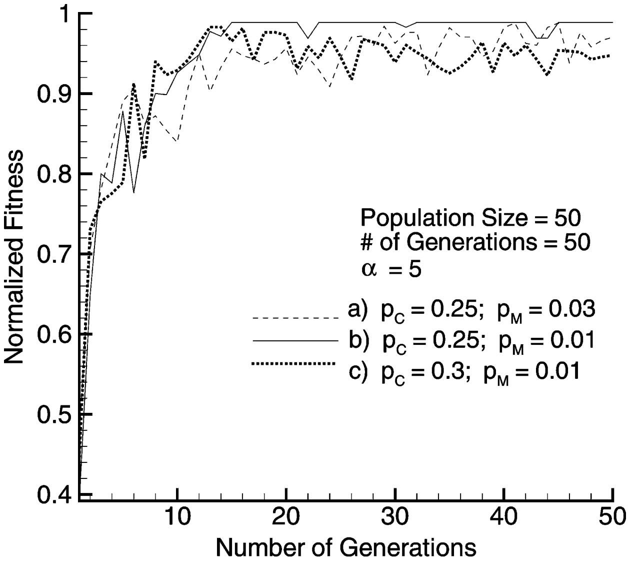

Figure 3. Development of the normalized total fitness of a population of 50 molecules for different GA parameters. The total fitness of populations a and c fluctuates stronger than in case b, indicating little diversity of population b. Populations a and c are better candidates for a set of GA parameters that ensures sufficient sampling of the search space.

fitness values close to 1, and the algorithm may converge into a suboptimal population with little probability of improving. Figures 2 and 3 show that the runs with parameters $p_c = 0.25$ and $p_m = 0.01$ (solid curves, Figures 2a and 3b) quickly converged to populations with a high normalized total fitness and exhibited a small fluctuation around this value. This is attributed to few changes in these populations after 20 generations. This set of parameters is unfavorable because it does not generate enough diversity in its populations to sample a variety of molecules. The dashed curves in Figures 2 and 3 show larger fluctuations. This can be attributed to a more random behavior of the algorithm, which frequently produces below-average individuals but also bears a higher possibility of finding above-average individuals.

Because the best molecules are sought as solutions of the optimization problem, it is not as important to find a good final population but to find good individuals in the course of running the algorithm. As indicated above, the parameter combination $p_c = 0.25$ and $p_m = 0.01$ did not promise to maintain a successful (high normal fitness) and diverse population, whereas other parameter combinations did (e.g., Figure 3a,c). For the decision on the set of GA parameters, a large number of parameter combinations were evaluated by trial and error while recording $P_{10,\text{best,avg}}$ and the number of function evaluations. Table 3 shows the performance of the GA for an average set of parameters (upper part) and the final set of parameters (lower part). The GA was run 1000 times for all sets of parameters that were tested to average over the different outcomes arising from the random search method. The frequency of occurrence of crossovers and mutations agrees well with their probabilities, for example, in the upper part, population size × number of generations × crossover probability = number of crossovers = 150. This value is close to the computational average of 149.794 over 1000 runs. The set of parameters $p_c = 0.25$ and $p_m = 0.01$ showed only an average performance ($P_{10,\text{best,avg}} = 55.49\%$) as expected from the discussion of Figures 2 and 3. The lower part of Table 3 shows the set of parameters that was chosen as the best performing set. As evidenced by the high mutation probability ($p_m = 0.03$), the total fitness of this GA fluctuated, but with almost 9 out of 10 best molecules, this GA exhibited a very good performance. As one would expect, better sampling of the search space requires more function evaluations, which is also observed in Table 3.

GA Based on DFT Calculations. Table 4 lists the 10 best molecules that were found by the GA, using Gaussian 98 with the B3LYP/6-31G(d) model chemistry for the calculation of $\Delta H_f°$. Five of these molecules were also found by the GCM-based MILP. Their ranks in Table 2 are listed in parentheses in Table 4. Note that the molecule from Table 1 that seems closest to the target value, 3,3,3-trifluoropropene ($CH_2{=}CH{-}CF_3$), was found by the DFT-based GA. The molecules ranked fourth, eighth, and ninth in Table 4 are also listed in Table 1, indicating that they can be synthesized. The list was compiled from the results (10 best molecules) of three runs of the GA. Each run took about 2.5−3 h of wall-clock time with evaluation times between 248.6 and 3617.2 CPU s/molecule. The Gaussian 98 evaluations run parallel in each generation.

A number of molecules presented problems to the Model Builder of Gaussian 98 because the subsequent geometry optimization did not converge. The cause lies in the nature of the Model Builder,[16] which was designed to account for *most* cases involving organic molecules. It is not guaranteed to give reasonable initial geometries in all cases. For molecules that created this problem (temporarily no result files available for the GA), the GA paused and the initial geometry was improved manually. After the result was submitted to the GA, the algorithm continued. Note that molecules ranked third and sixth in Table 4 have the same configuration but not the same conformation. Their molecular matrices (eq 2) are different, which caused the Model Builder to generate different initial geometries, converging into different local optima in the geometry optimization.

Table 3. GA Performance for Different Parameter Sets, Averaged over a Number of 1000 Runs for Each GA

GA parameter		average of 1000 runs	
population size	20	percentage of 10 best molecules	55.49%
no. of generations	30	no. of crossovers	149.794
crossover probability	0.25	no. of mutations	6.121
mutation probability	0.01	function evaluations after filtering	8.762
population size	50	percentage of 10 best molecules	88.56%
no. of generations	10	no. of crossovers	124.4
crossover probability	0.25	no. of mutations	14.952
mutation probability	0.03	function evaluations after filtering	15.637

Table 4. Ranking of 10 Molecules Closest to a Target Value of $\Delta H^\circ_{f,target} = -150$ kcal/mol, Calculated from the Quantum Chemical GA

rank	molecule	$\Delta H^\circ_{f,B3LYP/6-31G(d)}$ [kcal/mol]
1 (9)	$CH_2{=}CH{-}CF_3$	−150.004
2 (2)	$CF_2{=}C{=}CF_2$	−138.093
3	$CFH_2{-}CH_2{-}CF_2H$	−166.214
4	$CF_2{=}CF_2$	−166.382
5 (4)	$CF_2{=}CF{-}CH_3$	−132.754
6	$CF_2H{-}CH_2{-}CFH_2$ (conformer with 3)	−167.397
7 (1)	$CF{\equiv}C{-}CF_3$	−132.403
8 (3)	CF_3H	−168.408
9	$CH_3{-}CF_2{-}CH_3$	−130.983
10	$CFH_2{-}CF_2{-}CH_3$	−169.939

