A QM-CAMD approach to solvent design for optimal reaction rates

Heiko Struebing, Stephan Obermeier, Eirini Siougkrou, Claire S. Adjiman and Amparo Galindo[1]

Centre for Process Systems Engineering, Department of Chemical Engineering
Imperial College London, London SW7 2AZ, U.K.

Abstract

The choice of solvent in which to carry out liquid-phase organic reactions often has a large impact on reaction rates and selectivity and is thus a key decision in process design. A systematic methodology for solvent design that does not require any experimental data on the effect of solvents on reaction kinetics is presented. It combines quantum mechanical computations for the reaction rate constant in various solvents with a computer-aided molecular design (CAMD) formulation. A surrogate model is used to derive an integrated design formulation that combines kinetics and other considerations such as phase equilibria, as predicted by group contribution methods. The derivation of the mixed-integer nonlinear formulation is presented step-by-step. In the application of the methodology to a classic S_N2 reaction, the Menschutkin reaction, the reaction rate is used as the key performance objective. The results highlight the trade-offs between different chemical and physical properties such as reaction rate constant, solvent density and solid reactant solubility and lead to the identification of several promising solvents to enhance reaction performance.

Keywords

Quantum mechanics, Computer-Aided Molecular Design, Surrogate model, Phase equilibrium, Kinetics, Group contribution

[1]Author to whom all correspondence should be addressed. Email: a.galindo@imperial.ac.uk

1. Introduction

Solvents are widely used in chemical processes and especially in the pharmaceutical and agrochemical industries, where they can be involved in all steps of production, such as reaction, separation and formulation. In fine chemicals manufacturing, a key function of the solvents used in reactive steps is often to dissolve solid reactants (Kolář et al., 2002). In polymerization technology, solvents are often used to control the reaction temperature and the viscosity of the reaction mixture (Whelan, 1994). More generally, the selectivity, yield and rate of organic reactions can be enormously influenced by the solvent used (Carlson et al., 1985; Reichardt and Welton, 2010). The choice of solvent can therefore be used as a design variable to improve reactor performance, in addition to other design considerations such as temperature or pressure. Beyond the optimization of a single processing step, it can be beneficial to consider the selection of reaction and separation solvents simultaneously (Elgue et al., 2004; Zhou et al., 2015b). From this perspective, it is thus desirable to develop systematic methods to solvent selection or design that can take into account, within a unified framework, the impact of the solvent on economic performance. A particular challenge in developing such a design platform is how to evaluate quantitatively the impact of a solvent on the reaction rate and how to integrate this often demanding computation with other elements of the design methodology.

The reaction rate is, in its simplest form, a function of the rate constant, reaction order and, under the assumption of an ideal solution, the concentrations of the reactants. These quantities can be significantly affected by solvent choice. Indeed, the rate constant can vary by several orders of magnitude from one solvent to another. For example, the rate constant for the solvolysis of 2-chloro-2-methylpropane is $350,000$ times larger in water than in ethanol and the rate constant for the reaction between trimethylamine and trimethylsulfonium ion is 119 times larger in nitromethane than in water (Maki et al., 2009; Reichardt and Welton, 2010). If at least one of the reactants is a solid at the reaction temperature, the maximum achievable reactant concentration and hence the maximum reaction rate are affected by the choice of solvent: in some cases the reactants are highly soluble, (e.g., calcium chloride in water) while in other cases the reactants are nearly insoluble (e.g., eicosane in water) (Prausnitz et al., 1999), thereby significantly reducing the reaction rate. Similarly partial miscibility may be observed between

a liquid reactant and a solvent, limiting the achievable reaction rate. It is therefore essential to take into account the reaction rate constant and concentration limits during solvent design.

At present, solvent selection is frequently based on experience and intuition or on costly experimental investigations (Folić et al., 2007). This heuristic approach restricts the development of processes with improved economic and environmental performance. In view of this, the American Chemical Society Green Chemistry Institute Roundtable has identified the development of solvent-selection techniques as a key priority area (Jiménez-González et al., 2011). Computer-aided molecular design (CAMD) is a promising avenue of research for the development of systematic methodologies, since CAMD approaches have been successfully applied to solvent selection and design problems for separation processes (Gani et al., 1991; Odele and Macchietto, 1993; Achenie et al., 2002; Giovanoglou et al., 2003; Adjiman et al., 2014; Ng et al., 2015; Zhou et al., 2015b).

An underpinning concept in CAMD is to exploit the fact that a large number of chemical species, as defined by their molecular structure, can be generated from a small set of structural building groups. The suitability of these structures for a particular task or process can be evaluated with respect to a chosen criterion (for instance, maximizing the rate of a given reaction), while considering physical and chemical constraints, as well as process constraints of varying complexity. The application of CAMD techniques to the design of optimal reaction solvents can in principle lead to the identification of rate-enhancing solvents from a large number of candidate compounds, and serve as a guide to an experimental investigation of suitable solvents. This computational approach can accelerate the solvent-selection process and reduce the cost of developing new reaction routes.

Building on these ideas, several CAMD approaches have been proposed in recent years, with a focus on reactions. Three groups have proposed methodologies based on a combination of experimental data and predictive models. A hybrid experimental/computer-aided approach has been put forward (Folić et al., 2004, 2007, 2008), in which a predictive model of the rate constant as a function of solvent structure is obtained based on a limited set of experimental kinetic data (rate constants in six to eight solvents) and integrated into a CAMD framework to identify promising solvents from a set of hundreds to thousands of potential molecular structures. This approach has been extended to the design of gas-expanded liquids (GXLs) (Siougkrou et al.,

2014), identifying the optimal solvent to be combined with carbon dioxide (CO_2) based on its impact on overall process performance, including kinetics, solubility and separation considerations. While this work on GXLs presents an extension of the methodology to consider a broader set of design criteria and to include solvent mixtures, its current application is restricted to a limited set of organic solvents due to the lack of models that relate the composition of the GXL to its properties. An alternative approach to solvent design for reactions, which combines knowledge from industrial practice and physical insights as well as property prediction techniques, has been developed (Gani et al., 2005, 2008). This technique requires a significant amount of experimental data on the solvent and the reaction of interest. It has the benefit of taking many important design considerations into account but solvent effects on the reaction rate constant are either neglected or treated empirically. Zhou et al. (2014) recently proposed a screening approach based on the use of COSMO-RS (Klamt et al., 2010) descriptors to identify promising solvents from a given list of molecules. This was recently extended (Zhou et al., 2015b) to enable the design of solvents from functional groups, rather than their selection from a list; the approach was applied to the design of solvents for a Diels-Alder reaction, by considering the implications of solvent choice on process design (Zhou et al., 2015b), and the presence of multiple reactions and the impact of uncertainty (Zhou et al., 2015a). The impact of the solvent on catalyst solubility was also incorporated within this framework (McBride et al., 2016). This overall methodology is based on the regression of quantitative-structure property relations (QSPRs) linking the solvent descriptors to the reaction rate constants. This step of the approach requires kinetic data on the reactions of interest in various solvents.

Further to these techniques that combine experimenetal data and predictive models, two approaches have been proposed to date that do not require any experimental kinetic data. Stanescu and Achenie (2006) proposed a method in which promising solvents are obtained by applying a CAMD approach which is based on physical property constraints only. Once this set of candidate molecules has been found, each solvent in the set is screened in terms of the reaction rate constant that can be achieved, as predicted by density functional theory (DFT). Recently, following on from the work of Folić et al. (2007, 2008) and of direct relevance to our current paper, a fully predictive approach, QM-CAMD, which integrates quantum mechanical (QM) calculations within the CAMD problem, has been proposed and demonstrated on a model

S_N2 reaction (Struebing et al., 2013), achieving an increase in rate constant of 40% over the initial solvents selected. The reaction rate constants in a few solvents, as predicted by QM and conventional transition state theory (CTST), were used to build a surrogate model, which was then used in a CAMD formulation to design the solvent that maximizes the rate constant. The surrogate model was continuously improved through re-parameterization with an increasing number of solvents, based on the solvents designed by CAMD, until convergence was achieved. The application of the approach, however, was limited to the maximization of the rate constant, with constraints on some physical properties of the solvent. Further progress is required to develop comprehensive approaches that build on the strengths of the different methods proposed to date and that can take into account the many effects of solvents on reaction performance.

The aim of our current work is to present a detailed account of a systematic CAMD approach, QM-CAMD, for the *ab initio* design of reaction solvents, where a solvent is generally defined as a compound in excess that does not undergo chemical transformations during the course of the reaction. The methodology introduced by Struebing et al. (2013), in which the optimal solvent is identified based on maximizing the *reaction rate constant*, is extended by considering the reaction *rate*, a quantity that is of greater relevance to process performance. This requires the maximum achievable concentration of the reactants, and consequently their solubility, to be predicted, in addition to the reaction rate constant. We focus here on the solubility of solid reactants. Group contribution methods (Fredenslund et al., 1975; Constantinou et al., 1995) are used to determine the maximum amount of solid reactant(s) that can be dissolved in the reactor for a given candidate solvent. As in Struebing et al. (2013), the rate constants are predicted using QM calculations for a small set of solvents, and these computational data are used to parameterize a simpler surrogate model that relates the rate constant to a few solvent properties. These are in turn related to the solvent molecular structure via group-contribution (GC) methods (Constantinou et al., 1995; Marrero and Gani, 2001; Sheldon et al., 2005; Folić et al., 2007). The underlying multiscale model thus combines electronic structure methods with bulk thermodynamic property prediction and reactor design.

In the next section, an overview of the proposed QM-CAMD methodology is presented. This is followed by a more detailed explanation of specific aspects of the method, namely the property prediction methods and the CAMD formulation. The methodology is then applied

to a Menschutkin reaction (Menschutkin, 1890a; Barnard and Smith, 1981), which is used to investigate the performance of the approach for several scenarios. A summary of the current status of the approach and perspectives for future development are provided at the end of the paper.

2. The QM-CAMD solvent design methodology

The QM-CAMD methodology for the design of solvents that enhance reaction kinetics is illustrated in Figure 1. Initially, a small set of kinetic data for various solvents is used to construct a surrogate model for the prediction of the reaction rate constant. The kinetic data are obtained by combining QM calculations, the SMD continuum solvation model of the solvent effects (Marenich et al., 2009), and group-contribution (GC) techniques. By solving a CAMD problem which includes the surrogate model for the rate constant prediction as well as other relevant constraints, a large space of possible solvents is explored. Thanks to the use of a surrogate model, only a small number of computationally-intensive QM calculations is required. The surrogate model is improved iteratively by adding QM-computed kinetic constants from previously-designed solvents to the regression data, thereby increasing the statistical significance and reliability of the model. This approach allows the identification of one or more promising solvents and therefore helps to reduce the experimental effort required for solvent selection. In the remainder of this section, each step of the approach is described in more detail. The approach differs from that of Struebing et al. (2013) in Steps 4, 6 and 7.

2.1. Step 1: Define the design problem and set of initial solvents

In the first step, the reaction(s) to be studied are specified. Design constraints and objectives are also identified and may include physical property and process considerations. In addition, a set of initial solvents is specified, preferably with diverse physical properties and functional groups. Six to ten solvents are typically used in the initial set. In the current work, these are chosen by the user, although more systematic approaches (Wicaksono et al., 2014) can in principle be used. If several reactions are being considered (e.g., parallel or series reactions), it is not necessary to use the same set of initial solvents for all reactions.

