# Design of Solvents for Optimal Reaction Rate Constants

**Milica Folić, Claire S. Adjiman, and Efstratios N. Pistikopoulos**
Centre for Process Systems Engineering, Dept. of Chemical Engineering, Imperial College London, London SW7 2AZ, U.K.

*DOI 10.1002/aic.11146*
*Published online March 21, 2007 in Wiley InterScience (www.interscience.wiley.com).*

*A hybrid experimental/computer-aided methodology for the design of solvents for reactions is presented. It is based on the use of a few reaction rate measurements to build a reaction model, followed by the formulation and solution of an optimal computer-aided molecular design problem (CAMD) in which the reaction rate under given conditions is maximized. In order to verify the suitability of the solvent candidates identified in the CAMD step, and to assess the reliability of the model used, feedback can be introduced. When the reliability of the model is found to be insufficient, experimental rate data for the candidate solvents are obtained and added to the original data set to create an updated reaction model, which can be used to find new candidate solvents. Since very few measurements are used to build the reaction model, we perform a sensitivity analysis on the model to assess the impact of uncertainty. Using this information to generate scenarios, we then solve a stochastic optimization problem, which aims to determine the solvents that give the best average performance. The final output consists of a list of candidate solvents which can be targeted for experimentation. This methodology is illustrated, step by step, through application to a solvolysis reaction.* © 2007 American Institute of Chemical Engineers *AIChE J,* 53: 1240–1256, 2007
*Keywords: solvent design, design under uncertainty, CAMD*

## Introduction

Solvents are widely used as a reaction medium in the fine chemicals and other industries, where they serve to bring solid reactants together by dissolving them, to control temperature, and to enhance reaction rate. The effect of solvent choice on reaction rate can be dramatic. For instance, the solvolysis of 2-chloro-2-methylpropane is 335,000 times faster in water than in ethanol, while the reaction between trimethylamine and trymethylsulfonium ion is 119 times faster in nitromethane than in water.[1] In spite of the importance of solvent choice on productivity, there has been little work on systematic approaches to the selection of solvents for reactions. Thus, industrial practice is currently mostly based on experience and intuition when it comes to solvent choice during the development of new reaction routes.

Correspondence concerning this article should be addressed to C. S. Adjiman at c.adjiman@ic.ac.uk.



One set of "rules" is that developed by Hughes and Ingold to anticipate the effect of solvents in nucleophilic substitution and elimination reactions.[2,3] The assumption made by Hughes and Ingold was that they would consider only electrostatic interactions between the solute molecules (reactants and activated complex) and the solvent. A change to a more polar solvent tends to increase the reaction rate if the reactant is less dipolar than the activated complex, that is, the activated complex is more stabilized in the polar solvent than the reactants, and its free energy is lowered more than that of the reactants.

This reliance on heuristics is in striking contrast with the selection of solvents for separation, where several computer-aided molecular design (CAMD) approaches have been proposed in the last two decades. Several of these methods are described in Achenie et al.[4] These have been successfully applied to a variety of solvent-based separation problems, allowing a much larger number of solvent molecules to be considered during separation system design than is possible by experimentation alone. The development of CAMD meth-

ods for reactions along similar routes is highly desirable to speed-up and reduce the cost of process development.

CAMD is a synthesis activity, with the aim to identify a list of candidate molecules that perform a task (or a set of tasks) most effectively.[5] Molecular design methods are based on the fact that, from a small set of structural building groups, a large number of molecules can be generated and evaluated with respect to a certain performance index. One of the advantages of such an approach is that the search is not limited to or biased towards a particular set of compounds, and that the number of candidate molecular structures increases significantly with the number of building groups, and their maximum allowed composition in the designed molecule. The results of such computational tools serve as a guide to experimentations, focussing the search on promising solvents. Conversely, the experiments allow a verification of the models used.

In order to develop a CAMD approach for solvent for reactions, a model of solvent effects on reaction rate is needed. There have been many attempts to model such effects, from the empirical equation of Abraham and coworkers,[6–8] to the more recent *ab initio* approaches (see, for example,[9]). One tool for predicting the effect of solvent on reaction rates in solution is the "reaction fingerprint."[10] The "reaction fingerprint" is presented as the difference between the charge distribution of the reacting species and of the activated complex. Detailed quantum mechanical calculations are required to estimate the separation, creation or dispersion of charge in the transition from reactants to activated complex. This approach can only be applied to a predetermined set of solvent molecules, and, therefore, is most suitable for the verification of potential optimal solvents.

Recently, a method for solvent selection for the promotion of organic reactions that combines knowledge from industrial practice and physical insights has been proposed.[11] The solvent selection strategy involves allocating a score to each solvent in a database of the 75 most commonly used solvents. Computer-aided molecular design (CAMD) can also be used to generate a list of solvent candidates, ranked according to their score. Though this method has proven very effective for some application studies, it requires a significant amount of information on both solvents and reactions (for example, solvent association/dissociation, solvent reactivity, solubility of reactants/products, and so on), in order to build a table with solvent scores for each reaction.

Based on these considerations, our goal is to develop a systematic approach to solvent design for reactions. The methodology presented here is an extension of our previous work.[12,13] The basic premise is that since there is a lack of generic predictive models of solvent effects on reactions, an iterative strategy based on targeted experiments, model development, candidate generation by CAMD, and experimental verification must be adopted. The methodology we present is developed with the ultimate aim of plant-wide solvent selection in mind. In this context, it is important to focus on overall performance rather than the performance of single process units. This motivates the use of an optimization-based approach to CAMD, where trade-offs between different aspects of the process can be accounted for explicitly.

## Solvent Design Methodology

### *Overview*

The overall methodology proposed in this work is illustrated in Figure 1. For a given reaction, a small set of initial solvents is chosen. These solvents are selected to be diverse in terms of the types of interactions they can have with the species involved in the reaction. As a measure of this, we use the $E_N^T$ solvent polarity scale,[14] and we choose solvents that have $E_N^T$ values distributed over the entire physical range. The $E_T^N$ solvent polarity scale is based on the following relation

$$E_T^N = \frac{E_T(30)(\text{solvent}) - E_T(30)(\text{TMS})}{E_T(30)(\text{water}) - E_T(30)(\text{TMS})} \tag{1}$$

It ranges from 0.000 for TMS (tetramethylsilane), the least polar solvent, to 1.000 for water, the most polar solvent.

In addition, it is preferable to choose solvents with different functional groups. Wherever possible, literature data should be used at this stage to minimize experimental costs. In the absence of reliable data, experimental reaction rate constants for the solvents chosen are measured. Given the diversity of the solvents chosen, it is expected that the reaction rate constant will be large in some solvents and small in others. This information is then used to build a reaction model that predicts the reaction rate constant in other solvents based solely on their molecular structure.

Next, a computer-aided solvent design problem is formulated based on this model, and solved. The objective is to find candidate solvents which give high values of the reaction rate constant. The computer-aided design step thus serves two purposes: it identifies promising solvents and it guides experiments. Since we build the reaction model based on kinetic data in a few solvents only, and then use it for further extrapolation, there is uncertainty associated with the model coefficients. Therefore, we propose an alternative to deterministic optimization to take this into account. We use global sensitivity analysis to evaluate the effect of uncertain coefficient ranges on the model and to identify representative combinations of the coefficients. We then formulate a stochastic design problem using these scenarios. The solution of

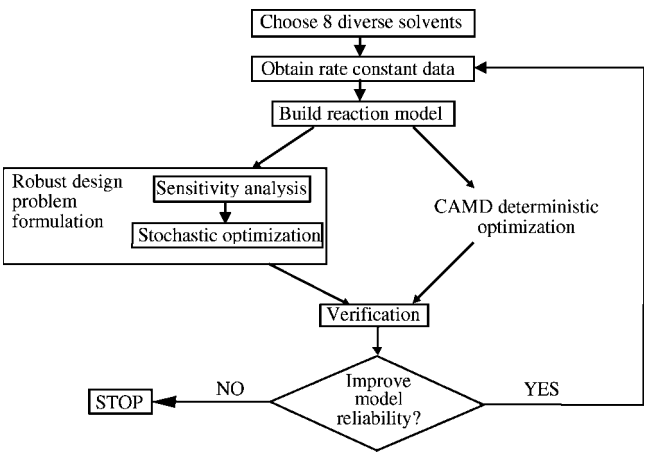


**Figure 1. Overview of the solvent design methodology.**

this problem gives the best solution in the presence of model uncertainty.

In the next step (verification), the predicted rate constants for the best candidate solvents identified are compared to experimental rate constants, to determine whether the reaction model needs improvement. If so, the measured rate constants for the candidate solvents are added to the set of initial solvents to build an updated reaction model. This procedure is repeated until the model is judged to be sufficiently reliable. Reaction model building, the molecular design step and model reliability are discussed in more detail in the subsequent sections.

### *Reaction modeling*

The key issue in the design of optimal solvents for reaction is to identify a relationship that links solvent properties to reaction rate in a quantitative way. Here, we use the multiparameter solvatochromic equation,[8] which correlates empirical solvatochromic parameters and Hildebrand solubility parameter with the logarithm of the reaction rate constant. The model, as illustrated in Figure 2, allows us to derive a ranked list of candidate solvents for a specific reaction from solvent building blocks (structural groups). We first choose a few diverse solvents and obtain experimental reaction rate constant values in those solvents. We also compute the required physical properties for the selected solvents via structure-property relationships. Once all the data are available, we regress the parameter values in the solvatochromic equation to obtain a model of solvent effects on the chosen reaction. We can then use this equation to predict the reaction rate constant in other solvents.

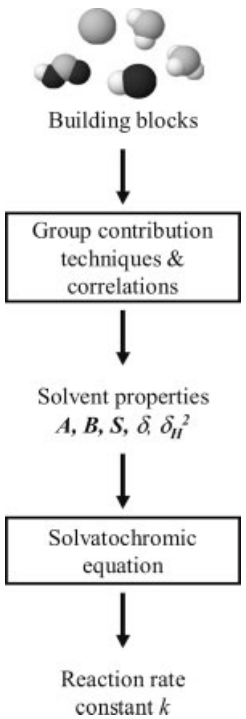


**Figure 2. Reaction model.**

### *Solvatochromic equation*

*Description of the Equation.* Abraham and co-workers[7,8,15,16] developed the "solvatochromic equation" for the prediction of solvent effects. It is a simple equation that gives a good balance between reliability and complexity. In its general form, it is given by

$$XYZ = XYZ_0 + s(\pi^* + d\delta) + a\alpha + b\beta + h\delta_H^2/100 \qquad (2)$$

where:

- the *XYZ* term can be the logarithm of a rate constant, equilibrium constant, free energy of solution, and so on;
- $\alpha$, $\beta$, $\pi^*$ are the Kamlet-Taft solvatochromic parameters[17,18,19]: $\alpha$ is the solvent hydrogen bond donor (HBD) acidity (the solvent's ability to donate a proton in a solvent-to-solute hydrogen bond), $\beta$ is the solvent hydrogen bond acceptor (HBA) basicity (the solvent's ability to donate an electron pair in a solute-to-solvent hydrogen bond) and $\pi^*$ is a measure of solvent dipolarity/polarisability (its ability to stabilize a charge or a dipole by virtue of its dielectric effect);
- $\delta$ is a "polarizability correction term" that reflects the fact that, generally, differences in solvent polarizability are significantly greater between solvent classes than within a class. It is equal to 1 for aromatics, 0.5 for polychlorinated aliphatics and 0 for other compounds;
- $\delta_H^2$ is the cohesive energy density which provides a measure of solvent-solvent interactions, and
- *s, d, a, b, h* are coefficients that measure the relative susceptibility of *XYZ* to the solvent properties. They are obtained by linear regression and are independent of the particular solvent being considered.

This equation has been successfully used to correlate octanol/water-partition coefficient,[20] as well as Gibbs free energies of transfer from gas to solvents,[21] with solvent parameters. The equation has also been used to quantify solvent effects on the solvolysis of t-butyl halides,[7,8,15] and on Diels-Alder reactions.[22] Both reaction classes have long been a key reference for theories of solvent effects on organic reaction rates. Abraham et al.[8] regressed solvatochromic reaction parameters for the solvolysis of t-butyl chloride in 21 aliphatic solvents and reported highly satisfactory statistics, including an $R^2$ (the square of the Pearson product moment correlation) value of 0.995, and a standard deviation of 0.242.