4. Case Study 2: Solvent Design

4.1. Problem Definition. The objective of this case study is to identify a solvent for the liquid−liquid extraction of benzene from cyclohexane. Solvent selection for liquid−liquid extraction or extractive distillation is based on various solvent criteria such as solvent selectivity, capacity, cost, safety requirements, wastewater load, environmental requirements, and several more. Out of these, we exemplarily chose capacity, selectivity, and environmental fate as guiding properties for the solvent design case study. The solvent capacity C_A is a measure of how well the solvent S can dissolve a solute A that is to be recovered from the mixture A−B, where B is the carrier. Solvent selectivity $S_{A,B}$ indicates the preference of S for A compared to B. For a "first-order" characterization of solvents, it is common practice to apply the limiting values of C_A and $S_{A,B}$ at infinite dilution of A and B (e.g., Hradetzky et al.[29]):

$$C_{\infty,A,S} = 1/\gamma_{\infty,A,S} \tag{13}$$

$$S_{\infty,A,B} = \gamma_{\infty,B,S}/\gamma_{\infty,A,S} \tag{14}$$

For most solvents, capacity and selectivity are competing properties (solvents with high selectivity show only low capacity and vice versa), which lead to the formulation of their product ω as a more realistic basis for evaluation:

$$\omega = C_{\infty,A,S}S_{\infty,A,B} \tag{15}$$

The environmental impact of a solvent has been correlated to the octanol−water partition coefficient K_{OW}.[30] It is a measure of the hydrophobicity because it describes the equilibrium partition between water and a nearly water-immiscible liquid phase. Also, 1-octanol is a good surrogate for the lipids in aquatic and animal biota and the organic matter in soils and sediments. Furthermore, increasing values of K_{OW} have been found to correlate with increasing bioaccumulation in the food chain (Lin and Sandler[31]). The same authors have correlated K_{OW} to the ratio of infinite-dilution activity coefficients in pure water and pure 1-octanol:

$$\log K_{OW,i} = b + a \log(\gamma_i^{W,\infty}/\gamma_i^{O,\infty}) \tag{16}$$

with $a = -0.68$ and $b = 0.91$. In the same paper, the authors evaluated log K_{OW} for 40 compounds using quantum chemical solvation calculations and a previously developed group contribution solvation (GCS) model,[8] which will be discussed in the next paragraph. On the basis of the ideas of this GCS model, they also devised a GCM for the rapid calculation of K_{OW} (GCSKOW). In our example problem, we evaluate $C_{\infty,A,S}$ and $S_{\infty,A,B}$ using the UNIFAC group contribution model and K_{OW} using quantum chemistry (GCS) or, for filtering molecules, with the GCSKOW model.

Infinite-Dilution Partition Coefficients. Lin and Sandler[8] developed a method to obtain infinite-dilution partition coefficents $\gamma^\infty_{S/1}/\gamma^\infty_{S/2}$ based on complex quantum chemical solvation calculations. These coefficients are a measure of how a solute S at infinite dilution partitions between solvents 1 and 2. Lin and Sandler's approach is based on the idea of combining the UNIQUAC[32] activity coefficient (γ) model with the free energy of solvation (ΔG^{sol}), which is available from quantum chemistry. ΔG^{sol} is found by assuming that a single solute molecule is placed into a solvent, which is modeled as an electric continuum represented by four physical constants: dielectric constant, ionization potential, refractive index, and density. Models based on this assumption are called continuum solvation models.[33] The solvation free energy change of a molecule placed in a fixed position into the continuum solvent ΔG^{*sol} (the asterisk indicates the fixed position) consists of four components, assuming molecular rotation and vibration effects are neglected:[8,33]

$$\Delta G^{*sol} = \Delta G^{cav} + \Delta G^{el} + \Delta G^{dis} + \Delta G^{rep} \tag{17}$$

with

$$\Delta G^{chg} = \Delta G^{el} + \Delta G^{dis} + \Delta G^{rep}$$

The cavitation contribution ΔG^{cav} is the work needed to form a sufficient cavity in the solvent for transferring a molecule from the gas phase into the solvated state. The electrostatic component ΔG^{el} represents the contribution from the electrostatic charge distribution that arises on the molecular surface and its electrostatic interaction with the solvent. The dispersion contribution ΔG^{dis} results from London dispersion attractions between the solute and solvent. The repulsion contribution ΔG^{rep} results from quantum-mechanical repulsions between the solute and solvent. For a more in-depth explanation of these terms, the reader is referred to the literature.[10,33] Note that the three latter components arise because of the charges of a molecule, while the first is due to the molecular size and shape. Lin and

Sandler summarize the latter three terms as the charging free energy $\Delta G^{\rm chg}$. The UNIQUAC activity coefficient model also distinguishes between a *combinatorial* term that is based on the size and shape of the interacting molecules and a *residual* term that accounts for the molecular interaction:

$$\ln \gamma_i = \ln \gamma_i^{\rm comb} + \ln \gamma_i^{\rm res}(u_{ii},u_{ij},u_{ji}) \tag{18}$$