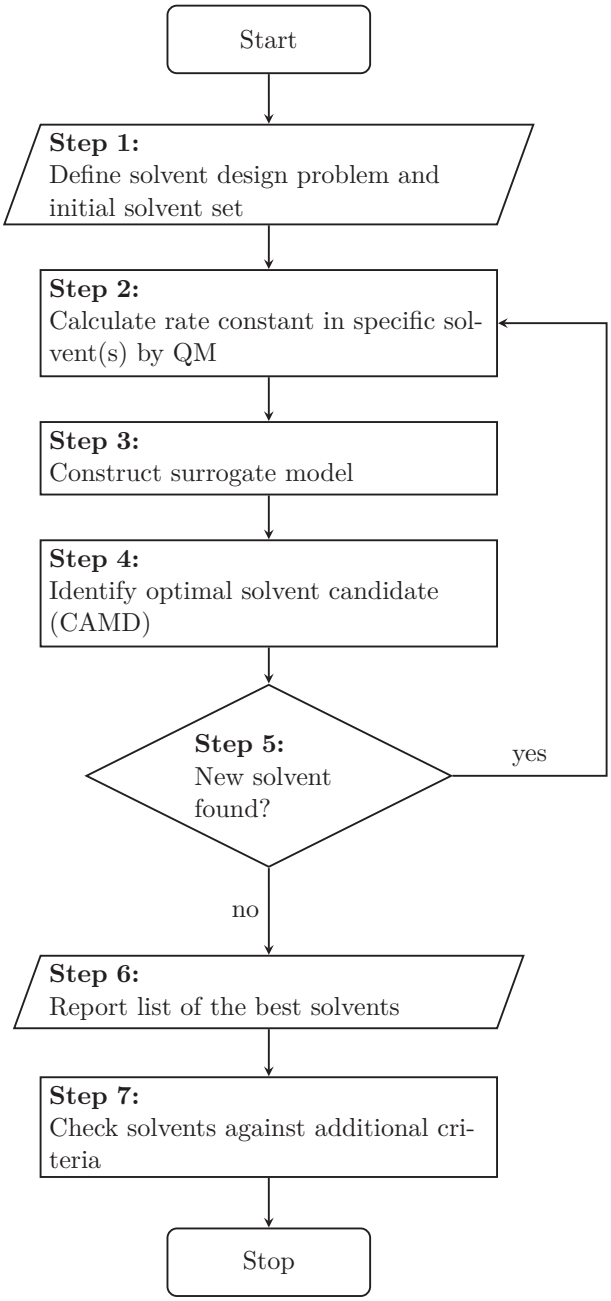

Figure 1: The QM-CAMD solvent design algorithm.

2.2. Step 2: Calculate the rate constant(s) for specific solvent(s) by QM

The aim of the second step is to obtain reliable values of the relevant rate constant(s) in the chosen solvents. In earlier work (Folić et al., 2007, 2008), such values were determined experimentally. In the QM-CAMD approach, quantum mechanical calculations are used. In the current implementation of the approach, the required liquid-phase reaction rate constants are computed based on CTST (Eyring, 1935; Evans and Polanyi, 1935). Under the common assumption that the activity coefficients of the reactants and transition states can be taken to be unity in all solvents, the choice of solvent affects the rate constant by changing the relative stability of the reactant(s) and transition state(s). Considering a reaction

$$\text{Reactants} \leftrightharpoons \text{TS} \rightarrow \text{Products},$$

where TS denotes the transition state, the rate constant k_j^{TST} in a solvent j is given by

$$k_j^{\mathrm{TST}} = \kappa \frac{k_B T}{h} \frac{1}{c^{\mathrm{o,L}}} \prod_{i=1}^{N_R+1} \left(q_i^{\mathrm{o,IG}}\right)^{\nu_i} \exp\left(\frac{-\Delta^{\ddagger} E^{\mathrm{el}}}{RT}\right) \exp\left(\frac{-\Delta^{\ddagger} \Delta G_j^{\mathrm{o,solv}}}{RT}\right), \tag{1}$$

where κ is the transmission coefficient; k_B is the Boltzmann constant; T is the reaction temperature in K; h is the Planck constant and $c^{\mathrm{o,L}}$ is the standard-state liquid-phase concentration ($1\,\mathrm{mol}\cdot\mathrm{dm}^{-3}$); $q_i^{\mathrm{o,IG}}$ denotes the ideal gas partition function for component i, where the index i runs over all N_R reactants and the transition state; ν_i is the stoichiometric coefficient for component i (a negative number for the reactants and +1 for the transition state); R denotes the ideal gas constant; E^{el} denotes the gas phase electronic energy; $\Delta G_j^{\mathrm{o,solv}}$ denotes the standard-state free energy of solvation in solvent j; and finally, $\Delta^{\ddagger}$ denotes the difference in energy between the transition state and the reactants, weighted by the stoichiometric coefficients:

$$\Delta^{\ddagger} E^{\mathrm{el}} = \sum_{i=1}^{N_R+1} \nu_i E_i^{\mathrm{el}}, \tag{2}$$

and

$$\Delta^{\ddagger} \Delta G_j^{\mathrm{o,solv}} = \sum_{i=1}^{N_R+1} \nu_i \Delta G_{j,i}^{\mathrm{o,solv}}. \tag{3}$$

Most of the relevant quantities can be derived from gas-phase quantum mechanical calculations. This includes the partition functions $q_i^{\mathrm{o,IG}}$, the electronic energies E_i^{el} and the transmission coefficient, which is computed here using the Wigner tunneling correction factor (Wigner, 1937),

an approach which is appropriate when the extent of tunnelling is small or when the system is at a high temperature. In order to compute the activation free energy of solvation, the SMD continuum solvation model (Marenich et al., 2009) as implemented in Gaussian 09 (Frisch et al.) is used. In this model, the solvent is treated as a continuum which affects the free energy of the solutes (reactants and transition state) through the presence of a potential field and through other effects such as dispersion, solvent re-arrangement to accommodate the solute, and specific interactions between solvent and solute. The effect of the potential field is quantified by computing its impact on the geometry and electrostatic potential of the solutes, while the other components are taken into account via a free energy correction based on an empirical expression that depends on a few solvent properties and the QM-derived solute geometry. The physical properties of the solvent j used in the SMD model are its Abraham solute hydrogen bond acidity A_j and basicity B_j (Abraham, 1993), the macroscopic surface tension at 298 K, γ_j, the refractive index at 298 K, $n_{D,j}$, the dielectric constant at 298 K, ϵ_j, the aromaticity, ϕ_j, and the halogenicity, ψ_j. The latter two properties are related to the fraction of aromatic atoms or halogen atoms, respectively, in the solvent. In order to obtain values for these properties for any solvent designed during the course of QM-CAMD, group contribution methods are used. These are presented in Section 3.3 in the case of A_j and B_j, and in the Supplementary Information for the other properties.

2.3. Step 3: Construct surrogate model

In Step 3 of the QM-CAMD algorithm, a surrogate model is built, based on the information from Step 2. In our current paper, we use the solvatochromic equation, a linear free energy relation (LFER), which was developed by Abraham et al. (1981, 1987a,b, 1988) for the prediction of the effects of solvents on various functions of free energy, e.g., the logarithm of kinetic rate constants, equilibrium constants, or free energies of solution. This linear relation offers a good balance in terms of computational cost and reliability and has been shown to be successful in correlating solvent effects. For instance, octanol/water partition coefficients (Kamlet et al., 1977), the kinetics of Diels-Alder reactions (Cativiela et al., 1997), and those of Menschutkin reactions (Folić et al., 2008) have been predicted successfully with this model. The general form

of the equation is given by

$$\log k_j^{\mathrm{CAMD}} = \boldsymbol{c}^T \boldsymbol{p}_j^1 \tag{4}$$

where k_j^{CAMD} is the rate constant in solvent j as calculated by this equation; $\boldsymbol{p}_j^1$ is a vector of size $N_p + 1$ with the first element equal to one and the remaining elements corresponding to N_p properties for solvent j; $\boldsymbol{c}$ is a vector of size $N_p + 1$ that consists of reaction-specific coefficients. The reaction-specific coefficients do not depend on solvent j and are obtained by linear regression to the set of quantum-mechanical rate constants k_j^{TST}, $j = 1, \ldots, N_s$, computed for N_s different solvents in Step 2. In order to estimate the reaction-specific coefficients, rate constant data in at least as many solvents as there are parameters in Equation (4) need to be used. Of course, the statistical significance of the coefficients, and hence the validity of the model, depends on the number of solvents used in the regression. Here, a small set initial solvents is used to limit the computational effort in the QM calculations. This set is progressively expanded as the algorithm progresses (cf. Step 5). Folić et al. (2007) and Wicaksono et al. (2014) showed that a solvatochromic equation regressed to a small set of solvents can be used to predict the rate constant well, as long as the chosen solvents provide sufficient coverage of the solvent design space (e.g., cover a wide range of polarities and represent diverse solvent classes, such as aromatics, nitrates, amides, alcohols, carboxylic acids and halogenated components). The required solvent properties are calculated using group contribution techniques (Folić et al., 2007; Marrero and Gani, 2001; Sheldon et al., 2005) that have been developed on the basis of extensive experimental data sets and are presented in Section 3.3.

2.4. Step 4: Obtain a candidate solvent by Computer-Aided Molecular Design

In the next step, the surrogate model for rate constant prediction obtained in Step 3 is integrated into an optimization-based molecular design problem. The CAMD formulation is extended from previous work (Folić et al., 2007; Struebing et al., 2013), with the addition of concentration through solubility, as computed with the UNIFAC activity coefficient model (Fredenslund et al., 1975), and molar volume, as computed with the GC method of Constantinou et al. (1995). This introduces a significant degree of non-linearity in the formulation, which becomes a mixed-integer non-linear programming (MINLP) problem. The objective of the CAMD problem considered here is to maximize the rate of a given reaction, based on the computed rate

constant and concentrations, and subject to a number of constraints, such as structure-property relations, chemical feasibility and molecular complexity, and design constraints. The detailed formulation of the CAMD problem is given in Section 3. The output of this step is a set of groups that defines the molecular structure of the optimal solvent, and the corresponding reaction rate. Because the GC methods used here do not distinguish between some isomers, the optimal set of groups may describe more than one solvent structure.

2.5. Step 5: Test for convergence

In Step 5, a convergence criterion is applied. Given that the surrogate model is initially based on kinetic data for a small set of solvents and that it is used for further extrapolation to a large set of solvents (typically thousands of molecules) in Step 4, the surrogate model has limited reliability at the beginning of the QM-CAMD algorithm. As a result, this step is used to determine whether the surrogate model should be further refined. If the solvent designed in Step 4 is already part of the solvent set that was used for regressing the surrogate model, no further calculations are needed and the algorithm proceeds to Step 6. If a new solvent was obtained in Step 4, the algorithm returns to Step 2, where the rate constant in this new solvent is calculated with the QM model and is added to the set of solvents used in Step 3 to regress the solvatochromic equation.

2.6. Step 6: Identify list of candidate solvents

In Step 6, once convergence has been achieved, the final formulation of the CAMD problem from Step 3 is used to generate a list of solvent candidates. This is achieved by solving the MINLP repeatedly, adding a new integer cut each time to eliminate the solvents already identified. In the case study considered in this paper, a list of the best five solvents is generated.

2.7. Step 7: Design validation

In Step 7, the list of candidate solvents is checked against criteria which were not considered in the CAMD formulation, such as the chemical stability and the reactivity of the solvent with the reactants. If one or more solvents are eliminated from consideration, appropriate integer cuts can be added to the CAMD problem in Step 6. Once the final check has been completed, a

selection of promising solvents is available, and kinetic experiments can be performed to validate the results.