Several methods have been proposed to measure the solvatochromic parameters ($\alpha$, $\beta$ and $\pi^*$) of a given compound. In their solvatochromic comparison method, Kamlet and Taft[17–19] obtained "solvent" parameters by comparing the electronic transition of different solutes (for example, 4-nitroaniline and N,N-diethyl-4-nitroaniline) in the compound of interest. In Abraham et al.,[23] a method for measuring solvatochromic parameters in which the compound of interest was treated as a solute was presented. Following the notation of Zissimos et al.[24] for these "solute" parameters, the hydrogen bond donor (HBD) acidity of the compound is denoted by *A*, its hydrogen bond acceptor (HBA) basicity is denoted by *B*, and its dipolarity/polarisability is referred to as *S*. Values of *A, B,* and *S* for many compounds can readily be found in the literature. Although these values are obtained by treating the compound as a solute, they provide a quantitative description

of the compound's interactions with other molecules. Furthermore, the solvatochromic parameters are in fact scales which show the relative strengths of the interactions. It can therefore be argued that, provided that the parameter scales are measured in a consistent way, the solvatochromic parameters derived either by the "solvent" method of measurement ($\alpha$, $\beta$ and $\pi^*$), or by the "solute" method of measurement ($A$, $B$, $S$) can be used to study how a given property varies from compound to compound. Combined with the cohesive energy density which measures the strength of the compound's interactions with itself, the parameters $A$, $B$, and $S$, can be used to correlate the dependence of a given property for a specific system (solute, reaction) in different compounds through the following solvatochromic equation

$$XYZ = XYZ_0 + s(S + d\delta) + aA + bB + h\delta_H^2/100 \qquad (3)$$

The description of a solvent using the $A$, $B$, and $S$ solvatochromic parameters has been tested with success to model solvation effects in the SM$x$ solvation model.[25] Given the wide availability of $A$, $B$, and $S$ values, and previous experience in using these parameters to describe solvent effects, the following solvatochromic equation is used to correlate the reaction rate constant for a given reaction in different solvents

$$\log k = \log k_0 + sS + d\delta + aA + bB + h\delta_H^2/100 \qquad (4)$$

The parameters $S$, $A$, $B$, $\delta$ and $\delta_H^2$ are properties of the solvent molecule only, and do not depend on the reaction being studied. On the other hand, the parameters $s$, $d$, $a$, $b$, $h$ and $\log k_0$, depend only on the reaction being studied and do not vary from solvent to solvent. The fact that the reaction parameters and solvent properties are independent of each other means that the solvatochromic equation can be used to predict rate constants for combinations of reaction and solvent for which no data are available.

In the following section, we describe the derivation of the reaction parameters in order to be able to predict $\log k$ values.

*Performance of the Solvatochromic Equation.* In order to be able to perform a linear regression to fit the reaction parameters in Eq. 4, experimental reaction rate constant data should be gathered. As there are six parameters, the absolute minimum number of solvents that must be used is six. The choice of the number of solvents is based on a trade-off between the cost of experiments to acquire the data and the statistical quality of the regression results. Statistically, results based on a small set of solvents are not very reliable and the accuracy of predictions upon extrapolation is low. On the other hand, including many solvents incurs additional costs, and it does not necessarily guarantee a significant improvement of the statistics.

To investigate the appropriate number of solvents, we performed three comparative studies, using different numbers of solvents from the same data set. All the solvent properties needed for this study were gathered from published experimental data.[6–8,26–29] Reaction rate data were collected for a solvolysis reaction. Solvolysis is a reaction that is induced by the solvent, and we studied solvolysis of t-butyl chloride. The rate of this reaction is determined by ionization of the covalent bond C—Cl, which involves the consecutive formation of three ion pairs: contact, loose, and solvent-separated

$$(CH_3)_3CCl \longleftrightarrow (CH_3)_3C^+ \ldots Cl^- \longleftrightarrow (CH_3)_3C^+ |Solv| Cl^- \rightarrow \text{Products} \qquad (5)$$

There have been many kinetic studies of this reaction,[6–8,26–29] and we gathered kinetic data for 24 solvents.

We performed three regressions: one with eight solvents (Regression 1), one with fifteen (Regression 2) and one with all 24 solvents (Regression 3). As already mentioned, the chosen solvents should represent various classes of chemicals (for example, an aromatic, a nitrate, an amide, an alcohol, a carboxylic acid, a halosubstituted compound), and cover a wide range of polarities. The eight solvents chosen for the first regression are listed (in bold) in Table 1, together with their polarities. For Regression 2, we complemented solvent list for Regression 1 with seven more solvents and the resulting list of fifteen solvents is shown in Table 1. For Regression 3, all the available solvent data points were used. The solvatochromic equations obtained are listed in Table 2.

The statistics obtained for all three regressions are shown in Table 3. These include values for $R^2$, the adjusted $R^2$, the standard error and the average absolute percentage error (AAPE) calculated after using Eq. 63, 64, and 65 to predict $\log k$ values for all 24 solvents in the data set. $R^2$ is the square of the Pearson product moment correlation coefficient, which measures the degree of linearity of fit and varies between 0 and 1. The AAPE is defined as:

$$\text{AAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|X_{pred,i} - X_{exp,i}|}{X_{exp,i}} \cdot 100\%$$

where $N$ is the number of compounds in the dataset, $X_{exp,i}$ is the experimental value of the property ($\log k$) for compound $i$ and $X_{pred,i}$ is the predicted value of the property for compound $i$.

**Table 1. List of Solvents Used in Regressions 1 and 2**

| Solvent | $E_T^N$ Value |
|---|---|
| **Phenol** | 0.95 |
| N-Methylformamide | 0.72 |
| Hexanol | 0.66 |
| **Acetic Acid** | 0.65 |
| **Ethanol** | 0.65 |
| **2-Propanol** | 0.55 |
| n-Butanol | 0.54 |
| 2-Methyl-1-Propanol | 0.49 |
| **Dimethylacetamide** | 0.40 |
| 1,2-Dichloroethane | 0.33 |
| **Acetophenone** | 0.31 |
| Cyclohexanone | 0.28 |
| **Tetrahydrofuran** | 0.21 |
| Ether | 0.12 |
| Benzene | 0.11 |

**Table 2. Three Solvatochromic Equations Obtained by Fitting to Three Different Data Sets**

| | |
|---|---|
| Regression 1 | $\log k = -15.00 + 0.51S + 0.64\delta + 7.98A + 3.60B + 1.92\delta_H^2/100$ (63) |
| Regression 2 | $\log k = -14.06 + 0.56S + 0.70\delta + 8.10A + 3.58B + 1.04\delta_H^2/100$ (64) |
| Regression 3 | $\log k = -14.31 + 0.57S + 1.17\delta + 8.29A + 3.61B + 1.15\delta_H^2/100$ (65) |

As can be seen in Table 3, the regression statistics worsen with the increase in the number of points used for regression. This can be attributed to the inherent limitations of the simple model used and/or experimental error. As for the prediction statistics, the AAPE values are slightly worse for regression 1 than regression 3, but the cost associated with obtaining the larger data set used in regression 3 is three times that for regression 1, and the expense does not result in significant improvements.

An analysis of the qualitative information obtained from the three models, such as solvent ranking, is instructive. Calculated values for log $k$ for each of the Eqs. 63, 64 and 65 are listed in Table 4. For each regression, a solvent ranking is also shown, with 1 denoting the solvent with the largest rate constant, and 24 the solvent with the smallest rate constant. The solvents are presented in a rank-ordered list based on the experimentally measured log $k$ values.

As shown in Table 4, the three regressions predict the same solvent as the best one (phenol), in agreement with experimental data. If we consider the rankings for the first ten experimentally ranked solvents, all three regressions predict 8 out of those 10 within the first 10, but not in the same order. Noting that there is a large gap between the log $k$ values for solvents 14 and 15 (2-methyl-2-propanol and dimethylacetamide), it is instructive to examine the first 14 solvents. There, the predictions from all three regressions predict the first 14 experimentally ranked solvents within the first 14.

Given the statistics and rankings presented for the three regressions performed, the solvatochromic equation appears to give good predictions of the effect of solvent on the logarithm of the reaction rate constant for the solvolysis reaction, even when a small set of diverse solvents is used for regression. A regression based on measurements in eight different solvents will be used to illustrate the approach in the remainder of this article.

### *Estimation of Solvent Properties*

*Methods Used.* For the prediction of the solvent properties used in Eq. 4 based on the solvent molecular structure, we use group contribution (GC) methods. The polarizability correction term $\delta$, can be calculated exactly based on molecular structure

$$\delta = \begin{cases} 1 & \text{for aromatics} \\ 0.5 & \text{for polychlorinated aliphatics} \\ 0 & \text{for all other compounds} \end{cases} \quad (6)$$

GC methods are based on the principles of transferability and additivity, and are widely used for property prediction in CAMD. Atom groups, such as $CH_2$ and $OH$, are used as building blocks. In order to make integration with existing CAMD approaches easy, we use the UNIFAC groups as building blocks for solvents.

All the properties are predicted directly from structure with the exception of $\delta_H$ (in J cm$^{-3}$). As discussed in Sheldon et al.[30], $\delta_H$ can be calculated based on the prediction of the molar volume of the solvent (in cm$^3$ mol$^{-1}$), $V_m$, and its enthalpy of vaporization (in kJ mol$^{-1}$), $\Delta H_V$, and it is defined as

$$\delta_H = \left[\frac{\Delta H_V - RT}{V_m}\right]^{0.5} \quad (7)$$

where $R$ is the gas constant (in kJ mol$^{-1}$ K$^{-1}$) and $T$ is the temperature (in K).

The group contribution method used for the prediction of $V_m$ is that of Constantinou et al.[31] They proposed a GC method using the standard UNIFAC groups for a first-level calculation and second-order groups to improve the accuracy of the estimation by considering additional information on the structure of the molecule, taking into account, for example, some proximity effects. We consider only first-order groups in this initial development of the methodology, partic-

**Table 3. Statistics for the Linear Regressions Performed with Three Different Sets of Solvent Data Points**

| | Regression 1 (8 Solvents) | Regression 2 (15 Solvents) | Regression 3 (24 Solvents) |
|---|---|---|---|
| $R^2$ | 0.995 | 0.948 | 0.937 |
| Adjusted $R^2$ | 0.981 | 0.919 | 0.919 |
| Standard Error | 0.343 | 0.644 | 0.598 |
| AAPE | 5.49% | 5.58% | 5.22% |