In eq 18, the interaction parameters u_{ii}, u_{ij}, and u_{ji} need to be determined from experimental data. Lin and Sandler[8] combined $\Delta G^{\rm chg}$ and the UNIQUAC model, eliminating the parameters u_{ii}, u_{ij}, and u_{ji} in order to avoid the need for experimental data. They arrived at the following expressions for the infinite-dilution activity coefficient of a solute S in a solvent 1 and for the infinite-dilution partition coefficient of a solute S in a solvent 1 and a solvent 2, respectively

$$RT \ln \gamma_{\rm S/1}^{\infty} = RT \ln \gamma_{\rm S/1}^{\infty,\rm comb} + RTq_{\rm S}(\tau_{\rm S} - \tau_1) + (\Delta G_{\rm S/1}^{\rm chg} - \Delta G_{\rm S/S}^{\rm chg}) \tag{19}$$

$$RT \ln(\gamma_{\rm S/1}^{\infty}/\gamma_{\rm S/2}^{\infty}) = RT \ln(\gamma_{\rm S/1}^{\infty,\rm comb}/\gamma_{\rm S/2}^{\infty,\rm comb}) + RTq_{\rm S}(\tau_2 - \tau_1) + (\Delta G_{\rm S/1}^{\rm chg} - \Delta G_{\rm S/2}^{\rm chg}) \tag{20}$$

with

$$\frac{\Delta G_{i/i}^{\rm chg}}{RT} = q_i(\tau_i - 1 + \ln \tau_i) \tag{21}$$

Equation 21 can be used to access τ_i from $\Delta G_{i/i}^{\rm chg}$ of a solvent molecule solvated in a dielectric continuum of itself. Here, q_i is the relative van der Waals surface of component i. The charging free energies $\Delta G_{{\rm S}/i}^{\rm chg}$ can be obtained from quantum chemical continuum solvation calculations of a single molecule S solvated in a solvent i. It should be noted that the molecular structure parameters r_i and q_i of the UNIQUAC-based combinatorial term in eqs 19 and 20 are obtained from optimized molecular geometries of solutes and solvents, that is, from quantum chemistry. Unfortunately, activity coefficients calculated from this model do not agree well with experimental activity coefficients.[8] Lin and Sandler adjusted a scale factor α that accounts for the size of each atom in the solvation calculation according to the functional group that the atom belongs to. These scale factors were adjusted for different solvents (water, acetonitrile, *n*-octanol, and *n*-hexane) until sufficient agreement with experimental data was achieved. For example, a carbon in a $-{\rm CH}_X$ group will be assigned a different α value than a carbon in a −CN group. Also, the α value of a carbon in a −CN group in water is different from a carbon in a −CN group in *n*-octanol. Lin and Sandler[8] found that infinite-dilution partition coefficients $\gamma_{\rm S/1}^{\infty}/\gamma_{\rm S/2}^{\infty}$ were in good agreement with experimental data. Combining eqs 20 and 16 yields a quantum chemistry based model for $K_{\rm OW}$ with only one adjustable parameter α per group.

Quantum Chemical Continuum Solvation Model. For generating starting geometries, the molecular matrix **M** (eq 2) was submitted to the Model Builder of Gaussian 98, followed by an inexpensive HF/STO-3G geometry optimization. For obtaining the charging free energies, the GCS method of Lin and Sandler[8] was used, including the same quantum chemistry software: General Atomic and Molecular Electronic Structure System[34] (GAMESS). The equilibrium geometry of the solute in a vacuum was obtained by a geometry optimization using HF with a Dunning−Hay double-ζ valence (GAMESS: DZV) basis[35] with added polarization functions (DZP). Furthermore, one additional set of diffuse and polarization functions was added with exponents of one-third of the exponent of each most diffuse DZP function, respectively, giving rise to a DZPsp(df) basis. The same vacuum equilibrium geometry was used in subsequent solvation calculations without further optimization. The solvation calculations were carried out at the same HF/DZPsp(df) level.

Bit String Representation and Objective Function for Case Study 2. The search space for this example includes saturated aliphatic molecules with up to five carbon atoms in the backbone and the following functional groups: N(−C≡N), O(>C═O), −OH, $-{\rm NO}_2$, and $-{\rm Cl}_n$ (n = 1−3). Each molecule can contain one of these functional groups or none; multiple functional groups are excluded. The bit string contains eight digits. Similar to the bit strings in the first example, the first three positions define the number of carbon atoms, positions four, five, and six form the code for the functional group type, and positions seven and eight encode the carbon atom to which the functional group is attached. The following constraints are implemented in the routine that translates the bit strings into the molecular matrix **M**: For avoiding redundancy, functional groups can only be attached to a carbon atom number C_i with

$$C_i \le [C_{\rm total}/2] + 1 \tag{22}$$

where $C_{\rm total}$ is the total number of carbon atoms in the molecule and [x] stands for the integer part of a real number x. For example, if the candidate molecule has five carbon atoms, the functional groups can only be attached to carbon atoms number one, two, or three. Groups N(−C≡N) and $-{\rm Cl}_3$ can only be attached to the first carbon atom, and O(>C═O) cannot be attached to the first carbon atom. This limitation of group O(>C═O) arises from Lin and Sandler's method, in which α values for ketones are parametrized, but not those for aldehydes. The objective function in this example is written in the form of eq 4 with a lower limit of $\omega = -1.5$, a slope of $\beta_\omega = 1.0$, an upper limit of log $K_{\rm OW}$ = 2.0 and a slope of $\beta_{\log K_{\rm OW}} = 7.0$. Thus, the algorithm favors individuals with high ω and low log $K_{\rm OW}$ values.

4.2. Results of Case Study 2. Accuracy of the Quantum Chemical Model. When attempting to reproduce data of Lin and Sandler,[31] we found that the accuracy of their model cannot fully be reproduced because the results of the continuum solvation calculation are sensitive to the Cartesian coordinates of the input geometry. For different initial guesses, the geometry optimization using the DZPsp(df) basis will find geometries with almost identical internal coordinates but with different Cartesian coordinates. In the continuum solvation calculation, the grid for the molecular cavity is set up based on Cartesian coordinates of the nuclei. Differences in the grid can lead to differences in the $\Delta G^{\rm el}$, $\Delta G^{\rm dis}$, and $\Delta G^{\rm rep}$ terms of eq 17 on the order of 0.1−0.2 kcal/mol each.[36] Because $K_{\rm OW}$ depends on the difference between two charging free energies (eqs 16 and 20), this uncertainty can be magnified or canceled when repeating the calculation with different initial Cartesian coordinates. Table 5 shows a comparison