3. The computer-aided molecular design (CAMD) problem

3.1. General formulation

The optimization problem that is solved in Step 4 of the QM-CAMD algorithm (Figure 1) is derived from the work of Folić et al. (2008). However, the introduction of solubility calculations and the change in the objective function to consider the reaction rate, rather than the rate constant, results in an MINLP problem. Furthermore, the set of building groups considered by Folić et al. (2008) is expanded by considering solvents that are represented by a single group, such as chloroform or acetonitrile. Although these solvents cannot be modelled via several groups, they are industrially important and should therefore be considered. This expansion of the design space thus makes the methodology more broadly applicable. The set G of all groups used in this work is given in the first column of Supplementary Information Table 1. The single-group molecule contributions reported in the table were derived from the experimental data sets of Lee (1996), Winget et al. (2010) and Lide (Lide, 2011). A further modification from Folić et al. (2008) is the implementation of integer cuts needed in Step 6 to generate a ranked list of possible solvent candidates. The general formulation of the MINLP problem can be written as follows:

$$
\begin{aligned}
\max_{\boldsymbol{p,n,y}} \quad & f(p) \\
\text{s.t.} \quad & g_1(p,n,y) \leq 0 \\
& g_2(n,y) \leq 0 \\
& g_3(n,y) \leq 0 \\
& d(p,n,y) \leq 0 \\
& p \in P \subset \mathbb{R}^m \\
& n \in N \subset \mathbb{R}^q \\
& y_i \in \{0,1\}^u \quad i = 1,\ldots,q,
\end{aligned}
\tag{5}
$$

where the objective function $f(p)$ is maximized with respect to the variables p, n and y, where p is an m-dimensional vector of continuous variables denoting continuous process variables, including physical properties, n is a q-dimensional vector of continuous variables denoting the number of groups of each type in the solvent molecule, y is a $q \times u$ matrix of binary variables used to represent the structure of the solvent molecules. Without loss of generality, equality constraints are subsumed within the set of inequalities. Furthermore, the constraints are partitioned into several subsets based on their physical interpretation: g_1 denotes a set of structure-property constraints and process or thermodynamic model equations, g_2 a set of chemical feasibility constraints, g_3 a set of molecular complexity equality constraints and d a set of design constraints related to physical properties and/or process performance. While the chemical feasibility constraints are essential to provide some assurance that only combinations of atom groups that can form a molecule are designed as solvent, the molecular complexity constraints are set by the user and can be treated as design constraints that limit the space of candidate solvents.

3.2. Objective function

In this work, the nonlinear objective function used is the rate of a given reaction, which is to be maximized. In the case of a bimolecular second-order reaction, $C_1 + C_2 \rightarrow P_1 + P_2$, the objective function is:

$$f(p) = k^{\mathrm{CAMD}}[C_1][C_2], \tag{6}$$

where $[C_1]$ and $[C_2]$ are the concentrations of reactant C_1 and reactant C_2, respectively, and k^{CAMD} denotes the rate constant obtained by applying the surrogate model, Equation (4). All three quantities in the objective function are implicit functions of the molecular structure of the solvent.

3.3. Structure-property and process constraints

The constraint set $g_1(p, n, y) \leq 0$ consists of all the structure-property relations, thermodynamic relations and process constraints required to evaluate the objective function and design constraints. Some, but not all, constraints depend explicitly on the solvent molecular structure. All relevant constraints are presented in this subsection for completeness.

3.3.1. Relating the rate constant to molecular structure

A form of the solvatochromic equation tailored to rate constant calculations is used; it includes five solvent properties ($N_p = 5$) and therefore six reaction-specific coefficients:

$$\log k^{\mathrm{CAMD}} = c_0 + c_A A + c_B B + c_S S + c_\delta \delta + c_H \delta_H^2, \tag{7}$$

where A, B, S, δ and δ_H^2 are the chosen solvent properties and are independent of the reaction. A, B and S are the so-called Abraham descriptors (hydrogen bond acidity, hydrogen bond basicity and polarizability/dipolarity respectively); δ denotes the polarizability correction term and accounts for the greater variation of the solvent polarizability between different molecular "classes" (e.g., aromatic versus aliphatic); δ_H^2 is the cohesive energy density.

Each property is related to the molecular structure of the solvent via a group contribution method. The hydrogen bond acidity, A_j, is obtained based on a method developed by Sheldon et al. (2005), using the revised coefficients proposed by Folić et al. (2007). It is given by:

$$A = \begin{cases} 0.010641 + \sum\limits_{i \in G} n_i \cdot A_i & \text{if} \quad y_A = 1, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where A_i is the acidity contribution of group i, $i \in G$, the set of all functional groups (cf. Supplementary Information Table 1 for the group contribution values), and the binary variable y_A is such that:

$$y_A = \begin{cases} 1 & \text{if} \quad 0.010641 + \sum\limits_{i \in G} n_i \cdot A_i \geq 0.029, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Equivalently, one can obtain the following algebraic expressions:

$$0 \leq A \leq M \cdot y_A, \tag{10}$$

$$-A + \sum_{i \in G} n_i \cdot A_i + y_A - 0.989359 \leq 0, \tag{11}$$

$$A - \sum_{i \in G} n_i \cdot A_i - 0.010641 \leq 0, \tag{12}$$

$$\sum_{i \in G} n_i \cdot A_i - M \cdot y_A - 0.018359 \leq 0, \tag{13}$$

$$M \cdot (y_A - 1) - \sum_{i \in G} n_i \cdot A_i + 0.018359 \leq 0, \tag{14}$$

where M is a sufficiently large positive number (in this work $M = 100$). The hydrogen bond basicity, B, is given by a similar relation (Sheldon et al., 2005; Folić et al., 2007):

$$B = \begin{cases} 0.12371 + \sum\limits_{i \in G} n_i \cdot B_i & \text{if} \quad y_B = 1, \\ 0 & \text{otherwise}, \end{cases} \tag{15}$$

where B_i denotes the basicity contribution of group i (cf. Supplementary Information Table 1) and y_B denotes a binary variable defined as:

$$y_B = \begin{cases} 1 \quad \text{if} \quad 0.12371 + \sum\limits_{i \in G} n_i \cdot B_i \geq 0.124, \\ 0 \quad \text{otherwise}. \end{cases} \tag{16}$$

Recasting the logic relations into algebraic functions yields the following equations:

$$0 \leq B \leq M \cdot y_B, \tag{17}$$

$$-B + \sum_{i \in G} n_i \cdot B_i + y_B - 0.87629 \leq 0, \tag{18}$$

$$B - \sum_{i \in G} n_i \cdot B_i - 0.12371 \leq 0, \tag{19}$$

$$\sum_{i \in G} n_i \cdot B_i - M \cdot y_B - 0.00029 \leq 0, \tag{20}$$

$$M \cdot (y_B - 1) - \sum_{i \in G} n_i \cdot B_i + 0.00029 \leq 0, \tag{21}$$

For the polarizability/dipolarity S, the group contribution method of Folić et al. (2007) is used:

$$S = 0.325675 + \sum_{i \in G} n_i \cdot S_i, \tag{22}$$

where S_i is the polarizability/dipolarity contribution of group i in G (cf. Supplementary Information Table 1).

The correction parameter for polarizability/dipolarity, δ, was introduced by Kamlet et al. (1977) and is used to identify differences in the polarizability of three molecular classes: non-halogenated aliphatics, polyhalogenated aliphatics and aromatics. It is defined as:

$$\delta = \begin{cases} 1 & \text{if the molecule is aromatic,} \\ 0.5 & \text{if the molecule is halogenated and aliphatic,} \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Although Kamlet et al. (1977) define a class of polyhalogenated aliphatic molecules for which $\delta = 0.5$, they appear to consider aliphatic solvents with a single halogen atom in this class. We adopt this approach here. Two binary variables are required to express δ in terms of algebraic equations. The first binary variable, y_1, indicates whether aromatic groups are present in the molecule:

$$y_1 = \begin{cases} 1 & \text{if } \sum\limits_{i \in G_{\mathrm{Ar}}} n_i \geq 1, \\ 0 & \text{otherwise,} \end{cases} \tag{24}$$

where G_{Ar} is the subset of aromatic groups ($G_{\mathrm{Ar}} \subset G$, cf. Supplementary Information Table 2. The second binary variable, y_2, indicates whether the molecule is halogenated and not aromatic:

$$y_2 = \begin{cases} 1 & \text{if } \sum\limits_{i \in G_{\mathrm{H}}} n_i \geq 1 \text{ and } y_1 = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

where G_{H} is the subset of non-aromatic halogen-containing groups ($G_{\mathrm{H}} \subset G$) (cf. Supplementary Information Table 2). This is described algebraically via Equations (26) to (30):

$$\sum_{i \in G_{\mathrm{Ar}}} n_i \leq \sum_{i \in G_{\mathrm{Ar}}} n_i^U \cdot y_1, \tag{26}$$

$$y_1 \leq \sum_{i \in G_{\mathrm{Ar}}} n_i, \tag{27}$$

$$y_2 \leq \sum_{i \in G_{\mathrm{H}}} n_i, \tag{28}$$

$$\sum_{i \in G_{\mathrm{H}}} n_i \leq (y_2 + y_1) \cdot \sum_{i \in G_{\mathrm{H}}} n_i^U, \tag{29}$$

$$y_1 + y_2 \leq 1, \tag{30}$$

where the superscript U denotes the upper bound on the variable. The last equation ensures that y_1 and y_2 cannot both be equal to 1. In the case of an aromatic solvent with a halogenated branch, the molecule is classified as aromatic and the maximum correction, $\delta = 1$, applied. The correction parameter δ is given by:

$$\delta = y_1 + 0.5 y_2. \tag{31}$$

The final solvent property required in the solvatochromic equation is the cohesive energy density δ_H^2, which is defined as the square of the Hildebrand solubility parameter δ_H and is a function of the liquid molar volume and enthalpy of vaporization of the solvent:

$$\delta_H^2 \left[\frac{\text{cal}}{\text{cm}^3}\right] = 0.238846 \cdot \frac{\Delta H_V - RT \cdot 10^{-3}}{V_m}, \tag{32}$$

where the correlation is multiplied with 0.238846 to convert the units of δ_H^2 from MPa to cal·cm^{-3}, R denotes the ideal gas constant in J·mol^{-1}·K^{-1} and T denotes the reaction temperature in K. The enthalpy of vaporization ΔH_V at a temperature of 298 K is determined by using the first-order group contribution technique proposed by Marrero and Gani (2001):

$$\Delta H_V \left[\frac{\text{kJ}}{\text{mol}}\right] = \sum_{i \in G} n_i \cdot H_{V,i} + 11.733 \left[\frac{\text{kJ}}{\text{mol}}\right], \tag{33}$$

where $H_{V,i}$ is the contribution of group i to the enthalpy of vaporization (cf. Supplementary Information Table 1). The liquid molar volume of the solvent is predicted using the group contribution approach proposed by Constantinou et al. (1995):

$$V_m \left[\frac{\text{m}^3}{\text{kmol}}\right] = \sum_{i \in G} n_i \cdot V_{m,i} + 0.01211 \left[\frac{\text{m}^3}{\text{kmol}}\right], \tag{34}$$

where $V_{m,i}$ is the liquid molar volume contribution of group i (cf. Supplementary Information Table 1). The nonlinear expression (32) is linearized (Maranas, 1996; Folić et al., 2007) by making use of the discrete nature of the n_i variables: these are expressed as linear combinations of binary variables so that the nonlinear products of continuous and binary variables can be replaced with linear inequalities.

3.3.2. Relating the reactant concentrations to the molecular structure of the solvent

The reactant concentrations are needed to calculate the reaction rate, in addition to the reaction rate constant. For a batch or a plug flow reactor, the maximum concentrations of the reactants, and hence reaction rate, that can be achieved correspond to the initial or inlet concentrations. The initial or inlet mixture can be assumed to consist of the reactants and solvent only, and thus to be free of products. The maximum concentrations depend on the mole fractions of the reactants and the molar volume of the mixture:

$$[i] = \frac{x_i}{V_{m,M}}, \qquad \forall i \in \mathrm{R}, \tag{35}$$

where $[i]$ denotes the molar concentration of component i, $V_{m,M}$ is the liquid molar volume of the mixture, and R is the set of reactants. In order to obtain the molar volume of the mixture, ideal behaviour is assumed so that a linear combination of the pure component molar volumes can be used:

$$V_{m,M} = \sum_{i \in \mathrm{C}} x_i V_{m,i}, \tag{36}$$

where C denotes the set of components in the mixture (the reactants and the solvent) and $V_{m,i}$ is given by Equation (34) (Constantinou et al., 1995). The error introduced with the ideal mixture assumption could in principle be avoided by using a group-contribution equation of state that provides a reliable prediction of the liquid phase densities (e.g., see Lymperiadis et al. (2007) and Papaioannou et al. (2014)). However, the group contribution parameter tables for such equations (Lymperiadis et al., 2008; Dufal et al., 2014) do not yet allow the prediction of the properties of all compounds considered in the case study of interest here.