**Table 4. List of Solvents, Experimental log $k$ Values, Predicted log $k$ Values for the Three Regressions and Associated Rankings**

| | | log $k_{pred}$/Rank | | |
|---|---|---|---|---|
| Solvent | log $k_{exp}$ | Reg 1 | Reg 2 | Reg 3 |
| Phenol | −4.66 | −4.86/**1** | −5.02/**1** | −4.49/**1** |
| Aniline | −6.15 | −7.58/**7** | −7.73/**7** | −7.28/**4** |
| N-Methylformamide | −6.54 | −4.90/**2** | −5.81/**2** | −5.72/**2** |
| Acetic Acid | −6.70 | −6.57/**3** | −6.29/**3** | −6.31/**3** |
| Ethanol | −7.06 | −6.98/**4** | −7.41/ **4** | −7.40/**5** |
| 1-Propanol | −7.33 | −7.36/**5** | −7.62/**5** | −7.63/**6** |
| Hexanol | −7.45 | −7.96/**12** | −7.95/**12** | −7.98/**12** |
| Octanol | −7.52 | −8.11/**14** | −8.03/**13** | −8.08/**13** |
| n-Butanol | −7.52 | −7.61/**8** | −7.76/**8** | −7.78/**8** |
| 2-Propanol | −7.74 | −7.67/**9** | −7.83/**9** | −7.85/**9** |
| Cyclohexanol | −8.07 | −7.53/**6** | −7.72/**6** | −7.74/**7** |
| Butan-2-ol | −8.10 | −7.76/**11** | −7.84/**11** | −7.87/**11** |
| 2-Methyl-1-Propanol | −8.30 | −7.72/**10** | −7.83/**10** | −7.85/**10** |
| 2-Methyl-2-Propanol | −8.39 | −8.08/**13** | −8.08/**14** | −8.13/**14** |
| Dimethylacetamide | −9.31 | −9.23/**15** | −9.29/**15** | −9.37/**15** |
| Cyclohexanone | −9.61 | −10.68/**18** | −10.57/**18** | −10.68/**18** |
| Acetophenone | −10.13 | −10.10/**16** | −9.98/**16** | −9.62/**16** |
| 1,2-dichloroethane | −10.64 | −11.28/**19** | −11.13/**19** | −11.01/**19** |
| Dioxane | −10.80 | −10.37/**17** | −10.30/**17** | −10.41/**17** |
| Tetrahydrofuran | −11.00 | −11.36/**20** | −11.16/**20** | −11.29/**20** |
| Chlorobenzene | −11.34 | −12.07/**23** | −11.82/**24** | −11.49/**23** |
| Ethyl Acetate | −11.50 | −11.53/**21** | −11.27/**21** | −11.41/**22** |
| Benzene | −12.16 | −11.98/**22** | −11.69/**22** | −11.37/**21** |
| Ether | −12.70 | −12.23/**24** | −11.79/**23** | −11.96/**24** |

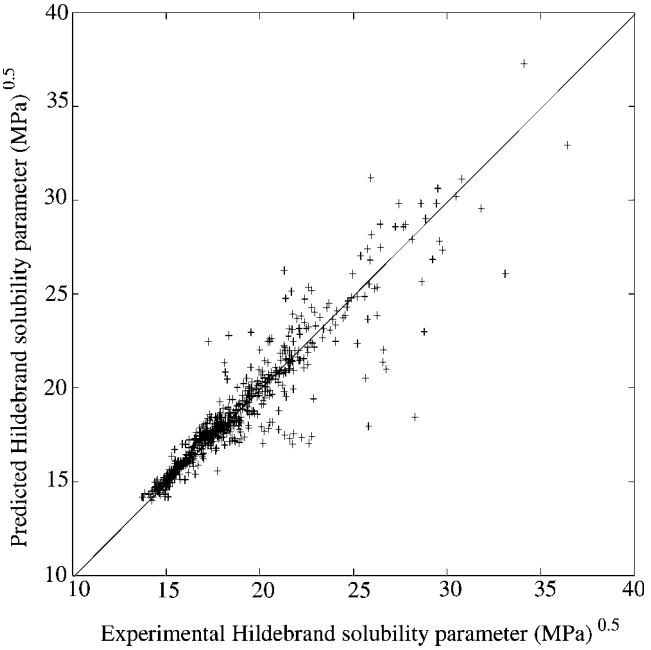


**Figure 3. Predicted vs. experimental values for $\delta_H$.**

ularly since we are considering solvent molecules which are usually relatively small. Reported statistics for predictions at first-order level are excellent and include an AAPE value of 1.16%, and a standard deviation of 0.00236 $m^3$ $kmol^{-1}$ for 312 data points. $V_m$ is predicted from

$$V_m[298\text{K}] = \sum_{i \in G} n_i V_{m,i} + 0.012 \tag{8}$$

where $G$ is the set of UNIFAC groups, $n_i$ is the number of groups of type $i$ in the molecule, and $V_{m,i}$ is the contribution to the liquid molar volume property value from group $i$.

The group contribution method of Constantinou and Gani[32] is used for the prediction of the standard enthalpy of vaporization

$$\Delta H_V[298\text{K}] = \sum_{i \in G} n_i H_{V,i} + 6.829 \tag{9}$$

where $n_i$ is the number of groups of type $i$ in the molecule, and $H_{V,i}$ is the contribution to the enthalpy of vaporization from group $i$. The reported statistics for predictions at first-order level include an AAPE value of 3.22%, and a standard deviation of 2.2 kJ $mol^{-1}$ for 225 data points.

The prediction of $\delta_H$, from Eq. 7, for the 872 data points, which are listed in the DIPPR database[33] was studied in Sheldon et al.[30] It is of good quality considering that the experimental $\delta_H$ values were not used to fit any coefficients. Predicted versus experimental values for $\delta_H$ are shown in Figure 3. The average absolute error (AAE) and the range of the property values are given in Table 5. The AAE is given by

$$\text{AAE} = \frac{1}{N} \sum_{i=1}^{N} |X_{pred,i} - X_{exp,i}| \tag{10}$$

The prediction of $A$ and $B$ is based on Sheldon et al.[30]

$$P = \begin{cases} \sum_{i \in G} n_i P_i + P_0 & \text{if } \sum_{i \in G} n_i P_i + P_0 > m \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $P$ is the value of the property (A or B), $P_i$ is the value of the coefficient (contribution) for group $i$ used in the calculation of the value of the property, $n_i$ is the number of groups of type $i$ in the molecule, and $P_0$ and $m$ are constants defined for the property of interest. In this work, we use an extended list of atom groups and obtain new coefficients for Eq. 11,

**Table 5. Average Absolute Error (AAE) for the Property Estimation Methods Used to Predict Solvent Properties $\delta_H$, $A$, $B$ and $S$**

| Property (Range) | $\delta_H$ (12.40–37.60) | A (0–0.90) | B (0–1.45) | S (0.08–1.72) |
|---|---|---|---|---|
| AAE | 1.13 $MP_a^{0.5}$ | 0.017 | 0.043 | 0.065 |

**Table 6. Atom Groups Used and their Contributions to Hydrogen Bond Acidity ($A$), Basicity ($B$), and Solvent Dipolarity/Polarizability ($S$)**

| Group $i$ | Number | $A_i$ | $B_i$ | $S_i$ |
|---|---|---|---|---|
| $CH_3$ | 1 | 0 | 0 | −0.12190 |
| $CH_2$ | 2 | −0.00120 | 0 | −0.00450 |
| $CH$ | 3 | 0 | 0 | 0.06046 |
| $C$ | 4 | −0.02910 | 0.0616 | 0.22053 |
| $CH_2{=}CH$ | 5 | 0 | 0 | −0.06890 |
| $CH{=}CH$ | 6 | 0 | 0 | 0 |
| $CH_2{=}C$ | 7 | 0 | 0.0272 | 0.06977 |
| $CH{=}C$ | 8 | 0 | 0.0203 | 0.11525 |
| $C{=}C$ | 9 | 0 | 0 | 0.25022 |
| $aCH$ | 10 | 0 | 0.0054 | 0.03740 |
| $aC$ | 11 | 0 | 0 | 0.12502 |
| $aCCH_3$ | 12 | −0.0050 | 0.0099 | 0.04876 |
| $aCCH_2$ | 13 | 0 | 0.0525 | 0.12852 |
| $aCCH$ | 14 | 0 | 0.0598 | 0.22046 |
| $OH$ | 15 | 0.31710 | 0.3621 | 0.20068 |
| $aCOH$ | 18 | 0.56550 | 0.1206 | 0.35341 |
| $CH_3CO$ | 19 | 0 | 0.3671 | 0.48527 |
| $CH_2CO$ | 20 | 0 | 0.3876 | 0.56702 |
| $CHO$ | 21 | 0 | 0.3049 | 0.48993 |
| $CH_3COO$ | 22 | −0.01680 | 0.3515 | 0.37393 |
| $CH_2COO$ | 23 | 0 | 0.3410 | 0.56216 |
| $CH_3O$ | 25 | −0.05930 | 0.2008 | 0.08413 |
| $CH_2O$ | 26 | 0 | 0.2923 | 0.19389 |
| $CH{-}O$ | 27 | 0 | 0.1955 | 0.24522 |
| $CH_2NH_2$ | 30 | 0.14357 | 0.4982 | 0.17117 |
| $CH_3NH$ | 32 | 0.27936 | 0.2730 | 0.37226 |
| $CH_2NH$ | 33 | 0.14119 | 0.5663 | 0.24524 |
| $CH_3N$ | 35 | 0 | 0.3965 | 0.35731 |
| $CH_2N$ | 36 | −0.03970 | 0.4687 | 0.30162 |
| $aCNH_2$ | 37 | 0.25598 | 0.2482 | 0.43739 |
| $CH_2CN$ | 42 | 0 | 0.2342 | 0.75871 |
| $COOH$ | 43 | 0.59926 | 0.3096 | 0.37009 |
| $CH_2Cl$ | 45 | 0 | 0.0123 | 0.18221 |
| $CHCl$ | 46 | 0 | 0.0329 | 0.17584 |
| $CHCl_2$ | 49 | 0.16627 | 0 | 0.26161 |
| $aCCl$ | 54 | 0 | −0.0637 | 0.11821 |
| $CH_2NO_2$ | 56 | 0 | 0.1783 | 0.75522 |
| $CHNO_2$ | 57 | 0 | 0.1963 | 0.6667 |
| $CH_2SH$ | 61 | 0 | 0.1123 | 0.16308 |
| $I$ | 64 | 0 | 0.0186 | 0.19669 |
| $Br$ | 65 | 0.01796 | 0.0046 | 0.17642 |
| $aCF$ | 71 | 0 | 0 | 0.04043 |
| $CH_2S$ | | 0 | 0.1963 | 0.31164 |

The 'Number' column refers to the group number in the UNIFAC classification (Poling et al., 2000). Group $CH_2S$ is not a UNIFAC group.

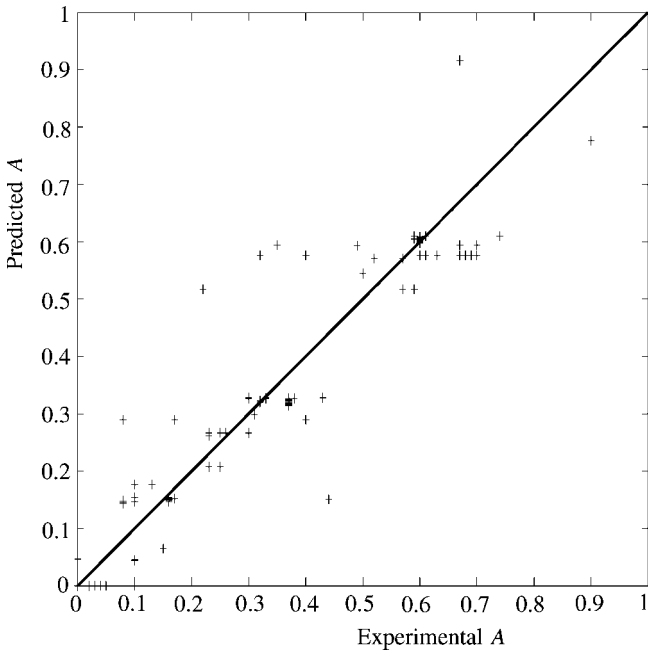


**Figure 4. Predicted vs. experimental values for *A*.**

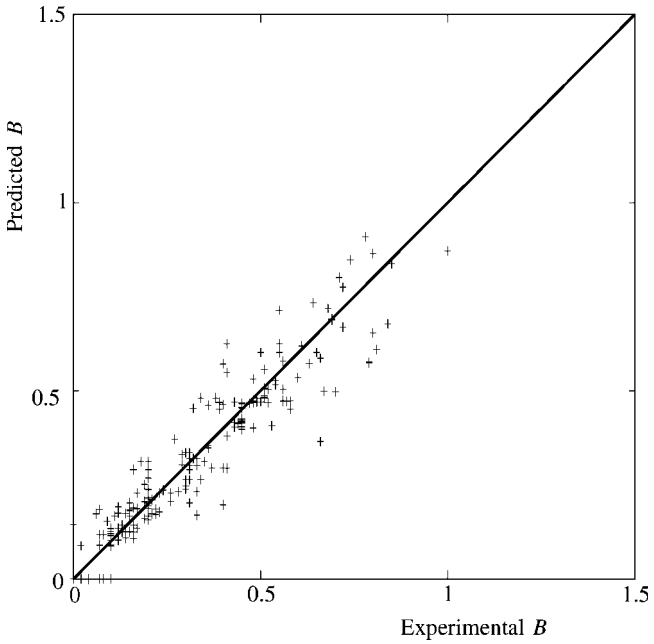


**Figure 5. Predicted vs. experimental values for *B*.**

by performing a regression on the compounds from an experimental database of 350 solvents, gathered from[34–36]. Group-contribution coefficients are, thus, available for 43 groups (see Table 6), allowing a wider variety of solvent molecules to be represented. Furthermore, the increased number of solvents in the database gives more confidence in the prediction techniques. The regressions for $A$ and $B$ were performed using a Numerical Algorithms Group[37] (NAG) statistical add-in in MS. The values obtained for $P_0$ and $m$ are 0.01064, and 0.029 for $A$, and 0.12371 and 0.124 for $B$, respectively. The range of property values and the average absolute errors obtained for each of the properties are reported in Table 5. Graphs of predicted vs. experimental values for $A$ and $B$ are shown in Figures 4 and 5, respectively, and the contributions $A_i$ and $B_i$ for each of the 43 groups are listed in Table 6.