Table 5. Comparison of Quantum Chemical K_{OW} Results from Solvation Calculations for Case Study 2

compound	$\Delta\Delta G^{\mathrm{chg}\ 31}_{\mathrm{Lin}}$ [kcal/mol]	log $K_{\mathrm{OW,Lin}}$	$\Delta\Delta G^{\mathrm{chg}}_{\mathrm{this\ work}}$ [kcal/mol]	log $K_{\mathrm{OW,this\ work}}$	log $K_{\mathrm{OW,exp}}$
methane	1.884	1.07	1.889	0.956	1.09[74]
ethane	3.004	1.84	2.994	1.674	1.81[75]
propane	3.77	2.39	3.981	2.322	2.36[74]
butane	4.55	2.96	4.832	2.902	2.89[74]
methanol	−0.927	−0.81	−1.163	−1.106	−0.77[75]
ethanol	−0.013	−0.17	−0.031	−0.363	−0.31[76]
propanol	0.762	0.39	1.034	0.340	0.25[76]
nitromethane	−0.022	−0.13	−0.001	−0.326	−0.33[76]
nitroethane	0.640	−0.33	0.734	0.150	0.18[76]
nitropropane	1.199	0.74	1.799	0.897	0.87[76]
acetonitrile	−0.650	−0.55	−0.663	−0.701	−0.34[76]
propionitrile	0.251	0.09	0.341	−0.031	0.16[76]
acetone	−0.185	−0.23	−0.102	−0.235	−0.24[76]
chloropropane	3.145	2.07	3.613	2.184	2.04[76]
dichloromethane	2.097	1.38	2.252	1.343	1.25[76]
1,1-dichloroethane	2.623	1.79	2.891	1.802	1.79[75]
chloroform	2.814	1.98	2.584	1.699	1.9[76]
1,1,1-trichloroethane	3.185	2.31	3.173	2.156	2.49[75]
tetrachloromethane	3.869	2.82	3.085	2.181	2.64[77]
2-propanol	0.297	0.1	1.102	0.403	0.05[76]
2-chloropropane	3.041	2.02	3.420	2.073	1.9[76]
rms		0.114		0.199	

between Lin and Sandler's results[31] and data that were reproduced for this work. The data can be reproduced only with a root-mean-square (rms) of 0.199 log K_{OW} units as opposed to the authors' rms of 0.114 (for the molecules shown). This difference in accuracy is attributed to the optmization of the group scale factor α (section 4.1), which was performed based on the authors' Cartesian coordinates for each optimized molecular geometry. These coordinates are different from the Cartesian coordinates obtained for this work because in the optimization procedure initial coordinates are generated automatically. For comparison, among other properties, K_{OW} data were used for the parametrization of the widely used COSMO-RS[37] statistical mechanics method, which is based on screening charge distributions obtained from quantum chemical COSMO[38] calculations. Eckert and Klamt[39] reported an rms deviation of 0.471 log K_{OW} units for a parametrization set of 301 compounds.

Best Molecules and GA Tuning Based on UNIFAC and GCSKOW Group Contribution Methods. Solving the fitness for the entire search space, which encompasses 52 molecules for this small example, yielded the 10 globally best compounds given in Table 6. On the basis of an analysis similar to the one in section 3.2, the following GA parameters were chosen for case study 2: *pop_size* = 30, $N_{\mathrm{generations}} = 20$, $p_c =$ 0.25, and $p_m = 0.03$.

GA Based on Continuum Solvation Calculations. The tuned GA generated almost the same list of 10 best candidate solvents as the one in the previous step. The resulting compounds are also listed in Table 6. The only difference is that 2-pentanone and 2-nitropropane switched ranks. The *ω* values in both cases are the same because capacity and selectivity were evaluated using UNIFAC in both cases. The log K_{OW} values differ because they were obtained from different methods. Although the quantum chemical method (rms = 0.199) is less accurate than the GCSKOW model (rms = 0.14), the trends are clearly the same. For comparison, experimental log K_{OW} values are provided. It is noteworthy that nitromethane, a classic solvent for the benzene−cyclohexane system,[40] is among the suggested candidate solvents. The difference in computational expense is striking: the GCM-based GA runs in less than 1 min, and the quantum chemistry based GA runs in 18−36 h.

Table 6. Case Study 2: the 10 Globally Best Molecules Based on GCM and the 10 Best Molecules Based on Quantum Chemical Solvation Calculations[a]

			log	log	log
rank	compound	ω_{UNIFAC}	$K_{\mathrm{OW,GCSKOW}}$	$K_{\mathrm{OW,GCS}}$	$K_{\mathrm{OW,exp}}$
1	acetonitrile	1.800	−0.380	−0.701	−0.34[76]
2	acetone	1.668	−0.114	−0.236	−0.24[76]
3	propionitrile	1.895	0.150	−0.031	0.16[76]
4	2-butanone	2.106	0.416	0.256	0.29[76]
5	nitromethane	0.860	−0.154	−0.326	−0.33[76]
6	nitroethane	1.264	0.375	0.150	0.18[76]
7	methanol	0.409	−0.697	−1.106	−0.77[75]
8	butyronitrile	1.962	0.679	0.605	0.60[76]
9	2-nitropropane	1.573	0.755		0.93[78]
10	2-pentanone	2.430	0.945		0.84[75]
9	2-pentanone	2.430		0.720	0.84[75]
10	2-nitropropane	1.573		0.729	0.93[78]

[a] The rankings are the same from 1 to 8, and the two last compounds switch ranks when changing the log K_{OW} evaluation method.

5. Summary and Conclusions

In this work, we explored the use of ab initio calculations for the property evaluation in molecular design. Two case studies were presented to assess the feasibility of the proposed approach. The first example involves the design of hydrofluorocarbons best matching a particular $\Delta H_f°$ target value, and the second example is a solvent design study for which we chose three criteria for solvent selection: capacity, selectivity, and environmental fate, represented by the octanol−water partition coefficient K_{OW}. In both examples, a GA was deployed as the optimization procedure, which calls a quantum chemical code as a subroutine to evaluate the properties of selected candidate molecules. Furthermore, GCMs were applied for various reasons: for tuning of the GA, for filtering of molecules whose deviation from the target value or upper bounds is considered too large, and as an additional property evaluation method (case study 2). After running the algorithm two or three times, lists of 10 molecules with the highest fitness values were compiled for each example.