The solubility of some of the reactants may limit the maximum achievable concentrations and should therefore be taken into account in calculating the mole fractions of reactant that can be present in the reaction medium. Strictly speaking, although the solvent is typically a component in excess, the reaction medium consists of all components in the mixture, including any reactants and products present. The multiple components in the reaction mixture are thus taken into account when calculating reactant solubilities using the UNIFAC model. In modelling reaction rate constants, however, it is assumed that the reaction medium is the solvent. This is in keeping with the standard assumption that rate constants are independent of composition.

It also reflects the fact that current models are not suited to a mixed reaction medium because the solvent properties that describe the reaction medium in the SMD model (cf. Step 2 in Section 2.2) and in the surrogate model (Equation (76)) are available only for a very limited set of binary mixtures (e.g., see Ràfols et al. (1995); Siougkrou et al. (2014)). In view of this, we ensure that the reaction medium is mostly solvent in the design problem formulation by limiting the reactant mole fractions so that the mole fraction of solvent is at least twice that of the most abundant reactant. This limit on the reactant concentrations is consistent with our objective to design a solvent-based reaction system. In our proposed formulation, solubility is thus treated as a constraint on the maximum achievable mole fraction only when the thermodynamic solubility of a given reactant in the mixture is lower than a threshold value, $x_{\max,\text{eq}}$. If it is greater than this threshold, solubility is not a limiting factor on the reaction rate and the reactant mole fraction is set to the threshold value.

To embed this in the optimization problem formulation, two mixtures and their associated mole fractions are considered in the model: the *equilibrium mixture*, for the calculation of phase equilibrium and hence solubility of component i, $x_{i,\text{eq}}$, and the *reactor mixture*, for the determination of the actual mole fraction of component i in the reactor, x_i, where

$$\sum_{i \in \text{C}} x_{i,\text{eq}} = 1, \tag{37}$$

$$\sum_{i \in \text{C}} x_i = 1. \tag{38}$$

For simplicity, we consider a ternary system involving components C_1, C_2 and C_3, where C_1 and C_2 are reactants and C_3 is the solvent. The model is developed for the case when only the solubility of C_2, a reactant which is solid at reaction conditions, may be a limiting factor but it can readily be extended to other cases. A relationship between some of the properties of the pure solid C_2 and its solubility in the solvent can be obtained from a thermodynamic analysis (Poling et al., 2000). To a good approximation, by assuming that the triple-point temperature is equal to the melting point temperature and the effect of the change in the molar heat capacity on solubility can be neglected (Poling et al., 2000), the solubility of C_2 is given by:

$$\ln\left(\gamma_{C_2} \cdot x_{C_2,\text{eq}}\right) = \frac{-\Delta H_{m,C_2}}{RT}\left(1 - \frac{T}{T_{m,C_2}}\right), \tag{39}$$

where γ_{C_2} is the activity coefficient of reactant C_2 in the liquid phase, at the reactor temperature, pressure and composition, $x_{C_2,\text{eq}}$ denotes the mole fraction of reactant C_2 at solid-liquid equilibrium (its solubility), T_{m,C_2} is its normal melting temperature, $\Delta H_{m,C_2}$ is its enthalpy of fusion, assumed to be constant, and T is the reaction temperature. For a given temperature, the right-hand side of Equation (39) is constant. The UNIFAC method (Fredenslund et al., 1975) is used to compute the activity coefficient of reactant C_2 in a ternary mixture of C_1, C_2 and C_3 as a function of the molecular structure of the components in the mixture. The full set of UNIFAC equations can be found in Poling et al. (2000). A user-specified value, $x^*_{C_1}$, is used for the mole fraction of reactant C_1 in the equilibrium mixture, $x_{C_1,\text{eq}}$ as well as in the reactor mixture, x_{C_1}:

$$x_{C_1,\text{eq}} = x_{C_1} = x^*_{C_1}, \tag{40}$$

The mole fraction, x_{C_2}, of C_2 in the reactor is defined as:

$$x_{C_2} = \begin{cases} x_{C_2,\text{eq}} & \text{if} \quad x_{C_2,\text{eq}} \leq x_{\text{max,eq}}, \\ x_{\text{max,eq}} & \text{otherwise.} \end{cases} \tag{41}$$

To represent this choice algebraically, the binary variable y_{eq} is introduced and defined as:

$$y_{\text{eq}} = \begin{cases} 0 & \text{if} \quad x_{C_2,\text{eq}} \leq x_{\text{max,eq}}, \\ 1 & \text{otherwise.} \end{cases} \tag{42}$$

The value of the variable y_{eq} is then given by the following set of equations:

$$x_{C_2,\text{eq}} - y_{\text{eq}} - x_{\text{max,eq}} \leq 0, \tag{43}$$

and

$$-x_{C_2,\text{eq}} + (y_{\text{eq}} - 1) + x_{\text{max,eq}} \leq 0, \tag{44}$$

Finally, the value of x_{C_2} is given by

$$x_{C_2} = x_{\text{max,eq}} \cdot y_{\text{eq}} + x_{C_2,\text{eq}} \cdot (1 - y_{\text{eq}}) \,. \tag{45}$$

This last equation can readily be re-formulated to eliminate the bilinear term $x_{C_2,\text{eq}} \cdot y_{\text{eq}}$.

3.4. Chemical feasibility constraints

The chemical feasibility constraints, $g_2(n, y) \leq 0$, ensure that only chemically-meaningful combinations of the atom groups are put forward as candidate solvents.

3.4.1. Constraining variable n_i to an integer

The number of groups of type i appearing in the designed molecule is defined by a continuous variable n_i, but only integer values of n_i are meaningful. Therefore, a binary representation is introduced to convert the continuous variable n_i to a pseudo-integer variable:

$$\sum_{k=1}^{K} 2^{k-1} y_{i,k} - n_i = 0, \quad \forall i \in G \tag{46}$$

where $y_{i,k}$, $k = 1, \ldots, K$, constitutes a set of binary variables defining the occurrence of group i. The value of K is chosen based on the desired maximum number of groups to appear in the molecule.

3.4.2. Cyclic and acyclic structures

In our formulation, three types of molecular structures are allowed as solvent candidates; these are acyclic, bicyclic and monocyclic structures. Here, a compound is deemed to be acyclic if the combination of its groups can form an acyclic graph. If a group contains a cycle (e.g., if one were to include a group representing cyclopentyl in the design space) but is included in a molecule which contains no cycles involving multiple groups, the compound is considered to be acyclic. Furthermore, only aromatic cycles are considered in the design of bicyclic and monocyclic structures – alicyclic compounds built from the aliphatic groups are not permitted. The binary variables y_a, y_b and y_m are used to define the molecular structure being designed, with $y_a = 1$ indicating an acyclic compound, $y_m = 1$ a monocyclic compound and $y_b = 1$ a bicyclic compound. Since only one type of structure can be designed, the following constraint is included:

$$y_a + y_m + y_b = 1. \tag{47}$$

A continuous variable m that can represent all three types of structures is needed to define further constraints, such as the octet rule of Odele and Macchietto (1993). The continuous

variable m is defined as:

$$m = \begin{cases} 1 & \text{for an acyclic molecule,} \\ 0 & \text{for a monocyclic molecule,} \\ -1 & \text{for a bicyclic molecule.} \end{cases} \tag{48}$$

The variable m can be described as a function of the binary variables y_a and y_b:

$$m - (y_a - y_b) = 0. \tag{49}$$

3.4.3. Aromatic molecules

As mentioned previously, only aromatic compounds are considered when designing cyclic molecules. Hence, the cyclic backbone must consist of aromatic groups, as defined by the set G_{Ar} shown in Supplementary Information Table 2. The knowledge that exactly six aromatic groups are necessary to create a monocyclic molecule and ten aromatic groups are necessary to create a bicyclic molecule leads to the following constraint:

$$\sum_{i \in G_{\mathrm{Ar}}} n_i - 6y_m - 10y_b = 0, \tag{50}$$

where n_i is the number of groups i appearing in the solvent.

3.4.4. Octet rule and molecular groups

Odele and Macchietto (1993) proposed the octet rule to ensure that the designed molecule does not have any free bonds:

$$\sum_{i \in G} (2 - v_i)\, n_i - 2m = 0, \tag{51}$$

where v_i is the valency of group i (the values of valency are listed in Supplementary Information Table 1).

The formulation used in previous works (e.g., Folić et al. (2007, 2008)) is extended so that solvents that are represented by a single (molecular) group are allowed to appear among acyclic structures. Such molecules are typically too small for their properties to be calculated accurately by GC methods. Molecular groups have a valency of 0 and therefore do not contribute to the octet rule. Without further constraints, it would thus be possible for a molecular group to appear alongside a combination of groups forming a chemically-feasible molecule, while satisfying

Equation (51). The following constraints prevent such an occurrence by ensuring that when a molecular group is selected, no other groups are included in the design:

$$\sum_{i \in G_1} n_i \leq 1, \tag{52}$$

$$\sum_{i \in G} n_i - \sum_{j \in G_1} n_j \leq \left(1 - \sum_{j \in G_1} n_j\right) n_{\mathrm{G,max}}, \tag{53}$$

where i denotes a structural group in the solvent set G and j denotes a molecular group in set $G_1 \subset G$ (cf. Supplementary Information Table 2).

3.4.5. Modified bonding rule

In GC techniques, bonding between any two given groups can only take place using a single covalent bond because atom groups that contain double or triple bonds are defined as separate groups, as in the CH_2=C group, for example. A group with a valency of two must be bonded to two distinct groups, which may or may not differ in terms of their chemical identities. Thus, a CH_2 group could be bonded to two CH_3 groups or to a CH_3 group and an OH group. Its two free bonds should not, however, be bonded to same CH_2, despite the fact that such an arrangement is feasible under the octet rule. A constraint is therefore included to ensure that adjacent groups can only be connected by a single covalent bond. A modified bonding rule which is based on the bonding rule of Odele and Macchietto (1993), but provides an extension to account for molecular groups, is proposed here:

$$n_j \left(v_j - 1\right) + 2\left(m - \sum_{i \in G_1} n_i\right) - \sum_{i \in G} n_i \leq 0, \quad \forall j \in G. \tag{54}$$

3.5. Chemical complexity constraints

The size and functionality of the molecules designed can be restricted by the user in order to eliminate combinations of groups that are unlikely to form good solvent molecules, or to decrease the size of the design space. This is achieved by defining a set of chemical complexity constraints, $g_3\left(n, y\right) \leq 0$. While it is essential to place a limit on the size of the molecules designed (e.g., in the form of a limit on the number of atom groups in the candidate solvents), the inclusion of other constraints in the design problem is optional. All constraints used in the case study discussed in our current work are presented here for completeness.

3.5.1. Molecular size

First, the maximum size of the solvent to be designed is defined. A lower bound on the number of groups $n_{\mathrm{G,min}}$ and an upper bound on the number of groups $n_{\mathrm{G,max}}$ are imposed by the following constraints:

$$n_{\mathrm{G,min}} - \sum_{i \in G} n_i \leq 0, \tag{55}$$

$$\sum_{i \in G} n_i \leq n_{\mathrm{G,max}}. \tag{56}$$

3.5.2. Number of groups of each type

The number of occurrences of each group type i is restricted by a specific upper bound n_i^U. A tighter upper bound tailored to each group type i can be derived for each n_i variable from the structures in which the group can take part (e.g., cyclic or not) and can be related to the corresponding binary variables (y_a, y_b and y_m). The upper bounds on n_i are given by the general constraint:

$$n_i - f_i(y_a, y_b, y_m) \leq 0, \quad \forall i \in G, \tag{57}$$

where the linear functions $f_i(y_a, y_b, y_m)$ are defined in Supplementary Information Table 3.

In addition to the upper bound constraint on each group type i, the number of groups from a subset of the group types can also be limited by an upper bound. Thus, the number of groups drawn from the subset of main groups G_{M} (cf. Supplementary Information Table 2) is restricted by

$$\sum_{i \in G_{\mathrm{M}}} n_i \leq n_{\mathrm{G},a}^U \cdot y_a + n_{\mathrm{G},b}^U \cdot y_b + n_{\mathrm{G},m}^U \cdot y_m, \tag{58}$$

where $n_{\mathrm{G},a}^U$, $n_{\mathrm{G},b}^U$, and $n_{\mathrm{G},m}^U$ denote the maximum number of main groups in acyclic, monocyclic and bicyclic molecules, respectively.