The experimental data set for $A$ contains 235 points at exactly zero-zero. The proposed method predicts 234 of these exactly, with an additional 7 solvents predicted as having a zero value for $A$, where the experimental value is greater than zero. Figure 4 appears to show few points because, in addition to the points at exactly zero-zero, many of the other points overlap. This is due to the relatively few groups that have a non-zero contribution to the prediction of $A$, as can be seen in Table 6.

The experimental data set for $B$ contains 20 points at zero-zero, 19 of which are predicted exactly. An additional 17 solvents are predicted as having a zero value for $B$ whereas the experimental value is greater than zero.

Finally, we have developed a group contribution approach to predict values of solvent dipolarity/polarizability $S$, using the same solvent database as for $A$ and $B$. The prediction is given by

$$S = \sum_{i \in G} n_i S_i + 0.326 \qquad (12)$$

where $n_i$ is the number of groups of type $i$, and $S_i$ is the contribution of group $i$ to the dipolarity/polarizability property value. The range of property values and the AAE obtained are reported in Table 5. A graph of predicted versus experimental values is given in Figure 6 and the contributions $S_i$ for each of the 43 groups are listed in Table 6.

*Impact on Rate Constant Predictions.* In the context of the solvolysis reaction, we investigate the effect of using predicted solvent property values on the predictions of the reaction rate constant. Using predicted instead of measured sol-

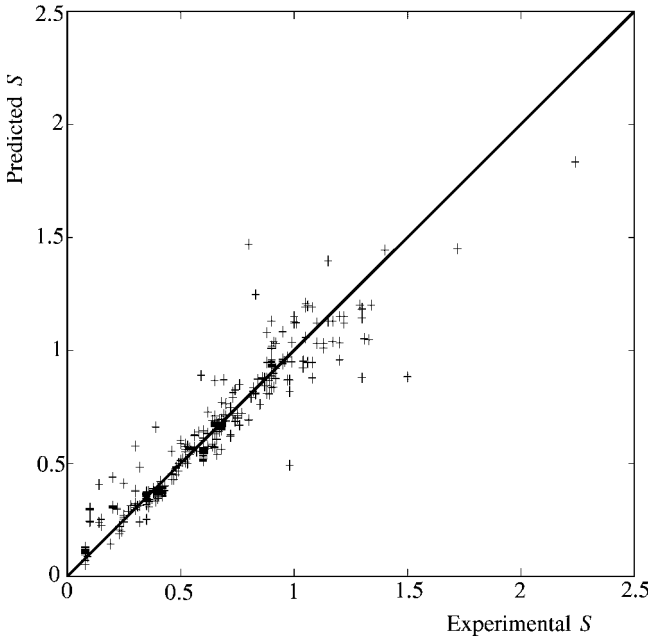


**Figure 6. Predicted vs. experimental values for *S*.**

**Table 7. List of Solvents Added to the Set in Table 4 and the Corresponding Measured log $k$ Values**

| Solvent | log $k_{exp}$ |
|---|---|
| Glycerol | −4.10 |
| 1,2-Ethanediol | −4.60 |
| Propane-1,3-diol | −5.29 |
| Propane-1,2-ethanediol | −5.51 |
| Diethylene glycol | −5.69 |
| Butane-1,4-diol | −5.99 |
| Triethylene glycol | −6.07 |
| Butane-1,2-diol | −6.14 |
| Butane-2,3-diol | −6.21 |
| Pentane-1,5-diol | −6.32 |
| N-Methylacetamide | −7.33 |
| Acetic anhydride | −8.32 |
| Dimethylformamide | −8.46 |
| gamma-Butyrolactone | −8.58 |
| Acetone | −9.61 |
| Heptane | −15.80 |
| Pentane | −16.00 |

vent parameters allows us to use experimental kinetic data for all 41 solvents reported in the literature, instead of 24. The additional solvents and the corresponding experimental log $k$ values are listed in Table 7. This large increase in the number of solvents in the data set is a great advantage, since it allows the exploration of a wider space of chemical compounds (for example, no diols were present in the initial set).

In order to study the effect of using *predicted* solvent properties, we have repeated Regression 1 with the eight solvents listed in bold in Table 1. This yielded values of 0.99 for $R^2$ and 0.48 for standard error. Upon extrapolation to the original data set of 24 solvents, an AAPE of 17.71 % was obtained, and for extrapolation to the whole data set of 41 solvents, an AAPE of 25.72 %. Comparing these statistics with those obtained for Regression 1 (Table 3), there is clearly a deterioration in AAPE. However, the regression statistics are only slightly worse.

Since the increased number of solvents in the data set gives more flexibility in the choice of solvents to be used for regression, we have performed a regression (Regression 4) on a different set of 8 solvents. These solvents, and their $E_T^N$ values are given in Table 8. To quantify the degree of diversity, we compared the variances ($\sigma^2$) of the solvent $E_T^N$ values for the two sets. The new proposed set has a value 30% higher than the initial set (0.087 compared to 0.066). Therefore, this subset is likely to be more representative. The following equation was obtained through regression.

$$\log k = -15.77 + 2.99S + 0.96\delta + 5.35A + 2.08B + 1.44\delta_H^2/100 \qquad (13)$$

The statistics obtained are reported in Table 9. It can be seen that there is an improvement, compared to the set of 8 solvents

**Table 8. List of Solvents Used in Regression 4**

| Solvent | $E_T^N$ Value |
|---|---|
| 1,2-Ethanediol | 0.79 |
| Propane-1,3-diol | 0.75 |
| Diethylene glycol | 0.71 |
| Acetic acid | 0.65 |
| Dimethylacetamide | 0.40 |
| Chlorobenzene | 0.19 |
| Benzene | 0.11 |
| Pentane | 0.01 |

**Table 9. Statistics Obtained for Regression 4**

| | Regression 4 |
|---|---|
| $R^2$ | 0.93 |
| Adjusted $R^2$ | 0.76 |
| Standard Error | 1.85 |
| AAPE | 16.65% |

previously used, in AAPE value, with a decrease of almost 10%. With this in mind, the higher value of standard error obtained for this regression indicates there is more variability in the chosen solvents, and not necessarily less reliable estimates.

The values of log $k$ calculated through Eq. 13, and the associated ranking are listed in Table 10, along with the measured log $k$ values, for the 14 best solvents tested experimentally. It can be seen that, qualitatively, the model is still very satisfactory. Glycerol is ranked first by prediction and by experiment. If we consider the first 14 solvents, for which the log $k$ values are close together, we can see that the two sets of solvents match well. This shows that the proposed model can be used to guide experiments. Although the experimental and predicted ranking differ in the details, the top 14 solvents identified using Eq. 13 are all good candidates for this reaction. The use of the predicted values for solvent properties gives more flexibility, allows a broader space of chemical compounds to be explored, and, hence, brings innovation to the problem of solvent selection.

## Computer-Aided Solvent Design

### *Deterministic design formulation*

Having developed a model of solvent effects for a given reaction based on a few data points, we now focus on generating new solvent candidates. The objective considered is to maximize log $k$ as calculated from Eq. 4. The constraints consist of the group contribution methods for solvent property prediction, chemical feasibility and complexity constraints and design constraints. Chemical feasibility and complexity constraints include standard constraints on the maximum and minimum number of groups in the molecule, constraints on the maximum number of main and functional groups, as well as constraints that forbid or limit the occurrence of some groups together. The octet rule[39] and the bonding rule (as modified by Buxton et al.,[40]) are included

**Table 10. Predicted Log $k$ Values and Ranking for the 15 Solvents with Highest Measured Log $k$ Values**

| Solvent | log $k_{exp}$ | log $k_{pred}$ | Predicted Ranking |
|---|---|---|---|
| Glycerol | −4.1 | −0.67 | 1 |
| 1,2-Ethanediol | −4.6 | −4.57 | 3 |
| Phenol | −4.66 | −6.04 | 10 |
| Propane-1,3-diol | −5.29 | −5.25 | 5 |
| Propane-1,2-diol | −5.51 | −5.54 | 6 |
| Diethylene glycol | −5.69 | −4.56 | 4 |
| Butane-1,4-diol | −5.99 | −5.67 | 7 |
| Triethylene glycol | −6.07 | −3.87 | 2 |
| Butane-1,2-diol | −6.14 | −5.92 | 8 |
| Aniline | −6.15 | −7.69 | 15 |
| Butane-2,3-diol | −6.21 | −6.21 | 11 |
| Pentane-1,5-diol | −6.32 | −5.98 | 9 |
| N-Methylformamide | −6.54 | −7.35 | 13 |
| Acetic acid | −6.7 | −6.98 | 12 |

as well to ensure there are no free attachments and no double bonds between atom groups. The choice of constraints which limit the complexity of the molecule formed is based on the recognition that solvents are usually medium-size molecules, because they must have a useful liquid range. In addition, the group contribution methods used here do not account for proximity effects beyond the make-up of atom group and, are, therefore, most reliable when used for medium-size molecules. The CAMD formulation is a mixed-integer linear-programming (MILP) problem. Integer cuts are included to allow the generation of successive solutions, giving a ranked list of candidate solvents. The general formulation is given as follows

$$\begin{aligned} &\max_{k,p,n,y} \log k \\ &\text{s.t.} \quad h_1(k,p,n,y) = 0 \\ &\qquad g_1(p,n,y) \le 0 \\ &\qquad h_2(n,y) = 0 \\ &\qquad g_2(n,y) \le 0 \\ &\qquad p \in \mathbb{R}^m \\ &\qquad n \in \mathbb{R}^r \\ &\qquad k \in \mathbb{R} \\ &\qquad y \in \{0,1\}^q \end{aligned} \tag{14}$$

where $\log k$ is the objective function, $h_1$ is a set of structure-property equality constraints; $h_2$ a set of chemical feasibility and complexity equality constraints; $g_1$ a set of structure-property inequality constraints; $g_2$ a set of chemical feasibility and complexity inequality constraints; $k$ is a scalar; $p$ is an $m$-dimensional vector of continuous variables denoting the properties; $n$ is a $r$-dimensional vector of continuous variables denoting the number of groups in the molecule; and $y$ is a set of binary variables (for example, used to constrain the $n$ variables to integer values).

A more detailed formulation of the particular problem is given in the remainder of this section.

*Objective Function*

$$\max \quad \log k = \log k_0 + sS + d\delta + aA + bB + h\delta_H^2/100 \tag{15}$$

with the reaction parameters obtained from Step 1.

*Structure-Property Constraints.* To calculate $A$, the hydrogen bond acidity, a binary variable $y_A$ is defined as:

$$y_A = \begin{cases} 1 & \text{if} \sum_{i \in G} n_i A_i \ge 0.018 \\ 0 & \text{otherwise} \end{cases}$$

where $G$ is the set of 43 groups listed in Table 6.