In case study 1, all molecules in Table 4 are known to exist in the ground state because they were obtained by DFT ab initio calculations. All of their geometry optimizations converged into local minima. Because the thermal energy contribution from 0 to 298 K is small compared to the dissociation energy $\sum D_0$, DFT results support the existence of these compounds. However, no information is obtained on whether a compound can be synthesized.

In case study 2, all molecules in the search space are known to exist. Nitromethane, a known extraction solvent of the benzene−cyclohexane system, was identified as a candidate solvent. The fitness function (eqs 3 and 4) can readily be extended to more than one property, thus accounting for more than one design objective.

Searching for candidates via GAs allows for the incorporation of multiple evaluation methods within one design application. Any property evaluation method can be used in this approach because the GA treats the property evaluation as a black box. The applicability of the proposed approach depends on the properties that are accessible for calculation from quantum chemical models. Because of computational expense, quantum models are typically used to calculate very few molecules, frequently just one, and obtain a bulk property from other relationships such as statistical mechanics. This limits their range of applicability. For example, a diffusion coefficient results from the interaction of many particles and cannot readily be obtained by quantum chemical models. Presently, a meaningful application for engineering applications is also limited by the accuracy of the existing quantum chemical methods, which in many instances still lies within the range of GCMs and does not necessarily warrant the computational expense of quantum chemistry. However, GCMs depend on experimental data for group parametrization. This disadvantage will likely be overcome by quantum chemical methods in the future.[41] For example, Eckert and Klamt[39] have used COSMO-RS to predict vapor−liquid, liquid−liquid, and solid−liquid equilibrium data, partition coefficients, and vapor pressures. In COSMO-RS, only atoms are parametrized, not functional groups. With the advent of more accurate and more efficient ab initio methods, opportunities for a fruitful combination of optimal molecular design and quantum chemical property prediction are likely to emerge. The proposed approach can be a valuable tool for automatically building databases of molecules that combine a number of desired properties.

Acknowledgment

Financial support by the National Science Foundation Career Award CTS-9701771 is gratefully acknowledged.

Appendix: MILP Formulation of the GCM-Based Design Problem for Case Study 1

As an objective, one seeks to minimize the deviation of the molecule's heat of formation from a given target value:

$$\min \left| 1 - \frac{\Delta H_f^\circ}{\Delta H_{f,\mathrm{target}}^\circ} \right|$$

The nonlinear objective is recast in a linear form[2] as follows:

$$\min s$$

subject to

$$s \geq \left(1 - \frac{\Delta H_f^\circ(L_l, \mathrm{FC_{tot}})}{\Delta H_{f,\mathrm{target}}^\circ} \right), \quad s \geq \left(\frac{\Delta H_f^\circ(L_l, \mathrm{FC_{tot}})}{\Delta H_{f,\mathrm{target}}^\circ} - 1 \right)$$

The dependencies on L_l and $\mathrm{FC_{tot}}$ are given in eqs 24 and 25. Molecules are modeled as a molecular graph with up to three possible vertexes. Each vertex can assume one of the types listed in Joback and Reid,[6] excluding the fluorine atoms. Information on how the vertexes are connected is encoded by the *adjacency matrix* $\mathbf{A} = a_{i,j,k}$,[42] whose elements are binary variables with

$$a_{i,j,k} = \begin{cases} 1 & \text{if vertex } i \text{ is connected to vertex } j, \\ & \quad \text{forming a bond of multiplicity } k \\ 0 & \text{otherwise} \end{cases}$$
$$\text{and } i, j, k = 1{-}3$$

Note that i and j range from 1 to 3 because the molecular graph is limited to three vertexes and k ranges from 1 to 3 because carbon atoms can be connected by single, double, or triple bonds. Similarly, we define a *vertex type binary* $y_{i,l}$, which determines the type of group that occupies a vertex:

$$y_{i,l} = \begin{cases} 1 & \text{if vertex } i \text{ is of type } l \\ 0 & \text{otherwise} \end{cases}$$
$$\text{and } i = 1{-}3 \text{ and } l = -\mathrm{CH_3}, ..., \equiv\mathrm{C}-$$

The indices l span over all of the groups[6] that contain carbon atoms. The number of fluorine atoms is calculated by evaluating the number of unoccupied sites on a vertex that is occupied by a carbon-containing group. For example, if $y_{1,=\mathrm{CH}-} = y_{2,=\mathrm{C}<} = a_{122} = 1$ and all other binary variables are zero, there is one site unoccupied on the first vertex and two on the second one. The third vertex has no carbon group; therefore, no fluorine atoms can be attached to it. This combination of binary variables yields $\mathrm{FHC{=}CF_2}$, 1,2,2-trifluoroethylene. This can be accomplished by writing the following constraints:

$$\mathrm{FC}_i = \sum_l 4y_{i,l} - \sum_{k=1}^{3} \Big(\sum_{j=1}^{j<i} k a_{i,j,k} + \sum_{j>i}^{3} k a_{i,j,k} \Big) - \sum_l y_{i,l} H_l$$
$$\text{with } i = 1{-}3 \quad (23)$$

$$\mathrm{FC_{tot}} = \sum_{i=1}^{3} \mathrm{FC}_i \qquad (24)$$

Here, FC_i is the number of fluorine atoms attached to vertex i and parameter H_l contains the number of hydrogen atoms pertaining to group l. $\mathrm{FC_{tot}}$ represents the total number of fluorine atoms in the molecule. The first term in eq 23 sets up four available sites on vertex i if the vertex is occupied. The second term subtracts the number of bonds that are formed with other carbon atoms, and the third term subtracts the bonds that are formed with hydrogen atoms. If any sites remain unoccupied, fluorine atoms are attached to vertex i. Equation 24 adds the fluorine atoms of the entire molecule. Because fluorine atoms are not treated as a group, eq 12 is rewritten as

$$\Delta H_{\mathrm{f}}^{\circ} = 68.29 + \sum_{l} L_l \Delta h_{\mathrm{f},l}^{\circ} + \mathrm{FC}_{\mathrm{tot}} \Delta h_{\mathrm{f},-\mathrm{F}}^{\circ} \qquad (25)$$

where L_l is the number of group l in the molecule. At least one vertex should be occupied:

$$\sum_{i=1}^{3} \sum_{l} y_{i,l} \geq 1 \qquad (26)$$

At most one group l can be present at a vertex i or the vertex remains empty:

$$\sum_{l} y_{i,l} \leq 1 \qquad i = 1-3 \qquad (27)$$

A molecular tree graph has exactly one vertex more than it has edges; i.e., the molecule has exactly one carbon group more than it has connections between these groups:

$$\sum_{i=1}^{3} \sum_{j=i+1}^{3} \sum_{k=1}^{3} a_{i,j,k} = \sum_{i=1}^{3} \sum_{l} y_{i,l} - 1 \qquad (28)$$