A further subset of G on which an upper bound constraint is imposed is the set of functional groups G_{F} (cf. Supplementary Information Table 2). This is done to reflect the limited accuracy of many first-order GC methods in predicting the properties of multifunctional molecules. The number of functional groups which appear in the candidate solvents is restricted by imposing a constant upper bound, $n_{i,\mathrm{F}}^U$, on each group type i, and by applying an overall constraint

that ensures that there are no groups from G_{F} in bicyclic molecules, and a limited number in monocyclic and acyclic molecules:

$$\sum_{i \in G_{\mathrm{F}}} \frac{n_i}{n_{i,\mathrm{F}}^U} \leq y_a + y_m, \tag{59}$$

where $n_{i,\mathrm{F}}^U$ is an upper bound on the number of occurrences of functional group i (cf., Supplementary Information Table 4).

Finally, to help ensure the chemical stability of the molecule designed, the number of carbon-carbon double bonds is restricted so that at most one such bond appears in the designed molecule:

$$n_{\mathrm{CH_2{=}CH}} + n_{\mathrm{CH{=}CH}} + n_{\mathrm{CH_2{=}C}} + n_{\mathrm{CH{=}C}} + n_{\mathrm{C{=}C}} \leq 1. \tag{60}$$

3.5.3. Branched molecules

When a monocyclic molecular structure is designed, side chains can only occur if one branched aromatic group such as aC, aCCH or $\mathrm{aCCH_2}$ is included in the combination of groups that make up the molecule. Three binary variables are introduced to define further constraints on the size and the functionality of molecules that contain at least one of these groups. Three binary variables are defined to indicate the presence of each of these groups:

$$y_i = \begin{cases} 1 & \text{if an aromatic group of type } i \text{ occurs in the solvent molecule,} \\ 0 & \text{otherwise.} \end{cases} \tag{61}$$

$$i \in \{\mathrm{aC}, \mathrm{aCCH}, \mathrm{aCCH_2}\}$$

The values of these binary variables are obtained by imposing two constraints per variable, in a generalization of the formulation proposed by Folić et al. (2008):

$$n_i - 0.9 - n_i^U \cdot y_i \leq 0, \qquad i \in \{\mathrm{aC}, \mathrm{aCCH}, \mathrm{aCCH2}\}, \tag{62}$$

$$y_i - n_i \leq 0, \qquad i \in \{\mathrm{aC}, \mathrm{aCCH}, \mathrm{aCCH2}\}. \tag{63}$$

A further binary variable, $y_{\mathrm{m,aC}}$, is introduced to identify whether an aC group appears in a monocyclic compound. Its value is defined as:

$$y_{\mathrm{m,aC}} = \begin{cases} 1 & \text{if } y_{\mathrm{aC}} + y_m = 2, \\ 0 & \text{otherwise.} \end{cases} \tag{64}$$

Equivalently,

$$y_m + y_{\mathrm{aC}} - 1 - y_{\mathrm{m,aC}} \leq 0, \tag{65}$$

$$2 \cdot y_{\mathrm{m,aC}} - y_m - y_{\mathrm{aC}} \leq 0. \tag{66}$$

In our current work, the aC group is constrained to appear at most once in monocyclic and twice in bicyclic molecules as follows:

$$2y_b + y_{\mathrm{m,aC}} - n_{\mathrm{aC}} = 0. \tag{67}$$

The complexity of the solvent molecules is further constrained by allowing at most one of aC, aCCH and $\mathrm{aCCH_2}$ in a monocyclic molecule, thereby limiting the maximum number of side chains to one. This condition is implemented by specifying:

$$y_{m,aC} + y_{\mathrm{aCCH}} + y_{\mathrm{aCCH_2}} \leq 1. \tag{68}$$

The side chains of monocyclic molecules consist of chain-ending groups G_{CE} and of non-chain-ending groups G_{NCE} (cf. Supplementary Information Table 2), where chain-ending groups G_{CE} can be attached directly to the aromatic backbone of the molecule or to a non-chain-ending group G_{NCE}. An aCCH group leads to two branches that normally consist of both chain-ending and non-chain-ending groups. To reduce the complexity of the designed molecules one of these branches is restricted to be a $\mathrm{CH_3}$ group by:

$$y_{\mathrm{aCCH}} \leq n_{\mathrm{CH_3}}. \tag{69}$$

Chain-ending groups can only appear up to three times in an acyclic molecule and once in an aromatic molecule. This is described by:

$$\sum_{i \in G_{\mathrm{CE}}} n_i \leq 3y_a + y_{m,aC} + y_{\mathrm{aCCH}} + y_{\mathrm{aCCH_2}}. \tag{70}$$

Non-chain-ending groups are constrained to occur at most three times in acyclic molecules and once in aromatic molecules by the following equation:

$$\sum_{i \in G_{\mathrm{NCE}}} n_i \leq 3y_a + y_{m,aC} + y_{\mathrm{aCCH_2}}. \tag{71}$$

We note that more complex molecules can easily be generated by removing some of chemical complexity constraints or by increasing the limits on the number of groups. The chemical complexity constraints provide a high degree of flexibility for the user, making it possible to increase or decrease the number of molecules in the design space based on preferences or heuristics.

3.6. Design constraints

The set of design constraints $d(p, n, y) \leq 0$ is used to impose restrictions on the design space, for instance based on the physical properties of the candidate solvent, its health and safety performance, and other aspects relevant to the overall performance of the process (Gani et al., 2005). Since the solvent must be in the liquid phase at reaction conditions, the normal melting point of the solvent, T_m, and its boiling point, T_b, are constrained:

$$T_m \leq T_{m,max},$$
$$T_{b,min} \leq T_b.$$

where $T_{m,max}$ is the user-specified upper bound on the melting point and $T_{b,min}$ is the user-specified lower bound on the boiling point. The first-order GC method proposed by Marrero and Gani (2001) can be used to obtain a linear inequality based on the dimensionless equivalent melting point $T_{m,e} = \exp(T_m/T_{m,0})$, where $T_{m,0} = 147.450\,\mathrm{K}$ is the reference value used in the GC method of Marrero and Gani (2001):

$$T_{m,e} = \sum_{i \in G} n_i T_{m,i} \leq \exp(T_{m,max}/T_{m,0}). \tag{72}$$

An analogous method is used to constrain the dimensionless equivalent boiling point $T_{b,e} = \exp(T_b/T_{b,0})$, where $T_{b,0} = 222.543\,\mathrm{K}$ is the reference value used in the GC method of Marrero and Gani (2001):

$$T_{b,e} = \sum_{i \in G} n_i T_{b,i} \geq \exp(T_{b,min}/T_{b,0}). \tag{73}$$

The contributions $T_{m,i}$ and $T_{b,i}$ are presented in Supplementary Information Table 1.

4. Case study - A Menschutkin reaction

4.1. Problem specification

Menschutkin reactions belong to the important class of nucleophilic substitution reactions, specifically S_N2 reactions, and occur when a tertiary amine reacts with a halogenated compound to form a quaternary ammonium salt. Menschutkin reactions have been shown to exhibit significant solvent effects (Menschutkin, 1890a,b; Truong et al., 1997; Amovilli et al., 1998; Castejon and Wiberg, 1999; Struebing et al., 2013). The Menschutkin reaction of pyridine (C_1)

and phenacyl bromide (C_2), which is illustrated in Figure 2, is chosen to demonstrate the proposed design methodology. It has been studied both experimentally (Barnard and Smith, 1981; Ganase, 2015) and computationally (Struebing et al., 2013); given the breadth of information available on this reaction, it provides an ideal test case for methodological developments. Here, the reaction is investigated at standard pressure (1 atm) and at a reaction temperature of 298 K. At these conditions, pyridine (reactant C_1) is a liquid but phenacyl bromide (reactant C_2) is a solid.

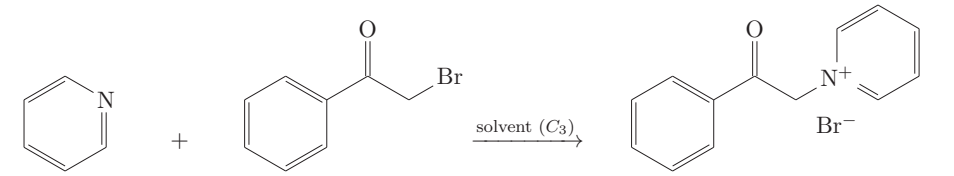

Figure 2: The chosen Menschutkin reaction: pyridine and phenacyl bromide react to form phenacyl pyridinium bromide salt (Barnard and Smith, 1981).

The rate of Menschutkin reactions has been shown to follow second-order kinetics (Pearson et al., 1952; Halvorsen and Songstad, 1978; Forster and Laird, 1982; Hwang et al., 1983; Ganase, 2015). Hence, it is defined by the following equation:

$$r = k \cdot [C_1] \cdot [C_2]. \tag{74}$$

In the formulation of the CAMD problem in Step 4 of the QM-CAMD approach (Figure 1), the rate constant is calculated using the solvatochromic equation, Equation (7). The QM calculations required to parameterize the solvatochromic equation are performed using the M05-2X/6-31G(d) level of theory and density functional together with the SMD continuum solvation model (Marenich et al., 2009), as implemented in Gaussian 09 (Frisch et al.); this approach has been shown to provide good agreement with experimental values (Struebing et al., 2013). To compute the concentrations of the reactants in the reaction mixture (cf Equations (35) and (36)), the liquid molar volume of the solvent is calculated by GC (Constantinou et al., 1995). However, pyridine is too small to be treated accurately by group contribution techniques, and the experimental value $V_{m,C_1} = 0.08048\,\mathrm{L \cdot mol^{-1}}$ is used (Liessmann et al., 1995).

The melting temperature T_{m,C_2} and the enthalpy of fusion $\Delta H_{m,C_2}$ of phenacyl bromide,

required to calculate the solubility of phenacyl bromide via Equation (39), are obtained using the third-order group contribution method proposed by Marrero and Gani (2001). No experimental data are available for the enthalpy of fusion. The predicted normal melting temperature of $T_{m,C_2} = 320.53\,\mathrm{K}$ is in a good agreement with experimental measurements, which vary between 320.15 K and 324.15 K (Chemspider). The term on the right-hand side of Equation (39) depends only on the properties of the solute and on the reaction temperature. Here, this term is constant since an isothermal reaction is assumed. To assess the suitability of the solubility model, the solubility of phenacyl bromide in water was computed and found to be in good agreement with experimental data (Römpp encyclopedia online - Version 3.5.).

To define the design problem fully, we specify all parameters relevant to the definition of the design space and the constraints. The atom groups used are listed in Supplementary Information Table 1. Groups marked with an asterisk are excluded from the design space due to reactivity concerns, either due to the presence of double bonds (e.g., CH=CH) or to specific affinity for one of the reactants (e.g., Br or CH_2NH). A total of 33 atom groups is thus considered, including 3 molecular groups (acetonitrile, chloroform and nitromethane).

Given that molecular groups are considered, the minimum number of groups must be given as $n_{\mathrm{G,min}} = 1$. The maximum number of group is limited to $n_{\mathrm{G,max}} = 7$. This reduces the complexity of the solvents that may be designed and eliminates the bicyclic structure. Based on this choice, we set $K = 3$, so that the number of groups of any given type i is constrained to a maximum of 7, i.e., $n^U_{\mathrm{G},a} = 7$, $n_{\mathrm{Gb,max}} = 0$, and $n_{\mathrm{Gm,max}} = 1$.

In principle, the melting point maximum and boiling point minimum for the solvent can be set to ensure a sufficiently large liquid range around 298 K by choosing values such that $T_{m,max} < 298$ K and $T_{b,min} > 298$ K. However, to allow for uncertainty in the predictions of the GC methods and avoid excluding potentially interesting solvents, the melting point upper bound is set to 317 K and the boiling point lower bound is set to 292 K. Therefore, the dimensionless equivalent values used in Equations (72) and (73) are $T_{m,max} = 8.6$ and $T_{b,min} = 3.7$, respectively.