Equations 16 and 17 determine the value of $y_A$

$$\sum_{i \in G} n_i A_i - 0.018 - M y_A \le 0 \tag{16}$$

$$M(y_A - 1) - \sum_{i \in G} n_i A_i + 0.018 \le 0 \tag{17}$$

where $M$ is a large enough positive number. We use a value of 100. Then $A$ is

$$A = \begin{cases} \sum_{i \in G} n_i A_i + 0.010641 & \text{if } y_A = 1 \\ 0 & \text{otherwise} \end{cases}$$

Constraints 18, 19, and 20 determine the value of $A$ for a given $y_A$

$$-A + \sum_{i \in G} n_i A_i + 0.010641 + (y_A - 1) \le 0 \tag{18}$$

$$0 \le A \le M y_A \tag{19}$$

$$A - \sum_{i \in G} n_i A_i - 0.010641 \le 0 \tag{20}$$

To calculate $B$, the hydrogen bond basicity, a binary variable $y_B$ is defined as

$$y_B = \begin{cases} 1 & \text{if } \sum_{i \in G} n_i B_i \ge 0.00003 \\ 0 & \text{otherwise} \end{cases}$$

Equations 21 and 22 determine the value of $y_B$

$$\sum_{i \in G} n_i B_i - 0.00003 - M y_B \le 0 \tag{21}$$

$$M(y_B - 1) - \sum_{i \in G} n_i B_i + 0.00003 \le 0 \tag{22}$$

where $M$ is as previously defined. Then $B$ is

$$B = \begin{cases} \sum_{i \in G} n_i B_i + 0.12371 & \text{if } y_B = 1 \\ 0 & \text{otherwise} \end{cases}$$

Constraints 23, 24, and 25 determine the value of $B$ for a given $y_B$

$$-B + \sum_{i \in G} n_i B_i + 0.12371 + (y_B - 1) \le 0 \tag{23}$$

$$0 \le B \le M y_B \tag{24}$$

$$B - \sum_{i \in G} n_i B_i - 0.12371 \le 0 \tag{25}$$

$\delta$ is the solvent "polarizability correction" parameter, equal to 1 for aromatics, 0.5 for polychlorinated aliphatics, and 0 for other compounds. $y_1$ and $y_2$ are binary variables defined as

$$y_1 = \begin{cases} 1 & \text{if } \sum_{i \in G_A} n_i \ge 1 \\ 0 & \text{otherwise} \end{cases}$$

$$y_2 = \begin{cases} 1 & \text{if } \sum_{i \in G_H} h_i n_i \ge 2 \\ 0 & \text{otherwise} \end{cases}$$

where $G_A$ is the set of aromatic groups; $h_i$ is the number of halogen atoms in group $i$, and $G_H$ is the set of halogen containing groups. Constraints 26 and 27 determine the value of $y_1$, and constraints 28 and 29 determine the value of $y_2$

$$\sum_{i\in G_A} n_i \le \left(\sum_{i\in G_A} n_i\right)_{max} y_1 \tag{26}$$

$$\sum_{i\in G_A} n_i \ge y_1 \tag{27}$$

$$2y_2 \le \sum_{i\in G_H} h_i n_i \tag{28}$$

$$y_2\left(\sum_{i\in G_H} h_i n_i\right)_{max} \ge \sum_{i\in G_H} h_i n_i - 0.26\left(\sum_{i\in G_H} h_i n_i\right)_{max} \tag{29}$$

where *max* denotes the maximum possible values of the summations. Then

$$\delta = y_1 + 0.5y_2 \tag{30}$$

$S$ is Abraham's dipolarity/polarizability property for the solvent, and $S_i$ is the coefficient for group $i$

$$S = \sum_{i\in G} n_i S_i + 0.326 \tag{31}$$

$\Delta H_V$ is the solvent's enthalpy of vaporization at a temperature of 298K, and at a pressure equal to the vapor pressure of the compound at this temperature, and $H_{V,i}$ is the coefficient for group $i$

$$\Delta H_V[298K] = \sum_{i\in G} n_i H_{V,i} + 6.829 \tag{32}$$

$V_m$ is the solvent's liquid molar volume, and $V_{m,i}$ is the coefficient for group $i$

$$V_m = \left(\sum_{i\in G} n_i V_{m,i} + 0.012\right) \tag{33}$$

$\delta_H$ is the Hildebrand solubility parameter

$$\delta_H^2 = 0.239\frac{\Delta H_V - RT}{V_m} \tag{34}$$

We multiply the correlation with 0.239 to convert the units of $\delta_H^2$ (MPa to cal cm$^{-3}$). This nonlinear expression is linearized using the method proposed by Maranas[38] for the reformulation of structure-property functions, as shown in the Appendix.

*Chemical Feasibility and Complexity Constraints.* The type of molecule designed (acyclic, bicyclic, and monocyclic) is represented by three binary variables. $y_5 = 1$ gives an acyclic molecule, $y_6 = 1$ a bicyclic solvent molecule, and $y_7 = 1$ a monocyclic molecule[39]

$$y_5 + y_6 + y_7 = 1 \tag{35}$$

$m$ is a continuous variable representing the type of molecule. For a monocyclic molecule, $m = 0$, for an acyclic molecule $m = 1$, and for a bicyclic molecule $m = -1$. This is expressed in terms of the binary variables as

$$m - (y_5 - y_6) = 0 \tag{36}$$

The octet rule[39] ensures that the solvent molecule designed is structurally feasible and that there are no remaining free attachments in the molecule. It is based on the valency ($v_i$) of different structural groups

$$\sum_{i\in G}(2 - v_i)n_i - 2m = 0 \tag{37}$$

The following constraint ensures that the molecule contains at least two groups

$$2 - \sum_{i\in G} n_i \le 0 \tag{38}$$

The following constraint ensures that the molecule contains no more than 10 groups

$$\sum_{i\in G} n_i \le 10 \tag{39}$$

In an aromatic molecule, the number of aromatic groups must equal 6 if the molecule is monocyclic, or 10 if it is bicyclic

$$\sum_{i\in G_A} n_i - 6y_7 - 10y_6 = 0 \tag{40}$$

In a monocyclic aromatic molecule, we also allow the occurrence of side-chains which consist of at most two non-aromatic groups. For the purpose of building such molecules, we define three new binary variables, $y_{aC}$, $y_{aCCH}$ and $y_{aCCH2}$. These are defined as

$$y_j = \begin{cases} 1 & \text{if group } j \text{ is present in the molecule} \\ 0 & \text{otherwise} \end{cases}$$

$$j = \{aC, aCCH, aCCH_2\}$$

To limit the complexity of the molecules designed, only one of the three groups (*aC, aCCH, aCCH*$_2$) is allowed to appear in a molecule

$$y_{aC} + y_{aCCH} + y_{aCCH2} = 1 \tag{41}$$

The composition of the side-chains is limited to only certain aliphatic groups. We categorize these groups into non-chain-ending (which are the groups allowed on the "first position," directly linked to the aromatic group) and chain-ending (which are the groups allowed on the "second position," linked to one of the non-chain-ending groups). Non-chain-ending groups belong to set Nceg, and chain-ending groups to set Ceg. Both sets are shown in Table 11.

The *aCCH* group leads to the presence of two side chains. At least one of these chains must be the $CH_3$ group:

$$y_{aCCH} \le n_{CH_3} \tag{42}$$

In a bicyclic molecule the number of aromatic carbon groups (*aC*) must be greater than or equal to 2

$$2y_6 - n_{aC} \le 0 \tag{43}$$

It is important to differentiate between an *aC* group present in a monocyclic or a bicyclic aromatic. If the *aC* group

**Table 11. List of the Groups Allowed in the Side-Chains of Monocyclic Aromatic Molecules**

| Set 'Nceg' | Set 'Ceg' |
|---|---|
| $CHCH$ | $CH_3CO$ |
| $CH_2CO$ | $CH_2CH$ |
| $CH_2COO$ | $CH_3COO$ |
| $CH_2O$ | $CH_3O$ |
| $CH_2NH$ | $CHO$ |
| $CH_3N$ | $CH_2NH_2$ |
| $CHNO_2$ | $COOH$ |
| $CH_2SH$ | $CH_2CN$ |
| | $CH_2Cl$ |
| | $CH_2NO_2$ |
| | $I$ |
| | $Br$ |

is present in the molecule, binary variable $y_{aC}$ is equal to one, whether the molecule is mono or bicyclic. However, a side-chain is only allowed if the molecule is monocyclic. Therefore, we introduce a binary variable $y_M$ that is defined as follows

$$y_M = \begin{cases} 1 & \text{if } y_{aC} + y_7 = 2 \text{ (that is a monocyclic molecule with an } aC \text{ group)} \\ 0 & \text{otherwise} \end{cases}$$

The following constraints determine the value of $y_M$

$$0 \le y_M \le y_{aC} \tag{44}$$

$$y_7 - 1 + y_{aC} \le y_M \le y_7 \tag{45}$$

The binary variables $y_{i,k}$ ensure that the values of the continuous variables $n_i$ are integer

$$\sum_{k=0}^{K} 2^k y_{i,k} - n_i = 0, \quad \forall i \in G \tag{46}$$

The following set of constraints (one for each group $n_j$) represents the bonding rule (as modified by Buxton et al.[40] which states that two adjacent groups in a molecule may not be linked by more than one bond. This ensures that double bonds do not form

$$n_j(v_j - 1) + 2m - \sum_{i \in G} n_i \le 0, \quad \forall j \in G \tag{47}$$

The following constraint states that the presence of functional groups is allowed only when an acyclic or a monocyclic molecule is designed, and that their number per molecule is restricted to $n_{FA,max}$ or $n_{FM,max}$, respectively

$$\sum_{i \in G_F} n_i - n_{FA,max} y_5 - n_{FM,max} y_7 \le 0 \tag{48}$$

where $G_F$ is the set of functional groups. The functional groups are groups that contain an atom other than $C$ and $H$.

The maximum number of occurrences for all the groups is limited as follows

$$n_i \le n_i^U, \quad \forall i \in G \tag{49}$$

Values/expressions for $n_i^U$ for each group $i$ are shown in Table 12.

We include integer cuts to ensure that any combination $y^p$ of binary variables $y_{i,k}$ (where $k = 1, \ldots, K$ as in Eq. 46) is not generated twice, so that we can, therefore, generate successive solutions. For candidate $p$, we define $Z^p = \{i : y_i^p = 0\}$ and $NZ^p = \{i : y_i^p = 1\}$, then the constraint is given by

$$\sum_{i \in NZ^p} \sum_{k=1}^{K} y_{i,k} - \sum_{i \in Z^p} \sum_{k=1}^{K} y_{i,k} \le |N|^p - 1 \tag{50}$$

*Design Constraints.* The solvent obtained should be liquid at room-temperature. The following correlation, taken from[41], is used to predict $T_m$ at the level of first-order groups

$$\exp(T_m / T_{m0}) = \sum_{i \in G} n_i T_{m,i} \tag{51}$$

where $T_{m,i}$ is the contribution of group $i$, $i \in G$, towards predicting $T_m$.

The design constraint is chosen to account for the average absolute error in the predictions reported to be 24.90 K.