The molecule should be connected; i.e., when vertexes i and j are occupied, there should be a connection of bonds leading from i to j. This can be achieved by enforcing that if vertex j is occupied, there is at least one vertex i with $i < j$ connected to j via an edge (bond) $a_{i,j,k}$:

$$\sum_{i=1}^{j-1} \sum_{k=1}^{3} a_{i,j,k} = \sum_{l} y_{j,l} \qquad j = 2 \text{ and } 3 \qquad (29)$$

Vertexes can only be connected by one type of bond at a time:

$$\sum_{k=1}^{3} a_{i,j,k} \leq 1 \qquad i = 1-3 \text{ and } j = i+1, ..., 3 \qquad (30)$$

The number of bonds of type k that a group l can form *with other carbon groups* is limited and varies between the groups. To account for this, the parameter $\mathrm{CB}_{l,k}$ is introduced. For example, $\mathrm{CB}_{>\mathrm{C}<,1} = 2$ because this model is restricted to unbranched hydrofluorocarbons. For double and triple bonds, $\mathrm{CB}_{l,k}$ must balance the number of edges with bond multiplicity k:

$$\sum_{j=1}^{i-1} a_{i,j,k} + \sum_{j=i+1}^{3} a_{i,j,k} = \sum_{l} y_{i,l} \mathrm{CB}_{l,k} \qquad i = 1-3 \text{ and } k = 2 \text{ and } 3 \qquad (31)$$

For single bonds, the number of edges $a_{i,j,1}$ can be smaller than $\mathrm{CB}_{l,1}$ because the vertex i can be an ending vertex of the molecule. In this case, not all of the possible connections $\mathrm{CB}_{l,1}$ to other carbon groups would be used. Instead, these connections could be used to attach fluorine atoms or hydrogen atoms:

$$\sum_{j=1}^{i-1} a_{i,j,1} + \sum_{j=i+1}^{3} a_{i,j,1} \leq \sum_{l} y_{i,l} \mathrm{CB}_{l,1} \qquad i = 1-3 \qquad (32)$$

The following constraint uses the continuous variable L_l to count the number of each group l in the molecule:

$$L_l = \sum_{i=1}^{3} y_{i,l} \qquad l = -\mathrm{CH}_3, ..., \equiv\mathrm{C}- \qquad (33)$$

In order to reduce the number of degenerate solutions, some constraints are added to tighten the formulation. The following two constraints enforce that vertexes 1 and 3 are ending vertexes by restraining their maximal number of adjacent groups to 1:

$$\sum_{j=2}^{3} \sum_{k=1}^{3} a_{1,j,k} \leq 1 \qquad (34)$$

$$\sum_{i=1}^{2} \sum_{k=1}^{3} a_{i,3,k} \leq 1 \qquad (35)$$

Finally, two constraints are added to force vertex 2 to be occupied if vertex 3 is occupied

$$\sum_{l} y_{2,l} \geq \sum_{l} y_{3,l} \qquad (36)$$

and to force vertex 1 to be occupied if vertex 2 is occupied.

$$\sum_{l} y_{1,l} \geq \sum_{l} y_{2,l} \qquad (37)$$

This formulation solves the MILP problem to global optimality. For generating more than one globally optimal molecule, efficient integer cuts[43] are incorporated. If $y_{i,l}^{\mathrm{sol}}$ is the optimal solution, then the constraint

$$\sum_{(i,l):y_{i,l}^{\mathrm{sol}}=1} y_{i,l} + \sum_{(i,l):y_{i,l}^{\mathrm{sol}}=0} (1 - y_{i,l}) \leq (\sum_{i,l} 1) - 1 \qquad (38)$$

makes $y_{i,l}^{\mathrm{sol}}$ infeasible when the problem is solved again. Thus, looping the solution procedure n times and accumulating $n - 1$ integer cuts in the process generates a list of the n best solutions to the problem.

Literature Cited

(1) Churie, N.; Achenie, L. Novel Mathematical Programming Model for Computer Aided Molecular Design. *Ind. Eng. Chem. Res.* **1996**, *35*, 3788.

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

(3) Venkatasubramanian, V.; Chan, K.; Caruthers, J. Computer-Aided Molecular Design Using Genetic Algorithms. *Comput. Chem. Eng.* **1994**, *18* (9), 833.

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

(5) van Krevelen, D. W. *Properties of Polymers: their correlation with chemical structure; their numerical estimation and prediction from additive group contributions*, 3rd ed.; Elsevier: Amsterdam, The Netherlands, 1990.

(6) Joback, K.; Reid, R. Estimation of Pure-Component Properties From Group Contributions. *Chem. Eng. Commun.* **1987**, *57*, 223.

(7) Fredenslund, A.; Gmehling, J.; Rasmussen, P. *Vapor−Liquid Equilibria Using UNIFAC*; Elsevier: Amsterdam, The Netherlands, 1977.

(8) Lin, S.; Sandler, S. Infinite Dilution Activity Coefficients from Ab Initio Solvation Calculations. *AIChE J.* **1999**, *45* (12), 2606.

(9) Poling, B.; Prausnitz, J.; O'Connel, J. *The Properties of Gases and Liquids*, 5th ed.; McGraw-Hill: New York, 2001.

(10) Levine, I. *Quantum Chemistry*, 5th ed.; Prentice Hall: Engelwood Cliffs, NJ, 2000.