4.2. Application of the QM-CAMD algorithm

To investigate the proposed QM-CAMD algorithm and formulation, several case studies are considered in which the initial set of solvents, the design space and the constraints on the

mole fractions of the reactants, $x_{\max,\text{eq}}$ and $x^*_{C_1}$, are varied. The application of the algorithm illustrated in Figure 1 is first described in detail by presenting the first iteration of one case study step by step. The MINLP in Step 4 is solved using the BARON 15.9.22 solver (Tawarmalani and Sahinidis, 2005) in GAMS 24.6.1 (GAMS Development Corporation, 2014).

4.2.1. Case A/Base case

First, the solvent design problem is specified for the base case, case A. The design space is restricted by removing the building group OH due to inaccuracies in QM rate constant calculations of alcohols (Struebing et al., 2013). Furthermore, the upper bound on the solubility $x_{\max,\text{eq}}$ and the mole fraction of pyridine $x^*_{C_1}$ are set as follows:

$$x_{\max,\text{eq}} = x^*_{C_1} = 0.25. \tag{75}$$

The choice of 0.25 ensures that the solvent is always in excess and the impact of this value on the solvents designed can readily be assessed by carrying out a sensitivity analysis.

An initial set of six common hydroxyl-free solvents is used, covering a large range of dielectric constant values. The solvents and their respective dielectric constants are: toluene ($\varepsilon = 2.38$), chlorobenzene ($\varepsilon = 5.70$), ethyl acetate ($\varepsilon = 5.99$), tetrahydrofuran ($\varepsilon = 7.43$), acetone ($\varepsilon = 20.49$) and acetonitrile ($\varepsilon = 35.69$) (Winget et al., 2010). To limit the number of QM calculations the minimum number of solvents (six) is used.

In Step 2, the pure solvent properties (refractive index, Abraham's hydrogen bond acidity, Abraham's hydrogen bacicity, macroscopic surface tension, dielectric constant, aromaticity and electronegative halogenicity) required to determine the QM reaction rate constant, k_j^{TST}, in each solvent j are computed using the GC techniques presented in Section 3.3 and in the Supplementary Information. The energy and geometry of each reactant and of the transition-state are then determined by considering the isolated molecule or activated complex surrounded by the solvent of interest, as described via the SMD model. Furthermore, additional solvent properties needed for the surrogate model (Abraham's dipolarity/polarizability parameter, the polarizability correction parameter and the cohesive energy density) are computed using the GC techniques presented in Section 3.3. The predicted rate constants k_j^{TST}, for all j, and the solvent properties needed for the surrogate model are summarized in an illustration of the first steps of

the algorithm in Table 1. It can be seen that for the chosen initial set of solvents, the reaction rate constants vary over two orders of magnitude.

In Step 3, the parameters of the solvatochromic equation are determined by a linear regression to the data computed for the initial solvent set. The resulting expression is

$$\log k^{\mathrm{CAMD}} = -18.82 - 87.31A + 6.98B + 6.46S + 1.80\delta + 10.33\frac{\delta_H^2}{100}. \tag{76}$$

Next, Equation (76) is incorporated into the CAMD problem which is solved in Step 4. The resulting optimal solvent candidate contains $1 \times \mathrm{CH_2NO_2}$ and $1 \times \mathrm{I}$ structural groups, i.e. iodonitromethane, with a predicted rate constant of $k^{\mathrm{CAMD}} = 2.013 \cdot 10^{10}\,\mathrm{L \cdot mol^{-1} \cdot s^{-1}}$.

In Step 5, the convergence test is applied and since iodonitromethane has not been used in the regression of the surrogate model, the solvent is added to the solvent set to update the surrogate model. A second iteration is initiated and the algorithm returns to Step 2 as illustrated in Table 1. Here, the rate constant in iodonitromethane as predicted by the QM model is $k^{\mathrm{TST}} = 3.700 \cdot 10^{-3}\,\mathrm{L \cdot mol^{-1} \cdot s^{-1}}$, 13 orders of magnitude lower than the corresponding rate constant obtained with the solvatochromic model, Equation (7). This large discrepancy is due to the very low statistical significance of the surrogate model, since its parameters were regressed based on only six data points.

As the algorithm proceeds until the convergence criterion is satisfied, the surrogate model is updated after each iteration and becomes more and more reliable. In the last iteration, the best solvent is determined by the CAMD approach to consist of $1 \times \mathrm{CH_3NO_2}$, i.e. nitromethane, with a CAMD rate constant of $k^{\mathrm{CAMD}} = 3.192 \cdot 10^{-3}\,\mathrm{L \cdot mol^{-1} \cdot s^{-1}}$. The QM-derived rate constant in nitromethane is $k^{\mathrm{TST}}_{CH_3NO_2} = 3.021 \cdot 10^{-3}\,\mathrm{L \cdot mol^{-1} \cdot s^{-1}}$; the error between k^{CAMD} and k^{TST} is of only 5.5 %, which illustrates the increased reliability of the surrogate model.

The solvent designed at each iteration and its performance metrics are presented in Table 2. Furthermore, the ranked list of the five best solvents found, which is obtained by using integer cuts at the final iteration, is also shown in Table 2. Since the MINLP model is solved to global optimality, these are the five best solvents within the entire design space *b*ased on the surrogate model at this final iteration. This may not correspond to the five best solvents according to the QM model due to differences in the two models of reaction kinetics. The final optimal solvent for the base case is nitromethane, although the QM-derived rate constant of iodonitromethane

Table 1: Illustration of the first steps in the QM-CAMD algorithm of the base case.

Step 1: Define solvent problem and initial set

↓

Iteration 1; Step 2: Calculate k^{TST} and solvent properties

Solvents	k^{TST} in $\frac{\mathrm{L}}{\mathrm{mol\cdot s}}$	A	B	S	δ	$\frac{\delta_H^2}{100}$ in $\frac{\mathrm{cal}}{\mathrm{cm}^3}$
Toluene	1.632E-05	0.000	0.151	0.516	1.0	0.760
Chlorobenzene	2.822E-04	0.000	0.000	0.631	1.0	0.910
Ethyl acetate	3.473E-04	0.000	0.475	0.574	0.0	0.808
Tetrahydrofuran	5.810E-04	0.000	0.480	0.520	0.0	0.860
Acetonitrile	2.342E-03	0.070	0.320	0.900	0.0	1.381
Acetone	1.544E-03	0.000	0.491	0.689	0.0	0.788

↓

Iteration 1; Step 3: Regress solvatochromic equation to data

$$\log k^{\mathrm{CAMD}} = -18.82 - 87.31A + 6.98B + 6.46S + 1.80\delta + 10.33\frac{\delta_H^2}{100}$$

↓

Iteration 1; Step 4: Identify optimal solvent candidate

Solvent: $1 \times \mathrm{CH_2NO_2}$, $1 \times \mathrm{I}$

↓

Iteration 1; Step 5: Convergence test → Fail

↓

Iteration 2; Step 2: Computing solvent properties, k^{TST} and updating solvent set

Solvent	n_{D}^2	A	B	γ in $\frac{\mathrm{cal}}{\mathrm{mol\cdot \AA^2}}$	ε	ϕ	ψ
$\mathrm{CH_2INO_2}$	2.162	0.000	0.321	74.122	22.620	0.000	0.000

→ $k^{\mathrm{TST}} = 3.700\text{E-}03\ \frac{\mathrm{L}}{\mathrm{mol\cdot s}}$

↓

Solvents	k^{TST} in $\frac{\mathrm{L}}{\mathrm{mol\cdot s}}$	A	B	S	δ	$\frac{\delta_H^2}{100}$ in $\frac{\mathrm{cal}}{\mathrm{cm}^3}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Acetone	1.544E-03	0.000	0.491	0.689	0.0	0.788
$\mathrm{CH_2INO_2}$	3.700E-03	0.000	0.321	1.278	0.5	1.718

⋮

is higher than the QM rate constant of nitromethane. This cannot be attributed to a lower solubility of phenacyl bromide in iodonitromethane as the mole fraction of phenacyl bromide at the optimal solution is below the thermodynamic solubility limit, i.e., $x_{C_2} = x_{\max,\text{eq}}$. In fact, phenacyl bromide is highly soluble in all the solvents tested by the algorithm. Nitromethane leads to a higher reaction rate due to the higher reactant concentrations, c_{C_1} and c_{C_2}, that can be achieved in nitromethane for given mole fractions, as result of the smaller liquid molar volume of nitromethane relative to iodonitromethane. This reflects the fact that a reaction mixture with a smaller liquid molar volume either requires a smaller reactor to yield the same productivity or results in higher productivity in an existing reactor, than a reaction mixture with a higher molar volume. The list of 5 solvents obtained with the final surrogate model (iteration 5 in Table 2, solvents 5.1 to 5.5) is ranked in order of decreasing reaction rate as predicted by the surrogate model. It can be seen that the reported reaction rate constants, computed with the QM model, do not follow this order. For example, the QM-derived rate constant in nitroethane is lower than the reaction rate in iodonitromethane and acetonitrile, although nitroethane is the second best solvent identified by solving the CAMD problem. This uncertainty in the rankings is caused by discrepancies between the two models.

This case study can be used to investigate the computational cost of the QM-CAMD approach and the relative cost of different steps. The CPU time for the solution of one MINLP is 2.4 CPU hours on average on a single core of a dual 12 core Intel Xeon, 2.70GHz. The total CPU time dedicated to Step 4 and Step 6 (generation of the final list of solvents using integer cuts) for this case study is 12 CPU hours. On the other hand, the CPU time required for one calculation of k^{TST} is typically about 30 hours on a twin quad core Xeon 2.8Ghz. Given that 11 rate constant calculations are carried out during the course of the solution, QM calculations require approximately 330 CPU hours in total. Although the k^{TST} calculations are carried out in parallel across 8 cores, it is clear that they dominate the CPU and real time costs. The direct integration of these calculations in the CAMD problem formulation, which would in addition require the calculation of the gradients of k^{TST} with respect to the solvent properties, would thus be prohibitively expensive. Instead, the proposed QM-CAMD approach makes the inclusion of kinetics in solvent design computationally tractable through the use of a surrogate model.

Table 2: Summary of results for case A/base case. The notation $i.1$, $i.2$, etc. denotes that integer cuts have been used at iteration i.

Iter.	Solvent	x_{C_1}	x_{C_2}	c_{C_1} $\frac{\text{mol}}{\text{L}}$	c_{C_2} $\frac{\text{mol}}{\text{L}}$	k^{TST} $\frac{\text{L}}{\text{mol·s}}$	r^{TST} $\frac{\text{mol}}{\text{L·s}}$
1	iodonitromethane	0.25	0.25	2.778	2.778	3.700E-03	2.855E-02
2	nitroethane	0.25	0.25	2.806	2.806	2.172E-03	1.710E-02
3	propanenitrile	0.25	0.25	2.815	2.815	1.731E-03	1.372E-02
4	nitromethane	0.25	0.25	3.127	3.127	3.021E-03	2.954E-02
5.1	nitromethane	0.25	0.25	3.127	3.127	3.021E-03	2.954E-02
5.2	nitroethane	0.25	0.25	2.806	2.806	2.172E-03	1.710E-02
5.3	iodonitromethane	0.25	0.25	2.778	2.778	3.700E-03	2.855E-02
5.4	acetonitrile	0.25	0.25	3.155	3.155	2.346E-03	2.335E-02
5.5	1-nitropropane	0.25	0.25	2.569	2.569	1.927E-03	1.271E-02

Case B/Inclusion of OH groups. In the second instance of the design problem, case B, the constraint that the initial solvent set and the design space should not contain OH groups is lifted to investigate its impact on the outcome of the design. The initial set now contains an alcohol, ethanol ($\varepsilon = 24.55$ (Ritzoulis and Fidantsi, 2000)), instead of acetone. All other aspects of the problem are identical to Case A and the results are presented in Table 3 in the same manner as Case A to facilitate their analysis. It is found that the solubility of phenacyl bromide is not a limiting factor in solvent in this particular case, since in all solvents considered the thermodynamic solubility is again greater than the upper bound of mole fraction for this compound, $x_{\text{max,eq}} = 0.25$. Further, the impact of the density on the reaction rate is negligible in this case due to the large differences between the QM-derived rate constants in the various solvents. The optimal solvent candidate is found to comprise $1 \times CH_2NO_2$ and $1 \times$ OH building groups, i.e., nitromethanol. It is predicted to exhibit a significantly higher reaction rate compared to other solvents. However, when checking for chemical stability and reactivity in Step 7 of the QM-CAMD algorithm, two of the final solvents, nitromethanol and hypoiodous, acid are found to be unstable molecules. Consequently, the alcohol 2-nitroethanol is found to be the best solvent that fulfils the requirement of chemical stability. Although alcohols appear to be very promising solvents, the limited accuracy of the QM model used suggests further investigaton of their suitability is needed. Nitromethane is once again identified as a viable alternative, as the third possible solvent in Case B.