**Table 12. Upper Bounds $n_i^U$ on the Occurrence of Each Group**

| Group $i$ | $n_i^U$ |
|---|---|
| $CH_3$ | $7y_5 + y_{aCCH}$ |
| $CH_2$ | $7y_5$ |
| $CH$ | $3y_5$ |
| $C$ | $y_5$ |
| $CH_2{=}CH$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH{=}CH$ | $y_5 + y_M + y_{aCCH_2}$ |
| $CH_2{=}C$ | $y_5$ |
| $CH{=}C$ | $y_5$ |
| $C{=}C$ | $y_5$ |
| $aCH$ | $6y_7 + 8y_6$ |
| $aC$ | $y_7 + 2y_6$ |
| $aCCH_3$ | $6y_7 + 8y_6$ |
| $aCCH_2$ | $y_7$ |
| $aCCH$ | $y_7$ |
| $OH$ | $3y_5$ |
| $aCOH$ | $6y_7 + 8y_6$ |
| $CH_3CO$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH_2CO$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CHO$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH_3COO$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH_2COO$ | $y_5 + y_M + y_{aCCH_2}$ |
| $CH_3O$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH_2O$ | $y_5 + y_M + y_{aCCH_2}$ |
| $CH{-}O$ | $y_5$ |
| $CH_2NH_2$ | $2y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH_3NH$ | $y_5$ |
| $CH_2NH$ | $y_5 + y_M + y_{aCCH_2}$ |
| $CH_3N$ | $y_5 + y_M + y_{aCCH_2}$ |
| $CH_2N$ | $y_5$ |
| $aCNH_2$ | $6y_7 + 8y_6$ |
| $CH_2CN$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $COOH$ | $y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CH_2Cl$ | $2y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CHCl$ | $y_5$ |
| $CHCl_2$ | $y_5$ |
| $aCCl$ | $6y_7 + 8y_6$ |
| $CH_2NO_2$ | $2y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $CHNO_2$ | $y_5 + y_M + y_{aCCH_2}$ |
| $CH_2SH$ | $y_5 + y_M + y_{aCCH_2}$ |
| $I$ | $2y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $Br$ | $2y_5 + y_M + y_{aCCH} + y_{aCCH_2}$ |
| $aCF$ | $6y_7 + 8y_6$ |
| $CH_2S$ | $y_5$ |

Therefore, instead of imposing a bound of 298 K on the melting point, we relax it to 315 K. Once we substitute the value for $T_{m0}$ in Eq. 51, and reformulate it into an equivalent linear equation, we obtain the following set of constraints

$$T_m = \sum_{i \in G} n_i (T_m)_i \tag{52}$$

$$T_m \leq 8.6 \tag{53}$$

### *Application to the solvolysis reaction*

To apply the proposed approach to the solvolysis reaction, the reaction coefficients derived in Regression 4 are used (Eq. 13). The deterministic CAMD optimization problem involves constraints 16–31, 35–53, and 57–62. Its solution yields a ranked list of candidate solvents that are predicted to perform optimally when used as a reaction medium for the solvolysis of t-butyl chloride. To limit the complexity of the molecules designed, $K$ in Eq. 46 is set to 3, so there can be at most seven groups of the same kind in the molecule.

The optimization problem formulation is implemented in GAMS[42] and solved using CPLEX. Based on the parameters chosen for this study and the chemical and complexity constraints, the search space consists of 8,443 feasible molecular structures. A larger space can be explored if some of these restrictions are lifted. In spite of the large dimensionality of the problem, it is solved in a matter of seconds on a single CPU. Thus, the complexity of the molecules that can be designed is limited by the reliability of the prediction methods used and by physical constraints, such as the boiling point of the solvent, rather than by computational requirements.

In Table 13, we show a list of the 10 best ranked solvents, and their objective function values. The best three solvents are shown in Figure 7.

Nine out of the ten candidate solvents in Table 13 are alcohols—either diols or triols, which points to this chemical class as a source of good solvents for the solvolysis reaction. This is in agreement with experimental results, in which alcohols are clustered at the top of the experimental ranking provided in Tables 4 and 7 and are, thus, proven to give good performance.

## Solvent Design under Uncertainty

Since we build our model based on kinetic data in a few solvents only, we expect there is uncertainty associated with the reaction parameters. This is confirmed by the very wide 95% confidence interval calculated for each of the parameters (for example, see Table 14). Therefore, we investigate the impact of uncertainty on model reliability and determine the optimal solvent candidate given this uncertainty. The first task is to perform a global sensitivity analysis in order to identify the key reaction parameters, and the most frequently identified "optimal" solvents across the uncertain parameter space. Then, we formulate and solve a stochastic (scenario-based) optimization problem for a small number of representative scenarios distributed throughout the solution map (that is, the uncertainty space). These two steps are explained in more detail in the next sections, and illustrated on the solvolysis reaction, using Eq. 13 as the reaction model.

**Table 13. Ranked List of Top 10 Solvents Obtained in the Deterministic Design Step for the Solvolysis Reaction**

| Solvent Rank | Solvent Name | Objective Function |
|---|---|---|
| 1 | Glycerol | −0.51 |
| 2 | 1,1,2-Propanetriol | −0.80 |
| 3 | 1,1,2-Butanetriol | −1.47 |
| 4 | 1,2-Ethanediol | −4.40 |
| 5 | 1,2-Dihidroxy-2-propene | −4.61 |
| 6 | 1,1-Ethanediol | −4.72 |
| 7 | 1,2-Dihidroxypropene | −4.73 |
| 8 | 1,3-Dihidroxypropene | −4.85 |
| 9 | o-(2-nitrol-3-butenyl)aniline | −5.00 |
| 10 | 1,1-Dihidroxy-2-propene | −5.03 |

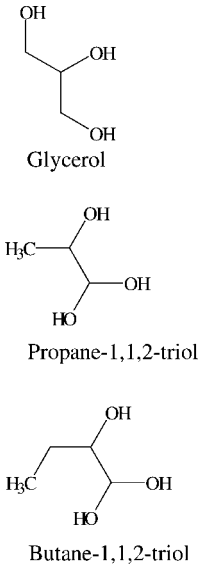


**Figure 7. Structures of the three best molecules generated in the design step.**

### *Sensitivity analysis*

In order to analyze the model's sensitivity, we start by solving the design problem for different values of the reaction parameters ($a$, $b$, $s$, $d$ and $h$; we treat log $k_0$ as a constant). Parameter values are changed one-at-a-time in the first instance. The effect of varying several parameters at a time is also considered. We find the optimal design for each parameter combination and undertake a comparative study of the optimal designs generated for different uncertain parameter realizations.

**Table 14. Lower and Upper Bound of the 95% Confidence Intervals for the Reaction Parameters in Eq. 13**

| | 95 % Confidence Interval | |
|---|---|---|
| Parameter | Lower Bound | Upper Bound |
| $s$ | −22.8 | 28.8 |
| $d$ | −11.8 | 13.7 |
| $a$ | −29.0 | 39.7 |
| $b$ | −21.5 | 25.7 |
| $h$ | −11.3 | 14.2 |

The use of global sensitivity methods[43] allows us to explore the entire uncertain parameter space for the set of parameters being varied, rather than perturbations around the nominal parameter values. When varying one parameter at a time, we use a uniform distribution to sample the space. When varying several parameters simultaneously, we explore the uncertain parameter space using an approach proposed by Sobol'[44,45]. In this technique, a low discrepancy or quasi-random sampling sequence is generated. One of the key ideas behind the Sobol' approach is to produce a sequence whose projection into any direction yields no overlapping points. Thus, in two dimensions $a$ and $b$, a sequence of 128 points will give 128 different values of $a$, and 128 different values of $b$. It is common to choose a number of sampling points which is a power of two to get as uniform a distribution of points as possible over the uncertainty space. For every parameter sample, we solve the molecular design step and identify the best solvent candidate. As a measure of sensitivity, we consider the number of designs generated in the molecular design step compared to the width of the parameter range.

We perform sensitivity analyses by varying one, two and five parameters at a time. In each case, we study the impact of the sample size on the results and specifically on the number of different solvent molecules designed.

For the one-dimensional (1-D) studies a uniform distribution is used, and the sample size is increased by 500 points from run to run. Convergence to a fixed set of top solvents is achieved with only 1,000 points.

For 2- and 5-D problems, we use the Sobol' sequence to sample the uncertainty space, starting with 1024 ($2^{10}$) sampled points, and increasing the number in each run by 1024 points. Graphs showing the change in the number of designs with the number of sampled points for a representative 2-D case and for the five-dimensional case are presented in Figures 8 and 9, respectively. It is shown in Figure 8 that the number of designs does not change after 3072 ($2^{12}$) points. The number of designs continues to increase after 10,240 points when five parameters are varied simultaneously (Figure 9), but the increase is relatively slow. Molecules generated from the first 2,048 points appear frequently as optimal molecules, whereas molecules generated from later points appear to be the optimal solvent for very few parameter realizations. The number of designs generated is small considering that over 8,000 molecules could be generated.

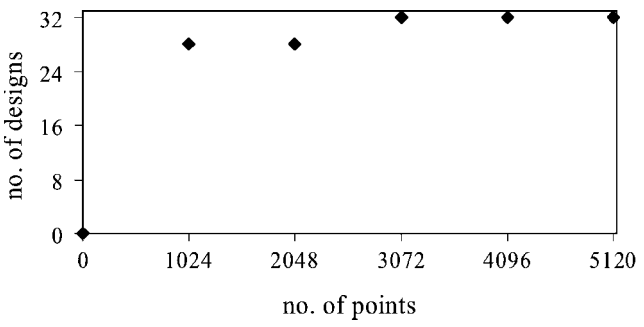


**Figure 8. Number of generated designs vs. the number of sampling points for 2-D uncertainty space *s–h*.**

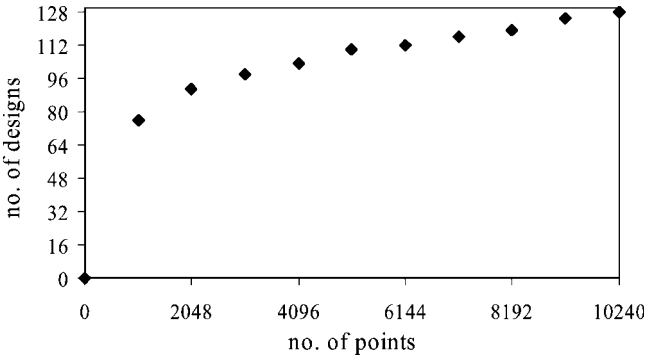


**Figure 9. Number of generated designs vs. the number of sampling points for 5-D uncertainty space.**

The numbers of designs generated from the 1-D problems, together with the width of the range of each parameter, are shown in Table 15. As we can see, the design is almost insensitive to parameters $d$ and $a$ and slightly more sensitive to parameter $b$. Parameters $s$ and $h$ give the largest numbers of distinct designs. This is to be expected since these parameter multiply the solvent dipolarity/polarizability and cohesive energy density, respectively, and those are the solvent properties with the highest values. Therefore, they contribute more to the value of log $k$. From a physical point of view, for a solvolysis reaction where the rate limiting step is the formation of a solvent-separated ion pair, these two properties are the most influential, explaining the high sensitivity of the design.

Overall, thirteen different solvents are generated in this set of parametric studies. They are listed in Table 16, together with the run in which they were generated. To make the comparison easier, the molecular structures (atom groups) are given instead of the molecule names. The molecules are rank ordered, with the molecule designed most frequently appearing first. The cumulative occurrence fraction of each solvent is plotted in Figure 10. The occurrence fraction is defined as the ratio of the number of times the molecule appears as optimal in all of the optimization runs over the total number of optimizations performed.

As can be seen in Table 16 and Figure 10, glycerol (molecule 1) appears no matter which parameter is varied, and it appears most frequently. This would point to glycerol as the main candidate for experimental verification or for direct use as a solvent for solvolysis of t-butyl chloride. This is also the

**Table 15. Sampled Parameter Range and Number of Distinct Designs from 1-D Sensitivity Runs for All Five Parameters**

| Parameter Varied | Range | No. of Designs Generated |
|---|---|---|
| $s$ | 51.5 | 6 |
| $d$ | 25.4 | 2 |
| $a$ | 68.7 | 2 |
| $b$ | 47.2 | 3 |
| $h$ | 25.5 | 6 |

**Table 16. List of the Design Solutions for the 1-D Sensitivity Runs**

| Mol. no. | Mol. Structure | Run |
|---|---|---|
| 1 | $CH_2 \times 2$, $CH \times 1$, $OH \times 3$ | *s,d,a,b,h* |
| 2 | $CH_2 \times 1$, $CH_2C \times 1$, $CH_2NO_2 \times 2$ | *s,a* |
| 3 | $CH_2CH \times 1$, $aCH \times 4$, $aCCH_2 \times 1$, $aCNH_2 \times 1$, $CHNO_2 \times 1$ | *d,h* |
| 4 | $aCH \times 2$, $aCCH_2 \times 1$, $aCCl \times 3$, $I \times 1$ | *b* |
| 5 | $aCH \times 5$, $aCOH \times 1$ | *b* |
| 6 | $CH_2CH \times 1$, $aCH \times 4$, $aCCH_2 \times 1$, $CH_3N \times 1$, $aCNH_2 \times 1$ | *h* |
| 7 | $CH_3 \times 1$, $CH \times 2$, $OH \times 1$ | *s* |
| 8 | $CH_3 \times 4$, $CH_2 \times 3$, $CH \times 1$, $CHC \times 1$ | *s* |
| 9 | $CH_3 \times 3$, $CH_2 \times 1$, $CH \times 1$, $C \times 1$, $CH_2C \times 1$, $CH_2NH_2 \times 1$, $CHCl_2 \times 1$ | *h* |
| 10 | $CH_3 \times 3$, $CH_2 \times 1$, $CH \times 2$, $C \times 1$, $CH_2CH \times 1$, $CH_2NH_2 \times 1$, $CHCl_2 \times 1$ | *h* |
| 11 | $CH_3 \times 5$, $CH_2 \times 1$, $CH \times 3$, $CH_2NH \times 1$ | *s* |
| 12 | $CH_3 \times 1$, $CH_2 \times 1$, $CH \times 2$, $OH \times 3$ | *h* |
| 13 | $CH_3 \times 5$, $CH \times 3$, $CHC \times 1$, $OH \times 1$ | *s* |

top solvent identified in the deterministic run (see Table 13) at nominal parameter values.