(11) Szabo, A.; Ostlund, N. *Modern Quantum Chemistry*; McGraw-Hill: New York, 1989.

(12) Parr, R.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.

(13) Labanowski, J. *Simplified Introduction to Ab Initio Basis Sets*; Ohio Supercomputer Center: Columbus, OH, 1996 [e-mail, jkl@osc.edu, jkl@ohstpy.bitnet; URL(several), http://www.ccl.net/cca/documents/basis-sets/basis.html].

(14) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Robb, G. E. S. M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; R. E. S.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Tomasi, O. F. J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.9; Gaussian, Inc.: Pittsburgh, PA, 1998.

(15) *Gaussian 98*, *User's Reference*, 2nd ed.; Gaussian, Inc.: Pittsburgh, PA, 1994−1999.

(16) Pople, J.; Gordon, M. Molecular orbital theory of the electronic structure of organic compounds. I. Substituent effects and dipole moments. *J. Am. Chem. Soc.* **1967**, *89* (17), 4253.

(17) Minoux, M. *Mathematical Programming—Theory and Algorithms*; Wiley: New York, 1986.

(18) Holland, J. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.

(19) Michalewicz, Z. *Genetic Algorithms + Data Structures = Evolution Programs*; Springer-Verlag: Berlin, 1994.

(20) Goldberg, D. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(21) Curtiss, L.; Raghavachari, K.; Redfern, P.; Pople, J. Assessment of Gaussian-2 and Density Functional Theories for the computation of Enthalpies of Formation. *J. Chem. Phys.* **1997**, *106* (3), 1063.

(22) Chase, M.; Davies, C.; Downey, J. R.; Frurip, D.; McDonald, R.; Syverud, A. JANAF Thermochemical Table. *Journal of Physical and Chemical Reference Data*, *Supplement No. 1*, 3rd ed.; 1985; p 14.

(23) Bauschlicher, C. A comparison of the accuracy of different functionals. *Chem. Phys. Lett.* **1995**, *246*, 40.

(24) Curtiss, L.; Raghavachari, K.; Trucks, G.; Pople, J. Gaussian-2 Theory for Molecular Energies of First- and Second-row Compounds. *J. Chem. Phys.* **1991**, *94* (11), 7221.

(25) Camarda, K.; Maranas, C. Optimization in Polymer Design Using Connectivity Indices. *Ind. Eng. Chem. Res.* **1999**, *38*, 1884.

(26) Raman, V.; Maranas, C. Optimization in product design with properties correlated with topological indeces. *Comput. Chem. Eng.* **1998**, *22* (6), 747.

(27) *GAMS: A User's Guide*; GAMS Development Corp.: Washington, DC, 1998.

(28) *GAMS: The Solver Manuals*; GAMS Development Corp.: Washington, DC, 1998.

(29) Hradetzky, G.; Hammerl, I.; Bittrich, H.-J.; Wehner, K.; Kisan, W. *Selective Solvents, Data on dimethylformamide−N-methylcaprolactam−N-methylpyrrolidone*; Physical Sciences Data 31; Elsevier: Amsterdam, The Netherlands, 1989.

(30) van Leeuwen, C.; Hermens, J. *Risk Assessment of Chemicals: An Introduction*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995.

(31) Lin, S.; Sandler, S. Prediction of Octanol−Water Partition Coefficients Using a Group Contribution Solvation Model. *Ind. Eng. Chem. Res.* **1999**, *38*, 4081.

(32) Abrams, D.; Prausnitz, J. Statistical Thermodynamics of Liquid Mixtures: A New Expression for the Excess Gibbs Energy of Partly or Completely Miscible Systems. *AIChE J.* **1975**, *21* (1), 116.

(33) Tomasi, J.; Persico, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94*, 2027.

(34) Schmidt, M.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. General Atomic and Molecular Electronic Structure System. *J. Comput. Chem.* **1993**, *14*, 1347.

(35) Dunning, T., Jr.; Hay, P. Gaussian Basis Sets for Molecular Calculations. In *Methods of Electronic Structure Theory*; Shaefer, H., III, Ed.; Plenum Press: New York, 1977; pp 1−27.

(36) Mennucci, B. (University of Pisa). Personal communication, 2002.

(37) Klamt, A. Conductor-Like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99* (7), 2224.

(38) Klamt, A.; Schüürmann, G. COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, *5*, 799.

(39) Eckert, F.; Klamt, A. Fast Solvent Screening via Quantum Chemistry: COSMO-RS Approach. *AIChE J.* **2002**, *48* (2), 369.

(40) Perry, R., Ed. *Perry's Chemical Engineers' Handbook*, 6th ed.; McGraw-Hill: New York, 1984.

(41) Sandler, S. I. Quantum mechanics: a new tool for engineering thermodynamics. *Fluid Phase Equilib.* **2003**, *210*, 147.

(42) Trinajsticć, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, 1992.

(43) Floudas, C. *Nonlinear and Mixed-Integer Optimization*; Oxford University Press: New York, 1995.

(44) Lias, S.; Karpas, Z.; Liebman, J. Halomethylenes: effects of halogen substitution on absolute heats of formation. *J. Am. Chem. Soc.* **1985**, *107*, 6089.

(45) Neugebauer, C.; Margrave, J. The heats of formation of CHF_3 and CH_2F_2. *J. Phys. Chem.* **1958**, *62*, 1043.

(46) Goy, C.; Lord, A.; Pritchard, H. Kinetics and thermodynamics of the reaction between iodine and fluoroform and the heat of formation of trifluoromethyl iodide. *J. Phys. Chem.* **1967**, *71*, 1086.

(47) Neugebauer, C.; Margrave, J. *Heats of formation of the fluoromethanes and fluoroethylenes*; Technical Report; 1957; pp 1−45.

(48) Chase, M., Jr. NIST-JANAF Thermochemical Tables. *Journal of Physical and Chemical Reference Data, Monograph 9*; 1998; p 1.

(49) *Selected Values of Properties of Chemical Compounds*; Data Project; Thermodynamics Research Center, Texas A&M University: College Station, TX, 1983.

(50) Kolesov, V.; Papina, T. Standard enthalpy of formation of vinyl fluoride. *J. Phys. Chem. (Engl. Transl.)* **1970**, *44*, 611.