Table 3: Summary of results for case B/no restriction to the initial set of solvents. The notation $i.1$, $i.2$, etc. denotes that integer cuts have been used at iteration i.

Iter.	Solvent	x_{C_1}	x_{C_2}	c_{C_1} $\frac{\text{mol}}{\text{L}}$	c_{C_2} $\frac{\text{mol}}{\text{L}}$	k^{TST} $\frac{\text{L}}{\text{mol·s}}$	r^{TST} $\frac{\text{mol}}{\text{L·s}}$
1	isohexane	0.25	0.25	2.113	2.113	1.157E-04	5.163E-04
2	nitromethanol*	0.25	0.25	3.173	3.173	9.009E-03	9.070E-02
3	acetic acid	0.25	0.25	2.998	2.998	1.238E-04	1.112E-03
4.1	nitromethanol*	0.25	0.25	3.173	3.173	9.009E-03	9.070E-02
4.2	2-nitroethanol	0.25	0.25	2.874	2.874	7.061E-03	5.831E-02
4.3	hypoiodous acid*	0.25	0.25	3.295	3.295	5.849E-03	6.350E-02
4.4	nitromethane	0.25	0.25	3.127	3.127	3.021E-03	2.954E-02
4.5	3-nitropropan-1-ol	0.25	0.25	2.626	2.626	5.864E-03	4.043E-02

Case C/Solubility limit. In Cases A and B, the solubility of the solid reactant, phenacyl bromide, in the designed solvents was not found to be a limiting factor in maximizing the reaction rate. Case C is formulated to study the impact of solubility limitations, by allowing a greater proportion of reactants to be present in the mixture (up to 0.9 mole fraction in total), making the predictions of the kinetic models less reliable. The objective of this case study is purely methodological: it illustrates how the proposed mathematical formulation allows the solubility limit to be taken into account; however, for this specific reaction, this means that the solvent is no longer in excess and therefore that the results (the specific solvents designed) should be treated with caution. The only differences from the base case are that $x_{\text{max,eq}}$ is set to 0.8 and the mole fraction of pyridine $x^*_{C_1}$ to 0.1. The same design space is used as in Case A (the OH building group is not permitted to occur in the designed solvent), and the surrogate model is regressed to the same initial set of solvents. The results are presented in Table 4 and include the mole fraction of phenacyl bromide present in the mixture, as calculated via Equation (45). For all solvents tested by the algorithm, the solubility of phenacyl bromide at 298 K is less than 0.8 so that the solubility constraint is active, thereby demonstrating the effectiveness of the proposed formulation. The mole fraction is predicted to be approximately 0.57 in most solvents, expect in nitromethane which yields a higher mole fraction, $x_{C_2} = 0.62$. As a result, a higher reaction rate is obtained in nitromethane than in 1,1-dichloro-3-nitropropane, although 1,1-dichloro-3-nitropropane is found to have a higher rate constant than nitromethane. Overall, halogenated and nitro compounds appear to give good performance, but we note that it is not

known whether excess phenacyl bromide has an impact on the reaction rate constant.

Table 4: Summary of results for case C/base case with $x_{eq,max} = 0.8$. The notation $i.1$, $i.2$, etc. denotes that integer cuts have been used at iteration i.

Iter.	Solvent	x_{C_1}	x_{C_2}	c_{C_1} $\frac{\text{mol}}{\text{L}}$	c_{C_2} $\frac{\text{mol}}{\text{L}}$	k^{TST} $\frac{\text{L}}{\text{mol·s}}$	r^{TST} $\frac{\text{mol}}{\text{L·s}}$
1	iodonitromethane	0.1	0.573	0.928	5.319	3.700E-03	1.826E-02
2	2-methyl-1-nitropropane	0.1	0.545	0.852	4.647	1.659E-03	6.572E-03
3	nitroethane	0.1	0.568	0.936	5.311	2.172E-03	1.080E-02
4	1,1-dichloro-2-nitroethane	0.1	0.568	0.879	4.990	4.235E-03	1.858E-02
5.1	1,1-dichloro-2-nitroethane	0.1	0.568	0.879	4.990	4.235E-03	1.858E-02
5.2	1,1-dichloro-3-nitropropane	0.1	0.565	0.839	4.743	3.754E-03	1.494E-02
5.3	nitromethane	0.1	0.620	0.954	5.910	3.021E-03	1.703E-02
5.4	nitroethane	0.1	0.568	0.936	5.311	2.172E-03	1.080E-02
5.5	iodonitromethane	0.1	0.573	0.928	5.319	3.700E-03	1.826E-02

A comparison of the results of the three cases indicates that nitromethane is a promising solvent for this reaction: it is systematically selected amongst the top five candidates. Its performance has been verified experimentally (Struebing et al., 2013). There is consistency in the results of the three case studies, with nitrogen-containing compounds appearing repeatedly, together with solvents containing iodine or chlorine. The results indicate that few iterations are required to converge to a set of high-performance solvents, so that only about 10 calculations of the rate constant with the QM and SMD model are needed to investigate the much larger design space. Considering the much higher cost of QM calculations relative to CAMD solution, this represents a significant time savings.

5. Concluding remarks

The QM-CAMD methodology for the design of optimal solvents for reactions (Struebing et al., 2013) has been extended to take into account the impact of solid reactant solubility and solvent density on solvent choice. The formulation of an MINLP that embeds these considerations has been presented in detail. The solubility of solid reactants is modelled by making appropriate assumptions on the solid phase and modelling the non-ideality of the liquid phase via the UNIFAC group contribution method. The concentrations of the various compounds in the reaction mixture are obtained by assuming a negligible volume of mixing. Overall, this approach allows the use of the reaction rate as an objective, instead of the reaction rate constant. All

required solvent properties are predicted by GC techniques, while QM calculations embedding a continuum solvation model are used to compute rate constants in various solvents, making the approach independent of any experimental data and widely applicable. A surrogate model for the QM calculations, in the form of the linear solvatochromic equation, is embedded in the MINLP to ensure computational tractability. The application of the proposed methodology to a S_N2-Menschutkin reaction, a classic reaction for the study of solvent effects, shows that it leads to the identification, from a large design space, of promising solvents that enhance the reaction rate. In spite of the simplicity of the surrogate model and of the uncertainty inherent in the group contribution techniques used, the results are found to be in good agreement with experimental data (Struebing et al., 2013) and to provide a first-principles guide to solvent design. Furthermore, the case study demonstrates the strong interplay between kinetics, other solvent properties and thermodynamic factors that affect process performance, such as solid reactant solubility and solvent density.

The proposed QM-CAMD framework offers the possibility to add further constraints that are also important in solvent selection. For instance, in the case of competing reactions, selectivity can be accounted for by computing the rate of each reaction in different solvents, with a corresponding surrogate model per reaction. The solubility of liquid reactants should also be considered, and this could be achieved in the first instance by embedding miscibility constraints for each reactant with the solvent, as proposed by Gani et al. (1991). Broader process design aspects can also be readily integrated within the proposed framework, including economic and environmental criteria. Despite the use of global optimisation algorithm to solve the nonconvex MINLP, the approach as presented may not always identify globally optimal solvents due to the discrepancies between the surrogate and detailed models of reaction kinetics. Although the surrogate model has the benefit of being linear and simple to derive, its accuracy is limited because of the small data set used in its regression and because of the inherently nonlinear relationship between rate constant and solvent molecular structure. The diversity of the high-performance solvents identified by the QM-CAMD approach could thus be further enhanced by adopting a more accurate surrogate model, offering an interesting avenue for further development of the methodology. As it stands, the approach provides a valuable framework to link reactor design and solvent design prior to any experimental investigations.

Acknowledgments

The authors gratefully acknowledge financial support from the Engineering and Physical Sciences Research Council (EPSRC) of the UK (grants EP/E016340, EP/J014958/1 and EP/J003840/1) as well as access to computational resources and support from the High Performance Computing Cluster at Imperial College London.

Data statement

Additional data underlying this article can be accessed in the Supplementary Information file and at http://dx.doi.org/10.5281/zenodo.51697, and used under the Creative Commons Attribution licence.

References

Abraham, M.H., 1993. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. Chemical Society Reviews 22, 73–83.

Abraham, M.H., Doherty, R.M., Kamlet, M.J., Harris, J.M., Taft, R.W., 1987a. 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. Journal of the Chemical Society, Perkin Transactions 2 , 913–920.

Abraham, M.H., Doherty, R.M., Kamlet, M.J., Harris, J.M., Taft, R.W., 1987b. Linear solvation energy relationships. Part 38. An analysis of the use of solvent parameters in the correlation of rate constants, with special reference to the solvolysis of t-butyl chloride. Journal of the Chemical Society, Perkin Transactions 2 , 1097–1101.

Abraham, M.H., Grellier, P.L., Nasehzadeh, A., Walker, R.A.C., 1988. Substitution at saturated carbon. Part 26. A complete analysis of solvent effects on initial states and transition states for the solvolysis of the t-butyl halides in terms of G, H, and S using the unified method. Journal of the Chemical Society, Perkin Transactions 2 , 1717–1724.

Abraham, M.H., Taft, R.W., Kamlet, M.J., 1981. Linear solvation energy relationships. 15. Heterolytic decomposition of the tert-butyl halides. The Journal of Organic Chemistry 46, 3053–3056.

Achenie, L., Venkatasubramanian, V., Gani, R., 2002. Computer Aided Molecular Design: Theory and Practice. Elsevier Science, Amsterdam.

Adjiman, C.S., Galindo, A., Jackson, G., 2014. Molecules matter: The expanding envelope of process design, in: Eden, M., Siirola, J.D., Towler, G.P. (Eds.), 8th International Conference on Foundations of Computer-Aided Process Design – FOCAPD, Elsevier B.V, Washington, USA. pp. 55–64.

Amovilli, C., Mennucci, B., Floris, F.M., 1998. MCSCF study of the S_N2 Menshutkin reaction in aqueous solution within the polarizable continuum model. The Journal of Physical Chemistry B 102, 3023–3028.

Barnard, P.W.C., Smith, B.V., 1981. The Menschutkin reaction: A group experiment in a kinetic study. Journal of Chemical Education 58, 282–285.

Carlson, R., Lundstedt, T., Albano, C., 1985. Screening of suitable solvents in organic synthesis. Strategies for solvent selection. Acta Chemica Scandinavica B39, 79–91.

Castejon, H., Wiberg, K.B., 1999. Solvent effects on methyl transfer reactions. 1. the Menshutkin reaction. Journal of the American Chemical Society 121, 2139–2146.

Cativiela, C., Garcia, J.I., Gil, J., Martinez, R.M., Mayoral, J.A., Salvatella, L., Urieta, J.S., Mainar, A.M., Abraham, M.H., 1997. Solvent effects on Diels-Alder reactions. The use of aqueous mixtures of fluorinated alcohols and the study of reactions of acrylonitrile. Journal of the Chemical Society, Perkin Transactions 2 , 653–660.