In Table 16, we can see that approximately half the molecules are aliphatic, and half are aromatic. There is structural similarity between several molecules (such as molecules 9 and 10, or 1 and 12), which can be used to reduce the number of candidate chemicals even further. It is also clear that sampling the uncertainty space of each parameter independently results in the generation of certain characteristic type of designs which are specific to each parameter (with the exception of glycerol, which is common to all the parameters).

For a more realistic assessment of the impact of uncertainty, the results obtained when varying two parameters at a time are shown in Table 17. As expected, there is an increase of number of designs generated. The increase is larger if we vary two parameters that describe different interactions in the system. A relatively small number of designs is obtained when combining the two hydrogen-bonding indicators (*a* and *b*), or the two polarity indicators (*s* and *d*). The largest number of designs is obtained when varying simultaneously parameters *s* and *h*, which is consistent with the larger number of designs obtained for those two parameters in the 1-D runs. The model is most sensitive to these two parameters, and their combination gives the largest number of different designs.

Finally, to further assess the impact of uncertainty, we vary all the parameters at the same time. For this 5-D analysis, we obtain 128 designs in 10,240 optimizations. Glycerol, which is the optimal solvent identified in the deterministic optimization step (see Table 13), appears over 25% of the time.

This set of sensitivity runs highlights the most important parameters and parameter interactions, and points to some key solvent candidates.

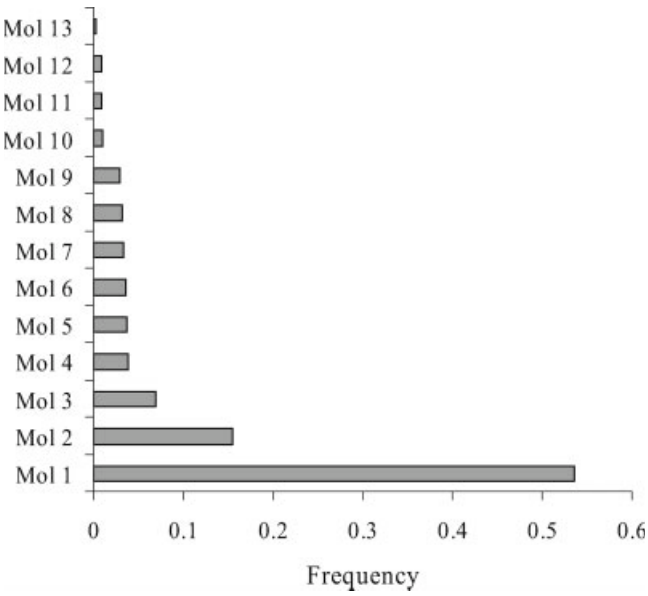


**Figure 10. Molecule number vs. its cumulative occurrence fraction for 1-D problems.**

### *Stochastic optimization*

Based on the results of the sensitivity analysis, we can proceed to formulate and solve a stochastic (scenario-based) optimization problem for the uncertain parameter space. We formulate the problem for a small number of representative scenarios that are distributed throughout the solution map of designs for different parameter realizations as obtained via sensitivity analysis.

The procedure for identifying a suitable set of scenarios is based on the fact that, because of the linear nature of the problem, the optimal candidate solvents found on the solution map are clustered. We first compute the convex hull of the area corresponding to each solvent candidate. Then, we can divide each convex hull into subareas. The number of subareas allows a compromise between good coverage of the space and the number of scenarios. The center of mass of

**Table 17. Results for 2-D Sensitivity Analysis Runs**

| Parameters Varied | Size of Parameter Space | No. of mols. Generated |
|---|---|---|
| *s* and *d* | $51.5 \times 25.4$ | 4 |
| *a* and *b* | $68.7 \times 47.2$ | 9 |
| *s* and *a* | $51.5 \times 68.7$ | 15 |
| *s* and *b* | $51.5 \times 47.2$ | 16 |
| *s* and *h* | $51.5 \times 25.5$ | 32 |

The size is defined as the area/volume defined by the 95% confidence intervals on the parameters varied.

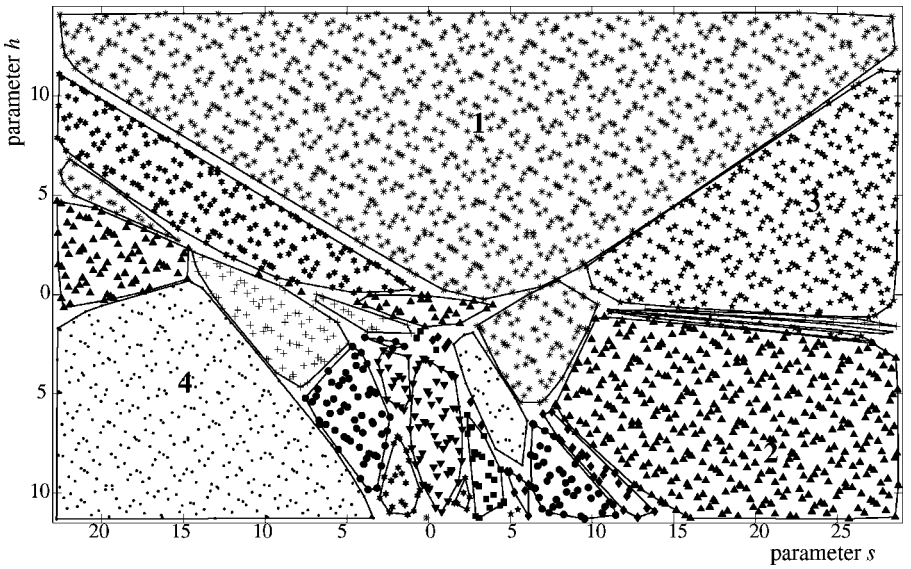


**Figure 11. Parameter space *s–h* divided into clustered areas corresponding to different design solutions.**

The lines indicate the convex hulls used for the case study. All 32 molecules generated are presented using different symbols. The largest areas are numbered and correspond to the following molecules: 1: glycerol, 2: 8-nitro-2-(nitromethyl)oct-1-ene, 3: 4-nitro-2-(nitromethyl)but-1-ene. 4: 2-methoxy-3,4,5-trimethylhexane.

each of the subareas is then found and added to the list of scenarios. To weigh the scenarios in the objective function, we attribute a weight factor proportional to the size each subarea to each center of mass.

Once the key scenarios have been determined, we formulate and solve a stochastic optimization problem. The objective is to find the molecule that has the highest expected value of the reaction rate constant averaged over all the weighted scenarios

$$\max_{k,n,y} \frac{1}{M} \sum_{i=1}^{M} w_i \log k_i$$

$$\text{s.t.} \quad \text{constraints } 16-31, 35-53, 57-62$$

$$\log k_i = \log k_0 + s_i S + d_i \delta + a_i A + b_i B + h_i \delta_H^2/100 \qquad i = 1, \ldots, M \tag{54}$$

where $w_i$ is the weight factor for scenario $i$; $s_i$, $d_i$, $a_i$, $b_i$, $h_i$ denote the reaction parameter values for scenario $i$; $k_i$ the corresponding reaction rate constant, and $M$ the number of scenarios.

Since we have identified parameters $s$ and $h$ as the key parameters, we first focus on their combination to obtain a robust solution in the 2-D uncertainty space. Other parameters are kept at their nominal values. The solution map for the 2-D space is shown in Figure 11. For those solvent candidates covering a sufficiently large area, three subareas and their centers of mass are identified. For other solvent candidates, a single area is used. By solving the stochastic optimization problem with the 58 resulting scenarios, we identify glycerol as the solvent with the maximum average rate constant. The implementation of the integer cut constraint in the problem formulation allows the generation of a list of candidate solvents, which is identical to that found in the deterministic CAMD step and shown in Table 13.

Next, the 5-D solution map is used to identify scenarios. Because many of the convex hulls are very small, a single scenario (the center of mass) is generated for each design identified. There are, therefore, 128 scenarios. The solution of this problem yields glycerol as the top solvent, and the same list of top 10 solvents is identified (Table 13). This list is generated within a few seconds on a single processor.

The consistency of the results indicates that, despite the large uncertainty in the parameters, the model is sufficiently reliable for use in solvent design. Experiments and plant-wide studies can, therefore, be undertaken to assess the solvents listed in Table 13. In some cases, the stochastic design or experimental verification of the top solvents may show significant deviations from the predictions of the deterministic optimization. The solvent candidate may be significantly worse, or better, than predicted. In this case, reaction rate measurements for the new solvents found should be added to the set of solvents used for regression (for example see[13]). This results in an improved reaction model, which can then be used for computer-aided solvent design. This procedure can be repeated until the results of experimental verification and the model predictions are in qualitative agreement.

## Concluding Remarks

We have proposed a systematic approach to solvent design for enhanced reaction rate constant, which combines experimentation, modeling and optimization. We have shown that, by making use of computational tools, the number of experiments required can be limited, thus reducing development costs. A simple model, based on the solvatochromic equation and group contribution techniques, is used. Its application to

a solvolysis reaction shows that it provides adequate quantitative predictions, and a good qualitative assessment of the suitability of a wide range of solvents. The model is used to identify optimal candidate solvents, based on a deterministic and a stochastic problem formulation. The scenarios for the stochastic problem are identified from a global sensitivity analysis, which pinpoints key parameters and the most likely solvent candidates. In spite of the simplicity of the model and the significant uncertainty, consistent results are obtained, indicating the robustness of the approach for this case study.

This promising extension of CAMD to solvent design for reactions is built on a framework which fits in with other CAMD applications, such as separations design, and can, therefore, be integrated in process-wide design methods. However, before tackling such problems, it is desirable to develop further the proposed methodology. In particular, for other reactions, the presence of uncertainty may necessitate a revision of the model based on further experimentation. The methodology should also be extended to more complex reaction schemes, such as multistep, concurrent and two-phase reactions.

Finally, this work shows that the solvatochromic equation could be used successfully in other molecular design applications. The equation has been used to predict partition coefficients, Gibbs free energies of transfer and many other properties of complex systems. Using the group contribution methods developed here to predict the solvatochromic parameters of a large set of compounds, it becomes straightforward to embed the equation in standard CAMD tools.

## Acknowledgments

The authors wish to thank Professor M. H. Abraham for providing some of the solvatochromic parameter data and useful comments. M. Folić gratefully acknowledges financial support from the ORS scheme.