(51) Wagman, D.; Evans, W.; Parker, V.; Schumm, R.; Halow, I.; Bailey, S. Churney, K.; Nuttall, R. NBS Tables of Chemical Thermodynamic Properties. *J. Phys. Chem. Ref. Data, Suppl. No. 2* **1982**.

(52) Kolesov, V.; Martynov, A.; Shtekher, S.; Skuratov, S. Standard enthalpies of formation of 1,1-difluoroethylene and of trifluoroethylene. *Russ. J. Phys. Chem. (Engl. Transl.)* **1962**, *36*, 1118.

(53) Neugebauer, C.; Margrave, J. The heats of formation of tetrafluoroethylene, tetrafluoromethane and 1,1-difluoroethylene. *J. Phys. Chem.* **1956**, *60*, 1318.

(54) Kolesov, V.; Zenkov, I.; Skuratov, S. The standard enthalpy of formation of tetrafluoroethylene. *Russ. J. Phys. Chem. (Engl. Transl.)* **1962**, *36*, 45.

(55) Wartenberg, H. V.; Schiefer, J. Bildungswärmen von Fluor-Chlor-Kohlenstoff-Verbindungen. *Z. Anorg. Chem.* **1955**, *278*, 326.

(56) Kirkbride, F.; Davidson, F. Heats of formation of gaseous fluoro- and fluorochloro-carbons. *Nature (London)* **1954**, *174*, 79.

(57) *TRC Thermodynamic Tables—Non-Hydrocarbons*; Thermodynamics Research Center, Texas A&M University: College Station, TX, 1989.

(58) *Thermochemical Data and Structures of Organic Compounds*; Thermodynamics Research Center, Texas A&M University: College Station, TX, 1994; Vol. 1.

(59) Kolesov, V.; Papina, T. Thermochemistry of Haloethanes. *Russ. Chem. Rev.* **1983**, *52*, 425.

(60) Wu, E.; Rodgers, A. Thermochemistry of gas-phase equilibrium $CF_3CH_3 + I_2 = CF_3CH_2I + HI$. The carbon−hydrogen bond dissociation energy in 1,1,1-trifluoroethane and the heat of formation of the 2,2,2-trifluoroethyl radical. *J. Phys. Chem.* **1974**, *78*, 2315.

(61) Kolesov, V.; Martynov, A.; Skuratov, S. Standard enthalpy of formation of 1,1,1-trifluoroethane. *Russ. J. Phys. Chem. (Engl. Transl.)* **1965**, *39*, 223.

(62) *Selected Values of Properties of Chemical Compounds*; Data Project (loose-leaf data sheets, extant); Thermodynamics Research Center, Texas A&M University: College Station, TX, 1980.

(63) Chen, S.; Rodgers, A.; Chao, J.; Wilhoit, R.; Zwolinski, B. Ideal Gas Thermodynamic Properties of Six Fluoroethanes. *J. Phys. Chem. Ref. Data* **1975**, *4*, 441.

(64) Walker, L.; Sinke, G.; Perettie, D.; Janz, G. Enthalpy of formation of trifluoroacetonitrile. *J. Am. Chem. Soc.* **1970**, *92*, 4525.

(65) Coomber, J.; Whittle, E. Bond dissociation energies from equilibrium studies. Part 2. $-$D(CF_3-CF_3) and enthalpy of formation of C_2F_6. *Trans. Faraday Soc.* **1967**, *63*, 1394.

(66) Sinke, G. The heat of reaction of nitrogen trifluoride and hexafluoroethane. *J. Phys. Chem.* **1966**, *70*, 1326.

(67) Alfassi, Z.; Golden, D.; Benson, S. The thermochemistry of the isomerization of 3-halopropenes (allyl halides) to 1-halopropenes; entropy and enthalpy of formation contribution of the Cd$-$(H)(X) group. *J. Chem. Thermodyn.* **1973**, *5*, 411.

(68) Kolesov, V.; Martinov, A.; Skuratov, S. Standard enthalpies of formation of 1,1,1-trifluoropropene. *Zh. Fiz. Khim.* **1967**, *41*, 913.

(69) Cox, J.; Pilcher, G. *Thermochemistry of Organic and Organometallic Compounds*; Academic Press: New York, 1970; pp 1$-$636.

(70) Papina, T.; Kolesov, V.; Golovanova, Y. Standard enthalpies of formation of 1,2-Dichlorohexafluoropropane and hexafluoropropane. *Russ. J. Phys. Chem. (Engl. Transl.)* **1987**, *61*, 1168.

(71) Williamson, A.; LeBreton, P.; Beauchamp, J. Photoionization mass spectrometry of 2-fluoropropane and 2,2-difluoropropane. A novel determination of the proton affinity of vinyl fluoride and 1,1-difluoroethylene. *J. Am. Chem. Soc.* **1976**, *98*, 2705.

(72) Pedley, J.; Naylor, R.; Kirby, S. *Thermochemical Data of Organic Compounds*; Chapman and Hall: London, 1986.

(73) Kolesov, V.; Talakin, O.; Skuratov, S. Standard enthalpy of formation of perfluoropropane and enthalpies of formation of normal perfluoroalkanes. *Vestn. Mosk. Univ. Khim.* **1967**, *22*, 38.

(74) Sangster, J. *Octanol$-$Water Partition Coefficients*; Wiley and Sons: New York, 1997.

(75) Suzuki, T.; Kudo, Y. Automatic log *P* estimation based on combined additive modeling methods. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 155.

(76) Sangster, J. Octanol$-$Water Partition Coefficients of Simple Organic Compounds. *J. Phys. Chem. Ref. Dat.* **1989**, *18* (3), 1111.

(77) Isnard, P.; Lambert, S. Aqueous Solubility and *n*-Octanol/Water Partition Coefficient Correlations. *Chemosphere* **1989**, *18*, 1837.

(78) Howard, P., Meylan, W., Eds. *Handbook of Physical Properties of Organic Chemicals*; Lewis Publishers: Boca Raton, FL, 1997.

Received for review February 10, 2003
Revised manuscript received December 12, 2003
Accepted February 20, 2004

IE030120Q