Chemspider, . Phenacyl bromide. Available at: http://www.chemspider.com. Accessed March 13, 2014.

Constantinou, L., Gani, R., O'Connell, J.P., 1995. Estimation of the acentric factor and the liquid molar volume at 298 K using a new group contribution method. Fluid Phase Equilibria 103, 11–22.

Dufal, S., Papaioannou, V., Sadeqzadeh, M., Pogiatzis, T., Chremos, A., Adjiman, C.S., Jackson, G., Galindo, A., 2014. Prediction of thermodynamic properties and phase behavior of fluids and mixtures with the SAFT-γ Mie group-contribution equation of state. Journal of Chemical & Engineering Data 59, 3272–3288.

Elgue, S., Prat, L., Cognet, P., Cabassud, M., Le Lann, J.M., Cézerac, J., 2004. Influence of solvent choice on the optimisation of a reaction-separation operation: application to a Beckmann rearrangement reaction. Separation and Purification Technology 34, 273–281.

Evans, M.G., Polanyi, M., 1935. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. Transactions of the Faraday Society 31, 875–894.

Eyring, H., 1935. The activated complex in chemical reactions. The Journal of Chemical Physics 3, 107–115.

Folić, M., Adjiman, C.S., Pistikopoulos, E.N., 2004. The design of solvents for optimal reaction rates. Computer Aided Chemical Engineering 18, 175–180.

Folić, M., Adjiman, C.S., Pistikopoulos, E.N., 2007. Design of solvents for optimal reaction rate constants. AIChE Journal 53, 1240–1256.

Folić, M., Adjiman, C.S., Pistikopoulos, E.N., 2008. Computer-aided solvent design for reactions: Maximizing product formation. Industrial & Engineering Chemistry Research 47, 5190–5202.

Forster, W., Laird, R.M., 1982. The mechanism of alkylation reactions. Part 1. The effect of substituents on the reaction of phenacyl bromide with pyridine in methanol. Journal of the Chemical Society, Perkin Transactions 2 , 135–138.

Fredenslund, A., Jones, R.L., Prausnitz, J.M., 1975. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE Journal 21, 1086–1099.

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G.A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H.P., Izmaylov, A.F., Bloino, J., Zheng, G., Sonnenberg, J.L., Hada, M., Ehara,

M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, Jr., J.A., Peralta, J.E., Ogliaro, F., Bearpark, M., Heyd, J.J., Brothers, E., Kudin, K.N., Staroverov, V.N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J.C., Iyengar, S.S., Tomasi, J., Cossi, M., Rega, N., Millam, J.M., Klene, M., Knox, J.E., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A.J., Cammi, R., Pomelli, C., Ochterski, J.W., Martin, R.L., Morokuma, K., Zakrzewski, V.G., Voth, G.A., Salvador, P., Dannenberg, J.J., Dapprich, S., Daniels, A.D., Farkas, ., Foresman, J.B., Ortiz, J.V., Cioslowski, J., Fox, D.J., . Gaussian 09 Revision D.01. Gaussian Inc. Wallingford CT; 2009.

GAMS Development Corporation, 2014. General Algebraic Modeling System (GAMS). URL: `http://www.gams.com`.

Ganase, Z., 2015. An Experimental Study on the Effects of Solvents on the Rate and Selectivity of Organic Reactions. Ph.D. thesis. Imperial College London.

Gani, R., Gómez, P.A., Folić, M., Jiménez-González, C., Constable, D.J.C., 2008. Solvents in organic synthesis: Replacement and multi-step reaction systems. Computers & Chemical Engineering 32, 2420–2444.

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

Gani, R., Nielsen, B., Fredenslund, A., 1991. A group contribution approach to computer-aided molecular design. AIChE Journal 37, 1318Ű1332.

Giovanoglou, A., Barlatier, J., Adjiman, C.S., Pistikopoulos, E.N., Cordiner, J.L., 2003. Optimal solvent design for batch separation based on economic performance. AIChE Journal 49, 3095–3109.

Halvorsen, A., Songstad, J., 1978. The reactivity of 2-bromo-1-phenylethanone (phenacyl bromide) toward nucleophilic species. Journal of the Chemical Society, Chemical Communications , 327–328.

Hwang, J.U., Chung, J.J., Yoh, S.D., Jee, J.G., 1983. Kinetics for the reaction of phenacyl bromide with pyridine in acetone under high pressure. Bulletion of the Korean Chemical Society 4, 237–240.

Jiménez-González, C., Poechlauer, P., Broxterman, Q.B., Yang, B., am Ende, D., Baird, J., Bertsch, C., Hannah, R.E., Dell'Orco, P., Noorman, H., Yee, S., Reintjens, R., Wells, A., Massonneau, V., Manley, J., 2011. Key green engineering research areas for sustainable manufacturing: A perspective from pharmaceutical and fine chemicals manufacturers. Organic Process Research & Development 15, 900–911.

Kamlet, M.J., Abboud, J.L.M., Taft, R.W., 1977. The solvatochromic comparison method. 6. The π^* scale of solvent polarities. Journal of the American Chemical Society 99, 6027–6038.

Klamt, A., Eckert, F., Arlt, W., 2010. COSMO-RS: An alternative to simulation for calculating thermodynamic properties of liquid mixtures. Annual Review of Chemical and Biomolecular Engineering 1, 101–122.

Kolář, P., Shen, J., Tsuboi, A., Ishikawa, T., 2002. Solvent selection for pharmaceuticals. Fluid Phase Equilibria 194-197, 771–782.

Lee, S.B., 1996. A new linear solvation energy relationship for the solubility of liquids in water. Journal of Pharmaceutical Sciences 85, 348–350.

Lide, D.R., 2011. CRC Handbook of Chemistry and Physics. Taylor and Francis Group, LLC., Washington, D.C. Available at: http://www.hbcpnetbase.com/. Accessed: July 28, 2011.

Liessmann, G., Schmidt, W., Reiffarth, S., 1995. Recommended thermophysical data. Data compilation of the Saechsische Olefinwerke Boehlen Germany .

Lymperiadis, A., Adjiman, C.S., Galindo, A., Jackson, G., 2007. A group contribution method for associating chain molecules based on the statistical associating fluid theory (SAFT-γ). The Journal of Chemical Physics 127, 234903.

Lymperiadis, A., Adjiman, C.S., Jackson, G., Galindo, A., 2008. A generalisation of the SAFT-γ group contribution method for groups comprising multiple spherical segments. Fluid Phase Equilibria 274, 85–104.

Maki, B.E., Patterson, E.V., Cramer, C.J., Scheidt, K.A., 2009. Impact of solvent polarity on N-heterocyclic carbene-catalyzed β-protonations of homoenolate equivalents. Organic Letters 11, 3942–3945.

Maranas, C.D., 1996. Optimal computer-aided molecular design: A polymer design case study. Industrial & Engineering Chemistry Research 35, 3403–3414.

Marenich, A.V., Cramer, C.J., Truhlar, D.G., 2009. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. The Journal of Physical Chemistry B 113, 6378–6396.

Marrero, J., Gani, R., 2001. Group-contribution based estimation of pure component properties. Fluid Phase Equilibria 183-184, 183–208.

McBride, K., Gaide, T., Vorholt, A., Behr, A., Sundmacher, K., 2016. Thermomorphic solvent selection for homogeneous catalyst recovery based on COSMO-RS. Chemical Engineering and Processing: Process Intensification 99, 97–106.

Menschutkin, N., 1890a. Beiträge zur Kenntnis der Affinitätskoeffizienten der Alkylhaloide und der organischen Amine. Zeitschrift für Physikalische Chemie 5, 589–600.

Menschutkin, N., 1890b. Über die Affinitätskoeffizienten der Alkylhaloide und der Amine. Zeitschrift für Physikalische Chemie 6, 41–57.

Ng, L.Y., Chong, F.K., Chemmangattuvalappil, N.G., 2015. Challenges and opportunities in computer-aided molecular design. Computers & Chemical Engineering 81, 115–129.

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

Papaioannou, V., Lafitte, T., Avendaño, C., Adjiman, C.S., Jackson, G., Müller, E.A., Galindo, A., 2014. Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments. The Journal of Chemical Physics 140, 054107.

Pearson, R.G., Langer, S.H., Williams, F.V., McGuire, W.J., 1952. Mechanism of the reaction of α-haloketones with weakly basic nucleophilic reagents. Journal of the American Chemical Society 74, 5130–5132.

Poling, B., Prausnitz, J., O'Connell, J.P., 2000. The Properties of Gases and Liquids. McGraw-Hill, New York.

Prausnitz, J.M., Lichtenthaler, R.N., de Azevedo, E.G., 1999. Molecular Thermodynamics of Fluid-Phase Equilibria. Prentice-Hall, New Jersey.

Ràfols, C., Rosés, M., Bosch, E., 1995. Solute-solvent and solvent-solvent interactions in binary solvent mixtures. 1. A comparison of several preferential solvation models for describing E(t)-(30) polarity of dipolar hydrogen-bond acceptor-cosolvent mixtures. Journal of the Chemical Society, Perkin Transactions 2 , 1607–1615.

Reichardt, C., Welton, T., 2010. Solvents and Solvent Effects in Organic Chemistry. Wiley, Weinheim.

Ritzoulis, G., Fidantsi, A., 2000. Relative permittivities, refractive indices, and densities for the binary mixtures N,N-dimethylacetamide with methanol, ethanol, 1-butanol, and 2-propanol at 298.15 K. Journal of Chemical & Engineering Data 45, 207–209.

Römpp encyclopedia online - Version 3.5., 2009. URL: `https://www.thieme.de/en/thieme-connect/roempp-online-5005.htm`.

Sheldon, T.J., Adjiman, C.S., Cordiner, J.L., 2005. 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 231, 27–37.

Siougkrou, E., Galindo, A., Adjiman, C.S., 2014. On the optimal design of gas-expanded liquids based on process performance. Chemical Engineering Science 115. doi:`10.1016/j.ces.2013.12.025`.

Stanescu, I., Achenie, L.E.K., 2006. A theoretical study of solvent effects on Kolbe-Schmitt reaction kinetics. Chemical Engineering Science 61, 6199–6212.

Struebing, H., Ganase, Z., Karamertzanis, P.G., Siougkrou, E., Haycock, P., Piccione, P.M., Armstrong, A., Galindo, A., Adjiman, C.S., 2013. Computer-aided molecular design of solvents for accelerated reaction kinetics. Nature Chemistry 5, 952–957.

Tawarmalani, M., Sahinidis, N.V., 2005. A polyhedral branch-and-cut approach to global optimization. Mathematical Programming 103, 225–249.

Truong, T.N., Truong, T.T.T., Stefanovich, E.V., 1997. A general methodology for quantum modeling of free-energy profile of reactions in solution: An application to the Menshutkin NH_3+CH_3Cl reaction in water. The Journal of Chemical Physics 6, 1881–1889.

Whelan, T., 1994. Polymer Technology Dictionary. Springer, Amsterdam.

Wicaksono, D.S., Mhamdi, A., Marquardt, W., 2014. Computer-aided screening of solvents for optimal reaction rates. Chemical Engineering Science 115, 167–176.

Wigner, E., 1937. Calculation of the rate of elementary association reactions. The Journal of Chemical Physics 5, 720–725.

Winget, P., Dolney, D.M., Giesen, D.J., Cramer, C.J., Truhlar, D.G., 2010. Minnesota solvent descriptor database. Available at:
http://comp.chem.umn.edu/solvation/mnsddb.pdf. Accessed February 12, 2014.

Zhou, T., Lyu, Z., Qi, Z., Sundmacher, K., 2015a. Robust design of optimal solvents for chemical reactions – A combined experimental and computational strategy. Chemical Engineering Science 137, 613–625.

Zhou, T., McBride, K., Zhang, X., Qi, Z., Sundmacher, K., 2015b. Integrated solvent and process design exemplified for a Diels-Alder reaction. AIChE Journal 61, 147–158.

Zhou, T., Qi, Z., Sundmacher, K., 2014. Model-based method for the screening of solvents for chemical reactions. Chemical Engineering Science 115, 177–185.