## Literature Cited

1. Reichardt C. Solvents and Solvent Effects in Organic Chemistry. Weinheim: VCM Publishers; 1988.
2. Hughes ED, Ingold CK. Mechanism of substitution at a saturated carbon atom. Part IV, a discussion of constitutional and solvent effects on the mechanism, kinetics, velocity and orientation of substitution. *J of the Chem Soc*. 1935;806–809.
3. Hughes ED, Ingold CK. Mechanism and kinetics of substitution at a saturated carbon atom. *Trans of the Faraday Society*. 1941;37:603–657.
4. Achenie LEK, Gani R, Venkatasubramanian V, (ed). *Computer-Aided Molecular Design: Theory and Practice*. Amsterdam: Elsevier Science; 2002.
5. Gani R, Brignole EA. Molecular design of solvents for liquid extraction based on UNIFAC. *Fluid Phase Equilibria*. 1983;13:331–340.
6. Abraham MH. Substitution on saturated carbon. Part XIV. Solvent effects on the free energies of ions, ion-pairs, non-electrolytes, and transition states in some $S_N$ and $S_E$ reactions. *J of the Chem Soc - Perkin Trans II*. 1972;1343–1357.
7. Abraham MH, Doherty RM, Kamlet MJ, Harris JM, Taft RW. Linear solvation energy relationships. 15. Heterolytic decomposition of the *tert*-Butyl halides. *The J of Organic Chemistry*. 1981;46: 3053–3056.
8. Abraham MH, Doherty RM, Kamlet MJ, Harris JM, Taft RW. Linear solvation energy relationships. Part 37. An analysis of contributions of dipolarity-polarisability, nucleophilic assistance, electrophilic assistance, and cavity terms to solvent effects on t-Butyl halide solvolysis rates. *J of the Chem Soc - Perkin Trans II*. 1987;913–920.
9. Cramer CJ. *Essentials of Computational Chemistry - Theories and Models*. Chichester: John Wiley and Sons; 2005.
10. Modi A, Aumond JP, Mavrovouniotis ML, Stephanopoulos G. Rapid plant-wide screening of solvents for batch processes. *Comp & Chem Eng*. 1996;20:S375–S380.
11. Gani R, Jiménez-González C, Constable DJC. Method for selection of solvents for promotion of organic reactions. *Comp & Chem Eng*. 2005;194–197:1661–1676.
12. Folić M, Adjiman CS, Pistikopoulos EN. The design of solvents for optimal reaction rates. In: Barbosa A, Matos H. *Computer-Aided Chemical Engineering*. Amsterdam: Elsevier Science; 2004;18:175–181.
13. Folić M, Adjiman CS, Pistikopoulos EN. A computer-aided methodology for optimal solvent design for reactions with experimental verification. In: Puigjaner L. *Computer-Aided Chemical Engineering*. Amsterdam: Elsevier Science; 2005;20B:1651–1657.
14. Reichardt C, Harbusch-Görnert E. Erweiterung, Korrektur und Neudefinition der $E_T$-Lösungs-mittepolaritässkala mit Hilfe eines Lipophilen Penta-*tert*-butyl-substituierten Pyridinium-N-phenolat-betainfarbstoffes. *Liebigs Annalen der Chemie*. 1983;5:721–743.
15. Abraham MH, Doherty RM, Kamlet MJ, Harris JM, Taft RW. Linear solvation energy relationships. Part 38. An analysis of the use of solvent parameters in the correlation of the rate constants, with special reference to the solvolysis of t-Butyl chloride. *J of the Chem Soc - Perkin Trans II*. 1987;1097–1101.
16. Abraham MH, Grellier PL, Nasehzadeh A, Walker RAC. Substitution at saturated carbon. Part 26. A complete analysis of solvent effects on initial states for the solvolysis of the *t*-butyl halides in terms of *G*, *H* and *S* using the unified method. *J of the Chem Soc - Perkin Trans II*. 1988;1717–1724.
17. Kamlet MJ, Taft RW. The solvatochromic comparison method. I. The $\beta$-scale of solvent hydrogen-bond acceptor (HBA) basicities. *J of the A Chem Soc* 1976;98:377–383.
18. Kamlet MJ, Taft RW. The solvatochromic comparison method. 2. The $\alpha$-scale of solvent hydrogen-bond donor (HBD) acidities. *J of the A Chem Soc* 1976;98:2886–2894.
19. Kamlet MJ, Abboud MJ, Abraham MH, Taft RW. The solvatochromic comparison method. 6. The $\pi$*-scale of solvent polarities. *J of the A Chem Soc*. 1977;99:6027–6038.
20. Kamlet MJ, Abraham MH, Doherty RM, Taft RW. Solubility properties in polymers and biological media. 4. Correlation of octanol/water partition coefficients with solvatochromic parameters. *J of the A Chem Soc* 1984;106:464–466.
21. Taft RW, Abraham MH, Doherty RM, Kamlet MJ. Linear solvation energy relationships. 29. Solution properties of some tetraalkylammonium halide ion pairs and dissociated ions. *J of the A Chem Soc*. 1985;107:3105–3110.
22. Cativiela C, Garcia JI, Gil J, Martinez RM, Mayoral JA, Salvatella L, Urieta JS, Mainar AM, Abraham MH. Solvent effects on Diels-Alder reactions. The use of aqueous mixtures of fluorinated alcohols and the study of reactions of acrylonitrile. *J of the Chem Soc - Perkin Trans II*. 1997;653–660.
23. Abraham MH, Grellier PL, Hamerton I, McGill RA, Prior DV, Whiting GS. Solvation of gaseous non-electrolytes. *Faraday Discussions of the Chemical Society*. 1988;85:101–115.
24. Zissimos AM, Abraham MH, Du CM, Valko K, Bevan C, Reynolds D, Wood J, Tam KY. Calculation of Abraham descriptors from experimental data from seven HPLC systems; Evaluation of five different methods of calculation. *J of the Chem Soc - Perkin Trans II*. 2002;2001–2010.
25. Li J, Hawkins D, Cramer CJ, Truhlar DG. Universal reaction field model based on *ab initio* Hartree-Fock theory. *Chem Phys Letts*. 1998;288:293–298.
26. Gonçalves RMC, Simões AMN, Leitão RASE, Albuquerque LMPC. Correlation of rate constants for the solvolysis of *tert*-butyl halides, effect of temperature. *J of Chem Res (S)*. 1992;330–331.
27. Albuquerque LMPC, Moita MLCJ, Gonçalves RMC. Application of multiparameteric equations and factor analysis to the solvolytic reactions of *tert*-alkyl halides. *J of Phys Organic Chem*. 2001;14:139–145.
28. Dvorko GF, Zaliznyi VV, Ponomarev NE. Kinetics and mechanism of monomolecular heterolysis of commercial organohalogen com-

pounds: XXIX. Solvent effects on the activation parameters of heterolysis of *tert*-butyl chloride. *Russian J of Gen Chem* 2002;72:1414–1428.
29. Dvorko GF, Zaliznyi VV, Ponomarev NE. Kinetics and mechanism of monomolecular heterolysis of commercial organohalogen compounds: XXX. Correlation analysis of solvent effects in heterolysis of *tert*-butyl chloride. *Russian J of Gen Chem* 2002;72:1549–1555.
30. Sheldon TJ, Adjiman CS, Cordiner JL. Pure component properties from group contribution: Hydrogen-bond basicity, hydrogen-bond acidity, Hildebrand solubility parameter, macroscopic surface tension, dipole moment, refractive index and dielectric constant. *Fluid Phase Equilibria*. 2004;231:27–37.
31. Constantinou L, Gani R, O'Connell JP. Estimation of the accentric factor and the liquid molar volume at 298 K using a new group contribution method. *Fluid Phase Equilibria*. 1995;103:11–22.
32. Constantinou L, Gani R. New group contribution method for estimating properties of pure compounds. *AIChE J*. 1994;40:1697–1710.
33. Daubert TE, Danner RP. *Physical and Thermodynamic Properties of Pure Compounds: Data Compilation*. New York: Hemisphere; 2002.
34. Abraham MH. Scales of solute hydrogen-bonding: Their construction and application to physicochemical and biochemical processes. *Chem Soc Rev*. 1993;73–83.
35. Abraham MH. Hydrogen bonding. 31. Construction of a scale of solute effective or summation hydrogen-bond basicity. *J of Phys Org Chem*. 1993;6:660–684.
36. Abraham MH. Private Communication; 2004.
37. Numerical Algorithms Group. Statistical Add-Ins for Excel. http://www.nag.co.uk; 2000.
38. Maranas CD. Optimal computer aided molecular design: A polymer design case study. *Ind and En Chem Res* 1996;35:3788–3794.
39. Odele O, Macchietto S. Computer aided molecular design: A novel method for optimal solvent selection. *Fluid Phase Equilibria*. 1993;82:47–54.
40. Buxton A, Livingston AG, Pistikopoulos EN. Optimal design of solvent blends for environmental impact minimization. *AIChE J* 1999;45:817–843.
41. Marrero J, Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilibria*. 2001;183–184:183–208.
42. Brooke A, Kendrick D, Meeraus A, Raman R. GAMS A User's Guide. GAMS Development Corporation. http://www.gams.com, 1998.
43. Saltelli A, Chan K, Scott EM. (ed). *Sensitivity Analysis*. Chichester: John Wiley & Sons, 2000.
44. Sobol' IM. On the distribution of points in a cube and the approximate evaluation of integrals. *Comp Maths and Mathematical Phys*. 1967;7:86–112.
45. Sobol' IM. Uniformly distributed sequences with additional uniformity properties. *USSR Comp Maths and Mathematical Phys*. 1976;16:236–242.

## Appendix

### *Linearization of structure-property functions*

The correlation for $\delta_H^2$, given by Eq. 7, which contains the ratio of two linear expressions in $n_i$, can be linearized by expressing all the integer variables as linear combinations of binary variables (Eq. 46), and replacing the nonlinear products of continuous and binary variables with linear inequalities.[38]

We start by rearranging the expression to represent $\delta_H^2$ as follows

$$\delta_H^2 = \left[\frac{\sum_{i\in G}(n_i H_{V,i} + 6.829) - RT}{\sum_{i\in G} n_i V_{m,i} + 0.012}\right] \tag{A1}$$

$$\Leftrightarrow \delta_H^2\left(\sum_{i\in G} n_i V_{m,i} + 0.012\right) = \sum_{i\in G} n_i H_{V,i} + 6.829 - RT \tag{A2}$$

Introducing a new variable, $deln_i = \delta_H^2\, n_i$, this is equivalent to

$$\sum_{i\in G} deln_i V_{m,i} + \delta_H^2 0.012 = \sum_{i\in G} n_i H_{Vi} + 6.829 - RT \tag{A3}$$

$$\Leftrightarrow \begin{cases} \delta_H^2 = \left[\dfrac{\sum_{i\in G} n_i H_{Vi} + 6.829 - RT - \sum_{i\in G} deln_i V_{m,i}}{0.012}\right] \\ \sum_{i\in G} n_i H_{V,i} + 6.829 - RT \geq \delta_H^{2,L}\left(\sum_{i\in G} n_i V_{m,i} + 0.012\right) \\ \sum_{i\in G} n_i H_{V,i} + 6.829 - RT \leq \delta_H^{2,U}\left(\sum_{i\in G} n_i V_{m,i} + 0.012\right) \end{cases} \tag{A4}$$

where $\delta_H^{2,U}$ and $\delta_H^{2,L}$ are the upper and lower bound for the property $\delta_H^2$, respectively.

New variable $deln_i$ is the product of continuous variables $n_i$ and $\delta_H^2$. Since variable $n_i$ can be expressed as a linear combination of binary variables (Eq. 46), the following constraints can be written

$$deln_i = \sum_{j=0}^{K} 2^k ydel_{i,k}, \quad \forall i \in G \tag{A5}$$

$$\delta_H^2 - \delta_H^{2,U}(1 - y_{i,k}) \leq ydel_{i,k} \leq \delta_H^2 - \delta_H^{2,L}(1 - y_{i,k}), \quad i \in G, k \in K \tag{A6}$$

$$\delta_H^{2,L} y_{i,k} \leq ydel_{i,k} \leq \delta_H^{2,U} y_{i,k}, \quad i \in G, k \in K \tag{A7}$$

Finally, $\delta_H^2$ is in MPa, and it needs to be converted in cal cm$^{-3}$.

$$\delta_H^2(\text{cal cm}^{-3}) = 0.239\delta_H^2(\text{MPa}) \tag{A8}$$

*Manuscript received Jun. 11, 2006, and revision received Jan. 29, 2007.*